

De novo assembly of transcriptomes from near-isogenic maize lines for novel defence gene discovery

Submitted in partial fulfilment of the degree:

MSc Bioinformatics

Centre for Bioinformatics and Computational Biology

Department of Biochemistry, Genetics and Microbiology

University of Pretoria

Tanya Welgemoed

13044542

11/20/2018

Supervisor: Prof Dave Berger

Co-supervisor: Dr Rian Pierneef

I, Tanya Welgemoed declare that the thesis/dissertation, which I hereby submit for the degree MSc Bioinformatics at the University of Pretoria, is my own work and has not previously been submitted by me for a degree at this or any other tertiary institution.

SIGNATURE: 

DATE: 04/09/2018.....

Acknowledgments

I would also like to acknowledge the following:

Felix Middleton from Pannar who did the backcross breeding to create the near isogenic lines used in this study.

Lieven Sterck, Yves van de Peer and the rest of the VIB group for their guidance and assistance with the differential expression analysis and gene annotation in Chapter 4.

The group members Molecular Plant-Pathogen Interactions (MPPI) and Cereal Foliar Pathogen Research (CFPR) for their support.

My supervisors, Prof Dave Berger and Dr Rian Pierneef, for their guidance and effort.

Azille Schulze who screened for MaYMV and sequenced the survey samples in Appendix B.

Gerhard Pietersen who gave much advice on the work with the virus in Appendix B.

The financial assistance of the National Research Foundation (NRF) towards this research is hereby acknowledged. Opinions expressed and conclusions arrived at, are those of the author and are not necessarily to be attributed to the NRF.

Preface

Maize is an economically important crop world-wide. This is especially true in South Africa where it is a staple crop and production contributes to food security. A major threat to maize production is a fungal foliar disease, grey leaf spot, that in South Africa is caused by *Cercospora zeina*. Controlling this disease to prevent the current yield losses is therefore important for food security. Ideally genetically resistant crops would be used, but there are only a few known defence genes against this disease. This study addressed this gap by developing a *de novo* assembly pipeline to identify maize genes involved in quantitative resistance against *C. zeina*. The study was done on three *C. zeina*-challenged and three control samples from a B73 maize line that had been introgressed with a quantitative trait locus for resistance against the disease. It was hypothesised that there would be a novel gene present in the introgressed QTL region that conferred the resistance against *C. zeina* infection.

The research presented in this dissertation is based on bioinformatics and experimental work conducted in the Molecular Plant-Pathogen Interactions (MPPI) group at the University of Pretoria, South Africa. The dissertation is presented in the form of six chapters.

Chapter 1 provides a review of next-generation RNA sequencing (RNA-seq) and its applications in gene discovery for plants. This includes an overview of the two strategies for assembling the RNA-seq reads (reference genome mapping and *de novo* assembly), their relative advantages and disadvantages, and the bioinformatics tools used to accomplish this. This chapter also discusses how the strategies can be combined in order to obtain greater results. The chapter concludes with an overview of the negative impact of *C. zeina* on maize production in South Africa.

Chapter 2 describes the bioinformatics pipeline used for the gene discovery. The study was done on a maize line that had a QTL for resistance against grey leaf spot introgressed. RNA-seq reads from the maize leaves was mapped against the *C. zeina* and maize reference genomes to exclude known genes. The unmapped reads were *de novo* assembled, protein coding sequences predicted and compared against the NCBI non-redundant protein database (nr) (NCB_Resource_Coordinators 2016). A list of novel plant genes that are potentially present in the QTL region was produced.

Chapter 3 discusses the differential expression analysis and annotation of the potential novel genes in the maize line with the introgressed QTL. The transcripts were tested for significant differential expression between the *C. zeina* challenged leaves and the control

leaves. The results were compared to the differential expression in the maize line without the QTL, which enabled the identification of genes whose regulation was affected by the QTL. The plant genes were annotated with the protein domains present and their associated gene ontology. A shortlist of novel plant genes that could be responsible for the grey leaf spot resistance QTL's effect was determined.

In **Chapter 4** the presence of a top candidate gene for causing the QTL effect was experimentally validated with PCR and Sanger sequencing. The gene was amplified from DNA obtained from the maize line with the QTL, the maize line from which the QTL originated and a fine-mapping population containing different regions of the QTL. It was not present in the maize line without the QTL thus showing that it is a novel gene located in the QTL region. An interesting result was that in all cases the non-QTL version of the gene was present as well.

Chapter 5 contains a general discussion of the impact and context of the findings of this project. An overview of the top candidate for causing the QTL effect was given and a potential origin for this allele was suggested. The role of the *Arabidopsis* ortholog in the anti-fungal defence response and the differences between the QTL and non-QTL version of the gene was outlined. Future research directions to functionally characterize the novel gene and to confirm whether or not it was created by a gene duplication was suggested.

Appendix A contains the full list of plant genes identified during the study. It also contains information on the additional sequences identified during the BLASTP comparison against nr.

Appendix B describes a new branch of the study created by an unexpected additional result. The bioinformatics pipeline in **Chapter 2** also identified coding sequences from Maize Yellow Mosaic virus as being present in the maize leaves. This is the first indication that this virus is present in South Africa and therefore this result was followed up on. Aligning the RNA-seq reads against the reference genome of the virus confirmed that the whole genome was present. This was repeated with other South African RNA-seq datasets that were available where the presence of the virus was confirmed in many of them. Stored RNA from the maize plants that were RNA-sequenced for the bioinformatics pipeline and RNA collected from a survey of diseased maize plants across South Africa was screened for the virus. A reverse-transcriptase PCR (RT-PCR) specific for Maize Yellow Mosaic virus and Sanger sequencing of some of the amplicons proved the presence of this virus in South Africa. In addition, the RT-PCR screening was done on RNA samples collected in Tanzania as well.

Appendix C details the important scripts and commands used during the bioinformatics analyses.

The research outputs generated from this body of this body of work to date are:

Welgemoed, T., (2017). *De novo* assembly of maize transcriptome data for defence gene discovery. Presented at SABPA 2018 symposium, DowDuPont Plant Sciences Symposium, Durban, South Africa. (Oral Presentation)

Welgemoed, T., (2017). *De novo* assembly of maize transcriptome data for defence gene discovery. Presented at GRI symposium 2017, Pretoria, South Africa. (Poster Presentation)

Read DA, Featherstone J, Rees DJG, Thompson GD, Roberts R, Flett BC, Mashingaidze K, Berger DK, Welgemoed, T., Pietersen G, Schulze SE, Kiula B, Kullaya A, Mbega E (2018) First report of maize yellow mosaic virus (MaYMV) on maize (*Zea mays*) in Tanzania. J Plant Pathol:1-1. (Published 15/08/2018, available at <https://rdcu.be/4Bjn>)

Welgemoed, T., Pierneef, R., Schulze, A., Pietersen, G., Berger D.K. Next generation sequencing reveals historical and widespread occurrence of Maize Yellow Mosaic Virus in South Africa. (Submitted to Scientific Reports)

Abstract

Maize is an economically important crop world-wide and is a staple crop in South Africa. A major threat to maize production is a fungal foliar disease, grey leaf spot, that is caused by *Cercospora zeina*. Controlling this disease is therefore important for food security, but there are only a few known defence genes. This study addressed this gap by developing a *de novo* assembly pipeline to identify maize genes involved in quantitative resistance against *C. zeina*. The pipeline used six RNA-seq datasets (three *C. zeina*-challenged and three control samples) from a B73 maize line that had been introgressed with a quantitative trait locus for resistance against the disease. The reads that did not map to the B73 or *C. zeina* genome were expected to contain novel defence transcripts. These were *de novo* assembled, and identified as plant, fungal or bacterial. The plant proteins were examined for potential defence functions and differential expression (DE) between *C. zeina*-challenged and control leaves to select candidate genes. Several potentially interesting genes were identified for further investigation. The most promising candidate is an allele of a G-type lectin S-receptor-like serine/threonine-protein kinase whose presence in the QTL region was validated with PCR.

List of Abbreviations

° C	degrees Celsius
B73-QTL	B73 with QTL10 introgressed
bp	base pair
BLAST	Basic Local Alignment Search Tool
BLASTp	search a protein database using a protein query
BSA	bulk segregant analysis
CMP	counts per million
DE	differentially expressed
eTMP	effective transcripts per million
E-value	expected value
FDR	false discovery rate
FPKM	fragments per kilobase of transcript per million mapped reads
GEO	Gene Expression Omnibus
GLS	Grey leaf spot
GO	gene ontology
K-mer	length of segment k
mRNA	messenger RNA
µl	microliter
MSIQ	Multiple RNA-seq Samples for accurate Isoform Quantification
ng	Nanograms
nr	NCBI non-redundant database

ORMAN	optimal resolution of multimapping ambiguity of RNA-seq reads
%	Percentage
PCR	polymerase chain reaction
PCR-RFLP	polymerase chain reaction restriction fragment length polymorphism
QTL	quantitative trait loci
RPKM	reads per kilobase of transcript of million mapped reads
RE	restriction enzyme
rRNA	ribosomal RNA
RNA-seq	RNA sequencing
SNP	single nucleotide polymorphism
SSR	simple sequence repeat
TMP	transcripts per million
UP	University of Pretoria

Table of Contents

Acknowledgments.....	iii
Preface	iv
Abstract.....	viii
List of Abbreviations.....	ix
Chapter 1: General Introduction	1
Literature review	1
Introduction	1
RNA-sequencing	2
Reference genome based mapping	3
<i>De novo</i> assembly	8
Comparison of <i>de novo</i> and mapping strategies	12
Combination of <i>de novo</i> and mapping strategies.....	14
Bioinformatics tools	16
Conclusion	17
Project Rationale	18
Aims and Objectives.....	20
Aims.....	20
Objectives (Fig. 1.5).....	20
Chapter 2: Candidate defence genes identified through designed bioinformatics pipeline...	22
Introduction	22
Methods	23
Plant Material.....	23
Data available for project	26
Bioinformatics Pipeline.....	26
Results	29
Discussion.....	43

Chapter 3: Differential expression analysis and annotation of candidate genes	45
Introduction	45
Methods	46
Plant Material	46
Data available for project	47
Differential expression analysis and annotation.....	47
Results	50
Discussion and Conclusion.....	55
Chapter 4: Experimental validation of selected candidate defence gene identified through the designed bioinformatics pipeline	57
Introduction	57
Methods	58
Plant material	58
PCR-RFLP and sequencing	59
Results	60
Discussion and conclusions.....	66
Chapter 5: General discussion and Conclusions	67
References	71
Appendix A	78
Appendix B: Next generation sequencing reveals historical and widespread occurrence of Maize Yellow Mosaic Virus in South Africa and Tanzania	144
Abstract	144
Introduction	144
Methods	146
Results	148
Discussion and Conclusion.....	161
Appendix C	172
Quality trimming the reads.....	173
Mapping to the <i>C. zeina</i> and B73 reference genomes	173
<i>De novo</i> assembly of the unmapped reads and total reads	174

Estimating the abundances of each transcript in the <i>C. zeina</i> challenged transcripts and the control transcripts	175
Rscript to test for differential expression between the <i>C. zeina</i> challenged transcripts and the control transcripts	175
BLAT command and AWK script to split differentially expressed transcripts	178

List of Figures

Chapter 1

Figure 1.1. An illustration of reference genome mapping of RNA-seq reads.....	4
Figure 1.2. An illustration of reads from different isoforms of an alternatively spliced transcript.....	5
Figure 1.3. Fragmentation creates overlapping fragments.....	9
Figure 1.4. An illustration of <i>de novo</i> assembly using a de Bruijn Chapter.....	2
Figure 1.5. Flow diagram of the methodology.....	25

Chapter 2

Figure 2.1. Breeding of B73-QTL by Felix Middleton (chromosome 10 shown).....	29
Figure 2.2. Diagram of the layout of B73 and B73QTL plants (represented as blocks) at Hildesheim farm in Greytown used as material for the RNA-seq.....	30
Figure 2.3. An example of the leaf material harvested for sequencing.....	30
Figure 2.4. An overview of the bioinformatics pipeline.....	32
Figure 2.5. The number of raw reads prior to trimming and the high-quality reads retained after trimming for each sample.....	35
Figure 2.6. A comparison of the lengths of the proteins produced from each parameter against the lengths of the matching B73 protein.....	36
Figure 2.7. Mapping and <i>de novo</i> assembly of the reads.....	47
Figure 2.8. Identification of the predicted proteins through BLASTp against nr.....	48

Chapter 3

Figure 3.1. Heat map showing differential expression between <i>C. zeina</i> challenged and control leaves.....	56
Figure 3.2. Examples of the log ₂ enrichment of gene ontology terms in the <i>C. zeina</i> challenged leaves.....	57
Figure 3.3. A biplot showing the variation caused by the top two principal components.....	57

Chapter 4

Figure 4.1. A diagram of the QTL10 region showing the four SSR markers used for marker assisted breeding.....85

Figure 4.2. PCR-RFLP to distinguish between the B73 and CML444 allele in different genotypes.....87

Figure 4.3. The expected products amplified with the Lectin primers in B73 and B73-QTL..88

Figure 4.4. A representation of the QTL10 region.....91

Figure 4.5. Control using the pdm327 plasmid showing restriction digests were not inhibited by ampliqon taq or faulty enzyme.....91

Figure 4.6. Portion the of sequencing reaction showing the expected sequence, sequence from the sequencing reaction and trace data for both B73 and CML444.....92

Appendix A

Figure A.1. The predicted proteins identified by BLASTP against the NCBI nr database that matched fungal sequences.....136

Figure A.2. An alignment of the *de novo* assembled transcripts against the B73 version of the G-type Lectin gene at the DNA level.....65

Figure A.3. An alignment of the *de novo* assembled transcripts against the B73 version of the G-type Lectin gene at the amino acid level.....66

Figure A.4. An alignment of the *de novo* assembled transcripts against the B73 version of the Thaumatin gene at the DNA level.....67

Figure A.5. An alignment of the *de novo* assembled transcripts against the B73 version of the thaumatin gene at the amino acid level.....67

Figure A.6. An alignment of the *de novo* assembled transcripts against the synthetic transcription factor and the B73 version of the transcription factor gene at the DNA level...68

Figure A.7. An alignment of the *de novo* assembled transcripts against the synthetic transcription factor at the amino acid level.....69

Figure A.8. An alignment of the *de novo* assembled transcripts against the B73 version of the chitinase gene at the DNA level.....70

Figure A.9. An alignment of the *de novo* assembled transcript against the B73 version of the chitinase gene at the amino acid level.....70

Figure A.10. An alignment of the de novo assembled transcript against the B73 version of the Apoptosis gene at the DNA level.....	73
Figure A.11. An alignment of the de novo assembled transcript against the B73 version of the Apoptosis gene at the amino acid level.....	73
Figure A.12. An alignment of the de novo assembled transcripts against the B73 Serine/Threonine kinase (Ser/Thr_kinase) at the DNA level.....	78
Figure A.13. An alignment of the de novo assembled transcripts against the B73 Serine/Threonine kinase (Ser/Thr_kinase) at the amino acid level.....	78
Figure A.14. An alignment of the de novo assembled transcripts against the B73 Leucine-rich repeat (LRR) at the DNA level.....	80
Figure A.15. An alignment of the de novo assembled transcripts against the B73 Leucine rich repeat (LRR) at the amino acid level.....	80
Appendix B	
Figure B.1. RT-PCR diagnostic showing the presence of MaYMV in RNA stored from RNA-seq field trial.....	186
Figure B.2. Actin RT-PCR to confirm RNA and cDNA integrity for samples that were RT-PCR negative for MaYMV.....	187
Figure B.3. A map showing the locations in South Africa where the positive RNA samples from the RNA-seq data and the positive survey samples had been collected.....	188
Figure B.4. Confirmation by Sanger sequencing that the expected MaYMV RT-PCR product was amplified from maize inbred B73 sample BR1B.....	190
Figure B.5. Representative maize leaf symptoms of samples that were positive for MaYMV with the RT-PCR assay using MaYMV-F and MaYMV-R primers.....	193
Figure B.6. Mild mosaic symptoms on leaf samples from maize inbred line SC Malawi shown to contain MaYMV in RNA-seq data.....	194
Figure B.7. Phylogentic analysis of African, American, and Chinese Maize Yellow Mosaic Virus isolates.....	195

List of Tables

Chapter 1

Table 1.1. A summary and comparison of three studies that used reference genome mapping.....	8
--	---

Table 1.2. A summary and comparison of three studies that used <i>de novo</i> assembly.....	13
---	----

Table 1.3. A comparison of the advantages and disadvantages of <i>de novo</i> assembly and reference genome mapping in assembling short transcriptome sequence reads.....	15
---	----

Table 1.4. A summary and comparison of three studies that used a combined strategy of <i>de novo</i> assembly and reference genome mapping.....	18
---	----

Chapter 2

Table 2.1. The RNA-seq reads produced from the B73-QTL leaf samples.....	37
--	----

Table 2.2. The accuracy obtained from each Trinity optimization run.....	38
--	----

Table 2.3. The set of plant proteins present in the unmapped assembly.....	40
--	----

Chapter 3

Table 3.1. Shortlist of defence gene candidates identified.....	59
---	----

Chapter 4

Table 4.1. The primer sequences used in the PCR-RFLP to distinguish between the candidate and B73 alleles.....	86
--	----

Appendix A

Table A.1. The list of novel transcripts from maize B73-QTL samples with BLASTP hits to plant sequences, sorted by the log2 Fold Change (FC) of their expression upon <i>C. zeina</i> infection.....	105
--	-----

Table A.2. The top 30 novel transcripts from maize B73-QTL samples with BLASTP hits to <i>E.turcicum</i> (<i>Setosphaeria turcica</i>) sequences.....	136
---	-----

Table A.3. The predicted proteins from the unmapped assembly identified by BLASTP against the NCBI nr database that matched putative plant viruses.....	141
---	-----

Table A.4. The predicted proteins from the unmapped assembly identified by BLASTP against the NCBI nr database that matched putative plant bacteria.....142

Table A.5. The predicted proteins from the control leaves identified in the blast that were neither plant nor fungal nor phylogenetic and sorted by species.....144

Table A.6. The predicted proteins from the *C. zeina* challenged leaves identified in the blast that were neither plant nor fungal nor phylogenetic and sorted by species.....146

Appendix B

Table B.1 MaYMV proteins identified from unmapped de novo assembled RNA in maize B73 plant 1 grown in field in KwaZulu-Natal, South Africa.....182

Table B.2. Maize RNA-seq datasets screened for the presence of MaYMV.....183

Table B.3 Comparison of predicted MaYMV proteins between MaYMV RSA BR1A from South Africa and MaYMV Yunnan 11 reference from China.....190

Table B.4. The percent nucleotide similarity for complete genomes between MaYMV from South Africa and other worldwide isolates.....191

Chapter 1: General Introduction

Literature review

Introduction

The bedrock of research has always been the production of data followed by the interpretation thereof. The advances in scientific technology correspondingly often aim to increase the amount and quality of the data produced for the researcher. Next generation sequencing technologies specifically, has revealed many new options for researchers by vastly increasing the amount of genomic and transcriptomic data that can be obtained. RNA sequencing (RNA-seq) in particular provides a higher coverage at a low cost method for transcriptomics (Schliesky *et al.* 2012). The options provided by RNA-seq are especially of interest when studying plants as they tend to have very large, repetitive genomes which often makes it easier to work with RNA data rather than DNA data (Martin *et al.* 2013).

A research field that has benefitted greatly from these new avenues provided by next generation sequencing is gene discovery. Genes may be difficult to define. The concept of a gene has existed as the means of heredity in organisms long before a mechanistic understanding of what a gene is was investigated and the definition of a gene has evolved along with our understanding of genetics (Gerstein *et al.* 2007). One of the current ways to define a gene is as “a union of genomic sequences encoding a coherent set of potentially overlapping functional products” (Gerstein *et al.* 2007) and in most cases the product or intermediate stage of the product is RNA. One way to discover genes therefore is through analysing the transcriptome, which can more easily be done through RNA-seq (Nagalakshmi *et al.* 2008).

However, the reads produced by next-generation RNA-seq are very short and generally range from only 30 bp up to 300 bp long. This is much shorter than the 1,000 bp reads produced by the previous Sanger sequencing platform and means that individual RNA-seq reads do not span a complete gene. This is a limitation in gene discovery since multiple reads that were sequenced from the same gene must first be reassembled to identify novel genes and variants. This can be done by *de novo* assembling the reads or by mapping them against a reference genome.

The subject of this review is to give an overview of the principals behind RNA-seq, reference-genome mapping and *de novo* assembly, examples of the bioinformatics tools used and the applications of the two strategies as an approach to discover novel genes in plants.

RNA-sequencing

RNA-seq is a high-throughput technique used to sequence a cDNA library. Any high-throughput sequencing technology, e.g. Illumina, can theoretically be used, but most of these platforms only operate on a specific size range of molecules and the RNA sample must first be fragmented into the appropriate sizes (Wang *et al.* 2009). The libraries are made by purifying the RNA, fragmenting the RNA (usually chemically, with the method depending on the desired insert size), synthesising the cDNA using reverse transcriptase and primers (usually random) and ligating adaptors to the ends of the cDNA (Wang *et al.* 2009).

The sequencing can be done from either one or both ends of the fragment, respectively creating single-end and paired-end reads. The approximate distance between the read pairs can thus also be determined from the insert size. Gene expression studies in relatively non-repetitive genomes do not require paired-end sequencing, but in repetitive genomes it helps to resolve the origin of the read and the additional information from paired-end sequencing is also preferable when studying alternative splicing (Van Verk *et al.* 2013). The cDNA libraries can also be to be strand-specific - meaning that the original strands are marked - which helps to identify transcripts and in reconstructing the genes structure (Levin *et al.* 2010). The two strands are generally distinguished from each other by ligating specific, orientated adaptors to the 5' and 3' ends or modifying the complementary strand during its synthesis (Levin *et al.* 2010; van Dijk *et al.* 2014).

The class of RNA of interest needs to be isolated from the total RNA which in the majority of cases means isolating messenger RNA (mRNA). This can be done by using oligo dT primers to selectively increase the mRNAs or by using Duplex-Specific Nuclease to degrade ribosomal RNA and tRNA structures which has the added benefit of not excluding mRNAs without poly(A) tails (Zhao *et al.* 2014). Other types of RNA, such as small RNA and microRNA, are more consistent than mRNA in size and length and can be isolated by using electrophoresis or chromatography to separate out the appropriate size for the desired class of RNA (Chionh *et al.* 2013).

The quality of the isolated RNA and the type of reads generated are important considerations. The simplest assessment of RNA quality is through gel electrophoresis. Ribosomal RNA (rRNA) is the most abundant RNA and dominates the visualization, although messenger RNA can be seen as a smear between the rRNA subunits. In intact RNA, the rRNA is separated into clear bands where the small subunit is twice as bright as the large subunit and their presence and ratio can be seen on an electrophoresis gel. This can also be automatically determined through a Bioanalyzer or Experion machine. A

Bioanalyzer interprets an electropherogram through an algorithm and gives the result as the RNA Integrity Number (RIN) which is scored on a 1 to 10 scale with 10 having the most intact RNA (Schroeder *et al.* 2006). An Experion uses the same principle, but uses a different algorithm and gives the result as the RNA quality indicator, which is functionally equivalent to RIN and is also on a 1 to 10 scale with 10 being the most intact (Denisov *et al.* 2008). Degraded RNA can interfere with sequencing and it is generally recommended that the RNA sample should have an RIN of at least 8 (Van Verk *et al.* 2013).

The relative abundance of reads produced is usually measured in reads per kilobase of transcript of million mapped reads (RPKM) or fragments per kilobase of transcript per million mapped reads (FPKM). The main difference between them is that in FPKM the fragment is the biological object produced during RNA-seq and could be single or paired reads (Trapnell *et al.* 2010). RPKM and FPKM are useful for measuring absolute expression, but counts per million (CPM), which is the number of reads divided by the size of the library in millions, is used to measure differential expression (Law *et al.* 2014). Alternatively, there is transcripts per million (TPM) or effective transcripts per million (eTPM). TPM measures the abundance of each isoform and eTPM measures the abundance of each effective transcript which is determined based on the distribution of reads (Huang *et al.* 2014).

RNA-seq data must be analysed through computational methods which raises the question of how the results can be validated. Ideally the results can be experimentally verified. This was used in one example (Petitot *et al.* 2016) who amplified their candidate effector genes with RT-qPCR to show that their genes were expressed as they had predicted from the RNA-seq data. Another validation method is to compare the assembled transcripts to known sequences from closely related species. This was used in another example (Xu *et al.* 2015) who *de novo* assembled a radish transcriptome from RNA-seq reads and validated their assembly by showing high homology between their transcripts and transcripts from 5 related Brassicaceae species, including the model species *Arabidopsis thaliana*.

Reference genome based mapping

Genes are longer than short RNA-seq reads and therefore multiple reads from different parts of each gene will be produced. To study a gene, it must first be determined which reads originated from that gene and how they relate to each other. If a reference genome is available this can be answered by aligning the reads to the reference and assembling adjacent reads together (Fig. 1.1). However, there are some complications to this assembly technique. Firstly, in repeated or similar sequences a read will have a match in each sequence - thereby making it unclear which alignment is the correct one - and secondly,

reads from spliced transcripts do not match the unspliced reference gene (Trapnell and Salzberg 2009).

Reads with multiple possible alignments are called ‘multireads’ and there are several methods to resolve which location to use. The most common solution is to assume none of the locations were more likely to be the source and choose one of the alignments at random, but this is inaccurate if the reads were more likely to have originated from specific locations (Zhao *et al.* 2016). Another solution is simply to bypass the problem by not including those reads in the analysis, but this causes information from duplicated and related genes to be overlooked (Treangen and Salzberg 2012; Zhao *et al.* 2016). Alternatively, an upper limit on the amount of alignments can be set and all the alignments under that can be used, but this brings in an additional complication of determining an appropriate cut-off value (Zhao *et al.* 2016). In a case where the repetitive sequences are very short micro-repeat sequences, it can improve accuracy to ‘mask’ them to exclude the repeat sequences in the genome before the alignment (Hillier *et al.* 2008).

Eukaryotic mRNA undergoes a process called splicing where parts of the RNA called introns are spliced out and parts called exons remain. Frequently alternative splicing occurs where the splicing is regulated to splice the mRNA in different places, resulting in different isoforms. This means there will be gaps in the alignment between the splice sites and that reads that span the splice junction will not align to the reference (Fig. 1.1). Mapping assemblers compensate for this by using annotations of known splice sites and using algorithms to identify novel splice sites (Trapnell *et al.* 2009). RNA-seq data can be used to identify new splice sites even at low coverage, but the results are not consistent between different assemblers (Seqc/Maqc-III_Consortium 2014).

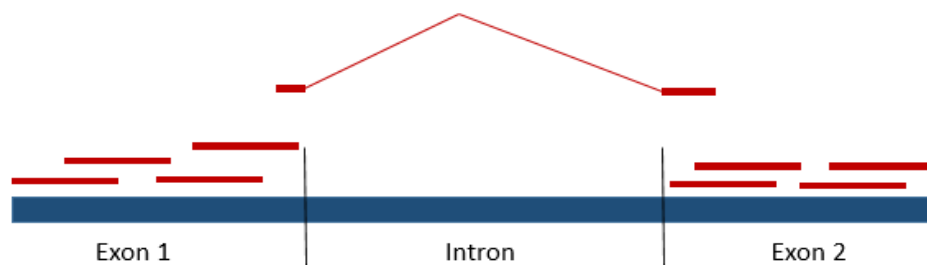


Figure 1.1. An illustration of reference genome mapping of RNA-seq reads. The reads of a transcript align to the matching sequence the transcript was transcribed from and adjacent reads can be assembled into the transcript. As the sequenced mRNA transcript consists of exons spliced together the reads map to the exons and reads from the exon junctions map partly to each exon and not to the intervening intron sequence.

Multireads are also a problem between the alternatively spliced mRNA transcripts, as a consequence of alternative splicing. The different isoforms created by the splicing will have sequences in common with each other that makes it unclear which isoform a read from such an area originated from (Fig. 1.2). Isoform quantification methods solve this by modelling the reads. Originally the read counts were modelled as Poisson variables and maximum likelihood estimates used to determine the correct alignment (Jiang and Wong 2009), but many improved models have been developed since. The recently developed Multiple RNA-seq Samples for accurate Isoform Quantification (MSIQ) is the first model that compensates for the noise and bias present in RNA-seq datasets. This is accomplished by using RNA-seq on multiple samples which should have the same transcriptome, comparing them to each other to estimate which is the most representative and reweighting the quantification accordingly (Li *et al.* 2016).

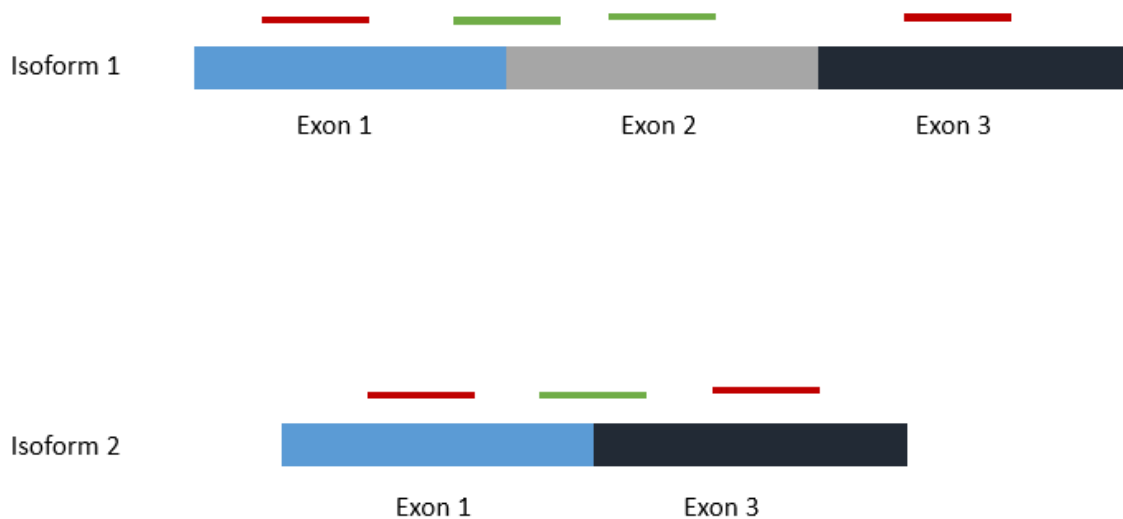


Figure 1.2. An illustration of reads from different isoforms of an alternatively spliced transcript. The green reads are from an area unique to an isoform and the red reads are from areas the isoforms have in common.

This strategy was used to discover genes in the plants, *Vitis vinifera* (Zenoni *et al.* 2010), *Oryza sativa* (Lu *et al.* 2010) and *Triticum aestivum* (Ramirez - Gonzalez *et al.* 2015) (Table 1.1). Zenoni *et al.* (2010) and Lu *et al.* (2010) mapped the RNA-seq reads to the reference genome which allowed them to identify potential novel genes from the reads that did not map near known genes. Ramirez-Gonzalez *et al.* (2015) did not have a complete genome for their organism and had to map their reads against individual *T. aestivum* gene models as well. Zenoni *et al.* (2010) and Lu *et al.* (2010) had also aligned their reads against splice junctions to discover novel isoforms of the known genes as well. All three studies used

the alignments to identify SNPs where the sequenced organisms had different alleles than the respective reference genome.

The three examples of reference genome mapping (Table 1.1) illustrate this strategy's advantages and disadvantages. An advantage of reference genome mapping is its ability to accurately identify transcripts at a range of expression levels including very low levels. This can be seen in the example of Zenoni *et al.* (2010) who could accurately assemble even the rarely expressed transcripts. Another advantage is its ability to compare changes across the entire genome as demonstrated by Zenoni *et al.* (2010) who could compare the difference in transcriptional regulation at each developmental stage and in the example of Ramirez - Gonzalez *et al.* (2015) who could use the alignments to identify SNPs. On the other hand, the main disadvantage of reference genome mapping is the reliance on the availability and the quality of the references and annotations. This can be seen in the three example. Firstly, Zenoni *et al.* (2010) lacked a validated alternative splicing database for *V. vinifera* and had to first create a junctions-database out of all the possible splice junctions in the annotated transcripts. Secondly, Ramirez - Gonzalez *et al.* (2015) lacked a complete reference of the same variety and had to make several alignments. Finally, Lu *et al.* (2010) who worked on a model plant and did not have such problems.

Table 1.1. A summary and comparison of three studies that used reference genome mapping.

Plant Sequenced	Tissue sampled	Extraction method	Mapping Program	Transcripts Assembled	Purpose of Study	Study
<i>Vitis vinifera</i> (Grapevine)	Berries	(Rezaian and Krake 1987)	ELAND	17,324	To validate RNA-seq as a method and to analyse the berries' unique development process	Zenoni <i>et al.</i> 2010
<i>Oryza sativa</i> (Rice)	Shoots	TRIzol reagent (Invitrogen)	Sequence Search and Alignment by Hashing Algorithm	15,708	To improve the rice genome annotations	Lu <i>et al.</i> 2010
<i>Triticum aestivum</i> (Bread Wheat)	Leaves	TRIzol reagent (Invitrogen)	Burrows-Wheeler Aligner	n/a*	To identify new SNP markers for use in wheat breeding	Ramirez-Gonzalez <i>et al.</i> 2015

*They were interested in SNPs within the genes, not whole transcripts and did not assemble them.

De novo assembly

The other assembly approach is *de novo* assembly where the reads are compared to each other instead of to a reference. When the transcripts are fragmented the fragments from the same transcript will have overlaps with each other. This is because there are many copies of each transcript in which fragments break off at random sites (Fig. 1.3). The transcript can be reassembled by joining these overlapping reads together.

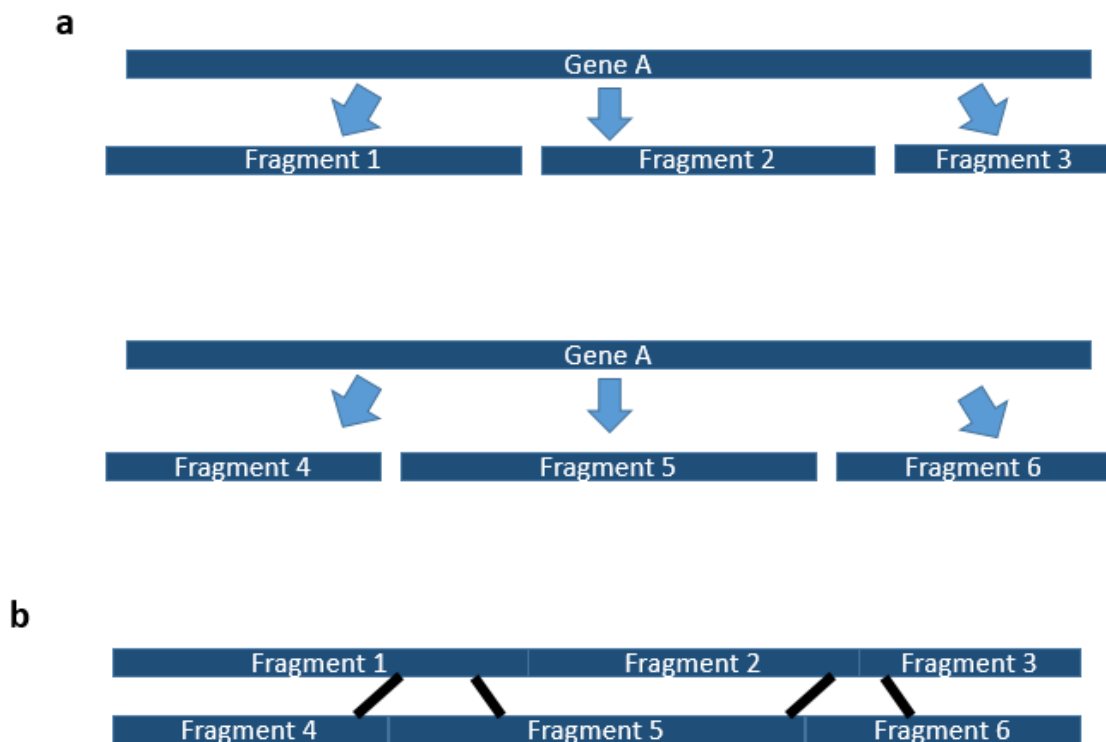


Figure 1.3. Fragmentation creates overlapping fragments. a. The fragmentation is at random sites so different copies of Gene A produces different sized fragments. b. Joining overlapping fragments re-assembles Gene A.

Since every read must be compared to every other read it is exponentially more computationally intensive than comparing every read to a reference. *De novo* assemblers can make this problem more manageable to approach by using a de Bruijn graph. In most modern assemblers the reads are further divided into shorter overlapping segments of length k (k -mers) and the de Bruijn graph is used to determine which reads to join together (Fig. 1.4) (Compeau *et al.* 2011).

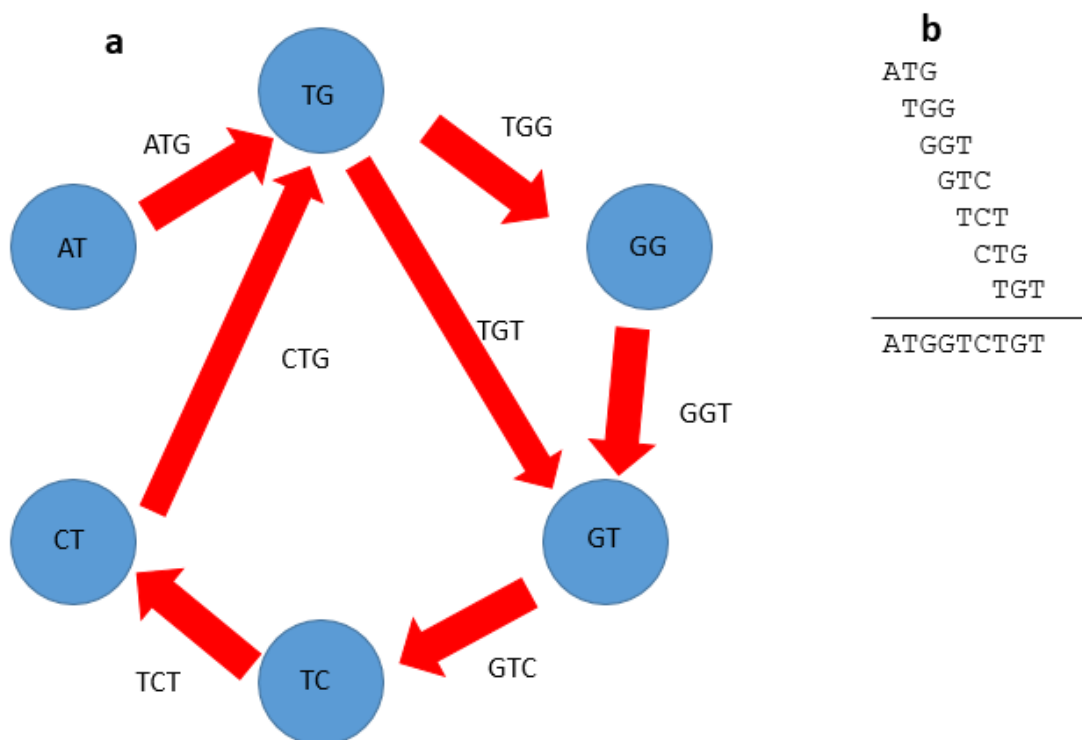


Figure 1.4. An illustration of de novo assembly using a de Bruijn graph (based on a diagram from Compeau et al. 2011). Each possible, unique k-mer (a 3-mer in this example) sequence is represented an edge of the graph and its prefixes and suffixes are represented as nodes. a) The de Bruijn graph is solved by finding the path through the graph that visits every edge exactly once. b) The k-mers from successive edges on the path are aligned and assembled together to extend the sequence to the end of the transcript.

The problem of reads mapping to multiple locations also affects *de novo* assembly. A *de novo* assembler which encounters indistinguishable repeats either risk mistakenly joining contigs from the separate repeats or avoids this by fragmenting the assembly (Treangen and Salzberg 2012). This problem can be overcome by sequencing paired reads and using the known distance and orientation between the pair to distinguish repeats or generating longer paired reads that span the repeat (Treangen and Salzberg 2012). Alternatively, the repeat structure could be estimated beforehand or statistically determined based on the coverage of the regions from libraries with different size fragments (Treangen and Salzberg 2012).

De novo assembly was used to discover genes in the plants *Cicer arietinum* (Garg et al. 2011), a commercially used *Eucalyptus grandis* × *Eucalyptus urophylla* hybrid (Mizrachi et al. 2010) and *Medicago sativa* (Yang et al. 2011) (Table 1.2). *C. arietinum* and *M. sativa* had very limited genome information available and *de novo* assembly was required for both plants to determine which genes were present. The *Eucalyptus* genome was in the process of being assembled and Mizrachi et al. (2010) used *de novo* assembly to discover genes as an aid to annotating the reference genome.

The three examples of *de novo* assembly (Table 1.2) illustrate this strategy's advantages and disadvantages. The main advantage of *de novo* assembly is that reference genomes and annotations are not required. This can be seen in all three examples where the authors used *de novo* assembly to create genomic resources for plant species that did not have good quality genomic resources available. On the other hand, *de novo* assemblers are more limited by low expression levels. This can be seen in the case of Yang *et al.* (2011) who had to use other sequences to compensate for transcripts they could not detect due to differential expression and in the case of Mizrachi *et al.* (2010) who attributed not being able to find coding sequences in a portion of their transcripts to low expression of those transcripts.

Table 1.2. A summary and comparison of three studies that used *de novo* assembly.

Plant Sequenced	Tissues sampled	Extraction method	Purpose of Study	Assembly Program	Transcripts Assembled	Study
<i>Cicer arietinum</i> (Chickpea)	Seeds, roots, shoots, mature leaves and flower buds.	TRI Reagent (Sigma Life Science, USA)	To develop a <i>de novo</i> assembly strategy and generate the first complete chickpea transcriptome	Velvet and Oases	35,409	Garg <i>et al.</i> 2011
<i>Eucalyptus grandis</i> × <i>Eucalyptus urophylla</i> (Eucalyptus)	Immature xylem, xylem, inner bark surface, shoot tips, young and mature leaves.	(Chang <i>et al.</i> 1993)	To create a high-quality expressed gene catalogue to aid in genome annotation and genomics	Velvet	6,208	Mizrachi <i>et al.</i> 2010
<i>Medicago sativa</i> (Alfalfa)	Stem internodes.	CTAB	To improve the limited genomic resources for alfalfa by creating the first alfalfa gene index	Velvet	124,025	Yang <i>et al.</i> 2011

Comparison of *de novo* and mapping strategies

The different reference genome mappers use the same principle, but they do not use same methods to accomplish this. The same also applies to different *de novo* assemblers. This makes a direct comparison difficult, but there have been studies that compared several assemblers of both kinds that shows their relative performance in general (Table 1.3).

Vijay *et al.* (2013) used a simulation with test data to compare the *de novo* assembly by Trinity and SOAPdenovo-Trans to the reference-based assembly by Stampy. They evaluated the strategies based on the proportion of bases from the reference transcriptome that are present in assembly (recovery) and the proportion of bases that are accurately placed (accuracy). The mapping assembly outperformed the *de novo* assemblers in terms of recovery on 5 and 15% diverged references. On the other hand, the mapping assembly performed worse on a 30% diverged reference and the *de novo* assemblers performed better on the whole in terms of accuracy. The recovery of the mapping assembly was less affected by transcriptome complexity than the *de novo* assembly. Polymorphisms and errors increased the recovery by a small amount and decreased the accuracy of the mapping assembly, but in the *de novo* assemblies the accuracy was not significantly affected and the recovery decreased.

Lu *et al.* (2013) also compared *de novo* assemblies by trans-ABYSS, Oases and Trinity to reference-based assemblies by Scripture and Cufflinks. They concluded that each assembler captured transcripts that the others did not and the performance of an assembler using either strategy was dependant on the conditions. In general, the reference genome mapping approach had required less computational time and memory, and was more specific. On the other hand, the *de novo* assemblers had detected more transcripts and isoforms, and only performed as well as reference-based assemblers when working with genes with high expression levels. This is because genes expressed at low levels may not be completely covered by the reads and there are no overlapping ends in these gaps that can be used to extend the sequence. Reference-based mapping assemblers are less sensitive to expression level than *de novo* assemblers as they also take genomic position into account and therefore can group nearby reads together despite gaps (Lu *et al.* 2013). The level of expression which is too low to assemble would therefore depend on the coverage and the cut-off would be lower for *de novo* assemblers.

Table 1.3. A comparison of the advantages and disadvantages of de novo assembly and reference genome mapping in assembling short transcriptome sequence reads.

De novo assembly		Reference genome mapping	
Advantages	Disadvantages	Advantages	Disadvantages
Does not require previous genomic information	Needs more computational resources	Less computationally expensive	Requires previous genomic information
Detects more transcripts and isoforms	Genes needs to be expressed at a high level	Genes do not need to be expressed at a high level	Detects less transcripts and isoforms

Combination of *de novo* and mapping strategies

The advantages and disadvantages of the two strategies are thus complementary. This means that combining them can be very useful to gain a more holistic assembly that includes transcripts that would have been missed otherwise. The combinatory strategies can also provide extra information as the reads mapping to the reference can be identified as originating from specific known loci and the other reads will still be *de novo* assembled and identified as originating from novel loci.

This combinatory strategy was used to discover genes in the *Meloidogyne graminicola*-*Oryza sativa* pathosystem (Petitot *et al.* 2016), the plant *Fragaria* × *ananassa* (Sánchez-Sevilla *et al.* 2014) and the plant *Brassica rapa* (Devisetty *et al.* 2014) (Table 1.4). Petitot *et al.* (2016) RNA-sequenced the parasite and host during infection and then mapped the *O. sativa* reads against its reference genome to isolate the *M. graminicola* (which did not have a reference sequence) reads and *de novo* assembled them. This allowed Petitot *et al.* (2016) to exclude the host genes they were not interested in and discover parasite genes expressed during infection. Sánchez-Sevilla *et al.* (2014) mapped their reads against the draft *Fragaria* × *ananassa* reference genome, but also used *de novo* assembly to extend genes that had been truncated in the draft. Devisetty *et al.* (2014) had mapped their reads against the *B. rapa* reference genome to discover SNPs also *de novo* assembled the reads to include novel genes.

The three examples of combining the two assembly strategies (Table 1.4) illustrate how this can create new opportunities for analysis. The combinatorial approach is especially useful when identifying specific novel transcripts as the reference genome mapping can be used to exclude the other reads and the transcripts of interest assembled *de novo*. This can be seen in the example by Petitot *et al.* (2016) who used mapping to exclude the reads from the host so that they could *de novo* assemble the parasite transcripts. Another advantage is that because the two strategies have different strengths and weakness a more comprehensive and complete assembly can be made by merging their results as done by Devisetty *et al.* (2014) and Sánchez-Sevilla *et al.* (2014).

Table 1.4. A summary and comparison of three studies that used a combined strategy of de novo assembly and reference genome mapping.

Plant Sequenced	Tissues sampled	Extraction method	Mapping program	De novo assembler	Transcripts Assembled	Purpose of Study	Study
Meloidogyne graminicola-Oryza sativa (Nematode infected rice)	Root tips	RNeasy Plant mini kit (Qiagen)	TopHat	Oases	66,396	To identify effector genes used by a parasitic nematode to infect rice plants	Petitot <i>et al.</i> 2016
Fragaria × ananassa (Strawberry)	Ripe fruits	Differential 2-butoxyethanol precipitation-based method	TopHat	Trinity	33,458	To identify the regulatory gene responsible for γ-decalactone content	Sánchez-Sevilla <i>et al.</i> 2014
Brassica rapa (Field mustard)	Whole seedlings, roots, vegetative and inflorescence meristems, internode and leaves.	RNAeasy Plant Mini Kit (Qiagen)	TopHat	Velvet and Oases	44,239	To identify coding SNPs and to improve the outdated <i>B. rapa</i> genome annotations.	Devisetty <i>et al.</i> 2014

Bioinformatics tools

The usefulness and difficulty of working with RNA-seq data has led to the development of many different software programs specializing in the different needs of researchers.

There are two categories of alignment programs used to map reads to a reference. Gapped aligners take into account potential splice sites and are used to map reads to a genomic reference, and ungapped aligners do not take into account potential splice sites and are used to map reads to a transcriptomic reference (Conesa *et al.* 2016). A transcriptomic reference is made from already spliced sequences so introns will not cause gaps in the alignment. Usually a separate program is used to assemble the aligned reads into transcripts.

Bowtie is a popular example of an ungapped aligner. Bowtie creates index files of the reference sequences and uses them to align the reads to the reference (Langmead *et al.* 2009; Langmead and Salzberg 2012). TopHat is a popular example of a gapped aligner (Trapnell *et al.* 2009). TopHat first uses Bowtie to align the reads to the reference genome. TopHat then searches for potential splicing donor and acceptor sites within the reference and predicts which donor and acceptors sites in neighbouring contigs could form an intron. The unmapped reads are aligned to the edges of the possible introns to find reads that span the splice junctions. StringTie (Pertea *et al.* 2015) is an example of an alignment assembler. StringTie simultaneously assembles the reads by grouping them into clusters and assembling the transcripts in the cluster based on a splice graph and uses a maximum flow algorithm to estimate their expression levels.

Most of the software for *de novo* assembly of RNA-seq reads started out as modifications of programs designed to assemble genomic short sequence reads (Paszkiewicz and Studholme 2010). For example the genome assemblers SOAPdenovo, ABySS and Velvet have been used to assemble the transcriptomes of various organisms (Zhao *et al.* 2011). There have also been programs developed specifically for RNA-seq, for example Trinity (Grabherr *et al.* 2011). Zhao *et al.* (2011) compared the performance of these assemblers on test data and found Trinity had the highest percentage of reads that could be mapped back to transcripts, Oases had the highest maximum memory requirement and ABySS the lowest, Trinity had the longest runtime and SOAPdenovo the shortest and Trinity produced the most accurate unique transcripts and SOAPdenovo the least.

There are also programs that attempt to resolve the problems caused by repeat sequences. One of the first programs designed to find and mask repeat elements was the library-based Censor, which aligns the query against a library of known repetitive elements (Jurka *et al.* 1996). The most widely used currently is RepeatMasker, which aligns the query sequences

to a library file of repeat consensus sequences for the organism (Tarailo - Graovac and Chen 2009). Alternatively, the optimal resolution of multimapping ambiguity of RNA-seq reads (ORMAN) approach can be used if there are annotations or transcriptome data available as well (Dao *et al.* 2013). This involves all the canonical exons being used to make a gene model where the substrings are partial transcripts, with mismatched regions smaller than 15 nucleotides being considered a new transcript with an indel, then distributing the minimum set of partial transcripts that covers all the reads and the reads to the set of partial transcripts according to the most likely distribution. Dao *et al.* (2013) tested their method on simulated data and it was able to place the multiple mapping reads and improved downstream analysis.

Conclusion

RNA-seq is a versatile tool, but the different strategies to process the data have their own benefits and shortcomings. Which of these advantages and disadvantages of each strategy are more relevant would depend on the data being analysed and the purpose of the analysis. For example, detecting genes with a low expression is more of a concern in a dataset with low coverage and the requirement of supplying genomic information is less of a concern in a model plant whose reference is accurate and complete. Using *de novo* assembly would also often be a matter of necessity when there are no reference genomes available and large plant genomes could make reference genome mapping less efficient than in other organisms.

The applications provided by a combinatorial approach are well suited for plant gene discovery. Firstly, novel genes are usually present in a genotype different from the model genome used as a reference and mapping assemblers are dependent on an accurate reference. This means that if the novel gene is too different from the known alleles or not present at all in the model genotype the reference mapper would not identify it, but including a *de novo* assembly step on the unmapped reads could solve this issue. Secondly, the ability to use reference mapping to filter out known genes is also very useful when the aim is to identify specific genes instead of a wide range of genes. Thirdly, creating a more complete assembly would be useful for discovering rare genes or variants. On the other hand, since both strategies are used it is essentially twice the work and would require more time and computational resources.

As sequencing technologies advance RNA-seq studies will become progressively more accessible. This will provide more opportunities for these strategies to be used, and potentially adapted to output produced by the next generations of sequencing technology.

Project Rationale

Maize is one of the three main cereal crops that are globally used as a food source and is especially important in regions such as Africa where it is a staple food (Ranum *et al.* 2014). It contributes significantly to food security and is a widespread crop in smallholder farming, but its production is often hampered in developing countries by various factors including diseases (Shiferaw *et al.* 2011).

Grey leaf spot (GLS) is a foliar fungal disease that greatly reduces global maize yields (Zheng *et al.* 2010; Ward *et al.* 1999). It was first identified in America as a maize disease caused by the fungus *Cercospora zeae-maydis* (Tehon and Daniels 1925). Isolates from diseased leaves in Southern Africa however revealed that the GLS was caused in South Africa by another *Cercospora* species; *Cercospora zeina* (Meisel *et al.* 2009).

GLS can be controlled through a combination of crop rotation, fungicide application and cultivating resistant hybrids (Latterell and Rossi 1983; Ward *et al.* 1999; Mallowa *et al.* 2015). Tillage also prevents the spread of GLS, but does not have a significant influence if other sources of inoculum are readily available (Ward *et al.* 1997b; Ward *et al.* 1997a; Asea *et al.* 2002). Increasing conventional tillage will therefore not only be more harmful to the environment (Coolman and Hoyt 1993), but is not likely to be an effective control measure in areas where GLS is already endemic.

Incorporating GLS resistant hybrids have had good results in mitigating yield losses. (Asea *et al.* 2009; Wallhead 2012; Tefera *et al.* 2013). GLS resistance is a quantitative trait with generally a high heritability (Benson *et al.* 2015) and many QTLs that confer GLS resistance have been identified. Despite this, the molecular basis of the resistance is still poorly understood.

This study will address the current shortfall by identifying novel defence expressed genes within QTL10 of the maize line CML444 that are associated with GLS resistance. This QTL was identified to increase GLS resistance (Berger *et al.*, 2014) and was backcrossed into the public, susceptible B73 line to produce the near isogenic line B73+QTL10 (B73-QTL). RNA-seq data (Wang *et al.* 2009) from B73-QTL will be analysed to identify these novel genes. RNA-seq reads have been used in the past as an effective method for gene discovery (Xu *et al.* 2015; Garg *et al.* 2011; Mizrahi *et al.* 2010; Song *et al.* 2015).

Xu *et al.* (2015) *de novo* assembled radish genes which were experimentally validated by amplifying them with RT-qPCR and by showing high homology between their transcripts and closely related species. Mizrahi *et al.* (2010) *de novo* assembled *Eucalyptus* genes and

validated their results by comparing a subset of their assembled transcripts to corresponding sequences in NCBI and independently realigning their reads to the contigs. Song *et al.* (2015) identified the gene causing a QTL in rice using reference genome mapping and used *in vivo* and *in vitro* assays to identify it as a histone H4 acetyltransferase.

As B73-QTL is identical to B73 apart from the QTL region any reads that could not be mapped to either the B73 version 4 or *C. zeina* genomes should hypothetically be from the QTL region and the novel genes therein. Therefore, the hypothesis is that a *de novo* assembly of the unmapped reads would reveal novel defence genes.

Aims and Objectives

Aims

1. To identify novel expressed genes in the QTL10 region that are associated with GLS resistance.
2. To identify genes in B73-QTL that are differentially expressed in response to GLS.
3. To experimentally validate the candidate genes.

Objectives (Fig. 1.5)

1. To map the RNA-seq reads against the B73 v4 reference genome and the *C. zeina* draft genome using TopHat2.
2. To analyse the unmapped reads using the following pipeline:
 - a. To assemble the unmapped reads using Trinity.
 - b. To find likely coding regions within the assembled transcripts using TransDecoder.
 - c. To filter out any *C. zeina* transcripts by comparing the predicted proteins to the predicted *C. zeina* polypeptides.
 - d. To filter out any B73 transcripts by comparing them to B73 proteins.
 - e. To identify the proteins by comparing them to proteins in the plant, fungi and bacteria RefSeq protein database.
3. To use edgeR to test for differential expression in response to GLS.
4. To annotated the plant sequences using PLAZA.
5. To select likely candidate genes in the annotated plant genes.
6. To design primers for the candidate genes using Primer Designer.
7. To validate the candidate genes, DNA extracted from B73-QTL will be amplified with PCR using the designed primers. The PCR products will be sequenced using Sanger sequencing to confirm the candidate genes were amplified.

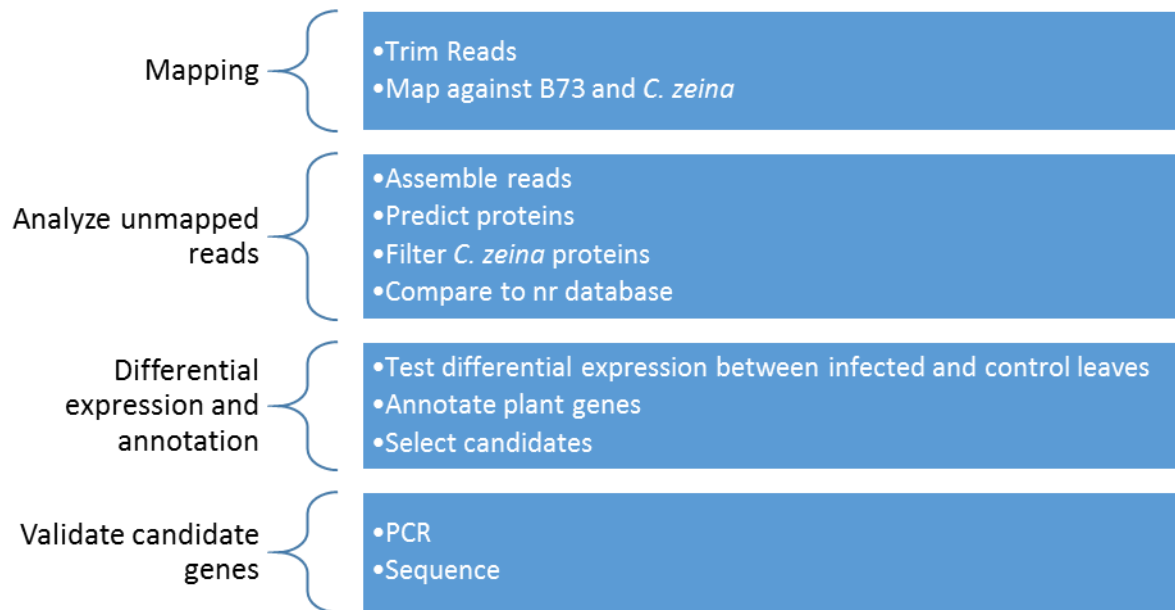


Figure 1.5. Flow diagram of the methodology.

Chapter 2: Candidate defence genes identified through designed bioinformatics pipeline

T. Welgemoed, R. Pierneef and D.K. Berger

Author Contributions

T.W.: Conducted the bioinformatics and data analyses. Wrote the chapter.

R.P.: Guided the bioinformatics analyses, suggested methods. Reviewed chapter and suggested improvements.

D.K.B.: Investigator of the funding project, acquired funding, provided day-to-day institutional supervision. Reviewed chapter and suggested improvements. Led discussions on data interpretation.

D.K.B. designed the study.

Introduction

As was outlined in Chapter 1, *de novo* assembly remains a useful component in gene discovery and cataloguing even when a reference genome is available. It is especially useful in organisms such as maize where many genes are not shared between different genotypes. A comparison of other maize lines against the B73 reference found that only 88.3% are not variable between genotypes (Swanson-Wagner *et al.* 2010) and there are at least 2000 genes between the B73 and Mo17 maize lines which differ in copy number or presence/absence (Springer *et al.* 2009). For example, Jin *et al.* (2016) used RNA-seq to identify maize transcripts that varied from the reference genome in 368 diverse inbred lines. They *de novo* assembled the reads from each line and aligned them to the B73 v2 reference genome to exclude B73 sequences. Jin *et al.* (2016) then filtered the transcripts to remove contaminating non-maize sequences and low coverage transcripts, which resulted in a total of 2,355 novel transcripts

In another example, Hansey *et al.* (2012) used RNA-seq data to catalogue variation in maize genes and to assemble novel maize genes. They RNA-sequenced 21 lines of genetically diverse maize seedlings and mapped the reads to the B73 v2 reference genome. The mapped reads for each line comprised of between 78% and 81% of the total reads and was used to call 350,710 SNPs between the maize lines. Hansey *et al.* (2012) then *de novo* assembled the unmapped reads and obtained 1,321 novel transcripts that did not match B73 sequences.

In a third example, Lai *et al.* (2010) performed a similar study on six commercially significant lines of maize. They mapped the RNA-seq reads of each line against the B73 reference genome and *de novo* assembled the unmapped reads. Lai *et al.* (2010) obtained 130,053 SNPs in coding regions and 157 *de novo* assembled maize transcripts that were absent in the B73 line.

The current project also used a strategy of alignment to exclude known genes and *de novo* assembly to capture novel the transcripts. RNA-seq reads were mapped against the B73 version 4 reference genome (Jiao *et al.* 2017) and a draft of the *C. zeina* genome (Wingfield *et al.* 2017). B73-QTL is nearly identical to B73 apart from the QTL region and therefore any reads that could not be mapped to either the B73 or *C. zeina* reference should hypothetically be from the QTL region and the novel genes therein.

There are two categories of aligners used to map reads to a reference. Gapped aligners are used to map reads to a genomic reference and ungapped mappers are used to map reads to a transcriptomic reference (Conesa *et al.* 2016). A genomic and transcriptomic reference will differ because eukaryotic mRNA undergoes a process called splicing where parts of the RNA called introns are spliced out and parts called exons remain. This means there will be gaps in the alignment between the splice sites and that reads that span the splice junction will not align to the reference. The algorithms used by gapped aligners detect and adapt to these gaps. In this study the gapped RNA-seq mapping tool, TopHat2 (Kim *et al.* 2013) was used as the only reference available for *C. zeina* is a draft genome and hence a gapped aligner was required.

The *de novo* assembly was done using Trinity (Grabherr *et al.* 2011) which is a software package designed for the *de novo* assembly of full-length transcripts based on RNA-seq reads. Compared to other assemblers Trinity, despite its comparatively long runtime, performs well overall and provides additional information on the different isoforms which is very useful during gene annotation (Vijay *et al.* 2013; Zhao *et al.* 2011).

The hypothesis was therefore that a *de novo* assembly of the unmapped reads would reveal novel defence genes. The designed bioinformatics pipeline assembled the unmapped reads using *de novo* assembly, identified the open reading frames within the assembled transcripts, filtered out non-maize sequences and compared them to similar proteins.

Methods

Plant Material

The QTL originated from the subtropical maize line CML444. It was identified to increase GLS resistance in a 2009 field trial in Hildesheim as 10G2_GLS and in another independent

2009 field trial in Baynesfield as 10H_GLS (Berger *et al.*, 2014). This region on chromosome 10, henceforth called QTL10, was backcrossed into the public, susceptible B73 line to produce the near isogenic line B73+QTL10 (B73-QTL) (Fig. 2.1). During the backcross breeding four SSR markers (Berger *et al.* 2014) were used where two markers flanked the QTL region (umc1084-10 and bnlgl450-10) and another two were located within the region (umc1196-10 and umc1176-10).

A multiplex PCR with primers for these four markers was used to screen progeny for the presence of QTL10. The multiplex PCR product was diluted 1:40 and formamide and GeneScan TM 500 LIZ TM dye Size Standard (Applied Biosystems, CA, USA) was added in a 10:0.12 ratio. The background was selected for B73 origin using 384 SNP markers across the genome. Capillary electrophoresis in the University of Pretoria (UP) sequencing facility and the GeneMarker v1.95 (SoftGenetics, PA, USA) software was used to determine the genotypes.

Three biological replicates of B73-QTL (LGT61807, LGT61819 and LGT61841) and three of B73 (LGT61811, LGT61820 and LGT61837) were RNA-sequenced.

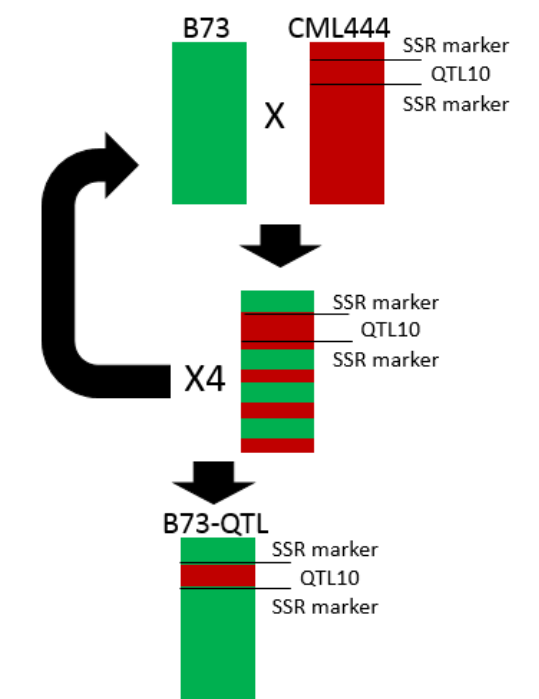


Figure 2.1. Breeding of B73-QTL by Felix Middleton (chromosome 10 shown). B73 was used as the recurrent parent and CML444 used as the donor parent for introgression of QTL10. SSR markers (Berger *et al.* 2014) were used to select progeny with the QTL10 region and 384 SNP markers spread across genome was used to select for B73 background. The progeny was backcrossed another four times, still selecting for QTL10.

The two lines had been planted on the 28 December 2012 at the Hildesheim farm in Greytown in a randomized block design with three replicate blocks of ten plants per row (Fig. 2.2). The area had a high GLS disease pressure already present due to natural inoculum of *C. zeina*. This field trial was also reported in studies by (Christie *et al.* 2017) and Swart *et al.* (2017), but the current study used the B73-QTL data whereas they used the B73 data.

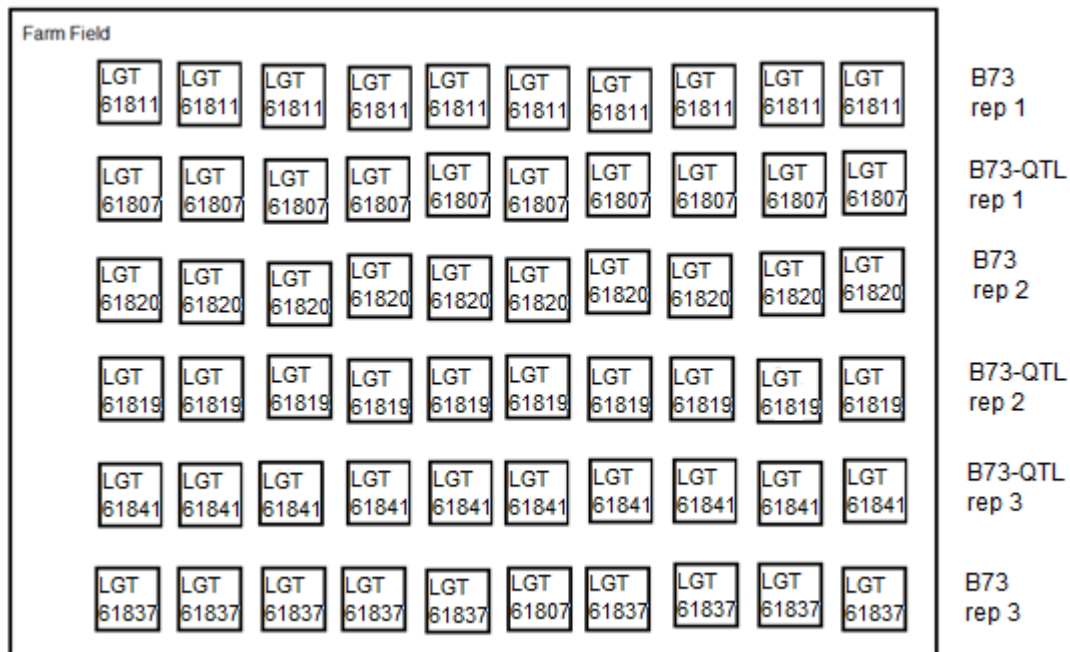
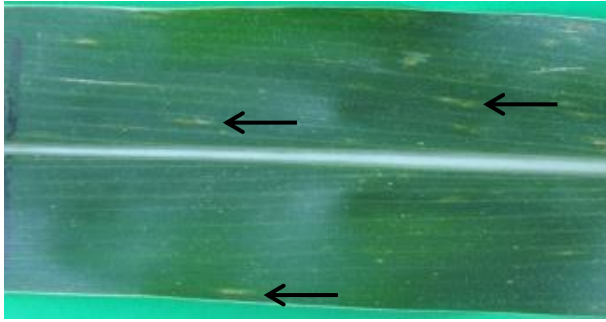


Figure 2.2. Diagram of the layout of B73 and B73-QTL plants (represented as blocks) at Hildesheim farm in Greytown used as material for the RNA-seq. They were planted in three replicate rows of 10 plants per row. Upper and lower leaf samples from one B73 plant and one B73-QTL plant in each replicate were harvested and subjected to RNA-seq separately. Data from the B73-QTL plants was analysed in this study in Chapter 2 & 3. Data from B73 plants published in Swart *et al* (2017) was included for comparison of differential expression in Chapter 3.

Samples the size of a CD cover (~30.65 cm²) were cut from the middles of two leaves above and two leaves below the cob from each maize plant in the replicates at 77 days after planting. In the leaves below the cob the infection had progressed and they developed GLS symptoms. In the leaves above the cob the infection had not progressed into causing GLS yet (Fig. 2.3). There are therefore two conditions, *C. zeina* challenged and control, and three biological replicates of the conditions.

i. Upper leaf: early *C. zeina* infection



ii. Lower leaf: severe *C. zeina* infection



Figure 2.3. An example of the leaf material harvested for sequencing. This shows the upper (control) and lower (*C. zeina* challenged) leaves harvested from one of the B73 replicates from the supplementary of Christie *et al.* (2017).

Data available for project

The RNA was extracted from each leaf sample at UP and stored on dry ice. The RNA-seq was done by BGI Tech Solutions Co., Ltd. (Beijing Genome Institute; Hong Kong) using Illumina HiSeq™ 2000 with a 100 bp paired-end module (Illumina Inc., San Diego, USA). This resulted in six datasets of RNA-seq reads consisting of three replicates of the control B73-QTL leaves and three replicates of the *C. zeina* challenged B73-QTL leaves (Table 2.1). In the names of the datasets were indicated the identity of the plant (B10 for B73-QTL), the replicate number of the plant (R1 for replicate 1, etc.) and the condition of the leaf (A for control or B for *C. zeina* challenged). The RNA-seq data from the B73 leaves from (Swart *et al.* 2017) are also available in the Gene Expression Omnibus GEO (GEO) under GSE94442.

Bioinformatics Pipeline

The candidate genes were identified using computational methods using the resources of the Centre for Bioinformatics and Computational Biology (CBCB) at UP for the *in silico* analysis.

The bioinformatics pipeline involves two main phases: mapping the B73-QTL reads to the references, and analysing the unmapped reads to identify novel plant genes (Fig. 2.4).

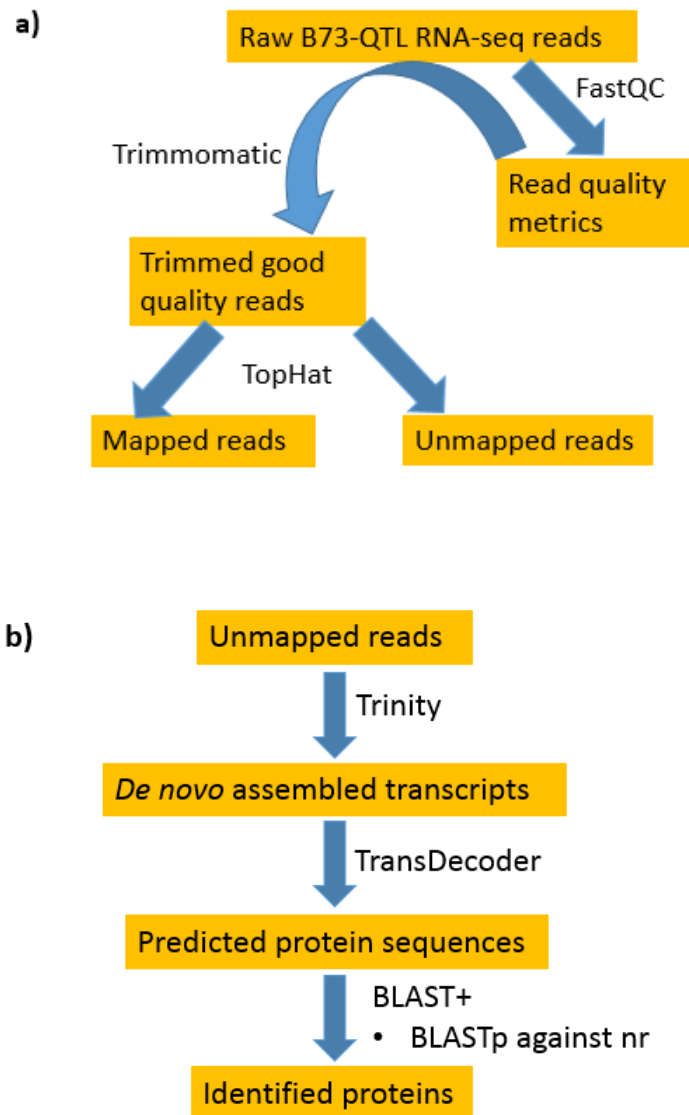


Figure 2.4. An overview of the bioinformatics pipeline. a) The raw reads are assessed with FastQC, trimmed with Trimmomatic based on the FastQC results, aligned to the references using TopHat. b) The unmapped reads are assembled using Trinity, their protein sequences predicted with TransDecoder and compared against the non-redundant protein database using BLASTp.

The B73-QTL FASTQ read files from BGI was examined with FastQC (Andrews 2016) for low quality bases and sequences and remaining adapter sequences.

Strict quality filtering was used to trim the reads through Trimmomatic (Bolger *et al.* 2014a). Firstly, the first 15 bases containing the random primer sequences and adapter sequences were trimmed from each sequence (Appendix C, File C.1). Secondly, low quality bases were trimmed using a sliding window of five bases and trimming the sequence when the average quality in the window dropped below a Q score of 28 (Appendix C, File C.1). Lower quality reads would be less likely to map and it was important to keep the pool of unmapped reads that would be used in the *de novo* assembly of good quality. Finally, the sequences that

were trimmed to less than 50 bp were discarded to achieve better performance during the alignment (Appendix C, File C.1). Trimmomatic produces four FASTQ output files for each set of forward and reverse FASTQ read files: a paired forward file and a paired reverse file for reads whose pairs are retained, and an unpaired forward and an unpaired reverse file for reads whose pairs are not retained.

The retained paired forward and reverse FASTQ files were mapped against the B73 version 4 reference genome (Jiao *et al.* 2017) and the *C. zeina* draft genome (Wingfield *et al.* 2017) using TopHat2 (Kim *et al.* 2013) with the default settings and the GTF annotation for B73 supplied (Appendix C, File C.2).. TopHat2 can detect all splice versions of the gene, including novel splicing events. This was expected to exclude the B73 genes thoroughly and accurately.

The *C. zeina* reference is available on GenBank under the accession MVDW01000014.1. The B73 references was obtained from ftp://ftp.ensemblgenomes.org/pub/plants/release-34/fasta/zea_mays/ and ftp://ftp.ensemblgenomes.org/pub/plants/release-40/gff3/zea_mays/. The B73 and *C. zeina* sequences were combined into a single reference file to enable the reads to be mapped to both references simultaneously.

The parameters for Trinity were optimized by first assembling the reads that mapped to B73 sequences. The assembly parameters were varied and the assembled transcripts compared to the known B73 proteins. The parameters chosen to vary were normalization (which increases runtime, but slightly lowers full-length transcription reconstruction), k-mer size (where larger k-mers increases specificity, but decreases coverage), jaccard clip (which prevents the assembly of fusion transcripts from gene dense regions), and the group pairs distance (which controls the maximum expected length between fragment pairs). In the default settings the reads are normalized, the k-mer size was 25, jaccard clip was not used and the group pairs distance was 500. The accuracy of each optimization run was determined by using BLAST to compare the coding sequences in the assemblies to known B73 sequences. and calculating the percentage of predicted sequences that was true sequences. A python script determined the number of true sequences by counting the number of predicted proteins that matched with 100% ID over at least 75% of their length. A histogram of the lengths of the sequences predicted from the assemblies and the lengths of their matches in B73 was created in R.

The unmapped reads produced by TopHat2 were converted to FASTA format using FASTX-toolkit (Gordon and Hannon 2010) and *de novo* assembled using Trinity (Haas *et al.* 2013) with the optimised setting of k-mer size to 31 (Appendix C, File C.3). The three replicates from the *C. zeina* challenged leaves were pooled together and the 3 replicates

from the control leaves were pooled during assembly. This resulted in a FASTA output file containing the assembled full-length transcript sequences present in the *C. zeina* challenged leaves and another in the control leaves.

The open reading frames in the Trinity FASTA file were determined using the program TransDecoder (Haas *et al.* 2013) which was developed for use with Trinity. It first identifies all the possible open reading frames above a specified length, in this case 50 amino acids, in the transcripts and then predicts and translates the most likely candidate protein coding regions. The likelihood is determined through a log-likelihood score based on the nucleotide position in the codon and relative frequency in transcript sequences. TransDecoder produced FASTA output files of the possible coding sequences, the most likely coding sequences and of their amino acid translations.

The predicted protein sequences were compared to the NCBI nr protein database using BLASTp with an e-value cut-off of 0.01. This was done to filter out any non-maize transcripts and identify whether any were similar to known plant defence proteins.

Results

After quality filtering the reads were reduced from 16-13 million pairs per sample to 11-8 million pairs per sample (Fig. 2.5). The majority of the reads were mapped to the B73 and *C. zeina* genomes, as expected (Fig. 2.7).

The reads which mapped against the B73 genes were reassembled during the optimization runs (Table 2.2). Not normalizing the reads before assembly resulted in the highest number of B73 sequences produced, but also the highest percentage of spuriously assembled transcripts. Decreasing the k-mer size to 19 both lowered the number of B73 sequences and increased the percentage of spurious transcripts compared to the default settings (Table 2.2). Increasing the k-mer to 31 lowered the amount of B73 sequences, but had the lowest percentage of spuriously assembled transcripts (Table 2.2). Changing the k-mer size to 31 was chosen as the optimal parameter based on the increased accuracy. Both using the jaccard clip setting and increasing the group pairs distance slightly decreased the number of B73 sequences and increased the percentage of spurious transcripts (Table 2.2). The proteins predicted from each of the optimization assemblies were shorter than the matching B73 proteins, but not normalizing the reads and using jaccard clip both increased the amount of fragmented sequences (Fig. 2.6).

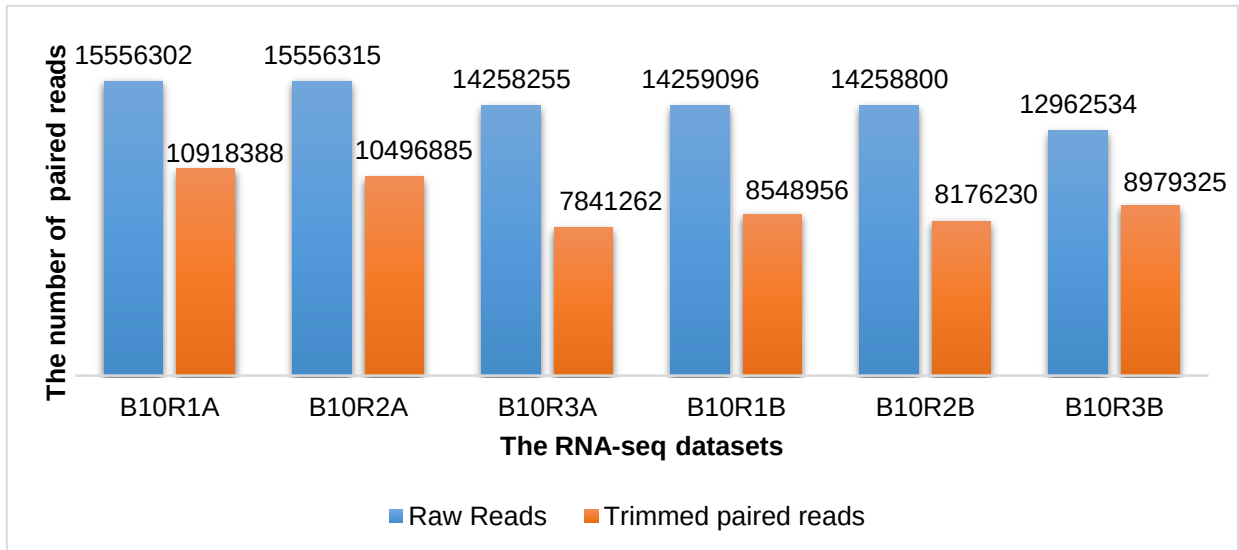


Figure 2.5. The number of raw reads prior to trimming and the high-quality reads retained after trimming for each sample.

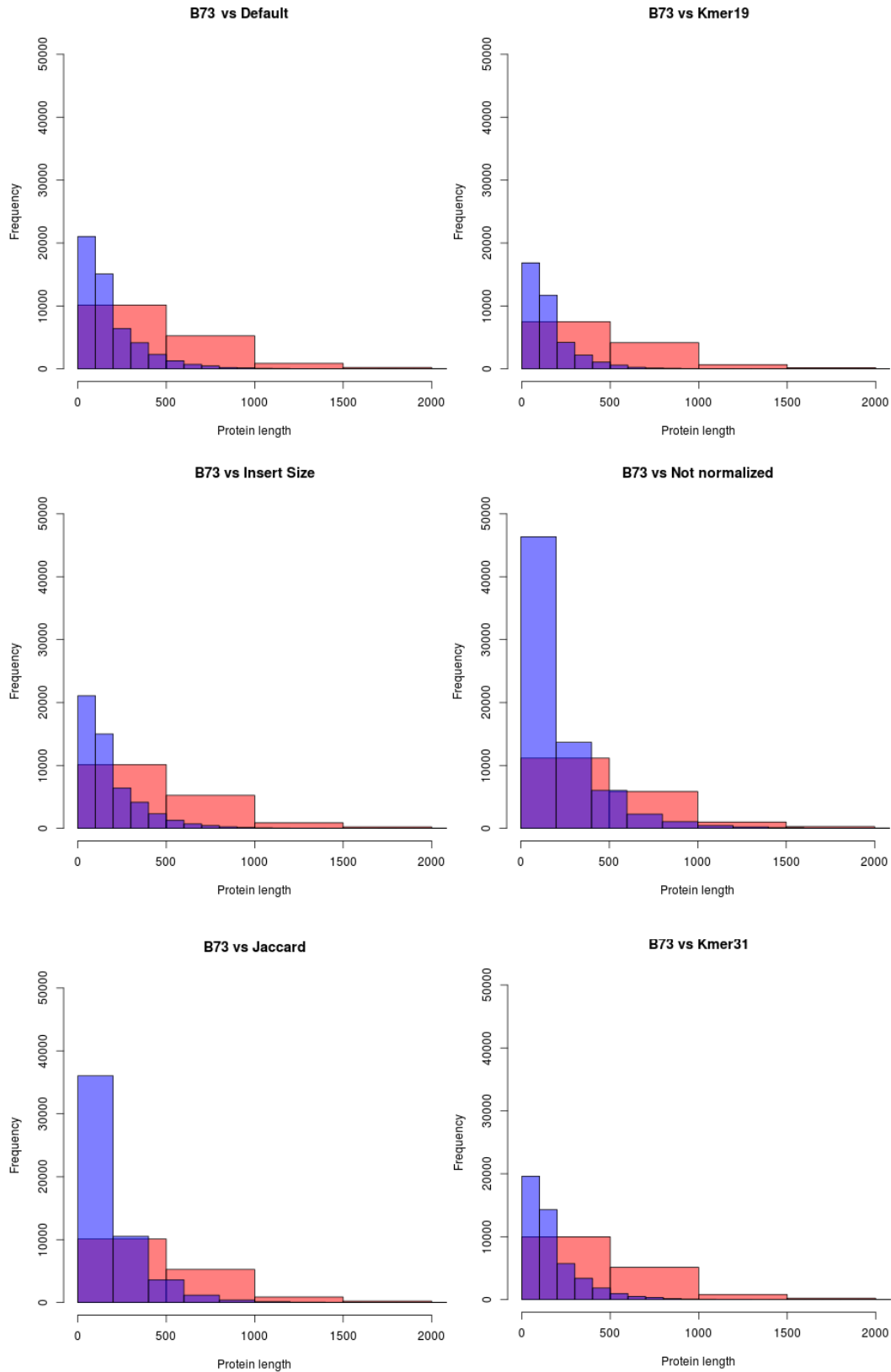


Figure 2.6. A comparison of the lengths of the proteins produced from each parameter (Blue) against the lengths of the matching B73 protein (Pink). Figure created in R.

Table 2.1. The RNA-seq reads produced from the B73-QTL leaf samples.

Sample Name ^a	Leaf ^b	Rep ^c	Insert size (bp) ^d	Read (bp) ^e	length	Clean Reads ^f	Clean bases ^g	Q20(%) ^h	GC(%) ⁱ
B10R1A	Upper	1	200	90		27869614	2508265260	97.51	53.71
B10R1B	Lower	1	200	90		25479860	2293187400	96.82	54.26
B10R2A	Upper	2	200	90		27306096	2457548640	97.37	54.00
B10R2B	Lower	2	200	90		25251354	2272621860	96.29	54.90
B10R3A	Upper	3	200	90		25058218	2255239620	96.21	53.76
B10R3B	Lower	3	200	90		24194948	2177545320	97.56	54.51

a. The name of the RNA-seq sample.

b. Indicates whether the upper or lower leaf of the maize was sequenced.

c. Indicates from which replicate maize plant the leaf was taken.

d. The length of the sequencing fragment.

e. The length of the total reads.

f. The number of total reads after removing adapter sequences, and contaminating and low-quality reads.

g. The number of total clean bases.

h. The percentage of nucleotides with quality higher than 20.

i. The percentage of nucleotides that are G/C.

Table 2.2. The accuracy obtained from each Trinity optimization run.

Varied parameters	Transcripts	Predicted proteins	Hits against B73	B73 Hits
				Predicted %
Default	61,046	52,063	30,372	58.34%
Not Normalized	87,006	70,298	34,279	48.76%
Kmer = 19	50,483	37,239	20,678	55.53%
Kmer = 31	55,047	47,108	29,194	61.97%
Jaccard clip	60,974	52,024	30,298	58.24%
Insert size = 300	60,934	52,095	30,350	58.26%
Kmer = 31 and Jaccard clip	60,786	52,145	30,360	58.22%

The varying parameters are highlighted in yellow. The k-mer sizes used are 19 and 31 (the largest allowed by Trinity is 32 and odd k-mers cannot be self-complementary). Jaccard clip is only worthwhile with high gene density such as in fungal genomes and might help distinguish remaining fungal transcripts. The insert size of the sequenced reads should be included by a group pair's distance of 300 bp.

- The number of transcripts de novo assembled.
- The proteins predicted from the transcripts.
- The number of proteins which matched a genuine B73 sequence.
- The percentage of the proteins produced which were genuine sequences.

In the optimised *de novo* assembly on the unmapped reads there were 5,938 transcripts assembled in the pooled *C. zeina* challenged leaves, 4,403 which were predicted to have protein coding sequences (Fig. 2.7). There were 927 transcripts assembled in the control leaves, 696 of which were predicted to have protein coding sequences (Fig. 2.7). The RNA-seq data came from a field trial, however. The plants were likely exposed to other fungi and microorganisms present in the environment and additional filtering was required to exclude them from the *de novo* assembled transcripts. Comparing the predicted proteins against the NCBI nr protein database allowed the sequences from fungi and microorganisms to be identified and removed. The fungal sequences (Fig. 2.8) were identified mainly as *E. turcicum* (Table A.2), with some other fungal genera also present (Fig. A.1). The potential candidates were narrowed down to 95 plant proteins in the *C. zeina* challenged leaves and 128 in the control leaves (Fig. 2.8). The sequences that were neither plant nor fungal were identified as plant viruses (Table A.3), plant bacteria (Table A.4) or not plant pathogens (Table A.5-6). The BLAST search identified 93% of the predicted proteins in the *C. zeina* challenged leaves and 59% of the predicted proteins in the control leaves.

Many of the plant proteins were fragments where more than one matched different portions of the same gene and the actual amount of total genes present is lower. There was a set of 74 plant proteins present in the *C. zeina* challenged leaves and 99 in the control leaves (Table 2.3). There were 17 proteins that were expressed in both leaf conditions, but the remainder was specific to one leaf condition (Table 2.3).

Table 2.3. The set of plant proteins present in the unmapped assembly.

Assembled transcript	Description of closest match	Closest match	%ID	Alignment Length	Mis-matches	Gap Openings	E-value	Leaf
TRINITY_DN554_c0_g1_i1 m.144	1-aminocyclopropane-1-carboxylate oxidase homolog [Zea mays]	ref XP_020400957.1	95	63	3	0	3E-33	Control
TRINITY_DN1938_c0_g1_i1 m.11440	36.4 kDa proline-rich protein-like precursor [Zea mays]	ref NP_001288371.1	98	51	1	0	2E-24	C. zeina challenged
TRINITY_DN816_c0_g1_i2 m.7181	40S ribosomal protein S3-1, partial [Noccaea caerulescens]	gb JAU88929.1	74	94	24	0	1E-44	C. zeina challenged
TRINITY_DN79_c0_g1_i1 m.1012	40S Ribosomal protein SA family protein [Klebsormidium flaccidum]	dbj GAQ80485.1	80	101	20	0	5E-51	Control
TRINITY_DN3392_c0_g1_i1 m.1465	60S ribosomal protein L7A family protein [Klebsormidium flaccidum]	dbj GAQ82738.1	49	108	54	1	8E-31	C. zeina challenged
TRINITY_DN3515_c0_g1_i1 m.520	60S ribosomal protein L8 [Anthurium amnicola]	gb JAT43573.1	83	121	21	0	3E-65	C. zeina challenged
TRINITY_DN603_c0_g1_i1 m.13313	abscisic stress ripening protein 1 [Zea mays]	gb ACG35620.1	100	103	0	0	3E-60	C. zeina challenged
TRINITY_DN2458_c0_g1_i1 m.6152	Acid phosphatase/vanadium-dependent haloperoxidase-related protein [Zea mays]	gb AQK91329.1	95	22	1	0	1E-02	C. zeina challenged
TRINITY_DN856_c0_g1_i2 m.7207	actin 5-like protein, partial [Miscanthus sinensis]	gb AFC88820.1	99	138	2	0	6E-95	C. zeina challenged
TRINITY_DN3042_c0_g1_i1 m.12256	Actin, non-muscle 6.2, partial [Noccaea caerulescens]	gb JAU73821.1	97	76	2	0	1E-47	C. zeina challenged
TRINITY_DN856_c0_g1_i1 m.7202	Actin, plasmodial isoform, partial [Anthurium amnicola]	gb JAT65395.1	98	180	4	0	2E-128	C. zeina challenged
TRINITY_DN781_c0_g1_i1 m.8139	actin-like protein, partial [Miscanthus sinensis]	gb AFC88819.1	100	62	0	0	2E-36	C. zeina challenged
TRINITY_DN850_c0_g1_i1 m.582	adenylate kinase [Zea mays]	gb ACG27659.1	93	30	2	0	3E-07	Control
TRINITY_DN1869_c0_g1_i1 m.6706	amidase precursor [Zea mays]	ref NP_001150232.2	100	72	0	0	5E-38	Both
TRINITY_DN5708_c0_g1_i1 m.12112	AN12 Zinc finger A20 and AN1 domain-containing protein [Zea mays]	ref NP_001106259.1	100	87	0	0	7E-53	C. zeina challenged
TRINITY_DN6_c0_g1_i1 m.1023	anthranilate synthase component II [Zea mays]	ref NP_001152611.1	100	160	0	0	7E-114	Control
TRINITY_DN3476_c0_g1_i1 m.4622	BnaCnng41770D [Brassica napus]	emb CDY63262.1	39	90	51	1	3E-11	C. zeina challenged

Assembled transcript	Description of closest match	Closest match	%ID	Alignment Length	Mis-matches	Gap Openings	E-value	Leaf
TRINITY_DN2238_c0_g1_i1 m.8564	cysteine proteinase 1 precursor [Zea mays]	ref NP_001105685.1	100	79	0	0	2E-47	C. zeina challenged
TRINITY_DN532_c0_g1_i1 m.3215	dehydrin 13 [Zea mays]	gb ACG47851.1	100	56	0	0	2E-23	C. zeina challenged
TRINITY_DN437_c0_g1_i1 m.202	E3 ubiquitin-protein ligase RGLG1 [Zea mays]	gb ONM09641.1	60	53	18	2	6E-07	Control
TRINITY_DN479_c0_g1_i1 m.221	expressed protein [Oryza sativa Japonica Group]	gb ABA97426.1	60	77	26	3	3E-13	Control
TRINITY_DN281_c0_g1_i1 m.11883	ferredoxin-binding subunit [Zea mays]	gb AAF67002.1 AF147726_1	99	86	1	0	3E-56	C. zeina challenged
TRINITY_DN799_c0_g1_i1 m.7975	Glucose-repressible gene protein, partial [Anthurium amnicola]	gb JAT62563.1	97	68	2	0	3E-36	C. zeina challenged
TRINITY_DN3012_c0_g1_i1 m.12378	glutathione transferase20 [Zea mays]	ref NP_001104986.1	100	140	0	0	9E-96	C. zeina challenged
TRINITY_DN2453_c0_g1_i1 m.6089	Glyceraldehyde-3-phosphate dehydrogenase 2, cytosolic [Zea mays]	gb AQK78434.1	100	70	0	0	4E-40	C. zeina challenged
TRINITY_DN658_c0_g1_i1 m.2190	glycine-rich protein [Zea mays]	gb AQK78348.1	96	82	3	0	1E-46	Control
TRINITY_DN5596_c0_g1_i1 m.1938	glycine-rich RNA-binding protein 2 [Zea mays]	gb ACG26657.1	92	61	5	0	2E-31	C. zeina challenged
TRINITY_DN3882_c0_g1_i1 m.15432	G-type lectin S-receptor-like serine/threonine-protein kinase B120 [Zea mays]	gb AQK46433.1	86	188	27	0	1E-110	Both
TRINITY_DN2684_c0_g1_i1 m.15624	high-affinity potassium transporter [Zea mays]	gb AHH35062.1	94	71	3	1	2E-29	Both
TRINITY_DN6010_c0_g1_i1 m.3269	Histone H2AX [Zea mays]	gb AQL03536.1	84	49	8	0	2E-18	C. zeina challenged
TRINITY_DN6061_c0_g1_i1 m.3273	hypothetical protein [Zea mays]	gb ACG30731.1	100	44	0	0	7E-09	C. zeina challenged
TRINITY_DN2934_c0_g1_i1 m.13016	hypothetical protein GLYMA_16G028800 [Glycine max]	gb KRH06533.1	64	70	25	0	6E-22	C. zeina challenged
TRINITY_DN370_c0_g1_i1 m.9590	hypothetical protein Osl_11203 [Oryza sativa Indica Group]	gb EAY89670.1	65	26	9	0	2E-03	C. zeina challenged
TRINITY_DN4069_c0_g1_i1 m.5199	hypothetical protein SORBI_003G029600 [Sorghum bicolor]	gb KXG31609.1	68	50	16	0	3E-14	C. zeina challenged

Assembled transcript	Description of closest match	Closest match	%ID	Alignment Length	Mis-matches	Gap Openings	E-value	Leaf
TRINITY_DN1629_c0_g1_i1 m.8841	hypothetical protein SORBIDRAFT_01g037483 [Sorghum bicolor]	ref XP_002467980.1	97	624	17	0	0E+00	C. zeina challenged
TRINITY_DN5601_c0_g1_i1 m.4426	hypothetical protein SORBIDRAFT_03g041490 [Sorghum bicolor]	ref XP_002458852.1	89	122	13	0	2E-69	C. zeina challenged
TRINITY_DN759_c0_g1_i1 m.7990	hypothetical protein SORBIDRAFT_06g001580 [Sorghum bicolor]	ref XP_002446089.1	86	163	23	0	1E-98	C. zeina challenged
TRINITY_DN1098_c0_g1_i1 m.9469	IDD9 [Zea mays subsp. mays]	gb AJF23770.1	98	85	0	2	2E-40	Both
TRINITY_DN890_c0_g1_i1 m.549	KH domain-containing protein [Zea mays]	ref NP_001332948.1	100	184	0	0	3E-126	Control
TRINITY_DN813_c0_g1_i1 m.443	mitochondrial protein - maize	pir A24522	100	81	0	0	5E-49	Control
TRINITY_DN644_c0_g1_i1 m.2102	O-fucosyltransferase family protein [Zea mays]	gb ONM55849.1	66	47	16	0	2E-10	Control
TRINITY_DN220_c0_g1_i1 m.1797	Os03g0718150, partial [Oryza sativa Japonica Group]	dbj BAS86095.1	72	96	27	0	2E-36	Control
TRINITY_DN171_c0_g1_i1 m.10226	Os03g0801600, partial [Oryza sativa Japonica Group]	dbj BAS86895.1	71	75	17	3	4E-22	C. zeina challenged
TRINITY_DN527_c0_g1_i1 m.56	Os04g0633950 [Oryza sativa Japonica Group]	dbj BAS91198.1	51	80	37	1	1E-07	Control
TRINITY_DN152_c0_g1_i1 m.1352	Os05g0438700, partial [Oryza sativa Japonica Group]	dbj BAS94223.1	65	62	22	0	6E-09	Control
TRINITY_DN905_c0_g1_i1 m.1708	Os07g0102150, partial [Oryza sativa Japonica Group]	dbj BAS99686.1	60	82	30	2	7E-13	Control
TRINITY_DN589_c0_g1_i1 m.123	pentatricopeptide repeat*-59040 [Zea mays]	gb AQK40903.1	87	98	6	3	1E-29	Control
TRINITY_DN4529_c0_g1_i1 m.3827	Polyubiquitin 11 [Zostera marina]	gb KMZ73934.1	100	93	0	0	2E-57	C. zeina challenged
TRINITY_DN4187_c0_g1_i1 m.14936	Polyubiquitin, partial [Noccaea caerulescens]	gb JAU22157.1	98	94	2	0	1E-57	C. zeina challenged
TRINITY_DN1294_c0_g1_i1 m.299	predicted protein [Hordeum vulgare subsp. vulgare]	dbj BAJ93702.1	84	74	12	0	3E-34	C. zeina challenged
TRINITY_DN3787_c0_g1_i1 m.10949	predicted protein [Hordeum vulgare subsp. vulgare]	dbj BAJ94184.1	88	73	9	0	1E-41	C. zeina challenged

Assembled transcript	Description of closest match	Closest match	%ID	Alignment Length	Mis-matches	Gap Openings	E-value	Leaf
TRINITY_DN1949_c0_g1_i1 m.11446	predicted protein [Hordeum vulgare subsp. vulgare]	dbj BAJ96087.1	99	83	1	0	6E-50	C. zeina challenged
TRINITY_DN227_c0_g1_i1 m.11782	predicted protein [Hordeum vulgare subsp. vulgare]	dbj BAJ96848.1	89	79	9	0	1E-41	C. zeina challenged
TRINITY_DN245_c0_g1_i1 m.1772	predicted protein [Hordeum vulgare subsp. vulgare]	dbj BAK00513.1	83	65	11	0	5E-28	Control
TRINITY_DN867_c0_g1_i3 m.6820	predicted protein [Hordeum vulgare subsp. vulgare]	dbj BAK00960.1	100	272	1	0	0E+00	C. zeina challenged
TRINITY_DN6038_c0_g1_i1 m.3431	predicted protein [Hordeum vulgare subsp. vulgare]	dbj BAK02581.1	74	61	14	1	2E-21	C. zeina challenged
TRINITY_DN4288_c0_g1_i1 m.15167	predicted protein [Hordeum vulgare subsp. vulgare]	dbj BAK05398.1	92	75	6	0	6E-42	C. zeina challenged
TRINITY_DN2649_c0_g1_i1 m.15460	predicted protein [Micromonas commoda]	ref XP_002508800.1	73	64	17	0	2E-26	C. zeina challenged
TRINITY_DN3940_c0_g1_i1 m.12582	PREDICTED: 40S ribosomal protein S20-1 [Oryza sativa Japonica Group]	ref XP_015614818.1	100	70	0	0	1E-41	Both
TRINITY_DN462_c0_g1_i1 m.4794	PREDICTED: 60S ribosomal protein L4-1-like [Erythranthe guttata]	ref XP_012831933.1	80	116	22	1	2E-58	C. zeina challenged
TRINITY_DN569_c0_g1_i1 m.3141	PREDICTED: chlorophyll a-b binding protein 1D [Glycine max]	ref XP_006587015.1	99	74	1	0	2E-44	C. zeina challenged
TRINITY_DN5470_c0_g1_i1 m.13534	PREDICTED: fructokinase-1-like [Phoenix dactylifera]	ref XP_008790605.1	40	78	25	1	6E-06	C. zeina challenged
TRINITY_DN3805_c0_g1_i1 m.15236	PREDICTED: laccase-2-like [Pyrus x bretschneideri]	ref XP_009350642.1	70	124	36	1	2E-47	C. zeina challenged
TRINITY_DN4167_c0_g1_i1 m.14800	PREDICTED: probable cysteine protease RD21B [Nicotiana attenuata]	ref XP_019239500.1	38	78	43	3	2E-05	C. zeina challenged
TRINITY_DN927_c0_g1_i1 m.1645	PREDICTED: serine hydroxymethyltransferase, mitochondrial-like [Malus domestica]	ref XP_008384294.1	98	99	2	0	4E-65	Control
TRINITY_DN485_c0_g1_i2 m.4943	PREDICTED: ubiquitin [Brassica napus]	ref XP_013745166.1	99	83	1	0	3E-50	C. zeina challenged
TRINITY_DN3639_c0_g1_i1 m.14696	PREDICTED: ubiquitin [Elaeis guineensis]	ref XP_019708906.1	100	68	0	0	5E-40	C. zeina challenged
TRINITY_DN5689_c0_g1_i1 m.4214	prenylated Rab receptor 2 [Zea mays]	ref NP_001151719.1	100	73	0	0	4E-38	C. zeina challenged
TRINITY_DN356_c0_g1_i1 m.1132	probable E3 ubiquitin-protein ligase ARI8 [Zea mays]	ref NP_001315295.1	100	83	0	0	2E-46	Control

Assembled transcript	Description of closest match	Closest match	%ID	Alignment Length	Mis-matches	Gap Openings	E-value	Leaf
TRINITY_DN827_c0_g1_i1 m.591	Protein ALTERED XYLOGLUCAN 4-like [Dichantherium oligosanthes]	gb OEL28198.1	88	43	5	0	2E-18	Control
TRINITY_DN809_c0_g1_i1 m.577	Protein FATTY ACID EXPORT 1 chloroplastic [Zea mays]	gb AQK44205.1	94	68	4	0	8E-34	Control
TRINITY_DN915_c0_g1_i1 m.1634	Protein NRT1/ PTR FAMILY 2.13 [Dichantherium oligosanthes]	gb OEL16316.1	71	35	10	0	9E-05	Control
TRINITY_DN436_c0_g1_i1 m.373	Protein S-acyltransferase 10 [Zea mays]	gb AQK54205.1	42	89	43	2	6E-08	Control
TRINITY_DN5447_c0_g1_i1 m.13539	Putative leucine-rich repeat receptor-like protein kinase family protein [Zea mays]	gb AQK46158.1	89	74	8	0	4E-40	C. zeina challenged
TRINITY_DN305_c0_g1_i1 m.1066	Putative leucine-rich repeat receptor-like protein kinase family protein [Zea mays]	gb AQK46176.1	84	56	9	0	8E-27	Control
TRINITY_DN926_c0_g1_i1 m.1572	Putative metacaspase family protein, partial [Zea mays]	gb AQL05567.1	76	29	7	0	3E-04	Control
TRINITY_DN807_c0_g1_i1 m.512	putative polyprotein [Zea mays]	gb AAL75999.1 AF466646_7	61	89	30	1	2E-27	Control
TRINITY_DN886_c0_g1_i1 m.460	putative protein kinase superfamily protein [Zea mays]	ref NP_001146603.1	100	106	0	0	3E-69	Control
TRINITY_DN696_c0_g1_i1 m.13285	pyruvate orthophosphate dikinase1 [Zea mays]	gb AQK85875.1	100	128	0	0	2E-84	C. zeina challenged
TRINITY_DN333_c0_g1_i1 m.9609	pyruvate orthophosphate dikinase1 [Zea mays]	gb AQK85899.1	100	63	0	0	2E-35	C. zeina challenged
TRINITY_DN2609_c0_g1_i1 m.15465	Receptor of activated protein kinase C 1A, component of 40S small ribosomal subunit [Ostreococcus lucimarinus CCE9901]	ref XP_001421096.1	80	84	13	1	1E-35	C. zeina challenged
TRINITY_DN1912_c0_g1_i1 m.11277	Receptor-like serine/threonine-protein kinase SD1-8 [Zea mays]	gb ONM13388.1	88	98	12	0	2E-53	C. zeina challenged
TRINITY_DN681_c0_g1_i1 m.2245	Receptor-like serine/threonine-protein kinase SD1-8 [Zea mays]	gb ONM13390.1	91	109	6	3	1E-53	Control
TRINITY_DN5521_c0_g1_i1 m.2036	RecName: Full=Ferredoxin-5, chloroplastic; AltName: Full=Ferredoxin V; Short=Fd V; Flags: Precursor	sp P27789.1 FER5_MAIZE	100	58	0	0	4E-32	C. zeina challenged

Assembled transcript	Description of closest match	Closest match	%ID	Alignment Length	Mis-matches	Gap Openings	E-value	Leaf
TRINITY_DN703_c0_g1_i1 m.745	ribosomal protein S9 component of cytosolic 80S ribosome and 40S small subunit [Coccomyxa subellipsoidea C-169]	ref XP_005648482.1	84	95	15	0	3E-48	Control
TRINITY_DN303_c0_g1_i1 m.1121	RNA polymerase sigma factor 1, partial [Zea mays]	gb AAC78467.1	100	62	0	0	7E-32	Control
TRINITY_DN75_c0_g1_i1 m.1047	salivary acidic proline-rich phosphoprotein 1/2 [Zea mays]	ref XP_008650711.1	100	73	0	0	7E-41	Control
TRINITY_DN5514_c0_g1_i1 m.1876	thiamine thiazole synthase 1, chloroplastic precursor [Zea mays]	ref NP_001105696.1	100	73	0	0	4E-42	C. zeina challenged
TRINITY_DN768_c0_g1_i1 m.779	Transcription factor bHLH68 [Zea mays]	gb AQK44493.1	72	89	6	2	3E-28	Control
TRINITY_DN240_c0_g1_i1 m.11825	ubiquitin-like protein, partial [Vanilla planifolia]	gb AEI29169.1	99	73	1	0	9E-43	C. zeina challenged
TRINITY_DN294_c0_g1_i1 m.11835	uncharacterized protein LOC100126943 [Zea mays]	ref NP_001295443.1	100	325	0	0	0E+00	C. zeina challenged
TRINITY_DN1791_c0_g1_i1 m.7808	uncharacterized protein LOC100193213 [Zea mays]	ref NP_001131838.1	100	241	0	0	2E-172	C. zeina challenged
TRINITY_DN1718_c0_g1_i1 m.7714	uncharacterized protein LOC100193608 [Zea mays]	ref NP_001132182.1	100	228	0	0	5E-166	C. zeina challenged
TRINITY_DN6107_c0_g1_i1 m.13205	uncharacterized protein LOC100193614 [Zea mays]	ref NP_001132187.1	97	77	2	0	1E-46	C. zeina challenged
TRINITY_DN1165_c0_g1_i1 m.9755	uncharacterized protein LOC100273155 [Zea mays]	ref NP_001141073.1	100	96	0	0	3E-59	C. zeina challenged
TRINITY_DN1384_c0_g1_i1 m.10091	uncharacterized protein LOC100279124 [Zea mays]	ref NP_001145637.2	100	80	0	0	4E-51	C. zeina challenged
TRINITY_DN4394_c0_g1_i1 m.6224	uncharacterized protein LOC100383110 [Zea mays]	ref NP_001169249.1	100	99	0	0	8E-59	C. zeina challenged
TRINITY_DN512_c0_g1_i1 m.3191	uncharacterized protein LOC103638366 [Zea mays]	ref XP_008659520.1	72	47	13	0	2E-10	C. zeina challenged
TRINITY_DN109_c0_g1_i1 m.10238	unknown [Medicago truncatula]	gb ACJ84693.1	77	79	18	0	6E-37	C. zeina challenged
TRINITY_DN3906_c0_g1_i1 m.12547	unknown [Zea mays]	gb ACF82437.1	100	64	0	0	1E-36	C. zeina challenged
TRINITY_DN546_c0_g1_i2 m.3118	unknown [Zea mays]	gb ACF84747.1	100	253	0	0	0E+00	Both

Assembled transcript	Description of closest match	Closest match	%ID	Alignment Length	Mis-matches	Gap Openings	E-value	Leaf
TRINITY_DN2427_c0_g1_i1 m.6138	unknown [Zea mays]	gb ACF85568.1	100	97	0	0	9E-61	Both
TRINITY_DN932_c0_g1_i1 m.1532	unknown [Zea mays]	gb ACL54350.1	100	107	0	0	1E-64	Control
TRINITY_DN280_c0_g1_i1 m.1762	unknown [Zea mays]	gb ACL54835.1	100	119	0	0	1E-79	Control
TRINITY_DN172_c0_g1_i2 m.1520	unknown [Zea mays]	gb ACN26848.1	100	23	0	0	2E-05	Control
TRINITY_DN617_c0_g1_i1 m.13296	unknown [Zea mays]	gb ACN29187.1	100	294	0	0	0E+00	Both
TRINITY_DN3312_c0_g1_i1 m.1388	unknown [Zea mays]	gb ACN33910.1	100	240	0	0	3E-169	Both
TRINITY_DN388_c0_g1_i1 m.1110	unknown [Zea mays]	gb ACN35719.1	84	101	7	2	2E-40	Control
TRINITY_DN6010_c0_g1_i1 m.3270	unknown, partial [Glycine max]	gb ACU16929.1	50	56	27	1	4E-08	C. zeina challenged
TRINITY_DN374_c0_g1_i1 m.1081	unnamed protein product [Ostreococcus tauri]	ref XP_003075092.1	43	81	46	0	4E-08	Control
TRINITY_DN299_c0_g1_i1 m.1855	VQ motif family protein [Zea mays]	gb ACG44706.1	100	113	0	0	4E-74	Control

In cases with multiple matched against the same reference sequence only the match with the lowest e-value is shown. All of the sequences in the C. zeina challenged leaves and sequences unique to the control leaves are shown.

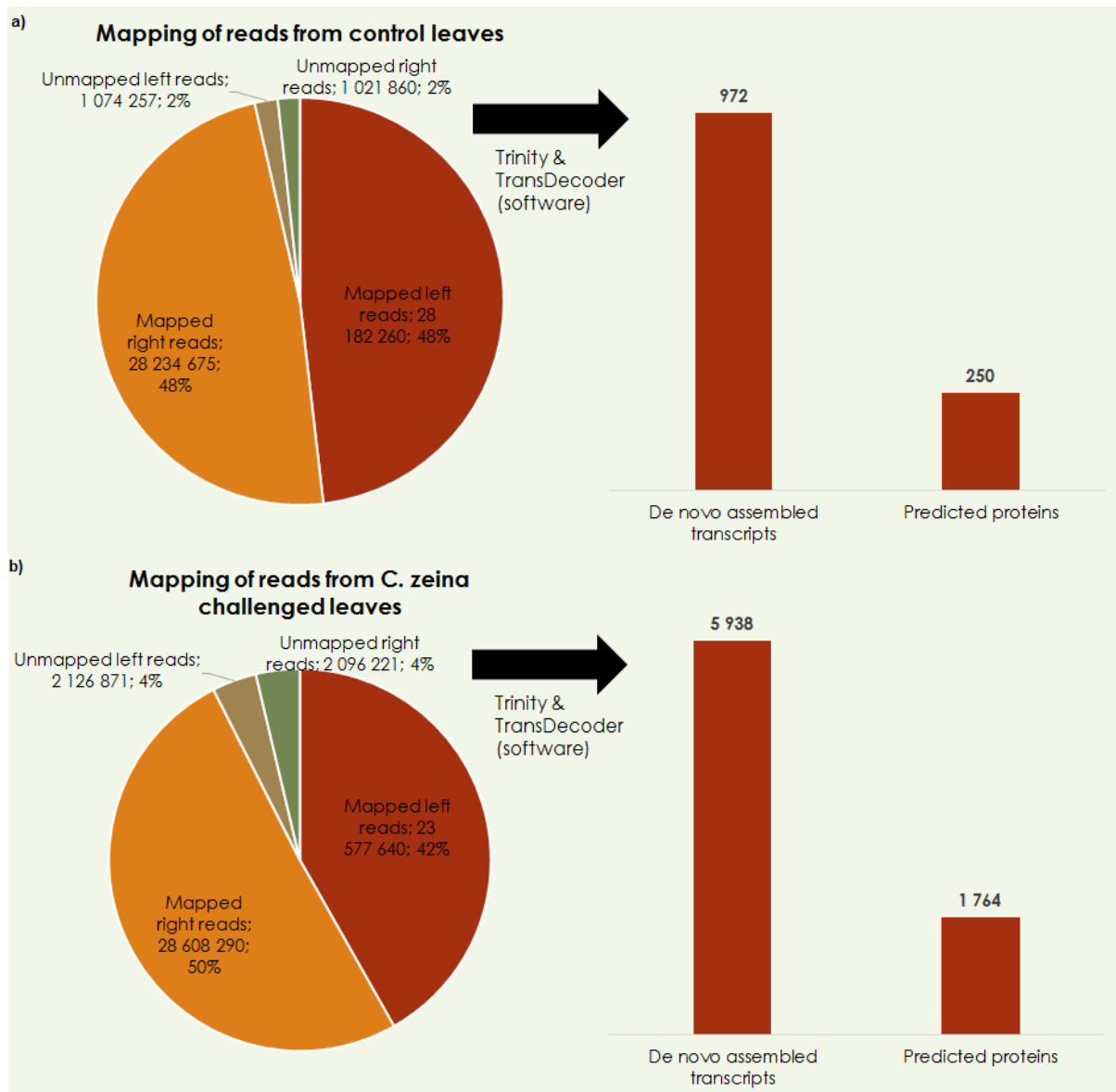


Figure 2.7. Mapping and de novo assembly of the reads. a) The percentage of reads from the control replicates that were unmapped and the number of transcripts assembled and protein coding sequences predicted from the unmapped reads. b) The percentage of reads from the *C. zeina* challenged replicates that were unmapped and the number of transcripts assembled and protein coding sequences predicted from the unmapped reads.

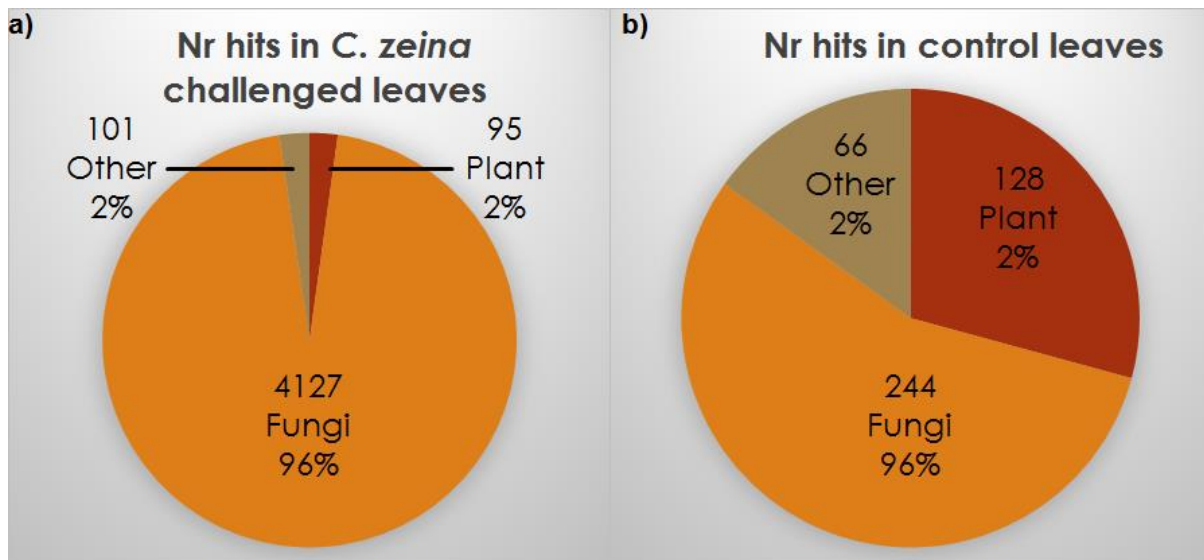


Figure 2.8. Identification of the predicted proteins through BLASTp against nr. a) The number of plant, fungal or other proteins present in the *C. zeina* challenged leaves. b) The number of plant, fungal or other proteins present in the control leaves.

Discussion

The majority of reads mapped against the B73 and *C. zeina* reference genomes, as expected. There were many more transcripts assembled in the *C. zeina* challenged leaves than in the control leaves (Fig. 2.7). The disparity was caused through an increase in fungal sequences present (Fig. 2.8). The most numerous was *Exserohilum turcicum* (*Setosphaeria turcica*) (Fig A1). This fungus causes the maize foliar disease northern corn leaf blight and is also wide-spread in South Africa (Kloppers and Tweer 2009) and could have been present in the field trial. It is also likely that *C. zeina* sequences which the draft reference genome was not comprehensive enough to include matched against conserved regions in the other fungi.

Based on the positions of the SSR markers the QTL region is approximately in the 3,000 kb range. Using the average gene density in maize of 18.7 per 1000 kb (Jiao *et al.* 2017) an estimate of 56 genes are expected to be in this region. The number of plant genes (Table 2.3) exclude genes not expressed under the experimental conditions and genes shared with B73. This puts the amount of genes produced by the pipeline in a realistic range. The *de novo* assembly of the reads that did not match B73 sequences successfully revealed the presence of non-B73 genes present in B73-QTL.

The plant genes identified included some that were shared between the two conditions, but the majority were only expressed in one of the conditions (Table 2.3). This pipeline only detected which novel genes were present in each condition however, not their expression. An analysis of the differential expression of the transcripts between the two conditions is needed to gain insight into the response of the potential candidate genes to the infection. It

is likely that the gene conferring the GLS resistance would be expressed in response to *C. zeina*, but it could also be a constitutive gene.

Chapter 3: Differential expression analysis and annotation of candidate genes

T. Welgemoed, R. Pierneef, L. Sterck, Y. Van de Peer and D.K. Berger

Author Contributions

T.W.: Conducted the bioinformatics and data analyses. Wrote the chapter.

R.P.: Guided the bioinformatics analyses, suggested methods. Reviewed chapter and suggested improvements.

D.K.B.: Principal Investigator of the funding project, acquired funding, provided day-to-day institutional supervision. Reviewed chapter and suggested improvements. Led discussions on data interpretation.

L.S (Bioinformatics and Systems Biology, VIB-UGent Center for Plant Systems Biology): Guided the bioinformatics analyses and suggested methods.

Y.VdP: Suggested methods. Hosts the Bioinformatics and Systems Biology, VIB-UGent Center for Plant Systems Biology where the bioinformatics analyses were carried out.

D.K.B: designed the study.

Introduction

There are many examples in maize specifically for using RNA-seq data for gene discovery or differential expression analysis. Kakumanu *et al.* (2012) identified genes and pathways involved in embryo abortion in maize caused by drought stress using RNA-seq. Chettoor *et al.* (2014) used RNA-seq to find differentially expressed genome features between male and female gametophytes. Regulski *et al.* (2013) studied epigenetics in two inbred maize lines to create a genome map of the methylation sites and the effect on alternatively spliced transcripts. Li *et al.* (2010) examined which genes were expressed in maize leaves at different stages during development of photosynthetic differentiation.

The regulation of the expression of several categories of plant antifungal defence genes play an important role in the defence response. For example, defensins are induced in plants challenged by fungal infection and many transgenic plants that overexpress a defensin have increased resistance to pathogenic fungi or oomycetes (De Coninck *et al.* 2013; Lacerda *et al.* 2014). The transgenic examples included a maize plant that overexpressed the defensin ZmDEF1 to improve resistance to the oomycete *Phytophthora parasitica* (De Coninck *et al.* 2013). Another example, thaumatin-like proteins, have long been established to have an

antifungal activity (Vigers *et al.* 1992) and are induced in maize challenged by pathogenic fungi (Shu *et al.* 2017). Some lectin receptor kinases are involved in pathogen recognition (Singh and Zimmerli 2013) and are upregulated in response to fungi (Sanabria *et al.* 2012). Chitinases are also directly involved in attacking fungi and are induced by fungal infection (Shu *et al.* 2017; Balasubramanian *et al.* 2012).

It is therefore likely that a gene involved in defence would most likely be induced by *C. zeina* infection. A candidate gene upregulated in the *C. zeina* challenged leaves compared to the control leaves thus have evidence for it playing a role in the defence pathway. For this reason, the regulation of the genes between the two conditions were investigated to provide support for potential candidates. It is also possible however, that the gene responsible for the QTL effect could potentially be a gene that is constantly active. For example, the lectin Pi-d2 in rice confers resistance against the fungal pathogen *Magnaporthe grisea*, but is constitutively expressed (Chen *et al.* 2006). Another example indicated that maize rachis tissue can limit the spread of the fungal pathogen *Aspergillus flavus* with constitutively expressed antimicrobial phenolics (Pechanova *et al.* 2011). It could also be a purely structural gene that causes the QTL effect. For instance, the distance between veins in maize leaves is associated with resistance against *C. zeina* (Benson *et al.* 2015). Transcripts that were found to not have significant differential expression were therefore not automatically excluded.

Methods

Plant Material

The same maize leaves as in Chapter 2 were used for this analysis. Briefly, the QTL originates from the subtropical maize line CML444. It was identified to increase GLS resistance in a 2009 field trial in Hildesheim as 10G2_GLS and in another independent 2009 field trial in Baynesfield as 10H_GLS (Berger *et al.*, 2014). This region on chromosome 10, henceforth called QTL10, was backcrossed into the public, susceptible B73 line to produce the near isogenic line B73+QTL10 (B73-QTL). During the backcross breeding four SSR markers (Berger *et al.* 2014) were used where two markers flanked the QTL region (umc1084-10 and bnlg1450-10) and another two were located within the region (umc1196-10 and umc1176-10). A multiplex PCR with primers for these four markers was used to screen progeny for the presence of QTL10. The background was selected for B73 origin using 384 SNP markers across the genome. Three biological replicates of B73-QTL (LGT61807, LGT61819 and LGT61841) and three of B73 (LGT61811, LGT61820 and LGT61837) were RNA-sequenced.

The two lines were planted on the 28th December 2012 at the Hildesheim farm in Greytown in a randomized block design with three replicate blocks of ten plants per row. The field site had relatively high GLS disease pressure due to natural inoculum of *C. zeina*. This field trial was also reported in studies by (Christie *et al.* 2017) and Swart *et al.* (2017), but the current study used the B73-QTL data whereas they used the B73 data.

Samples the size of a CD cover (~30.65 cm²) were cut from the middles of two leaves above and two leaves below the cob from each maize plant in the replicates at 77 days after planting. The leaves below the cob had developed GLS symptoms and the leaves above the cob had not developed symptoms yet. There are therefore two conditions, *C. zeina* challenged and control, and three biological replicates of the conditions.

Data available for project

The RNA-seq data was also the same as used in chapter 2. Briefly, the RNA was extracted from each leaf sample at UP and stored on dry ice. The RNA-seq was done by BGI Tech Solutions Co., Ltd. (Beijing Genome Institute; Hong Kong) using Illumina HiSeq™ 2000 with a 100 bp paired-end module (Illumina Inc., San Diego, USA). This resulted in six datasets of RNA-seq reads consisting of three replicates of the control B73-QTL leaves and three replicates of the *C. zeina* challenged B73-QTL leaves (Table 2.1). In the names of the datasets were indicated the identity of the plant (B10 for B73-QTL), the replicate number of the plant (R1 for replicate 1, etc.) and the condition of the leaf (A for control or B for *C. zeina* challenged). The RNA-seq data from the B73 leaves from (Swart *et al.* 2017) are also available in the Gene Expression Omnibus GEO (GEO) under GSE94442.

Differential expression analysis and annotation

The differential expression analysis compared the expression of the same transcript across the different conditions and thus required that the same transcript is assembled in each assembly. The two separate *de novo* assemblies could not be compared because of this reason and a master assembly of all the trimmed reads in the six datasets (three biological replicates of the *C. zeina* challenged leaves and three biological replicates of the control leaves) was assembled using the same method as the two previous assemblies (Appendix C, File C.3).

The quality trimmed reads of the three replicates of the *C. zeina* challenged leaves (B10R1B, B10R2B and B10R3B) and of the control leaves (B10R1A, B10R2A and B10R3A) were *de novo* assembled together to create the master assembly. The same quality trimmed reads and assembly method as in the bioinformatics pipeline in chapter 2 were used (Appendix C, File C.1). Briefly, the quality trimming was done using Trimmomatic (Bolger *et al.* 2014b) to trim low quality and adaptor sequences from the FASTQ read files. The first 15

bases were also trimmed from the beginning of each sequence to remove the random primer sequences. The low quality bases were trimmed using a sliding window of five bases and trimming the sequence when the average quality in the window dropped below a Q score of 28. The sequences that were trimmed to less than 50 and their paired reads were discarded. The k-mer size was set to 31 and the strand-specificity indicated.

The respective counts of the master assembly transcripts in the datasets were estimated using salmon software (Patro *et al.* 2017), which is a fast transcriptome-wide quantifier that accounts for GC content bias and has high accuracy (Appendix C, File C.4). The FASTA file of *de novo* assembled master assembly transcripts was indexed using the salmon quasi-type index. The quality trimmed reads from each individual replicate was quantified against the index with the quasi-mapping mode of salmon. A table of counts was created from the quantified abundances using the `abundance_estimates_to_matrix.pl` perl script in the Trinity utilities (Haas *et al.* 2013). The table had rows with the name of each of the transcripts in the master assembly and six columns of the abundances of the six datasets.

The abundances of the transcripts in the table of counts were tested for significant differences between the *C. zeina* challenged and control conditions in the R package edgeR (Robinson *et al.* 2010) (Appendix C, File C.5). The default settings and trimmed mean of M-values normalization (TMM) was used. TMM is a normalization used in RNA-seq to adjust the library sizes of the RNA-seq samples to the same scale (Robinson and Oshlack 2010). The dispersion between the transcripts in three biological replicates was estimated using tag-wise dispersion. The results were written to a table showing the name, the log₂ fold change between the control leaves and the *C. zeina* challenged, log counts per million, p-value and false discovery rate (FDR) for each transcript.

The identities of the differentially expressed genes were determined by comparing them to the B73 v.4 (Jiao *et al.* 2017) and *C. zeina* (Wingfield *et al.* 2017) reference genomes using BLAT (Kent 2002). Due to the high amount of *E. turcicum* present in the previous BLAST results its reference genome was also included (Condon *et al.* 2013), obtained from <https://genome.jgi.doe.gov/Settu3/Settu3.download.html>. The differential expression analysis results were split using a custom AWK script based on this identification with a cut-off where the match to B73, *C. zeina* or *E. turcicum* required at least 80% identity across 90% of the transcript's length. The remaining differentially expressed (DE) transcripts which did not match to any of the references were compared against nr using BLASTx. This resulted in a table showing only the results from unique transcripts.

The B73 and unique transcripts were uploaded into the Monocots PLAZA 4.0 workbench (Van Bel *et al.* 2018) for annotation. During the import process the sequences were

compared to the PLAZA maize database using BLAST and the information on the closest match retrieved. Unique transcripts that had matched other plants according to nr were also compared to the PLAZA database for that plant to confirm the maize gene was the closest match. B73 transcripts with more than 95% identity were excluded as originating from the B73 background of B73-QTL. This resulted in a table showing only the results of variants of B73 transcripts that are potential novel alleles. The retrieved information included the protein domains present and the gene ontologies (GO) in the biological process category associated with that record.

A previous study by Swart *et al.* (2017) had examined differential expression in B73 between *C. zeina* infected leaves and control leaves. The leaves were harvested from the same field trial as the current study and were RNA-sequenced along with the B73-QTL leaves. The B73 RNA-seq datasets were deposited in GEO under GSE94442 and their differential expression analysis results were deposited as a supplementary file under this accession. A list of names of the B73 genes which had significant ($FDR < 0.05$) differential expression were copied from this file and compared to the differentially expressed genes identified in B73-QTL. This enabled the identification of gene regulation unique to the B73-QTL.

Swart *et al.* (2017) had however used the B73 v.3 reference genome, which used an older naming system for the maize genes. The current study used the v.4 reference genome and the genes in the results were named according to the new naming system. This required the names to be converted to the same system in order to be able to compare the two studies. A custom python script converted the list of the DE genes identified by Swart *et al.* (2017) to their equivalent names in v.4 based on a reference document from MaizeGDB (Andorf *et al.* 2016) that gave the v.3 and v.4 names for each gene. The script then compared the list of the converted names from Swart *et al.* (2017) to the list of DE genes identified in B73-QTL to determine which were shared between B73 and B73-QTL.

The master assembly transcripts were compared to the two unmapped assemblies using BLAST to determine which of these also originated from the pool of unmapped reads. The blast results were filtered with a python script and hits that had over 90% identity across over 90% of its length were taken to be the same sequence.

Another python script then consolidated the information from all the analyses done into a single table. This added columns indicating whether or not that transcript was also DE in B73, the protein domains and GO terms it was annotated with, and the matching transcript(s) in the unmapped assemblies.

Results

The differential expression analysis identified which genes in B73-QTL were induced or repressed in the *C. zeina* challenged leaves compared with the control (Fig. 3.1). Using a FDR cut-off of 0.05 there were 5,236 B73 and B73 variant genes (Table A.1) that were significantly differentially regulated, nine *C. zeina* genes, 200 *E. turcicum* genes and 147 other genes. The 147 other genes contained 47 plant genes (Table A.1).

The GO enrichment (calculated with PLAZA using the hypergeometric distribution) of the B73, B73 variant and plant genes that had higher expression in the *C. zeina* challenged leaves showed the overrepresentation of the GO terms present. Notably the GO terms “Defence response to fungus, incompatible interaction”, “Defence response by callose deposition in cell wall”, “Negative regulation of cell death”, “Response to stress”, “regulation of translation in response to stress”, “Response to stress” and “Negative regulation of cell death” were enriched in the *C. zeina* challenged as would be expected in a leaf stressed by fungal invasion (Fig. 3.2). The under-represented GO terms notably included “DNA repair”, “Cell proliferation” and “Cell proliferation” as would be expected in a leaf focussing less on growth (Fig. 3.2).

A principal component analysis of the expression of the DE genes values normalized for library size was done in R. The three *C. zeina* challenged replicates clustered together and the three B73 replicates clustered together, with the *C. zeina* challenged replicates clustering closer than the control replicates (Fig. 3.3). The *C. zeina* cluster and B73 cluster were separate (Fig. 3.3) showing that the replicates in each condition was more similar to each other than to replicates in the other conditions.

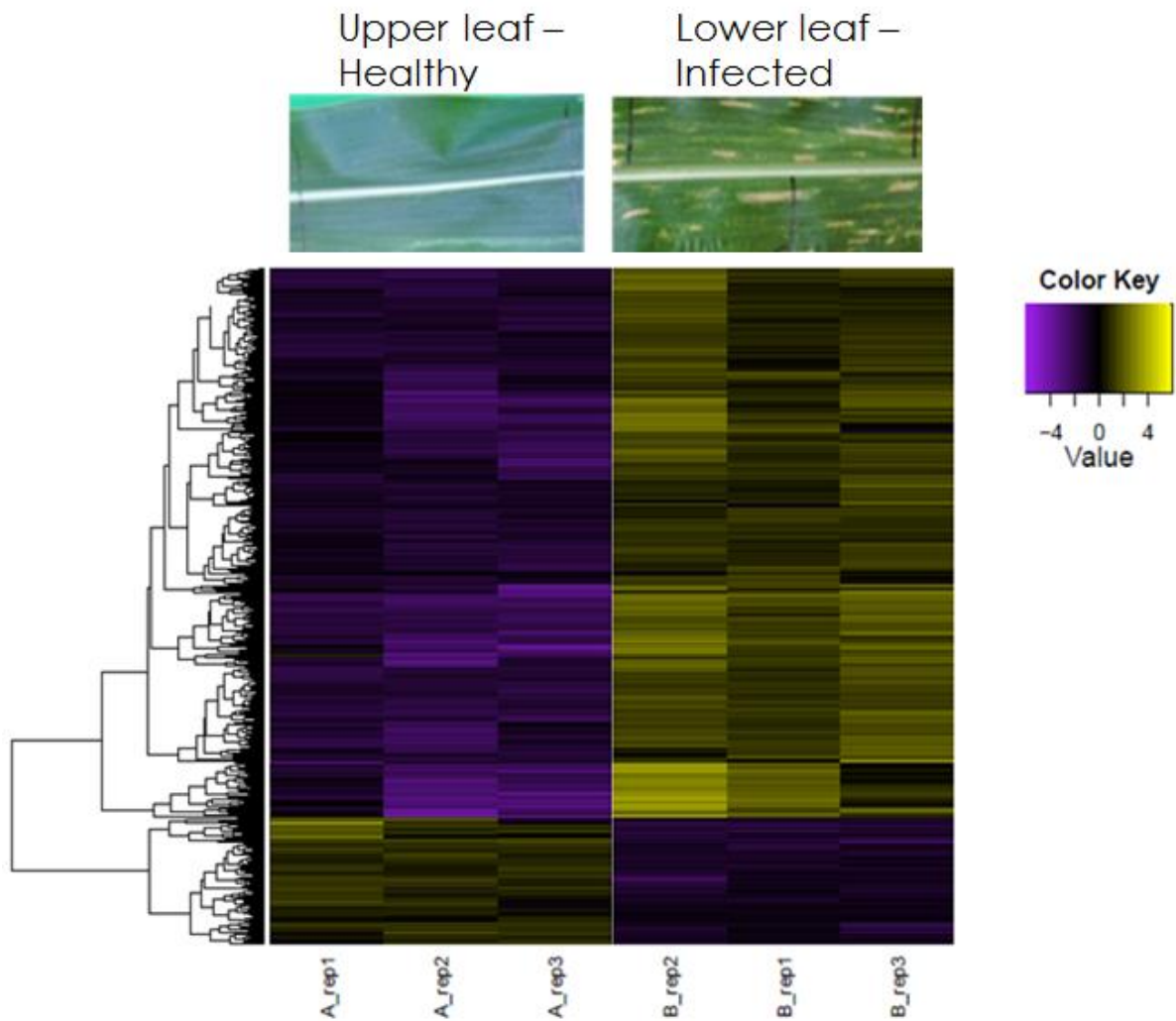


Figure 3.1. Heat map showing differential expression between *C. zeina* challenged and control leaves. The values are the log2 of the amount of the transcript present (in transcripts per million). The most significant ($FDR < 0.01$) transcripts are shown. The heatmap was generated through Trinity and the transcripts with higher expression is shown in yellow and lower expression in purple.

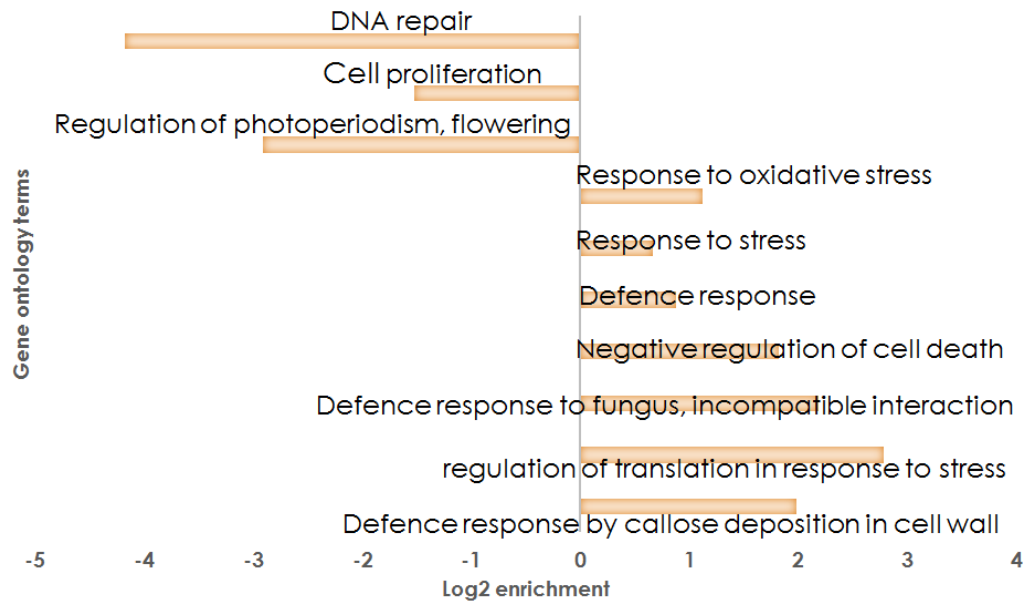


Figure 3.2. Examples of the log₂ enrichment calculated with PLAZA of gene ontology terms of the maize genes in the *C. zeina* challenged leaves.

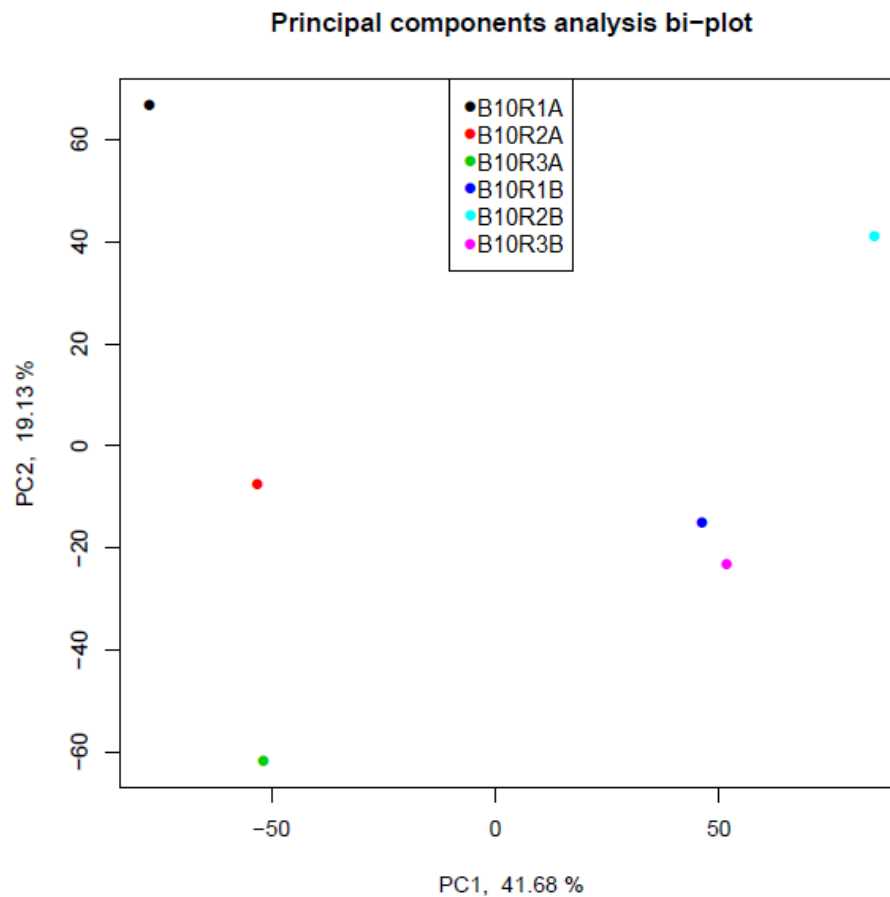


Figure 3.3. A biplot showing the variation caused by the top two principal components. Shown are the control replicates (B10R1A, B10R2A and B10R3A) and the *C. zeina* challenged replicates (B10R1B, B10R2B and B10R3B).

The information on the differential expression in B73 and B73-QTL, the annotation, the GO and the genomic location for each plant gene were taken in consideration to screen them in order to create the shortlist of potential candidates. Specifically, non-B73 genes whose proteins have annotations that are known to be associated with defence and genes that have GO terms indicating a role in pathways involved in defence were selected. Then whether or not the gene was located within the QTL10 region and induced by *C. zeina* was used as corroborating evidence. This resulted in seven candidates consisting of both novel genes and novel variants of B73 genes (Table 3.1). Those which were variants of B73 genes were all differentially expressed in B73-QTL, but their counterparts in B73 was not.

The candidates still had good, but not identical similarity with their closest matches at both the nucleotide and amino acid level when they were globally aligned for the G-type lectin (Fig. A.2-3), thaumatin (Fig A.4-5), transcription factor (Fig. A.6-7), chitinase (Fig. A.8-9), apoptosis (Fig. A.10-11), serine/threonine kinase (Fig. 12-13) and leucine rich repeat (Fig. A.14-15). The apoptosis (Fig. A.10-11), thaumatin (Fig. A.4-5) and chitinase candidates (Fig. A.8-9) were only present in the master assembly done in this chapter and not in the unmapped assemblies done in Chapter 2 which could have been caused by insufficient coverage in the separate assemblies or too many reads mapping to the B73 reference. The other candidates were present in all of the assemblies and the different assemblies were consistent with each other when they overlapped (Fig. 3.4-17).

Table 3.1. Shortlist of defence gene candidates identified from the unmapped assembly and differential expression analysis.

Name	Fold Change ^a	FDR (0.05)	Potential defence Function	%ID (DNA level) to B73 version
G-type lectin S-receptor-like serine/threonine-protein kinase	1.9	0.02	Recognition of pathogens	94
Novel endochitinase	3.6	6×10 ⁻⁸	Innate immune response	n/a
Novel transcription factor	3.8	0.02	Improved regulation of defence response	n/a
Novel Thaumatin	0.2	0.01	Antifungal activity	n/a
Receptor-like serine/threonine-protein kinase	Not differentially expressed	ns*	Plant-pathogen interaction	88
Leucine-rich repeat receptor-like	Not differentially expressed	ns*	Plant-pathogen interaction	89
Allele of Nuclear pore complex protein NUP88	0.01	2×10 ⁻⁴	Systemic acquired resistance	92.75

* not significant

a. The fold change in the infected leaves over the control leaves.

Discussion and Conclusion

A selection of unique genes and unique alleles that could be responsible for the QTL effect was successfully identified (Table 3.1). First is a novel allele of a maize G-type lectin S-receptor-like serine/threonine-protein kinase (AQK46433.1). G-type lectins can be involved in pathogen recognition (Singh and Zimmerli 2013) and this variant could enable more rapid recognition of *C. zeina* during infection.

Second is a novel chitinase that bears similarity to both the maize chitinases Chitinase2 (NM_001148760.2) and DNA basic endochitinase C (NM_001154745). Chitinases are used by the plant to attack fungal cell walls (Balasubramanian *et al.* 2012) and the novel chitinase could be part of a stronger response against *C. zeina*.

Third is a novel transcription factor, which most closely matched an artificial maize transcription factor (AIB05919.1). The artificial transcription factor was based on the B73 basic helix-loop-helix transcription factor GRMZM2G016039 which was codon optimized to increase expression. The B73 gene was not included in the B74 v4 reference and GenBank record is obsolete. The novel transcription factor could alter the regulation of the defence response to be more effective.

Fourth is a novel thaumatin. The predicted protein sequence for this transcript contains the thaumatin domain which has antifungal activity (Balasubramanian *et al.* 2012). The novel thaumatin does not have close matches outside of the thaumatin domain to other maize thaumatins and could contribute to a more effective antifungal response.

Fifth and sixth are respectively novel alleles of a maize receptor-like serine/threonine-protein kinase (ONM13388.1) and a leucine-rich repeat receptor-like (AQK46158.1). Both can be involved in plant-pathogen interaction (Afzal *et al.* 2008; DeYoung and Innes 2006) and the variants could coordinate a more effective defence response from the maize plant.

Last is a novel allele of the maize nuclear pore complex protein NUP88 (AQK45998). The ortholog of this gene in *A. thaliana* (AT5G05680) was shown to be involved in systemic acquired resistance (Cheng *et al.* 2009). The novel allele could improve the signalling during the defence response against *C. zeina*.

The top candidate (the gene most likely to be responsible for the QTL effect) is the novel allele of the G-type lectin S-receptor-like serine/threonine-protein kinase B120. This and the leucine-rich repeat receptor-like gene were the only allele candidates where the B73 version was located within the QTL region, but the G-type lectin candidate was the only one of the two which also had significantly higher expression in the *C. zeina* challenged leaves. The other novel allele candidates could still potentially be copies that are located within the QTL

region, but this is less likely than an allele in the same region. The novel candidates do not have versions in the known B73 genes and so their location cannot be predicted from the location in B73. The novel candidates and alleles would however, have to originate from the QTL region of B73-QTL if they can be shown not to be present in pure B73.

In the first principal component the *C. zeina* challenged replicates are clustered together and the replicates in the control leaves are clustered together (Fig. 3.3). The variation in the data was therefore mostly caused by the two different conditions. In the second principal component the replicates are not clustered together (Fig. 3.3). There was therefore also some variation caused by differences between the replicates, but less than between the conditions.

The low amount of *C. zeina* differentially expressed genes present was surprising considering the large amount of *C. zeina* reads present. This was likely primarily caused by the absence of the *C. zeina* transcripts in the control leaves influencing the statistical support. To a lesser extent the *de novo* assembly optimization being focussed on maize genes would also lead to fewer *C. zeina* transcripts being assembled. There was also more variation between the control replicates than between the *C. zeina* challenged replicates. These leaves had a considerable amount of *C. zeina* pathogenesis transcripts being expressed within each of them as well which could have shifted the average transcripts to contain a greater amount of similar sequences.

Chapter 4: Experimental validation of selected candidate defence gene identified through the designed bioinformatics pipeline

T. Welgemoed, R. Pierneef F. Middleton and D.K. Berger

Author Contributions

T.W.: Conducted the laboratory procedures and data analyses. Wrote the chapter.

R.P.: Reviewed chapter and suggested improvements.

F.M. (Pannar): Created the fine mapping population.

D.K.B.: Principal Investigator of the funding project, acquired funding, provided day-to-day institutional supervision. Reviewed chapter and suggested improvements. Led discussions on data interpretation.

D.K.B. designed the study.

Introduction

During Chapter 3 a list of *de novo* assembled genes that were potentially responsible for the GLS resistance QTL was created. Validation is required to show computationally predicted genes are true genes and not artefacts caused by the *de novo* assembly. As was briefly discussed in Chapter 1 this can be done using PCR to amplify the predicted sequence (Petitot *et al.* 2016; Xu *et al.* 2015) or by comparing the predicted sequences to related sequences (Xu *et al.* 2015; Mizrachi *et al.* 2010). Further examples include Yang *et al.* (2017) who identified a novel allele of caffeoyl-CoA O-methyltransferase in maize that conferred resistance against GLS. They used PCR to validate the sequence of this allele. In another example Gaines *et al.* (2014) who identified genes in the weed *Lolium rigidum* that were involved in herbicide resistance. They validated these genes with quantitative PCR. In a third example Vidal *et al.* (2013) who identified novel nitrate induced genes in *A. thaliana* and also used the quantitative PCR validation strategy.

The sequences identified in this project need to be validated. They were compared against other maize sequences in Chapter 3. The candidates were thus already validated to an extent to be true maize genes. Experimental validation however, could demonstrate that the candidate was located within the QTL region in addition to further confirming its identity.

The top candidate, the G-type lectin S-receptor-like serine/threonine-protein kinase allele, was chosen for experimental validation. The aim was to use PCR-Restriction Fragment

Length Polymorphism (PCR-RFLP) to prove the presence of the *de novo* assembled B73-QTL candidate allele in B73-QTL and its absence in B73. The candidate allele will be present only in B73-QTL if it is a unique gene located within the QTL region. The PCR products of the B73 and B73-QTL alleles were sequenced to confirm their identity. One of the nucleotide base substitutions between the novel B73-QTL allele and the B73 allele created a BamHI restriction site unique to the candidate allele. The different lengths of products present after digestion with the enzyme were used to distinguish between the alleles.

Methods

Plant material

The same maize leaves as in Chapters 2 and 3 were used for this analysis. Briefly, the QTL originates from the subtropical maize line CML444. It was identified to increase GLS resistance in a 2009 field trial in Hildesheim as 10G2_GLS and in another independent 2009 field trial in Baynesfield as 10H_GLS (Berger *et al.*, 2014). This region on chromosome 10, henceforth called QTL10, was backcrossed into the public, susceptible B73 line to produce the near isogenic line B73+QTL10 (B73-QTL). During the backcross breeding four SSR markers (Berger *et al.* 2014) were used where two markers flanked the QTL region (umc1084-10 and bnlg1450-10) and another two were located within the region (umc1196-10 and umc1176-10). A multiplex PCR with primers for these four markers was used to screen progeny for the presence of QTL10. The GeneMarker v1.95 (SoftGenetics, PA, USA) software was used to determine the genotypes. The background was selected for B73 origin using 384 SNP markers across the genome.

Additionally, more maize plants from a fine-mapping population of 77 plants were used in this experimental validation. This population was created through backcrossing B73-QTL to B73 four times. The population was also screened using the four SSR markers (Fig. 4.1), but this time selecting for the presence of a single cross-over event within the QTL region instead of the entire region. This resulted in the QTL10 region being further sub-divided into B73 and CML444 regions. Therefore, using the PCR-RFLP to screen these genotypes for the presence of the B73 or candidate allele will determine between which markers within the QTL region the allele is located.

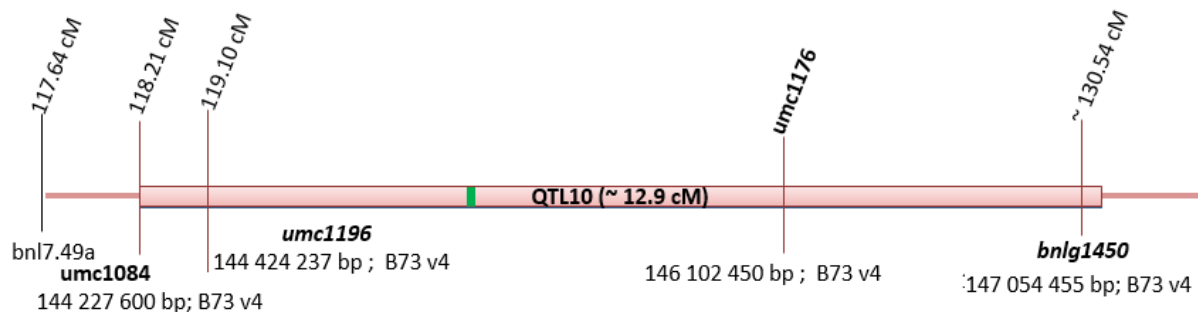


Figure 4.1. A diagram of the QTL10 region showing the four SSR markers used for marker assisted breeding.

PCR-RFLP and sequencing

The *de novo* assembled allele was aligned with the B73 allele (Fig A.2.). The restriction enzyme recognition sites in the sequences were visualised with CLC Main Workbench 8.0.1 (www.qiagenbioinformatics.com/) to select a unique restriction site to distinguish between them. A nucleotide substitution between the two versions at position 932 created a recognition site for the restriction enzyme (RE) BamHI exclusive to the *de novo* assembled candidate.

Primer Designer was used to design primers around the BamHI restriction site. The primer was designed against the transcript TRINITY_DN12415_c0_g1_i1 in the combined assembly from Chapter 3. It was the longest transcript present in this location and had consensus with the other, shorter transcripts (Fig A.2). The primers (Table 4.1) were specific for both B73 and candidate alleles and would produce equal sized products, but only the sequences containing the restriction enzyme site would be cut by the enzyme (Fig 4.1).

Stored DNA extracted from CML444, B73 and seven genotypes from the fine-mapping population developed from crosses between B73 and CML444 were screened. The PCR reactions were set up in a total volume of 12.5 µl, which consisted of: 1× ampliqon taq, 0.2 µm of each primer, 1-2 µl of the DNA, and sterile distilled water. The cycling conditions were as follows: 3 min at 95 °C followed by 30 cycles of 30 s at 95 °C, 30 s at 55 °C, and 30 s at 72 °C, and a final extension step of 5 min at 72 °C. The PCR products from CML444 and B73 were forward sequenced to confirm their identity.

The PCR products were digested with FastDigest reagents (ThermoFisher) to distinguish between the alleles. The reactions were set up in a total volume of 15 µl, which consisted of 10X FastDigest Green buffer, 0.5 µl BamHI FastDigest enzyme, 1 µl of PCR product and sterile distilled water. The reactions were incubated for 1 hour at 37 °C. The PCR product was expected to be a size of 430 bp and if cut by the enzyme to cleave into two segments that are respectively 138 and 296 bp.

The plasmid pdm327 (Kamo *et al.* 2000) was also digested (120 ng) for 15 min at 37 °C using the same protocol to set up the reaction, but using three different aliquots of BamHI enzyme. Duplicate reactions were also set up as a control that had the same ampliqon taq added as was used during the PCR reaction. The reactions with the plasmid were used as a control to determine if digestion had been inhibited by a faulty enzyme or by the dyes and buffers in the ampliqon taq.

The products of the PCR and restriction digestion reactions were visualised through electrophoresis on at 120 V for 30 minutes on a 1% agarose gel.

Table 4.1. The primer sequences used in the PCR-RFLP to distinguish between the candidate and B73 alleles.

Primer	Sequence (5'-3')
Lectin-F	TACCAGGCCAACAATACCTC
Lectin-R	CACACTGGTCGAAGGTTCTA

Results

The PCR reaction was expected to amplify a 430 bp product for both the B73 and CML444 versions of the allele (Fig. 4.3). The CML444 allele is expected to cleave into 138 bp and 296 fragments after digestion (Fig. 4.3). A product with the expected size range was amplified from the CML444, B73, B73-QTL and fine-mapping DNA samples (Fig. 4.2). The products from CML444 and B73-QTL were digested into fragments in the expected size range, but some undigested product also remained in both samples (Fig. 4.3). The remaining undigested product suggest that the B73 sequence is present as well, potentially through heterozygosity or gene duplication. The product from B73 remained undigested as expected. The presence of the candidate allele in both B73-QTL and CML444 and its absence in B73 proves that it is a true CML444 allele that is located within the QTL region.

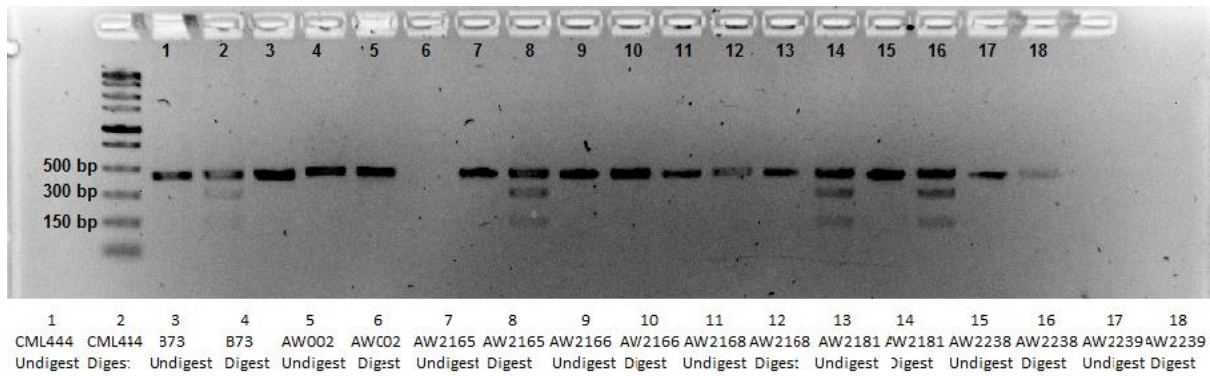


Figure 4.2. PCR-RFLP to distinguish between the B73 and CML444 allele in different genotypes. Fast DNA Ladder (New England Biolabs) was used as the marker. Lane 1, CML444 before digestion; lane 2, CML444 after digestion; lane 3, B73 before digestion; lane 4 B73 after digestion; lane 5 AW002 before digestion; lane 6 AW002 after digestion; lane 7, AW2165 before digestion; lane 8; AW2165 after digestion; lane 9, AW2166 before digestion; lane 10, AW2166 after digestion; lane 11, AW2168 before digestion; lane 12, AW2168 after digestion; lane 13, AW2181 before digestion; lane 14, AW2181 after digestion; lane 15, AW2238 before digestion; lane 16, AW2238 after digestion; lane 17, AW2239 before digestion; lane 18, AW2239 after digestion.

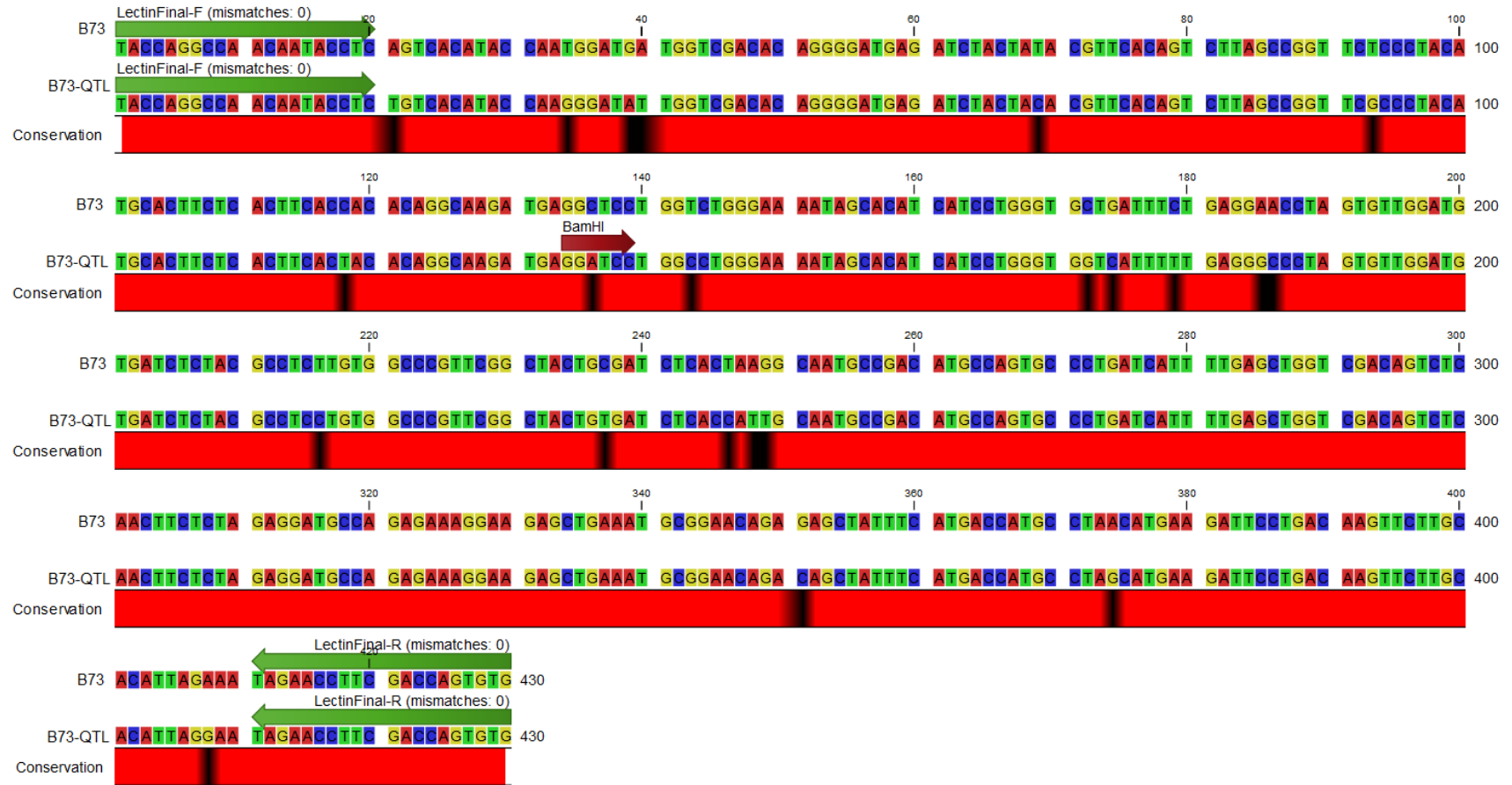


Figure 4.3. The expected products amplified with the Lectin primers in B73 and B73-QTL. The first sequence in the alignment shows the B73 sequence and the second shows the de novo assembled B73-QTL sequence. The primer binding sites are indicated on both sequences with a green arrow and the unique BamHI restriction site is indicated on B73-QTL with a red arrow. The conservation bar shows the conserved bases where red indicates conserved bases and black indicates non-conserved bases. Image created in CLC Main work bench 8.0.1 (<https://www.qiagenbioinformatics.com/>).

The B73 allele is located on chromosome 10 at position 144,818,522 to position 144,822,103 in the B73 genome, presuming the candidate is an alternate allele it should also be located here. This position would be between the umc1196 and umc1176 markers used in the fine-mapping population (Fig. 4.1).

In the genotypes selected to have no crossover events between the umc1196 and umc1176 markers (CML444, AW002, AW2238 and AW2181) the region between the markers originated from CML444 (Fig. 4.4). In these genotypes there was digestion, indicating that the candidate allele is present (Fig. 4.2). AW002 was unsuccessful in this set of reactions, but had the same results as CML444 in another (not shown). AW2181 was selected to have a crossover event between the umc1176 and bnlgl450 markers (Fig. 4.4) and the sequence changed to from CML444 to B73 origin after umc1176. This showed that the candidate allele was located upstream from the bnlgl450 marker and was most likely between the umc1196 and umc1176 markers.

In the genotypes where the region between the umc1196 and umc1176 markers were selected to originate from B73 (AW2239/B73 and AW2166; Fig. 4.4) there was no digestion (Fig. 4.2). AW2166 was selected to have a crossover event between the umc1176 and bnlgl450 markers (Fig. 4.4) and the sequence changed from B73 to CML444 origin after umc1176. This showed that the candidate allele was not downstream from the bnlgl450 marker.

AW2168 was selected to have a crossover event between the markers umc1196 and umc1176 (Fig. 4.4) and the sequence changed from CML444 to B73 origin after the umc1176 marker. There was digestion, indicating that the CML444 allele was present (Fig. 4.2). In the opposite genotype (AW2165) there is also a crossover event between these two markers, but the sequence changed from B73 to CML444 origin after the umc1176 marker (Fig. 4.4). This genotype did not have digestion, indicating that the CML444 allele was not present (Fig. 4.2). This showed that the candidate allele was between the umc1196 and umc1176 markers as expected.

The samples also had undigested product remaining in every digestion reaction. This indicates that the B73 copy was retained in all the progeny that was tested. In the control restriction digest reactions, the plasmid was fully linearized which showed that this was not a result of the reaction being inhibited by either a faulty enzyme or the dyes and buffers in the ampliqon taq (Fig. 4.5). The BamHI restriction enzyme used is also not sensitive to methylation and so the undigested copies could not have been caused by the RE site being methylated in some of the nuclei. This confirmed that the B73 copy was indeed present alongside the CML444 copy.

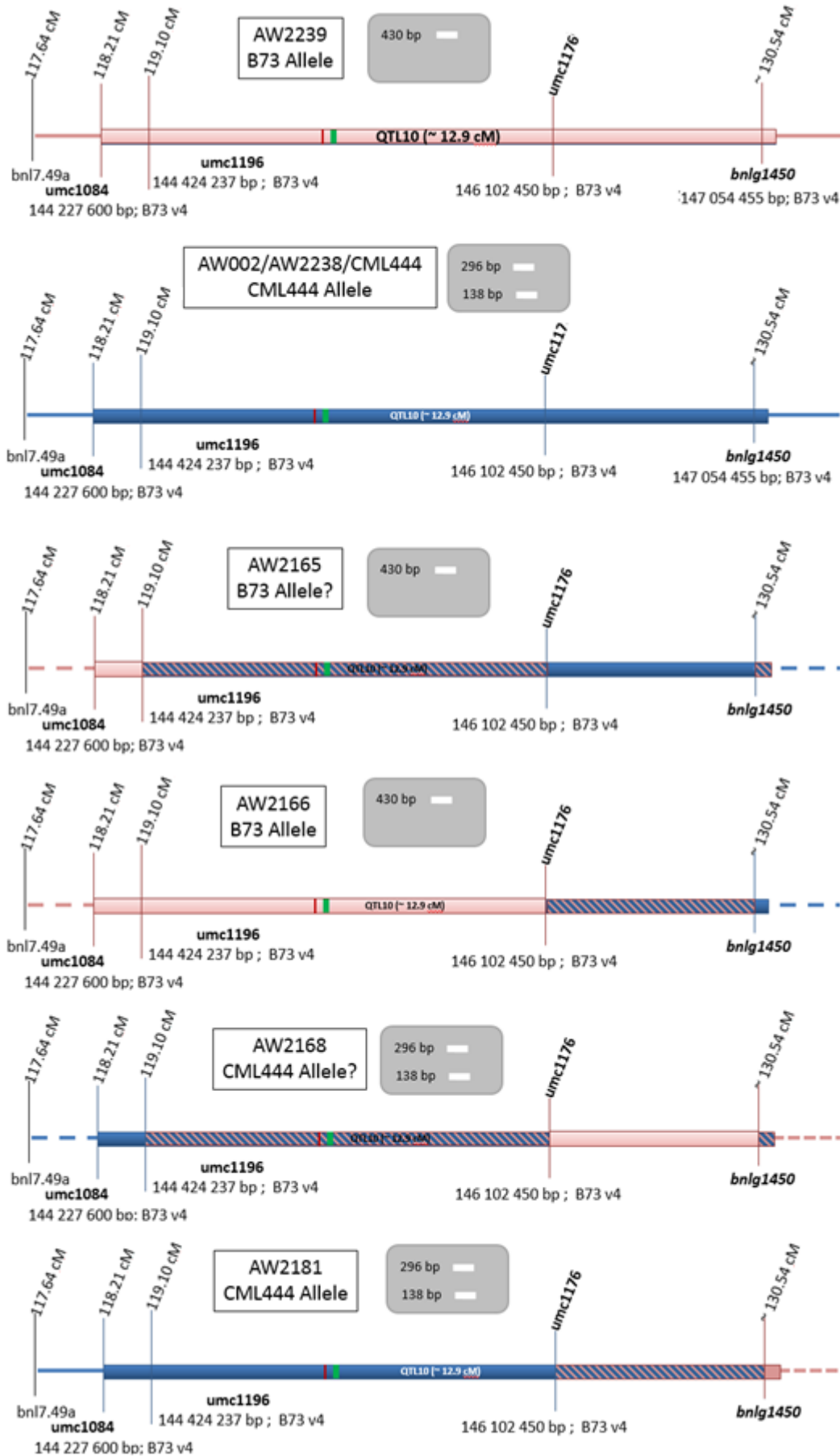


Figure 4.4. A representation of the QTL10 region. The lines show the markers used for selection between CML444 and B73 (Berger et al. 2014). Pink areas and lines show regions where the sequence is from B73, blue shows regions where the sequence is from CML444 and diagonal stripes and dashes shows regions that at some point between the two markers changed from either B73 or CML444 to the other. The green area shows the approximate position of the peak marker closest to the QTL peak and the red area shows the approximate location of the candidate allele. The name of each genotype and the expected allele of the G-type lectin present is given at the top.

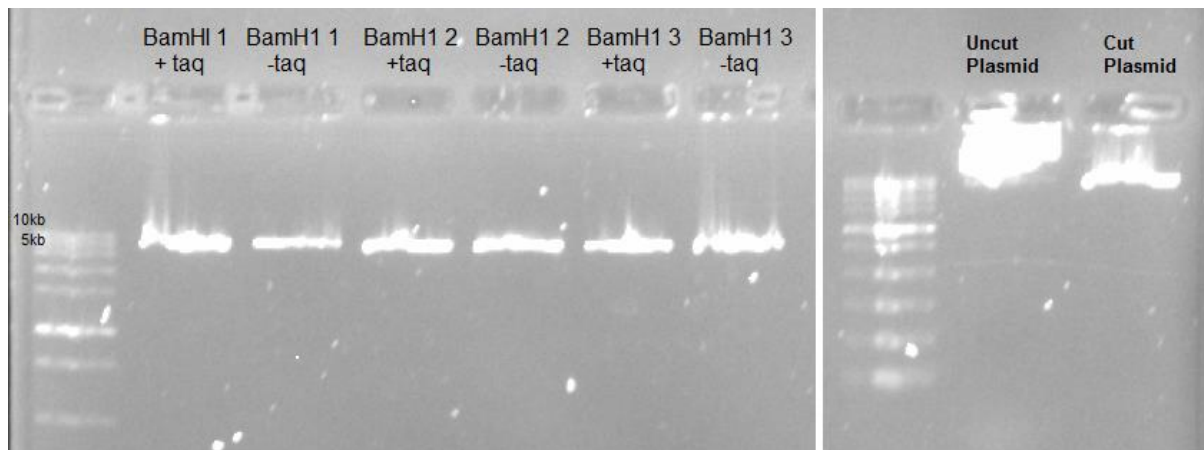


Figure 4.5. Control using the pdm327 plasmid showing restriction digests were not inhibited by ampliqon taq or faulty enzyme. To the left is shown the test reactions, which aliquot of BamHI RE was used and whether ampliqon taq was present or absent is indicated on the top of the lanes. To the right is shown a size comparison between the linearized and circular plasmid, indicated at the top of the lane. Fast DNA Ladder (New England Biolabs) was used as the markers for both.

The sequencing data from the B73 PCR product fully matched the expect sequence, confirming that the correct product had been amplified (Fig. 4.6). The sequencing data from the CML444 PCR product largely matched the *de novo* assembled sequence, but at several of the substituted nucleotide positions there were a double peak in the trace data where the identity of the nucleotide base is unclear, including in the BamHI recognition site (Fig. 4.6). The base present at these locations could have been either the one expected for B73 or for CML444. The B73 sequencing reaction was carried out at the same time and under same conditions, but had clear peaks in the trace data, including at the substituted positions.

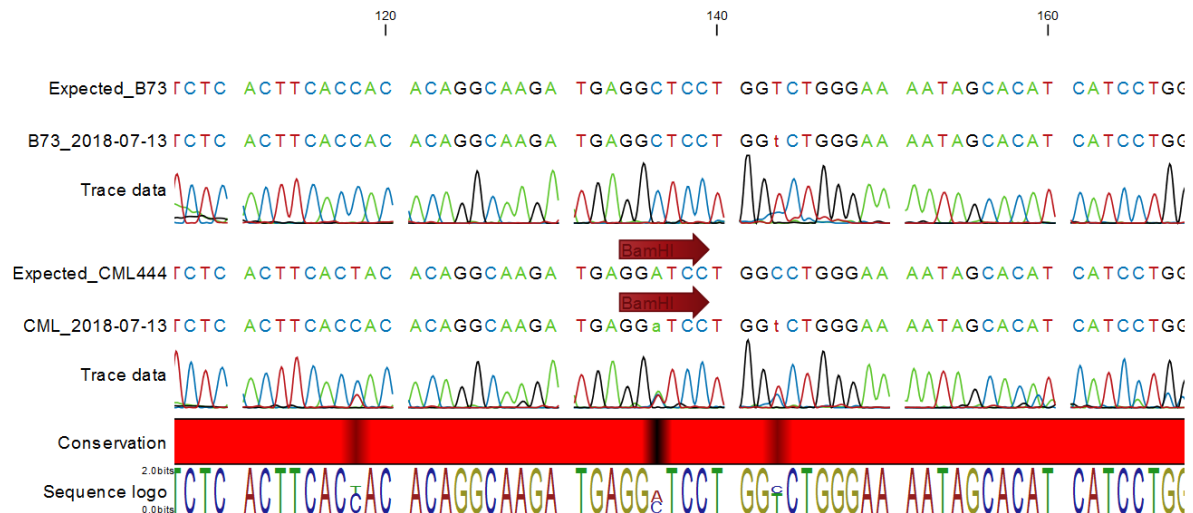


Figure 4.6. Portion of sequencing reaction showing the expected sequence, sequence from the sequencing reaction and trace data for both B73 and CML444. In the conservation bar red indicates conserved bases and black indicates non-conserved bases.

Image created CLC main workbench 8.0.1 (<https://www.qiagenbioinformatics.com/>).

Discussion and conclusions

The chosen candidate, (the novel allele of the G-type lectin S-receptor-like serine/threonine-protein kinase) was shown to be present in only B73-QTL, CML444 and fine-mapping genotypes including the QTL region. This proves that it is a genuine CML444 allele unique to QTL10 region.

The results from the fine-mapping population showed that the candidate was located between the umc1196 and umc1176 markers. In the AW2168 and AW2165 genotypes the sequence between these two markers changed between CML444 and B73. The location of the B73 version of the allele is closer to the umc1196 marker than umc1176 marker and it was expected for the allele to be CML444 version when umc1196 originated from CML444 and vice versa. The opposite result was obtained during the PCR-FLP. Nothing conclusive can be drawn from this as it is unclear where exactly the crossover event between the two markers occurred, but it suggests the allele is further down-stream than expected.

The candidate allele seems to be present in addition to the B73 allele, instead of replacing it in all of the samples. This was also supported by the sequencing data, where the double peaks in the trace could represent both sequences being present in the reaction. As this included the three pure CML444 lines (CML444, AW2238, and AW002) this was also not an artefact of the cross-breeding introducing heterozygosity from B73. The B73 version in CML444 could be due to that line maintaining the heterozygosity in its population or because of a gene duplication.

Chapter 5: General discussion and Conclusions

The pipeline allowed the identification of novel, expressed genes located in the QTL10 region of CML444. Mapping against the B73 reference genome and assembling the unmapped reads reduced the maize sequences to a small set that contained transcripts from the QTL region. The presence of at least one gene in this region was successfully validated through PCR-RFLP.

Theoretically mapping against the B73 reference genome and annotation should have excluded all transcripts from other genomic locations, but this did not work perfectly in practice. Some of the assembled transcripts were found to be identical to older B73 sequences. These had been excluded from the B73 v.4 references and therefore were missed during the mapping process. The filtering step against the nr database (NCB_Resource_Coordinators 2016), which had a broader range of reference sequences, allowed this to be identified.

The filtering against the nr database (NCB_Resource_Coordinators 2016) also revealed a large amount of fungi and other sequences present. It was expected that other fungi and plant bacteria or viruses could have infected the leaves during the field-trial and be included in the RNA-seq data. These were found to be present, but other miscellaneous sequences were also included (Table A.3-6).

The non-redundant nature of the database means that while there is a large representation of sequences, there could have been a closer match to an excluded similar sequence. The BLAST algorithm also uses local alignments and consequently portions conserved between different species can create additional matches. The sequences that were not plant or plant pathogen sequences most likely consisted of such partial matches and incorrectly assembled transcripts that coincidentally matched an nr entry.

The differential expression analysis successfully identified which genes were significantly induced by *C. zeina* infection. A limitation of this study was that the leaves under experimental conditions were older than the leaves under control conditions. This means that some of the regulatory differences would be caused by the difference in developmental stage in addition to the difference in the tested conditions. The same applies to the analysis of B73 by Swart *et al.* (2017) where the *C. zeina* challenged leaves were also older than the control leaves. This enabled the comparison between the results for B73-QTL and B73 to be valid as any purely developmental regulatory differences between the older and younger

leaves would be present in both B73-QTL and B73. Genes which only had higher expression in B73-QTL would consequently have been influenced by the presence of the QTL region.

The transcripts were successfully annotated with their protein domains and GO terms. This included a subset of the transcripts which could be involved in a defence process or have antifungal activity. This resulted in the selected shortlist of candidates which could potentially cause the QTL effect. Similarity by itself does not mean that the sequences are related and even very phylogenetically distant sequences can have a high BLAST score (Koski and Golding 2001). Full-length alignments of the candidates against the closest match were also made to highlight discrepancies and InterProScan (Jones *et al.* 2014) independently confirmed that the candidates contain the relevant domains instead of only relying on BLAST to show which the transcripts were similar to the identified sequence.

The G-type lectin S-receptor-like serine/threonine-protein kinase allele was selected as the top candidate. This gene has an ortholog in *A. thaliana* also called G-type lectin S-receptor-like serine/threonine-protein kinase B120 (AT4G21390). There are also orthologs present in *Oryza sativa*, *Setaria italica* and *Sorghum bicolor*, but all are putative genes. To date no studies have been done on the functionality of this specific gene in maize, but the *A. thaliana* ortholog is known to respond to fungal pathogens (Qutob *et al.* 2006) and oxidative stress (Stanley Kim *et al.* 2005) and likely function similarly in maize.

The *de novo* assembled G-type lectin allele contains nucleotide substitutions throughout the sequence compared to the B73 allele (Fig. A.5). Many of these substitutions will cause non-synonymous mutations in functional areas of the protein. In the protein there are substitutions at two sites in the dimerization interface, at six sites in the D-mannose binding lectin domain and at four sites in the protein kinase domain. A follow up study is required to determine the exact effects of these changes, but alterations in the functional domains suggests that there are functional differences between the two alleles. This is also supported by the candidate allele in B73-QTL having significantly higher expression in the infected leaves, but the B73 allele in B73 is not.

It is still required to determine if the G-type lectin allele or one of the other candidates is the gene causing the QTL effect. This could be done through a knockout study in B73-QTL or by introducing only the single gene into B73. For example, Yang *et al.* (2017) determined which of their candidates conferred resistance by transforming each of them into a susceptible maize line and evaluating which transgenic line gained the resistance.

Several candidates, including the top candidate, were alternative alleles to a gene in the susceptible line. In a case where one of these was determined to confer the *C. zeina* resistance there must be a functional difference to the susceptible B73 allele. It would

require investigation to determine why the defence mechanism is more effective in this allele than the B73 version. This could be because of the higher induction in B73-QTL than in B73. Such candidates which had significantly higher expression in response to *C. zeina* could have altered promoter sequences as well. This could be investigated by sequencing the DNA upstream from the coding sequence. For example, it was determined using this method that the maize genes GRMZM2G021277 and GRMZM2G108338 had an indel upstream from their coding sequences that caused them to be significantly differently expressed (Wen *et al.* 2014). The novel transcription factor could also cause regulatory changes in the defence response, in which case chromatin immunoprecipitation could be used to identify which genes it targets.

The resistant allele could also have structural differences that change how it interacts with other protein partners. Three of the candidates, namely the G-type lectin serine/threonine-protein kinase, receptor-like serine/threonine-protein kinase and leucine-rich repeat receptor-like candidates, are likely to be involved in pathogen recognition (Afzal *et al.* 2008; DeYoung and Innes 2006; Singh and Zimmerli 2013). A protein interaction study could confirm if these candidates interact with *C. zeina* proteins and if they interact with more or different proteins than the B73 versions. Protein interaction studies and protein structure prediction could also be used to determine if any of these or the other candidates interact differently with internal maize proteins.

The results from the restriction digestions indicate that both versions of the alleles are present in CML444. This could be explained through all of the tested samples being heterozygous at this locus. The maize lines were inbred so this is unlikely unless there was a strong pressure to maintain a copy of the B73 allele. Alternatively, the gene could have been duplicated in CML444 and the copy had undergone subfunctionalization or neofunctionalization to create the novel version.

Maize went through whole genome duplications that were reduced back to diploidy and the redundant genes have in many cases been adapted to new functions (Hughes *et al.* 2014; Schnable *et al.* 2011). For example, a maize QTL for resistance against corn ear-worm is caused by such a case (Zhang *et al.* 2003). Two genes originated from a duplication event act in combination to confer the resistance (Zhang *et al.* 2003). This might also be occurring in the QTL10 region. It could be that instead of the CML444 allele being more effective than the B73 allele they improve each other's functionality when defending against *C. zeina* infection.

It will require further work to test these hypotheses. Further sequencing and PCR-RFLP screening could be used to place the location of the alleles in the genome more precisely

thus determining if they are heterozygotes in the same location on opposite strands or if they are duplicates located in different positions in the genome. In any future studies on this allele it should also be taken into account that the CML444 version of the gene might, like in the previous example, require the B73 version to be present as well to provide an equally effective defence response.

Similar studies have been done on other maize QTLs for grey leaf spot resistance. A study on a QTL on chromosome 8 by Zhang *et al.* (2017) had narrowed down the gene/element responsible for the QTL effect to one of four candidates in the QTL region: A receptor-like protein kinase, an upstream open reading frame, an ATP-binding cassette transporter or an uncharacterized protein. Another study on a QTL on chromosome 9 by Yang *et al.* (2017) had successfully identified the gene responsible for the QTL effect to be conferred by an alternative allele of Caffeoyl-CoA O-methyltransferase.

In conclusion, this study like the previous examples narrowed down the genes responsible for a QTL effect. A promising top candidate was proposed that can be followed up in future studies to confirm it is the gene responsible for the effect in QTL10.

References

- Afzal AJ, Wood AJ, Lightfoot DA (2008) Plant receptor-like serine threonine kinases: roles in signaling and plant defense. *Molecular Plant-Microbe Interactions* 21 (5):507-517
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *Journal of molecular biology* 215 (3):403-410
- Andorf CM, Cannon EK, Portwood JL, Gardiner JM, Harper LC, Schaeffer ML, Braun BL, Campbell DA, Vinnakota AG, Sribalusu VV (2016) MaizeGDB update: new tools, data and interface for the maize model organism database. *Nucleic acids research* 44 (D1):D1195-D1201
- Andrews S (2016) FastQC: A quality control application for FastQ data. <http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc>.
- Asea G, Bigirwa G, Adipala E, Oweru SAP, Pratt RC, Lipps PE (2002) Effect of *Cercospora zeae-maydis* infested maize residue on progress and spread of grey leaf spot of maize in central Uganda. *Annals of Applied Biology* 140 (2):177-185. doi:10.1111/j.1744-7348.2002.tb00171.x
- Asea G, Vivek BS, Bigirwa G, Lipps PE, Pratt RC (2009) Validation of consensus quantitative trait loci associated with resistance to multiple foliar pathogens of maize. *Phytopathology* 99 (5):540-547
- Balasubramanian V, Vashisht D, Cletus J, Sakthivel N (2012) Plant β -1, 3-glucanases: their biological functions and transgenic expression against phytopathogenic fungi. *Biotechnology letters* 34 (11):1983-1990
- Benson JM, Poland JA, Benson BM, Stromberg EL, Nelson RJ (2015) Resistance to gray leaf spot of maize: genetic architecture and mechanisms elucidated through nested association mapping and near-isogenic line analysis. *PLoS Genet* 11 (3):e1005045
- Berger D, Carstens M, Korsman J, Middleton F, Klopppers F, Tongoona P, Myburg A (2014) Mapping QTL conferring resistance in maize to gray leaf spot disease caused by *Cercospora zeina*. *BMC Genetics* 15 (1):60
- Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*:btu170
- Chen X, Shang J, Chen D, Lei C, Zou Y, Zhai W, Liu G, Xu J, Ling Z, Cao G, Ma B, Wang Y, Zhao X, Li S, Zhu L (2006) A B-lectin receptor kinase gene conferring rice blast resistance. *The Plant Journal* 46 (5):794-804. doi:doi:10.1111/j.1365-3113X.2006.02739.x
- Cheng YT, Germain H, Wiermer M, Bi D, Xu F, Garcia AV, Wirthmueller L, Despres C, Parker JE, Zhang Y, Li X (2009) Nuclear pore complex component MOS7/Nup88 is required for innate immunity and nuclear accumulation of defense regulators in *Arabidopsis*. *The Plant cell* 21 (8):2503-2516. doi:10.1105/tpc.108.064519
- Chettoor AM, Givan SA, Cole RA, Coker CT, Unger-Wallace E, Vejlupkova Z, Vollbrecht E, Fowler JE, Evans MM (2014) Discovery of novel transcripts and gametophytic functions via RNA-seq analysis of maize gametophytic transcriptomes. *Genome biology* 15 (7):414
- Chionh YH, Ho C-H, Pruksakorn D, Babu IR, Ng CS, Hia F, McBee ME, Su D, Pang YLJ, Gu C (2013) A multidimensional platform for the purification of non-coding RNA species. *Nucleic acids research*:gkt668
- Christie N, Myburg AA, Joubert F, Murray SL, Carstens M, Lin YC, Meyer J, Crampton BG, Christensen SA, Ntuli JF (2017) Systems genetics reveals a transcriptional network associated with susceptibility in the maize–grey leaf spot pathosystem. *The Plant Journal* 89 (4):746-763
- Compeau PE, Pevzner PA, Tesler G (2011) How to apply de Bruijn graphs to genome assembly. *Nature biotechnology* 29 (11):987-991

- Condon BJ, Leng Y, Wu D, Bushley KE, Ohm RA, Otilar R, Martin J, Schackwitz W, Grimwood J, MohdZainudin N, Xue C, Wang R, Manning VA, Dhillon B, Tu ZJ, Steffenson BJ, Salamov A, Sun H, Lowry S, LaButti K, Han J, Copeland A, Lindquist E, Barry K, Schmutz J, Baker SE, Ciuffetti LM, Grigoriev IV, Zhong S, Turgeon BG (2013) Comparative genome structure, secondary metabolite, and effector coding capacity across *Cochliobolus* pathogens. *PLoS Genet* 9 (1):e1003233. doi:10.1371/journal.pgen.1003233
- Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, Szcześniak MW, Gaffney DJ, Elo LL, Zhang X (2016) A survey of best practices for RNA-seq data analysis. *Genome biology* 17 (1):13
- Coolman R, Hoyt G (1993) The effects of reduced tillage on the soil environment. *HortTechnology* 3 (2):143-145
- Dao P, Numanagić I, Lin Y-Y, Hach F, Karakoc E, Donmez N, Collins C, Eichler EE, Sahinalp SC (2013) Optimal resolution of ambiguous RNA-Seq multimappings in the presence of novel isoforms. *Bioinformatics*:btt591
- De Coninck B, Cammue BP, Thevissen K (2013) Modes of antifungal action and in planta functions of plant defensins and defensin-like peptides. *Fungal Biology Reviews* 26 (4):109-120
- Denisov V, Strong W, Walder M, Gingrich J, Wintz H (2008) Development and validation of RQI: an RNA quality indicator for the Experion automated electrophoresis system. *Bio-Rad Bulletin* 5761
- Devisetty UK, Covington MF, Tat AV, Lekkala S, Maloof JN (2014) Polymorphism identification and improved genome annotation of *Brassica rapa* through deep RNA sequencing. *G3: Genes| Genomes| Genetics* 4 (11):2065-2078
- DeYoung BJ, Innes RW (2006) Plant NBS-LRR proteins in pathogen sensing and host defense. *Nature immunology* 7 (12):1243-1249
- Gaines TA, Lorentz L, Figge A, Herrmann J, Maiwald F, Ott MC, Han H, Busi R, Yu Q, Powles SB (2014) RNA - Seq transcriptome analysis to identify genes involved in metabolism - based diclofop resistance in *Lolium rigidum*. *The Plant Journal* 78 (5):865-876
- Garg R, Patel RK, Tyagi AK, Jain M (2011) De novo assembly of chickpea transcriptome using short reads for gene discovery and marker identification. *DNA research* 18 (1):53-63
- Gerstein MB, Bruce C, Rozowsky JS, Zheng D, Du J, Korb J, Emanuelsson O, Zhang ZD, Weissman S, Snyder M (2007) What is a gene, post-ENCODE? History and updated definition. *Genome research* 17 (6):669-681
- Gordon A, Hannon G (2010) Fastx-toolkit. FASTQ/A short-reads pre-processing tools. Unpublished http://hannonlabcshtedu/fastx_toolkit
- Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature biotechnology* 29 (7):644-652
- Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature protocols* 8 (8):1494-1512
- Hansey CN, Vaillancourt B, Sekhon RS, De Leon N, Kaeppler SM, Buell CR (2012) Maize (*Zea mays* L.) genome diversity as revealed by RNA-sequencing. *PloS one* 7 (3):e33071
- Hillier LW, Marth GT, Quinlan AR, Dooling D, Fewell G, Barnett D, Fox P, Glasscock JI, Hickenbotham M, Huang W (2008) Whole-genome sequencing and variant discovery in *C. elegans*. *Nature methods* 5 (2):183
- Huang Y, Hu Y, Liu J (2014) Piecing the puzzle together: a revisit to transcript reconstruction problem in RNA-seq. *BMC bioinformatics* 15 (Suppl 9):S3

- Hughes TE, Langdale JA, Kelly S (2014) The impact of widespread regulatory neofunctionalization on homeolog gene evolution following whole genome duplication in maize. *Genome research*:gr. 172684.172114
- Jiang H, Wong WH (2009) Statistical inferences for isoform expression in RNA-Seq. *Bioinformatics* 25 (8):1026-1032
- Jiao Y, Peluso P, Shi J, Liang T, Stitzer MC, Wang B, Campbell MS, Stein JC, Wei X, Chin C-S, Guill K, Regulski M, Kumari S, Olson A, Gent J, Schneider KL, Wolfgruber TK, May MR, Springer NM, Antoniou E, McCombie WR, Presting GG, McMullen M, Ross-Ibarra J, Dawe RK, Hastie A, Rank DR, Ware D (2017) Improved maize reference genome with single-molecule technologies. *Nature advance online publication*. doi:10.1038/nature22971
- <http://www.nature.com/nature/journal/vaop/ncurrent/abs/nature22971.html#supplementary-information>
- Jin M, Liu H, He C, Fu J, Xiao Y, Wang Y, Xie W, Wang G, Yan J (2016) Maize pan-transcriptome provides novel insights into genome complexity and quantitative trait variation. *Scientific reports* 6:18936
- Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics* 30 (9):1236-1240
- Jurka J, Klonowski P, Dagman V, Pelton P (1996) CENSOR—a program for identification and elimination of repetitive elements from DNA sequences. *Computers & chemistry* 20 (1):119-121
- Kakumanu A, Ambavaram MMR, Klumas C, Krishnan A, Batlang U, Myers E, Grene R, Pereira A (2012) Effects of Drought on Gene Expression in Maize Reproductive and Leaf Meristem Tissue Revealed by RNA-Seq. *Plant Physiology* 160 (2):846-867. doi:10.1104/pp.112.200444
- Kamo K, Blowers A, McElroy D (2000) Effect of the cauliflower mosaic virus 35S, actin, and ubiquitin promoters on uidA expression from a bar-uidA fusion gene in transgenic *Gladiolus* plants. *In Vitro Cellular & Developmental Biology-Plant* 36 (1):13-20
- Kent WJ (2002) BLAT—the BLAST-like alignment tool. *Genome research* 12 (4):656-664
- Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome biology* 14 (4):R36
- Kloppers R, Tweer S (2009) Northern corn leaf blight fact sheet. PANNAR Seed (Pty) Ltd [\[https://www.plantwise.org/FullTextPDF/2011/20117800335pdf\]](https://www.plantwise.org/FullTextPDF/2011/20117800335pdf) Accessed 21/08/2018
- Koski LB, Golding GB (2001) The closest BLAST hit is often not the nearest neighbor. *Journal of molecular evolution* 52 (6):540-542
- Lacerda A, Vasconcelos ÉAR, Pelegrini PB, Grossi-de-Sa MF (2014) Antifungal defensins and their role in plant defense. *Frontiers in microbiology* 5:116
- Lai J, Li R, Xu X, Jin W, Xu M, Zhao H, Xiang Z, Song W, Ying K, Zhang M (2010) Genome-wide patterns of genetic variation among elite maize inbred lines. *Nature genetics* 42 (11):1027
- Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nature methods* 9 (4):357-359
- Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* 10 (3):R25-R25. doi:10.1186/gb-2009-10-3-r25
- Latterell FM, Rossi AE (1983) Gray leaf spot of corn: a disease on the move. *Plant Disease* 67 (8):842-847
- Law CW, Chen Y, Shi W, Smyth GK (2014) Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome biology* 15 (2):1
- Levin JZ, Yassour M, Adiconis X, Nusbaum C, Thompson DA, Friedman N, Gnirke A, Regev A (2010) Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nature methods* 7 (9):709-715. doi:10.1038/nmeth.1491

- Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25 (14):1754-1760. doi:10.1093/bioinformatics/btp324
- Li H, Ruan J, Durbin R (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research* 18 (11):1851-1858. doi:10.1101/gr.078212.108
- Li P, Ponnala L, Gandotra N, Wang L, Si Y, Tausta SL, Kebrom TH, Provart N, Patel R, Myers CR (2010) The developmental dynamics of the maize leaf transcriptome. *Nature genetics* 42 (12):1060
- Li WV, Zhao A, Zhang S, Li JJ (2016) MSIQ: Joint Modeling of Multiple RNA-seq Samples for Accurate Isoform Quantification. arXiv preprint arXiv:160305915
- Lu B, Zeng Z, Shi T (2013) Comparative study of de novo assembly and genome-guided assembly strategies for transcriptome reconstruction based on RNA-Seq. *Science China Life sciences* 56 (2):143-155
- Lu T, Lu G, Fan D, Zhu C, Li W, Zhao Q, Feng Q, Zhao Y, Guo Y, Li W, Huang X, Han B (2010) Function annotation of the rice transcriptome at single-nucleotide resolution by RNA-seq. *Genome Research* 20 (9):1238-1249. doi:10.1101/gr.106120.110
- Mallowa SO, Esker PD, Paul PA, Bradley CA, Chapara VR, Conley SP, Robertson AE (2015) Effect of maize hybrid and foliar fungicides on yield under low foliar disease severity conditions. *Phytopathology* 105 (8):1080-1089
- Martin L, Fei Z, Giovannoni J, Rose JKC (2013) Catalyzing plant science research with RNA-seq. *Frontiers in plant science* 4:66
- Meisel B, Korsman J, Kloppers FJ, Berger DK (2009) *Cercospora zeina* is the causal agent of grey leaf spot disease of maize in southern Africa. *European journal of plant pathology* 124 (4):577-583
- Mizrachi E, Hefer CA, Ranik M, Joubert F, Myburg AA (2010) De novo assembled expressed gene catalog of a fast-growing Eucalyptus tree produced by Illumina mRNA-Seq. *BMC genomics* 11 (1):1
- Muller MF, Barnes I, Kunene NT, Crampton BG, Bluhm BH, Phillips SM, Olivier NA, Berger DK (2016) *Cercospora zeina* from Maize in South Africa Exhibits High Genetic Diversity and Lack of Regional Population Differentiation. *Phytopathology* 106 (10):1194-1205. doi:10.1094/PHYTO-02-16-0084-FI
- Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* 320 (5881):1344-1349
- NCB_Resource_Coordinators (2016) Database resources of the national center for biotechnology information. *Nucleic acids research* 44 (Database issue):D7
- Paszkiwicz K, Studholme DJ (2010) De novo assembly of short sequence reads. *Briefings in Bioinformatics* 11 (5):457-472. doi:10.1093/bib/bbq020
- Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C (2017) Salmon provides fast and bias-aware quantification of transcript expression. *Nature methods* 14 (4):417
- Pechanova O, Pechan T, Williams WP, Luthe DS (2011) Proteomic analysis of the maize rachis: potential roles of constitutive and induced proteins in resistance to *Aspergillus flavus* infection and aflatoxin accumulation. *Proteomics* 11 (1):114-127
- Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature biotechnology* 33 (3):290-295
- Petitot AS, Dereeper A, Agbessi M, Da Silva C, Guy J, Ardisson M, Fernandez D (2016) Dual RNA - seq reveals *Meloidogyne graminicola* transcriptome and candidate effectors during the interaction with rice plants. *Molecular plant pathology* 17 (6):860-874
- Qutob D, Kemmerling B, Brunner F, Küfner I, Engelhardt S, Gust AA, Lubracki B, Seitz HU, Stahl D, Rauhut T, Glawischnig E, Schween G, Lacombe B, Watanabe N, Lam E, Schlichting R, Scheel D, Nau K, Dodt G, Hubert D, Gijzen M, Nürnberger T (2006) Phytotoxicity and Innate Immune Responses Induced by Nep1-Like Proteins. *The Plant cell* 18 (12):3721-3744. doi:10.1105/tpc.106.044180

- Ramirez - Gonzalez RH, Segovia V, Bird N, Fenwick P, Holdgate S, Berry S, Jack P, Caccamo M, Uauy C (2015) RNA - Seq bulked segregant analysis enables the identification of high - resolution genetic markers for breeding in hexaploid wheat. *Plant biotechnology journal* 13 (5):613-624
- Ranum P, Peña-Rosas JP, Garcia-Casal MN (2014) Global maize production, utilization, and consumption. *Annals of the New York Academy of Sciences* 1312 (1):105-112. doi:10.1111/nyas.12396
- Regulski M, Lu Z, Kendall J, Donoghue MT, Reinders J, Llaca V, Deschamps S, Smith A, Levy D, McCombie WR (2013) The maize methylome influences mRNA splice sites and reveals widespread paramutation-like switches guided by small RNA. *Genome research*:gr. 153510.153112
- Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26 (1):139-140
- Robinson MD, Oshlack A (2010) A scaling normalization method for differential expression analysis of RNA-seq data. *Genome biology* 11 (3):R25
- Sanabria NM, van Heerden H, Dubery IA (2012) Molecular characterisation and regulation of a *Nicotiana tabacum* S-domain receptor-like kinase gene induced during an early rapid response to lipopolysaccharides. *Gene* 501 (1):39-48
- Sánchez-Sevilla JF, Cruz-Rus E, Valpuesta V, Botella MA, Amaya I (2014) Deciphering gamma-decalactone biosynthesis in strawberry fruit using a combination of genetic mapping, RNA-Seq and eQTL analyses. *BMC Genomics* 15 (1):218. doi:10.1186/1471-2164-15-218
- Schliesky S, Gowik U, Weber APM, Braeutigam A (2012) RNA-seq assembly – Are we there yet? *Frontiers in Plant Science* 3. doi:10.3389/fpls.2012.00220
- Schnable JC, Springer NM, Freeling M (2011) Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. *Proceedings of the National Academy of Sciences* 108 (10):4069-4074
- Schroeder A, Mueller O, Stocker S, Salowsky R, Leiber M, Gassmann M, Lightfoot S, Menzel W, Granzow M, Ragg T (2006) The RIN: an RNA integrity number for assigning integrity values to RNA measurements. *BMC molecular biology* 7 (1):3
- Seqc/Maqc-III Consortium (2014) A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nature biotechnology* 32 (9):903-914
- Shiferaw B, Prasanna BM, Hellin J, Bänziger M (2011) Crops that feed the world 6. Past successes and future challenges to the role played by maize in global food security. *Food Security* 3 (3):307-327
- Shu X, Livingston III DP, Woloshuk CP, Payne GA (2017) Comparative histological and transcriptional analysis of maize kernels infected with *Aspergillus flavus* and *Fusarium verticillioides*. *Frontiers in plant science* 8:2075
- Singh P, Zimmerli LZ (2013) Lectin receptor kinases in plant innate immunity. *Frontiers in plant science* 4:124
- Song XJ, Kuroha T, Ayano M, Furuta T, Nagai K, Komeda N, Segami S, Miura K, Ogawa D, Kamura T (2015) Rare allele of a previously unidentified histone H4 acetyltransferase enhances grain weight, yield, and plant biomass in rice. *Proceedings of the National Academy of Sciences* 112 (1):76-81
- Springer NM, Ying K, Fu Y, Ji T, Yeh C-T, Jia Y, Wu W, Richmond T, Kitzman J, Rosenbaum H (2009) Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. *PLoS genetics* 5 (11):e1000734
- Stanley Kim H, Yu Y, Snesrud EC, Moy LP, Linford LD, Haas BJ, Nierman WC, Quackenbush J (2005) Transcriptional divergence of the duplicated oxidative stress-responsive genes in the *Arabidopsis* genome. *The Plant Journal* 41 (2):212-220. doi:doi:10.1111/j.1365-3113X.2004.02295.x

- Swanson-Wagner RA, Eichten SR, Kumari S, Tiffin P, Stein JC, Ware D, Springer NM (2010) Pervasive gene content variation and copy number variation in maize and its undomesticated progenitor. *Genome research*
- Swart V, Crampton BG, Ridenour JB, Bluhm BH, Olivier NA, Meyer JJM, Berger DK (2017) Complementation of CTB7 in the Maize Pathogen *Cercospora zeina* Overcomes the Lack of In Vitro Cercosporin Production. *Molecular plant-microbe interactions* : MPMI 30 (9):710-724. doi:10.1094/mpmi-03-17-0054-r
- Tarailo - Graovac M, Chen N (2009) Using RepeatMasker to identify repetitive elements in genomic sequences. *Current Protocols in Bioinformatics*:4.10. 11-14.10. 14
- Tefera T, Mugo S, Beyene Y, Karaya H, Tende R (2013) Grain yield, stem borer and disease resistance of new maize hybrids in Kenya. *African Journal of Biotechnology* 10 (23):4777-4783
- Tehon L, Daniels E (1925) Notes on the parasitic fungi of Illinois: II. *Mycologia* 17 (6):240-249
- Trapnell C, Pachter L, Salzberg SL (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 25. doi:10.1093/bioinformatics/btp120
- Trapnell C, Salzberg SL (2009) How to map billions of short reads onto genomes. *Nature biotechnology* 27 (5):455-457. doi:10.1038/nbt0509-455
- Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, Van Baren MJ, Salzberg SL, Wold BJ, 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 TJ, Salzberg SL (2012) Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews Genetics* 13 (1):36-46
- van Dijk EL, Jaszczyszyn Y, Thermes C (2014) Library preparation methods for next-generation sequencing: tone down the bias. *Experimental cell research* 322 (1):12-20
- Van Verk MC, Hickman R, Pieterse CM, Van Wees SC (2013) RNA-Seq: revelation of the messengers. *Trends in plant science* 18 (4):175-179
- Van Bel M, Diels T, Vancaester E, Kreft L, Botzki A, Van de Peer Y, Coppens F, Vandepoele K (2018) PLAZA 4.0: an integrative resource for functional, evolutionary and comparative plant genomics. *Nucleic Acids Research* 46 (D1):D1190-D1196. doi:10.1093/nar/gkx1002
- Vidal EA, Moyano TC, Krouk G, Katari MS, Tanurdzic M, McCombie WR, Coruzzi GM, Gutiérrez RA (2013) Integrated RNA-seq and sRNA-seq analysis identifies novel nitrate-responsive genes in *Arabidopsis thaliana* roots. *BMC genomics* 14 (1):701
- Vigers AJ, Wiedemann S, Roberts WK, Legrand M, Selitrennikoff CP, Fritig B (1992) Thaumatin-like pathogenesis-related proteins are antifungal. *Plant Science* 83 (2):155-161
- Vijay N, Poelstra JW, Künstner A, Wolf JB (2013) Challenges and strategies in transcriptome assembly and differential gene expression quantification. A comprehensive in silico assessment of RNA - seq experiments. *Molecular ecology* 22 (3):620-634
- Wallhead M (2012) Foliar fungicide effects on gray leaf spot and yield of hybrid corn as influenced by application timing, hybrid characteristics and production practices. The Ohio State University,
- Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 10 (1):57-63
- Ward J, Nowell D, Laing M, Whitwell M (1997a) Susceptibility of maize hybrids to grey leaf spot under two tillage systems in South Africa. *South African Journal of Plant and Soil* 14 (4):169-176
- Ward JMJ, Laing MD, Cairns ALP (1997b) Management practices to reduce gray leaf spot of maize. *Crop Science* 37:1257+
- Ward JMJ, Stromberg EL, Nowell DC, Nutter FW (1999) Gray leaf Spot: A Disease of Global Importance in Maize Production. *Plant Disease* 83 (10):884-895. doi:10.1094/PDIS.1999.83.10.884

- Wen W, Li D, Li X, Gao Y, Li W, Li H, Liu J, Liu H, Chen W, Luo J (2014) Metabolome-based genome-wide association study of maize kernel leads to novel biochemical insights. *Nature communications* 5:3438
- Westermann AJ, Gorski SA, Vogel J (2012) Dual RNA-seq of pathogen and host. *Nature Reviews Microbiology* 10 (9):618-630
- Xu L, Wang Y, Liu W, Wang J, Zhu X, Zhang K, Yu R, Wang R, Xie Y, Zhang W, Gong Y, Liu L (2015) De novo sequencing of root transcriptome reveals complex cadmium-responsive regulatory networks in radish (*Raphanus sativus* L.). *Plant science : an international journal of experimental plant biology* 236:313-323. doi:10.1016/j.plantsci.2015.04.015
- Yang Q, He Y, Kabahuma M, Chaya T, Kelly A, Borrego E, Bian Y, El Kasmi F, Yang L, Teixeira P (2017) A gene encoding maize caffeoyl-CoA O-methyltransferase confers quantitative resistance to multiple pathogens. *Nature* 201:7
- Yang SS, Tu ZJ, Cheung F, Xu WW, Lamb JF, Jung H-JG, Vance CP, Gronwald JW (2011) Using RNA-Seq for gene identification, polymorphism detection and transcript profiling in two alfalfa genotypes with divergent cell wall composition in stems. *BMC genomics* 12 (1):199
- Zenoni S, Ferrarini A, Giacomelli E, Xumerle L, Fasoli M, Malerba G, Bellin D, Pezzotti M, Delledonne M (2010) Characterization of Transcriptional Complexity during Berry Development in *Vitis vinifera* Using RNA-Seq. *Plant Physiology* 152 (4):1787-1795. doi:10.1104/pp.109.149716
- Zhang P, Wang Y, Zhang J, Maddock S, Snook M, Peterson T (2003) A maize QTL for silk maysin levels contains duplicated Myb-homologous genes which jointly regulate flavone biosynthesis. *Plant molecular biology* 52 (1):1-15
- Zhang X, Yang Q, Rucker E, Thomason W, Balint-Kurti P (2017) Fine mapping of a quantitative resistance gene for gray leaf spot of maize (*Zea mays* L.) derived from teosinte (*Z. mays* ssp. *parviglumis*). *Theoretical and Applied Genetics* 130 (6):1285-1295. doi:10.1007/s00122-017-2888-2
- Zhao Q-Y, Wang Y, Kong Y-M, Luo D, Li X, Hao P (2011) Optimizing de novo transcriptome assembly from short-read RNA-Seq data: a comparative study. *BMC Bioinformatics* 12 (14):1-12. doi:10.1186/1471-2105-12-s14-s2
- Zhao S, Zhang B, Zhang Y, Gordon W, Du S, Paradis T, Vincent M, von Schack D (2016) Bioinformatics for RNA - Seq Data Analysis. *BIOINFORMATICS-UPDATED FEATURES AND APPLICATIONS*:125
- Zhao W, He X, Hoadley KA, Parker JS, Hayes DN, Perou CM (2014) Comparison of RNA-Seq by poly (A) capture, ribosomal RNA depletion, and DNA microarray for expression profiling. *BMC genomics* 15 (1):1
- Zheng Y, Li B, Zhang X, Lin Y, Liu Z (2010) Occurrence and damage of maize gray leaf spot and its GPS monitoring. *Plant Diseases and Pests* 1 (6):1-3, 11

Appendix A

Table A.1. The list of novel transcripts from maize B73-QTL samples with BLASTP hits to plant sequences, sorted by the log2 Fold Change (FC) of their expression upon *C. zeina* infection.

Assembled transcript	In Un-mapped? ^a	In B73? ^b	In QTL?	Description of closest match ^c	Closest match	%ID ^d	E-value ^e	FC	FDR ^f
TRINITY_DN32513_c0_g1_i1	Yes	No	No	'hypothetical protein BVRB_042430; partial [Beta vulgaris subsp. vulgaris]'	KMS64810.1	57	2.00E-05	- 4.11	2.35E-12
TRINITY_DN12566_c0_g1_i5	No	No	No	'Acyl-CoA synthetase long-chain family member 3 [Zea mays]'	ONM20005.1	95	1.00E-168	- 3.45	7.14E-07
TRINITY_DN12566_c0_g1_i2	No	Yes	No	'protein DA1-related 2 isoform X1 [Zea mays]'	XP_008660302.2	98	0.00E+00	- 2.73	3.24E-04
TRINITY_DN19757_c0_g1_i1	Yes	No	No	'bHLH transcription factor; partial [Zea mays]'	AIB05919.1	100	5.00E-41	- 1.91	2.11E-02
TRINITY_DN12960_c0_g1_i2	No	No	No	'chitinase [Zea mays subsp. parviglumis]'	ACJ62113.1	86	6.00E-115	- 1.86	5.79E-08
TRINITY_DN13942_c2_g1_i7	No	No	No	'uncharacterized protein LOC100275588 [Zea mays]'	NP_001143110.2	100	3.00E-155	- 1.74	8.81E-03
TRINITY_DN2521_c0_g1_i1	Yes	No	No	'uncharacterized protein LOC100276952 [Zea mays]'	NP_001144112.1	99	9.00E-45	- 1.53	6.29E-04
TRINITY_DN12860_c0_g1_i3	Yes	No	No	'tetraspanin family protein [Zea mays]'	NP_001150827.1	87	4.00E-87	- 1.47	2.04E-02
TRINITY_DN7656_c0_g1_i1	No	Yes	No	'Aldose reductase [Zea mays]'	ONM22382.1	86	2.00E-72	- 1.42	9.76E-03

Table A.1 (continued). The list of novel transcripts from maize B73-QTL samples with BLASTP hits to plant sequences, sorted by the log2 Fold Change (FC) of their expression upon *C. zeina* infection.

Assembled transcript	In Un-mapped? ^a	In B73? ^b	In QTL?	Description of closest match ^c	Closest match	%ID ^d	E-value ^e	FC	FDR ^f
TRINITY_DN13955_c1_g1_i20	No	No	No	'hypothetical protein [Zea mays]'	ACG31286.1	67	7.00E-09	-1.41	1.52E-02
TRINITY_DN31347_c0_g1_i1	Yes	No	No	'hypothetical protein SORBI_3001G366100 [Sorghum bicolor]'	OQU92564.1	83	1E-62	-1.39	2.40E-06
TRINITY_DN12089_c0_g1_i2	No	Yes	No	'uncharacterized protein LOC100273868 [Zea mays]'	NP_001141736.1	100	2E-94	-1.33	8.02E-04
TRINITY_DN7644_c0_g1_i3	No	No	No	'hypothetical protein ZEAMMB73_Zm00001d021568 [Zea mays]'	ONM57527.1	98	3E-35	-1.29	1.62E-03
TRINITY_DN11949_c0_g1_i5	No	No	No	'hypothetical protein PAHAL_c00555; partial [Panicum hallii]'	PAN16187.1	84	3E-80	-1.15	1.32E-02
TRINITY_DN4610_c0_g1_i2	No	No	No	'Glycerophosphodiester phosphodiesterase GDPD2 [Zea mays]'	AQK71219.1	88	1E-151	-1.09	4.62E-02
TRINITY_DN11666_c0_g1_i2	No	No	Yes	'NAD(P)-binding Rossmann-fold superfamily protein [Zea mays]'	ONM13850.1	99	8E-44	-1.08	1.00E-02
TRINITY_DN14908_c0_g1_i1	Yes	No	No	'uncharacterized LOC100193608 [Zea mays]'	NP_001132182.1	100	2E-160	-0.90	1.68E-02
TRINITY_DN13727_c0_g1_i7	No	No	No	'predicted protein [Hordeum vulgare subsp. vulgare]'	BAJ87396.1	99	6E-57	-0.90	2.78E-02
TRINITY_DN8286_c	No	Yes	No	'sugar carrier protein C [Zea	NP_001148007.2	97	0E+00	-0.87	3.02E-

Table A.1 (continued). The list of novel transcripts from maize B73-QTL samples with BLASTP hits to plant sequences, sorted by the log2 Fold Change (FC) of their expression upon *C. zeina* infection.

Assembled transcript	In Un-mapped? ^a	In B73? ^b	In QTL?	Description of closest match ^c	Closest match	%ID ^d	E-value ^e	FC	FDR ^f
0_g1_i1				maysj'					03
TRINITY_DN13879_c3_g5_i3	No	Yes	No	'unknown [Zea mays]'	ACN31288.1	85	1E-29	0.74	3.98E-02
TRINITY_DN10925_c0_g1_i3	No	Yes	No	'plastid transcriptionally active 6 [Zea mays]'	AQK56991.1	96	5E-84	0.89	1.13E-02
TRINITY_DN12013_c0_g1_i4	No	No	No	'thioredoxin-like protein 4B [Zea mays]'	NP_001147284.1	100	7E-103	0.93	2.36E-02
TRINITY_DN14007_c1_g3_i2	No	No	No	'uncharacterized protein LOC100384090 [Zea mays]'	XP_020402771.1	100	2E-95	0.95	6.62E-03
TRINITY_DN13306_c1_g2_i3	No	No	No	'uncharacterized protein LOC100273072 [Zea mays]'	NP_001347092.1	98	5E-106	0.99	2.62E-02
TRINITY_DN9994_c1_g2_i1	Yes	Yes	No	'unknown [Zea mays]'	ACN30811.1	100	4E-74	1.03	6.31E-04
TRINITY_DN12319_c0_g1_i4	No	Yes	No	'uncharacterized protein LOC100281641 [Zea mays]'	NP_001148032.1	100	9E-66	1.05	1.67E-04
TRINITY_DN13655_c1_g2_i1	No	No	No	'pyridoxal kinase isoform X2 [Zea mays]'	XP_023157656.1	100	0E+00	1.09	4.39E-05
TRINITY_DN12319_c0_g1_i2	No	No	No	'50S ribosomal protein L28 [Zea mays]'	ACG35998.1	99	1E-60	1.10	3.50E-04
TRINITY_DN17084_c0_g1_i1	No	Yes	No	'1;4-dihydroxy-2-naphthoyl-CoA synthase; peroxisomal-like [Jatropha curcas]'	XP_012085866.1	92	1E-20	1.13	2.92E-02

Table A.1 (continued). The list of novel transcripts from maize B73-QTL samples with BLASTP hits to plant sequences, sorted by the log2 Fold Change (FC) of their expression upon *C. zeina* infection.

Assembled transcript	In Un-mapped? ^a	In B73? ^b	In QTL?	Description of closest match ^c	Closest match	%ID ^d	E-value ^e	FC	FDR ^f
TRINITY_DN43926_c0_g1_i1	No	Yes	No	'hypothetical protein ZEAMMB73_Zm00001d009085 [Zea mays]'	AQK91319.1	100	3E-29	1.14	1.07E-05
TRINITY_DN7575_c0_g1_i1	No	Yes	No	'uncharacterized protein LOC100383971 [Zea mays]'	NP_001170055.1	99	2E-131	1.21	3.22E-03
TRINITY_DN13942_c2_g1_i2	No	No	No	'uncharacterized protein LOC100275588 [Zea mays]'	NP_001143110.2	100	5E-155	1.25	4.68E-02
TRINITY_DN18076_c0_g1_i1	Yes	No	No	'uncharacterized protein LOC100383110 [Zea mays]'	NP_001169249.1	100	1E-41	1.32	3.29E-02
TRINITY_DN13655_c1_g2_i3	No	No	No	'pyridoxal kinase isoform X2 [Zea mays]'	XP_023157656.1	100	0E+00	1.34	5.02E-06
TRINITY_DN13942_c2_g1_i9	No	No	No	'snRNA-activating protein complex subunit [Zea mays]'	AQK98750.1	100	2E-129	1.35	4.98E-02
TRINITY_DN3435_c0_g1_i1	No	No	No	'Remorin family protein [Zea mays]'	ONM54720.1	90	4E-07	1.43	2.57E-02
TRINITY_DN13306_c1_g2_i8	No	Yes	No	'uncharacterized protein LOC100273072 [Zea mays]'	NP_001347092.1	98	1E-105	1.49	2.72E-04
TRINITY_DN10633_c0_g1_i2	No	Yes	No	'Glucose-1-phosphate adenylyltransferase large subunit 1; chloroplastic/amyloplastic [Triticum urartu]'	EMS53815.1	86	6E-26	1.53	1.11E-02
TRINITY_DN13933_c	No	Yes	No	'uncharacterized protein LOC100281496 isoform X1	XP_008643622.1	99	0E+00	1.66	3.48E-

Table A.1 (continued). The list of novel transcripts from maize B73-QTL samples with BLASTP hits to plant sequences, sorted by the log2 Fold Change (FC) of their expression upon *C. zeina* infection.

Assembled transcript	In Un-mapped? ^a	In B73? ^b	In QTL?	Description of closest match ^c	Closest match	%ID ^d	E-value ^e	FC	FDR ^f
0_g1_i3				[Zea mays]					02
TRINITY_DN12566_c0_g1_i9	No	No	No	'Acyl-CoA synthetase long-chain family member 3 [Zea mays]'	ONM19995.1	95	4E-169	1.80	3.19E-05
TRINITY_DN12043_c0_g1_i3	No	Yes	No	'unknown [Zea mays]'	ACN25352.1	84	7E-111	1.94	4.48E-02
TRINITY_DN12566_c0_g1_i6	No	No	No	'protein DA1-related 2 isoform X1 [Zea mays]'	XP_008660302.2	97	0E+00	2.12	7.65E-06
TRINITY_DN42734_c0_g1_i1	Yes	No	No	'high-affinity potassium transporter [Zea mays]'	AHH35062.1	100	8E-44	2.15	1.19E-02
TRINITY_DN13995_c0_g1_i8	No	No	No	'Eukaryotic translation initiation factor 3 subunit E [Zea mays]'	ONM20648.1	91	3E-131	2.84	3.75E-02
TRINITY_DN14009_c0_g1_i20	No	No	No	'CCT motif family protein [Zea mays]'	NP_001151864.1	100	3E-28	2.92	6.29E-06
TRINITY_DN13822_c5_g1_i10	No	No	No	'uncharacterized protein LOC100384367 [Zea mays]'	NP_001170384.1	88	2E-16	6.76	2.25E-06
TRINITY_DN14031_c0_g1_i11	No	No	No	'hypothetical protein SETIT_000911mg [Setaria italica]'	KQL07844.1	92	2E-166	7.66	1.91E-08
TRINITY_DN1098_c0_g1_i1 m.9469	Only in Unmapped	No	No	IDD9 [Zea mays subsp. mays]	gb AJF23770.1	98	2E-40	ns*	ns*
TRINITY_DN109_c	Only in	No	No	unknown [Medicago truncatula]	gb ACJ84693.1	77	6E-37	ns*	ns*

Table A.1 (continued). The list of novel transcripts from maize B73-QTL samples with BLASTP hits to plant sequences, sorted by the log₂ Fold Change (FC) of their expression upon *C. zeina* infection.

Assembled transcript	In Un-mapped? ^a	In B73? ^b	In QTL?	Description of closest match ^c	Closest match	%ID ^d	E-value ^e	FC	FDR ^f
0_g1_i1 m.10238	Unmapped								
TRINITY_DN1294_c0_g1_i1 m.299	Only in Unmapped	No	No	predicted protein [Hordeum vulgare subsp. vulgare]	dbj BAJ93702.1	84	3E-34	ns*	ns*
TRINITY_DN171_c0_g1_i1 m.10226	Only in Unmapped	No	No	Os03g0801600, partial [Oryza sativa Japonica Group]	dbj BAS86895.1	71	4E-22	ns*	ns*
TRINITY_DN1869_c0_g1_i1 m.6706	Only in Unmapped	No	No	amidase precursor [Zea mays]	ref NP_001150232.2	100	5E-38	ns*	ns*
TRINITY_DN1912_c0_g1_i1 m.11277	Only in Unmapped	No	No	Receptor-like serine/threonine-protein kinase SD1-8 [Zea mays]	gb ONM13388.1	88	2E-53	ns*	ns*
TRINITY_DN1949_c0_g1_i1 m.11446	Only in Unmapped	No	No	predicted protein [Hordeum vulgare subsp. vulgare]	dbj BAJ96087.1	99	6E-50	ns*	ns*
TRINITY_DN2238_c0_g1_i1 m.8564	Only in Unmapped	No	No	cysteine proteinase 1 precursor [Zea mays]	ref NP_001105685.1	100	2E-47	ns*	ns*
TRINITY_DN227_c0_g1_i1 m.11782	Only in Unmapped	No	No	predicted protein [Hordeum vulgare subsp. vulgare]	dbj BAJ96848.1	89	1E-41	ns*	ns*
TRINITY_DN240_c0_g1_i1 m.11825	Only in Unmapped	No	No	ubiquitin-like protein, partial [Vanilla planifolia]	gb AEI29169.1	99	9E-43	ns*	ns*
TRINITY_DN2453_c0_g1_i1 m.6089	Only in Unmapped	No	No	Glyceraldehyde-3-phosphate dehydrogenase 2, cytosolic [Zea mays]	gb AQK78434.1	100	4E-40	ns*	ns*
TRINITY_DN2458_c	Only in	No	No	Acid phosphatase/vanadium-dependent haloperoxidase-	gb AQK91329.1	95	1E-02	ns*	ns*

Table A.1 (continued). The list of novel transcripts from maize B73-QTL samples with BLASTP hits to plant sequences, sorted by the log2 Fold Change (FC) of their expression upon *C. zeina* infection.

Assembled transcript	In Un-mapped? ^a	In B73? ^b	In QTL?	Description of closest match ^c	Closest match	%ID ^d	E-value ^e	FC	FDR ^f
0_g1_i1 m.6152	Unmapped			related protein [Zea mays]					
TRINITY_DN2609_c0_g1_i1 m.15465	Only in Unmapped	No	No	Receptor of activated protein kinase C 1A, component of 40S small ribosomal subunit [Ostreococcus lucimarinus CCE9901]	ref XP_001421096.1	80	1E-35	ns*	ns*
TRINITY_DN2649_c0_g1_i1 m.15460	Only in Unmapped	No	No	predicted protein [Micromonas commoda]	ref XP_002508800.1	73	2E-26	ns*	ns*
TRINITY_DN2684_c0_g1_i1 m.15624	Only in Unmapped	No	No	high-affinity potassium transporter [Zea mays]	gb AHH35062.1	94	2E-29	ns*	ns*
TRINITY_DN3012_c0_g1_i1 m.12378	Only in Unmapped	No	No	glutathione transferase20 [Zea mays]	ref NP_001104986.1	100	9E-96	ns*	ns*
TRINITY_DN3042_c0_g1_i1 m.12256	Only in Unmapped	No	No	Actin, non-muscle 6.2, partial [Noccaea caerulea]	gb JAU73821.1	97	1E-47	ns*	ns*
TRINITY_DN3312_c0_g1_i1 m.1388	Only in Unmapped	No	No	unknown [Zea mays]	gb ACN33910.1	100	3E-169	ns*	ns*
TRINITY_DN333_c0_g1_i1 m.9609	Only in Unmapped	Yes	No	pyruvate orthophosphate dikinase1 [Zea mays]	gb AQK85899.1	100	2E-35	ns*	ns*
TRINITY_DN3392_c0_g1_i1 m.1465	Only in Unmapped	No	No	60S ribosomal protein L7A family protein [Klebsormidium flaccidum]	dbj GAQ82738.1	49	8E-31	ns*	ns*
TRINITY_DN3476_c0_g1_i1 m.4622	Only in Unmapped	No	No	BnaCnng41770D [Brassica napus]	emb CDY63262.1	39	3E-11	ns*	ns*

Table A.1 (continued). The list of novel transcripts from maize B73-QTL samples with BLASTP hits to plant sequences, sorted by the log2 Fold Change (FC) of their expression upon *C. zeina* infection.

Assembled transcript	In Un-mapped? ^a	In B73? ^b	In QTL?	Description of closest match ^c	Closest match	%ID ^d	E-value ^e	FC	FDR ^f
TRINITY_DN3515_c0_g1_i1 m.520	Only in Unmapped	No	No	60S ribosomal protein L8 [Anthurium amnicola]	gb JAT43573.1	83	3E-65	ns*	ns*
TRINITY_DN3639_c0_g1_i1 m.14696	Only in Unmapped	No	No	PREDICTED: ubiquitin [Elaeis guineensis]	ref XP_019708906.1	100	5E-40	ns*	ns*
TRINITY_DN3787_c0_g1_i1 m.10949	Only in Unmapped	No	No	predicted protein [Hordeum vulgare subsp. vulgare]	dbj BAJ94184.1	88	1E-41	ns*	ns*
TRINITY_DN3805_c0_g1_i1 m.15236	Only in Unmapped	No	No	PREDICTED: laccase-2-like [Pyrus x bretschneideri]	ref XP_009350642.1	70	2E-47	ns*	ns*
TRINITY_DN3812_c0_g1_i1 m.15396	Only in Unmapped	No	No	AN12 Zinc finger A20 and AN1 domain-containing protein [Zea mays]	ref NP_001106259.1	100	2E-42	ns*	ns*
TRINITY_DN3882_c0_g1_i1 m.15432	Only in Unmapped	No	No	G-type lectin S-receptor-like serine/threonine-protein kinase B120 [Zea mays]	gb AQK46433.1	86	1E-110	ns*	ns*
TRINITY_DN3906_c0_g1_i1 m.12547	Only in Unmapped	No	No	unknown [Zea mays]	gb ACF82437.1	100	1E-36	ns*	ns*
TRINITY_DN3940_c0_g1_i1 m.12582	Only in Unmapped	No	No	PREDICTED: 40S ribosomal protein S20-1 [Oryza sativa Japonica Group]	ref XP_015614818.1	100	1E-41	ns*	ns*
TRINITY_DN4163_c0_g1_i1 m.14842	Only in Unmapped	No	No	G-type lectin S-receptor-like serine/threonine-protein kinase B120 [Zea mays]	gb AQK46433.1	94	3E-21	ns*	ns*
TRINITY_DN4167_c	Only in	No	No	PREDICTED: probable cysteine protease RD21B	ref XP_019239500.1	38	2E-05	ns*	ns*

Table A.1 (continued). The list of novel transcripts from maize B73-QTL samples with BLASTP hits to plant sequences, sorted by the log2 Fold Change (FC) of their expression upon *C. zeina* infection.

Assembled transcript	In Un-mapped? ^a	In B73? ^b	In QTL?	Description of closest match ^c	Closest match	%ID ^d	E-value ^e	FC	FDR ^f
0_g1_i1 m.14800	Unmapped			[<i>Nicotiana attenuata</i>]					
TRINITY_DN4187_c0_g1_i1 m.14936	Only in Unmapped	No	No	Polyubiquitin, partial [<i>Noccaea caerulescens</i>]	gb JAU22157.1	98	1E-57	ns*	ns*
TRINITY_DN4187_c0_g1_i1 m.14936	Only in Unmapped	No	No	Polyubiquitin, partial [<i>Noccaea caerulescens</i>]	gb JAU22157.1	95	5E-03	ns*	ns*
TRINITY_DN4288_c0_g1_i1 m.15167	Only in Unmapped	No	No	predicted protein [<i>Hordeum vulgare</i> subsp. <i>vulgare</i>]	dbj BAK05398.1	92	6E-42	ns*	ns*
TRINITY_DN4529_c0_g1_i1 m.3827	Only in Unmapped	No	No	Polyubiquitin 11 [<i>Zostera marina</i>]	gb KMZ73934.1	100	2E-57	ns*	ns*
TRINITY_DN4529_c0_g1_i1 m.3827	Only in Unmapped	No	No	Polyubiquitin 11 [<i>Zostera marina</i>]	gb KMZ73934.1	97	3E-44	ns*	ns*
TRINITY_DN462_c0_g1_i1 m.4794	Only in Unmapped	No	No	PREDICTED: 60S ribosomal protein L4-1-like [<i>Erythranthe guttata</i>]	ref XP_012831933.1	80	2E-58	ns*	ns*
TRINITY_DN485_c0_g1_i2 m.4943	Only in Unmapped	No	No	PREDICTED: ubiquitin [<i>Brassica napus</i>]	ref XP_013745166.1	99	3E-50	ns*	ns*
TRINITY_DN532_c0_g1_i1 m.3215	Only in Unmapped	No	No	dehydrin 13 [<i>Zea mays</i>]	gb ACG47851.1	100	2E-23	ns*	ns*
TRINITY_DN5447_c0_g1_i1 m.13539	Only in Unmapped	No	No	Putative leucine-rich repeat receptor-like protein kinase family protein [<i>Zea mays</i>]	gb AQK46158.1	89	4E-40	ns*	ns*
TRINITY_DN546_c	Only in	No	No	unknown [<i>Zea mays</i>]	gb ACF84747.1	100	5E-24	ns*	ns*

Table A.1 (continued). The list of novel transcripts from maize B73-QTL samples with BLASTP hits to plant sequences, sorted by the log2 Fold Change (FC) of their expression upon *C. zeina* infection.

Assembled transcript	In Un-mapped? ^a	In B73? ^b	In QTL?	Description of closest match ^c	Closest match	%ID ^d	E-value ^e	FC	FDR ^f
0_g1_i1 m.3112	Unmapped								
TRINITY_DN546_c0_g1_i2 m.3118	Only in Unmapped	No	No	unknown [Zea mays]	gb ACF84747.1	100	0E+00	ns*	ns*
TRINITY_DN5470_c0_g1_i1 m.13534	Only in Unmapped	No	No	PREDICTED: fructokinase-1-like [Phoenix dactylifera]	ref XP_008790605.1	40	6E-06	ns*	ns*
TRINITY_DN5514_c0_g1_i1 m.1876	Only in Unmapped	Yes	No	thiamine thiazole synthase 1, chloroplastic precursor [Zea mays]	ref NP_001105696.1	100	4E-42	ns*	ns*
TRINITY_DN5521_c0_g1_i1 m.2036	Only in Unmapped	No	No	RecName: Full=Ferredoxin-5, chloroplastic; AltName: Full=Ferredoxin V; Short=Fd V; Flags: Precursor	sp P27789.1 FER5_MAIZE	100	4E-32	ns*	ns*
TRINITY_DN5596_c0_g1_i1 m.1938	Only in Unmapped	No	No	glycine-rich RNA-binding protein 2 [Zea mays]	gb ACG26657.1	92	2E-31	ns*	ns*
TRINITY_DN569_c0_g1_i1 m.3141	Only in Unmapped	Yes	No	PREDICTED: chlorophyll a-b binding protein 1D [Glycine max]	ref XP_006587015.1	99	2E-44	ns*	ns*
TRINITY_DN5708_c0_g1_i1 m.12112	Only in Unmapped	No	No	AN12 Zinc finger A20 and AN1 domain-containing protein [Zea mays]	ref NP_001106259.1	100	7E-53	ns*	ns*
TRINITY_DN6010_c0_g1_i1 m.3269	Only in Unmapped	No	No	Histone H2AX [Zea mays]	gb AQL03536.1	84	2E-18	ns*	ns*
TRINITY_DN6010_c	Only in	No	No	unknown, partial [Glycine max]	gb ACU16929.1	50	4E-08	ns*	ns*

Table A.1 (continued). The list of novel transcripts from maize B73-QTL samples with BLASTP hits to plant sequences, sorted by the log2 Fold Change (FC) of their expression upon *C. zeina* infection.

Assembled transcript	In Un-mapped? ^a	In B73? ^b	In QTL?	Description of closest match ^c	Closest match	%ID ^d	E-value ^e	FC	FDR ^f
0_g1_i1 m.3270	Unmapped								
TRINITY_DN6038_c0_g1_i1 m.3431	Only in Unmapped	No	No	predicted protein [Hordeum vulgare subsp. vulgare]	dbj BAK02581.1	74	2E-21	ns*	ns*
TRINITY_DN603_c0_g1_i1 m.13313	Only in Unmapped	Yes	No	abscisic stress ripening protein 1 [Zea mays]	gb ACG35620.1	100	3E-60	ns*	ns*
TRINITY_DN603_c0_g1_i2 m.13316	Only in Unmapped	No	No	abscisic stress ripening protein 1 [Zea mays]	gb ACG35620.1	100	3E-60	ns*	ns*
TRINITY_DN617_c0_g1_i1 m.13296	Only in Unmapped	No	No	unknown [Zea mays]	gb ACN29187.1	100	0E+00	ns*	ns*
TRINITY_DN696_c0_g1_i1 m.13285	Only in Unmapped	No	No	pyruvate orthophosphate dikinase1 [Zea mays]	gb AQK85875.1	100	2E-84	ns*	ns*
TRINITY_DN781_c0_g1_i1 m.8139	Only in Unmapped	No	No	actin-like protein, partial [Miscanthus sinensis]	gb AFC88819.1	100	2E-36	ns*	ns*
TRINITY_DN799_c0_g1_i1 m.7975	Only in Unmapped	No	No	Glucose-repressible gene protein, partial [Anthurium amnicola]	gb JAT62563.1	97	3E-36	ns*	ns*
TRINITY_DN799_c0_g1_i2 m.7979	Only in Unmapped	No	No	Glucose-repressible gene protein, partial [Anthurium amnicola]	gb JAT62563.1	91	1E-34	ns*	ns*
TRINITY_DN816_c0_g1_i1 m.7180	Only in Unmapped	No	No	40S ribosomal protein S3-1, partial [Noccaea caerulea]	gb JAU88929.1	73	4E-44	ns*	ns*
TRINITY_DN816_c	Only in	No	No	40S ribosomal protein S3-1,	gb JAU88929.1	74	1E-44	ns*	ns*

Table A.1 (continued). The list of novel transcripts from maize B73-QTL samples with BLASTP hits to plant sequences, sorted by the log2 Fold Change (FC) of their expression upon *C. zeina* infection.

Assembled transcript	In Un-mapped? ^a	In B73? ^b	In QTL?	Description of closest match ^c	Closest match	%ID ^d	E-value ^e	FC	FDR ^f
0_g1_i2 m.7181	Unmapped			partial [Noccaea caerulescens]					
TRINITY_DN856_c0_g1_i1 m.7202	Only in Unmapped	No	No	Actin, plasmodial isoform, partial [Anthurium amnicola]	gb JAT65395.1	98	2E-128	ns*	ns*
TRINITY_DN856_c0_g1_i2 m.7207	Only in Unmapped	No	No	actin 5-like protein, partial [Miscanthus sinensis]	gb AFC88820.1	99	6E-95	ns*	ns*
TRINITY_DN863_c0_g1_i1 m.6900	Only in Unmapped	No	No	Actin, plasmodial isoform, partial [Anthurium amnicola]	gb JAT65395.1	99	3E-74	ns*	ns*
TRINITY_DN863_c0_g1_i2 m.6902	Only in Unmapped	No	No	Actin, plasmodial isoform, partial [Anthurium amnicola]	gb JAT65395.1	99	8E-41	ns*	ns*
TRINITY_DN863_c0_g1_i3 m.6903	Only in Unmapped	No	No	Actin, plasmodial isoform, partial [Anthurium amnicola]	gb JAT65395.1	99	7E-73	ns*	ns*
TRINITY_DN863_c0_g1_i4 m.6905	Only in Unmapped	No	No	Actin, plasmodial isoform, partial [Anthurium amnicola]	gb JAT65395.1	99	7E-57	ns*	ns*
TRINITY_DN863_c0_g1_i5 m.6906	Only in Unmapped	No	No	Actin, plasmodial isoform, partial [Anthurium amnicola]	gb JAT65395.1	99	2E-90	ns*	ns*
TRINITY_DN863_c0_g1_i6 m.6908	Only in Unmapped	No	No	Actin, plasmodial isoform, partial [Anthurium amnicola]	gb JAT65395.1	99	9E-41	ns*	ns*
TRINITY_DN867_c0_g1_i1 m.6817	Only in Unmapped	No	No	predicted protein [Hordeum vulgare subsp. vulgare]	dbj BAK00960.1	100	1E-161	ns*	ns*
TRINITY_DN867_c0_g1_i3 m.6820	Only in Unmapped	No	No	predicted protein [Hordeum vulgare subsp. vulgare]	dbj BAK00960.1	100	0E+00	ns*	ns*
TRINITY_DN867_c	Only in	No	No	predicted protein [Hordeum	dbj BAK00960.1	100	0E+00	ns*	ns*

Table A.1 (continued). The list of novel transcripts from maize B73-QTL samples with BLASTP hits to plant sequences, sorted by the log2 Fold Change (FC) of their expression upon *C. zeina* infection.

Assembled transcript	In Un-mapped? ^a	In B73? ^b	In QTL?	Description of closest match ^c	Closest match	%ID ^d	E-value ^e	FC	FDR ^f
0_g1_i4 m.6822	Unmapped			vulgare subsp. vulgare]					
TRINITY_DN867_c0_g1_i5 m.6824	Only in Unmapped	No	No	predicted protein [Hordeum vulgare subsp. vulgare]	dbj BAK00960.1	100	0E+00	ns*	ns*
TRINITY_DN871_c0_g1_i3 m.7318	Only in Unmapped	No	No	Glucose-repressible gene protein, partial [Anthurium amnicola]	gb JAT62563.1	73	1E-03	ns*	ns*
TRINITY_DN106_c0_g1_i1 m.1471	Only in Unmapped	No	No	KH domain-containing protein [Zea mays]	ref NP_001332948.1	100	2E-113	ns*	ns*
TRINITY_DN106_c0_g1_i1 m.1471	Only in Unmapped	No	No	KH domain-containing protein [Zea mays]	ref NP_001332948.1	46	2E-10	ns*	ns*
TRINITY_DN106_c0_g1_i2 m.1476	Only in Unmapped	No	No	KH domain-containing protein [Zea mays]	ref NP_001332948.1	100	2E-113	ns*	ns*
TRINITY_DN106_c0_g1_i2 m.1476	Only in Unmapped	No	No	KH domain-containing protein [Zea mays]	ref NP_001332948.1	46	2E-10	ns*	ns*
TRINITY_DN132_c0_g1_i1 m.1363	Only in Unmapped	No	No	unknown [Zea mays]	gb ACN29187.1	100	6E-126	ns*	ns*
TRINITY_DN132_c0_g1_i2 m.1365	Only in Unmapped	No	No	unknown [Zea mays]	gb ACN29187.1	100	0E+00	ns*	ns*
TRINITY_DN172_c0_g1_i2 m.1520	Only in Unmapped	No	No	unknown [Zea mays]	gb ACN26848.1	100	2E-05	ns*	ns*
TRINITY_DN200_c0_g1_i1 m.1829	Only in Unmapped	No	No	unknown [Zea mays]	gb ACF84747.1	100	0E+00	ns*	ns*

Table A.1 (continued). The list of novel transcripts from maize B73-QTL samples with BLASTP hits to plant sequences, sorted by the log2 Fold Change (FC) of their expression upon *C. zeina* infection.

Assembled transcript	In Un-mapped? ^a	In B73? ^b	In QTL?	Description of closest match ^c	Closest match	%ID ^d	E-value ^e	FC	FDR ^f
TRINITY_DN200_c0_g1_i2 m.1837	Only in Unmapped	No	No	unknown [Zea mays]	gb ACF84747.1	100	2E-23	ns*	ns*
TRINITY_DN200_c0_g1_i4 m.1844	Only in Unmapped	No	No	unknown [Zea mays]	gb ACF84747.1	99	5E-46	ns*	ns*
TRINITY_DN220_c0_g1_i1 m.1797	Only in Unmapped	Yes	No	Os03g0718150, partial [Oryza sativa Japonica Group]	dbj BAS86095.1	72	2E-36	ns*	ns*
TRINITY_DN245_c0_g1_i1 m.1772	Only in Unmapped	No	No	predicted protein [Hordeum vulgare subsp. vulgare]	dbj BAK00513.1	83	5E-28	ns*	ns*
TRINITY_DN280_c0_g1_i1 m.1762	Only in Unmapped	No	No	unknown [Zea mays]	gb ACL54835.1	100	1E-79	ns*	ns*
TRINITY_DN303_c0_g1_i1 m.1121	Only in Unmapped	Yes	No	RNA polymerase sigma factor 1, partial [Zea mays]	gb AAC78467.1	100	7E-32	ns*	ns*
TRINITY_DN305_c0_g1_i1 m.1066	Only in Unmapped	No	No	Putative leucine-rich repeat receptor-like protein kinase family protein [Zea mays]	gb AQK46176.1	84	8E-27	ns*	ns*
TRINITY_DN356_c0_g1_i1 m.1132	Only in Unmapped	No	No	probable E3 ubiquitin-protein ligase ARI8 [Zea mays]	ref NP_001315295.1	100	2E-46	ns*	ns*
TRINITY_DN374_c0_g1_i1 m.1081	Only in Unmapped	No	No	unnamed protein product [Ostreococcus tauri]	ref XP_003075092.1	43	4E-08	ns*	ns*
TRINITY_DN388_c0_g1_i1 m.1110	Only in Unmapped	No	No	unknown [Zea mays]	gb ACN35719.1	84	2E-40	ns*	ns*
TRINITY_DN436_c	Only in	No	No	Protein S-acyltransferase 10	gb AQK54205.1	42	6E-08	ns*	ns*

Table A.1 (continued). The list of novel transcripts from maize B73-QTL samples with BLASTP hits to plant sequences, sorted by the log2 Fold Change (FC) of their expression upon *C. zeina* infection.

Assembled transcript	In Un-mapped? ^a	In B73? ^b	In QTL?	Description of closest match ^c	Closest match	%ID ^d	E-value ^e	FC	FDR ^f
0_g1_i1 m.373	Unmapped			[Zea mays]					
TRINITY_DN437_c0_g1_i1 m.202	Only in Unmapped	Yes	No	E3 ubiquitin-protein ligase RGLG1 [Zea mays]	gb ONM09641.1	60	6E-07	ns*	ns*
TRINITY_DN491_c0_g1_i1 m.230	Only in Unmapped	No	No	KH domain-containing protein [Zea mays]	ref NP_001332948.1	100	2E-22	ns*	ns*
TRINITY_DN527_c0_g1_i1 m.56	Only in Unmapped	No	No	Os04g0633950 [Oryza sativa Japonica Group]	dbj BAS91198.1	51	1E-07	ns*	ns*
TRINITY_DN553_c0_g1_i1 m.170	Only in Unmapped	No	No	amidase precursor [Zea mays]	ref NP_001150232.2	100	4E-45	ns*	ns*
TRINITY_DN554_c0_g1_i1 m.144	Only in Unmapped	No	No	1-aminocyclopropane-1-carboxylate oxidase homolog [Zea mays]	ref XP_020400957.1	95	3E-33	ns*	ns*
TRINITY_DN592_c0_g1_i1 m.50	Only in Unmapped	No	No	unknown [Zea mays]	gb ACN33910.1	100	2E-53	ns*	ns*
TRINITY_DN644_c0_g1_i1 m.2102	Only in Unmapped	No	No	O-fucosyltransferase family protein [Zea mays]	gb ONM55849.1	66	2E-10	ns*	ns*
TRINITY_DN658_c0_g1_i1 m.2190	Only in Unmapped	No	No	glycine-rich protein [Zea mays]	gb AQK78348.1	96	1E-46	ns*	ns*
TRINITY_DN665_c0_g1_i1 m.2095	Only in Unmapped	Yes	No	PREDICTED: 40S ribosomal protein S20-1 [Oryza sativa Japonica Group]	ref XP_015614818.1	100	4E-41	ns*	ns*
TRINITY_DN681_c	Only in	No	No	Receptor-like serine/threonine-protein kinase SD1-8 [Zea	gb ONM13390.1	91	1E-53	ns*	ns*

Table A.1 (continued). The list of novel transcripts from maize B73-QTL samples with BLASTP hits to plant sequences, sorted by the log2 Fold Change (FC) of their expression upon *C. zeina* infection.

Assembled transcript	In Un-mapped? ^a	In B73? ^b	In QTL?	Description of closest match ^c	Closest match	%ID ^d	E-value ^e	FC	FDR ^f
0_g1_i1 m.2245	Unmapped			mays]					
TRINITY_DN694_c0_g1_i1 m.2041	Only in Unmapped	No	No	high-affinity potassium transporter [Zea mays]	gb AHH35062.1	100	4E-48	ns*	ns*
TRINITY_DN6_c0_g1_i1 m.1023	Only in Unmapped	No	No	anthranilate synthase component II [Zea mays]	ref NP_001152611.1	100	7E-114	ns*	ns*
TRINITY_DN703_c0_g1_i1 m.745	Only in Unmapped	No	No	ribosomal protein S9 component of cytosolic 80S ribosome and 40S small subunit [Coccomyxa subellipsoidea C-169]	ref XP_005648482.1	84	3E-48	ns*	ns*
TRINITY_DN740_c0_g1_i1 m.634	Only in Unmapped	No	No	unknown [Zea mays]	gb ACN33910.1	100	2E-154	ns*	ns*
TRINITY_DN746_c0_g1_i1 m.787	Only in Unmapped	No	No	high-affinity potassium transporter [Zea mays]	gb AHH35062.1	100	2E-94	ns*	ns*
TRINITY_DN768_c0_g1_i1 m.779	Only in Unmapped	No	No	Transcription factor bHLH68 [Zea mays]	gb AQK44493.1	72	3E-28	ns*	ns*
TRINITY_DN79_c0_g1_i1 m.1012	Only in Unmapped	No	No	40S Ribosomal protein SA family protein [Klebsormidium flaccidum]	dbj GAQ80485.1	80	5E-51	ns*	ns*
TRINITY_DN807_c0_g1_i1 m.512	Only in Unmapped	No	No	putative polyprotein [Zea mays]	gb AAL75999.1 AF466646_7	61	2E-27	ns*	ns*
TRINITY_DN809_c0_g1_i1 m.577	Only in Unmapped	No	No	Protein FATTY ACID EXPORT 1 chloroplastic [Zea mays]	gb AQK44205.1	94	8E-34	ns*	ns*

Table A.1 (continued). The list of novel transcripts from maize B73-QTL samples with BLASTP hits to plant sequences, sorted by the log2 Fold Change (FC) of their expression upon *C. zeina* infection.

Assembled transcript	In Un-mapped? ^a	In B73? ^b	In QTL?	Description of closest match ^c	Closest match	%ID ^d	E-value ^e	FC	FDR ^f
TRINITY_DN813_c0_g1_i1 m.443	Only in Unmapped	No	No	mitochondrial protein - maize	pir A24522	100	5E-49	ns*	ns*
TRINITY_DN827_c0_g1_i1 m.591	Only in Unmapped	No	No	Protein ALTERED XYLOGLUCAN 4-like [Dichantherium oligosanthes]	gb OEL28198.1	88	2E-18	ns*	ns*
TRINITY_DN827_c0_g1_i1 m.591	Only in Unmapped	No	No	Protein ALTERED XYLOGLUCAN 4-like [Dichantherium oligosanthes]	gb OEL28198.1	64	1E-11	ns*	ns*
TRINITY_DN850_c0_g1_i1 m.582	Only in Unmapped	Yes	No	adenylate kinase [Zea mays]	gb ACG27659.1	93	3E-07	ns*	ns*
TRINITY_DN868_c0_g1_i1 m.628	Only in Unmapped	No	Yes	G-type lectin S-receptor-like serine/threonine-protein kinase B120 [Zea mays]	gb AQK46433.1	85	2E-108	ns*	ns*
TRINITY_DN886_c0_g1_i1 m.460	Only in Unmapped	No	No	putative protein kinase superfamily protein [Zea mays]	ref NP_001146603.1	100	3E-69	ns*	ns*
TRINITY_DN890_c0_g1_i1 m.549	Only in Unmapped	No	No	KH domain-containing protein [Zea mays]	ref NP_001332948.1	100	3E-126	ns*	ns*
TRINITY_DN890_c0_g1_i1 m.549	Only in Unmapped	No	No	KH domain-containing protein [Zea mays]	ref NP_001332948.1	42	3E-16	ns*	ns*
TRINITY_DN905_c0_g1_i1 m.1708	Only in Unmapped	No	No	Os07g0102150, partial [Oryza sativa Japonica Group]	dbj BAS99686.1	60	7E-13	ns*	ns*
TRINITY_DN915_c0_g1_i1 m.1634	Only in Unmapped	No	No	Protein NRT1/ PTR FAMILY 2.13 [Dichantherium	gb OEL16316.1	71	9E-05	ns*	ns*

Table A.1 (continued). The list of novel transcripts from maize B73-QTL samples with BLASTP hits to plant sequences, sorted by the log2 Fold Change (FC) of their expression upon *C. zeina* infection.

Assembled transcript	In Un-mapped? ^a	In B73? ^b	In QTL?	Description of closest match ^c	Closest match	%ID ^d	E-value ^e	FC	FDR ^f
oligosanthes]									
TRINITY_DN927_c0_g1_i1 m.1645	Only in Unmapped	No	No	PREDICTED: serine hydroxymethyltransferase, mitochondrial-like [Malus domestica]	ref XP_008384294.1	98	4E-65	ns*	ns*
TRINITY_DN932_c0_g1_i1 m.1532	Only in Unmapped	No	No	unknown [Zea mays]	gb ACL54350.1	100	1E-64	ns*	ns*
TRINITY_DN944_c0_g1_i1 m.1618	Only in Unmapped	Yes	No	IDD9 [Zea mays subsp. mays]	gb AJF23770.1	98	6E-40	ns*	ns*
TRINITY_DN103_c0_g1_i1 m.1490	Only in Unmapped	No	No	hypothetical protein LE_TR4279_c0_g1_i1_g.14007, partial [Noccaea caerulescens]	gb JAU59768.1	48	1E-12	ns*	ns*
TRINITY_DN105_c0_g1_i1 m.1340	Only in Unmapped	No	No	uncharacterized protein LOC100279124 [Zea mays]	ref NP_001145637.2	100	1E-130	ns*	ns*
TRINITY_DN125_c0_g1_i1 m.1387	Only in Unmapped	No	No	hypothetical protein SORBI_009G142200 [Sorghum bicolor]	gb KXG22031.1	83	9E-20	ns*	ns*
TRINITY_DN125_c0_g1_i2 m.1393	Only in Unmapped	No	No	hypothetical protein SORBI_009G142200 [Sorghum bicolor]	gb KXG22031.1	83	9E-20	ns*	ns*
TRINITY_DN178_c0_g1_i1 m.1405	Only in Unmapped	No	No	uncharacterized protein LOC103638366 [Zea mays]	ref XP_008659520.1	72	2E-10	ns*	ns*

Table A.1 (continued). The list of novel transcripts from maize B73-QTL samples with BLASTP hits to plant sequences, sorted by the log2 Fold Change (FC) of their expression upon *C. zeina* infection.

Assembled transcript	In Un-mapped? ^a	In B73? ^b	In QTL?	Description of closest match ^c	Closest match	%ID ^d	E-value ^e	FC	FDR ^f
TRINITY_DN192_c0_g1_i1 m.1312	Only in Unmapped	No	No	hypothetical protein SORBIDRAFT_06g001580 [Sorghum bicolor]	ref XP_002446089.1	85	1E-97	ns*	ns*
TRINITY_DN200_c0_g1_i3 m.1841	Only in Unmapped	No	No	uncharacterized protein LOC100284429 [Zea mays]	ref NP_001150796.1	99	3E-114	ns*	ns*
TRINITY_DN220_c0_g1_i2 m.1799	Only in Unmapped	No	No	hypothetical protein GA_TR19271_c0_g1_i1_g.62644, partial [Noccaea caerulescens]	gb JAU23678.1	69	1E-32	ns*	ns*
TRINITY_DN220_c0_g1_i4 m.1806	Only in Unmapped	Yes	No	hypothetical protein GA_TR19271_c0_g1_i1_g.62644, partial [Noccaea caerulescens]	gb JAU23678.1	79	2E-23	ns*	ns*
TRINITY_DN244_c0_g1_i1 m.1989	Only in Unmapped	No	No	uncharacterized protein LOC100192087 [Zea mays]	ref NP_001130982.1	99	9E-75	ns*	ns*
TRINITY_DN260_c0_g1_i1 m.1759	Only in Unmapped	No	No	uncharacterized protein LOC100381671 [Zea mays]	ref NP_001167955.1	64	5E-09	ns*	ns*
TRINITY_DN281_c0_g1_i1 m.1962	Only in Unmapped	No	No	hypothetical protein GA_TR8221_c0_g1_i1_g.26592, partial [Noccaea caerulescens]	gb JAU09906.1	45	8E-06	ns*	ns*
TRINITY_DN284_c0_g1_i1 m.1761	Only in Unmapped	No	No	uncharacterized protein LOC100383368 [Zea mays]	ref NP_001335990.1	100	2E-49	ns*	ns*
TRINITY_DN313_c	Only in	No	No	hypothetical protein	ref XP_002467980.1	97	0E+00	ns*	ns*

Table A.1 (continued). The list of novel transcripts from maize B73-QTL samples with BLASTP hits to plant sequences, sorted by the log2 Fold Change (FC) of their expression upon *C. zeina* infection.

Assembled transcript	In Un-mapped? ^a	In B73? ^b	In QTL?	Description of closest match ^c	Closest match	%ID ^d	E-value ^e	FC	FDR ^f
0_g1_i1 m.1288	Unmapped			SORBIDRAFT_01g037483 [Sorghum bicolor]					
TRINITY_DN319_c0_g1_i1 m.1116	Only in Unmapped	No	No	uncharacterized protein LOC109942331 [Zea mays]	ref XP_020399838.1	85	2E-10	ns*	ns*
TRINITY_DN326_c0_g1_i1 m.1243	Only in Unmapped	No	No	uncharacterized protein LOC100279124 [Zea mays]	ref NP_001145637.2	94	5E-27	ns*	ns*
TRINITY_DN339_c0_g1_i1 m.1218	Only in Unmapped	No	No	uncharacterized protein LOC103631768 [Zea mays]	ref XP_008651447.1	95	5E-03	ns*	ns*
TRINITY_DN353_c0_g1_i1 m.1220	Only in Unmapped	No	No	uncharacterized protein LOC100383368 [Zea mays]	ref NP_001335990.1	100	1E-49	ns*	ns*
TRINITY_DN35_c0_g1_i1 m.1050	Only in Unmapped	No	No	hypothetical protein GA_TR7912_c1_g1_i1_g.25584, partial [Noccaea caerulescens]	gb JAU08974.1	41	8E-05	ns*	ns*
TRINITY_DN38_c0_g1_i2 m.918	Only in Unmapped	No	No	uncharacterized protein LOC100217222 [Zea mays]	ref NP_001137050.1	65	8E-05	ns*	ns*
TRINITY_DN38_c0_g1_i2 m.916	Only in Unmapped	No	No	uncharacterized protein LOC109944984 [Zea mays]	ref XP_020406344.1	73	3E-13	ns*	ns*
TRINITY_DN38_c0_g1_i2 m.915	Only in Unmapped	No	No	LOW QUALITY PROTEIN: uncharacterized protein LOC109939299 [Zea mays]	ref XP_020397089.1	68	4E-04	ns*	ns*
TRINITY_DN41_c0_g1_i1 m.904	Only in Unmapped	No	No	hypothetical protein BAE44_0014613	gb OEL24367.1	39	3E-03	ns*	ns*

Table A.1 (continued). The list of novel transcripts from maize B73-QTL samples with BLASTP hits to plant sequences, sorted by the log2 Fold Change (FC) of their expression upon *C. zeina* infection.

Assembled transcript	In Un-mapped? ^a	In B73? ^b	In QTL?	Description of closest match ^c	Closest match	%ID ^d	E-value ^e	FC	FDR ^f
[Dichanthelium oligosanthos]									
TRINITY_DN426_c0_g1_i1 m.278	Only in Unmapped	No	No	uncharacterized protein LOC100273155 [Zea mays]	ref NP_001141073.1	100	2E-24	ns*	ns*
TRINITY_DN426_c0_g1_i1 m.278	Only in Unmapped	No	No	uncharacterized protein LOC100273155 [Zea mays]	ref NP_001141073.1	45	3E-03	ns*	ns*
TRINITY_DN42_c0_g1_i1 m.951	Only in Unmapped	No	No	uncharacterized protein LOC100193463 [Zea mays]	ref NP_001132051.1	100	2E-120	ns*	ns*
TRINITY_DN441_c0_g1_i1 m.337	Only in Unmapped	No	No	hypothetical protein ZOSMA_461G00010 [Zostera marina]	gb KMZ62450.1	36	3E-07	ns*	ns*
TRINITY_DN449_c0_g1_i1 m.239	Only in Unmapped	No	No	hypothetical protein GA_TR8221_c0_g1_i1_g.26592, partial [Noccaea caerulescens]	gb JAU09906.1	49	1E-09	ns*	ns*
TRINITY_DN455_c0_g1_i1 m.413	Only in Unmapped	No	No	hypothetical protein SETIT_0360912mg [Setaria italica]	gb KQK92612.1	96	4E-48	ns*	ns*
TRINITY_DN461_c0_g1_i1 m.347	Only in Unmapped	No	No	hypothetical protein LR48_Vigan06g101100 [Vigna angularis]	gb KOM45705.1	58	6E-17	ns*	ns*
TRINITY_DN470_c0_g1_i1 m.233	Only in Unmapped	No	No	uncharacterized protein LOC100280971 [Zea mays]	ref NP_001147363.1	100	1E-80	ns*	ns*
TRINITY_DN483_c	Only in	No	No	hypothetical protein MP_TR7759_c	gb JAU99948.1	45	3E-11	ns*	ns*

Table A.1 (continued). The list of novel transcripts from maize B73-QTL samples with BLASTP hits to plant sequences, sorted by the log2 Fold Change (FC) of their expression upon *C. zeina* infection.

Assembled transcript	In Un-mapped? ^a	In B73? ^b	In QTL?	Description of closest match ^c	Closest match	%ID ^d	E-value ^e	FC	FDR ^f
0_g1_i1 m.362	Unmapped			0_g1_i1_g.22033, partial [Noccaea caerulescens]					
TRINITY_DN499_c0_g1_i1 m.260	Only in Unmapped	No	No	uncharacterized protein LOC100382529 [Zea mays]	ref NP_001168737.1	100	1E-28	ns*	ns*
TRINITY_DN523_c0_g1_i1 m.166	Only in Unmapped	No	No	hypothetical protein LC_TR2636_c11_g1_i1_g.8998, partial [Noccaea caerulescens]	gb JAU43961.1	71	2E-36	ns*	ns*
TRINITY_DN540_c0_g1_i1 m.4	Only in Unmapped	No	No	uncharacterized protein LOC100273902 [Zea mays]	ref NP_001141766.1	89	1E-72	ns*	ns*
TRINITY_DN604_c0_g1_i1 m.2082	Only in Unmapped	No	No	hypothetical protein SORBIDRAFT_01g037483 [Sorghum bicolor]	ref XP_002467980.1	98	1E-150	ns*	ns*
TRINITY_DN60_c0_g1_i1 m.894	Only in Unmapped	No	No	uncharacterized protein LOC103648814 [Zea mays]	ref XP_008671435.1	80	9E-08	ns*	ns*
TRINITY_DN634_c0_g1_i1 m.2108	Only in Unmapped	No	No	uncharacterized protein LOC100126943 [Zea mays]	ref NP_001295443.1	100	1E-33	ns*	ns*
TRINITY_DN654_c0_g1_i1 m.2046	Only in Unmapped	No	No	uncharacterized protein LOC100278878 [Zea mays]	ref NP_001145484.1	100	3E-88	ns*	ns*
TRINITY_DN675_c0_g1_i1 m.2113	Only in Unmapped	No	No	hypothetical protein Csa_6G445175 [Cucumis sativus]	gb KGN48152.1	55	3E-17	ns*	ns*
TRINITY_DN67_c	Only in	No	No	hypothetical protein LE_TR14379_c	gb JAU54570.1	49	7E-05	ns*	ns*

Table A.1 (continued). The list of novel transcripts from maize B73-QTL samples with BLASTP hits to plant sequences, sorted by the log2 Fold Change (FC) of their expression upon *C. zeina* infection.

Assembled transcript	In Un-mapped? ^a	In B73? ^b	In QTL?	Description of closest match ^c	Closest match	%ID ^d	E-value ^e	FC	FDR ^f
0_g1_i1 m.883	Unmapped			0_g1_i1_g.45562, partial [Noccaea caerulescens]					
TRINITY_DN680_c 0_g1_i1 m.2129	Only in Unmapped	Yes	No	uncharacterized protein LOC103164613 isoform X2 [Zea mays]	ref XP_020393129.1	87	5E-37	ns*	ns*
TRINITY_DN689_c 0_g1_i1 m.2201	Only in Unmapped	No	No	hypothetical protein LE_TR1209_c 0_g1_i1_g.3131, partial [Noccaea caerulescens]	gb JAU75215.1	54	4E-07	ns*	ns*
TRINITY_DN697_c 0_g1_i1 m.2162	Only in Unmapped	No	No	uncharacterized protein LOC100273155 [Zea mays]	ref NP_001141073.1	100	8E-41	ns*	ns*
TRINITY_DN697_c 0_g1_i1 m.2162	Only in Unmapped	No	No	uncharacterized protein LOC100273155 [Zea mays]	ref NP_001141073.1	39	1E-03	ns*	ns*
TRINITY_DN702_c 0_g1_i1 m.668	Only in Unmapped	No	No	hypothetical protein ZEAMMB73_Zm00001d044025 [Zea mays]	gb ONM39753.1	89	2E-19	ns*	ns*
TRINITY_DN71_c 0_g1_i1 m.1055	Only in Unmapped	No	No	hypothetical protein SORBIDRAFT_03g041490 [Sorghum bicolor]	ref XP_002458852.1	90	7E-55	ns*	ns*
TRINITY_DN71_c 0_g1_i2 m.1056	Only in Unmapped	No	No	hypothetical protein SORBIDRAFT_03g041490 [Sorghum bicolor]	ref XP_002458852.1	90	7E-55	ns*	ns*
TRINITY_DN757_c 0_g1_i1 m.788	Only in Unmapped	No	No	uncharacterized protein LOC100193213 [Zea mays]	ref NP_001131838.1	100	2E-172	ns*	ns*

Table A.1 (continued). The list of novel transcripts from maize B73-QTL samples with BLASTP hits to plant sequences, sorted by the log2 Fold Change (FC) of their expression upon *C. zeina* infection.

Assembled transcript	In Un-mapped? ^a	In B73? ^b	In QTL?	Description of closest match ^c	Closest match	%ID ^d	E-value ^e	FC	FDR ^f
TRINITY_DN763_c0_g1_i1 m.854	Only in Unmapped	No	No	hypothetical protein POPTR_0001s45830g [Populus trichocarpa]	ref XP_002298985.2	45	1E-06	ns*	ns*
TRINITY_DN775_c0_g1_i1 m.686	Only in Unmapped	No	No	uncharacterized protein LOC100272848 [Zea mays]	ref NP_001140773.1	100	1E-34	ns*	ns*
TRINITY_DN776_c0_g1_i1 m.831	Only in Unmapped	No	No	uncharacterized protein LOC100277025 isoform X1 [Zea mays]	ref XP_008647666.1	49	2E-09	ns*	ns*
TRINITY_DN834_c0_g1_i1 m.559	Only in Unmapped	No	No	hypothetical protein ZeamMp074 (mitochondrion) [Zea mays subsp. mays]	ref YP_588336.1	74	6E-15	ns*	ns*
TRINITY_DN848_c0_g1_i1 m.422	Only in Unmapped	No	No	hypothetical protein LE_TR19268_c0_g1_i1_g.61701, partial [Noccaea caerulescens]	gb JAU53817.1	39	1E-03	ns*	ns*
TRINITY_DN849_c0_g1_i1 m.548	Only in Unmapped	No	No	uncharacterized protein LOC100272327 isoform X2 [Zea mays]	ref XP_020406931.1	92	2E-42	ns*	ns*
TRINITY_DN860_c0_g1_i1 m.541	Only in Unmapped	No	No	hypothetical protein SORBIDRAFT_07g004510 [Sorghum bicolor]	ref XP_002445126.1	87	6E-88	ns*	ns*
TRINITY_DN882_c0_g1_i1 m.621	Only in Unmapped	No	No	hypothetical protein [Zea mays]	gb ACG48355.1	97	1E-07	ns*	ns*
TRINITY_DN888_c	Only in	No	No	uncharacterized protein	ref NP_001295443.1	100	0E+00	ns*	ns*

Table A.1 (continued). The list of novel transcripts from maize B73-QTL samples with BLASTP hits to plant sequences, sorted by the log2 Fold Change (FC) of their expression upon *C. zeina* infection.

Assembled transcript	In Un-mapped? ^a	In B73? ^b	In QTL?	Description of closest match ^c	Closest match	%ID ^d	E-value ^e	FC	FDR ^f
0_g1_i1 m.449	Unmapped			LOC100126943 [Zea mays]					
TRINITY_DN949_c0_g1_i1 m.1647	Only in Unmapped	No	No	hypothetical protein ZEAMMB73_Zm00001d025253 [Zea mays]	gb AQK43314.1	69	2E-61	ns*	ns*
TRINITY_DN954_c0_g1_i1 m.1611	Only in Unmapped	No	No	hypothetical protein MP_TR11143_c0_g1_i1_g.33573, partial [Noccaea caerulescens]	gb JAU82286.1	42	1E-07	ns*	ns*
TRINITY_DN990_c0_g1_i1 m.1733	Only in Unmapped	No	No	hypothetical protein LE_TR11712_c0_g1_i1_g.38412, partial [Noccaea caerulescens]	gb JAU73831.1	49	3E-14	ns*	ns*
TRINITY_DN990_c0_g1_i2 m.1735	Only in Unmapped	No	No	hypothetical protein LE_TR11712_c0_g1_i1_g.38412, partial [Noccaea caerulescens]	gb JAU73831.1	48	8E-15	ns*	ns*
TRINITY_DN1165_c0_g1_i1 m.9755	Only in Unmapped	No	No	uncharacterized protein LOC100273155 [Zea mays]	ref NP_001141073.1	100	3E-59	ns*	ns*
TRINITY_DN1384_c0_g1_i1 m.10091	Only in Unmapped	No	No	uncharacterized protein LOC100279124 [Zea mays]	ref NP_001145637.2	100	4E-51	ns*	ns*
TRINITY_DN1629_c0_g1_i1 m.8841	Only in Unmapped	No	No	hypothetical protein SORBIDRAFT_01g037483 [Sorghum bicolor]	ref XP_002467980.1	97	0E+00	ns*	ns*
TRINITY_DN1791_c	Only in	No	No	uncharacterized protein	ref NP_001131838.1	100	2E-172	ns*	ns*

Table A.1 (continued). The list of novel transcripts from maize B73-QTL samples with BLASTP hits to plant sequences, sorted by the log2 Fold Change (FC) of their expression upon *C. zeina* infection.

Assembled transcript	In Un-mapped? ^a	In B73? ^b	In QTL?	Description of closest match ^c	Closest match	%ID ^d	E-value ^e	FC	FDR ^f
0_g1_i1 m.7808	Unmapped			LOC100193213 [Zea mays]					
TRINITY_DN2167_c0_g1_i1 m.5	Only in Unmapped	No	No	uncharacterized protein LOC100383110 [Zea mays]	ref NP_001169249.1	100	4E-27	ns*	ns*
TRINITY_DN2934_c0_g1_i1 m.13016	Only in Unmapped	No	No	hypothetical protein GLYMA_16G028800 [Glycine max]	gb KRH06533.1	64	6E-22	ns*	ns*
TRINITY_DN294_c0_g1_i1 m.11835	Only in Unmapped	No	No	uncharacterized protein LOC100126943 [Zea mays]	ref NP_001295443.1	100	0E+00	ns*	ns*
TRINITY_DN370_c0_g1_i1 m.9590	Only in Unmapped	No	No	hypothetical protein OsI_11203 [Oryza sativa Indica Group]	gb EAY89670.1	65	2E-03	ns*	ns*
TRINITY_DN4069_c0_g1_i1 m.5199	Only in Unmapped	No	No	hypothetical protein SORBI_003G029600 [Sorghum bicolor]	gb KXG31609.1	68	3E-14	ns*	ns*
TRINITY_DN4394_c0_g1_i1 m.6224	Only in Unmapped	No	No	uncharacterized protein LOC100383110 [Zea mays]	ref NP_001169249.1	100	8E-59	ns*	ns*
TRINITY_DN512_c0_g1_i1 m.3191	Only in Unmapped	No	No	uncharacterized protein LOC103638366 [Zea mays]	ref XP_008659520.1	72	2E-10	ns*	ns*
TRINITY_DN5488_c0_g1_i1 m.13565	Only in Unmapped	No	No	hypothetical protein SORBIDRAFT_01g037483 [Sorghum bicolor]	ref XP_002467980.1	98	9E-163	ns*	ns*
TRINITY_DN5601_c0_g1_i1 m.4426	Only in Unmapped	No	No	hypothetical protein SORBIDRAFT_03g041490 [Sorghum bicolor]	ref XP_002458852.1	89	2E-69	ns*	ns*

Table A.1 (continued). The list of novel transcripts from maize B73-QTL samples with BLASTP hits to plant sequences, sorted by the log2 Fold Change (FC) of their expression upon *C. zeina* infection.

Assembled transcript	In Un-mapped? ^a	In B73? ^b	In QTL?	Description of closest match ^c	Closest match	%ID ^d	E-value ^e	FC	FDR ^f
TRINITY_DN6061_c0_g1_i1 m.3273	Only in Unmapped	No	No	hypothetical protein [Zea mays]	gb ACG30731.1	100	7E-09	ns*	ns*
TRINITY_DN759_c0_g1_i1 m.7990	Only in Unmapped	No	No	hypothetical protein SORBIDRAFT_06g001580 [Sorghum bicolor]	ref XP_002446089.1	86	1E-98	ns*	ns*

- a. Lists whether or not the transcript from the master assembly done for the DE analysis had a matching sequence in the unmapped assembly. Sequences from the unmapped assembly that were not present in the master assembly is listed as "Only in unmapped".
- b. Lists whether or not this gene was one of the genes identified to be differentially expressed in B73 by (Swart et al. 2017).
- c. The description retrieved from the accession of the closest match.
- d. The percentage identity the assembled transcript had with the closest match at the DNA level, excluding the transcripts only in the unmapped assembly which were at the amino acid level.
- e. The e-value of the closest match.
- f. The false discover rate of the differential expression analysis result.
- * Not significant.

Table A.2. The top 30 novel transcripts from maize B73-QTL samples with BLASTP hits to *E. turcicum* (*Setosphaeria turcica*) sequences.

Query	Description of closest match	Closest match	%ID*	Alignment Length	Mis-matches	Gap Openings	E-value
TRINITY_DN1470_c0_g1_i1 m.3686	hypothetical protein SETTUDRAFT_168881 [Setosphaeria turcica Et28A]	ref XP_008024616.1	100	748	0	0	0
TRINITY_DN456_c0_g1_i1 m.4779	hypothetical protein SETTUDRAFT_133966 [Setosphaeria turcica Et28A]	ref XP_008020973.1	100	645	0	0	0
TRINITY_DN807_c0_g1_i1 m.6923	hypothetical protein SETTUDRAFT_20167 [Setosphaeria turcica Et28A]	ref XP_008026870.1	100	621	0	0	0
TRINITY_DN4347_c0_g1_i1 m.6402	hypothetical protein SETTUDRAFT_131269 [Setosphaeria turcica Et28A]	ref XP_008026041.1	100	585	0	0	0
TRINITY_DN2514_c0_g1_i1 m.5620	hypothetical protein SETTUDRAFT_174629 [Setosphaeria turcica Et28A]	ref XP_008031544.1	100	583	0	0	0
TRINITY_DN2538_c0_g1_i1 m.5446	hypothetical protein SETTUDRAFT_163200 [Setosphaeria turcica Et28A]	ref XP_008025646.1	100	573	0	0	0
TRINITY_DN4333_c0_g1_i1 m.6324	hypothetical protein SETTUDRAFT_162001 [Setosphaeria turcica Et28A]	ref XP_008020533.1	100	548	0	0	0

Table A.2 (continued). The top 30 novel transcripts from maize B73-QTL samples with BLASTP hits to *E. turcicum* (*Setosphaeria turcica*) sequences.

Query	Description of closest match	Closest match	%ID	Alignment Length	Mis-matches	Gap Openings	E-value
TRINITY_DN1523_c0_g1_i1 m.10640	carbohydrate-binding module family 18 protein [Setosphaeria turcica Et28A]	ref XP_008030609.1	100	540	0	0	0
TRINITY_DN453_c0_g1_i1 m.4919	hypothetical protein SETTUDRAFT_27183 [Setosphaeria turcica Et28A]	ref XP_008023692.1	100	522	0	0	0
TRINITY_DN5206_c0_g1_i1 m.1073	hypothetical protein SETTUDRAFT_32906 [Setosphaeria turcica Et28A]	ref XP_008028079.1	100	514	0	0	0
TRINITY_DN5218_c0_g1_i1 m.1082	hypothetical protein SETTUDRAFT_131269 [Setosphaeria turcica Et28A]	ref XP_008026041.1	100	505	0	0	0
TRINITY_DN751_c0_g1_i1 m.8174	hypothetical protein SETTUDRAFT_96690 [Setosphaeria turcica Et28A]	ref XP_008029699.1	100	500	0	0	0
TRINITY_DN443_c0_g1_i1 m.4925	hypothetical protein SETTUDRAFT_165532 [Setosphaeria turcica Et28A]	ref XP_008030367.1	100	499	0	0	0
TRINITY_DN4396_c0_g1_i1 m.6365	hypothetical protein SETTUDRAFT_43347 [Setosphaeria turcica Et28A]	ref XP_008028880.1	100	487	0	0	0
TRINITY_DN1599_c0_g1_i1 m.10539	hypothetical protein SETTUDRAFT_162792	ref XP_008025209.1	100	476	0	0	0

Table A.2 (continued). The top 30 novel transcripts from maize B73-QTL samples with BLASTP hits to *E. turcicum* (*Setosphaeria turcica*) sequences.

Query	Description of closest match	Closest match	%ID	Alignment Length	Mis-matches	Gap Openings	E-value
	[Setosphaeria turcica Et28A]						
TRINITY_DN2615_c0_g1_i1 m.15575	hypothetical protein SETTUDRAFT_163780 [Setosphaeria turcica Et28A]	ref XP_008027530.1	100	476	0	0	0
TRINITY_DN5233_c0_g1_i1 m.1021	hypothetical protein SETTUDRAFT_100700 [Setosphaeria turcica Et28A]	ref XP_008031586.1	100	461	0	0	0
TRINITY_DN36_c0_g1_i1 m.9194	glycoside hydrolase family 13 protein [Setosphaeria turcica Et28A]	ref XP_008026287.1	100	456	0	0	0
TRINITY_DN1519_c0_g1_i1 m.10550	hypothetical protein SETTUDRAFT_164298 [Setosphaeria turcica Et28A]	ref XP_008028395.1	100	451	0	0	0
TRINITY_DN3307_c0_g1_i1 m.1440	glycoside hydrolase family 3 protein [Setosphaeria turcica Et28A]	ref XP_008028859.1	100	447	0	0	0
TRINITY_DN129_c0_g1_i1 m.10245	glycoside hydrolase family 128 protein [Setosphaeria turcica Et28A]	ref XP_008027440.1	100	432	0	0	0
TRINITY_DN2588_c0_g1_i1 m.5404	hypothetical protein SETTUDRAFT_164902 [Setosphaeria turcica Et28A]	ref XP_008029266.1	100	429	0	0	0
TRINITY_DN930_c0_g1_i1 m.11583	hypothetical protein SETTUDRAFT_162469	ref XP_008020920.1	100	426	0	0	0

Table A.2 (continued). The top 30 novel transcripts from maize B73-QTL samples with BLASTP hits to *E. turcicum* (*Setosphaeria turcica*) sequences.

Query	Description of closest match	Closest match	%ID	Alignment Length	Mis-matches	Gap Openings	E-value
	[Setosphaeria turcica Et28A]						
TRINITY_DN174_c0_g1_i2 m.10356	hypothetical protein SETTUDRAFT_23420 [Setosphaeria turcica Et28A]	ref XP_008030062.1	100	423	0	0	0
TRINITY_DN1577_c0_g1_i1 m.10561	hypothetical protein SETTUDRAFT_86594 [Setosphaeria turcica Et28A]	ref XP_008020473.1	100	416	0	0	0
TRINITY_DN319_c0_g1_i1 m.9593	hypothetical protein SETTUDRAFT_91803 [Setosphaeria turcica Et28A]	ref XP_008027706.1	100	410	0	0	0
TRINITY_DN1729_c0_g1_i1 m.7677	hypothetical protein SETTUDRAFT_162924 [Setosphaeria turcica Et28A]	ref XP_008025356.1	100	406	0	0	0
TRINITY_DN4403_c0_g1_i1 m.1653	hypothetical protein SETTUDRAFT_95885 [Setosphaeria turcica Et28A]	ref XP_008029577.1	100	398	0	0	0
TRINITY_DN643_c0_g1_i1 m.13249	hypothetical protein SETTUDRAFT_168107 [Setosphaeria turcica Et28A]	ref XP_008023525.1	100	392	0	0	0
TRINITY_DN666_c0_g1_i1 m.13230	hypothetical protein SETTUDRAFT_118721 [Setosphaeria turcica Et28A]	ref XP_008028327.1	100	390	0	0	0

*Percentage identity at the amino acid level.

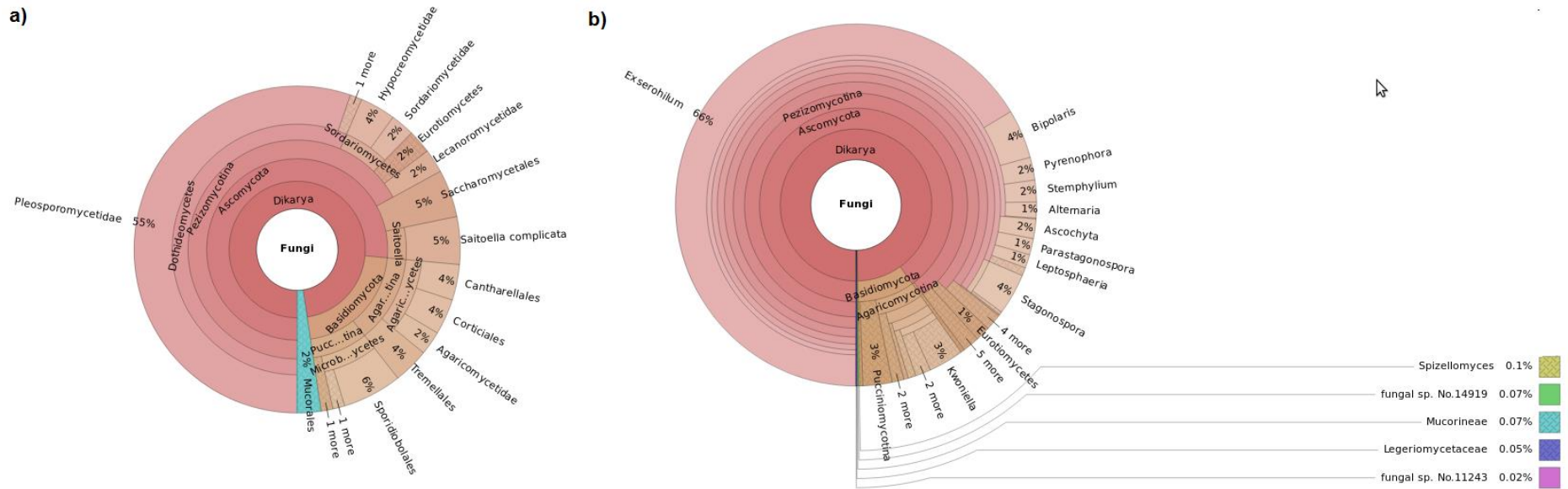


Figure A.1. The predicted proteins identified by BLASTP against the NCBI nr database that matched fungal sequences. The outer ring shows the percentage of each fungal genus present and the lower rings shows the lineages of the genera a) The relative fungal genii in the control leaves. b) The relative fungal genii in the *C. zeina* challenged leaves. Figure drawn using Krona (Ondov et al. 2011).

Table A.3. The predicted proteins from the unmapped assembly identified by BLASTP against the NCBI nr database that matched putative plant viruses.

Query	Description	Subject	%ID*	Alignment Length	Mis-matches	Gap Openings	E-value	Leaf
TRINITY_DN855_c0_g1_i1 m.578	P0 protein [Maize yellow mosaic virus]	gb ANI26462.1	97.22	108	3	0	3E-67	Control
TRINITY_DN556_c0_g1_i1 m.7	P2 [Maize yellow mosaic virus]	gb API68969.1	99.1	111	1	0	1E-73	Control
TRINITY_DN549_c0_g1_i1 m.147	P3-P5 readthrough protein [Maize yellow mosaic virus]	gb ANI26451.1	97.41	116	3	0	2E-77	Control
TRINITY_DN386_c0_g1_i1 m.1196	P1 [Maize yellow mosaic virus]	gb ANI26444.1	98.22	169	3	0	1E-112	Control
TRINITY_DN603_c0_g1_i1 m.2119	P3-P5 readthrough protein [Maize yellow mosaic virus]	gb ANI26454.1	96.67	300	10	0	0E+00	Control
TRINITY_DN635_c0_g1_i1 m.2163	P1 [Maize yellow mosaic virus]	gb ANI26438.1	98.62	290	4	0	0E+00	Control
TRINITY_DN330_c0_g1_i1 m.1092	capsid protein [Rose partitivirus]	gb ANQ45204.1	58.57	140	58	0	9E-53	Control
TRINITY_DN905_c0_g1_i1 m.11702	59 kDa protein [Pleospora typhicola fusarivirus 1]	ref YP_009182159.1	52.56	78	37	0	1E-16	<i>C. zeina</i> challenged
TRINITY_DN1998_c0_g1_i1 m.11383	59 kDa protein [Pleospora typhicola fusarivirus 1]	ref YP_009182159.1	71.26	87	25	0	3E-36	<i>C. zeina</i> challenged
TRINITY_DN859_c0_g1_i1 m.7213	capsid protein [Rose partitivirus]	gb ANQ45204.1	52.83	441	200	4	1E-160	<i>C. zeina</i> challenged

*Percentage identity at the amino acid level.

Table A.4. The predicted proteins from the unmapped assembly identified by BLASTP against the NCBI nr database that matched putative plant bacteria.

Query	Description	Subject	%ID*	Alignment Length	Mis-matches	Gap Openings	E-value	Leaf
TRINITY_DN741_c0_g1_i1 m.829	hypothetical protein BGC29_19500 [Acinetobacter baumannii]	gb ODP69016.1	70.59	68	20	0	2E-20	Control
TRINITY_DN82_c0_g1_i1 m.921	hypothetical protein [uncultured bacterium]	gb AOE07968.1	77.78	81	18	0	8E-32	Control
TRINITY_DN217_c0_g1_i1 m.1788	Cell wall-associated hydrolase family protein [Paenibacillus sp. P22]	emb CDN41077.1	100	105	0	0	3E-67	Control
TRINITY_DN213_c0_g2_i3 m.1917	hypothetical protein BN871_AB_00880 [Paenibacillus sp. P22]	emb CDN41090.1	95.88	194	8	0	2E-121	Control
TRINITY_DN202_c0_g1_i1 m.1850	conserved hypothetical protein [Thiomonas sp. CB2]	emb CDW93818.1	89.36	47	5	0	1E-20	Control
TRINITY_DN524_c0_g1_i1 m.19	hypothetical protein, partial [uncultured alpha proteobacterium HF0070_14E07]	gb ADI17229.1	74.67	75	19	0	9E-29	Control
TRINITY_DN403_c0_g1_i1 m.4892	hypothetical protein A7M48_19040, partial [Acinetobacter baumannii]	gb OIC32608.1	98.48	66	1	0	2E-36	<i>C. zeina</i> challenged
TRINITY_DN488_c0_g1_i1 m.4879	hypothetical protein [Acinetobacter baumannii]	ref WP_079863869.1	100	136	0	0	3E-91	<i>C. zeina</i> challenged
TRINITY_DN782_c0_g1_i1 m.8039	hypothetical protein, partial [Acinetobacter baumannii]	ref WP_071414375.1	84.81	79	12	0	1E-36	<i>C. zeina</i> challenged
TRINITY_DN511_c0_g1_i1 m.3170	hypothetical protein [Spirosoma rigui]	ref WP_080240697.1	71.05	76	18	1	8E-22	<i>C. zeina</i> challenged
TRINITY_DN1519_c0_g1_i1 m.10551	putative NAD-specific glutamate dehydrogenase [Mesorhizobium alhagi]	gb EHK58756.1	42.16	389	223	2	2E-69	<i>C. zeina</i> challenged

*Percentage identity at the amino acid level.

Table A.5. The predicted proteins from the control leaves identified in the blast that were neither plant nor fungal nor phylogenetic and sorted by species.

Query	Description	Subject	%ID*	Alignment Length	Mis-matches	E-value
TRINITY_DN3_c0_g1_i1 m.877	putative ribosomal protein, partial [Haematobia irritans]	gb JAV14374.1	43.28	67	38	8E-11
TRINITY_DN651_c0_g1_i1 m.2243	hypothetical protein, partial [Haematobia irritans]	gb JAV14341.1	37.93	87	53	1E-12
TRINITY_DN383_c0_g1_i1 m.1168	hypothetical protein, partial [Haematobia irritans]	gb JAV17896.1	51.25	80	39	3E-14
TRINITY_DN442_c0_g1_i1 m.369	hypothetical protein, partial [Haematobia irritans]	gb JAV14295.1	51.14	88	43	2E-22
TRINITY_DN808_c0_g1_i1 m.619	PREDICTED: regenerating islet-derived protein 4-like [Xenopus tropicalis]	ref XP_002934628.1	32.74	113	71	5E-16
TRINITY_DN881_c0_g1_i1 m.432	Chain A, Solution Structure Of Domain 3 From Human Serum Albumin Complexed To An Anti-Apoptotic Ligand Directed Against Bcl- XI And Bcl-2	pdb 1YSX A	100	73	0	2E-40
TRINITY_DN503_c0_g1_i1 m.30	albumin, isoform CRA_c [Homo sapiens]	gb EAX05661.1	100	71	0	4E-43
TRINITY_DN873_c0_g1_i1 m.599	albumin, isoform CRA_p [Homo sapiens]	gb EAX05674.1	100	83	0	1E-52
TRINITY_DN953_c0_g1_i1 m.1726	HP protein [Homo sapiens]	gb AAH58031.1	100	87	0	3E-54
TRINITY_DN676_c0_g1_i1 m.2152	ALB protein [Homo sapiens]	gb AAH41789.1	100	96	0	2E-58
TRINITY_DN947_c0_g1_i1 m.1652	ALB, partial [Homo sapiens]	gb AAQ89947.1	100	117	0	1E-79

Table A.5 (Continued). The predicted proteins from the control leaves identified in the blast that were neither plant nor fungal nor phylogenetic and sorted by species.

Query	Description	Subject	%ID	Alignment Length	Mis-matches	E-value
TRINITY_DN419_c0_g1_i1 m.273	PRO2675 [Homo sapiens]	gb AAF69644.1	100	130	0	5E-88
TRINITY_DN230_c0_g1_i1 m.2034	HP protein [Homo sapiens]	gb AAH58031.1	100	127	0	1E-89
TRINITY_DN258_c0_g1_i1 m.1819	albumin-like, partial [Homo sapiens]	gb AAA64922.1	100	144	0	2E-93
TRINITY_DN790_c0_g1_i1 m.859	PRO2044 [Homo sapiens]	gb AAF22034.1	100	255	0	0
TRINITY_DN308_c0_g1_i1 m.1063	TPA: Os05g0242100-like [Bos taurus]	tpg DAA18777.1	70.13	77	23	1E-24
TRINITY_DN308_c0_g1_i1 m.1063	TPA: Os05g0242100-like [Bos taurus]	tpg DAA18777.1	69.74	76	23	1E-24
TRINITY_DN308_c0_g1_i1 m.1063	TPA: Os05g0242100-like [Bos taurus]	tpg DAA18777.1	71.43	77	22	3E-26
TRINITY_DN308_c0_g1_i1 m.1063	TPA: Os05g0242100-like [Bos taurus]	tpg DAA18777.1	72.73	77	21	1E-26
TRINITY_DN574_c0_g1_i1 m.12	hypothetical protein, partial [Lygus hesperus]	gb JAG58402.1	98.21	56	1	5E-29
TRINITY_DN220_c0_g1_i3 m.1802	Acetolactate synthase small subunit, partial [Lygus hesperus]	gb JAG36151.1	67.13	143	47	4E-54
TRINITY_DN162_c0_g1_i1 m.1418	protein synthesis elongation factor 1-alpha, partial [Planoprotostelium aurantium]	gb AAB69703.1	97.97	148	3	2E-99
TRINITY_DN981_c0_g1_i1 m.1730	hypothetical protein H257_02818 [Aphanomyces astaci]	ref XP_009824939.1	45.16	93	51	6E-15
TRINITY_DN9_c0_g1_i1 m.889	PREDICTED: 40S ribosomal protein S18 [Diaphorina citri]	ref XP_008475493.1	70.11	87	25	3E-37
TRINITY_DN713_c0_g1_i1 m.751	hypothetical protein THAOC_17987, partial [Thalassiosira oceanica]	gb EJK61508.1	44.83	87	45	3E-11
TRINITY_DN821_c0_g1_i1 m.481	hypothetical protein CAPTEDRAFT_189241 [Capitella teleta]	gb ELU06605.1	67.57	74	24	2E-21

Table A.5 (Continued). The predicted proteins from the control leaves identified in the blast that were neither plant nor fungal nor phylogenetic and sorted by species.

Query	Description	Subject	%ID	Alignment Length	Mis-matches	E-value
TRINITY_DN344_c0_g1_i1 m.1298	PREDICTED: 40S ribosomal protein S23 [Amphimedon queenslandica]	ref XP_003382653.1	86.25	80	11	3E-42
TRINITY_DN709_c0_g1_i1 m.722	NAD-specific glutamate dehydrogenase, partial [Cerapachys biroii]	gb EZA53142.1	44.53	128	71	6E-21
TRINITY_DN217_c0_g1_i1 m.1792	hypothetical protein HMPREF0527_01622 [Lactobacillus jensenii SJ-7A-US]	gb EEX26613.1	83.67	49	8	1E-21
TRINITY_DN999_c0_g1_i1 m.1695	ribosomal protein S19 [Capsaspora owczarzaki ATCC 30864]	ref XP_004348972.1	66.13	62	21	3E-19
TRINITY_DN630_c0_g1_i1 m.2148	large subunit ribosomal protein L35Ae, cytoplasmic [Guillardia theta CCMP2712]	ref XP_005818688.1	62.07	87	31	5E-30
TRINITY_DN188_c0_g1_i1 m.1360	actin, partial [Thecamoeba sp. JRG-2011]	gb AEN94513.1	98.29	117	2	2E-76
TRINITY_DN213_c0_g2_i1 m.1915	Uncharacterised protein [Mycobacterium abscessus subsp. abscessus]	emb SHT70305.1	74.07	81	21	1E-31
TRINITY_DN202_c0_g1_i1 m.1848	hypothetical protein GBS1173_1869 [Streptococcus agalactiae CF01173]	emb CCQ81474.1	82.69	52	9	2E-20
TRINITY_DN56_c0_g1_i1 m.980	putative 60s acidic ribosomal protein p2 [Corethrella appendiculata]	gb JAB59602.1	72.88	59	16	1E-19
TRINITY_DN215_c0_g1_i1 m.1932	actin, partial [Vermistella sp. TT-2015a]	gb AJT35040.1	97.48	119	3	2E-79
TRINITY_DN575_c0_g1_i1 m.163	actin, partial [Vermistella sp. TT-2015b]	gb AJT35058.1	97.18	71	2	9E-43

*Percentage identity at the amino acid level.

Table A.6. The predicted proteins from the *C. zeina* challenged leaves identified in the blast that were neither plant nor fungal nor phylogenetic and sorted by species.

Query	Description	Subject	%ID*	Alignment Length	Mismatches	E-value
TRINITY_DN823_c0_g2_i1 m.6757	RecName: Full=Ubiquitin-60S ribosomal protein L40; Contains: RecName: Full=Ubiquitin; Contains: RecName: Full=60S ribosomal protein L40; Flags: Precursor	sp P0CH27.1 RL402_TRYC R	91.57	356	30	0
TRINITY_DN3438_c0_g1_i1 m.4708	glyceraldehyde 3-phosphate dehydrogenase [Schistosoma japonicum]	emb CAX72630.1	78.57	70	15	6E-32
TRINITY_DN867_c0_g1_i2 m.6819	actin-2 [Ascaris suum]	gb ERG81274.1	100	130	0	1E-90
TRINITY_DN342_c0_g1_i1 m.9637	hypothetical protein HELRODRAFT_88883, partial [Helobdella robusta]	ref XP_009028539.1	74.63	67	17	7E-23
TRINITY_DN388_c0_g1_i1 m.9582	hypothetical protein HELRODRAFT_156722 [Helobdella robusta]	ref XP_009014518.1	89.02	82	9	2E-45
TRINITY_DN485_c0_g1_i1 m.4939	hypothetical protein HELRODRAFT_115575 [Helobdella robusta]	ref XP_009028383.1	97.78	90	2	1E-54
TRINITY_DN284_c0_g1_i1 m.11755	PREDICTED: 60S ribosomal protein L23a-like [Limulus polyphemus]	ref XP_013792554.1	78.95	95	20	1E-45

Table A.6 (continued). The predicted proteins from the *C. zeina* challenged leaves identified in the blast that were neither plant nor fungal nor phylogenetic and sorted by species.

Query	Description	Subject	%ID	Alignment Length	Mismatches	E-value
TRINITY_DN1470_c0_g1_i1 m.3687	hypothetical protein, partial [<i>Culex tarsalis</i>]	gb JAV18991.1	49.44	718	357	1E-158
TRINITY_DN274_c0_g1_i1 m.11760	hypothetical protein, partial [<i>Haematobia irritans</i>]	gb JAV14342.1	35.71	224	143	2E-33
TRINITY_DN4204_c0_g1_i1 m.15061	PREDICTED: 40S ribosomal protein S20 [<i>Lingula anatina</i>]	ref XP_013404997.1	90.2	51	5	7E-23
TRINITY_DN1247_c0_g1_i1 m.248	PREDICTED: 40S ribosomal protein S26 [<i>Lepisosteus oculatus</i>]	ref XP_015199979.1	74.03	77	20	1E-34
TRINITY_DN2625_c0_g1_i1 m.15512	PREDICTED: tubulin beta-4B chain, partial [<i>Lepisosteus oculatus</i>]	ref XP_015222005.1	95.45	66	3	2E-39
TRINITY_DN763_c0_g1_i2 m.8239	beta-actin, partial [<i>Boa constrictor</i>]	gb ACN32770.1	99.28	138	1	2E-94
TRINITY_DN763_c0_g1_i1 m.8238	actin beta, partial [<i>Rousettus leschenaultii</i>]	gb AFJ05205.1	99	100	1	3E-64
TRINITY_DN5828_c0_g1_i1 m.2454	PREDICTED: keratin, type I cytoskeletal 18 [<i>Macaca nemestrina</i>]	ref XP_011727312.1	100	71	0	2E-42
TRINITY_DN681_c0_g1_i1 m.13453	ACTB protein, partial [<i>Homo sapiens</i>]	gb AAI13037.1	100	79	0	2E-49
TRINITY_DN4443_c0_g1_i1 m.1628	aldehyde dehydrogenase I, partial [<i>Homo sapiens</i>]	gb AAA35518.1	100	113	0	3E-74
TRINITY_DN3125_c0_g1_i1 m.14322	unnamed protein product [<i>Homo</i>]	dbj BAG63535.1	100	126	0	1E-85

Table A.6 (continued). The predicted proteins from the *C. zeina* challenged leaves identified in the blast that were neither plant nor fungal nor phylogenetic and sorted by species.

Query	Description	Subject	%ID	Alignment Length	Mismatches	E-value
sapiens]						
TRINITY_DN4200_c0_g1_i1 m.15107	PREDICTED: elongation factor 1-alpha 1, partial [Odobenus rosmarus divergens]	ref XP_004413149.1	100	95	0	3E-61
TRINITY_DN266_c0_g1_i1 m.11834	60S ribosomal protein L3, partial [Brachionus plicatilis]	emb CAQ37747.1	61.45	83	31	1E-28
TRINITY_DN1365_c0_g1_i1 m.10022	PREDICTED: 60S ribosomal protein L5-like [Saccoglossus kowalevskii]	ref XP_002736904.1	66.67	60	20	2E-21
TRINITY_DN77_c0_g1_i1 m.9177	PREDICTED: 40S ribosomal protein S5-like [Monodelphis domestica]	ref XP_001371286.2	97.96	49	1	5E-24
TRINITY_DN5565_c0_g1_i1 m.1961	60S ribosomal protein L30 [Sycon ciliatum]	emb CCQ18587.1	82.14	56	10	2E-27
TRINITY_DN3059_c0_g1_i1 m.12393	PREDICTED: actin-1, partial [Drosophila kikkawai]	ref XP_017021020.1	100	67	0	3E-42
TRINITY_DN139_c0_g1_i1 m.10297	hypothetical protein, partial [Lygus hesperus]	gb JAG61789.1	96.88	96	2	2E-53
TRINITY_DN2040_c0_g1_i1 m.12786	hypothetical protein, partial [Lygus hesperus]	gb JAG50099.1	91.3	92	8	8E-54
TRINITY_DN854_c0_g1_i1 m.6957	hypothetical protein, partial [Lygus hesperus]	gb JAG58402.1	98.13	107	2	7E-68
TRINITY_DN854_c0_g1_i2 m.6960	hypothetical protein, partial [Lygus hesperus]	gb JAG58402.1	99.07	107	1	3E-68

Table A.6 (continued). The predicted proteins from the *C. zeina* challenged leaves identified in the blast that were neither plant nor fungal nor phylogenetic and sorted by species.

Query	Description	Subject	%ID	Alignment Length	Mismatches	E-value
TRINITY_DN854_c0_g1_i3 m.6963	hypothetical protein, partial [Lygus hesperus]	gb JAG58402.1	99.07	107	1	3E-68
TRINITY_DN514_c0_g1_i1 m.3248	RecName: Full=40S ribosomal protein S18	sp Q8IT98.1 RS18_ARGIR	80.28	71	14	1E-35
TRINITY_DN2643_c0_g1_i1 m.15628	unnamed protein product [Oikopleura dioica]	emb CBY12791.1	98.65	74	1	2E-44
TRINITY_DN276_c0_g1_i1 m.11804	PREDICTED: 60S acidic ribosomal protein P2 [Thamnophis sirtalis]	ref XP_013913161.1	67.14	70	23	7E-25
TRINITY_DN378_c0_g1_i1 m.9613	hypothetical protein PBRA_008258 [Plasmodiophora brassicae]	emb CEP00946.1	53.98	113	52	4E-34
TRINITY_DN516_c0_g1_i1 m.3159	putative 60S acidic ribosomal protein P0-like protein [Pinctada fucata]	gb JAS03248.1	55.14	107	48	1E-34
TRINITY_DN3784_c0_g1_i1 m.10721	ribosomal protein L37e [Necator americanus]	ref XP_013290942.1	69.12	68	19	4E-24
TRINITY_DN721_c0_g1_i1 m.8110	actin 5C/42A, partial [Glomeris marginata]	emb CCA89752.1	100	70	0	9E-43
TRINITY_DN5046_c0_g1_i1 m.3056	elongation factor 1-alpha 1, partial [Panthera tigris altaica]	gb AGW80518.1	100	118	0	5E-80
TRINITY_DN842_c0_g1_i1 m.6913	ubiquitin/actin fusion protein 1, partial [Lotharella globosa]	gb AAP34631.1	97.18	142	4	2E-98
TRINITY_DN2536_c0_g1_i1 m.5496	glycoside hydrolase family 11	emb CDS82524.1	63.08	65	20	2E-16

Table A.6 (continued). The predicted proteins from the *C. zeina* challenged leaves identified in the blast that were neither plant nor fungal nor phylogenetic and sorted by species.

Query	Description	Subject	%ID	Alignment Length	Mismatches	E-value
enzyme [uncultured eukaryote]						
TRINITY_DN4294_c0_g1_i1 m.15196	SCF ubiquitin ligase [uncultured eukaryote]	emb CCG34069.1	100	71	0	4E-43
TRINITY_DN1231_c0_g1_i1 m.286	PREDICTED: elongation factor 1-alpha 1, partial [Chrysochloris asiatica]	ref XP_006835853.1	100	132	0	8E-90
TRINITY_DN5613_c0_g1_i1 m.4274	heat shock protein 70-2, partial [Laodelphax striatella]	gb AMD09928.1	98.61	72	1	9E-40
TRINITY_DN741_c0_g1_i1 m.8189	ribosomal protein S25 [Lepidochitona cinerea]	gb ACR24971.1	65.12	86	30	9E-31
TRINITY_DN481_c0_g1_i1 m.4937	hypothetical protein DFA_11760 [Dictyostelium fasciculatum]	ref XP_004350707.1	56.34	71	31	2E-18
TRINITY_DN920_c0_g1_i1 m.11481	elongation factor 1-alpha, partial (macronuclear) [Philasterides dicentrarchi]	gb AHJ61123.1	75.25	101	23	2E-46
TRINITY_DN4878_c0_g1_i1 m.8374	putative 60s ribosomal protein l8 [Nyssomyia neivai]	gb JAV08950.1	75.51	98	24	5E-50
TRINITY_DN1820_c0_g1_i1 m.6541	elongation factor 1-alpha, partial [Rhynchophorus ferrugineus]	gb AMK48604.1	92.04	113	9	1E-69
TRINITY_DN3124_c0_g1_i1 m.14339	PREDICTED: 60S acidic ribosomal protein P0-like [Amphimedon queenslandica]	ref XP_003384811.1	63.27	49	18	3E-12
TRINITY_DN593_c0_g1_i1 m.3158	60S ribosomal protein L3, putative [Perkinsus marinus ATCC 50983]	ref XP_002764639.1	76	75	18	3E-34

Table A.6 (continued). The predicted proteins from the *C. zeina* challenged leaves identified in the blast that were neither plant nor fungal nor phylogenetic and sorted by species.

Query	Description	Subject	%ID	Alignment Length	Mismatches	E-value
TRINITY_DN5829_c0_g1_i1 m.2384	elongation factor 1A, partial [Cavostelium apophysatum]	gb ABQ41398.1	94.52	73	4	3E-41
TRINITY_DN5697_c0_g1_i1 m.4364	elongation factor 1A, partial [Cavostelium apophysatum]	gb ABQ41398.1	87.63	97	12	6E-54
TRINITY_DN5132_c0_g1_i1 m.13991	elongation factor 1A, partial [Schizoplasmodiopsis amoeboides]	gb ABQ41409.1	95.28	106	5	5E-64
TRINITY_DN5021_c0_g1_i1 m.2987	elongation factor 1A, partial [Soliformovum irregulare]	gb ABQ41411.1	96	100	4	7E-60
TRINITY_DN4389_c0_g1_i1 m.6214	40S ribosomal protein S18 [Microcosmus squamiger]	emb CBN08641.1	73.91	69	18	2E-30
TRINITY_DN3507_c0_g1_i1 m.628	beta-tubulin [Symbiodinium sp. B1]	gb AIA81327.1	93.94	66	4	2E-40
TRINITY_DN857_c0_g1_i3 m.6988	LOW QUALITY PROTEIN: conserved hypothetical protein, partial [Brucella suis bv. 3 str. 686]	gb EEY31938.1	93.75	64	4	2E-34
TRINITY_DN5407_c0_g1_i1 m.13642	elongation factor 1 alpha, partial [Blastocystis sp. OSU25]	gb ACG60897.1	87.67	73	9	2E-37
TRINITY_DN3611_c0_g1_i1 m.14649	guanine nucleotide-binding protein subunit beta-2-like 1 [Saprolegnia parasitica CBS 223.65]	ref XP_012205266.1	86.49	74	10	3E-38
TRINITY_DN823_c0_g2_i3 m.6774	uncharacterized protein [Babesia microti strain RI]	ref XP_012648878.1	93.51	77	5	1E-39
TRINITY_DN823_c0_g2_i3 m.6774	uncharacterized protein [Babesia	ref XP_012648878.1	95.04	121	6	1E-73

Table A.6 (continued). The predicted proteins from the *C. zeina* challenged leaves identified in the blast that were neither plant nor fungal nor phylogenetic and sorted by species.

Query	Description	Subject	%ID	Alignment Length	Mismatches	E-value
	microti strain RI]					
TRINITY_DN823_c0_g2_i2 m.6767	uncharacterized protein [Babesia microti strain RI]	ref XP_012648878.1	94.12	153	9	1E-94
TRINITY_DN823_c0_g2_i3 m.6774	uncharacterized protein [Babesia microti strain RI]	ref XP_012648878.1	90.12	172	17	1E-109
TRINITY_DN823_c0_g2_i2 m.6767	uncharacterized protein [Babesia microti strain RI]	ref XP_012648878.1	89.32	206	22	8E-132
TRINITY_DN2551_c0_g1_i1 m.5506	Cathepsin L [Oxytricha trifallax]	gb EJY83513.1	49.57	115	55	4E-24
TRINITY_DN2887_c0_g1_i1 m.2304	CsbD family protein [Acanthamoeba castellanii str. Neff]	ref XP_004340963.1	43.43	99	56	1E-16
TRINITY_DN461_c0_g1_i1 m.4884	ribosomal protein L21e, putative [Acanthamoeba castellanii str. Neff]	ref XP_004333362.1	57.83	83	28	1E-19
TRINITY_DN5106_c0_g1_i1 m.13814	Actophorin, putative [Acanthamoeba castellanii str. Neff]	ref XP_004368291.1	61.45	83	31	3E-29
TRINITY_DN2758_c0_g1_i1 m.5830	ribosomal L40e family [Acanthamoeba castellanii str. Neff]	ref XP_004335122.1	95	100	5	2E-63
TRINITY_DN5441_c0_g1_i1 m.13673	putative phosphoglycerate kinase [Bifidobacterium bifidum CAG:234]	emb CDB23221.1	35.78	313	181	2E-21
TRINITY_DN678_c0_g1_i1 m.13341	recombinational DNA repair protein [Aeromonas diversa CDC	gb ENY73677.1	36.51	304	192	9E-31

Table A.6 (continued). The predicted proteins from the *C. zeina* challenged leaves identified in the blast that were neither plant nor fungal nor phytogenic and sorted by species.

Query	Description	Subject	%ID	Alignment Length	Mismatches	E-value
2478-85]						
TRINITY_DN351_c0_g1_i1 m.9572	hypothetical protein SAM00019534_070720 [Acytostelium subglobosum LB1]	ref XP_012752933.1	47.89	71	36	5E-11
TRINITY_DN5429_c0_g1_i1 m.13527	hypothetical protein SAM00019534_038040 [Acytostelium subglobosum LB1]	ref XP_012760150.1	70.37	81	24	6E-34
TRINITY_DN598_c0_g1_i1 m.3089	hypothetical protein SAM00019534_123360, partial [Acytostelium subglobosum LB1]	ref XP_012747850.1	100	66	0	4E-40
TRINITY_DN322_c0_g1_i1 m.9642	translation elongation factor 1- alpha, partial [Colponema edaphicum]	gb AIM46988.1	71.43	70	20	1E-29
TRINITY_DN5705_c0_g1_i1 m.12021	actin, partial [Vermistella sp. TT- 2015a]	gb AJT35040.1	98.57	70	1	3E-41
TRINITY_DN2708_c0_g1_i1 m.5729	40S ribosomal protein S15 [Exaiptasia pallida]	gb KXJ16102.1	79.43	141	29	1E-75
TRINITY_DN1653_c0_g1_i1 m.8987	PREDICTED: glyceraldehyde-3- phosphate dehydrogenase-like [Cebus capucinus imitator]	ref XP_017369010.1	100	77	0	2E-48

*Percentage identity at the amino acid level.

	10	20	30	40	50	60	70	80	90	100
B73_G-type_Lectin										
BU TRINITY_DN1039_c0_g1_i1	CGGCAGGTCCACTCATTACTGCAGAGTTCCATTCCATACCTCCATTTTCCATCTGCACCGCGAGTTCCATTCCATGCTTCATATCATAGGCCAGAGAACA									
BU TRINITY_DN3882_c0_g1_i1	-----									
BU TRINITY_DN4163_c0_g1_i1	GTA.....C.....T.....									
AU TRINITY_DN868_c0_g1_i1	-----									
MA TRINITY_DN12415_c0_g1_i1	-----									
MA TRINITY_DN4559_c0_g1_i1	-----A.....									

	110	120	130	140	150	160	170	180	190	200
B73_G-type_Lectin										
BU TRINITY_DN1039_c0_g1_i1	AGAGAAGCCCTGAAAGATCAATAGACAGATATGATGTATGTTCCCATCTTCATCCTCCTCGTAGTCTTGAGTTCGCTCTGCCAATCCGACGATCAGCTGA									
BU TRINITY_DN3882_c0_g1_i1	-----									
BU TRINITY_DN4163_c0_g1_i1	C.....									
AU TRINITY_DN868_c0_g1_i1	-----									
MA TRINITY_DN12415_c0_g1_i1	-----									
MA TRINITY_DN4559_c0_g1_i1	C.....									

	210	220	230	240	250	260	270	280	290	300
B73_G-type_Lectin										
BU TRINITY_DN1039_c0_g1_i1	CACATACAAAGCCGCTCTTCCCCAAAGAGACGCTCATCTCGGCCGCGGGGACTTCGCTCTTGGCTTCTTTTCTCTACCAGTTCAAGCAACAAGCTGTA									
BU TRINITY_DN3882_c0_g1_i1	-----									
BU TRINITY_DN4163_c0_g1_i1	CT.....C.....									
AU TRINITY_DN868_c0_g1_i1	-----									
MA TRINITY_DN12415_c0_g1_i1	-----									
MA TRINITY_DN4559_c0_g1_i1	CT.....C.....									

	310	320	330	340	350	360	370	380	390	400
B73_G-type_Lectin										
BU TRINITY_DN1039_c0_g1_i1	CATCGGGATTTGGTACAACAATGTCCCAGAGCGGGAGCGCACGGTGGTGTGGATCGCCAACCGTGACAGCCCAGTCACCACCCAGCTACATCAGTGAAG									
BU TRINITY_DN3882_c0_g1_i1	-----									
BU TRINITY_DN4163_c0_g1_i1	-----									
AU TRINITY_DN868_c0_g1_i1	-----									
MA TRINITY_DN12415_c0_g1_i1	-----									
MA TRINITY_DN4559_c0_g1_i1	-----									

	410	420	430	440	450	460	470	480	490	500
B73_G-type_Lectin										
BU TRINITY_DN1039_c0_g1_i1	CTCGCCGTCACAAACGATTCAGAACTGGTGCTGTCAGACTCCCAAGGCCACATCTTTTGGATGGCAACGAGTAGCACAAAGTGGAGGCGCTGGAGCTTTTG									
BU TRINITY_DN3882_c0_g1_i1	-----									
BU TRINITY_DN4163_c0_g1_i1	-----									
AU TRINITY_DN868_c0_g1_i1	-----									
MA TRINITY_DN12415_c0_g1_i1	-----									
MA TRINITY_DN4559_c0_g1_i1	A.....A..C.....									

	510	520	530	540	550	560	570	580	590	600
B73_G-type_Lectin	 									
BU TRINITY_DN1039_c0_g1_i1	CAGTGCTGCTCAACTCAGGGAACCTTTGCTCTCGATCGCCAAACGGCACGGACGTATGGCAAAGCTTCGATCACCCGACTGATACAGTGCTTCCAACCAT									
BU TRINITY_DN3882_c0_g1_i1	-----T-----C-----									
BU TRINITY_DN4163_c0_g1_i1	-----T-----C-----									
AU TRINITY_DN868_c0_g1_i1	-----T-----C-----									
MA TRINITY_DN12415_c0_g1_i1	-----T-----C-----									
MA TRINITY_DN4559_c0_g1_i1	T.....T.....C.....									
	610	620	630	640	650	660	670	680	690	700
B73_G-type_Lectin	 									
BU TRINITY_DN1039_c0_g1_i1	GAGGAGTATGATAAGATACAGGTACAAACCGGCTACGGGCTCTTTGCTGGAAGAGCCCTGACGATCCATCAACCGGGGACATCTCTTCTGGCATGGAT									
BU TRINITY_DN3882_c0_g1_i1	-----TT.GG.....C.....AC.....G.....G.....A.....T.A..T....									
BU TRINITY_DN4163_c0_g1_i1	-----TT.GG.....C.....AC.....G.....G.....A.....T.A..T....									
AU TRINITY_DN868_c0_g1_i1	-----TT.GG.....C.....AC.....G.....G.....A.....T.A..T....									
MA TRINITY_DN12415_c0_g1_i1	-----TT.GG.....C.....AC.....G.....G.....A.....T.A..T....									
MA TRINITY_DN4559_c0_g1_i1	-----TT.GG.....C.....AC.....G.....G.....A.....T.A..T....									
	710	720	730	740	750	760	770	780	790	800
B73_G-type_Lectin	 									
BU TRINITY_DN1039_c0_g1_i1	CCGGCCTCCTCAAACGTTCAAGGTGTTTCATTTGGAATGGGACATTGCCGTACTATCGTTCTTCTTTTAAATGATGTCCCAGCATCCGGCGGCACGTACC									
BU TRINITY_DN3882_c0_g1_i1	-----C.....A.....C.....T.TC....C...G..G.....T.C.T.....									
BU TRINITY_DN4163_c0_g1_i1	-----C.....A.....C.....T.TC....C...G..G.....T.C.T.....									
AU TRINITY_DN868_c0_g1_i1	-----C.....A.....C.....T.TC....C...G..G.....T.C.T.....									
MA TRINITY_DN12415_c0_g1_i1	-----C.....A.....C.....T.TC....C...G..G.....T.C.T.....									
MA TRINITY_DN4559_c0_g1_i1	-----C.....A.....C.....T.TC....C...G..G.....T.C.T.....									
	810	820	830	840	850	860	870	880	890	900
B73_G-type_Lectin	 									
BU TRINITY_DN1039_c0_g1_i1	AGGCCAACAATACCTCAGTCACATACCAATGGATGATGGTCGACACAGGGGATGAGATCTACTATACGTTACAGTCTTAGCCGGTTCTCCCTACATGCA									
BU TRINITY_DN3882_c0_g1_i1	-----T.....G...AT.....C.....G.....									
BU TRINITY_DN4163_c0_g1_i1	-----T.....G...AT.....C.....G.....									
AU TRINITY_DN868_c0_g1_i1	-----T.....G...AT.....C.....G.....									
MA TRINITY_DN12415_c0_g1_i1	-----T.....G...AT.....C.....G.....									
MA TRINITY_DN4559_c0_g1_i1	-----T.....G...AT.....C.....G.....									
	910	920	930	940	950	960	970	980	990	1000
B73_G-type_Lectin	 									
BU TRINITY_DN1039_c0_g1_i1	CTTCTCACTTCACCACACAGGCAAGATGAGGCTCCTGGTCTGGGAAATAGCACATCATCCTGGGTGCTGATTCTGAGGAACCTAGTGTGGATGTGAT									
BU TRINITY_DN3882_c0_g1_i1	-----T.....A.....C.....G.C...T....GC.....									
BU TRINITY_DN4163_c0_g1_i1	-----T.....A.....C.....G.C...T....GC.....									
AU TRINITY_DN868_c0_g1_i1	-----T.....A.....C.....G.C...T....GC.....									
MA TRINITY_DN12415_c0_g1_i1	-----T.....A.....C.....G.C...T....GC.....									
MA TRINITY_DN4559_c0_g1_i1	-----T.....A.....C.....G.C...T....GC.....									

	1010	1020	1030	1040	1050	1060	1070	1080	1090	1100
B73_G-type_Lectin		CTCTACGCCTCTTGTGGCCCGTTCGGCTACTGCGATCTCACTAAGGCAATGCCGACATGCCAGTGCCCTGATCATT	TTTGAGCTGGTCGACAGTCTCAACT							
BU TRINITY_DN1039_c0_g1_i1	C.....	T.....	C.TT.....	-----						
BU TRINITY_DN3882_c0_g1_i1	C.....	T.....	C.TT.....	-----						
BU TRINITY_DN4163_c0_g1_i1	C.....	T.....	C.TT.....	-----						
AU TRINITY_DN868_c0_g1_i1	C.....	T.....	C.TT.....	-----						
MA TRINITY_DN12415_c0_g1_i1	C.....	T.....	C.TT.....	-----						
MA TRINITY_DN4559_c0_g1_i1	C.....	T.....	C.TT.....	-----						

	1110	1120	1130	1140	1150	1160	1170	1180	1190	1200
B73_G-type_Lectin		TCTCTAGAGGATGCCAGAGAAAGGAAGAGCTGAAATGCGGAACAGAGAGCTATTTTCATGACCATGCCTAACATGAAGATTCTGACACAT								
BU TRINITY_DN1039_c0_g1_i1	-----T.....	C.....	G.....							
BU TRINITY_DN3882_c0_g1_i1	-----T.....	C.....	G.....							
BU TRINITY_DN4163_c0_g1_i1	-----T.....	C.....	G.....							
AU TRINITY_DN868_c0_g1_i1	-----T.....	C.....	G.....							
MA TRINITY_DN12415_c0_g1_i1	-----T.....	C.....	G.....							
MA TRINITY_DN4559_c0_g1_i1	-----T.....	C.....	G.....							

	1210	1220	1230	1240	1250	1260	1270	1280	1290	1300
B73_G-type_Lectin		TAGAAATAGAACCTTCGACCAGTGTGCAGCTGAGTGCGCCAGAACTGCACCTGTATGGCATACACCTATGCCACTCTAAGCACTGCTAGCACTACGGGC								
BU TRINITY_DN1039_c0_g1_i1	..G.....	C.....	T.....	C.....	G.....	AC.....	G..G.....			
BU TRINITY_DN3882_c0_g1_i1	..G.....	C.....	T.....	C.....	G.....	AC.....	G..G.....			
BU TRINITY_DN4163_c0_g1_i1	..G.....	C.....	T.....	C.....	G.....	AC.....	G..G.....			
AU TRINITY_DN868_c0_g1_i1	..G.....	C.....	T.....	C.....	G.....	AC.....	G..G.....			
MA TRINITY_DN12415_c0_g1_i1	..G.....	C.....	T.....	C.....	G.....	AC.....	G..G.....			
MA TRINITY_DN4559_c0_g1_i1	..G.....	C.....	T.....	C.....	G.....	AC.....	G..G.....			

	1310	1320	1330	1340	1350	1360	1370	1380	1390	1400
B73_G-type_Lectin		GATGCATCGAGGTGTTTGGTCTGGACTCAGGATCTCATTGACATGGAGAAAGCAAGTGTTTTTTAGAGAACTTGATACATCCGGCTTGGCGAATCTCCTG								
BU TRINITY_DN1039_c0_g1_i1	..CA.....	T.....	G.....	CA.C....	G.....					
BU TRINITY_DN3882_c0_g1_i1	..CA.....	T.....	G.....	CA.C....	G.....					
BU TRINITY_DN4163_c0_g1_i1	..CA.....	T.....	G.....	CA.C....	G.....					
AU TRINITY_DN868_c0_g1_i1	..CA.....	T.....	G.....	CA.C....	G.....					
MA TRINITY_DN12415_c0_g1_i1	..CA.....	T.....	G.....	CA.C....	G.....					
MA TRINITY_DN4559_c0_g1_i1	..CA.....	T.....	G.....	CA.C....	G.....					

	1410	1420	1430	1440	1450	1460	1470	1480	1490	1500
B73_G-type_Lectin		GTACGCTTGTGTCTTTCTCTTAGCGTGCCGTCTCTAAGCCCTTGCAACTCCCATTGGTAAGCCTAAGTTTACTTAAATGTATTATAATGCTGCTTTTCCT								
BU TRINITY_DN1039_c0_g1_i1	---ACCC.TGT.CCT..	CT-----								
BU TRINITY_DN3882_c0_g1_i1	---ACCC.TGT.CCT..	CT-----								
BU TRINITY_DN4163_c0_g1_i1	---ACCC.TGT.CCT..	CT-----								
AU TRINITY_DN868_c0_g1_i1	---ACCC.TGT.CCT..	CT-----								
MA TRINITY_DN12415_c0_g1_i1	---ACCC.TGT.CCT..	CT-----								
MA TRINITY_DN4559_c0_g1_i1	---ACCC.TGT.CCT..	CT-----								

	1510	1520	1530	1540	1550	1560	1570	1580	1590	1600
B73_G-type_Lectin										
BU TRINITY_DN1039_c0_g1_i1	CAATGGAAAGTATTTAGAGGTAAAGTATTTGCAGAGCAGTCAGGGGAACAATGTGAAAGTAAACTCACACAGTTGCAAGGAATGCTTGATGCAATAAAT									
BU TRINITY_DN3882_c0_g1_i1	-----									
BU TRINITY_DN4163_c0_g1_i1	-----									
AU TRINITY_DN868_c0_g1_i1	-----									
MA TRINITY_DN12415_c0_g1_i1	-----									
MA TRINITY_DN4559_c0_g1_i1	-----									

	1610	1620	1630	1640	1650	1660	1670	1680	1690	1700
B73_G-type_Lectin										
BU TRINITY_DN1039_c0_g1_i1	TCTAACAAGATACTTTCTTATGTACAGATCAGAAGAAGAGAAATTTTCAGAAAATTTTACTCTCAGCTATAGCATGTCTGCTGCTACTTACAATTGGTGCT									
BU TRINITY_DN3882_c0_g1_i1	-----									
BU TRINITY_DN4163_c0_g1_i1	-----									
AU TRINITY_DN868_c0_g1_i1	-----									
MA TRINITY_DN12415_c0_g1_i1	-----									
MA TRINITY_DN4559_c0_g1_i1	-----									

	1710	1720	1730	1740	1750	1760	1770	1780	1790	1800
B73_G-type_Lectin										
BU TRINITY_DN1039_c0_g1_i1	TCTTGGTTGGGCATGTAAATCAGAGGTATAGAGAAAATGAGCCCTCGTTTCGAAAATGCAGTGTTGATGTGTAAATTTTCAAGGAAAATAAAGTGATCC									
BU TRINITY_DN3882_c0_g1_i1	-----									
BU TRINITY_DN4163_c0_g1_i1	-----									
AU TRINITY_DN868_c0_g1_i1	-----									
MA TRINITY_DN12415_c0_g1_i1	-----									
MA TRINITY_DN4559_c0_g1_i1	-----									

	1810	1820	1830	1840	1850	1860	1870	1880	1890	1900
B73_G-type_Lectin										
BU TRINITY_DN1039_c0_g1_i1	ATTGATGTAACAGGCAAAATGGCACAAGAAGAAAATCCAGACGAGAATGATGCTAGAATACTTGAGTTCTACGGATGAAGCTGGGGGAAAGAATATAGAGT									
BU TRINITY_DN3882_c0_g1_i1	-----									
BU TRINITY_DN4163_c0_g1_i1	-----									
AU TRINITY_DN868_c0_g1_i1	-----									
MA TRINITY_DN12415_c0_g1_i1	-----									
MA TRINITY_DN4559_c0_g1_i1	-----									

	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
B73_G-type_Lectin										
BU TRINITY_DN1039_c0_g1_i1	TCCCATTTATTACCTTTGAAAACATTGTCACAGCAACAGACAATTTCTCTGACTGTAATATACTTGGGAAAGGGGTTTGGCAAAGTATACAAGGTACT									
BU TRINITY_DN3882_c0_g1_i1	-----									
BU TRINITY_DN4163_c0_g1_i1	-----									
AU TRINITY_DN868_c0_g1_i1	-----									
MA TRINITY_DN12415_c0_g1_i1	-----									
MA TRINITY_DN4559_c0_g1_i1	-----									

	2010	2020	2030	2040	2050	2060	2070	2080	2090	2100
B73_G-type_Lectin										
BU TRINITY_DN1039_c0_g1_i1	AGTAAAAAGCTTCACCTTGAAATTATATTACTCAAAAAATATATGATTTTGAATGATGATATCATCTAGGGAATGTTGGAAGGAACCAAGAAGTAGC									
BU TRINITY_DN3882_c0_g1_i1	-----									
BU TRINITY_DN4163_c0_g1_i1	-----									
AU TRINITY_DN868_c0_g1_i1	-----									
MA TRINITY_DN12415_c0_g1_i1	-----									
MA TRINITY_DN4559_c0_g1_i1	-----									

	2110	2120	2130	2140	2150	2160	2170	2180	2190	2200
B73_G-type_Lectin										
BU TRINITY_DN1039_c0_g1_i1	CATCAAGAGACTGAGCAAGAGTTCTGGGCAAGGTGCCGAAGAGTTTCAGGAATGAAGTAATCTTGATTGCAAAGTTGCAGCACAGAACCTAGTTAAGCTT									
BU TRINITY_DN3882_c0_g1_i1	-----									
BU TRINITY_DN4163_c0_g1_i1	-----									
AU TRINITY_DN868_c0_g1_i1	-----									
MA TRINITY_DN12415_c0_g1_i1	A.....G.T.....									
MA TRINITY_DN4559_c0_g1_i1	-----									

	2210	2220	2230	2240	2250	2260	2270	2280	2290	2300
B73_G-type_Lectin										
BU TRINITY_DN1039_c0_g1_i1	CTTGGTTGCTGTGTTTCATGAAGATGAGAAGCTGTTAGTTTATGAGTACTTGCCGAACAAAAGCTTAGATTACTTCTCTTTGGTACATTGAGTTTAAAA									
BU TRINITY_DN3882_c0_g1_i1	-----									
BU TRINITY_DN4163_c0_g1_i1	-----									
AU TRINITY_DN868_c0_g1_i1	-----									
MA TRINITY_DN12415_c0_g1_i1	T.....A.....									
MA TRINITY_DN4559_c0_g1_i1	-----									

	2310	2320	2330	2340	2350	2360	2370	2380	2390	2400
B73_G-type_Lectin										
BU TRINITY_DN1039_c0_g1_i1	CTTTCCATTAGTTAACAAAATCAACAAATATACATGCATATGCCTTCTGATATGTATGCCTTGTTTCATTTTATTCATATTAGATTCTGCAAGAAAGTC									
BU TRINITY_DN3882_c0_g1_i1	-----									
BU TRINITY_DN4163_c0_g1_i1	-----									
AU TRINITY_DN868_c0_g1_i1	-----									
MA TRINITY_DN12415_c0_g1_i1	-----C...T.....									
MA TRINITY_DN4559_c0_g1_i1	-----									

	2410	2420	2430	2440	2450	2460	2470	2480	2490	2500
B73_G-type_Lectin										
BU TRINITY_DN1039_c0_g1_i1	AACGCTTCAGTGGCCAACAAGGTTCAAGATAATCCACGGGGTAGCTAGAGGAATTATGTATCTTCATCAAGACTCGAGATTAACAATAATCCATAGGGAT									
BU TRINITY_DN3882_c0_g1_i1	-----									
BU TRINITY_DN4163_c0_g1_i1	-----									
AU TRINITY_DN868_c0_g1_i1	-----									
MA TRINITY_DN12415_c0_g1_i1	G.....A...									
MA TRINITY_DN4559_c0_g1_i1	-----									

```

                2510      2520      2530      2540      2550      2560
B73 G-type Lectin      ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
BU TRINITY_DN1039_c0_g1_i1      CTCAAAGCAAGCAACATCTTGTTAGATAAGGACATGAGCCCAAAAATATCAGATTTTGGCAT
BU TRINITY_DN3882_c0_g1_i1      -----
BU TRINITY_DN4163_c0_g1_i1      -----
AU TRINITY_DN868_c0_g1_i1      -----
MA TRINITY_DN12415_c0_g1_i1      .....G.....
MA TRINITY_DN4559_c0_g1_i1      -----

```

Figure A.2. An alignment of the de novo assembled transcripts against the B73 version of the G-type Lectin gene at the DNA level. This candidate was present in the unmapped assemblies for both the control (indicated here with AU prefixed to the name), the *C. zeina* challenged (indicated here with BU prefixed to the name) and the master assembly (indicated here with MA prefixed to the name). Conserved bases are indicated with dots.

```

                10      20      30      40      50      60      70      80      90      100
B73 G-type lectin      MMVVPiFILLVVLSSLQSDQLTHTKPLFPKETLISAGGDFALGFFSPTSSSNKLYIGIWIYNNVPERERTVWVIANRDSPTTTPATSVKLAVTNDSELV
BU TRINITY_DN1039      -----
BU TRINITY_DN3882      -----
BU TRINITY_DN4163      .....L.....S..D.....
AU TRINITY_DN868      -----
MA TRINITY_DN12415      -----
MA TRINITY_DN4559      .....L.....S..D.....

                110      120      130      140      150      160      170      180      190      200
B73 G-type lectin      LDSQGHIFWMATSSSTSGGAGAFVLLNSGNFVLRSPNGTDVWQSFDPHTDTPVLPPTMRSMIRYRSQPATGLFAWKSPDDPSTGDISSGMDPASSNVQVFI
BU TRINITY_DN1039      -----
BU TRINITY_DN3882      -----
BU TRINITY_DN4163      .....L.....FG.S.....R.....G.....N..FSL.....L.M..
AU TRINITY_DN868      -----
MA TRINITY_DN12415      .....L.....FG.S.....R.....G.....N..FSL.....L.M..
MA TRINITY_DN4559      .....N.....L.....FG.S.....

                210      220      230      240      250      260      270      280      290      300
B73 G-type lectin      WNGTLPYRSSFENDVPASGGTYQANNTSVTYQWMMVDTGDEIYYTFTVLAGSPYMHFSLHHTGKMRLLVWENSTSSWVLISEEPSVGCDLYASCGPFGY
BU TRINITY_DN1039      -----
BU TRINITY_DN3882      .....F...VV...SV.....GIL.....Y.....I.A.....V.F.G.....
BU TRINITY_DN4163      -----
AU TRINITY_DN868      .....F...VV...SV.....GIL.....Y.....I.A.....V.F.G.....
MA TRINITY_DN12415      -----
MA TRINITY_DN4559      -----

                310      320      330      340      350      360      370      380      390      400

```

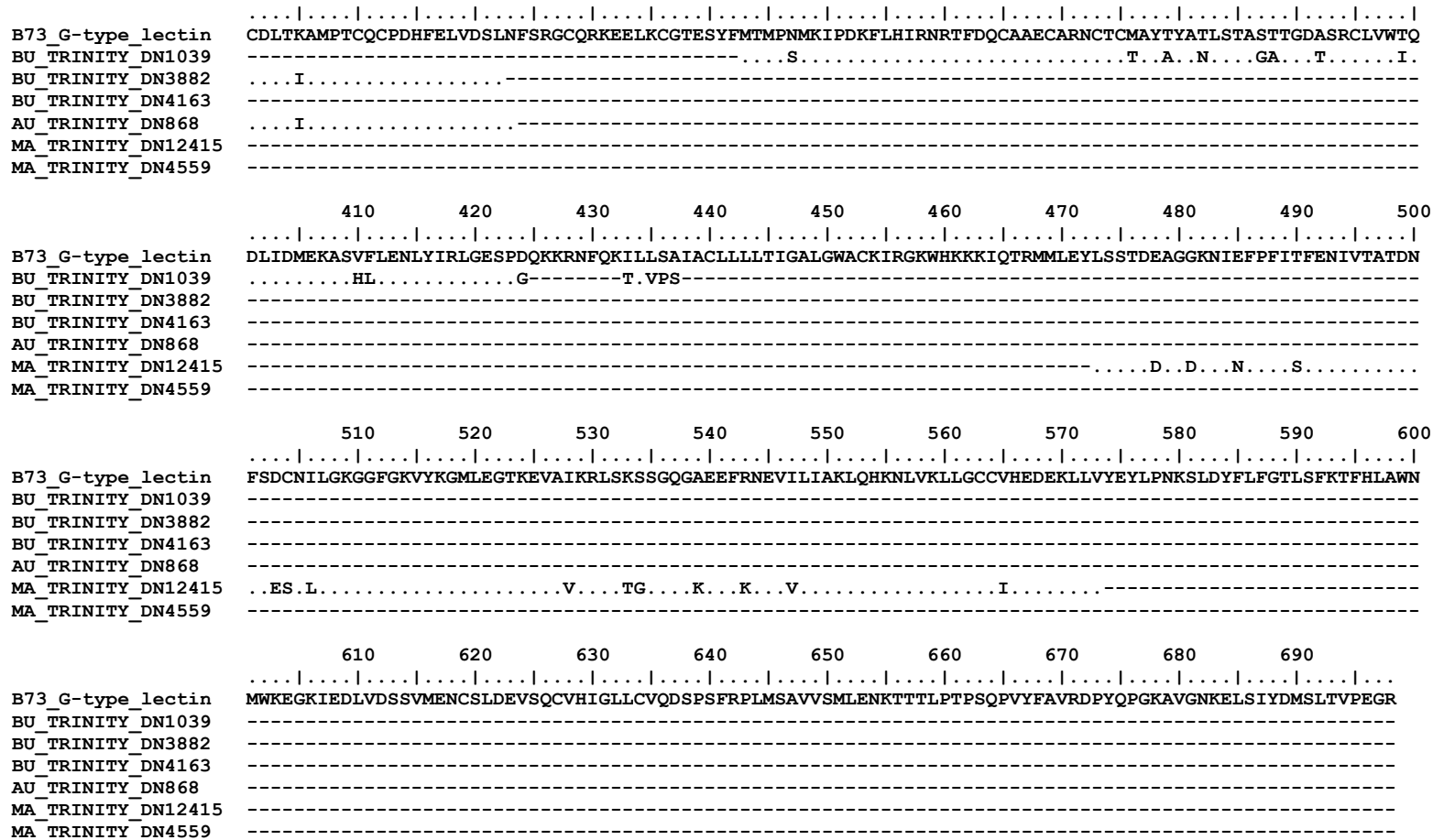


Figure A.3. An alignment of the de novo assembled transcripts against the B73 version of the G-type Lectin gene at the amino acid level. This candidate was present in the unmapped assemblies for both the control (indicated here with AU prefixed to the name), the *C. zeina* challenged (indicated here with BU prefixed to the name) and the master assembly (indicated here with MA prefixed to the name). Conserved bases are indicated with dots.

```

          10      20      30      40      50      60      70      80      90     100
      .....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|
B73_Thaumatn -----GATCTGAGA
TRINITY_DN7795_c0_g1_i1 CTACTAAAGCTACAACCACCGTACATGCTGGGCCTAGCTAGCTTCTTGATTTTACTTGCTATAAATACACGTACAGACAGATACCAGCGTC.....

          110     120     130     140     150     160     170     180     190     200
      .....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|
B73_Thaumatn TCGACATCGATCCACCGAGGTTTCTTGTCGTGCCGCCATGGCCATGGCAGTAGCTCACCTTCTTCCTCAGCTCTCCGCCGAAGGCAATGGAGCTAGCAG
TRINITY_DN7795_c0_g1_i1 .....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|

          210     220     230     240     250     260     270     280     290     300
      .....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|
B73_Thaumatn GAACAGAATGGCCATTCTCTTGATTCTTCAGCTTTTCTCTTGACGG-----CGCTAGGCGTGGATGGAGCGGCGTCGTTTC
TRINITY_DN7795_c0_g1_i1 .....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|
                                     TTCTTTGTTCATAACTGATTCAG.....

          310     320     330     340     350     360     370     380
      .....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|
B73_Thaumatn ATCTTCACCAACGCCTGCAAGCACGCGGTGTGGGTGGGCGCGCTCCACGGCGCCAGCTCGCCGCCGCTCGCGCGGTCTGGGCTTCTA
TRINITY_DN7795_c0_g1_i1 .....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|

```

Figure A.4. An alignment of the de novo assembled transcripts against the B73 version of the Thaumatn gene at the DNA level. This candidate was present only in the master assembly. Conserved bases are indicated with a dot.

```

          10      20      30      40      50      60      70      80      90     100
      .....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|
B73_Thaumatn -----MAMAVAHLLPQLSAEGNGASRNRMAILLILQLFLT-----ALGVDGAASF
TRINITY_DN7795_c0_g1_i1|m.1830 LLKLQPPYMLGLASFLILLAINTRTDYQRRSEIDIDPPRFLVRAA.....VLCSITDS.....

          110     120
      .....|.....|.....|.....|.....|.....|
B73_Thaumatn IFTNACKHAVVVGALHGASSPPLARSGF
TRINITY_DN7795_c0_g1_i1|m.1830 .....|.....|.....|.....|.....|.....|

```

Figure A.5. An alignment of the de novo assembled transcripts against the B73 version of the thaumatn gene at the amino acid level. This candidate was present only in the master assembly. Conserved bases are indicated with a dot.


```

      ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
B73_Chitinase CGCCAACCAGACGCCCTTCAACGGCTGA-----
TRINITY_DN12960_c0_g1_i2 .AAGG.....A...T.A.GGA..A...GGACGTACGTGGACAACAATAATTATATGGTTGCTGCAAGCCTTCACGTGTTTTTGTTCCTCTCTACCAAC
      710      720      730      740      750      760      770      780      790      800
      ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
B73_Chitinase -----
TRINITY_DN12960_c0_g1_i2 TAAGAATAAAATCAAGCTAGCAGCAAACAGGGGCAACCTGCTACCTTCGGCATGTGTCTACTTATCATATAGTGTGTCTATGTATGGCACAGTATCTCATC
      810      820      830      840      850      860      870      880      890      900
      ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
B73_Chitinase -----
TRINITY_DN12960_c0_g1_i2 TCTGGATGCATGATAACGAACATAACGCGTTAGCAGCATTCACCTTTACGTTTCCGGTGGTTCCACTTGGCATTATGACACCAGCAGTGAACAGATG

```

Figure A.8. An alignment of the de novo assembled transcripts against the B73 version of the chitinase gene at the DNA level. This candidate was present only in the master assembly. Conserved bases are indicated with a dot.

```

      10      20      30      40      50      60      70      80      90      100
B73_Chitinase MMRALAVVAMVATALFFAVPARAEQCGSQAGGALCPNCLCCSQFGWCGSTSDYCGSGCQSCGSGCGSTPNPPSSGGVASIIPESLFNQMLLHRNDAACP
Adjusted_TRINITY_DN12960_c0_g1 -----
TRINITY_DN12960_c0_g1_i2 -----RA.A.-----,GRPT.R

      110      120      130      140      150      160      170      180      190      200
B73_Chitinase ANGFTYTYAGFIAAANAFPGFGTTGAPD-----TSHETTGGWATAPDGPYAWGYCFKEEQGGASGPDYCEPSAQWPCAAGKKYYGRGPIQIS
Adjusted_TRINITY_DN12960_c0_g1 -----
TRINITY_DN12960_c0_g1_i2 S-ASWRL.A.L.QT-----VQKRELALAAFLAQ.....

      210      220      230      240      250      260      270      280      290      300
B73_Chitinase YNINYGPAGQAIGAGILANPDLVATDPTVSFETAVFWMTQSPKPSCHDVMTGQWTPSAADTAAGRLPGYGVVTNIINGGLECGHGADSR-VADRIGFY
Adjusted_TRINITY_DN12960_c0_g1 .....A...A...T.R.....LTS.....K.QSTDGAK..V.Y.
TRINITY_DN12960_c0_g1_i2 .....A...A...T.R.....LTS.....K.QSTDGAK..V.Y.

      310      320
B73_Chitinase KRYCDLLGVSYGDNLDCANQTPFNG
Adjusted_TRINITY_DN12960_c0_g1 .....M.R.G...VP.KD.K.YG.
TRINITY_DN12960_c0_g1_i2 .....M.R.G...VP.KD.K.YG.

```

Figure A.9 An alignment of the de novo assembled transcript against the B73 version of the chitinase gene at the amino acid level. This candidate was present only in the master assembly. Deleting one nucleotide base (adjusted sequence) produced a closer alignment indicating either a frame-shift mutation or sequencing error. Conserved bases are indicated with a dot.

	10	20	30	40	50	60	70	80	90	100
B73_Apoptosis										
TRINITY_DN13999_c1_g2_i13	CTTTGGGTTTGGGACTGGGAGCTCCATCGCTCTATGGTATGAACCACGTGTGTTTGTGACACCAATGTGTAGGATCCATATACTCTGTCTAGATTAATT									
	110	120	130	140	150	160	170	180	190	200
B73_Apoptosis										
TRINITY_DN13999_c1_g2_i13	-----ATGGACATGGC--ATTGCCCGT-TGTGAATGCCACAGCGCGGTGCTCGCCCGTGTCTCGGCTGCGTTCAATG AGCAGTGTGTGAATGCCACAGCATACTC..TC.TA.ATTA...AG.A..G.....A.....C.....									
	210	220	230	240	250	260	270	280	290	300
B73_Apoptosis										
TRINITY_DN13999_c1_g2_i13	CCCCACTTGGCCGTGCAGTCGTCTTCGGGGTCCATATCGATGGTAAATCCCTCTTTCCTCATGGATTTCATAAGACGGGTCGGCGGCTT-AGTGTCTCACA G...C.....C.....T.....T.....G-.....T..G.ATG.....T.....AT...T.....									
	310	320	330	340	350	360	370	380	390	400
B73_Apoptosis										
TRINITY_DN13999_c1_g2_i13	TTCAGTCGCGACGGGATTTTCCAGGGCACTTGGTCGTGGAAGGGCTTCTTATTGCAGTCATCGTGTTCAGCTCTCCAGGAAGAGCTACAAACCACCAAA CG....G..C.....T...T.....A.....C.....T.....									
	410	420	430	440	450	460	470	480	490	500
B73_Apoptosis										
TRINITY_DN13999_c1_g2_i13	GAAACCACTTACTGAAAAGGAGATTGATGAGCTATGTGATGAGTGGGAGCCAGACCCGCTATGCCCTCCAGTCAAGGAGGGGCTCGAATTGATACTCCA ...G....C.....									
	510	520	530	540	550	560	570	580	590	600
B73_Apoptosis										
TRINITY_DN13999_c1_g2_i13	ATGTTGGAAGTGCTGCTGGACCACATACGATTGTTGATGGTAAAGAAGTTGTGAACTTTGCATCAGCAAACTACCTCGGTTTAATTGGCAATGAGAAGA A.....C..A....									
	610	620	630	640	650	660	670	680	690	700
B73_Apoptosis										
TRINITY_DN13999_c1_g2_i13	TTATTGATTCTTGCAATTGGTTCACCTGGAGAAATATGGTGTGGGTCTTGTGGTCCAGTGGCTTTATGGAACAATTGATGTCCATCTTGACTGTGAGTC CA.....T.....T.....T.....C.....									
	710	720	730	740	750	760	770	780	790	800
B73_Apoptosis										
TRINITY_DN13999_c1_g2_i13	AAAAATAGCTAAATTCCTGGGCACTCCAGACTCCATTCTGTATTTCATATGGGATTCTTACAATATTCAGTGTGATACCTGCCTTTTGTAAAGAAAGGAGAT ...G.....TT...G.....T.....A.....C.....									
	810	820	830	840	850	860	870	880	890	900
B73_Apoptosis										
TRINITY_DN13999_c1_g2_i13	ATCATAGTCGCTGATGAGGGTGTTCACCTGGGCACTGCAAAATGGTCTCCATTTATCAAGAAGCACTGTGGTGTACTTCAAGCACAATGATATGGTTTCAC C.....C.....									
	910	920	930	940	950	960	970	980	990	1000
B73_Apoptosis										
TRINITY_DN13999_c1_g2_i13	TTGCAAGCACTTTTGAAAACTTACACATGGAATAAATGCTGCTGAAAAGATTAGACGCTACATTGTTGTAGAATCCATTACCAGAATCTGGCCAAAT T.G.....									

```

      1010      1020      1030      1040      1050      1060      1070      1080      1090      1100
      ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
B73_Apoptosis TGCCCCCTTGGATGAAATCGTCAGGTTGAAGGAGAAATATCGGTTCCGTGTTATTCTGGAGGAGCCATTCTTTTGGTGTGCTTGGCAAGTCTGGGCGT
TRINITY_DN13999_c1_g2_i13 .....G..T.....A.....T.....G.....

      1110      1120      1130      1140      1150      1160      1170      1180      1190      1200
      ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
B73_Apoptosis GGCCCTTGCTGAACATTATGGAGTTCTATTGAAAAAATTGATATAATTACTGCTGGAATGGGAAATGCATTAGCTACCGATGGTGGATTTGTACAGGAA
TRINITY_DN13999_c1_g2_i13 .....

      1210      1220      1230      1240      1250      1260      1270      1280      1290      1300
      ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
B73_Apoptosis GTGTCAGAGTTGTTGATCATCAGCGTCTAAGCAGCTCTGGATATGTTTTCTCTGCATCTCTGCCACCTTATCTTGCCACTGCTGCTGTTTCTGCTGTCAA
TRINITY_DN13999_c1_g2_i13 .....G.....C..T..

      1310      1320      1330      1340      1350      1360      1370      1380      1390      1400
      ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
B73_Apoptosis CTACCTGGAGCAGAATCCCGCAGTTCTTGCAAATCTAAGGAGCAATATTGCTCTTTTGATAAAGAATTATCAGATACTCCAGGGCTAGAAATTTTCAGC
TRINITY_DN13999_c1_g2_i13 ...T..T..G.A.....T.....C.....C.....T.....C....

      1410      1420      1430      1440      1450      1460      1470      1480      1490      1500
      ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
B73_Apoptosis CATGTTTTGTACCTATTGCTTCCTTAAGCTGAAGAAATCGACAGGTTCTCCTACCCTGACCTAGACCTTCTTGAAACTATTGCTGGCAGGGTCTTGA
TRINITY_DN13999_c1_g2_i13 .....C.....C.....G....T.....T.....A..A..C....

      1510      1520      1530      1540      1550      1560      1570      1580      1590      1600
      ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
B73_Apoptosis AGGAAGACTCAGTTCTCATTTGTGACATCAAGAAGTCAAATCTGGATAGGTGCAAACCTCCCGTTGGAATCCGCCTGTTTGTATCAGCTGGACATACTGA
TRINITY_DN13999_c1_g2_i13 .....T..G.....A.....T...A.....C.....

      1610      1620      1630      1640      1650      1660      1670      1680      1690      1700
      ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
B73_Apoptosis ATCCGACATCTCCAGGCTTTCCTCATCCTTGAAGCGAGTTTCTGCGGCAGTTCTTTTCAGACTGCTTTTGATCCACATCGGATACCCTTGAAGATGAAAGC
TRINITY_DN13999_c1_g2_i13 ..TA.....A.G.....AT.G.....T.A.....T.....-....C.....

      1710      1720      1730      1740      1750      1760      1770      1780      1790      1800
      ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
B73_Apoptosis CATCCATTTTCTACGCACTTTGTACCCTAGACCGTGTGTGTGTGTTTGGTAAATGTACATAACCTGTACATTTCTACATATTTATGAATTCACCTTGTTG
TRINITY_DN13999_c1_g2_i13 .....G.....T...G.....---.....

      1810      1820      1830      1840      1850      1860      1870      1880      1890      1900
      ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
B73_Apoptosis TCGTTTATACCCGTAGTACATGGGATTTTTTTTTTGTAGAACAGTGATGTTGGAGAATAGTGCGTTTGTGCTGTACATCTCGTTGATTTCCATGTAGTGA
TRINITY_DN13999_c1_g2_i13 .....

      1910      1920      1930      1940      1950
      ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
B73_Apoptosis TGTTTTCCCTGTATCCTCACCCACAAGACATTGGCCGTCTTATATTTTCAAA
TRINITY_DN13999_c1_g2_i13 .....

```

Figure A.10. An alignment of the de novo assembled transcript against the B73 version of the Apoptosis gene at the DNA level. This candidate was present only in the master assembly. Conserved bases are indicated with a dot.

```

          10      20      30      40      50      60      70      80      90     100
B73_Apoptosis      ....|....|....|....|....|....|....|....|....|....|....|....|
TRINITY_DN13999_c1_g2_i13|m.43  MDMALPVVNATAAVLARVSAAFNAPLARAVVFGVHIDGKSLFPHGFIRRVGGLVSHIQSRRDFPGHLVVEGLLIAVIVFQLSRKSYKPPKKPLTEKEIDE
-----

          110     120     130     140     150     160     170     180     190     200
B73_Apoptosis      ....|....|....|....|....|....|....|....|....|....|....|....|
TRINITY_DN13999_c1_g2_i13|m.43  LCDEWEPEPLCPPVKEGARIDTPMLESAAAGPHTIVDGKEVVNFASANYLGLIGNEKIIDSCIGSLEKYGVGSCGPRGFYGTIDVHLDCESKIAKFLGTPD
-----

          210     220     230     240     250     260     270     280     290     300
B73_Apoptosis      ....|....|....|....|....|....|....|....|....|....|....|....|
TRINITY_DN13999_c1_g2_i13|m.43  SILYSYGISTIFSVPFAFCKKGDIIVADEGVHWAVQNGLHLSRSTVVYFKHNDMVSLASTLEKLTHGNKRAEKIRRYIVVESIYQNSGQIAPLDEIVRLK
-----
                                     .A.....R.....

          310     320     330     340     350     360     370     380     390     400
B73_Apoptosis      ....|....|....|....|....|....|....|....|....|....|....|....|
TRINITY_DN13999_c1_g2_i13|m.43  EKYRFRVILEESHSGVVGKSGRGLAEHYGVPIEKIDIIITAGMGNALATDGGFCTGSGVRVVDHQRLSSSGYVFSASLPPYLATAAVSAVNYLEQNPAVLA
-----
                                     S.....E..S...

          410     420     430     440     450     460     470     480     490     500
B73_Apoptosis      ....|....|....|....|....|....|....|....|....|....|....|....|
TRINITY_DN13999_c1_g2_i13|m.43  NLRSNIALHKLSDTPGLEIFSHVLSPIVFLKLKKTGSPTTDLLETIAGRVLKEDSVLIVTSKKSNLDRCKLPVGGIRLFVSAGHTESDISRLSSSL
-----
                                     S.....DKA.....FV.....I.....L.....

          510
B73_Apoptosis      ....|....|..
TRINITY_DN13999_c1_g2_i13|m.43  KRVSAAVLSDCF
-----
                                     .....S....Y.

```

Figure A.11. An alignment of the de novo assembled transcript against the B73 version of the Apoptosis gene at the amino acid level. This candidate was present only in the master assembly. Conserved bases are indicated with a dot.

	10	20	30	40	50	60	70	80	90	100
B73_Ser/Thr_kinase										
MA_TRINITY_DN9878_c0_g1_i2	-----ACTAACAGCGTTTC									
BU_TRINITY_DN1912_c0_g1_i1	TTTGGTTTGTACTAGAGGAGGACCAATCAAGTCAGTCGGCAGACAACATATAAGGTAGACGTTTCGTTATTGTACCGCCTGTGCGGAA.....									
AU_TRINITY_DN681_c0_g1_i1 GAAC	-----									
	110	120	130	140	150	160	170	180	190	200
B73_Ser/Thr_kinase										
MA_TRINITY_DN9878_c0_g1_i2	ATGAATCATGACGAAGACGAACACCATGAGAAGCCACCACCATGGATACAGCCACCGCCACTGCCACCTACTCGTACAGTCTTCAATCCTGCTTTTCCTC									
BU_TRINITY_DN1912_c0_g1_i1	G...TC--C.TG.TT..A....C.....									
AU_TRINITY_DN681_c0_g1_i1 GAAC	-----									
	210	220	230	240	250	260	270	280	290	300
B73_Ser/Thr_kinase										
MA_TRINITY_DN9878_c0_g1_i2	GGAACTTTTGCCCGCGCACAAAGCTGCATCCGACATCCTCAGCAAGGGCAGTAACCTTACTAACGGCGAGACCCTGGTCTCCGCTAACGGGTCGTTTCACAT									
BU_TRINITY_DN1912_c0_g1_i1	T.....CT.T.....									
AU_TRINITY_DN681_c0_g1_i1 GAAC	-----									
	310	320	330	340	350	360	370	380	390	400
B73_Ser/Thr_kinase										
MA_TRINITY_DN9878_c0_g1_i2	TAGGGTCTTCACTCGCGGCGTGCCGGCAAGGAGGTACCTGGGCATCTGGTTACCGGTGGCCAACCTCCAGCAGCGACGCTGTATGCTGGGTGGCGAACCG									
BU_TRINITY_DN1912_c0_g1_i1	T.....A.....A.....T.....G.....A.....									
AU_TRINITY_DN681_c0_g1_i1 GAAC	-----									
	410	420	430	440	450	460	470	480	490	500
B73_Ser/Thr_kinase										
MA_TRINITY_DN9878_c0_g1_i2	GGACCTCCCTCTCGGCGACACCTCCGGCGTGCTGGTGATCAGCGACACGGGAAGCCTTGTCTGCTCGACGGTTCTGGCCGGACGGCGTGGTCTTCGAAC									
BU_TRINITY_DN1912_c0_g1_i1	T...A.....T.T.....C.....T.....C.....G.....									
AU_TRINITY_DN681_c0_g1_i1 GAAC	-----									
	510	520	530	540	550	560	570	580	590	600
B73_Ser/Thr_kinase										
MA_TRINITY_DN9878_c0_g1_i2	ACGACTGCAGGCGCGG-----CTTCCCCGACGGTGAAGCTACTCGAGTCCGGCAACCTCGTCTGCTCGACGGGAACGGCGGCGG---CGACGACTACG									
BU_TRINITY_DN1912_c0_g1_i1	C..C.....GCGCCG.....A.....G.....A...AG.AGT.....									
AU_TRINITY_DN681_c0_g1_i1 GAAC	-----G.....A...AG.AGT.....									
	610	620	630	640	650	660	670	680	690	700
B73_Ser/Thr_kinase										
MA_TRINITY_DN9878_c0_g1_i2	ACGTGGTGAAGCTGTGGCAGTCGTTTCGATCACCCGACGAACACCTTGCTCCCGGGCGCCAAGATCGGCATGAACCTGTGGAGCGGCGGGGGTGGTCACT									
BU_TRINITY_DN1912_c0_g1_i1	G...C.....C.....T.....C.....									
AU_TRINITY_DN681_c0_g1_i1 GAAC	G...C.....C.....T.....C.....									

	710	720	730	740	750	760	770	780	790	800	
B73_Ser/Thr_kinase											
MA_TRINITY_DN9878_c0_g1_i2	CACGTCGTGGCGGGACGCCGACGACCCCTCGACGGGGGAATTCCGGTACCGCATGGTCAGGCGAGGAGGGTTGCTGCCCCGAGATCGTGATGCTGGACTCG										
BU_TRINITY_DN1912_c0_g1_i1	T..T.....C.....A.....---										
AU_TRINITY_DN681_c0_g1_i1 GAAC	T..T.....C.....A.....---										
	810	820	830	840	850	860	870	880	890	900	
B73_Ser/Thr_kinase											
MA_TRINITY_DN9878_c0_g1_i2	AGCGATGCCATCAAGTACCGCACGGGAGTGTGGAACGGGCGGTGGTTCAGCGGCATCCCGGAGATGAACTCGTACTCGAACATGTTTCGTCTTCCATGTGA										
BU_TRINITY_DN1912_c0_g1_i1	C.....T.....G....										
AU_TRINITY_DN681_c0_g1_i1 GAAC	C.....T.....G....										
	910	920	930	940	950	960	970	980	990	1000	
B73_Ser/Thr_kinase											
MA_TRINITY_DN9878_c0_g1_i2	CGGTGAGCCAGAGCGAGGTCAGCTTCAGCTACGCCGCCAATGCCGCGCGCGCCATCCCTTTCCCGCGTCCTCCTCAACTACACGGCTGAGGCGGTGCG										
BU_TRINITY_DN1912_c0_g1_i1	...A...C.....A.....G.....G.....C.....										
AU_TRINITY_DN681_c0_g1_i1 GAAC	...A...C.....A.....G.....G.....C.....										
	1010	1020	1030	1040	1050	1060	1070	1080	1090	1100	
B73_Ser/Thr_kinase											
MA_TRINITY_DN9878_c0_g1_i2	CGTCGTGTGGGTGCCGACAAAGCGAGGGTGGGCAAACTTCTTCACGGGACCCCGAGAAGACTGCGACCCTACAAACAGGTGCGGGCACCTCTGGCGTGTGC										
BU_TRINITY_DN1912_c0_g1_i1	TG..T.....AC.....G..G....T.....T.....										
AU_TRINITY_DN681_c0_g1_i1 GAAC	TG..T.....AC.....G..G....T.....T.....										
	1110	1120	1130	1140	1150	1160	1170	1180	1190	1200	
B73_Ser/Thr_kinase											
MA_TRINITY_DN9878_c0_g1_i2	AATCAGACTGCAGCGTCGACGGCGTGGCCGTGCAGCTGTGTCCAGGGCTTCGTCCCCGTCTCGTCGTCGGAAGTGGGACGGGAGAGACCCATCTGGCGGGT										
BU_TRINITY_DN1912_c0_g1_i1	T.....A.....T.....										
AU_TRINITY_DN681_c0_g1_i1 GAAC	T.....A.....T.....										
	1210	1220	1230	1240	1250	1260	1270	1280	1290	1300	
B73_Ser/Thr_kinase											
MA_TRINITY_DN9878_c0_g1_i2	GCCGGCGGAACGTGTGCTGGACTGCGGGACAAATGGCACCACGGACGGGTTTGTCCGTTTGCCGGGAGTGAAGCTGCCCGACACGCTCAACTCGTCGCT										
BU_TRINITY_DN1912_c0_g1_i1											
AU_TRINITY_DN681_c0_g1_i1 GAAC											
	1310	1320	1330	1340	1350	1360	1370	1380	1390	1400	
B73_Ser/Thr_kinase											
MA_TRINITY_DN9878_c0_g1_i2	GGACACGAGCATCACGCTGGACGAGTGCAGGGCGAAGTGCTTGCCAACCTGCTCCTGCGTGGCGTATGCCGCGGCAGATGTGCAAGGCGGAGGCGATGAT										
BU_TRINITY_DN1912_c0_g1_i1	G.....										
AU_TRINITY_DN681_c0_g1_i1 GAAC											

	1410	1420	1430	1440	1450	1460	1470	1480	1490	1500
B73_Ser/Thr_kinase										
MA_TRINITY_DN9878_c0_g1_i2	GTTAGCACTGGATGCATCATGTGGCCTGAAAACCTCACTGACCTACGCTACGTGGCTGGAGGACAGACCTGTACCTACGGCAGGCTACTCCCCATCTG									
BU_TRINITY_DN1912_c0_g1_i1	...G.....									
AU_TRINITY_DN681_c0_g1_i1 GAAC	-----									
	1510	1520	1530	1540	1550	1560	1570	1580	1590	1600
B73_Ser/Thr_kinase										
MA_TRINITY_DN9878_c0_g1_i2	GGACAGGAAGGTCAAAGAAAGAAACGGTGGTTCTCATAGCAGCTGGATCAATATTGGGCTTCATTGTCCTTGTCATGCTGGCCATCTTCGCTGTGCAGGC									
BU_TRINITY_DN1912_c0_g1_i1	T.....T.....CG.....G.....C.....T.....G....A.....									
AU_TRINITY_DN681_c0_g1_i1 GAAC	-----									
	1610	1620	1630	1640	1650	1660	1670	1680	1690	1700
B73_Ser/Thr_kinase										
MA_TRINITY_DN9878_c0_g1_i2	GGTTCGGATAAGGCGTCGGAACCTTGATAATTCAAATGACGGAGGCAGTTGAAACAGCCCAAGATCCCTCTGTTTCTTCCATTGCTCTGGCTACTGTCAAG									
BU_TRINITY_DN1912_c0_g1_i1	T.....									
AU_TRINITY_DN681_c0_g1_i1 GAAC	-----									
	1710	1720	1730	1740	1750	1760	1770	1780	1790	1800
B73_Ser/Thr_kinase										
MA_TRINITY_DN9878_c0_g1_i2	AGTGCAACAAGGAATTTCTCCACAAGGAATGTGATAGGCGAAGGCACTTTCGGCATCGTATATGAGGGCAAGTTGCCCAGAGGGCATCCGCTCCTACACG									
BU_TRINITY_DN1912_c0_g1_i1										
AU_TRINITY_DN681_c0_g1_i1 GAAC	-----									
	1810	1820	1830	1840	1850	1860	1870	1880	1890	1900
B73_Ser/Thr_kinase										
MA_TRINITY_DN9878_c0_g1_i2	TGCTAGCCGGGAGAACCATTTGCCGTGAAGAGGCTGAAATCGATCGGCGATCTTCCGGACATAATCGTCAGGTATTTACAGAGAGAGATGCAGCTCATGTC									
BU_TRINITY_DN1912_c0_g1_i1	G.....C.....									
AU_TRINITY_DN681_c0_g1_i1 GAAC	-----									
	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
B73_Ser/Thr_kinase										
MA_TRINITY_DN9878_c0_g1_i2	CGGGCTCAAGCAGCACCGGAATGTGCTCCGCTCCTTGCCCTACTGCGACGAAGCCAGCGAACGGATCCTGGTGTACGAGTACATGCACAGGAGGAGCTTG									
BU_TRINITY_DN1912_c0_g1_i1										
AU_TRINITY_DN681_c0_g1_i1 GAAC	-----									
	2010	2020	2030	2040	2050	2060	2070	2080	2090	2100
B73_Ser/Thr_kinase										
MA_TRINITY_DN9878_c0_g1_i2	GACTCCTACATATTTCGGAACACCTAGAGAACGCGCGCTGCTGAACTGGCGCCGGAGGCTGCAGATCATTTCAGGGGATCGCCGACGGCGTGAAGCACCTCC									
BU_TRINITY_DN1912_c0_g1_i1	...G.....T.....									
AU_TRINITY_DN681_c0_g1_i1 GAAC	-----									

	2110	2120	2130	2140	2150	2160	2170	2180	2190	2200	
B73_Ser/Thr_kinase											
MA_TRINITY_DN9878_c0_g1_i2	ACGAGGGAGAAGGGTCGTC	CGGCCAACGTGATCCACCGGGATCTGAAGCCGGCCAATGTGCTGCTGGACGGCGGATGGCAGGCCAAGGTGGCCGACTTCGG									
BU_TRINITY_DN1912_c0_g1_i1	G.....										
AU_TRINITY_DN681_c0_g1_i1 GAAC	-----										
	2210	2220	2230	2240	2250	2260	2270	2280	2290	2300	
B73_Ser/Thr_kinase											
MA_TRINITY_DN9878_c0_g1_i2	GACAGCAAAATTGCTCGTCGCTGGAGCTACCGGAACTCGGACAAGGATAGGCACAGCGTAAGAAAGATCACCGCAGTGGGCTTGCTTCGAATCAGCTCG										
BU_TRINITY_DN1912_c0_g1_i1	C.....										
AU_TRINITY_DN681_c0_g1_i1 GAAC	-----										
	2310	2320	2330	2340	2350	2360	2370	2380	2390	2400	
B73_Ser/Thr_kinase											
MA_TRINITY_DN9878_c0_g1_i2	CTCTTTCTTCAGTGTATATATCTACTACTAATATTTACATCATCGTTTCTTTTCTTCTACTGGGAACTCACTCTGTTCTGAATGTTATGGTACC										
BU_TRINITY_DN1912_c0_g1_i1	-----										
AU_TRINITY_DN681_c0_g1_i1 GAAC	-----										
	2410	2420	2430	2440	2450	2460	2470	2480	2490	2500	
B73_Ser/Thr_kinase											
MA_TRINITY_DN9878_c0_g1_i2	GTGGATACATGGCTCCAGAGTACGTTTCAGAGCGACGGTAGCGAGACGACCTCAAGTGCACGTGTATAGCTTCGGAGTCACCTTAATGGAGACACTGAG										
BU_TRINITY_DN1912_c0_g1_i1	G.....										
AU_TRINITY_DN681_c0_g1_i1 GAAC	-----										
	2510	2520	2530	2540	2550	2560	2570	2580	2590	2600	
B73_Ser/Thr_kinase											
MA_TRINITY_DN9878_c0_g1_i2	CGGACGAAAGAACTGTGACACGCCAGGCCCTCGTCTCAGAAGCCTGGAGACTCTGGGTGGGCCGCTGCGTCACGGCACTCCTTGATCCAGCGGTTCACCCG										
BU_TRINITY_DN1912_c0_g1_i1	A.....AA.....										
AU_TRINITY_DN681_c0_g1_i1 GAAC	-----										
	2610	2620	2630	2640	2650	2660	2670	2680	2690	2700	
B73_Ser/Thr_kinase											
MA_TRINITY_DN9878_c0_g1_i2	GCGCCTGCCAAGCCCAGCTTGCGCAACTGCGCAGGTGCATCCAGGTTCGGCTCCTCTGCGTCCAGGAGAAGCCGGACGAGAGGCCCGCATGTCCGCGG										
BU_TRINITY_DN1912_c0_g1_i1	A.T.AA.T.....GG.C.T-----										
AU_TRINITY_DN681_c0_g1_i1 GAAC	-----										
	2710	2720	2730	2740	2750	2760	2770	2780	2790	2800	
B73_Ser/Thr_kinase											
MA_TRINITY_DN9878_c0_g1_i2	TTGTGAGATGCTAGGCAGCCCCTGCTCGGAGCTTGCCGAGCCCATGGTGCCGACGGTCGTCGGCAATGCAGCCTTGGCTACTCTCCTGGAGGCCGATCT										
BU_TRINITY_DN1912_c0_g1_i1	---.....G.....										
AU_TRINITY_DN681_c0_g1_i1 GAAC	-----										

```

                2810      2820      2830      2840      2850      2860      2870      2880      2890      2900
B73_Ser/Thr_kinase  ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
MA_TRINITY_DN9878_c0_g1_i2  CTCCAGACCGACGGTGTACGAGACAATCGATTTTAGATAAGAGATGACGCGCTCGCACTCGCTTGTAGCTTATAACACTGTTTATACTTTCGTA
BU_TRINITY_DN1912_c0_g1_i1  .....A.....TATAC.....
AU_TRINITY_DN681_c0_g1_i1 GAAC -----

                2910      2920      2930      2940      2950      2960      2970
B73_Ser/Thr_kinase  ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
MA_TRINITY_DN9878_c0_g1_i2  CTGTTGTTTTACACTGATTATAATTGTTATTAGTAACGTTTTTGT-----
BU_TRINITY_DN1912_c0_g1_i1  .....A...A.....GTTTGTGGTCGAAAAGACATGACGGTACTAAGG
AU_TRINITY_DN681_c0_g1_i1 GAAC -----

```

Figure A.12. An alignment of the de novo assembled transcripts against the B73 Serine/Threonine kinase (Ser/Thr_kinase) at the DNA level. This candidate was present in the unmapped assemblies for both the control (indicated here with AU prefixed to the name), the C. zeina challenged (indicated here with BU prefixed to the name) and the master assembly (indicated here with MA prefixed to the name). Conserved bases are indicated with dots.

```

                10      20      30      40      50      60      70      80      90      100
B73_Ser/Thr_kinase  ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
BU_TRINITY_DN1912_c0_g1_i1|m.1  SSNTTAGAASPTVKLLESGNLVLLDGN-GRDDYDVVKLWQSFDPHTNTLLPGAKIGMNLWSSGGWSLTSWRDADDPSTGEFRYAMVRRGGLLPEIVMLD
AU_TRINITY_DN681_c0_g1_i1|m.22  NTTAG.....D..V...G.....I.....P....T....-.....T..

                110      120      130      140      150      160      170      180      190      200
B73_Ser/Thr_kinase  ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
BU_TRINITY_DN1912_c0_g1_i1|m.1  SSDAIKYRTGVWNGRWFSGIPEMNSYSNMFVHVTVSQSEVSFSYAANAGAPPSLSRVLLNYTAEAVRVVWVPDKRGWANFFFTGPREDCHYNRCSHSGV
AU_TRINITY_DN681_c0_g1_i1|m.22  .....F....Q...P...Y...K.....D.....WL....D.....D.....

                210
B73_Ser/Thr_kinase  ....|....|...
BU_TRINITY_DN1912_c0_g1_i1|m.1  CNQTAASTAWPC
AU_TRINITY_DN681_c0_g1_i1|m.22  ..H.....T...

```

Figure A.13. An alignment of the de novo assembled transcripts against the B73 Serine/Threonine kinase (Ser/Thr_kinase) at the amino acid level. This candidate was present in the unmapped assemblies for both the control (indicated here with AU prefixed to the name), the C. zeina challenged (indicated here with BU prefixed to the name) and the master assembly (indicated here with MA prefixed to the name). Conserved bases are indicated with dots.

```

                10      20      30      40      50      60      70      80      90      100
B73_LRR             ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
MA_TRINITY_DN1034_c0_g1_i1  TCAAGCTTTTGGATCCCCACGCTCCCCACCCATTCTTCTGTAATGCGGTGTTCTCCAAGCTCGGGCTGCAGGTGTGGTCGCCATGGTTCAAAGGCGACC
AU_TRINITY_DN305_c0_g1_i1  A..G..CC..ATC.GTGAC...GTTTG..G.G.AGGCG...-..C.....G.....C.....G...CC.....
BU_TRINITY_DN5447_c0_g1_i1  A..G..CC..ATC.GTGAC...GTTTG..G.G.AGGCG...-..C.....G.....C.....G...CC.....

```

	110	120	130	140	150	160	170	180	190	200
B73_LRR		CGTGCGTCGGCGCGGCCACGGACGGCACGTCCATCTACGCCGACAAAATGAACCCGGGCATCAAGTGCGACTGCTCCGACCACAACAACACCATCTGCCA								
MA_TRINITY_DN1034_c0_g1_i1	A..A.....C.....A.....									
AU_TRINITY_DN305_c0_g1_i1	A..A.....C.....A.....									
BU_TRINITY_DN5447_c0_g1_i1	A..A.....C.....A.....									
	210	220	230	240	250	260	270	280	290	300
B73_LRR		CATTACCGAGCTGAAAATGAACAACTGGATGTGGTTGGTCTATACCTGAGGAGCTGAGGAACCTCAGCCACTTGACAACCTCTGAATTACAAACAGAAAT								
MA_TRINITY_DN1034_c0_g1_i1	...C.T.....	T.....								
AU_TRINITY_DN305_c0_g1_i1	...C.T.A...-CG.C..CT.CG..CAC.-	-----								
BU_TRINITY_DN5447_c0_g1_i1	...C.T.....	T.....------								
	310	320	330	340	350	360	370	380	390	400
B73_LRR		TACTTAACAGGGCCTTTGCCATCGTTCCCTGGAGAATTAAGTCTTTCGAGCAGATGTCGCTTTCTGGCAATGCATTATCAGGAACCTATTCCAAAGGAGC								
MA_TRINITY_DN1034_c0_g1_i1										
AU_TRINITY_DN305_c0_g1_i1	-----									
BU_TRINITY_DN5447_c0_g1_i1	-----									
	410	420	430	440	450	460	470	480	490	500
B73_LRR		TGGGGAACCTTGTGAATCTCATAATCCTGAGTTTAGGCACGAACAATTTAACTGGCTCCCTTCTCTGAATTGGGTAACTAGAGCATAT								
MA_TRINITY_DN1034_c0_g1_i1										
AU_TRINITY_DN305_c0_g1_i1	-----									
BU_TRINITY_DN5447_c0_g1_i1	-----									
	510	520	530	540	550	560	570	580	590	600
B73_LRR		GTACATTGATAGTGCTGGCTTAAGTGGTCTATACCATCATCATTTTCGAAACTTACACGATTGAAAACATTGTGCGCATCAGATAATGACTTTACTGGG								
MA_TRINITY_DN1034_c0_g1_i1	C.....									
AU_TRINITY_DN305_c0_g1_i1	-----									
BU_TRINITY_DN5447_c0_g1_i1	-----									
	610	620	630	640	650	660	670	680	690	700
B73_LRR		AAAATACCAGATTACATTGGAAGCTGGAGTAATTTAACAGATTTCGCGGTTTCAAGGCAATTCATTCCAAGGTCCACTTCCAAGCACTCTTGCTAATCTTG								
MA_TRINITY_DN1034_c0_g1_i1										
AU_TRINITY_DN305_c0_g1_i1	-----									
BU_TRINITY_DN5447_c0_g1_i1	-----									
	710	720	730	740	750	760				
B73_LRR		TTCAGCTAACAAGCTTACGAATAGGTGACATTGTAAGTGGGAGTTCTTCGCTGGCATTATCAGTA								
MA_TRINITY_DN1034_c0_g1_i1										
AU_TRINITY_DN305_c0_g1_i1	-----									
BU_TRINITY_DN5447_c0_g1_i1	-----									

Appendix B: Next generation sequencing reveals historical and widespread occurrence of Maize Yellow Mosaic Virus in South Africa

T. Welgemoed, R. Pierneef, A. Schulze, G. Pietersen and D.K. Berger

Author Contributions

T.W.: Conducted the bioinformatics pipeline and the laboratory procedures on the RSA field-trial and Tanzania RNA samples. Wrote the initial draft of the manuscript.

R.P.: Guided the design of the bioinformatics pipeline and suggested methods. Edited the manuscript.

A.S: Conducted the laboratory procedures on the RSA survey RNA samples and optimization reaction.

G.P: Suggested methods. Edited the manuscript.

D.K.B.: Principal Investigator of the funding project, acquired funding, provided day-to-day institutional supervision, and edited the manuscript. Led discussions on data interpretation.

D.K.B., G.P, R.P. and T.W. designed the study.

Abstract

Maize yellow mosaic virus (MaYMV) is a single-stranded RNA polerovirus first identified in China from samples collected in 2013. MaYMV was recently reported from West and East Africa, however it had not yet been reported from the major maize producing region of southern Africa. RNA-seq data from South African field-grown fungal-infected maize was mined for viral sequences by *de novo* assembly of reads that did not map to the maize or fungal genomes, prediction of protein coding sequences, and BLASTP searches against the NCBI database. Predicted proteins from the *de novo*-assembled unmapped reads matched MaYMV proteins with regions of 96-100% identity. A complete 5542 nucleotide MaYMV genome was assembled, and this showed 98-99% identity to African isolates, and 97% identity to the reference isolate from China. MaYMV was also detected in RNA-seq data from samples collected in 2009 and 2012 at a nearby site. A 2016/2017 RT-PCR survey of maize plants exhibiting virus-like symptoms, such as yellowing and mosaic patterns, revealed MaYMV in 39 samples from six maize production provinces in South Africa. This

report documents the earliest known MaYMV infection world-wide, and indicates that the virus is widespread throughout maize cropping systems in Africa.

Introduction

The pipeline in Chapter 2 resulted in the detection of some proteins that were not similar to plant or fungal proteins (Fig. 2.8, Table A.3) that were unexpectedly identified as belonging to Maize yellow mosaic virus (MaYMV). The presence of MaYMV in these and other samples collected between 2009 and 2017 was detected using RNA-seq and/or a virus specific RT-PCR, and demonstrate the widespread occurrence of this virus in South African maize.

MaYMV is the provisional name of a recently characterised maize RNA virus identified in China and placed in the *Polerovirus* genus (Chen *et al.* 2016). A parallel study in China discovered and sequenced the same virus, which they named Maize yellow dwarf virus (MYDV)-RMV2 (Wang *et al.* 2016), and other authors have called it MYDV-like polerovirus or MYDV-RMV (Massawe *et al.* 2018; Wamaitha *et al.* 2018). We utilise the commonly used MaYMV designation in this study. The *Polerovirus* genus belongs to the family *Luteoviridae*, and consists of single stranded RNA positive-strand viruses. Poleroviruses are generally aphid-transmitted and phloem limited (Hogenhout *et al.* 2008). The transmission method for MaYMV has not been confirmed yet, but it does possess the polerovirus P3–P5 read-through protein that is associated with aphid transmission (Chen *et al.* 2016). The presence of MaYMV in maize leaf material is associated with mild mosaic symptoms and yellowing of the leaves (Chen *et al.* 2016; Palanga *et al.* 2017).

MaYMV has been identified using next generation sequencing after extraction of RNA from maize leaves collected on three continents – Asia, Africa and South America (Chen *et al.* 2016; Palanga *et al.* 2017; Gonçalves *et al.* 2017). Samples collected in 2013 in the Yunnan and Anhui provinces were the source of the first reports in China (Chen *et al.* 2016; Wang *et al.* 2016). MaYMV was reported in maize fields surrounding the Agricultural Research Centre of Kamboinsé in Burkina Faso, West Africa in 2016 (Palanga *et al.* 2017). In East Africa, MaYMV was present in maize samples collected from 2012-2014 in Kenya (Wamaitha *et al.* 2018), from 2013-2016 in Kenya, Tanzania, Rwanda and Ethiopia (Massawe *et al.* 2018), from 2015 and 2016 in Ethiopia (Guadie *et al.* 2017) and from Tanzania in 2015 (Read *et al.* 2018). MaYMV was reported from Sao Paulo state in Brazil (2013 and 2017) and Los Rios region of Ecuador (2016) in South America (Gonçalves *et al.* 2017; Bernreiter *et al.* 2017).

MaYMV has also been reported from other grasses, namely sugar cane and the itch weed (*Rottboellia cochinchinensis*) in the northern Guinea savannah region of Nigeria (Yahaya *et al.* 2017), and from *Panicum milliaceum* in South Korea, and sorghum in Kenya and South Korea (Lim *et al.* 2017; Wamaitha *et al.* 2018). These reports suggest that the occurrence of the virus is relatively widespread, although to our knowledge there have not been any reports of the virus from southern Africa. Currently no data is available on the impact of MaYMV on agricultural yields.

Methods

The pipeline described in Chapter 2 was used on RNA-seq data from three B73 maize plants that were in the same field-trial. Aliquots of RNA from the samples that were RNA-sequenced had been stored at -80°C and were used for RT-PCR confirmation of MaYMV in the samples.

Additional maize material for which RNA-seq data was screened by mapping of Illumina reads to the MaYMV reference genome (KU248489) were the following samples collected at 103 days after planting (dap) on 24 March 2009 from the field trial at Baynesfield Estate, KwaZulu-Natal (described in (Christie *et al.* 2017)): two parental inbred lines CML444 and SC Malawi (pool of leaves from three plants each), three plants each of RIL165 and RIL387 (described in (Meyer *et al.* 2017)); and three biological replicates each of six gray leaf spot (GLS) susceptible RILs (RIL31, RIL259, RIL284, RIL329, RIL349, RIL362) and six GLS resistant RILs (RIL125, RIL166, RIL327, RIL366, RIL381, RIL387). RIL387 Maize leaf material of SC Malawi was also collected in 27 March 2012 (96 dap) from a field trial at Baynesfield estate, made up three biological replicate plants each from fungicide sprayed and unsprayed plots. RNA extraction and RNA-sequencing of RIL165 and RIL387 on an Illumina platform (200 bp inserts; 50 nt paired end reads) at BGI, Shenzhen, China has been described in (Meyer *et al.* 2017). The same protocols were followed for the other genotypes except that the 2009 CML444 and SC Malawi samples were Illumina sequenced (100 nt paired end reads) at Fasteris SA, Geneva, Switzerland; and the 2012 SC Malawi samples were sequenced at BGI (90 nt paired end reads).

Maize leaf material with putative symptoms of virus infection (e.g. mosaics, mottles, yellow streaking) was collected from maize fields during a survey carried out in the 2016/2017 maize growing season in the following provinces of South Africa: Free State, Gauteng, KwaZulu-Natal, Limpopo, Mpumalanga, North West, and Northern Cape. A total of 125 leaf samples were collected and a selection of 44 representing each province was screened for MaYMV by RT-PCR.

The presence of MaYMV in maize samples previously subjected to RNA sequencing (Table B.2) was evaluated by mapping the RNA-seq reads against a reference MaYMV genome (KU248489) using Bowtie2 with the default settings (Langmead and Salzberg 2012). Raw reads were used apart from CML444 samples GKV-4 and GKV-5 reads (Table 2) in which the first 10 bp and last 40 bp were trimmed with FASTX trimmer (Gordon and Hannon 2010).

The upper leaf sample (BR1A) of maize B73 inbred line from Greytown (2013) had the most complete RNA-seq read coverage of the MaYMV reference genome, and therefore this was used to assemble a MaYMV genome as a representative South African isolate. The reads from BR1A that mapped against a MaYMV reference genome (KU248489) using Bowtie were *de novo* assembled using Trinity. The assembled MaYMV genome was annotated using CLC Main Workbench v8.01 to identify the coding regions. The annotated MaYMV RSA BR1A genome sequence was submitted to GenBank under the accession MG570476.

RT-PCR was used to confirm the presence of MaYMV in the RNA-seq samples and to survey South African maize growing areas for MaYMV. RNA from the same samples that had been used for RNA-seq and had been stored at -80 °C, as well as RNA extracted from fresh material collected during the 2017 maize virus survey was used. RNA was extracted using the CTAB (cetyltrimethylammonium bromide) method as described in (White *et al.* 2008).

RT-PCR with the primer pair MaYMV-F and MaYMV-R (Chen *et al.* 2016) was carried out. First, the annealing temperature for the RT-PCR was optimized using a virus positive maize RNA sample with cDNA synthesis using the M-MLV RNase H- system (Promega, Madison, USA), and PCR using MyTaq. The gradient PCR reactions were set up in a total volume of 22 µl, which consisted of: 5 µl MyTaq Reaction buffer, 0.5 µm of each primer, 0.25 µl MyTaq DNA polymerase, 3 µl of the cDNA, and sterile distilled water. The cycling conditions were as follows: 1 min at 95 °C followed by 40 cycles of 15 s at 95 °C, 15 s at the annealing temperature (58, 56.1, 53.8 or 51.9 °C), and 10 s at 72 °C. The remaining samples were screened for MaYMV using the optimal annealing temperature and the OneTaq® One-Step RT-PCR Kit (New England BioLabs, Ipswich, USA). The RT-PCR reactions were set up in a total volume of 25 µl, which consisted of: 2× Quick-Load One Taq One-Step Reaction Mix, 0.5 µm of each primer, 1 µl of 25X One Taq One-Step Enzyme Mix, 1 µl of the RNA, and sterile distilled water. The cycling conditions were as follows: 15 min at 48 °C, 1 min at 94 °C followed by 40 cycles of 15 s at 94 °C, 30 s at 58 °C, and 45 s at 68 °C, and a final extension step of 68 °C for 5 min. PCR products were visualized by agarose gel electrophoresis. The positive control for RNA and cDNA quality of MaYMV-negative samples involved RT-PCR

with primers actinF and actinR (Van den Berg *et al.* 2004) that flank an intron in the maize actin gene. The RT-PCR protocol was the same as used for the MaYMV primers.

Selected RT-PCR products were Sanger sequenced using the BigDye® reagents, and processed on an ABI 3500xl Genetic Analyser (Applied Biosystems, Foster City, USA) using the methods described in (Swart *et al.* 2017). Prior to sequencing, single stranded DNA was removed from RT-PCR reactions as follows: 5U exonuclease I (Thermo, Vilnius, Lithuania) and 1U FastAP® (Thermo, Vilnius, Lithuania) was added to amplification products and reaction was carried out at 37 (degrees) and 85 (degrees), for 15 min each, respectively.

Whole genome sequences of MaYMV isolates available on GenBank and the one assembled in this study were aligned using ClustalW and a phylogenetic tree was constructed in Mega7 (Kumar *et al.* 2016) using the Maximum Likelihood method (bootstrapped 1000 times). Maize yellow dwarf virus-RMV (KC921392.1) and Barley virus G isolate Gimje (KT962089.1) were used as outgroups.

Results

Both RNA-seq datasets (BR1A-BR3A and BR1B-BR3B) from upper and lower leaves of the maize B73 plants contained reads that (i) did not map to the maize or fungal genomes, and (ii) assembled into RNA molecules that coded for proteins with BLASTp matches to MaYMV proteins with more than 96% identity and E values below $3.1E^{-43}$ (Table B.1). For example, the BR1B-BR3B sample was predicted to contain MaYMV virus sequences that corresponded to the P1-P2 fusion protein (RNA dependent RNA polymerase), the P3 coat protein, and the P3-P5 readthrough protein (Table B.1).

Reads from all six RNA-seq datasets from three maize B73 plants in the same field trial were mapped separately to one of the first reported reference genomes of MaYMV from China (Yunnan 11; KU248489)(Chen *et al.* 2016). Coverage of a large proportion of the MaYMV genome was found in lower (older) leaves of all three plants (samples BR1B, BR2B, BR3B), ranging from 3968 – 4972 nucleotides at greater than 5× coverage of the 5642 nucleotide reference genome (Table B.2). Furthermore, reads matching MaYMV were also found in upper (younger) leaves of all three plants, with sample BR1A covering 5208 nucleotides (92%) of the MaYMV reference genome at 5× coverage or more.

Table B.1 MaYMV proteins identified from unmapped de novo assembled RNA in maize B73 plant 1 grown in field in KwaZulu-Natal, South Africa.

RNA source	BLASTP query ^c	Description ^d	Subject ^e	ID ^f	Alignment Length	Mis-matches	Gap Openings	Query Start	Query End	Subject Start	Subject End	E-value
Lower leaves (BR1B-3B) ^a	TRINITY_DN4048_c0_g1_i1 m.6699	P1-P2 fusion protein (RNA dependent RNA polymerase) [Maize yellow mosaic virus]	ref YP_009249572.1	100	141	0	0	1	141	659	799	2E-91
	TRINITY_DN412_c0_g1_i1 m.6212	P3 (Coat Protein) [Maize yellow mosaic virus]	gb ANI26433.1	100	101	0	0	1	101	97	197	2E-65
	TRINITY_DN622_c0_g1_i1 m.15051	P1 [Maize yellow mosaic virus]	gb ANI26422.1	97, 6	290	7	0	1	290	333	622	0
	TRINITY_DN736_c0_g1_i1 m.8767	P3-P5 readthrough protein [Maize yellow mosaic virus]	gb ANI26457.1	96, 5	284	10	0	1	284	381	664	0
Upper leaves (BR1A-3A) ^b	TRINITY_DN598_c0_g1_i1 m.35	P5 [Maize yellow dwarf virus-RMV2]	gb ANI26454.1	96	74	3	0	1	74	99	172	3E-43

^a RNA was extracted from a pool of two lower leaves of 3 maize plants in the inbred line B73 with well-developed lesions of GLS disease (8% leaf surface area), and subjected to RNA sequencing (the same material and methods in Methods S1 file of Christie et al 2017). Illumina RNA-seq reads were mapped to both the maize B73 (ver 4) genome and the C. zeina genome. Unmapped reads were de novo assembled (Trinity software), proteins predicted (Transdecoder software) and BLASTp analysis against NCBI nr database was conducted.

^b RNA was extracted from a pool of two upper leaves of 3 maize plants in the inbred line B73 with few lesions of GLS disease (0.2% leaf surface area). The source and analysis was the same as (a).

^c Protein predicted from denovo assembled transcript from transcript set that did not map to maize or fungal genomes.

^d Description of BLASTp match in NCBI nr database

^e GI number of BLASTp match in NCBI nr database

^f Percent identity of BLASTp alignment

Table B.2. Maize RNA-seq datasets screened for the presence of MaYMV.

Maize line	Sample name ¹	Year of sampling	Leaf sample	Total Illumina read pairs in dataset ²	Read pairs aligned ³	Bases covered ⁴
B73	<i>BR1A</i> [#]	2013	Upper leaf, low GLS infection, plant 1	14 258 978	516	5208
B73	<i>BR1B</i> [#]	2013	Lower leaf, high GLS infection, plant 1	14 258 728	318	4972
B73	<i>BR2A</i> [#]	2013	Upper leaf, low GLS infection, plant 2	14 258 209	66	527
B73	<i>BR2B</i> [#]	2013	Lower leaf, high GLS infection, plant 2	14 258 325	278	4112
B73	<i>BR3A</i> [#]	2013	Upper leaf, low GLS infection, plant 3	14 258 813	46	74
B73	<i>BR3B</i> [#]	2013	Lower leaf, high GLS infection, plant 3	12 962 476	248	3968
SC Malawi	<i>C1</i> [*]	2012	low GLS infection, fungicide spray, rep 1	12 934 797	562	4994
SC Malawi	<i>C2</i> [*]	2012	low GLS infection, fungicide spray, rep 2	12 985 168	108	1744
SC Malawi	<i>C3</i> [*]	2012	low GLS infection, fungicide spray, rep 3	12 816 669	82	1232
SC Malawi	<i>I1</i> [*]	2012	GLS infection, no fungicide spray, rep 1	13 593 497	68	418
SC Malawi	<i>I2</i> [*]	2012	GLS infection, no fungicide spray, rep 2	13 594 242	192	2909
SC Malawi	<i>I3</i> [*]	2012	GLS infection, no fungicide spray, rep 3	13 664 116	331	4346
CML444	<i>GKV-4</i> [*]	2009	Pool from 6 plants exposed to GLS	21 267 484	25	0
SC Malawi	<i>GKV-5</i> [*]	2009	Pool from 6 plants exposed to GLS	32 501 665	56	28
R_RILs bulk	<i>BR1</i> [*]	2009	Pool of six R RILs, low GLS infection, rep 1	35 717 600	88	517
R_RILs bulk	<i>BR2</i> [*]	2009	Pool of six R RILs, low GLS infection, rep 2	39 830 558	251	2319
R_RILs bulk	<i>BR3</i> [*]	2009	Pool of six R RILs, low GLS infection, rep 3	37 444 703	117	651
S_RILs bulk	<i>BS1</i> [*]	2009	Pool of six R RILs, high GLS infection, rep 1	33 865 116	304	2670
S_RILs bulk	<i>BS2</i> [*]	2009	Pool of six R RILs, high GLS infection, rep 2	32 318 237	212	1733
S_RILs bulk	<i>BS3</i> [*]	2009	Pool of six R RILs, high GLS infection, rep 3	39 371 687	223	1934
RIL387	<i>R_1</i> [*]	2009	R RIL, low GLS infection, rep 1	31 199 633	0	0

Maize line	Sample name ¹	Year of sampling	Leaf sample	Total Illumina read pairs in dataset ²	Read pairs aligned ³	Bases covered ⁴
RIL387	<i>R_2*</i>	2009	R RIL, low GLS infection, rep 2	31 899 347	155	1211
RIL387	<i>R_3*</i>	2009	R RIL, low GLS infection, rep 3	29 652 427	0	0
RIL165	<i>S_1*</i>	2009	S RIL, high GLS infection, rep 1	35 384 885	40	155
RIL165	<i>S_2*</i>	2009	S RIL, high GLS infection, rep 2	31 151 129	152	1207
RIL165	<i>S_3*</i>	2009	S RIL, high GLS infection, rep 3	39 441 945	219	1937

¹ RNA-seq sample name. Collection sites, KwaZulu-Natal province: #Greytown. *Baynesfield.

² Total RNA-seq read pairs (after quality filtering) used as query against the MaYMV reference genome (KU248489).

³ Number of read pairs aligned to the MaYMV reference genome (KU248489).

⁴ Number of sequential nucleotide bases that match the MaYMV reference genome with at least 5× coverage.

The reads from maize B73 sample BR1A that mapped to the reference MaYMV genome were assembled *de novo* to produce a MaYMV genome of 5,542 bases as a representative South African isolate, named MaYMV RSA BR1A (Genbank accession # MG570476). All of the expected full-length proteins of a polerovirus with identities of 96-100% to proteins of MaYMV Yunnan 11 isolate (P0, P1-P2, P3, P4, P5) were predicted from the MaYMV RSA BR1A genome (Table B.3).

The complete MaYMV genome sequences from other countries and continents were aligned with the MaYMV RSA BR1A genome, which showed 98-99% similarity to MaYMV genomes from Africa, 96-97% similarity to MaYMV genomes from China, and 95% similarity to MaYMV genomes from South America, but only 72% similarity with Maize Yellow Dwarf Virus-RMV (Table B.4). Phylogenetic analysis based on complete viral genomes employing a maximum likelihood algorithm (with 1000 bootstraps) corroborated these findings with the South African MaYMV isolate most similar to a Kenyan isolate, and isolates from different continents forming different clades, and distinct from the outgroups Maize Yellow Dwarf Virus-RMV and Barley virus G isolate Gimje (Fig. B.2).

RNA-seq data from maize field trials held in 2009 and 2012 at Baynesfield Estate in KwaZulu-Natal province of South Africa were also assessed. This site is 93 km from the Greytown site where the 2013 sample was collected. The 2009 field trial was made up of two parental maize inbred lines CML444 and SC Malawi with recombinant inbred line (RIL) progeny from a cross between these parents (Berger *et al.* 2014). RNA-seq reads from 11 of the 14 samples from the 2009 field trial mapped to the reference genome of MaYMV, and seven of these samples had 1200 or more bases with greater than 5 X coverage, indicating widespread distribution of the virus in the field (Table B.2).

MaYMV was also detected in the RNA-seq data from six plants of the parental inbred line SC Malawi sampled in a 2012 field trial at Baynesfield. Interestingly, MaYMV coverage was greater in plants sampled at Baynesfield in 2012 compared to 2009, which may reflect a higher virus titre.

RT-PCR with primers diagnostic for MaYMV (Chen *et al.* 2016)(within the RNA dependent RNA polymerase and coat protein genes) was first tested at different annealing temperatures using RNA from a positive maize field sample. Specific amplification of the expected MaYMV RT-PCR product of 753 bp was obtained at all annealing temperatures from 51.9°C to 58°C, and therefore 58°C was used in subsequent experiments.

RT-PCR produced the diagnostic MaYMV product from stored RNA corresponding to maize B73 samples BR1A and BR1B (Fig. B.1), and Sanger sequencing of the BR1B sample confirmed that the expected MaYMV fragment had been amplified (Fig. B.4).

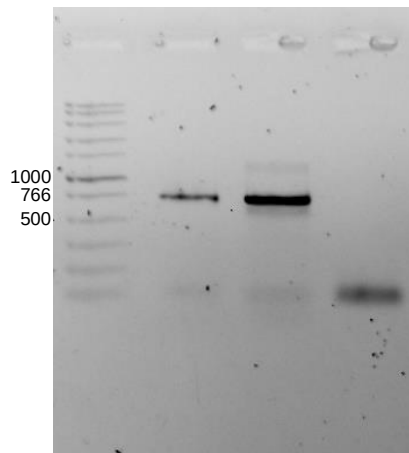


Figure B.1. RT-PCR diagnostic showing the presence of MaYMV in RNA stored from RNA-seq field trial. RNA from BR1A was used as the template in the first lane and BR1B in the second. The third shows the negative control.

Maize leaf samples exhibiting virus-like symptoms (mosaic, mottling, yellow streaks) were collected during the 2016/2017 season from the main maize producing provinces of South Africa (Fig. B.3), and 46 samples were screened for MaYMV by RT-PCR (Table B.3). A total of 39 samples were RT-PCR positive for MaYMV (Table B.3), representing isolates from all provinces sampled, namely Gauteng, KwaZulu-Natal, Limpopo, Mpumalanga, North West, and Northern Cape. RT-PCR of MaYMV negative samples targeting a maize actin gene confirmed that the cDNA was intact and that these were true negatives (Fig. B.2). Sanger sequencing of RT-PCR products from one representative sample from different provinces in the survey (17-4010: MP-highveld, 17-4101: NC, 17-4172: GT, 17-4203: MP-lowveld and 16-3379: NW) confirmed the expected MaYMV product (data not shown).

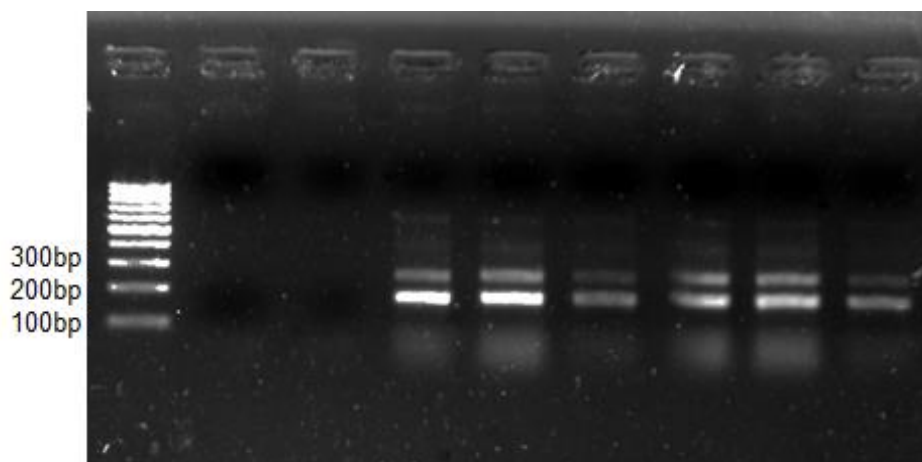


Figure B.2. Actin RT-PCR to confirm RNA and cDNA integrity for samples that were RT-PCR negative for MaYMV.

RT-PCR products amplified using the primer pair BananaActin-F and BananaActin-R were visualized by agarose gel electrophoresis, with an expected 169 bp cDNA actin product. Lane M, 100bp ladder; lane 1, RT no template control; lane 2, PCR no template control; lane 3, maize healthy control/RNA positive control; lane 4, 16-3308; lane 5, 16-3328; lane 6, 16-3224; lane 7, 16-3252; lane 8, 16-3256. The additional 277 bp product in lanes 3-8 is the actin gDNA product, since the primers flank an intron, indicating presence of some gDNA in the samples.



Figure B.3. A map showing the locations in South Africa where the positive RNA samples from the RNA-seq data and the positive survey samples had been collected.

5' TGCTCACATGTCTTCGAGAAGGAA 3' MaYMV-F primer

10 20 30 40 50 60 70 80 90 100

BR1A_assembled_sequence TGCTCACATGTCTTCGAGAAGGAAATCCTCGCACGCCCGGTTAACCAAAACAAAATGATTTACAAATTGGTTATGGTTACAAACCCGGCGAACGGTTCAT
KU248489
BR1B.MaYMV.F
BR1B.MaYMV.R

110 120 130 140 150 160 170 180 190 200

BR1A_assembled_sequence CCGGAGGTCTTGACGCGTTATCTTGACGCGTGTATGAGCGTGCTTCACGAGCTACGCCATGACCCCGAAACCGTGGAGCTTCTGTACAAGTGGCTGGTATC
KU248489
BR1B.MaYMV.F
BR1B.MaYMV.R

210 220 230 240 250 260 270 280 290 300

BR1A_assembled_sequence TCCAGTCCAGCAACAAAGGTTTGAAAAAGAGTTTCAAAGGTAGCCAGACACACGTGAGTTGCAAGTGCTGGAATCCTAGTCTCACACATAAGCAGCC
KU248489
BR1B.MaYMV.F
BR1B.MaYMV.R

310 320 330 340 350 360 370 380 390 400

BR1A_assembled_sequence ATAGATTGGAACCTCTTTTTCGCGGGTTCTTATAGGATCCTCGTTGCTGTCCCGTTAACCATCTTTGGCTTGTAACAAGATCTACCTATCTATCTCCTCGA
KU248489
BR1B.MaYMV.F
BR1B.MaYMV.R

410 420 430 440 450 460 470 480 490 500

BR1A_assembled_sequence ACGTGCGTTCAATTGTGAATGAATACGGGAGGTAGAAATGGACGCGAGAGCTAGGAACCGCCGACGCGCTCGCAATAATAACCGGGCCAGCCAGTGGTTG
KU248489
BR1B.MaYMV.F
BR1B.MaYMV.R

510 520 530 540 550 560 570 580 590 600

BR1A_assembled_sequence TTGTCGCGGCAAAATCCCGTCGAGGACGCCCTCGAAGACGAAGACGACCAAGTGGAAACACTGCAGGAAGACCTGGAGTCAGACGAGGCTCGCGGGAGAC
KU248489
BR1B.MaYMV.F
BR1B.MaYMV.R

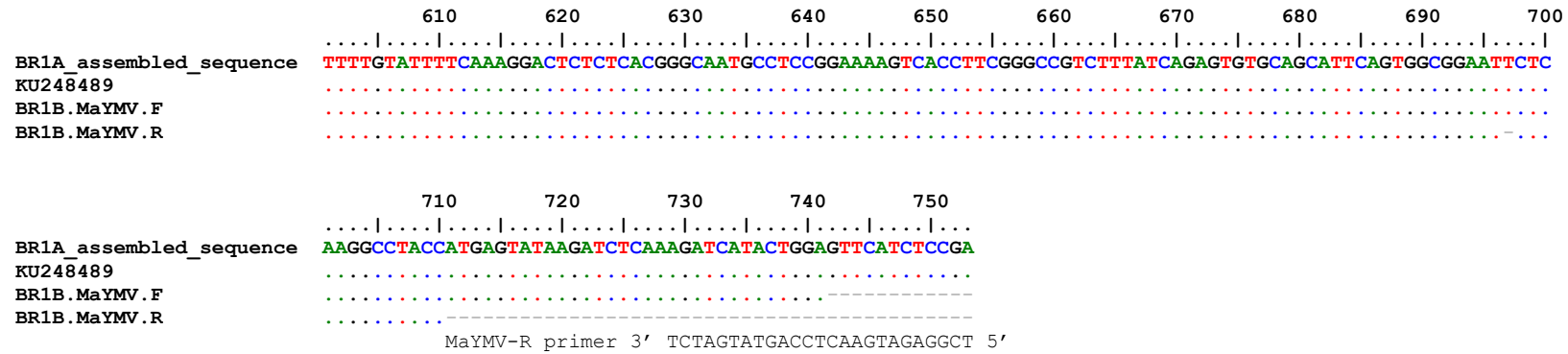


Figure B.4. Confirmation by Sanger sequencing that the expected MaYMV RT-PCR product was amplified from maize inbred B73 sample BR1B. Sequences derived from sequencing the 753bp RT-PCR product with the MaYMV-F or MaYMV-R primers were named BR1B.MaYMV.F or BR1B.MaYMV.R, respectively. These sequences were aligned to the corresponding sequence from the assembled MaYMV RSA BR1A genome (BR1A_assembled_sequence), and the reference sequence MaYMV Yunnan 11 from China (KU248489). The MaYMV-F and MaYMV-R primer sequences are reported in Chen et al. (2016). Consensus sites are shown as a dot and non-consensus sites are shown by an asterisk.

Table B.4 Comparison of predicted MaYMV proteins between MaYMV RSA BR1A from South Africa and MaYMV Yunnan 11 reference from China.

RSA BR1A vs Yunnan 11		RSA BR1A vs Yunnan 11		RSA BR1A	Yunnan 11
Name	Protein	% AA identity	% AA similarity (BLOSUM62)	Size (AA)	Size (AA)
Suppressor	P0	96%	97%	264	264
	P1	97%	98%	622	622
RdRP	P1-P2	98%	99%	1047	1047
CP	P3	100%	100%	197	197
MP	P4	98%	99%	192	192
RTD	P3-P5	97%	98%	664	664

Table B.4. The percent nucleotide similarity for complete genomes between MaYMV from South Africa and other worldwide isolates.

Region	Virus isolate and genome sequence accession #	%nucleotide similarity to MaYMV RSA BR1A (MG570476)	Sequence length (nt)
Africa	South Africa (BR1A)_MG570476		5542
	Kenya_MF974579.2	99%	5641
	Tanzania16/0076_MG664790.1	99%	5668
	Tanzania16/0092_MG664791.1	99%	5667
	Nigeria_KY684356.1	98%	5566
China	ChinaY7_KU291103.1	97%	5644
	ChinaY4_KU291100.1	97%	5641
	ChinaY6_KU291102.1	97%	5646
	ChinaY8_KU291104.1	97%	5642
	ChinaY11_KU248489.1	97%	5642
	China_MYDV-RMV2_KT992824.1	97%	5625
	ChinaY1_KU179221.1	97%	5634
	ChinaG1_KU291107.1	97%	5646
	ChinaY5_KU291101.1	97%	5635
	ChinaG2_KU291108.1	97%	5637
	ChinaBeijing_KY378940.1	97%	5642
	ChinaY2_KU248490.1	97%	5635
	ChinaY9_KU291105.1	96%	5624
	ChinaY10_KU291106.1	96%	5604
	ChinaY3_KU291099.1	96%	5563
America	Ecuador_KY052793	95%	5488
	Brazil_KY940544.1	95%	5637
Outgroups	Barley_virus_G_isolate_Gimje_KT962089.1	77%	5620
	Maize_yellow_dwarf_virus-RMV_KC921392.1	72%	5612

Maize leaf symptoms associated with MaYMV positive samples from the RT-PCR test from the country-wide survey included mosaic symptoms, and narrow or broad yellow streaking (Figure 5). These samples were collected based on scouting for virus-like symptoms in different fields across South Africa during the 2016/2017 survey. The particular maize genotypes were not known, however they are likely to be highly variable due to the diversity of farms, environments and altitudes (Table B.3).

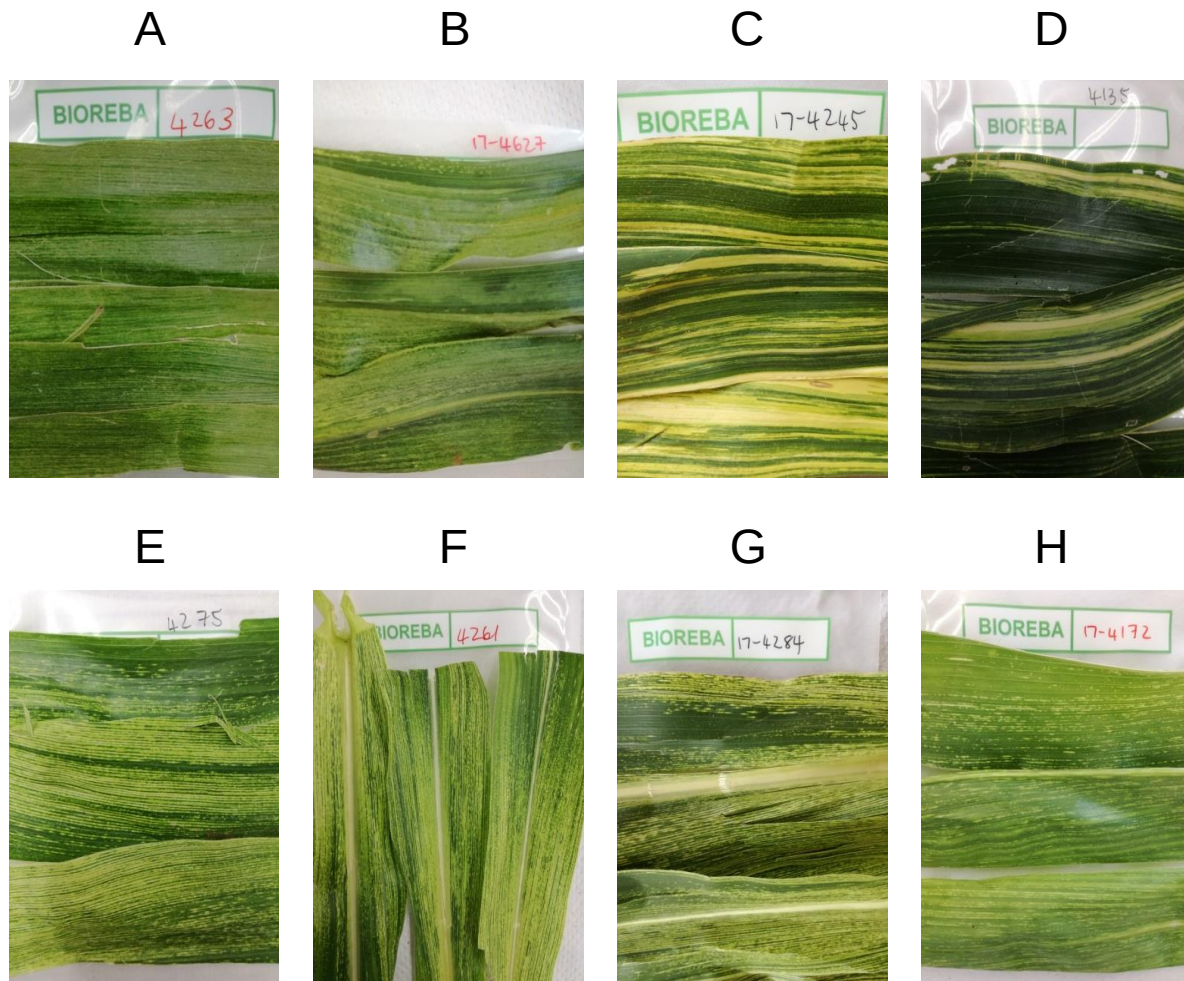


Figure B.5. Representative maize leaf symptoms of samples that were positive for MaYMV with the RT-PCR assay using MaYMV-F and MaYMV-R primers.

A, B: Mosaic symptoms; C-D: Yellow streaks; E-H: Narrow yellow streaks. A-H samples: 17-4263, 17-4267, 17-4245, 17-4135, 17-4275, 17-4261, 17-4284, 17-4172, respectively. Maize genotypes are not known.

Maize genotypes may influence virus symptoms as can be seen from the known maize genotypes that were MaYMV positive based on RNA-seq. Maize inbred line B73 sample BR1A from 2013 which had the highest number of RNA-seq reads mapping to the MaYMV reference genome and that was used to assemble the MaYMV RSA BR1A genome did not have apparent virus-like symptoms (Fig. B.2). Maize inbred line SC Malawi 2012 plants that were positive for MaYMV based on RNA-seq exhibited mosaic symptoms which were evident on close inspection (Fig. B.5). Six RILs that contributed leaf pieces to bulk samples

for RNA-seq (2009 sample BR1) that were MaYMV positive showed minor mosaic symptoms that was variable between genotypes (Fig. B.6). RNA-seq samples from leaves with high numbers of GLS lesions were not suitable for inspection for virus-like symptoms.

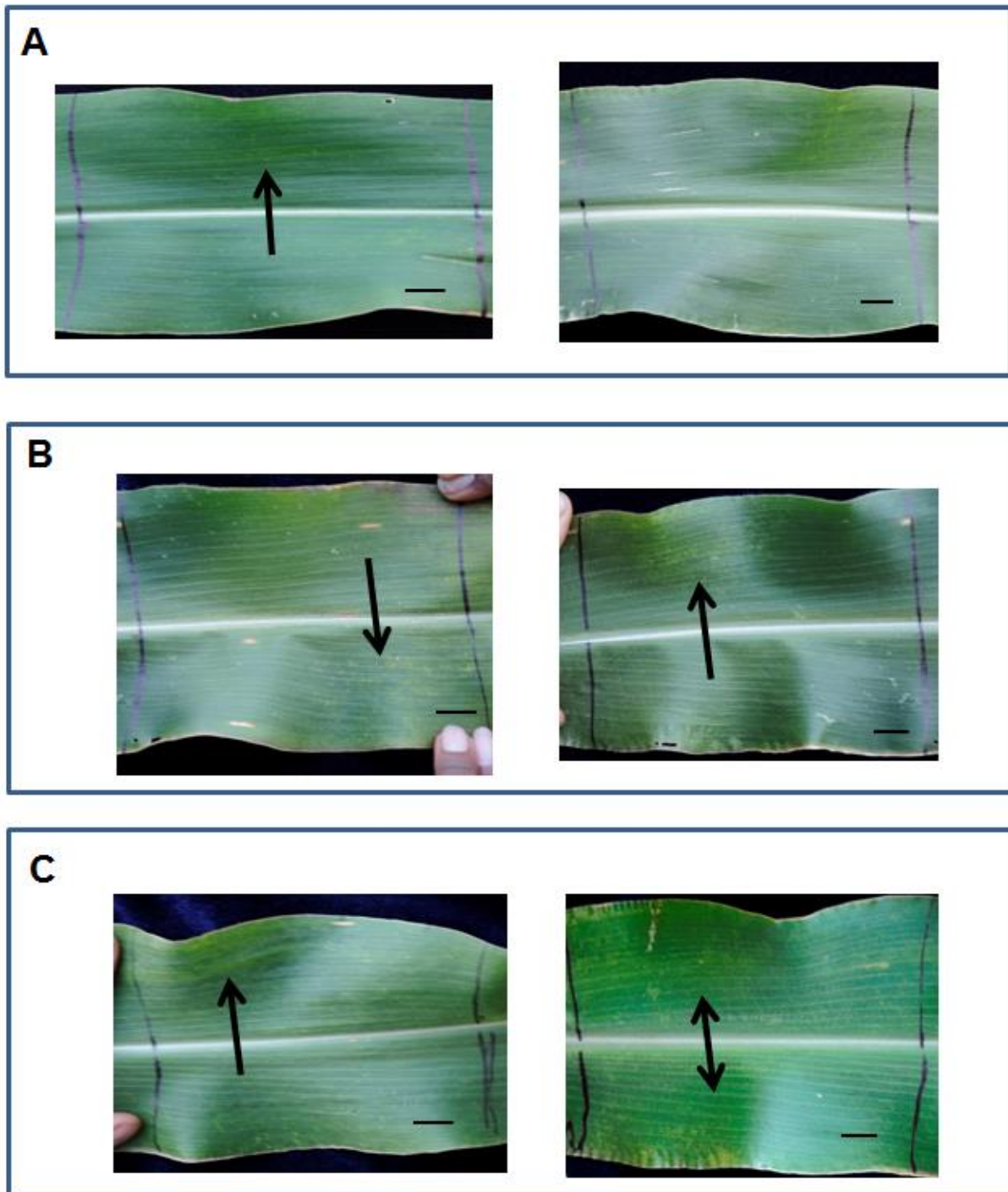


Figure B.6. Mild mosaic symptoms on leaf samples from maize inbred line SC Malawi shown to contain MaYMV in RNA-seq data.

Samples were collected from fungicide sprayed plots at Baynesfield in 2012. Panel A: Maize line SC Malawi pool of two leaf pieces combined to form sample C1; Panel B: SC Malawi C2; Panel C: SC Malawi C3. Arrows indicate mild mosaic symptoms. Scale bar = 1 cm

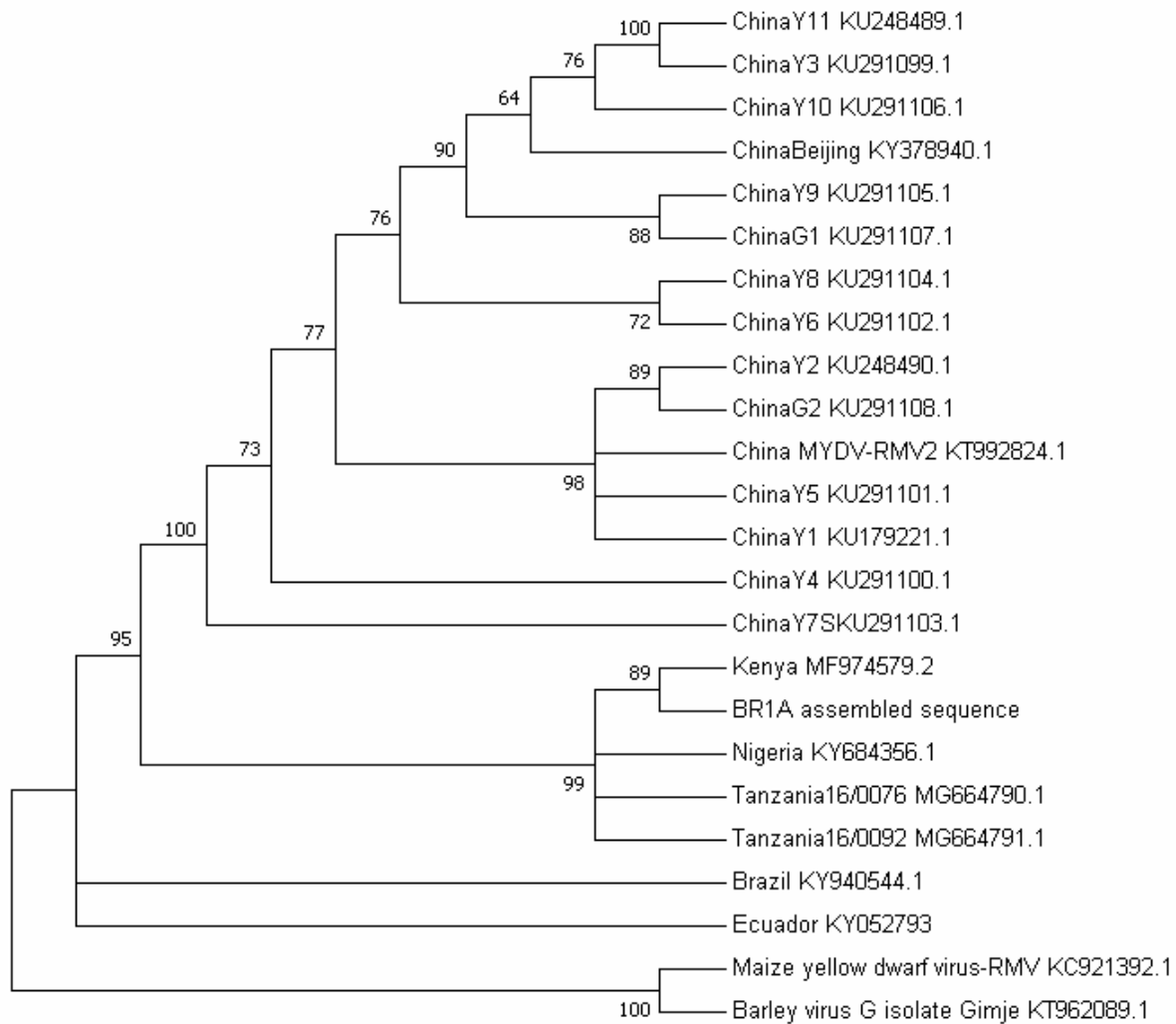


Figure B.7. Phylogenetic analysis of African, American, and Chinese Maize Yellow Mosaic Virus isolates. Barley virus G isolate Gimje and Maize yellow dwarf virus-RMV were outgroups. The evolutionary history was inferred by using the Maximum Likelihood method based on the Tamura-Nei model (Tamura and Nei 1993). The bootstrap consensus tree inferred from 1000 replicates (Felsenstein 1985) is taken to represent the evolutionary history of the taxa analyzed (Felsenstein 1985). Branches corresponding to partitions reproduced in less than 50% bootstrap replicates are collapsed. The analysis involved 24 whole genome sequences. All positions with less than 95% site coverage were eliminated. Evolutionary analyses were conducted in MEGA7 (Kumar et al. 2016).

Discussion and Conclusion

This study presents the first report of MaYMV in maize in South Africa, and demonstrates that it was present in KwaZulu-Natal province of the country from at least 2009, which represents the oldest report for Africa. Furthermore, the virus currently appears to be widespread throughout the main maize growing regions of the country, since maize leaf samples with virus-like symptoms (mosaic, yellow streaking) from all the six tested provinces Free State, Gauteng, KwaZulu-Natal, Limpopo, Mpumalanga, and North West were RT-PCR positive for MaYMV.

Next generation sequencing followed by a range of bioinformatics tools is increasingly being employed for diagnostics of known viruses and detection of unknown viruses in plant samples (Jones *et al.* 2017). The “discovery” phase of our study employed a variation of most previous approaches that map nucleic acid sequences (reads or assembled contigs) to databases for virus detection (Massawe *et al.* 2018; Visser *et al.* 2016). Our bioinformatics pipeline was designed to identify novel protein-coding genes in maize, with the following key steps: (i) recovery of RNA-seq reads that did not map to the host genome; and (ii) assembly of these reads into RNA contigs which were translated into predicted proteins and searched against the NCBI database. Matches to MaYMV proteins were obtained at high confidence. This approach was similar to modules of the pipeline “VirFind” developed by (Ho and Tzanetakis 2014), which the authors demonstrated by identifying known and novel viruses in more than 30 plant species.

Subsequent to the discovery phase, we further exploited maize RNA-seq data from different South African field trials collected in different seasons (2009, 2012, 2013) to detect MaYMV by mapping the quality-filtered RNA-seq reads to a reference MaYMV genome. The sample that had the highest depth and coverage was then used to assemble the MaYMV RSA BR1A genome sequence. Annotation of the genome confirmed that this virus was MaYMV since it encoded the expected full length proteins (P0 silencing suppressor; P1-P2 RNA dependent RNA polymerase, P3 coat protein, P4 movement protein, P3-P5 readthrough protein) with 96-100% amino acid sequence identity to the MaYMV reference isolate (KU248489). Similar approaches were followed to obtain full-length sequences of MaYMV from Kenya (Wamaita *et al.* 2018), Tanzania (Read *et al.* 2018), Ethiopia (Guadie *et al.* 2017), Brazil (Gonçalves *et al.* 2017), and Ecuador (Bernreiter *et al.* 2017).

Naming of MaYMV requires standardization since isolates that appear to be the same species of polerovirus (based on whole genome phylogenetic analysis and $\geq 95\%$ identity; Fig. B.7) have been named as either MaYMV (Chen *et al.* 2016) or MYDV-RMV2 (Wang *et al.* 2016) or MYDV-RMV (Wamaita *et al.* 2018) or MYDV-like polerovirus (Massawe *et al.* 2018). The name MYDV-RMV2 is based on sequence similarity to a virus isolated from wheat in Montana, USA, but also known to infect maize, previously known as Barley Yellow Dwarf Virus-RMV and renamed as MYDV-RMV (Krueger *et al.* 2013). However, since MaYMV isolates such as those reported in this paper from South Africa show less than 80% sequence similarity to MYDV-RMV Montana, we suggest that the name “MYDV-RMV” is not appropriate. For the purposes of this study we have used the name MaYMV for the South African isolates.

MaYMV has now been reported from Asia, Africa and South America (Chen *et al.* 2016; Yahaya *et al.* 2017; Gonçalves *et al.* 2017). Phylogenetic analysis of available MaYMV genome sequences including MaYMV RSA BR1A sheds some light on the possible movement and resultant distribution of the virus. MaYMV isolates for which whole genome sequences are available appear to group in continent-specific clades (Fig. B.7). For example, MaYMV RSA BR1A from South Africa groups with Nigerian, Kenyan and Tanzanian isolates (98-99% similarity), distinct from Chinese isolates (96-97% similarity), and South American isolates (95% similarity)(Table B.4). This indicates that MaYMV has been isolated for a sufficient period of time on each continent to develop continent-specific haplotypes, indicating that the virus may have present earlier than the first collection dates (2009 in Africa; this study; 2013 in China and Brazil; (Chen *et al.* 2016; Wang *et al.* 2016; Gonçalves *et al.* 2017)).

Several reasons may explain why MaYMV has only recently been identified. First, symptoms associated with MaYMV appear to vary greatly by maize genotype and range from mosaics to yellow streaking (Palanga *et al.* 2017; Wamaitha *et al.* 2018)(this study). Second, MaYMV has been found together with several other viruses as illustrated in a metagenomics study aimed at characterizing viruses associated with maize lethal necrosis disease (MLND)(Wamaitha *et al.* 2018). In cases of mixed virus infections, symptoms caused by MaYMV may have not been distinguished from symptoms caused by other viruses. Third, large scale screening with ELISA using antibodies to poleroviruses (such as BYDV-RMV; (Wang *et al.* 2016)) or RT-PCR with degenerate universal polerovirus primers (Palanga *et al.* 2017) may not have been carried out in many of these maize production regions. Finally, the advent of next generation sequencing has afforded a more comprehensive analysis of virus diversity in maize samples. The recent epidemic of MLND in Africa has precipitated several studies where MaYMV has been found by mining data from MLND surveys (Wamaitha *et al.* 2018; Massawe *et al.* 2018; Read *et al.* 2018), this study).

The results of our RT-PCR survey of 2016/2017 maize samples from six different maize producing provinces in South Africa indicates that MaYMV is widespread in the country in many different agro-ecological zones, indicating that it may have been introduced sometime prior to our first identification in RNA-seq data from a 2009 field trial. MaYMV was also shown to be widely distributed in Ethiopia (Guadie *et al.* 2017), Kenya (Massawe *et al.* 2018; Wamaitha *et al.* 2018), and Tanzania (Read *et al.* 2018). Despite the widespread distribution of MaYMV in South Africa, it does not seem to be major threat to maize production locally, based on two observations. First, the survey was conducted by manual scouting in maize fields for virus-like symptoms and only a few samples were found

scattered throughout individual fields, and there were no cases of large-scale symptoms (data not shown). Second, the symptoms of maize leaves found to contain MaYMV were variable from no visible symptoms in maize inbred line B73, to mild mosaic in inbred SC Malawi to a range of mosaic and yellow streaking in diverse genotypes across the country (Figs. 6.5, 6.6 and 6.7). This indicates that some maize genotypes including commercial hybrids grown in South Africa may harbour tolerance to MaYMV (defined as the ability to withstand viral load without an effect on yield; (Van der Linden *et al.* 2013)). Our study did not reveal any relationship between MaYMV presence and the fungal pathogen *Cercospora zeina* in maize, and thus these samples merely served as a serendipitous source of the virus.

The causal routes of introduction and spread of MaYMV into maize producing countries such as South Africa are not known, however three possible routes could be considered. First, if it is shown to be a seed-borne virus then it could have spread as a result of the worldwide distribution of maize seed (Ranum *et al.* 2014). If not seed-borne, it could also have reached Africa within leaves of whole plants that were introduced into West Africa from the America's as part of one of the initial introduction routes of maize into Africa (McCann, 2005). Second, poleroviruses are aphid transmitted, for example MYDV-RMV (Montana) was shown to be transmitted highly efficiently by the corn-leaf aphid *Rhopalosiphum maidis*, and by two other aphid species at lower efficiency (Krueger *et al.* 2013). *R. maidis* and other aphids are widespread in maize production areas throughout Africa (Hatting *et al.* 2000; Infonet-Biovision). Third, MaYMV has been reported from alternative hosts, such as sugarcane and itchgrass in Nigeria (Yahaya *et al.* 2017), and sorghum in Kenya (Wamaitha *et al.* 2018) and South Korea (Lim *et al.* 2017). Sugarcane and sorghum are also known to be hosts of aphid species (Singh *et al.* 2004) and therefore could serve as reservoirs of the virus. The host range of the sugarcane aphid, *Melanaphis sacchari*, includes other commercial Poaceae such as maize and sorghum (Singh *et al.* 2004), which makes it another potential vector.

This study has illustrated the power of next generation sequencing with libraries derived from total RNA for the identification of RNA viruses from crop field samples. We established that MaYMV is widespread in South Africa in addition to its presence in West Africa (Palanga *et al.* 2017; Yahaya *et al.* 2017), and widespread distribution in East Africa (Massawe *et al.* 2018; Wamaitha *et al.* 2018; Guadie *et al.* 2017; Read *et al.* 2018). Considering the importance of maize in other southern Africa countries such as Zambia, Zimbabwe, Botswana and Mozambique the monitoring for MaYMV in these countries should be an immediate priority.

References

- Afzal AJ, Wood AJ, Lightfoot DA (2008) Plant receptor-like serine threonine kinases: roles in signaling and plant defense. *Mol Plant-Microbe Interact* 21 (5):507-517
- Andorf CM, Cannon EK, Portwood JL, Gardiner JM, Harper LC, Schaeffer ML, Braun BL, Campbell DA, Vinnakota AG, Sribalusu VV (2016) MaizeGDB update: new tools, data and interface for the maize model organism database. *Nucleic Acids Res* 44 (D1):D1195-D1201
- Andrews S (2016) FastQC: A quality control application for FastQ data. <http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc>.
- Asea G, Bigirwa G, Adipala E, Oweru SAP, Pratt RC, Lipps PE (2002) Effect of *Cercospora zeae-maydis* infested maize residue on progress and spread of grey leaf spot of maize in central Uganda. *Ann Appl Biol* 140 (2):177-185
- Asea G, Vivek BS, Bigirwa G, Lipps PE, Pratt RC (2009) Validation of consensus quantitative trait loci associated with resistance to multiple foliar pathogens of maize. *Phytopathology* 99 (5):540-547
- Balasubramanian V, Vashisht D, Cletus J, Sakthivel N (2012) Plant β -1, 3-glucanases: their biological functions and transgenic expression against phytopathogenic fungi. *Biotechnol Lett* 34 (11):1983-1990
- Benson JM, Poland JA, Benson BM, Stromberg EL, Nelson RJ (2015) Resistance to gray leaf spot of maize: genetic architecture and mechanisms elucidated through nested association mapping and near-isogenic line analysis. *PLoS Genet* 11 (3):e1005045
- Berger D, Carstens M, Korsman J, Middleton F, Kloppers F, Tongoona P, Myburg A (2014) Mapping QTL conferring resistance in maize to gray leaf spot disease caused by *Cercospora zeina*. *BMC Genet* 15 (1):60
- Bernreiter A, Garcia Teijeiro R, Jarrin D, Garrido P, Ramos L (2017) First report of Maize yellow mosaic virus infecting maize in Ecuador. *New Disease Reports* 31:1
- Bolger AM, Lohse M, Usadel B (2014a) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*:btu170
- Bolger AM, Lohse M, Usadel B (2014b) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30 (15):2114-2120
- Chen S, Jiang G, Wu J, Liu Y, Qian Y, Zhou X (2016) Characterization of a Novel Polerovirus Infecting Maize in China. *Viruses* 8 (5):120
- Chen X, Shang J, Chen D, Lei C, Zou Y, Zhai W, Liu G, Xu J, Ling Z, Cao G, Ma B, Wang Y, Zhao X, Li S, Zhu L (2006) A B-lectin receptor kinase gene conferring rice blast resistance. *The Plant Journal* 46 (5):794-804
- Cheng YT, Germain H, Wiermer M, Bi D, Xu F, Garcia AV, Wirthmueller L, Despres C, Parker JE, Zhang Y, Li X (2009) Nuclear pore complex component MOS7/Nup88 is required for innate immunity and nuclear accumulation of defense regulators in *Arabidopsis*. *Plant Cell* 21 (8):2503-2516
- Chettoor AM, Givan SA, Cole RA, Coker CT, Unger-Wallace E, Vejlpkova Z, Vollbrecht E, Fowler JE, Evans MM (2014) Discovery of novel transcripts and gametophytic functions via RNA-seq analysis of maize gametophytic transcriptomes. *Genome biology* 15 (7):414
- Chionh YH, Ho C-H, Pruksakorn D, Babu IR, Ng CS, Hia F, McBee ME, Su D, Pang YLJ, Gu C (2013) A multidimensional platform for the purification of non-coding RNA species. *Nucleic Acids Res*:gkt668
- Christie N, Myburg AA, Joubert F, Murray SL, Carstens M, Lin YC, Meyer J, Crampton BG, Christensen SA, Ntuli JF (2017) Systems genetics reveals a transcriptional network associated with susceptibility in the maize–grey leaf spot pathosystem. *The Plant Journal* 89 (4):746-763
- Compeau PE, Pevzner PA, Tesler G (2011) How to apply de Bruijn graphs to genome assembly. *Nat Biotechnol* 29 (11):987-991
- Condon BJ, Leng Y, Wu D, Bushley KE, Ohm RA, Otiilar R, Martin J, Schackwitz W, Grimwood J, MohdZainudin N, Xue C, Wang R, Manning VA, Dhillon B, Tu ZJ,

- Steffenson BJ, Salamov A, Sun H, Lowry S, LaButti K, Han J, Copeland A, Lindquist E, Barry K, Schmutz J, Baker SE, Ciuffetti LM, Grigoriev IV, Zhong S, Turgeon BG (2013) Comparative genome structure, secondary metabolite, and effector coding capacity across *Cochliobolus* pathogens. *PLoS Genet* 9 (1):e1003233
- Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, Szczesniak MW, Gaffney DJ, Elo LL, Zhang X (2016) A survey of best practices for RNA-seq data analysis. *Genome biology* 17 (1):13
- Coolman R, Hoyt G (1993) The effects of reduced tillage on the soil environment. *HortTechnology* 3 (2):143-145
- Dao P, Numanagić I, Lin Y-Y, Hach F, Karakoc E, Donmez N, Collins C, Eichler EE, Sahinalp SC (2013) Optimal resolution of ambiguous RNA-Seq multimappings in the presence of novel isoforms. *Bioinformatics*:btt591
- De Coninck B, Cammue BP, Thevissen K (2013) Modes of antifungal action and in planta functions of plant defensins and defensin-like peptides. *Fungal Biology Reviews* 26 (4):109-120
- Denisov V, Strong W, Walder M, Gingrich J, Wintz H (2008) Development and validation of RQI: an RNA quality indicator for the Experion automated electrophoresis system. *Bio-Rad Bulletin* 5761
- Devisetty UK, Covington MF, Tat AV, Lekkala S, Maloof JN (2014) Polymorphism identification and improved genome annotation of *Brassica rapa* through deep RNA sequencing. *G3: Genes| Genomes| Genetics* 4 (11):2065-2078
- DeYoung BJ, Innes RW (2006) Plant NBS-LRR proteins in pathogen sensing and host defense. *Nat Immunol* 7 (12):1243-1249
- Felsenstein J (1985) Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* 39 (4):783-791
- Gaines TA, Lorentz L, Figge A, Herrmann J, Maiwald F, Ott MC, Han H, Busi R, Yu Q, Powles SB (2014) RNA - Seq transcriptome analysis to identify genes involved in metabolism - based diclofop resistance in *Lolium rigidum*. *The Plant Journal* 78 (5):865-876
- Garg R, Patel RK, Tyagi AK, Jain M (2011) De novo assembly of chickpea transcriptome using short reads for gene discovery and marker identification. *DNA Res* 18 (1):53-63
- Gerstein MB, Bruce C, Rozowsky JS, Zheng D, Du J, Korbel JO, Emanuelsson O, Zhang ZD, Weissman S, Snyder M (2007) What is a gene, post-ENCODE? History and updated definition. *Genome Res* 17 (6):669-681
- Gonçalves MC, Godinho M, Alves-Freitas DMT, Varsani A, Ribeiro SG (2017) First Report of Maize yellow mosaic virus Infecting Maize in Brazil. *Plant Dis* 101 (12):2156
- Gordon A, Hannon G (2010) Fastx-toolkit. FASTQ/A short-reads pre-processing tools. Unpublished http://hannonlabcsghedu/fastx_toolkit
- Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* 29 (7):644-652
- Guadie D, Abraham A, Tesfaye K, Winter S, Menzel W, Knierim D (2017) First Report of Maize yellow mosaic virus Infecting Maize (*Zea mays*) in Ethiopia. *Plant Disease* 102 (5):1044
- Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature protocols* 8 (8):1494-1512
- Hansey CN, Vaillancourt B, Sekhon RS, De Leon N, Kaeppler SM, Buell CR (2012) Maize (*Zea mays* L.) genome diversity as revealed by RNA-sequencing. *PloS one* 7 (3):e33071
- Hatting J, Poprawski T, Miller R (2000) Prevalences of fungal pathogens and other natural enemies of cereal aphids (Homoptera: Aphididae) in wheat under dryland and irrigated conditions in South Africa. *BioControl* 45 (2):179-199

- Hillier LW, Marth GT, Quinlan AR, Dooling D, Fewell G, Barnett D, Fox P, Glasscock JI, Hickenbotham M, Huang W (2008) Whole-genome sequencing and variant discovery in *C. elegans*. *Nat Methods* 5 (2):183
- Ho T, Tzanetakis IE (2014) Development of a virus detection and discovery pipeline using next generation sequencing. *Virology* 471:54-60
- Hogenhout SA, Ammar E-D, Whitfield AE, Redinbaugh MG (2008) Insect vector interactions with persistently transmitted viruses. *Annu Rev Phytopathol* 46:327-359
- Huang Y, Hu Y, Liu J (2014) Piecing the puzzle together: a revisit to transcript reconstruction problem in RNA-seq. *BMC Bioinformatics* 15 (Suppl 9):S3
- Hughes TE, Langdale JA, Kelly S (2014) The impact of widespread regulatory neofunctionalization on homeolog gene evolution following whole genome duplication in maize. *Genome Res*:gr. 172684.172114
<https://www.infonet-biovision.org/PlantHealth/Pests/Aphids>. Accessed 25/07/2018
- Jiang H, Wong WH (2009) Statistical inferences for isoform expression in RNA-Seq. *Bioinformatics* 25 (8):1026-1032
- Jiao Y, Peluso P, Shi J, Liang T, Stitzer MC, Wang B, Campbell MS, Stein JC, Wei X, Chin C-S, Guill K, Regulski M, Kumari S, Olson A, Gent J, Schneider KL, Wolfgruber TK, May MR, Springer NM, Antoniou E, McCombie WR, Presting GG, McMullen M, Ross-Ibarra J, Dawe RK, Hastie A, Rank DR, Ware D (2017) Improved maize reference genome with single-molecule technologies. *Nature advance online publication*
- Jin M, Liu H, He C, Fu J, Xiao Y, Wang Y, Xie W, Wang G, Yan J (2016) Maize pan-transcriptome provides novel insights into genome complexity and quantitative trait variation. *Scientific reports* 6:18936
- Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics* 30 (9):1236-1240
- Jones S, Baizan-Edge A, MacFarlane S, Torrance L (2017) Viral diagnostics in plants using next generation sequencing: computational analysis in practice. *Frontiers in plant science* 8:1770
- Jurka J, Klonowski P, Dagman V, Pelton P (1996) CENSOR—a program for identification and elimination of repetitive elements from DNA sequences. *Computers & chemistry* 20 (1):119-121
- Kakumanu A, Ambavaram MMR, Klumas C, Krishnan A, Batlang U, Myers E, Grene R, Pereira A (2012) Effects of Drought on Gene Expression in Maize Reproductive and Leaf Meristem Tissue Revealed by RNA-Seq. *Plant Physiol* 160 (2):846-867
- Kamo K, Blowers A, McElroy D (2000) Effect of the cauliflower mosaic virus 35S, actin, and ubiquitin promoters on uidA expression from a bar-uidA fusion gene in transgenic *Gladiolus* plants. *In Vitro Cellular & Developmental Biology-Plant* 36 (1):13-20
- Kent WJ (2002) BLAT—the BLAST-like alignment tool. *Genome Res* 12 (4):656-664
- Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome biology* 14 (4):R36
- Kloppers R, Tweer S (2009) Northern corn leaf blight fact sheet. PANNAR Seed (Pty) Ltd [<https://www.plantwise.org/FullTextPDF/2011/20117800335pdf>] Accessed 21/08/2018
- Koski LB, Golding GB (2001) The closest BLAST hit is often not the nearest neighbor. *J Mol Evol* 52 (6):540-542
- Krueger EN, Beckett RJ, Gray SM, Miller WA (2013) The complete nucleotide sequence of the genome of Barley yellow dwarf virus-RMV reveals it to be a new Polerovirus distantly related to other yellow dwarf viruses. *Frontiers in Microbiology* 4 (205):1-11
- Kumar S, Stecher G, Tamura K (2016) MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. *Mol Biol Evol* 33 (7):1870-1874
- Lacerda A, Vasconcelos ÉAR, Pelegri PB, Grossi-de-Sa MF (2014) Antifungal defensins and their role in plant defense. *Frontiers in microbiology* 5:116

- Lai J, Li R, Xu X, Jin W, Xu M, Zhao H, Xiang Z, Song W, Ying K, Zhang M (2010) Genome-wide patterns of genetic variation among elite maize inbred lines. *Nat Genet* 42 (11):1027
- Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nat Methods* 9 (4):357-359
- Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* 10 (3):R25-R25
- Latterell FM, Rossi AE (1983) Gray leaf spot of corn: a disease on the move. *Plant Dis* 67 (8):842-847
- Law CW, Chen Y, Shi W, Smyth GK (2014) Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome biology* 15 (2):1
- Levin JZ, Yassour M, Adiconis X, Nusbaum C, Thompson DA, Friedman N, Gnirke A, Regev A (2010) Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat Methods* 7 (9):709-715
- Li P, Ponnala L, Gandotra N, Wang L, Si Y, Tausta SL, Kebrom TH, Provart N, Patel R, Myers CR (2010) The developmental dynamics of the maize leaf transcriptome. *Nat Genet* 42 (12):1060
- Li WV, Zhao A, Zhang S, Li JJ (2016) MSIQ: Joint Modeling of Multiple RNA-seq Samples for Accurate Isoform Quantification. *arXiv preprint arXiv:160305915*
- Lim S, Yoon Y, Jang YW, Bae DH, Kim BS, Maharjan R, Yi H, Bae S, Lee YH, Lee BC, Park CY, Lee SH, Moon JS (2017) First Report of Maize yellow mosaic virus Infecting *Panicum miliaceum* and *Sorghum bicolor* in South Korea. *Plant Disease* 102 (3):689
- Lu B, Zeng Z, Shi T (2013) Comparative study of de novo assembly and genome-guided assembly strategies for transcriptome reconstruction based on RNA-Seq. *Science China Life sciences* 56 (2):143-155
- Lu T, Lu G, Fan D, Zhu C, Li W, Zhao Q, Feng Q, Zhao Y, Guo Y, Li W, Huang X, Han B (2010) Function annotation of the rice transcriptome at single-nucleotide resolution by RNA-seq. *Genome Res* 20 (9):1238-1249
- Mallowa SO, Esker PD, Paul PA, Bradley CA, Chapara VR, Conley SP, Robertson AE (2015) Effect of maize hybrid and foliar fungicides on yield under low foliar disease severity conditions. *Phytopathology* 105 (8):1080-1089
- Martin L, Fei Z, Giovannoni J, Rose JKC (2013) Catalyzing plant science research with RNA-seq. *Frontiers in plant science* 4:66
- Massawe DP, Stewart LR, Kamatenesi J, Asimwe T, Redinbaugh MG (2018) Complete sequence and diversity of a maize-associated Polerovirus in East Africa. *Virus Genes* 54 (3):432-437
- Meisel B, Korsman J, Kloppers FJ, Berger DK (2009) *Cercospora zeina* is the causal agent of grey leaf spot disease of maize in southern Africa. *Eur J Plant Pathol* 124 (4):577-583
- Meyer J, Berger DK, Christensen SA, Murray SL (2017) RNA-Seq analysis of resistant and susceptible sub-tropical maize lines reveals a role for kauralexins in resistance to grey leaf spot disease, caused by *Cercospora zeina*. *BMC plant biology* 17 (1):197
- Mizrachi E, Hefer CA, Ranik M, Joubert F, Myburg AA (2010) De novo assembled expressed gene catalog of a fast-growing Eucalyptus tree produced by Illumina mRNA-Seq. *BMC Genomics* 11 (1):1
- Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* 320 (5881):1344-1349
- NCB_Resource_Coordinators (2016) Database resources of the national center for biotechnology information. *Nucleic Acids Res* 44 (Database issue):D7
- Ondov BD, Bergman NH, Phillippy AM (2011) Interactive metagenomic visualization in a Web browser. *BMC Bioinformatics* 12:385

- Palanga E, Longué R, Koala M, Néya J, Traoré O, Martin DP, Peterschmitt M, Filloux D, Roumagnac P (2017) First report of maize yellow mosaic virus infecting maize in Burkina Faso. *New Disease Reports* 35:26
- Paszkiwicz K, Studholme DJ (2010) De novo assembly of short sequence reads. *Briefings in Bioinformatics* 11 (5):457-472
- Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C (2017) Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods* 14 (4):417
- Pechanova O, Pechan T, Williams WP, Luthe DS (2011) Proteomic analysis of the maize rachis: potential roles of constitutive and induced proteins in resistance to *Aspergillus flavus* infection and aflatoxin accumulation. *Proteomics* 11 (1):114-127
- Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* 33 (3):290-295
- Petitot AS, Dereeper A, Agbessi M, Da Silva C, Guy J, Ardisson M, Fernandez D (2016) Dual RNA - seq reveals *Meloidogyne graminicola* transcriptome and candidate effectors during the interaction with rice plants. *Mol Plant Pathol* 17 (6):860-874
- Qutob D, Kemmerling B, Brunner F, Küfner I, Engelhardt S, Gust AA, Lubracki B, Seitz HU, Stahl D, Rauhut T, Glawischnig E, Schween G, Lacombe B, Watanabe N, Lam E, Schlichting R, Scheel D, Nau K, Dodt G, Hubert D, Gijzen M, Nürnberger T (2006) Phytotoxicity and Innate Immune Responses Induced by Nep1-Like Proteins. *The Plant Cell* 18 (12):3721-3744
- Ramirez - Gonzalez RH, Segovia V, Bird N, Fenwick P, Holdgate S, Berry S, Jack P, Caccamo M, Uauy C (2015) RNA - Seq bulked segregant analysis enables the identification of high - resolution genetic markers for breeding in hexaploid wheat. *Plant Biotechnol J* 13 (5):613-624
- Ranum P, Peña-Rosas JP, Garcia-Casal MN (2014) Global maize production, utilization, and consumption. *Ann N Y Acad Sci* 1312 (1):105-112
- Read DA, Featherstone J, Rees DJG, Thompson GD, Roberts R, Flett BC, Mashigaidze K, Berger DK, Welgemoed T, Pietersen G, Schulze SE, Kiula B, Kullaya A, Mbega E (2018) First report of maize yellow mosaic virus (MaYMV) on maize (*Zea mays*) in Tanzania. *J Plant Pathol*:1-1
- Regulski M, Lu Z, Kendall J, Donoghue MT, Reinders J, Llaca V, Deschamps S, Smith A, Levy D, McCombie WR (2013) The maize methylome influences mRNA splice sites and reveals widespread paramutation-like switches guided by small RNA. *Genome Res*:gr. 153510.153112
- Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26 (1):139-140
- Robinson MD, Oshlack A (2010) A scaling normalization method for differential expression analysis of RNA-seq data. *Genome biology* 11 (3):R25
- Sanabria NM, van Heerden H, Dubery IA (2012) Molecular characterisation and regulation of a *Nicotiana tabacum* S-domain receptor-like kinase gene induced during an early rapid response to lipopolysaccharides. *Gene* 501 (1):39-48
- Sánchez-Sevilla JF, Cruz-Rus E, Valpuesta V, Botella MA, Amaya I (2014) Deciphering gamma-decalactone biosynthesis in strawberry fruit using a combination of genetic mapping, RNA-Seq and eQTL analyses. *BMC Genomics* 15 (1):218
- Schliesky S, Gowik U, Weber APM, Braeutigam A (2012) RNA-seq assembly – Are we there yet? *Frontiers in Plant Science* 3
- Schnable JC, Springer NM, Freeling M (2011) Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. *Proceedings of the National Academy of Sciences* 108 (10):4069-4074
- Schroeder A, Mueller O, Stocker S, Salowsky R, Leiber M, Gassmann M, Lightfoot S, Menzel W, Granzow M, Ragg T (2006) The RIN: an RNA integrity number for assigning integrity values to RNA measurements. *BMC Mol Biol* 7 (1):3

- Seqc/Maqc-III_Consortium (2014) A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nat Biotechnol* 32 (9):903-914
- Shiferaw B, Prasanna BM, Hellin J, Bänziger M (2011) Crops that feed the world 6. Past successes and future challenges to the role played by maize in global food security. *Food Security* 3 (3):307-327
- Shu X, Livingston III DP, Woloshuk CP, Payne GA (2017) Comparative histological and transcriptional analysis of maize kernels infected with *Aspergillus flavus* and *Fusarium verticillioides*. *Frontiers in plant science* 8:2075
- Singh B, Padmaja P, Seetharama N (2004) Biology and management of the sugarcane aphid, *Melanaphis sacchari* (Zehntner)(Homoptera: Aphididae), in sorghum: a review. *Crop Protect* 23 (9):739-755
- Singh P, Zimmerli LZ (2013) Lectin receptor kinases in plant innate immunity. *Frontiers in plant science* 4:124
- Song XJ, Kuroha T, Ayano M, Furuta T, Nagai K, Komeda N, Segami S, Miura K, Ogawa D, Kamura T (2015) Rare allele of a previously unidentified histone H4 acetyltransferase enhances grain weight, yield, and plant biomass in rice. *Proceedings of the National Academy of Sciences* 112 (1):76-81
- Springer NM, Ying K, Fu Y, Ji T, Yeh C-T, Jia Y, Wu W, Richmond T, Kitzman J, Rosenbaum H (2009) Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. *PLoS Genet* 5 (11):e1000734
- Stanley Kim H, Yu Y, Snesrud EC, Moy LP, Linford LD, Haas BJ, Nierman WC, Quackenbush J (2005) Transcriptional divergence of the duplicated oxidative stress-responsive genes in the *Arabidopsis* genome. *The Plant Journal* 41 (2):212-220
- Swanson-Wagner RA, Eichten SR, Kumari S, Tiffin P, Stein JC, Ware D, Springer NM (2010) Pervasive gene content variation and copy number variation in maize and its undomesticated progenitor. *Genome Res*
- Swart V, Crampton BG, Ridenour JB, Bluhm BH, Olivier NA, Meyer JJM, Berger DK (2017) Complementation of CTB7 in the Maize Pathogen *Cercospora zeina* Overcomes the Lack of In Vitro Cercosporin Production. *Molecular plant-microbe interactions : MPMI* 30 (9):710-724
- Tamura K, Nei M (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Molecular biology and evolution* 10 (3):512-526
- Tarailo - Graovac M, Chen N (2009) Using RepeatMasker to identify repetitive elements in genomic sequences. *Current Protocols in Bioinformatics*:4.10. 11-14.10. 14
- Tefera T, Mugo S, Beyene Y, Karaya H, Tende R (2013) Grain yield, stem borer and disease resistance of new maize hybrids in Kenya. *African Journal of Biotechnology* 10 (23):4777-4783
- Tehon L, Daniels E (1925) Notes on the parasitic fungi of Illinois: II. *Mycologia* 17 (6):240-249
- Trapnell C, Pachter L, Salzberg SL (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 25
- Trapnell C, Salzberg SL (2009) How to map billions of short reads onto genomes. *Nat Biotechnol* 27 (5):455-457
- Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, Van Baren MJ, Salzberg SL, Wold BJ, Pachter L (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* 28 (5):511-515
- Treangen TJ, Salzberg SL (2012) Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews Genetics* 13 (1):36-46
- Van den Berg N, Crampton BG, Hein I, Birch PR, Berger DK (2004) High-throughput screening of suppression subtractive hybridization cDNA libraries using DNA microarray analysis. *Biotechniques* 37 (5):818-824

- Van der Linden L, Bredenkamp J, Naidoo S, Fouche-Weich J, Denby KJ, Genin S, Marco Y, Berger DK (2013) Gene-for-gene tolerance to bacterial wilt in *Arabidopsis*. *Molecular plant-microbe interactions* : MPMI 26:398-406
- van Dijk EL, Jaszczyszyn Y, Thermes C (2014) Library preparation methods for next-generation sequencing: tone down the bias. *Exp Cell Res* 322 (1):12-20
- Van Verk MC, Hickman R, Pieterse CM, Van Wees SC (2013) RNA-Seq: revelation of the messengers. *Trends Plant Sci* 18 (4):175-179
- Van Bel M, Diels T, Vancaester E, Kreft L, Botzki A, Van de Peer Y, Coppens F, Vandepoele K (2018) PLAZA 4.0: an integrative resource for functional, evolutionary and comparative plant genomics. *Nucleic Acids Res* 46 (D1):D1190-D1196
- Vidal EA, Moyano TC, Krouk G, Katari MS, Tanurdzic M, McCombie WR, Coruzzi GM, Gutiérrez RA (2013) Integrated RNA-seq and sRNA-seq analysis identifies novel nitrate-responsive genes in *Arabidopsis thaliana* roots. *BMC Genomics* 14 (1):701
- Vigers AJ, Wiedemann S, Roberts WK, Legrand M, Selitrennikoff CP, Fritig B (1992) Thaumatin-like pathogenesis-related proteins are antifungal. *Plant Science* 83 (2):155-161
- Vijay N, Poelstra JW, Künstner A, Wolf JB (2013) Challenges and strategies in transcriptome assembly and differential gene expression quantification. A comprehensive in silico assessment of RNA - seq experiments. *Mol Ecol* 22 (3):620-634
- Visser M, Burger JT, Maree HJ (2016) Targeted virus detection in next-generation sequencing data using an automated e-probe based approach. *Virology* 495:122-128
- Wallhead M (2012) Foliar fungicide effects on gray leaf spot and yield of hybrid corn as influenced by application timing, hybrid characteristics and production practices. The Ohio State University,
- Wamaitha MJ, Nigam D, Maina S, Stomeo F, Wangai A, Njuguna JN, Holton TA, Wanjala BW, Wamalwa M, Lucas T (2018) Metagenomic analysis of viruses associated with maize lethal necrosis in Kenya. *Virology journal* 15 (1):90
- Wang F, Zhou BG, Gao ZL, Xu DF (2016) A New Species of the Genus Polerovirus Causing Symptoms Similar to Maize yellow dwarf virus-RMV of Maize in China. *Plant Dis* 100 (7):1508-1508
- Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 10 (1):57-63
- Ward J, Nowell D, Laing M, Whitwell M (1997a) Susceptibility of maize hybrids to grey leaf spot under two tillage systems in South Africa. *South African Journal of Plant and Soil* 14 (4):169-176
- Ward JMJ, Laing MD, Cairns ALP (1997b) Management practices to reduce gray leaf spot of maize. *Crop Sci* 37:1257+
- Ward JMJ, Stromberg EL, Nowell DC, Nutter FW (1999) Gray leaf Spot: A Disease of Global Importance in Maize Production. *Plant Dis* 83 (10):884-895
- Wen W, Li D, Li X, Gao Y, Li W, Li H, Liu J, Liu H, Chen W, Luo J (2014) Metabolome-based genome-wide association study of maize kernel leads to novel biochemical insights. *Nature communications* 5:3438
- White EJ, Venter M, Hiten NF, Burger JT (2008) Modified Cetyltrimethylammonium bromide method improves robustness and versatility: The benchmark for plant RNA extraction. *Biotechnology Journal: Healthcare Nutrition Technology* 3 (11):1424-1428
- Wingfield BD, Berger DK, Steenkamp ET, Lim H-J, Duong TA, Bluhm BH, De Beer ZW, De Vos L, Fourie G, Naidoo K (2017) Draft genome of *Cercospora zeina*, *Fusarium pininemorale*, *Hawksworthiomyces lignivorus*, *Huntia decipiens* and *Ophiostoma ips*. *IMA fungus* 8 (2):385-396
- Xu L, Wang Y, Liu W, Wang J, Zhu X, Zhang K, Yu R, Wang R, Xie Y, Zhang W, Gong Y, Liu L (2015) De novo sequencing of root transcriptome reveals complex cadmium-responsive regulatory networks in radish (*Raphanus sativus* L.). *Plant Sci* 236:313-323

- Yahaya A, Al Rwahnih M, Dangora D, Gregg L, Alegbejo M, Lava Kumar P, Alabi O (2017) First Report of Maize yellow mosaic virus Infecting Sugarcane (*Saccharum* spp.) and Itch Grass (*Rottboellia cochinchinensis*) in Nigeria. *Plant Disease* 101 (7):1335
- Yang Q, He Y, Kabahuma M, Chaya T, Kelly A, Borrego E, Bian Y, El Kasmi F, Yang L, Teixeira P (2017) A gene encoding maize caffeoyl-CoA O-methyltransferase confers quantitative resistance to multiple pathogens. *Nature* 201:7
- Yang SS, Tu ZJ, Cheung F, Xu WW, Lamb JF, Jung H-JG, Vance CP, Gronwald JW (2011) Using RNA-Seq for gene identification, polymorphism detection and transcript profiling in two alfalfa genotypes with divergent cell wall composition in stems. *BMC Genomics* 12 (1):199
- Zenoni S, Ferrarini A, Giacomelli E, Xumerle L, Fasoli M, Malerba G, Bellin D, Pezzotti M, Delledonne M (2010) Characterization of Transcriptional Complexity during Berry Development in *Vitis vinifera* Using RNA-Seq. *Plant Physiol* 152 (4):1787-1795
- Zhang P, Wang Y, Zhang J, Maddock S, Snook M, Peterson T (2003) A maize QTL for silk maysin levels contains duplicated Myb-homologous genes which jointly regulate flavone biosynthesis. *Plant Mol Biol* 52 (1):1-15
- Zhang X, Yang Q, Rucker E, Thomason W, Balint-Kurti P (2017) Fine mapping of a quantitative resistance gene for gray leaf spot of maize (*Zea mays* L.) derived from teosinte (*Z. mays* ssp. *parviglumis*). *Theor Appl Genet* 130 (6):1285-1295
- Zhao Q-Y, Wang Y, Kong Y-M, Luo D, Li X, Hao P (2011) Optimizing de novo transcriptome assembly from short-read RNA-Seq data: a comparative study. *BMC Bioinformatics* 12 (14):1-12
- Zhao S, Zhang B, Zhang Y, Gordon W, Du S, Paradis T, Vincent M, von Schack D (2016) Bioinformatics for RNA - Seq Data Analysis. *BIOINFORMATICS-UPDATED FEATURES AND APPLICATIONS*:125
- Zhao W, He X, Hoadley KA, Parker JS, Hayes DN, Perou CM (2014) Comparison of RNA-Seq by poly (A) capture, ribosomal RNA depletion, and DNA microarray for expression profiling. *BMC Genomics* 15 (1):1
- Zheng Y, Li B, Zhang X, Lin Y, Liu Z (2010) Occurrence and damage of maize gray leaf spot and its GPS monitoring. *Plant Diseases and Pests* 1 (6):1-3, 11

Appendix C

File C.1 Quality trimming the reads

```
java -jar /usr/local/src/Trimmomatic-0.33/trimmomatic-0.33.jar PE -threads 4 combined_1.fq
combined_2.fq B10R1A_combined_forward_paired.fq B10R1A_combined_forward_unpaired.fq
B10R1A_combined_reverse_paired.fq B10R1A_combined_reverse_unpaired.fq
ILLUMINACLIP:/usr/local/src/Trimmomatic-0.33/adapters/TruSeq3-PE-2.fa:2:30:10 HEADCROP:15
SLIDINGWINDOW:5:28 MINLEN:50
```

```
java -jar /usr/local/src/Trimmomatic-0.33/trimmomatic-0.33.jar PE -threads 4 FCC327PACXX-
MALajpTAADRAAPEI-22_L1_1.fq FCC327PACXX-MALajpTAADRAAPEI-22_L1_2.fq
B10R1B_combined_forward_paired.fq B10R1B_combined_forward_unpaired.fq
B10R1B_combined_reverse_paired.fq B10R1B_combined_reverse_unpaired.fq
ILLUMINACLIP:/usr/local/src/Trimmomatic-0.33/adapters/TruSeq3-PE-2.fa:2:30:10 HEADCROP:15
SLIDINGWINDOW:5:28 MINLEN:50
```

```
java -jar /usr/local/src/Trimmomatic-0.33/trimmomatic-0.33.jar PE -threads 4 combined_1.fq
combined_2.fq B10R2A_combined_forward_paired.fq B10R2A_combined_forward_unpaired.fq
B10R2A_combined_reverse_paired.fq B10R2A_combined_reverse_unpaired.fq
ILLUMINACLIP:/usr/local/src/Trimmomatic-0.33/adapters/TruSeq3-PE-2.fa:2:30:10 HEADCROP:15
SLIDINGWINDOW:5:28 MINLEN:50
```

```
java -jar /usr/local/src/Trimmomatic-0.33/trimmomatic-0.33.jar PE -threads 4 FCC30ACACXX-
MALajpTAAHRAAPEI-33_L1_1.fq FCC30ACACXX-MALajpTAAHRAAPEI-33_L1_2.fq
B10R2B_combined_forward_paired.fq B10R2B_combined_forward_unpaired.fq
B10R2B_combined_reverse_paired.fq B10R2B_combined_reverse_unpaired.fq
ILLUMINACLIP:/usr/local/src/Trimmomatic-0.33/adapters/TruSeq3-PE-2.fa:2:30:10 HEADCROP:15
SLIDINGWINDOW:5:28 MINLEN:50
```

```
java -jar /usr/local/src/Trimmomatic-0.33/trimmomatic-0.33.jar PE -threads 4 FCC327PACXX-
MALajpTAAKRAAPEI-36_L3_1.fq FCC327PACXX-MALajpTAAKRAAPEI-36_L3_2.fq
B10R3A_combined_forward_paired.fq B10R3A_combined_forward_unpaired.fq
B10R3A_combined_reverse_paired.fq B10R3A_combined_reverse_unpaired.fq
ILLUMINACLIP:/usr/local/src/Trimmomatic-0.33/adapters/TruSeq3-PE-2.fa:2:30:10 HEADCROP:15
SLIDINGWINDOW:5:28 MINLEN:50
```

```
java -jar /usr/local/src/Trimmomatic-0.33/trimmomatic-0.33.jar PE -threads 4 FCC30ACACXX-
MALajpTAALRAAPEI-37_L6_1.fq FCC30ACACXX-MALajpTAALRAAPEI-37_L6_2.fq
B10R3B_combined_forward_paired.fq B10R3B_combined_forward_unpaired.fq
B10R3B_combined_reverse_paired.fq B10R3B_combined_reverse_unpaired.fq
ILLUMINACLIP:/usr/local/src/Trimmomatic-0.33/adapters/TruSeq3-PE-2.fa:2:30:10 HEADCROP:15
SLIDINGWINDOW:5:28 MINLEN:50
```

File C.2 Mapping to the *C. zeina* and B73 reference genomes

```
tophat -G ~/Mapping/Zea_mays.AGPv4.34.gff3 --transcriptome-index=transcriptome_data/known
B73_and_Czeina
```

```
nohup tophat -o ~/Mapping/B10R1A -p 4 --transcriptome-index=transcriptome_data/known
B73_and_Czeina ~/B10.R1.A/75bp/B10R1A_combined_forward_paired_1.fq
~/B10.R1.A/75bp/B10R1A_combined_reverse_paired_2.fq &
```

```
tophat -o ~/Mapping/B10R1B -p 4 --transcriptome-index=transcriptome_data/known
B73_and_Czeina ~/B10.R1.B/for_mapping/B10R1B_paired_1.fq
~/B10.R1.B/for_mapping/B10R1B_paired_2.fq
```

```
nohup tophat -o ~/Mapping/B10R2A --transcriptome-index=transcriptome_data/known
B73_and_Czeina ~/B10.R2.A/B10R2A_combined_paired_1.fq
~/B10.R2.A/B10R2A_combined_paired_2.fq &
```

```
tophat -o ~/Mapping/B10R2B -p 2 --transcriptome-index=transcriptome_data/known
B73_and_Czeina ~/B10.R2.B/B10R2B_paired_1.fq ~/B10.R2.B/B10R2B_paired_2.fq
```

```
nohup tophat -o ~/Mapping/B10R3A --transcriptome-index=transcriptome_data/known
B73_and_Czeina ~/B10.R3.A/B10R3A_paired_1 ~/B10.R3.A/B10R3A_paired_2.fq &
```

```
tophat -o ~/Mapping/B10R3B -p 4 --transcriptome-index=transcriptome_data/known
B73_and_Czeina ~/B10.R3.B/B10R3B_paired_1.fq ~/B10.R3.B/B10R3B_paired_2.fq
```

File C.3 *De novo* assembly of the unmapped reads and total reads

```
Trinity --output A_Unmapped2_trinity --SS_lib_type RF --seqType fa --max_memory 60G --CPU 24 --
KMER_SIZE 31 --left B10R1A_unmapped_1.fa,B10R2A_unmapped_1.fa,B10R3A_unmapped_1.fa --
right B10R1A_unmapped_2.fa,B10R3A_unmapped_2.fa,B10R3A_unmapped_2.fa
```

```
Trinity --output B_Unmapped2_trinity --SS_lib_type RF --seqType fa --max_memory 60G --CPU 24 --
KMER_SIZE 31 --samples_file Replicate_B_relationship.txt
```

```
Trinity --output /scratch/cercomaize/MasterAssembly_trinity --SS_lib_type RF --seqType fa --
max_mem\
```

```
ory 60G --CPU 12 --KMER_SIZE 31 --left
/scratch/cercomaize/FastaReads/B10R1A_paired_1.fasta,/scratc\
```

```
h/cercomaize/FastaReads/B10R1B_paired_1.fasta,/scratch/cercomaize/FastaReads/B10R2A_paired
_1.fasta,\
```

```
/scratch/cercomaize/FastaReads/B10R2B_paired_1.fasta,/scratch/cercomaize/FastaReads/B10R3A_
paired_1\
```

```
.fasta,/scratch/cercomaize/FastaReads/B10R3B_paired_1.fasta --right
/scratch/cercomaize/FastaReads/\
```

```
B10R1A_paired_2.fasta,/scratch/cercomaize/FastaReads/B10R1B_paired_2.fasta,/scratch/cercomaiz
e/Fast\
```

```
aReads/B10R2A_paired_2.fasta,/scratch/cercomaize/FastaReads/B10R2B_paired_2.fasta,/scratch/c
ercomai\
```

```
ze/FastaReads/B10R3A_paired_2.fasta,/scratch/cercomaize/FastaReads/B10R3B_paired_2.fasta
```

```
TransDecoder.LongOrfs -m 50 -t B_Unmapped_Trinity.fasta
```

```
TransDecoder.Predict -t B_Unmapped_Trinity.fasta.fasta
```

```
TransDecoder.LongOrfs -m 50 -t A_Unmapped_Trinity.fasta
```

```
TransDecoder.Predict -t A_Unmapped_Trinity.fasta.fasta
```

File C.4 Estimating the abundances of each transcript in the *C. zeina* challenged transcripts and the control transcripts

```
salmon index -t MasterAssembly_Trinity2.fasta -i transcripts_index --type quasi -k 31
```

```
salmon quant -i transcripts_index -p 8 --writeUnmappedNames -l SR -1  
/ngsprojects/cercomaize/data_archive/Reads/B10R1A/combined_1.fq -2  
/ngsprojects/cercomaize/data_archive/Reads/B10R1A/combined_2.fq -o A1_transcripts_quant
```

```
salmon quant -i transcripts_index -p 8 --writeUnmappedNames -l SR -1  
/ngsprojects/cercomaize/data_archive/Reads/B10R2A/for_mapping/B10R2A_combined_paired_1.f  
q -2  
/ngsprojects/cercomaize/data_archive/Reads/B10R2A/for_mapping/B10R2A_combined_paired_2.f  
q -o A2_transcripts_quant
```

```
salmon quant -i transcripts_index -p 8 --writeUnmappedNames -l SR -1  
/ngsprojects/cercomaize/data_archive/Reads/B10R3A/B10R3A_paired_1.fq -2  
/ngsprojects/cercomaize/data_archive/Reads/B10R3A/B10R3A_paired_2.fq -o A3_transcripts_quant
```

```
salmon quant -i transcripts_index -p 8 --writeUnmappedNames -l SR -1  
/ngsprojects/cercomaize/data_archive/Reads/B10R1B/for_mapping/B10R1B_paired_1.fq -2  
/ngsprojects/cercomaize/data_archive/Reads/B10R1B/for_mapping/B10R1B_paired_2.fq -o  
B1_transcripts_quant
```

```
salmon quant -i transcripts_index -p 8 --writeUnmappedNames -l SR -1  
/ngsprojects/cercomaize/data_archive/Reads/B10R2B/B10R2B_paired_1.fq -2  
/ngsprojects/cercomaize/data_archive/Reads/B10R2B/B10R2B_paired_2.fq -o B2_transcripts_quant
```

```
salmon quant -i transcripts_index -p 8 --writeUnmappedNames -l SR -1  
/ngsprojects/cercomaize/data_archive/Reads/B10R3B/B10R3B_paired_1.fq -2  
/ngsprojects/cercomaize/data_archive/Reads/B10R3B/B10R3B_paired_2.fq -o B3_transcripts_quant
```

```
/trinity/2.5.1/util/abundance_estimates_to_matrix.pl --est_method salmon --gene_trans_map none  
--name_sample_by_basedir A1_transcripts_quant/quant.sf A2_transcripts_quant/quant.sf  
A3_transcripts_quant/quant.sf B1_transcripts_quant/quant.sf B2_transcripts_quant/quant.sf  
B3_transcripts_quant/quant.sf
```

File C.5 Rscript to test for differential expression between the *C. zeina* challenged transcripts and the control transcripts

```
/trinity/2.5.1/util/support_scripts/run_TMM_scale_matrix.pl --matrix  
salmon.isoform.TPM.not_cross_norm > salmon.isoform.TMM.EXPR.matrixCMD: R --vanilla -q <  
__tmp_runTMM.R 1>&2
```

```
> library(edgeR)
```

```
Loading required package: limma
```

```
>
```

```
> rnaseqMatrix = read.table("salmon.isoform.TPM.not_cross_norm", header=T, row.names=1,
com=", check.names=F)

> rnaseqMatrix = as.matrix(rnaseqMatrix)

> rnaseqMatrix = round(rnaseqMatrix)

> exp_study = DGEList(counts=rnaseqMatrix, group=factor(colnames(rnaseqMatrix)))

> exp_study = calcNormFactors(exp_study)

> exp_study$samples$eff.lib.size = exp_study$samples$lib.size * exp_study$samples$norm.factors

> write.table(exp_study$samples, file="salmon.isoform.TPM.not_cross_norm.TMM_info.txt",
quote=F, sep="\t", row.names=F)
```

```
/software/shared/apps/x86_64/trinity/2.5.1/Analysis/DifferentialExpression/run_DE_analysis.pl --
matrix salmon.isoform.counts.matrix --method edgeR --samples_file
/scratch/cercomaize/bioreps.txt
```

Got 6 samples, and got: 7 data fields.

Header: A_rep1 A_rep2 A_rep3 B_rep1 B_rep2 B_rep3

Next: TRINITY_DN5772_c0_g1_i1 471.708961 382.966430 378.431031 344.922220
266.568160 350.965196

```
$VAR1 = {
  'B_rep2' => 5,
  'A_rep2' => 2,
  'A_rep1' => 1,
  'B_rep3' => 6,
  'B_rep1' => 4,
  'A_rep3' => 3
};
```

```
$VAR1 = {
  'A' => [
    'A_rep1',
    'A_rep2',
    'A_rep3'
  ],
  'B' => [
    'B_rep1',
```

```
'B_rep2',
'B_rep3'
]
};
```

Contrasts to perform are: \$VAR1 = [

```
[
'A',
'B'
]
];
```

CMD: R --vanilla -q < salmon.isoform.counts.matrix.A_vs_B.A.vs.B.EdgeR.Rscript

```
> library(edgeR)
```

Loading required package: limma

```
>
```

```
> data =
read.table("/scratch/cercomaize/AbundanceEstimations/Salmon/salmon.isoform.counts.matrix",
header=T, row.names=1, com="")
```

```
> col_ordering = c(1,2,3,4,5,6)
```

```
> rnaseqMatrix = data[,col_ordering]
```

```
> rnaseqMatrix = round(rnaseqMatrix)
```

```
> rnaseqMatrix = rnaseqMatrix[rowSums(cpm(rnaseqMatrix) > 1) >= 2,]
```

```
> conditions = factor(c(rep("A", 3), rep("B", 3)))
```

```
>
```

```
> exp_study = DGEList(counts=rnaseqMatrix, group=conditions)
```

```
> exp_study = calcNormFactors(exp_study)
```

```
> exp_study = estimateCommonDisp(exp_study)
```

```
> exp_study = estimateTagwiseDisp(exp_study)
```

```
> et = exactTest(exp_study, pair=c("A", "B"))
```

```
> tTags = topTags(et,n=NULL)
```

```
> result_table = tTags$table
```

```
> result_table = data.frame(sampleA="A", sampleB="B", result_table)
```

```
> result_table$logFC = -1 * result_table$logFC
```

```
> write.table(result_table, file='salmon.isoform.counts.matrix.A_vs_B.edgeR.DE_results', sep=' ',
quote=F, row.names=T)

> write.table(rnaseqMatrix, file='salmon.isoform.counts.matrix.A_vs_B.edgeR.count_matrix',
sep=' ', quote=F, row.names=T)

>
source("/shared/clissoft/apps/x86_64/trinity/2.5.1/Analysis/DifferentialExpression/R/rnaseq_plot_fu
ncs.R")

> pdf("salmon.isoform.counts.matrix.A_vs_B.edgeR.DE_results.MA_n_Volcano.pdf")

> plot_MA_and_Volcano(rownames(result_table), result_table$logCPM, result_table$logFC,
result_table$FDR)

> dev.off()

null device
```

1

File C.6 BLAT command and AWK script to split differentially expressed transcripts
 blat B73_Czeina_and_Eturcicum.fa ../MasterAssembly2_trinity/Trinity.fasta
 MasterAssembly_blat_script.psl -q=rna -t=dna

```
sed '1,5d' MasterAssembly_blat_script.psl | awk '$1/$11 >= 0.90 && $13-$12 >= 0.8*$11' | sort -nk
10 > MasterAssembly_blat_awk_filtered.tsv
```

```
sed '1,5d' MasterAssembly_blat_script.psl | awk '$14 ~ /czeina/' | cut -f10 | sort | uniq -c | sort -nk1
> MasterAssembly_blat_CzeinaTranscripts.txt
```

```
sed '1,5d' MasterAssembly_blat_script.psl | awk '$14 == 1 || $14 == 2 || $14 == 3 || $14 == 4 || $14
== 5 || $14 == 6 || $14 == 7 || $14 == 8 || $14 == 9 || $14 == 10 || $14 ~ /B73V4_/ || $14 ~ /Mt/ ||
$14 ~ /Pt/' | cut -f10 | sort | uniq -c | sort -nk1 > MasterAssembly_blat_B73.txt
```

```
sed '1,5d' MasterAssembly_blat_script.psl | awk '$14 ~ /scaffold/' | cut -f10 | sort | uniq -c | sort -
nk1 > MasterAssembly_blat_Eturcicum.txt
```

```
awk '$14 ~ /czeina/' MasterAssembly_blat_awk_filtered.tsv | cut -f10 | sort | uniq -c | sort -nk1 >
MasterAssembly_blat_Filtered_CzeinaTranscripts.txt
```

```
awk '$14 == 1 || $14 == 2 || $14 == 3 || $14 == 4 || $14 == 5 || $14 == 6 || $14 == 7 || $14 == 8 ||
$14 == 9 || $14 == 10 || $14 ~ /B73V4_/ || $14 ~ /Mt/ || $14 ~ /Pt/'
MasterAssembly_blat_awk_filtered.tsv | cut -f10 | sort | uniq -c | sort -nk1 >
MasterAssembly_blat_Filtered_B73.txt
```

```
awk '$14 ~ /scaffold/' MasterAssembly_blat_awk_filtered.tsv | cut -f10 | sort | uniq -c | sort -nk1 >
MasterAssembly_blat_Filtered_Eturcicum.txt
```

```
awk '$1 ~ />/ {print $1}' ../MasterAssembly2_trinity/Trinity.fasta | sed 's/>/' | sort >
MasterAssembly2_IDs.txt
```

```
sort MasterAssembly_blat_Filtered_TotalTranscripts.txt >
MasterAssembly_blat_Filtered_TotalTranscripts_sorted.txt
```

```
diff -y MasterAssembly2_IDs.txt MasterAssembly_blat_Filtered_TotalTranscripts_sorted.txt | awk  
'$1 != $2 {print $1}' > MasterAssembly_blat_Filtered_UnalignedTranscripts.txt
```