

# Pathogenomics of *Phytophthora fragariae*, the causal agent of strawberry red core disease

A thesis submitted for the degree of Doctor of Philosophy  
School of Agriculture, Policy and Development

Thomas Michael Adams

January 2019

## Abstract

Strawberries are a valuable fruit crop for the UK and worldwide. The UK berry market was worth £1.2 billion in 2016, with strawberries as 50% of this market. One of the major threats to the industry is the water-borne disease strawberry red core, caused by the oomycete pathogen *Phytophthora fragariae*.

One way to increase the resilience of the industry to this pathogen is breeding for genetic resistance in plants. To do this effectively, a strong understanding of the pathogen, including identification of avirulence genes is required. To investigate this, ten isolates of *P. fragariae* and three isolates of the related raspberry pathogen *Phytophthora rubi* were sequenced with the Illumina MiSeq system. PacBio sequencing was used on an additional isolate of *P. fragariae*, BC-16, to produce an improved reference assembly of 91 Mb in 180 contigs. Following assembly of these genomes and gene prediction through RNA-Seq guided and *ab initio* methods, effector genes were identified. Further investigations utilising orthology analysis, variant calling and differential expression analysis allowed for the identification of candidate avirulence genes for races UK1, UK2 and UK3. A very strong candidate for the avirulence gene in the UK2 race was identified as an RxLR effector named g27513 in BC-16. Further investigations did not identify clear *cis* or *trans* DNA sequence polymorphisms causing the variation in expression across races. Three of the identified races (UK1, UK2 and UK3) clustered into a single population with low levels of nucleotide diversity and a mixture of clonal and sexual reproduction which may possibly be undergoing recovery from a recent genetic bottleneck.

The candidate genes provide future targets for identifying avirulence genes to aid in effector guided breeding approaches to produce disease resistant strawberry plants. This work also provides resources for further investigations of this pathosystem and the *Phytophthora* genus as a whole.

## **Declaration**

I confirm that this is my own work and the use of all material from other sources has been properly and fully acknowledged.

Thomas Michael Adams

## Acknowledgements

I would first like to thank my three official supervisors for offering me the opportunity to undertake this PhD work over the past three years: Dr. Richard Harrison and Dr. Charlotte Nellist at NIAB EMR and Prof. Jim Dunwell at the University of Reading. I additionally would like to thanks Dr. Helen Bates and Kieran Housley for performing genome sequencing.

Despite being a novice in bioinformatics when starting this PhD, I am grateful to Dr. Andrew Armitage for introducing me to the world of genome assembly, annotation and pathogenomics and to Dr. Maria Sobczyk for showing me the wonderful world of population genetics. I am grateful to Dr. Robert Vickerstaff for managing the NIAB EMR cluster system and for being a source of UNIX knowledge, without whom conducting the analyses would have been difficult.

Additionally, I am grateful to Dr. Charlotte Nellist for assisting me with culturing the pathogen and performing pot-based pathogenicity work, as well as for her initial race typing work before I began work on this PhD. I am also grateful for the assistance provided by Dr. Bethany Greenfield and Dr. Emma Cascant-López for their advice on working with RNA and conducting RT-qPCR experiments.

Thanks also go to kind colleagues who aided in proof-reading this thesis and providing advice. Particular thanks go to Dr. Charlotte Nellist and Dr. Richard Harrison who read the entire document. Thanks also go to: Dr. Andrew Armitage, Dr. Emma Cascant-López, Dr. Helen Cockerton and Dr. Michelle Hulin.

Most importantly, I would like to thank my family for their unwavering support and encouragement through the inevitable ups and downs of PhD studies. Particularly I would like to thank my partner Chloë Scott, who must be sick of hearing about *Phytophthora* by now, though I think the spare fruit helped!

Finally, I am grateful to the University of Reading and NIAB EMR for providing the funding for my PhD studies.



## Table of contents

<b>Chapter 1: General Introduction</b> .....	<b>1</b>
1.1 Strawberry as an important horticultural crop .....	1
1.2 The plant immune system and strategies of oomycete pathogens to evade defences .....	3
1.3 <i>Phytophthora fragariae</i> as a member of a genus of highly destructive oomycete pathogens of plants .....	11
1.4 Developments in genome and transcriptome sequencing technologies .....	19
1.5 Aims and Objectives .....	20
<b>Chapter 2: General Materials and Methods</b> .....	<b>22</b>
2.1 Sources of isolates of <i>Phytophthora fragariae</i> and <i>Phytophthora rubi</i> .....	22
2.2 Production of media for isolate and micropropagated plant growth .....	24
2.3 Isolate maintenance, subculturing and routine growth .....	25
2.4 Source of plant material .....	26
2.5 Ordering of primers .....	27
2.6 DNA extraction for sequencing .....	27
2.7 RNA extraction for sequencing and qRT-PCR .....	28
2.7.1 Extraction from inoculated plant tissue .....	28
2.7.2 Extraction from <i>in vitro</i> mycelium .....	28
2.7.3 Assessment of RNA quality .....	29
2.8 Sequencing of DNA and RNA .....	29
2.8.1 Illumina sequencing of DNA .....	29
2.8.2 PacBio SMRT sequencing of DNA .....	31
2.8.3 Oxford Nanopore Technologies sequencing .....	31
2.8.4 RNA sequencing .....	32
2.9 Availability of code and data .....	32
2.10 Computing resources .....	32
<b>Chapter 3: Characterisation of <i>Phytophthora fragariae</i> races and genome assembly of isolates of <i>P. fragariae</i> and <i>Phytophthora rubi</i></b> .....	<b>33</b>
3.1 Introduction .....	33

3.1.1 The NIAB EMR collection of <i>Phytophthora fragariae</i> and <i>Phytophthora rubi</i> .....	33
3.1.2 Assessment of the pathogenicity of <i>Phytophthora fragariae</i> and current knowledge of the pathogenicity races of the collection .....	33
3.1.3 Genome assembly of <i>Phytophthora</i> species.....	36
3.2 Aims.....	37
3.3 Materials and Methods .....	38
3.3.1 Isolate preparation .....	38
3.3.2 Inoculation procedure .....	38
3.3.3 Experimental conditions.....	38
3.3.4 Disease assessment.....	38
3.3.5 Preparation of raw data .....	39
3.3.6 <i>De novo</i> assembly of genomes.....	40
3.3.6.1 Assembly of Illumina data.....	40
3.3.6.2 Assembly of PacBio data.....	40
3.3.7 Removal of contaminant sequences from assemblies.....	42
3.3.8 Masking of repetitive sequence and transposon sequences .....	46
3.3.9 Availability of sequencing data.....	46
3.4 Results.....	48
3.4.1 Characterisation of the race of a selection of <i>Phytophthora fragariae</i> isolates allowed conversion of three Canadian races and one U.S.A. race to the U.K. race scheme .....	48
3.4.2 Assessment of quality of sequencing data showed the data was sufficient for genome assembly .....	52
3.4.3 <i>De novo</i> assembly of genomes.....	57
3.4.3.1 <i>De novo</i> assembly of genomes from Illumina data showed little variation at the genome level between the isolates.....	57
3.4.3.2 <i>De novo</i> assembly of a new reference assembly using PacBio sequencing data improved the contiguity of the genome assembly .....	59
3.4.4 Identification and removal of error and contaminant sequences identified bacterial contaminants and improved the assemblies .....	61
3.4.5 Assessment of genome completeness showed the assemblies represented the majority of the core genome of eukaryotic organisms.....	66

3.4.6 Identification and masking of regions of repetitive sequence showed a large proportion of the genome consisted of repeat rich regions .....	68
3.4.7 Comparison to published assemblies of <i>Phytophthora fragariae</i> demonstrated the improvement achieved by the PacBio assembly.....	70
3.5 Discussion .....	73
3.5.1 Pot based pathogenicity tests allowed assignment of additional isolates of <i>Phytophthora fragariae</i> to the U.K. race scheme .....	73
3.5.2 <i>De novo</i> assembly of genomes of isolates of <i>Phytophthora fragariae</i> and <i>Phytophthora rubi</i> showed similar genome sizes and completeness and allowed the creation of a new reference assembly .....	73
3.5.3 Comparison with other studied <i>Phytophthora</i> spp. showed similarities to a species in the same clade, but differences compared to a more distantly related species .....	75
3.5.4 Conclusions .....	76

<b>Chapter 4: Genome annotation and initial investigations for the identification of race-specific effectors .....</b>	<b>77</b>
4.1 Introduction .....	77
4.1.1 Inoculation methods of <i>in vitro</i> plants for root infecting <i>Phytophthora</i> spp. ....	77
4.1.2 Molecular detection of <i>Phytophthora fragariae</i> infection .....	78
4.1.3 Methods of gene prediction and the identification of effector genes .....	78
4.1.4 Changes in pathogen gene expression over time during infection by <i>Phytophthora</i> spp.....	79
4.1.5 Copy number variation in avirulence genes in <i>Phytophthora</i> spp. ....	80
4.1.6 Avirulence genes and presence/absence variation in <i>Phytophthora</i> spp.	81
4.2 Aims.....	82
4.3 Materials and Methods .....	84
4.3.1 Production of flooding solutions .....	84
4.3.2 Assessment of sporangia growth and production of zoospores.....	84
4.3.3 Methods of inoculation of micropropagated <i>Fragaria</i> × <i>ananassa</i> plants with <i>Phytophthora fragariae</i> .....	85
4.3.4 DNA extraction and testing for presence of <i>Phytophthora fragariae</i> in inoculated micropropagated <i>Fragaria</i> × <i>ananassa</i> plants .....	86

4.3.5 Design of novel primers for testing for the presence of <i>Phytophthora fragariae</i> in inoculated micropropagated <i>Fragaria × ananassa</i> plants .....	87
4.3.6 Preparation of inoculated micropropagated <i>Fragaria × ananassa</i> plant material for RNA extraction.....	88
4.3.7 RT-PCR analysis of extracted RNA to test for the presence of <i>Phytophthora fragariae</i> .....	89
4.3.8 Initial assessment of RNA-Seq data and cleaning of raw data .....	90
4.3.9 Alignment of RNA-Seq reads to genome assemblies .....	90
4.3.10 <i>Ab initio</i> annotation of <i>Phytophthora fragariae</i> and <i>Phytophthora rubi</i> genome assemblies.....	91
4.3.10.1 Initial gene prediction.....	91
4.3.10.2 Prediction of effector genes .....	92
4.3.10.3 Creation of a final gene model set .....	93
4.3.10.4 Submission of draft annotated assemblies to GenBank .....	94
4.3.11 Analysis of differential expression between time points from the inoculation time course experiment and <i>in vitro</i> grown mycelium .....	96
4.3.12 Assessment of copy number variation .....	96
4.3.13 Orthology analysis of all proteins predicted for all sequenced <i>Phytophthora fragariae</i> and <i>Phytophthora rubi</i> isolates .....	97
4.3.14 Availability of sequencing data and annotated assemblies.....	97
4.4 Results.....	99
4.4.1 Trial of methods to generate a zoospore suspension of <i>Phytophthora fragariae</i> were not sufficient for inoculation experiments .....	99
4.4.2 Development of an inoculation procedure for micropropagated <i>Fragaria × ananassa</i> plants with <i>Phytophthora fragariae</i> allowed the conducting of an <i>in vitro</i> pathogenicity time course experiment .....	99
4.4.3 Design of primers for assessing presence of <i>Phytophthora fragariae</i> from RNA extracted from inoculated <i>Fragaria × ananassa</i> micropropagated plants resulted in the development of a suitable PCR based assay for detection of <i>P. fragariae</i> .....	103
4.4.4 Generation of RNA-Seq data for the BC-16 isolate of <i>Phytophthora fragariae</i> from an inoculation time course resulted in the sequencing of three time points post inoculation .....	107

4.4.5 Assessment of quality of sequencing data demonstrated a likely increase in the amount of <i>Phytophthora fragariae</i> RNA-Seq reads from the inoculation time course.....	110
4.4.6 Alignment of RNA-Seq data to assembled genomes of <i>Phytophthora fragariae</i> and <i>Phytophthora rubi</i> showed an increased proportion of <i>P. fragariae</i> reads as the inoculation time course progressed .....	112
4.4.7 <i>Ab initio</i> gene prediction .....	115
4.4.7.1 Initial gene prediction showed similar numbers of genes in all isolates .....	115
4.4.7.2 Prediction of putative effector genes showed similar numbers of possible effector genes in all isolates .....	117
4.4.7.3 Creation of a final gene model set provided a set for further investigation to identify putative avirulence genes.....	121
4.4.7.4 Additional functional annotation of gene annotations provided additional evidence for possible involvement of genes in the pathogenicity process.....	125
4.4.8 Analysis of expression in the BC-16 isolate of <i>Phytophthora fragariae</i> showed wide ranging transcriptional reprofiling occurring upon infection .....	127
4.4.9 Assessment of gene copy number variation in Illumina-only genomes showed little evidence of copy number variation between isolates of three distinct pathogenicity races.....	138
4.4.10 Orthology analysis identified presence/absence variation between <i>Phytophthora fragariae</i> and <i>Phytophthora rubi</i> effectors with a single possible avirulence gene for race UK2 .....	140
4.5 Discussion .....	147
4.5.1 Development of a novel method for the inoculation of micropropagated <i>Fragaria</i> × <i>ananassa</i> plants with <i>Phytophthora fragariae</i> allowed for the extraction of RNA from a time course experiment.....	147
4.5.2 <i>Ab initio</i> annotation of assembled genomes and prediction of effector class genes produced a valuable resource for pathogenomic investigations.....	148
4.5.3 Analysis of expression patterns during the inoculation time course and in relation to <i>in vitro</i> grown mycelium showed the time course experiment had captured numerous stages of the infection process.....	150

4.5.4 Copy number variation analysis showed the reference assembly had captured the majority of copy number variation in <i>Phytophthora fragariae</i> , with one gene being of possible future interest.....	152
4.5.5 Orthology analysis identifies a candidate avirulence gene for race UK2	153
4.5.6 Conclusions .....	153

**Chapter 5: Population genetic investigations and identification of candidate avirulence genes in three pathogenicity races of *Phytophthora fragariae***

.....	<b>156</b>
5.1 Introduction.....	156
5.1.1 Pathogenomic investigations beyond presence/absence variations.....	156
5.1.2 Population genetics insights in filamentous plant pathogens.....	158
5.2 Aims.....	161
5.3 Materials and Methods .....	162
5.3.1 Alignment of raw Illumina sequencing reads to the FALCON assembly of BC-16 .....	162
5.3.2 Variant calling process and annotation of variant sites .....	162
5.3.3 Assessment of population structure.....	163
5.3.4 Identification of private and polarising variant sites in races UK1, UK2 and UK3.....	164
5.3.5 Identification of structural variants and larger INDELS .....	164
5.3.6 Assessments of linkage disequilibrium .....	164
5.3.7 Population genetics analyses with the Popgenome R package.....	165
5.3.8 Preparation of RNA for sequencing .....	167
5.3.8.1 Testing of primers against additional isolates .....	167
5.3.8.2 Preparation of infection time course samples .....	167
5.3.8.3 RT-PCR to assess which time points to sequence .....	167
5.3.9 Initial assessment of RNA-Seq data, cleaning of raw data and aligning to assemblies.....	168
5.3.10 Identifying candidate avirulence genes in the BC-1, BC-16 and NOV-9 isolates.....	168
5.3.11 Confirmation of RNA-Seq results by RT-qPCR.....	168
5.3.12 Identification of putative promotor regions of a candidate avirulence gene .....	172

5.3.12.1 Building of a coexpression network .....	172
5.3.12.2 Identification of conserved regions upstream of genes of interest ....	172
5.3.13 Investigation of changes in <i>cis</i> of a candidate avirulence gene .....	173
5.3.13.1 Assembly and repeat masking of Oxford Nanopore Technologies (ONT) reads.....	173
5.3.13.2 Identification of variant sites nearest to the candidate avirulence gene .....	174
5.3.14 Investigation of changes in <i>trans</i> of a candidate avirulence gene.....	175
5.3.15 Investigation for a possible RxLR family .....	176
5.3.16 Assessment of synteny between the BC-16 and NOV-9 assemblies...	176
5.3.17 Availability of sequencing data and annotated assemblies.....	176
5.4 Results.....	178
5.4.1 Alignment of all Illumina reads to the reference assembly of the BC-16 isolate of <i>Phytophthora fragariae</i> showed similar mapping rates for isolates of the same species.....	178
5.4.2 Identification of variant sites allowed the construction of a phylogenetic tree of <i>Phytophthora fragariae</i> and demonstrated species separation between <i>P. fragariae</i> and <i>Phytophthora rubi</i> .....	180
5.4.3 Annotation of SNP sites for their effects in the BC-16 isolate of <i>Phytophthora fragariae</i> showed less than half of SNP sites were in exonic regions.....	185
5.4.4 Investigation of a population structure within the sequenced isolates of <i>Phytophthora fragariae</i> and <i>Phytophthora rubi</i> showed all isolates of pathogenicity races UK1, UK2 and UK3 of <i>P. fragariae</i> grouped as a single population .....	185
5.4.5 Investigation of polarising and private variant sites failed to explain the observed differences in pathogenicity race in the UKR1-2-3 population.....	189
5.4.6 Structural variant site identification failed to explain the observed differences in pathogenicity race in the UKR1-2-3 population .....	191
5.4.7 Population Genetic Analyses .....	191
5.4.7.1 Investigation of linkage disequilibrium in the UKR1-2-3 population suggested a mixed reproductive strategy with clonal reproduction and occasional sexual reproduction .....	191

5.4.7.2 Investigation of historical recombination within and between species showed greater levels of historical recombination between species than within species .....	196
5.4.7.3 Investigation of regions showing high levels of population separation showed variation across the genome with apoplastic effectors and a single functional term significantly overrepresented in genes which showed population separation.....	198
5.4.7.4 Investigation of levels of variation in the UKR1-2-3 population and the population represented by the BC-23 and ONT-3 isolates showed lower variation in the UKR1-2-3 population despite the presence of a similar number of segregating sites in both populations.....	202
5.4.7.5 Investigation of levels and types of selection in the UKR1-2-3 population showed little selection compared to an outgroup and the action of diversifying and purifying selection within the UKR1-2-3 population.....	205
5.4.8 Sequencing of mRNA from additional <i>Phytophthora fragariae</i> isolates, BC-1 and NOV-9, for additional investigations into candidate avirulence genes .....	209
5.4.9 Investigation of expression data to identify candidate avirulence genes for each race.....	213
5.4.9.1 Alignment of RNA-Seq data and identification of differentially expressed genes from <i>in planta</i> time points compared to <i>in vitro</i> mycelium showed transcriptional reprofiling during the pathogenicity process .....	213
5.4.9.2 Identification of uniquely expressed transcripts in the BC-1, BC-16 and NOV-9 isolates showed a large number of genes showing expression in only a single isolate.....	220
5.4.9.3 Identification of uniquely differentially expressed transcripts showed a large number of genes showing differential expression in only a single isolate .....	222
5.4.9.4 Identification of candidate avirulence genes in the BC-1, BC-16 and NOV-9 isolates produced a small number of high to medium confidence candidate genes .....	224
5.4.9.5 Confirmation of RNA-Seq data using RT-qPCR reinforces the identification of the high quality avirulence gene candidate in BC-16.....	226



5.4.10 Further investigation of candidate avirulence gene identified in BC-16 .....	230
5.4.10.1 Construction of a coexpression network allowed the identification of genes which showed similar expression patterns, although attempts to identify a promoter sequence failed .....	230
5.4.10.2 Investigation of changes in <i>cis</i> showed a SNP site 14 Kb downstream of the stop codon of the strong candidate avirulence gene in BC-16.....	231
5.4.10.3 Investigation of changes in <i>trans</i> identified some candidate transcription factors and transcriptional regulators that may be involved in the differential expression of the strong candidate avirulence gene in BC-16 .....	240
5.5 Discussion .....	241
5.5.1 Identification of variant sites allowed the resolution of a population structure with <i>Phytophthora fragariae</i> .....	241
5.5.2 Investigation of variant sites within the UKR1-2-3 population showed no sequence differences correlating with pathogenicity race.....	241
5.5.3 Population genetic analyses showed the UKR1-2-3 population was mostly clonally propagated with occasional sexual recombination and showed low levels of variation, likely due to a recent genetic bottleneck.....	242
5.5.4 Additional RNA-Seq data allowed for the identification of candidate avirulence genes.....	245
5.5.5 Investigations of a high-quality candidate avirulence gene in BC-16 failed to show a cause for the lack of expression in BC-1 and NOV-9 .....	246
5.5.6 Conclusions .....	247
<b>Chapter 6: General Discussion.....</b>	<b>248</b>
6.1 Key findings .....	248
6.2 Future directions .....	256
<b>References .....</b>	<b>261</b>
<b>Appendix .....</b>	<b>283</b>

## List of figures

<b>Fig. 1.1:</b> Summary of the plant immune receptor network model .....	4
<b>Fig. 1.2:</b> A summary of mechanisms of pathogen recognition by a plant cell.....	6
<b>Fig. 1.3:</b> A summary of defence responses deployed by the plant cell following pathogen detection at the plasma membrane.....	7
<b>Fig. 1.4:</b> A phylogeny for the <i>Phytophthora</i> genus derived from nuclear genetic makers.....	13
<b>Fig. 1.5:</b> Expanded phylogeny for clade 7 <i>Phytophthora</i> species based on nuclear genetic markers .....	14
<b>Fig. 1.6:</b> The infection life cycle of <i>Phytophthora fragariae</i> on <i>Fragaria</i> × <i>ananassa</i> plants.....	17
<b>Fig. 3.1:</b> Below ground symptoms of <i>Fragaria</i> × <i>ananassa</i> plants infected with <i>Phytophthora fragariae</i> .....	35
<b>Fig. 3.2:</b> An overview of assembly methods used in this study .....	45
<b>Fig. 3.3:</b> Harvested roots from pathogenicity test.....	50
<b>Fig. 4.1:</b> Summary of the gene prediction procedure conducted for all assembled isolates of <i>Phytophthora fragariae</i> and <i>Phytophthora rubi</i> .....	95
<b>Fig. 4.2:</b> Agarose gel electrophoresis of a nested PCR to detect <i>Phytophthora fragariae</i> in inoculated <i>Fragaria</i> × <i>ananassa</i> micropropagated plants .....	101-102
<b>Fig. 4.3:</b> Agarose gel electrophoresis of three primer pairs testing for the presence of <i>Phytophthora fragariae</i> isolate BC-16 by detection of β-tubulin ..	105
<b>Fig. 4.4:</b> Alignment of Sanger sequencing results to the <i>Phytophthora fragariae</i> β-tubulin gene sequence .....	106
<b>Fig. 4.5:</b> Agarose gel electrophoresis of an RT-PCR of a representative sample of each time point from an inoculation time course of the BC-16 isolate of <i>Phytophthora fragariae</i> on the 'Hapil' cultivar of the host plant <i>Fragaria</i> × <i>ananassa</i> .....	108
<b>Fig. 4.6:</b> Principal component analysis following expression assessment in the BC-16 isolate of <i>Phytophthora fragariae</i> .....	131

<b>Fig. 4.7:</b> Venn diagram of all differentially expressed transcripts in the BC-16 isolate of <i>Phytophthora fragariae</i> .....	132
<b>Fig. 4.8:</b> Venn diagram of all differentially expressed RxLR effectors in the BC-16 isolate of <i>Phytophthora fragariae</i> .....	133
<b>Fig. 4.9:</b> Venn diagram of all differentially expressed crinker effectors in the BC-16 isolate of <i>Phytophthora fragariae</i> .....	134
<b>Fig. 4.10:</b> Venn diagram of all differentially expressed apoplastic effectors in the BC-16 isolate of <i>Phytophthora fragariae</i> .....	135
<b>Fig. 4.11:</b> Venn diagram of all differentially expressed secreted proteins in the BC-16 isolate of <i>Phytophthora fragariae</i> .....	136
<b>Fig. 4.12:</b> Venn diagram of the distribution of shared orthogroups within a subset of seven isolates of <i>Phytophthora fragariae</i> representing three different races .....	142
<b>Fig. 4.13:</b> Alignments of BLASTn hits of a rejected candidate avirulence gene from BC-16 against BC-1 and NOV-9.....	143
<b>Fig. 4.14:</b> Alignment of Illumina reads from BC-1 against the annotated BC-16 assembly .....	144
<b>Fig. 4.15:</b> Alignment of Illumina reads from NOV-9 against the annotated BC-16 assembly .....	145
<b>Fig. 4.16:</b> Summary of the <i>de novo</i> assembly of the BC-16 isolate of <i>Phytophthora fragariae</i> from long read PacBio data, featuring the location of effector genes and mapping rates of raw short reads.....	155
<b>Fig. 5.1:</b> Pairwise comparison heatmaps and principal component analyses of percentage similarity of shared biallelic SNP sites for all sequenced isolates of <i>Phytophthora fragariae</i> and <i>Phytophthora rubi</i> .....	182
<b>Fig. 5.2:</b> Neighbour joining tree of all sequenced <i>Phytophthora fragariae</i> isolates based on high quality, biallelic SNP sites .....	183
<b>Fig. 5.3:</b> Distruct plots of fastSTRUCTURE results conducted on high quality, biallelic SNP sites .....	187
<b>Fig. 5.4:</b> Distruct plots of fastSTRUCTURE results conducted on high quality, biallelic SNP sites .....	188
<b>Fig. 5.5:</b> Assessment of linkage disequilibrium decay by distance in the BC-16 isolate of <i>Phytophthora fragariae</i> .....	193

<b>Fig. 5.6:</b> Assessment of the recombination rate along contig 8 and depiction of a recombination hotspot .....	194
<b>Fig. 5.7:</b> Results of a four gamete test conducted on different populations of <i>Phytophthora fragariae</i> and <i>Phytophthora rubi</i> isolates .....	197
<b>Fig. 5.8:</b> Results of pairwise $F_{ST}$ , Hudson's $K_{ST}$ and average $D_{xy}$ analysis comparing the UKR1-2-3 <i>Phytophthora fragariae</i> population to the population consisting of the isolates BC-23 and ONT-3 of <i>P. fragariae</i> .....	200
<b>Fig. 5.9:</b> Measures of variation in the UKR1-2-3 population and the BC-23 and ONT-3 population of <i>Phytophthora fragariae</i> .....	204
<b>Fig. 5.10:</b> Assessment of selection levels and types in the UKR1-2-3 population of <i>Phytophthora fragariae</i> .....	208
<b>Fig. 5.11:</b> Agarose gel electrophoresis of the RT-PCR products of $\beta$ -tubulin primers on samples from a time course of BC-1 and NOV-9 infection on <i>Fragaria</i> $\times$ <i>ananassa</i> plants of the 'Hapil' cultivar.....	210
<b>Fig. 5.12:</b> Principal component analysis following expression assessment using RNA-Seq data from three isolates with BC-1 as a reference .....	215
<b>Fig. 5.13:</b> Principal component analysis following expression assessment using RNA-Seq data from three isolates with BC-16 as a reference.....	216
<b>Fig. 5.14:</b> Principal component analysis following expression assessment using RNA-Seq data from three isolates with NOV-9 as a reference .....	217
<b>Fig. 5.15:</b> Results of RT-qPCR analysis of three genes of interest in the A4, BC-1, BC-16 and NOV-9 isolates of <i>Phytophthora fragariae</i> .....	228
<b>Fig. 5.16:</b> Graphical representation of sequence variants in the flanking regions of the candidate avirulence gene in BC-16 and its orthologous gene in NOV-9 .....	233
<b>Fig. 5.17:</b> Alignment of the Illumina reads of NOV-9 to the BC-16 FALCON assembly at the identified SNP downstream of the strong candidate avirulence gene in BC-16.....	234
<b>Fig. 5.18:</b> Alignment of the Illumina reads of NOV-9 to the BC-16 FALCON assembly at the identified INDEL upstream of the strong candidate avirulence gene in BC-16.....	235
<b>Fig. 5.19:</b> Alignment of the Illumina reads of BC-16 to the NOV-9 ONT assembly at the identified INDEL upstream of the orthologue of the strong candidate avirulence gene in BC-16 .....	236

<b>Fig. 5.20:</b> Alignment of the Illumina reads of NOV-9 to the NOV-9 ONT assembly at the identified INDEL upstream of the orthologue of the strong candidate avirulence gene in BC-16 .....	237
<b>Fig. 5.21:</b> Syntenic alignment of the assemblies of BC-16 and NOV-9 .....	238
<b>Fig. A.1:</b> Agarose gel electrophoresis of the PCR products of $\beta$ -tubulin primers on BC-1 and NOV-9 gDNA .....	283
<b>Fig. A.2:</b> Alignment of the strong candidate avirulence gene in BC-16 to the orthologous gene in NOV-9 .....	284
<b>Fig. A.3:</b> Alignment of the Illumina reads of BC-16 to the BC-16 FALCON assembly at the identified SNP downstream of the strong candidate avirulence gene in BC-16.....	285
<b>Fig. A.4:</b> Alignment of the Illumina reads of BC-16 to the NOV-9 ONT assembly at the identified SNP downstream of the orthologue of the strong candidate avirulence gene in BC-16 .....	286
<b>Fig. A.5:</b> Alignment of the Illumina reads of NOV-9 to the NOV-9 ONT assembly at the identified SNP downstream of the orthologue of the strong candidate avirulence gene in BC-16 .....	287
<b>Fig. A.6:</b> Alignment of the Illumina reads of BC-16 to the BC-16 FALCON assembly at the identified INDEL upstream of the strong candidate avirulence gene in BC-16.....	288

## List of tables

<b>Table 2.1:</b> Isolation details and race structure of isolates of <i>Phytophthora fragariae</i> and <i>Phytophthora rubi</i> used in this study.....	23
<b>Table 3.1:</b> Details of the NCBI SRA codes for all sequencing reads uploaded for future release.....	47
<b>Table 3.2:</b> Results of pathogenicity test with deduced U.K. races.....	51
<b>Table 3.3:</b> Summary of all pathogenicity races for all investigated <i>Phytophthora fragariae</i> isolates .....	51
<b>Table 3.4:</b> Fastqc results from analysis of raw sequencing data.....	53-54
<b>Table 3.5:</b> Improved Fastqc results from analysis of trimmed sequencing data .....	55-56
<b>Table 3.6:</b> Assembly statistics from pre-assembly and post-assembly .....	58
<b>Table 3.7:</b> Assembly statistics for the <i>Phytophthora fragariae</i> isolate BC-16 utilising PacBio sequencing data .....	60
<b>Table 3.8:</b> Assembly statistics after contaminant removal by DeconSeq .....	63
<b>Table 3.9:</b> Assembly statistics after NCBI suggestions performed .....	64
<b>Table 3.10:</b> Assembly statistics following manual removal of contaminant sequences .....	65
<b>Table 3.11:</b> CEGMA and BUSCO statistics for genome assemblies.....	67
<b>Table 3.12:</b> Repeat masking and transposon masking statistics for the genome assemblies.....	69
<b>Table 3.13:</b> Comparison of statistics between BC-16 FALCON-Unzip assembly, the published <i>Phytophthora fragariae</i> assemblies and the assembly of the closely related soybean pathogen <i>Phytophthora sojae</i> and the late blight pathogen <i>Phytophthora infestans</i> .....	72
<b>Table 4.1:</b> Primer sequences used to assess the presence of <i>Phytophthora fragariae</i> through detection of $\beta$ -tubulin .....	87
<b>Table 4.2:</b> Details of the NCBI SRA codes for all sequencing reads uploaded for future release.....	98

<b>Table 4.3:</b> Quality control results for RNA samples extracted from <i>Fragaria × ananassa</i> cultivar 'Hapil' plants inoculated with the BC-16 isolate of <i>Phytophthora fragariae</i> before sequencing .....	109
<b>Table 4.4:</b> Quality control statistics of RNA samples extracted from <i>in vitro</i> grown mycelium of the BC-16 isolate of <i>Phytophthora fragariae</i> before sequencing	109
<b>Table 4.5:</b> Fastqc results from analysis of trimmed RNA-Seq data .....	111
<b>Table 4.6:</b> Mapping rates of RNA-Seq data from <i>Phytophthora fragariae in vitro</i> grown mycelium and time points post inoculation with the BC-16 isolate of <i>P. fragariae</i> on <i>Fragaria × ananassa</i> cultivar 'Hapil' plant roots against <i>de novo</i> assembled genomes of isolates of <i>P. fragariae</i> .....	114
<b>Table 4.7:</b> Mapping rates of RNA-Seq data from <i>Phytophthora rubi</i> against <i>de novo</i> assembled genomes of isolate of <i>P. rubi</i> .....	114
<b>Table 4.8:</b> The number of genes present from both guided and unguided gene predictions in all sequenced isolates .....	116
<b>Table 4.9:</b> Predicted effector proteins from RNA-Seq guided gene models ...	119
<b>Table 4.10:</b> The number of different classes of effector proteins predicted via various methods on unguided gene predictions in <i>Phytophthora fragariae</i> and <i>Phytophthora rubi</i> isolates .....	120
<b>Table 4.11:</b> Final numbers of genes, proteins and putative effector class genes in all sequenced isolates of <i>Phytophthora fragariae</i> and <i>Phytophthora rubi</i> ...	123
<b>Table 4.12:</b> A comparison of BUSCO predictions from gene annotations against BUSCO predictions from unannotated assemblies for all sequenced <i>Phytophthora fragariae</i> and <i>Phytophthora rubi</i> isolates .....	124
<b>Table 4.13:</b> Number of genes predicted to possess transmembrane helices or GPI anchors.....	126
<b>Table 4.14:</b> The number of genes showing expression levels greater than the threshold FPKM value of 5 in the BC-16 isolate of <i>Phytophthora fragariae</i> ....	137
<b>Table 4.15:</b> The number of transcripts identified as being differentially expressed in any <i>in planta</i> time point post inoculation versus <i>in vitro</i> grown mycelium in the BC-16 isolate of <i>Phytophthora fragariae</i> .....	137
<b>Table 4.16:</b> Assessment of copy number variation in three isolates of <i>Phytophthora fragariae</i> : BC-1, BC-16 and NOV-9 .....	139
<b>Table 4.17:</b> Summary of expanded and unique orthogroup investigation for races UK1 (BC-1), UK2 (BC-16) and UK3 (NOV-9) of <i>Phytophthora fragariae</i> .....	146

<b>Table 5.1:</b> Primer sequences used in the RT-qPCR experiment.....	171
<b>Table 5.2:</b> Details of the NCBI SRA codes for all sequencing reads uploaded for future release.....	177
<b>Table 5.3:</b> Mapping rates of DNA-Seq reads of all eleven sequenced isolates of <i>Phytophthora fragariae</i> and all three sequenced isolates of <i>Phytophthora rubi</i> , to the reference assembly of the BC-16 isolate of <i>P. fragariae</i> .....	179
<b>Table 5.4:</b> The classification of the effects of variant sites in the BC-16 genome assembly .....	185
<b>Table 5.5:</b> Details of all sites identified as private variants in the UKR1-2-3 population .....	190
<b>Table 5.6:</b> Details of gene models contained in the recombination hotspot on BC-16 contig 8.....	195
<b>Table 5.7:</b> Investigation of the enrichment of classes of genes showing evidence of population separation between the <i>Phytophthora fragariae</i> UKR1-2-3 population and the population of the isolates BC-23 and ONT-3.....	201
<b>Table 5.8:</b> Quality control results on extracted RNA prior to sequencing .....	211
<b>Table 5.9:</b> Fastqc results from analysis of trimmed RNA-Seq data .....	212
<b>Table 5.10:</b> The number of transcripts showing expression levels greater than the threshold FPKM value of 5 in the three sequenced isolates of <i>Phytophthora fragariae</i> : BC-1, BC-16 and NOV-9 .....	218
<b>Table 5.11:</b> The number of transcripts showing differential expression in <i>in planta</i> time points compared to <i>in vitro</i> mycelium for BC-1, BC-16 and NOV-9, with the assemblies of these isolates used as references.....	219
<b>Table 5.12:</b> Identification of uniquely expressed genes in the BC-1, BC-16 and NOV-9 isolates using the assemblies of these three isolates as a reference .	221
<b>Table 5.13:</b> Identification of uniquely differentially expressed genes in the BC-1, BC-16 and NOV-9 isolates using the assemblies of these three isolates as a reference .....	223
<b>Table 5.14:</b> Identified candidate avirulence genes in the BC-1, BC-16 and NOV-9 isolates of <i>Phytophthora fragariae</i> .....	225
<b>Table 5.15:</b> Results of statistical testing for all significant comparisons from RT-qPCR data .....	229



<b>Table 5.16:</b> A comparison of assembly statistics between the FALCON assembly of BC-16 and the ONT assembly of NOV-9 .....	239
<b>Table 5.17:</b> Numbers of candidate transcription factors identified in the BC-1, BC-16 and NOV-9 isolates .....	240
<b>Table A.1:</b> Investigation of the enrichment of classes of genes showing evidence of population separation between the <i>Phytophthora fragariae</i> UKR1-2-3 population and the population of the isolates BC-23 and ONT-3.....	289-291
<b>Table A.2:</b> Investigation of the enrichment of effector classes of genes showing evidence of diversifying selection in the <i>Phytophthora fragariae</i> UKR1-2-3 population .....	292
<b>Table A.3:</b> Investigation of the enrichment of InterProScan terms of genes showing evidence of diversifying selection in the <i>Phytophthora fragariae</i> UKR1-2-3 population .....	293-297
<b>Table A.4:</b> Mapping rates of RNA-Seq data from <i>Phytophthora fragariae in vitro</i> grown mycelium and time points post inoculation with the BC-1 and NOV-9 isolates of <i>P. fragariae</i> on <i>Fragaria × ananassa</i> cultivar ‘Hapil’ plant roots against <i>de novo</i> assembled genomes of the BC-1, BC-16 and NOV-9 isolates of <i>P. fragariae</i> .....	298
<b>Table A.5:</b> All ANOVA and Tukey-HSD results for the strong candidate avirulence gene in BC-16.....	299-301
<b>Table A.6:</b> All ANOVA and Tukey-HSD results for an RxLR effector predicted to peak at 24 hpi .....	302-304
<b>Table A.7:</b> All ANOVA and Tukey-HSD results for an RxLR effector predicted to peak at 48 hpi .....	305-307
<b>Table A.8:</b> Gene names of all genes identified as putative candidate avirulence genes.....	308-315

## Abbreviations

<b>AFHRC</b>	Atlantic Food and Horticulture Research Centre
<b>AFLP</b>	Amplification Fragment Length Polymorphism
<b>ANOVA</b>	Analysis of Variance
<b>ATS</b>	<i>Arabidopsis thaliana</i> salts
<b>BBA</b>	Broad Bean Agar
<b>BLAST</b>	Basic Local Alignment Search Tool
<b>BUSCO</b>	Benchmarking Universal Single-Copy Orthologs
<b>BWA</b>	Burrows-Wheeler Aligner
<b>CAX</b>	Canadian race scheme assignment, where x is the race number
<b>cDNA</b>	Complementary Deoxyribonucleic Acid
<b>CEGMA</b>	Core Eukaryotic Genes Mapping Approach
<b>CNV</b>	Copy Number Variation
<b>CRN</b>	Crinkler Effector
<b>DEFRA</b>	Department for the Environment, Food and Rural Affairs
<b>DEPC</b>	Diethyl pyrocarbonate
<b>DET</b>	Differentially Expressed Transcript
<b>DNA</b>	Deoxyribonucleic Acid
<b>DNA-Seq</b>	Deoxyribonucleic Acid Sequencing
<b>dNTP</b>	Deoxyribonucleotide Triphosphate
<b>DREME</b>	Discriminative Regular Expression Motif Elicitation
<b>EBA</b>	Edamame Bean Agar
<b>EPPO</b>	European and Mediterranean Plant Protection Organisation
<b>ETI</b>	Effector Triggered Immunity
<b>FPKM</b>	Fragments Per Kilobase of transcript per Million mapped reads
<b>GAG</b>	Genome Annotation Generator
<b>GATK</b>	Genome Analysis Toolkit
<b>GBA</b>	Green Bean Agar
<b>gDNA</b>	Genomic Deoxyribonucleic Acid
<b>GPI</b>	Glycophosphatidylinositol
<b>HMM</b>	Hidden Markov Model
<b>HPC</b>	High Performance Computing

<b>HR</b>	Hypersensitivity Response
<b>HSD</b>	Honestly Significant Difference
<b>IDT</b>	Integrated DNA Technologies
<b>IGV</b>	Integrative Genomics Viewer
<b>INDEL</b>	Insertion or deletion
<b>ITS</b>	Internal Transcribed Spacer
<b>JHI</b>	James Hutton Institute
<b>KBA</b>	Kidney Bean Agar
<b>KMC</b>	<i>K</i> -mer Counter
<b>LD</b>	Linkage Disequilibrium
<b>LD<sub>50</sub></b>	Half decay distance of linkage disequilibrium
<b>LFC</b>	Log <sub>2</sub> Fold Change
<b>MAFFT</b>	Multiple Alignment using Fast Fourier Transform
<b>MBSA</b>	Mung Bean Shoot Agar
<b>MEME</b>	Multiple EM for Motif Elicitation
<b>MHAP</b>	MinHash Alignment Process
<b>MKT</b>	McDonald-Kreitman Test
<b>MNP</b>	Multiple Nucleotide Polymorphism
<b>mRNA</b>	Messenger Ribonucleic Acid
<b>MYA</b>	Million Years Ago
<b>NCBI</b>	National Centre for Biotechnology Information
<b>NGS</b>	Next Generation Sequencing
<b>NIAB EMR</b>	National Institute of Agricultural Botany East Malling Research
<b>NLR</b>	Nucleotide-binding domain and Leucine-rich Repeat containing receptor
<b>NRC</b>	NLR-Required for Cell death protein
<b>ONT</b>	Oxford Nanopore Technologies
<b>ORF</b>	Open Reading Frame
<b>PacBio</b>	Pacific Biosciences
<b>PAMP</b>	Pathogen Associated Molecular Pattern
<b>PBA</b>	Pea Broth Agar
<b>PB</b>	Pea Broth
<b>PCR</b>	Polymerase Chain Reaction
<b>PenSeq</b>	Pathogen enrichment sequencing

<b>PHPS</b>	Plant Health Propagation Scheme
<b>PICh</b>	Proteomics of Isolated Chromatin segments
<b>PSI-BLAST</b>	Position-Specific Iterative Basic Local Alignment Search Tool
<b>PTI</b>	Pattern Triggered Immunity
<b>PVX</b>	Potato Virus X
<b>qPCR</b>	Quantitative Polymerase Chain Reaction
<b>Regex</b>	Regular expression
<b>RLCK</b>	Receptor-Like Cytoplasmic Kinase
<b>RLK</b>	Receptor Like Kinase
<b>RLP</b>	Receptor Like Protein
<b>RNA</b>	Ribonucleic Acid
<b>RNA-Seq</b>	Ribonucleic Acid Sequencing
<b>ROS</b>	Reactive Oxygen Species
<b><i>Rpf1</i></b>	<i>Resistance to Phytophthora fragariae 1</i>
<b><i>Rpf2</i></b>	<i>Resistance to Phytophthora fragariae 2</i>
<b><i>Rpf3</i></b>	<i>Resistance to Phytophthora fragariae 3</i>
<b>rRNA</b>	Ribosomal Ribonucleic Acid
<b>RT-PCR</b>	Reverse Transcription Polymerase Chain Reaction
<b>RT-qPCR</b>	Quantitative Reverse Transcription Polymerase Chain Reaction
<b>SAMtools</b>	Sequence Alignment/Map tools
<b>SBA</b>	Soya Bean Agar
<b>SEA</b>	Strawberry Extract Agar
<b>SGS</b>	Second-Generation Sequencing
<b>SMA</b>	Strawberry Material Agar
<b>SMRT</b>	Single Molecule Real Time
<b>SNP</b>	Single Nucleotide Polymorphism
<b>SPAdes</b>	St. Petersburg genome assembler
<b>SRA</b>	Sequencing Read Archive
<b>SSR</b>	Simple Sequence Repeat
<b>STAR</b>	Spliced Transcripts Alignment to a Reference
<b>TAE</b>	Tris-Acetate-EDTA
<b>TGS</b>	Third-Generation Sequencing
<b>UKx</b>	UK race scheme assignment, where x is the race number
<b>USx</b>	US race scheme assignment, where x is the race number

**WGCNA**    Weighted correlation network analysis  
**Y2H**        Yeast-2-Hybrid

## Chapter 1: General Introduction

### 1.1 Strawberry as an important horticultural crop

Strawberry is an important soft fruit crop in the family *Rosaceae*, which includes various cultivated fruit species; such as apple (*Malus × domestica*), plum (*Prunus domestica*), cherry (*Prunus avium*) and peach (*Prunus persica*). The genus *Fragaria* is thought to be comparatively recent in origin. Data from chloroplast genome sequencing, combined with molecular clock estimates, suggested that the origin of the *Fragaria* genus to be 1.52 - 4.44 million years ago (MYA) and the origin of the first octoploid species to be 0.19 - 0.86 MYA (Njuguna *et al.*, 2013).

The octoploid, hermaphroditic species *Fragaria × ananassa* is the most commonly commercially cultivated species. This species is surprisingly recent in origin and is one of the youngest domesticated crop plants. It is thought that a hybridisation event occurred between two wild octoploid species, *Fragaria virginiana* from North America and *Fragaria chiloensis* from South America, in Southern France in the early 18<sup>th</sup> century (Hancock, 1999).

The precise composition of the octoploid genome has long been controversial and despite recent advances still remains unclear. The current model is based on the use of short sequences of single nucleotide polymorphisms (SNPs) located on the same chromosome, known as HaploSNPs, which exploit homoeologous sequence variants to reduce marker ploidy. This allowed the identification of a highly distinct A-subgenome showing high similarity to the wild diploid woodland strawberry *Fragaria vesca*, a reasonably distinct b-subgenome that is similar to the wild Japanese species *Fragaria iinumae* and no clear pattern for the remaining 14 chromosomes which appeared to be phylogenetic orphans. Therefore a genome structure of AA,bb,X-X,X-X was suggested (Sargent *et al.*, 2016). In this model, the authors propose that an ancestor of *F. vesca* subsp. *bracteata* served as the maternal partner in a cross with a hexaploid species descended from an ancestor of *F. iinumae* and the other unknown ancestor(s).

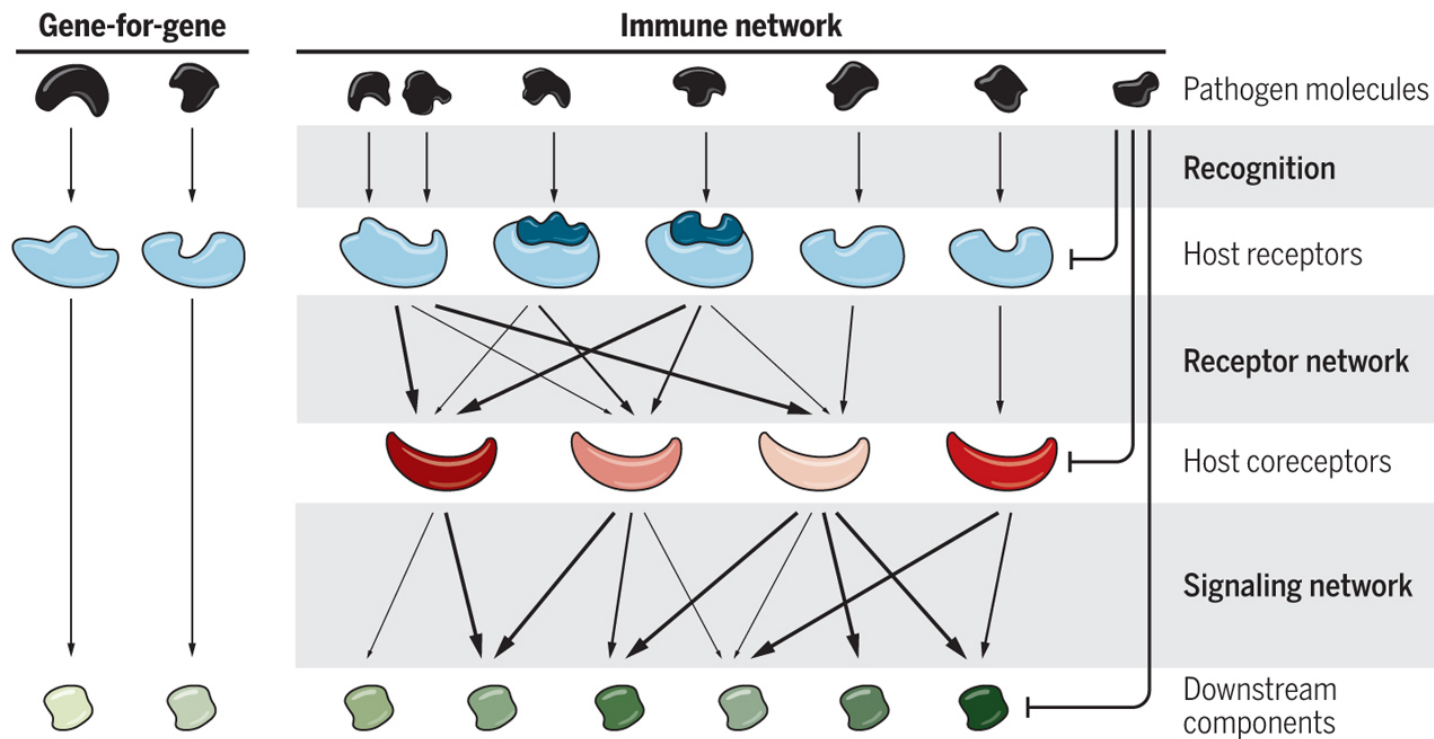
Despite this, in studies of strawberry genomics it is common to refer to the homoeologues of each chromosome as A, B, C and D for convenience (van Dijk *et al.*, 2014). It is also thought that the inheritance behaviour in this species is mostly disomic, based on work with simple sequence repeat (SSR) markers. However, some large linkage groups composed solely of markers in the coupling phase suggested that some partial polysomic behaviour still occurs, further suggesting that diploidisation is an ongoing process in *F. × ananassa* (Rousseau-Gueutin *et al.*, 2008). The genome of this organism has proven difficult to assemble due to its octoploid nature and that it is an obligate outcrosser, however a genome assembly has been produced with an approximate size of 698 Mb, representing 80% of the size of quadrupling the diploid genomes (~ 200 Mb each; Hirakawa *et al.*, 2014).

It is thought that the history of the human consumption of strawberries is both long and widespread across the Northern Hemisphere. The archaeobotanical evidence for consumption in pre-history is limited (Walshaw, 2009). Finds of strawberry achenes at pre-Columbian sites in eastern North America provided some evidence of consumption of wild strawberries by Native Americans, though there was no evidence of cultivation in these regions (Gremillion and Sobolik, 1996; Pauketat *et al.*, 2002). *F. chiloensis* was domesticated by the Picunche and Mapuche people of Chile over 1,000 years ago (Finn *et al.*, 2013). In Europe, the diploid woodland strawberry *F. vesca* has been cultivated since at least the Roman period and the hexaploid *Fragaria moschata* has been cultivated since the 16th century (Wilhelm and Sagen, 1974). Today, strawberry ranks among the most widely produced soft fruit crops, with 9,118,336 tonnes produced in 2016 worldwide; in 2016, the U.K. produced 118,179 tonnes, ranking 14<sup>th</sup> worldwide (<http://www.fao.org/faostat>, last accessed 03/01/2019). The U.K. Department for Environment, Food and Rural Affairs (DEFRA) estimated the farm gate value of home production marketed for the calendar year 2016 in the U.K. at £253,100,000 (DEFRA, 2018).

## 1.2 The plant immune system and strategies of oomycete pathogens to evade defences

Plants are constantly exposed to microbes in the environment, including those which are beneficial, commensal and harmful. In response to harmful microbes, plants have evolved an intricate defence system in order to protect themselves. The first explanation put forward to explain this system was a gene for gene model proposed by Flor (1942). This proposed that single, complementary genes in the pathogen and the plant controlled the resistance response. The gene-for-gene model was then refined to a 'Z-scheme' comprising of two layers of plant immunity: pattern triggered immunity (PTI) and effector triggered immunity (ETI). In this model, PTI is triggered by conserved pathogen associated molecular patterns (PAMPs). This resistance can then be overcome by pathogen effectors. In turn, the plant can evolve resistance (*R*) genes and trigger ETI. The Z-scheme model predicts that this resistance can occur over several rounds in an evolutionary arms race (Jones and Dangl, 2006; Hein *et al.*, 2009). Later work suggested that in fact PTI and ETI are not distinct processes but in fact represent a continuum of immune responses, with some effector proteins behaving as PAMPs against non-host species (Thomma *et al.*, 2011). Additionally, as demonstrated in a recent review, the plant immune system is far more complex than a collection of gene-for-gene interactions. The current model of the plant immune system describes a process with sensor receptors identifying molecules derived from pathogens either directly or indirectly. This is followed by the involvement of components such as: coreceptors, helper receptors or receptor-like regulatory scaffolds. Finally, these complexes recruit distinct, but overlapping, signalling components which result in resistance to the pathogen (Wu *et al.*, 2018; Fig. 1.1).



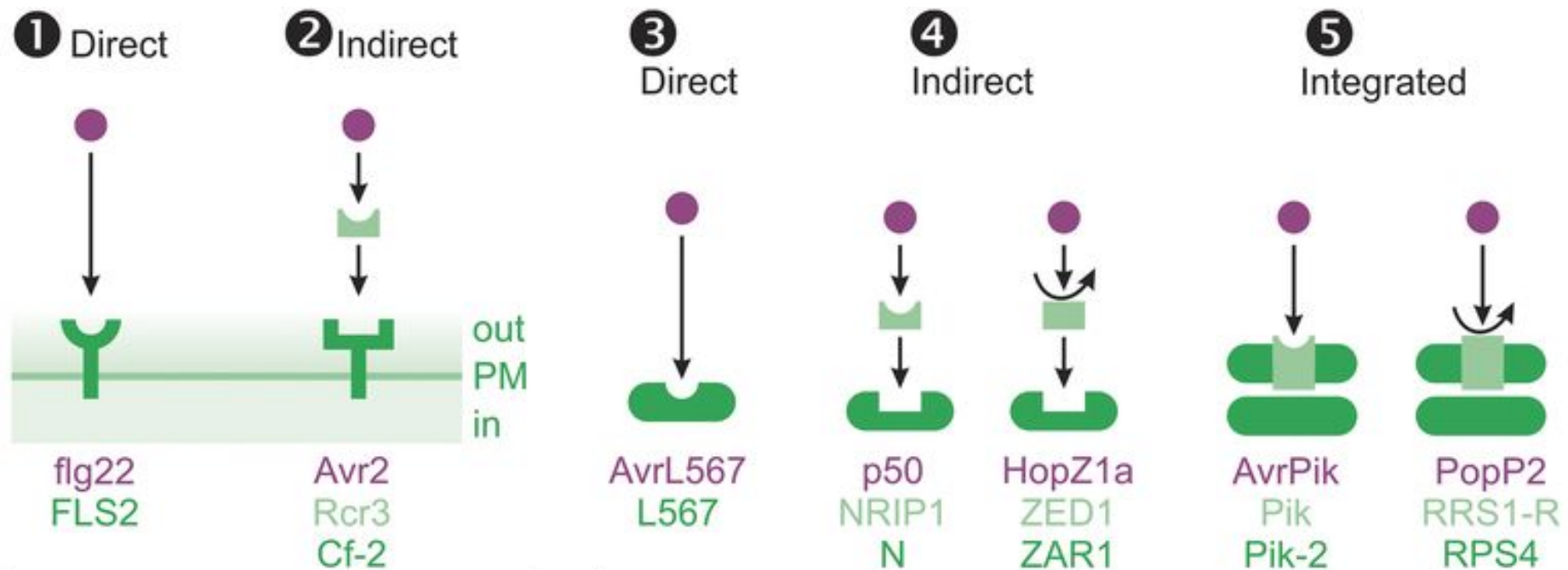


**Fig. 1.1: Summary of the plant immune receptor network model. Left:** The traditional gene-for-gene model of resistance, where a single effector from the pathogen is detected by a single resistance gene in the plant. This results in the activation of downstream resistance components. **Right:** The current immune network hypothesis, demonstrating a number of different recognition mechanisms, along with instances of cross-talk with co-receptors prior to the activation of downstream resistance components. Mechanisms depicted from left to right are: the detection of two effectors by a resistance gene being required for resistance; the guard model, where a resistance gene identifies an effector either by its action on an integrated decoy domain or a guardee protein; direct detection of effectors. Finally, the action of effectors to suppress immune responses is depicted (Adapted from: Wu *et al.*, 2018).

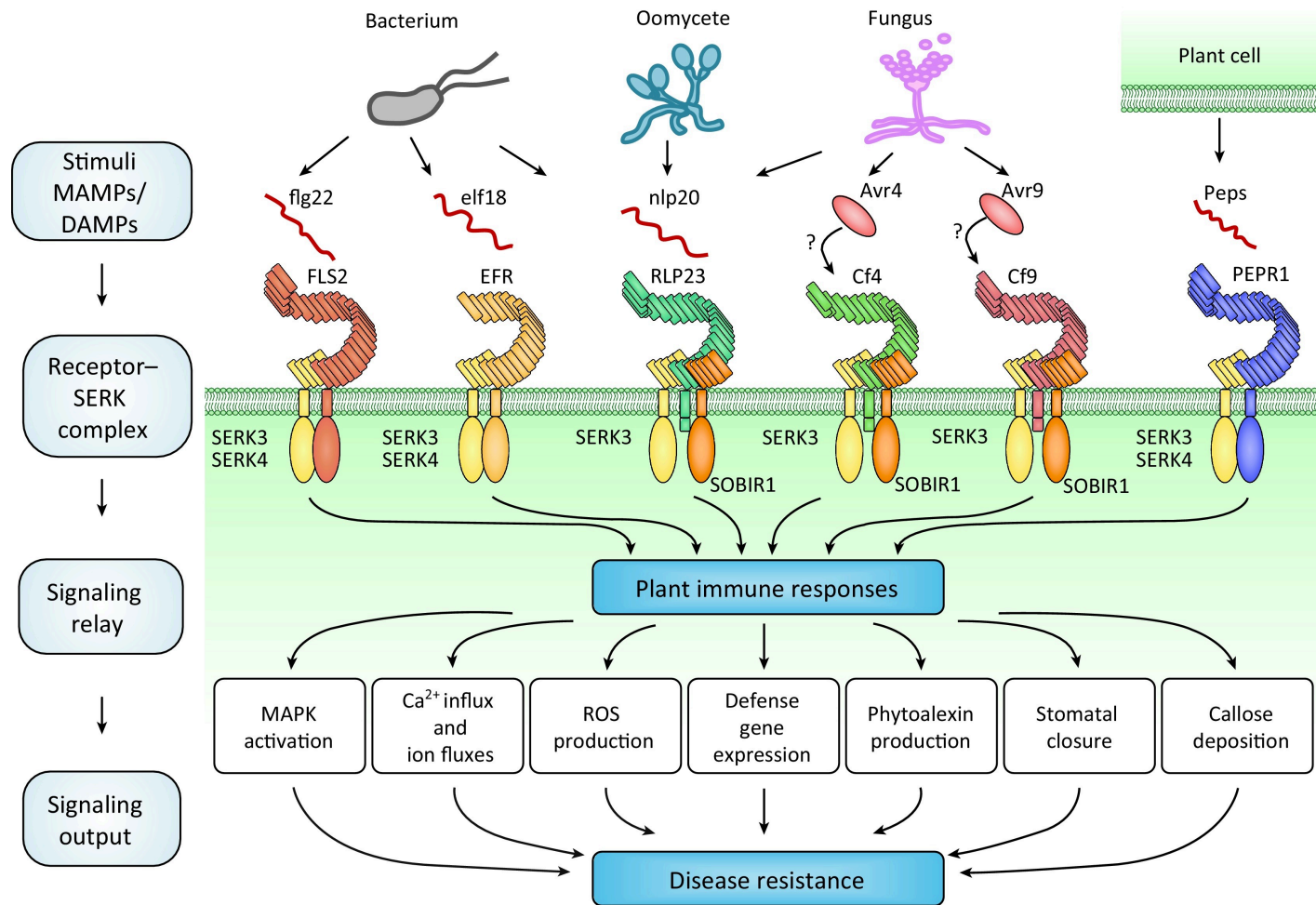
Detection of pathogenic microbes by the plant can occur in two subcellular locations. Firstly, PAMPs, such as bacterial flagellin and fungal chitin can be detected at the cell surface by receptor-like kinases (RLKs) and receptor-like proteins (RLPs). Secondly, should this first defence mechanism fail, intracellular nucleotide-binding domain and leucine-rich repeat containing (NLR) receptors can detect pathogenicity proteins produced by the pathogen and trigger a resistance response (Dodds and Rathjen, 2010; Fig. 1.2 and 1.3).

RLKs and RLPs form complexes with additional receptor-like cytoplasmic kinases (RLCKs) both before and after pathogen perception at the cell plasma membrane. These complexes lead to downstream resistance responses, including: the activation of intracellular kinases, the production of reactive oxygen species (ROS), callose deposition in the cell wall, the closure of stomata and transcriptional reprogramming of defence genes (Couto and Zipfel, 2016; Ma *et al.*, 2016; Fig. 1.2 and 1.3).

NLR proteins also usually do not act alone, instead they form configurations ranging from NLR pairs to complex networks. Pairs of NLRs most commonly function through the lifting of repression of the helper NLR by the sensor NLR upon pathogen detection (Jones *et al.*, 2016). Broadly, NLRs usually act by one of three methods: direct detection, indirect detection and via an integrated method including both a guard/guardee model and an integrated decoy model (Kourelis and van der Hoorn, 2018; Fig. 1.2 and 1.3). The concept of NLR networks is newer than that for RLKs and RLPs, however it has recently been demonstrated for the association of NLR-required for cell death proteins (NRCs) and NRC-dependent NLRs in asterid plants such as tomato (*Solanum lycopersicum*) and tobacco (*Nicotiana tabacum*; Wu *et al.*, 2017), as well as the involvement of paralogues of the *Arabidopsis thaliana* NLR ACTIVATED DISEASE RESISTANCE PROTEIN 1 (ADR1) helper NLR family being involved in the function of several NLRs (Bonardi *et al.*, 2011).



**Fig. 1.2: A summary of mechanisms of pathogen recognition by a plant cell.** Illustration of various mechanisms of detection of pathogens by plants, both extracellular (1 and 2) and intracellular (3 – 5). **1:** A pathogen associated molecular pattern (PAMP) is directly detected by a plant receptor-like kinase (RLK) protein. **2:** A pathogen effector is either bound by a host protein or modifies the host, this is detected by a plant RLK. **3:** Direct recognition, a pathogen effector is detected directly by a host resistance (R) protein. **4:** Indirect detection of an effector. **Left:** an effector is bound by a plant protein; this binding is detected by a plant R protein. **Right:** an effector modifies a plant protein; this modification is detected by a plant R protein. **5:** Integrated methods of detection. **Left:** A pathogen effector is bound by a plant protein which is associated with an R protein. This triggers a resistance response. **Right:** A pathogen effector modifies a plant protein which is associated with an R protein. This triggers a resistance response. (Adapted from Kourelis and van der Hoorn, 2018).



**Fig. 1.3: A summary of defence responses deployed by the plant cell following pathogen detection at the plasma membrane.**

Illustration of a summary of defence responses deployed by the plant cell following detection of pathogens at the cell membrane (Adapted from Ma *et al.*, 2016).

Like all biological networks, these networks require control in order to avoid autoactivation but also to ensure an appropriate response upon pathogen detection. Several processes are involved in the network control: gene expression, receptor complex formation and the activation of signal transduction (Wu *et al.*, 2018). For instance, in a selection of domesticated and wild relatives of rice (*Oryza sativa*), the genes which encode paired NLRs are predominantly arranged in a head-to-head formation (Stein *et al.*, 2018). This suggests that they may be under the control of similar gene expression regulatory elements. However, the precise mechanisms for how the expression of genetically unlinked receptors occurs have not been fully elucidated (Wu *et al.*, 2018). However, there is evidence for the involvement of several key regulatory factors from studies in a wide range of plant species, including: the action of transcription factors, regulation of chromatin modification, alternative splicing of transcripts of NLRs and the actions of small RNAs controlling the NLR transcript dosage (Zhang *et al.*, 2016; Lai and Eulgem, 2018). Complex formation is also controlled for cell surface receptors, both through the recruitment of other receptor proteins upon ligand detection to aid in signalling (Smakowska-Luzan *et al.*, 2018) and through negative regulation of signalling via the recruitment of RLCKs, the recruitment of pseudokinases, the action of phosphatases and the action of ubiquitin E3 ligases (Couto and Zipfel, 2016).

Plant pathogens can evade these defences through the action of effector proteins, which target a variety of components of the disease resistance process (Dodds and Rathjen, 2010). These proteins show high levels of redundancy, but can activate NLR mediated immunity. Effector genes are also expressed at different points of the infection process, suggesting they fulfil a wide variety of functions. Work in *Phytophthora capsici* identified four classes of RxLR effectors and two classes of crinkler effectors (CRNs) based on their expression patterns (Jupe *et al.*, 2013; Stam *et al.*, 2013). It can therefore be argued that effector diversification can be viewed as an evolutionary “gamble” which allows pathogens to develop new mechanisms to overcome plant immunity. Hence, plants and pathogens are in a complex coevolutionary arms race (Wu *et al.*, 2018).

Oomycete plant pathogens produce a variety of effector gene classes to allow the infection of host plants. These effectors can act in both the apoplast, extra-haustorial matrix and plant cell cytoplasm (Kamoun, 2006). However, determining the precise function of these effector proteins can prove difficult due to the lack of homology to proteins whose function has been characterised, likely due to the rapid diversification of these proteins in order to improve their effectiveness and avoid host recognition (Franceschetti *et al.*, 2017). Pathogens of the *Phytophthora* species contain two key classes of translocated effector: RxLR effectors and Crinkler effectors (CRNs). Previous studies in other *Phytophthora* spp. have identified 563 RxLRs and 196 CRNs in *Phytophthora infestans*; 350 RxLRs and 100 CRNs in *P. sojae*; and 350 RxLRs and 19 CRNs in *Phytophthora ramorum* (Tyler *et al.*, 2006; Haas *et al.*, 2009)

RxLR effectors are so named for the presence of the conserved N-terminal Arg-Xaa-Leu-Arg (RxLR) motif in their amino acid sequences. This motif has recently been shown to be cleaved in AVR3a from *P. infestans* prior to secretion from the cell, with likely acetylation of the N-terminus (Wawra *et al.*, 2017). It is likely that this is the case for other RxLR effectors and since they appear to be present in a variety of *Phytophthora* spp. it is likely this is a conserved mechanism allowing export from the cell. It has also been shown that in a majority of RxLR effectors, the RxLR motif is followed within 25 residues by a Glu-Glu-Arg (EER) motif. This allowed for an improvement in the sensitivity of bioinformatic approaches to identify these effectors via a hidden markov model (HMM; Whisson *et al.*, 2007). Additional studies of these effectors in *P. sojae* and *P. ramorum* identified three additional domains: Trp (W), Tyr (Y) and Leu (L) with the respective amino acid at a highly conserved residue in the sequence (Jiang *et al.*, 2008). Analysis of the crystal structure of AVR3a11 from *P. capsici* and PexRD2 from *P. infestans* showed that the W and Y motifs formed a specific fold within a three- $\alpha$ -helical bundle and were stacked against one another (Boutemy *et al.*, 2011). The authors also created a HMM and demonstrated that under half of the RxLRs studied would adopt this fold.

CRNs are another key class of effector protein in *Phytophthora* spp.. These proteins were first identified through the expression of the gene in a potato virus

X (PVX) vector in plant cells, resulting in necrosis and crinkling in systemic leaves (Torto *et al.*, 2003). These proteins possess two key domains: a highly conserved N-terminal ~50-amino-acid Leu-Xaa-Leu-Phe-Leu-Ala-Lys (LxLFLAK) motif, referred to as the LFLAK domain, and a His-Val-Leu-Val-Val-Pro (HVLVVP) motif, referred to as the DWL domain. Some of these proteins contain a DI domain between those two domains (Haas *et al.*, 2009; Stam *et al.*, 2013). The LFLAK domain has been shown to be involved in entry into the host cell and the majority of these effectors appear to target the host cell nucleus (Orsomando *et al.*, 2011). Interestingly, these proteins do not appear to display classical signals of secretion, however two HMM models have recently been developed to aid in their bioinformatic identification (Armitage *et al.*, 2018b). These models were trained from 271 CRNs predicted for *P. infestans*, *P. sojae*, *P. ramorum* and *P. capsici* (Tyler *et al.*, 2006; Haas *et al.*, 2009; Stam *et al.*, 2013).

However, these two classes of genes do not describe the entire suite of effectors for *Phytophthora* spp.. Recent progress includes the development of a machine learning model trained on a set of 84 apoplastic plant and pathogen proteins, with 1,773 proteins making up a negative training set to aid in bioinformatic identification of these genes (Sperschneider *et al.*, 2018).

Whilst a large number of effector genes from *Phytophthora* spp. have been characterised as suppressors of host resistance, the precise intracellular function of many of these genes has yet to be elucidated. Despite this, for a few effectors, the molecular function has been determined. One example is AVR3a from *P. infestans*. Through the use of yeast-two-hybrid (Y2H) experiments alongside *in planta* studies, this effector was shown to interact with the potato (*Solanum tuberosum*) U-box E3 ligase CMPG1, resulting in the suppression of INF1-triggered cell death (Bos *et al.*, 2010). Also, in this pathogen, the RxLR effector PiTG\_03192 was shown to prevent the movement of two potato NAC transcription factors from the endoplasmic reticulum to the nucleus, through the use of Y2H and fluorescent microscopy (McLellan *et al.*, 2013).

### **1.3 *Phytophthora fragariae* as a member of a genus of highly destructive oomycete pathogens of plants**

Oomycetes are a group of filamentous organisms that initially were considered to be fungi. However, the development of molecular phylogenetic approaches lead to the formation of kingdom level phylogenies of eukaryotic organisms which initially grouped oomycetes as stramenopiles (also known as heterokonts) along with many photosynthetic algal species (Baldauf *et al.*, 2000). Interestingly, following the sequencing of the genomes of *P. sojae* and *P. ramorum*, 855 genes were identified with possible heritage from either an ancestral red alga or a previously engulfed cyanobacterial species. Additionally, some genes still appeared to display plastid localisation sequences, despite *Phytophthora* species no longer possessing a plastid organelle. This suggested that oomycetes did indeed possess a photosynthetic ancestor, as predicted by the chromalveolate hypothesis (Tyler *et al.*, 2006). However, it remains unclear which development came first in oomycete evolution: the loss of the ability to photosynthesise or the transition to a parasitic lifestyle.

The *Phytophthora* ('the plant-destroyer' in Greek) genus contains a large number of highly destructive pathogens in both agricultural and ecological contexts (Erwin and Ribeiro, 1996). The most well-known member of this genus is the causative agent of late blight disease, the disease involved in the Irish potato famine, *P. infestans*. Notoriously this famine resulted in approximately one million deaths and the emigration of 1.5 million people from Ireland in the mid-19<sup>th</sup> century. Today, this disease is still a major problem for potato production (Turner, 2005). More recently, *P. ramorum* emerged in the mid-1990s as the cause of the sudden oak death disease in Oregon and California devastating tanoak (*Lithocarpus densiflorus*) and coast live oak (*Quercus agrifolia*) and is now present in both North America and Europe affecting over forty plant genera (Rizzo *et al.*, 2005). The number of species identified in this genus has increased rapidly over the last few decades, from about 58 in 1996 (Erwin and Ribeiro, 1996) to over 150 in 2017 (Yang *et al.*, 2017). Prior to the development of molecular methods for phylogenetic classification, Waterhouse (1963) developed a morphological classification scheme consisting of six clades using three



morphological characteristics. More recently, an examination of more than 180 *Phytophthora* taxa led to the creation of a system of classification consisting of ten major clades with twenty-four subclades (Yang *et al.*, 2017; Fig. 1.4). The species studied in this thesis, *Phytophthora fragariae*, resides in subclade 7a alongside the closely related raspberry pathogen *Phytophthora rubi*. None of the other species in subclade 7a have been well studied, however *P. sojae* resides in subclade 7b (Yang *et al.*, 2017; Fig. 1.5).

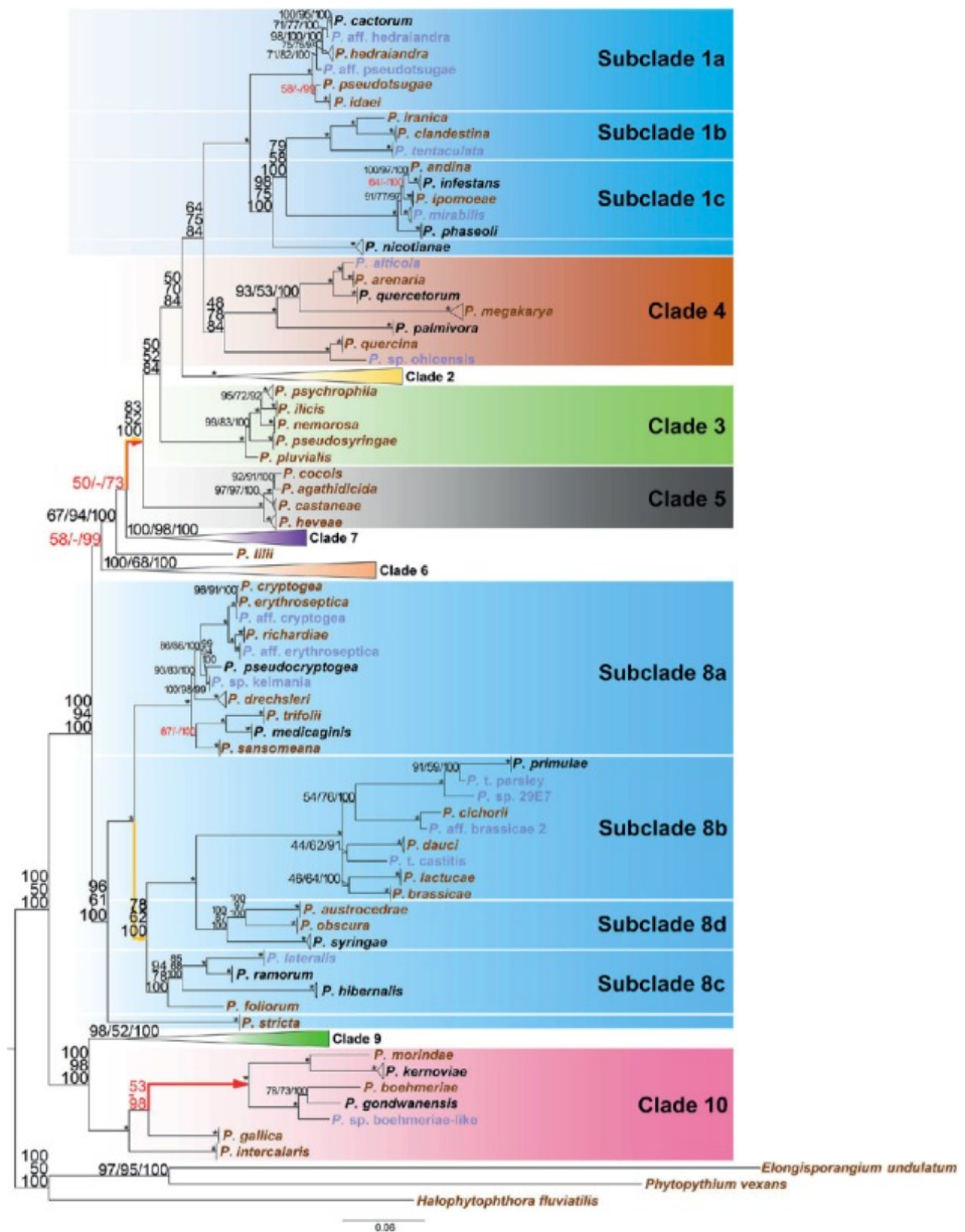
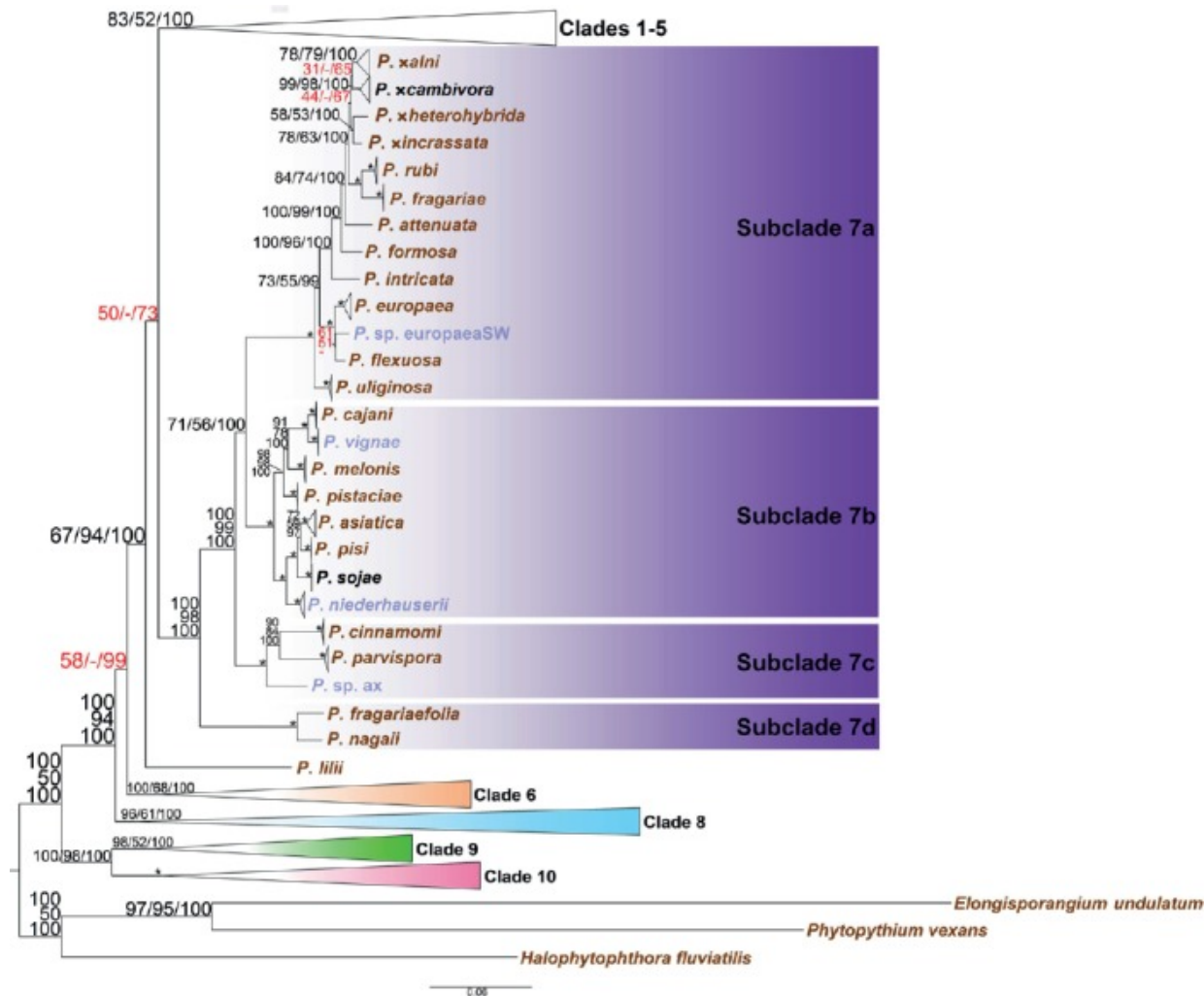


Fig. 1.4: A phylogeny for the *Phytophthora* genus derived from nuclear genetic makers. (Adapted from Yang *et al.*, 2017).

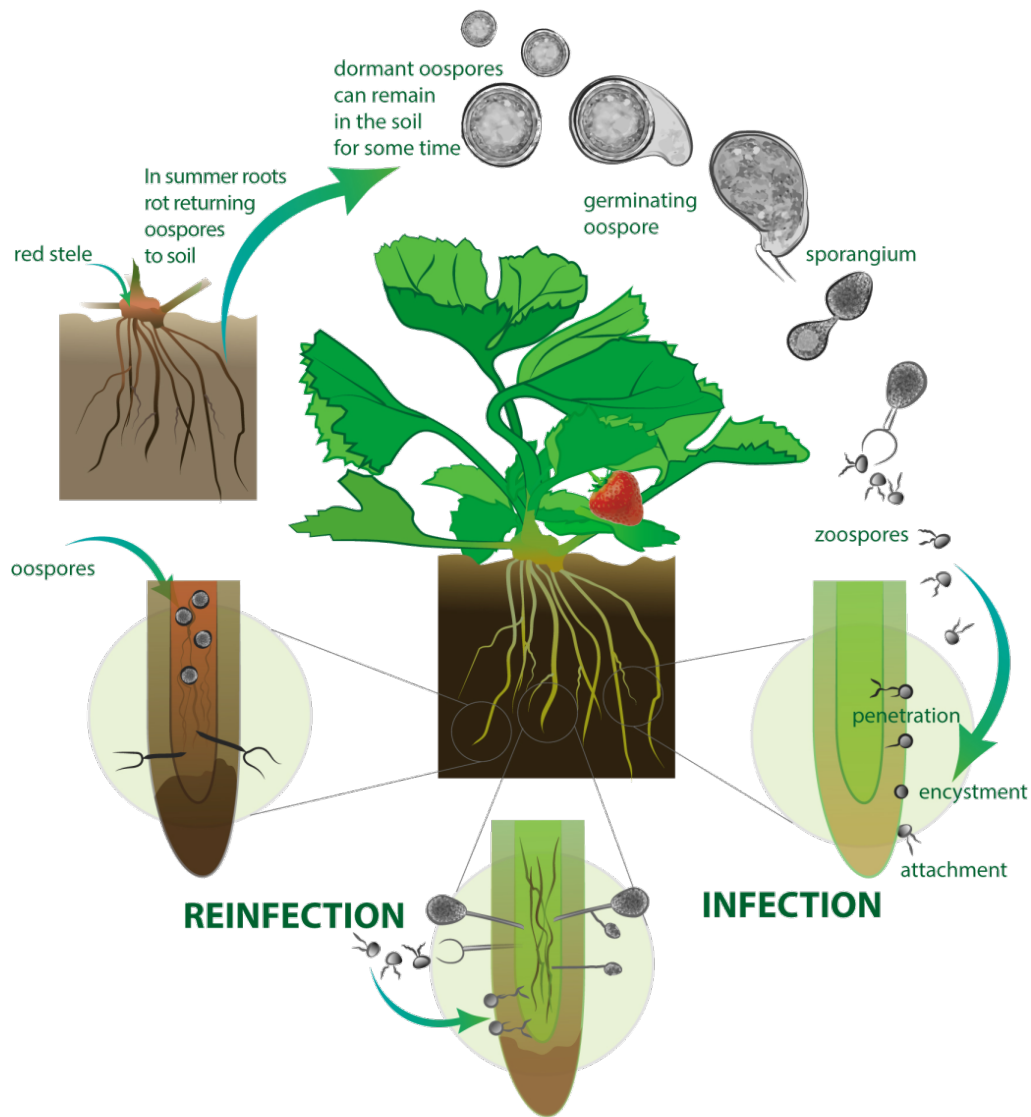


**Fig. 1.5: Expanded phylogeny for clade 7 *Phytophthora* species based on nuclear genetic markers.** (Figure taken from Yang *et al.*, 2017).

*P. fragariae* is pathogenic on strawberry, causing the disease known as red core root rot (also known as red stele root rot) and was first reported in Lanarkshire, Scotland in 1920, where a thriving strawberry industry was transformed into 'a skeleton of its former self' (Alcock and Howells, 1936). The disease is now present in the majority of strawberry growing regions, including: Australia, Canada, Japan, Lebanon, Russia, the U.S.A. and the majority of Europe except the southern Mediterranean regions (van de Weg, 1997a; EFSA Panel on Plant Health (PLH), 2014). This pathogen is currently listed by the European and Mediterranean Plant Protection Organisation (EPPO) as an A2 pest and recommended to member states that it be treated as a quarantine pest. In most member states, this results in a policy of mandatory reporting by growers and the prohibition of strawberry growth in any affected fields (van de Weg, 1997a; EPPO, 2018). The initial identification of this pathogen incorrectly identified it as a fungus (see section 1.3 for assignment of oomycetes as stramenopiles rather than fungi) and established the name *P. fragariae* (Hickman, 1941). Later, following investigations of a *Phytophthora* causing disease on raspberry (*Rubus idaeus*), these organisms were classified as *Phytophthora fragariae* var. *fragariae* and *Phytophthora fragariae* var. *rubi* as only a small number of phenotypic and electrophoretic differences were observed between isolates of both organisms (Wilcox *et al.*, 1993). However, more recently the analysis of isozymes and cytochrome oxidase sequences lead to the separation of the two organisms into the separate species of *P. fragariae* and *P. rubi* (Man in't Veld, 2007). This was recently confirmed through population genetics analyses based on whole genome sequencing data (Tabima *et al.*, 2018).

*P. fragariae* attacks the roots of strawberry (*F. × ananassa*) plants, leading to a decline in vigor and eventual plant collapse, particularly under drought conditions (Hickman, 1941). The oospores of the pathogen are generated in the roots of infected plants. During the summer, these roots rot and release the oospores into the soil, where they can remain viable for several decades. Under favourable conditions, these oospores germinate and produce zoospore filled sporangia. These sporangia subsequently burst, releasing the zoospores into the soil water which the zoospores move through until they reach a plant root. Upon reaching a root, the zoospores encyst and penetrate the root. Following establishment,

sporangia and oospores are produced. The infection process damages the root, discolouring it to produce the characteristic 'red core' symptom. As the roots are damaged, they rot away from the tips, producing the 'rat's tails' symptom (EFSA Panel on Plant Health (PLH), 2014; Fig. 1.6).



**Fig. 1.6: The infection life cycle of *Phytophthora fragariae* on *Fragaria x ananassa* plants.** A previously clean field is infested through the introduction of an infected plant or infested soil. Following root rot, oospores are released into the soil. When conditions are favourable, these germinate to produce zoospore filled sporangia. These burst and zoospores move through soil water to uninfected plant roots. The zoospores encyst and penetrate the root tip where they establish and produce sporangia and oospores. This process damages the root leading to a discoloured root core, the 'red core' symptom, and eventual root rot, producing the 'rat's tails' symptom (Image courtesy of Penny Greaves).

Following the initial identification of *P. fragariae*, new varieties were bred which showed resistance to the disease. However, in the 1940s previously resistant varieties in the breeding programme at Auchincruive, Ayrshire, Scotland, U.K. showed symptoms of red core disease and the existence of physiological races in *P. fragariae* was proposed (Reid, 1948). Following this, two races were identified in the U.S.A. (Scott *et al.*, 1950). Further analysis of isolates from Canada, the U.K. and New Zealand through pathogenicity tests on various cultivars of *F. × ananassa* plants revealed at least twelve races distributed both worldwide (Hickman, 1962) and within the U. K. (Montgomerie, 1964). These results were later refined to eleven races in the U.K. and twelve when including isolates from the U.S.A. and Canada (Montgomerie, 1967). Later work disputed the validity of these races due to the differential *F. × ananassa* cultivars used, additionally microscopic examination of root tissue improved the assessment of plant resistance (van de Weg *et al.*, 1996). This method allowed the improvement of the differential series of cultivars used for pathogenicity testing (van de Weg *et al.*, 1997). This resulted in the proposal of a gene-for-gene model of resistance, as described in section 1.2. Initially, this model consisted of five races and also applied to isolates from the U.K. and Germany (van de Weg, 1997b). The model has since been updated with additional data to comprise eleven hypothesised resistance genes (W. E. van de Weg, personal communication).

Control of *P. fragariae* is important for commercial production of strawberry fruit due to the devastating disease it causes. There are several cultural control methods that can be deployed, briefly these are: selection of the growing site and growth medium to avoid the presence of potentially viable oospores, ensure good drainage and avoid heavy soils as the zoospores require the presence of soil water to infect new hosts, preparation of the site prior to planting to break up compacted layers or using mulched raised beds. Additionally, plants can be sourced from schemes such as the DEFRA plant health propagation scheme (PHPS) to ensure healthy, disease-free material is planted. It is also recommended to ensure plants can establish strong roots by keeping plants moist prior to planting and ensuring adequate irrigation, without causing water logging (Perry and Raffle, 2004).

Chemical control was historically performed through soil fumigation to destroy viable oospores, though the efficiency of this is questionable. Within the European Union, this practise is banned (EFSA Panel on Plant Health (PLH), 2014). The fungicide Fenomenal™ (Bayer CropScience Ltd.) is recommended for use as a preventative agent against red core through the inhibition of mitochondrial respiration, spore germination, mycelial development and sporulation. Paraat (BASF) may have some control action against *P. fragariae*, though it is not specifically recommended for use against the pathogen (A. Berrie, personal communication, 2015).

#### **1.4 Developments in genome and transcriptome sequencing technologies**

The era of genome sequencing began with the sequencing of the 5.4 kbp bacteriophage  $\phi$ X174 as the first genome to be sequenced in 1977 (Sanger *et al.*, 1977a). This utilised 'Sanger sequencing' technology, briefly this method relies on polymerase chain reaction (PCR) termination with dideoxynucleotides and visualisation on agarose gels (Sanger *et al.*, 1977b). The throughput of this method was improved in the 1990s by removing the need for the use of agarose gels with capillary sequencing machines, leading to the sequencing of the first eukaryotic genome, that of the model yeast *Saccharomyces cerevisiae* (Goffeau *et al.*, 1996). Genome sequencing was revolutionised by the development of what was then termed next generation sequencing (NGS), but now tends to be referred to as second-generation sequencing (SGS) technologies. These technologies improved the read-length, throughput and cost per base, superseding even Moore's law, a common descriptor of the rapid decrease of the cost of computing power, halving every two years. For instance, the initial sequencing of the human genome between 1990 and 2001 cost approximately \$3 billion, in 2013 the cost to sequence a complete human genome had plummeted to around \$5,000 (Hayden, 2014). A commonly used SGS method is based on Illumina sequencing by synthesis chemistry and is available as desktop devices such as the MiSeq. This machine can generate up to 8 Gb of data with read lengths of up to 250 bp for low cost. It is also possible to sequence RNA by these methods, with the addition of a reverse transcription step prior to the preparation of genome sequencing libraries



<https://emea.illumina.com/techniques/sequencing.html>,

last accessed 03/01/2019).

Recently, long read technologies have been developed to allow for the improvement of *de novo* assembly contiguity. These technologies are referred to as third-generation sequencing (TGS) technologies. There are currently two major technologies in this area: Pacific Biosciences (PacBio) single molecule real time (SMRT) sequencing developed by Pacific Biosciences and nanopore sequencing developed by Oxford Nanopore Technologies (ONT). PacBio SMRT sequencing captures sequence information during the replication of template DNA created by ligating adaptors to both end of the double-stranded DNA. A DNA polymerase can then begin reading from either adaptor and add fluorescent-labeled nucleotides, whose emission spectra is detected. This produces a 'movie' of light pulses which is decoded to produce the nucleotide sequence. As this is recorded in real time, kinetic variation can be used to infer base modifications such as methylation. However, this technology does show a higher per base error rate than shorter read technologies (Rhoads and Au, 2015). An alternative method that does not rely on sequencing by synthesis has been developed by ONT. This technology utilises nanopores embedded in a membrane over an electrical detection grid. This detects DNA bases by changes in the ionic current as the molecule passes through the pore with the aid of a motor protein (Clarke *et al.*, 2009). Accuracy can be improved through the sequencing of both the template strand and its complement strand (termed 2D reads) as opposed to just the template strand (termed 1D reads). Similar to PacBio sequencing however, this technology displays a higher error rate than shorter read technologies (Goodwin *et al.*, 2015).

### **1.5 Aims and Objectives**

This project focuses on improving the understanding of the genomic basis of the gene-for-gene model for resistance of *F. × ananassa* plants to the devastating plant pathogen *P. fragariae* and to investigate the population dynamics of this pathogen. These investigations will aid in the breeding of resistant cultivars as

well as informing long-term control strategies for this economically important pathogen.

Specifically, this will be addressed by investigating different questions:

- What are the genes that are detected by the resistance genes in the *F. × ananassa* plants and result in the resistance to infection by this pathogen?
- What level of diversity is there within the population of *P. fragariae*?
- Does the population dynamics of *P. fragariae* provide any hints about the future evolutionary direction of this pathogen?

Phenotyping studies were first performed in order to confirm the pathogenicity races of a selection of the isolates used in this study. A collection of isolates of both *P. fragariae* and *P. rubi* were then whole genome sequenced in order to provide a selection of assemblies. Additionally, a novel reference assembly was generated for *P. fragariae* using long-read PacBio SMRT sequencing (Chapter 3). Further to this, *ab initio* gene prediction was performed using RNA-Seq data from a novel method designed in this study. The patterns of expression of these genes were investigated and assessments of copy number variation and presence/absence variation was conducted to identify candidate avirulence genes (Chapter 4). Additionally, a novel panel of variant sites was identified, this allowed for the conducting of population genetics analyses. Finally, RNA-Seq of additional isolates provided additional candidate avirulence genes (Chapter 5).

## **Chapter 2: General Materials and Methods**

### **2.1 Sources of isolates of *Phytophthora fragariae* and *Phytophthora rubi***

A selection of ten isolates of *Phytophthora fragariae* were sourced from the Atlantic Food and Horticulture Research Centre (AFHRC), Nova Scotia, Canada. An additional isolate of *P. fragariae*, SCRP245, was sourced from the James Hutton Institute (JHI), Dundee, Scotland. All three *Phytophthora rubi* isolates were also sourced from the JHI (Table 2.1).

**Table 2.1: Isolation details and race structure of isolates of *Phytophthora fragariae* and *Phytophthora rubi* used in this study**

Species	Isolate	Location Isolated	Date Isolated	Isolated by	Pathogenicity Race
<i>Phytophthora fragariae</i>	A4	Unknown	17/12/2001	Unknown	US4
<i>Phytophthora fragariae</i>	BC-1	Commercial Strawberry Field, Delta, BC, Canada	05/01/2007	N. L. Nickerson	CA1 (UK1)
<i>Phytophthora fragariae</i>	BC-16	Commercial Strawberry Field, Ladner, BC, Canada	05/01/2007	N. L. Nickerson	CA3 (UK2)
<i>Phytophthora fragariae</i>	BC-23	Commercial Strawberry Field, Aldergrove, BC, Canada	30/01/2012	N. L. Nickerson	CA5
<i>Phytophthora fragariae</i>	NOV-5	Commercial Strawberry Field, Nine Mile River, Hants County, NS, Canada	17/12/2001	N. L. Nickerson	CA1
<i>Phytophthora fragariae</i>	NOV-9	Commercial Strawberry Field, Billtown, Kings County, NS, Canada	05/01/2007	N. L. Nickerson	CA2 (UK3)
<i>Phytophthora fragariae</i>	NOV-27	Commercial Strawberry Field, Cambridge Station, Kings County, NS, Canada	19/12/2001	N. L. Nickerson	CA2
<i>Phytophthora fragariae</i>	NOV-71	Commercial Strawberry Field, Middle Clyde River, Shelburne County, NS, Canada	05/01/2007	N. L. Nickerson	CA2
<i>Phytophthora fragariae</i>	NOV-77	Commercial Strawberry Field, Nine Mile River, Hants County, NS, Canada	30/01/2012	N. L. Nickerson	CA5
<i>Phytophthora fragariae</i>	ONT-3	Commercial Strawberry Field, Fort Erie, ON, Canada	05/01/2007	N. L. Nickerson	CA4
<i>Phytophthora fragariae</i>	SCR245	Kent, England, United Kingdom	1945	Unknown	Unknown
<i>Phytophthora rubi</i>	SCR249	Germany	1985	Unknown	1
<i>Phytophthora rubi</i>	SCR324	Scotland, United Kingdom	1991	Unknown	1
<i>Phytophthora rubi</i>	SCR333	Scotland, United Kingdom	1985	Unknown	3

## 2.2 Production of media for isolate and micropropagated plant growth

Kidney bean agar (KBA) was produced by a method similar to Maas (1972). Briefly, 35 g of dried red kidney beans (*Phaseolus vulgaris*) were soaked in 500 mL dH<sub>2</sub>O overnight and homogenised with a handheld blender. The solution was then autoclaved for 20 minutes at 121°C alongside 10 g 1.5% agar in 500 mL dH<sub>2</sub>O. The solutions were then mixed prior to dispensing approximately 20 mL into 90 mm triple vent petri dishes (Thermo Scientific, Waltham, MA, U.S.A.). Pea broth agar (PBA) was produced in a method similar to Campbell *et al.* (1989) but with the addition of 10 g/L sucrose (Fisher Scientific, Loughborough, U.K.) alongside the addition of 10 g/L 1.5% agar (Fisher Scientific, Fairlawns, NJ, U.S.A.). The resulting solution was then autoclaved at 121°C for 20 minutes prior to dispensing approximately 20 mL into petri dishes for plates or into 15 mL falcon tubes (Fisher Scientific, Loughborough, U.K.) for slopes.

Broad Bean (*Vicia faba*) Agar (BBA), Green Bean (*P. vulgaris*) Agar (GBA), Soybean (*Glycine max*) Agar (SBA), Edamame Bean (*G. max*) Agar (EBA) and Mung Bean (*Vigna radiata*) Shoot Agar (MBSA) were produced using the following recipe. 120 g of frozen beans (except for MBSA when fresh bean sprouts were used) were placed in 500 mL of dH<sub>2</sub>O and homogenised using a handheld blender. Following this, another 500 mL of dH<sub>2</sub>O was added along with 10 g of 1.5% agar and the solution was autoclaved at 121°C for 20 minutes. Strawberry (*Fragaria × ananassa*) Material Agar was made using a similar method. Whole micropropagated strawberry plants of the 'Hapil' cultivar were added instead of frozen beans. Strawberry extract agar was produced in a similar way to PBA, except for the substitution of frozen peas with micropropagated strawberry plants of the 'Hapil' cultivar. Oatmeal agar was produced by dissolving 12 g of 1.5% agar in 400 mL of dH<sub>2</sub>O in a microwave oven. 60 g of large flake quaker oats were added to 600 mL dH<sub>2</sub>O and homogenised using a handheld blender. The two solutions were then combined and mixed prior to autoclaving at 121°C for 15 minutes to sterilise before dispensing into 15 mL falcon tubes for slopes or petri dishes for plates.

Pea Broth (PB) for *in vitro* mycelium growth was produced in a similar method to Campbell *et al.* (1989) with the addition of 10 g/L sucrose (Fisher Scientific, Loughborough, U.K.) prior to dispensing into 90 mm triple vent petri dishes (Thermo Scientific, Waltham, MA, U.S.A.).

ATS (*Arabidopsis thaliana* salts) media was made of the following components, as described by Taylor *et al.* (2016): 4.5 g of Phytagel (Sigma-Aldrich, Gillingham, U.K.), 5 mL of 1 M KNO<sub>3</sub> (Fisher Scientific, Loughborough, U.K.), 2.5 mL of 1 M KPO<sub>4</sub> (Sigma Chemical Co., St. Louis, MO, U.S.A.), 3 mL of 1 M MgSO<sub>4</sub> (Fisher Scientific, Loughborough, U.K.), 3 mL of 1 M Ca(NO<sub>3</sub>)<sub>2</sub> (BDH laboratory reagents, Poole, U.K.), 2.5 mL of 20 mM Fe-EDTA; which consisted of 7.34 g of EDFS (Ethylenediaminetetraacetic acid iron(III) sodium salt; Sigma-Aldrich, Gillingham, U.K.) in 1 L of dH<sub>2</sub>O, 1 mL of a Microelements mix; consisting of 70 mM H<sub>3</sub>BO<sub>3</sub> (Sigma Chemical Co., St. Louis, MO, U.S.A.), 14 mM MnCl<sub>2</sub> (Sigma Chemical Co., St. Louis, MO, U.S.A.), 0.5 mM CuSO<sub>4</sub> (Fisher Scientific, Loughborough, U.K.), 1 mM ZnSO<sub>4</sub> (Fisher Scientific, Loughborough, U.K.), 0.2 mM Na<sub>2</sub>MoO<sub>4</sub> (Sigma Chemical Co., St. Louis, MO, U.S.A.), 10 mM NaCl (Fisher Scientific, Loughborough, U.K.) and 0.01 mM CoCl<sub>2</sub> (Sigma Chemical Co., St. Louis, MO, U.S.A.). This mixture was made up to 200 mL with sterile, distilled water (Elga, High Wycombe, U.K.) before being stirred with a magnetic stirrer bar in order to dissolve the Phytagel. The stirrer bar was then removed and the mixture was made up to 1 L and autoclaved at 121°C for 20 minutes. The mixture was then dispensed into 120 mm x 120 mm or 100 mm x 100 mm square plates depending on availability (Gosselin, Borre, France).

### **2.3 Isolate maintenance, subculturing and routine growth**

Medium term storage of isolates was performed by growing the isolates on PBA and oatmeal agar slopes, which were subsequently covered with autoclaved liquid paraffin (Fisher Scientific, Loughborough, U.K.), sealed with Parafilm® (Bemis, Oshkosh, WI, U.S.A.) and stored at 4°C. Additionally, plugs taken from PBA plates with a small cork borer to cut out a large number of plugs for storage in dH<sub>2</sub>O in 50 mL Falcon tubes (Fisher Scientific, Loughborough, U.K.) at 4°C. Both slopes and plugs were stored in triplicate in two separate locations.

Longer term storage was performed using overgrown plates of *P. fragariae* or *P. rubi* on KBA. Discs of media were cut from these plates with a small, sterilised cork borer and five plugs were placed in a 5% dimethyl sulfoxide (DMSO) solution in 2 mL cryotubes (Alpha Laboratories Ltd. Eastleigh, U.K.). These were then stored in a liquid nitrogen cooled cryostore (Tooley, 1988; Houseknecht *et al.*, 2012).

All work using *P. fragariae* and *P. rubi* was carried out in a Tri-MAT Class-II microbiological safety cabinet (Contained Air Solutions, Manchester, U.K.). Routine subculturing was performed by growing isolates from medium term storage on 90 mm triple vent petri dishes (Thermo Scientific, Waltham, MA, U.S.A.) containing PBA and sealed with Parafilm®. Plates were then grown at 20°C, in the dark, in a Panasonic MIR-254 cooled incubator (Osaka, Japan) for between seven and fourteen days, before the mycelium had reached the edge of the petri dish. Mycelium of *P. fragariae* was subcultured from plates grown out of medium-term storage by transferring two pieces of between 1 and 4 mm<sup>2</sup> onto fresh plates of KBA in 90 mm triple vent petri dishes, sealed with parafilm®.

*In vitro* mycelial growth of *P. fragariae* and *P. rubi* was performed on PB plates. These plates were inoculated with approximately five pieces of small (between 1 and 4 mm<sup>2</sup>) pieces of mycelium and media, sealed with Parafilm®. These were grown at 20°C for four or five days in constant darkness.

#### **2.4 Source of plant material**

Micropropagated plants were propagated by GenTech (Dundee, U.K.) and supplied as *in vitro* cultures. The plants were subbed on to ATS plates (see section 2.2) which had half of the media removed using a flame sterilised scalpel, with flame sterilised metal tweezers. These plates were then sealed with Sellotape (Winsford, U.K.) and grown in a controlled environment room at a constant 20°C with 16 hours of fluorescent light per day. Plants were again subcultured either once every three weeks, a maximum of three times, or on the day prior to usage in *in vitro* inoculation experiments.

For pot-based pathogenicity experiments, runners were taken from stock plants maintained as part of the NIAB EMR collection. Runners from these plants were pinned down in a sieved mixture of peat-based compost and sand in a 1:1 volume, autoclaved at 121°C for four hours (van de Weg *et al.* 1996). Runners were then rooted for 18-21 days using sterilised pins. After rooting, plants were cut free from the mother plant.

## **2.5 Ordering of primers**

All primers were ordered from Integrated DNA Technologies (Leuven, Belgium) as 25 nM DNA Oligos with standard desalting purification. Primers were delivered as lyophilised DNA and resuspended to a concentration of 100 µM with dH<sub>2</sub>O and stored at -20°C.

## **2.6 DNA extraction for sequencing**

DNA extraction for Illumina sequencing was performed by Dr. Charlotte F. Nellist on 300 mg of freeze-dried mycelium using a Macherey-Nagel NucleoSpin® Plant II kit (Fisher Scientific, Loughborough, U.K.). The manufacturer's protocol was modified by doubling the volume of lysis buffer PL1, increasing the incubation period after addition of RNase A by five minutes, using the alternative method of pre-centrifuging the crude lysate for five minutes at 11,000x *g* before loading onto the NucleoSpin® Filter column, doubling the volume of buffer PC to adjust the DNA binding conditions and performing two consecutive elution steps with 35 µL of buffer PE pre-heated to 70°C.

DNA extraction for PacBio Single Molecule Real Time (SMRT) sequencing and Oxford Nanopore Technologies (ONT) sequencing was also performed by Dr. Charlotte F. Nellist with the Genomic-Tip DNA kit (QIAGEN Inc., Venlo, The Netherlands) following the manufacturer's instructions.



## 2.7 RNA extraction for sequencing and qRT-PCR

### 2.7.1 Extraction from inoculated plant tissue

Extractions from the root tissue of inoculated plants were performed using a similar method to Yu *et al.* (2012). Briefly, an RNA extraction buffer was created using RNase free stocks of 1 M Tris-HCl pH 8.0 (Gibco by Life Technologies, Carlsbad, CA, U.S.A.), 5 M NaCl, 0.5 M EDTA (both from Ambion, Foster City, CA, U.S.A.), dH<sub>2</sub>O (Fisher BioReagents, Fairlawns, NJ, U.S.A.), PVP (BDH, Poole, U.K.) and CTAB (Sigma Chemical Co., St. Louis, MO, U.S.A.), as described in the 3% CTAB<sub>3</sub> protocol in Yu *et al.* (2012). This buffer was heated to 65°C in two 700 µL aliquots, with 10 µL β-mercaptoethanol (Sigma Chemical Co., St. Louis, MO, U.S.A.) added just before extraction. 0.01 g of PVPP (Sigma Chemical Co., St. Louis, MO, U.S.A.) was added per 0.1 g of frozen material before grinding the sample in a mortar and pestle under liquid nitrogen. Prior to extraction, mortars and pestles were cleaned with RNaseZap™ RNase decontamination solution (ThermoFisher Scientific, Waltham, MA, U.S.A.) and baked for 2 hours at 230°C to deactivate any RNase present to avoid degradation of the RNA. The rest of the protocol was performed as described in Yu *et al.* (2012), except that 60 µL DEPC-treated H<sub>2</sub>O (Fisher BioReagents, Fairlawns, NJ, U.S.A.) was used to elute the RNA.

### 2.7.2 Extraction from *in vitro* mycelium

*In vitro* grown mycelium was prepared as described above. Mycelium was dried on a Q100 90 mm filter paper (Fisher Scientific, Loughborough, U.K.) using a Büchner funnel and a Büchner flask attached to a 2016 VacuGene Vacuum Blotting Pump (LKB, Stockholm, Sweden). Plates were unsealed by removing the parafilm and poured onto the filter paper. To ensure all the mycelium was collected, dH<sub>2</sub>O was used to rinse out the petri dish. Once all the media had passed through the filter, the vacuum pump was disengaged and more dH<sub>2</sub>O was added to the funnel. Sterilised tweezers were used to gather the mycelium together before re engaging the vacuum pump to dry the mycelium. Dried

mycelium was transferred to a nuclease free 2 mL Eppendorf tube (Stevenage, U.K.) and flash-frozen in liquid nitrogen.

Total RNA was extracted using the QIAGEN RNEasy Plant Mini kit (Venlo, The Netherlands), using buffer RLC to avoid the risk of RLT causing solidification of secondary metabolites and using the optional additional spin to remove residual ethanol (C<sub>2</sub>H<sub>5</sub>OH). Otherwise, the protocol was performed following the manufacturer's instructions.

### **2.7.3 Assessment of RNA quality**

Extracted RNA was initially assessed on a NanoDrop 1000 spectrophotometer (Thermo Scientific, Waltham, MA, U.S.A.) using the 260/280 and 260/230 ratios as initial indicators of quality and the presence of contaminating extraction salts, proteins and carbohydrates. RNA concentration was then quantified using a Qubit (Thermo Fisher Scientific, Waltham, MA, U.S.A.) with the high sensitivity RNA kit. Finally, the extracted RNA was analysed on a TapeStation 4200 (Agilent Genomics, Santa Clara, CA, U.S.A.) in order to assess RNA integrity via the RIN<sup>e</sup> metric.

## **2.8 Sequencing of DNA and RNA**

### **2.8.1 Illumina sequencing of DNA**

Illumina sequencing was performed by Dr. Helen J. Bates. PCR-free libraries were prepared by sonication of the DNA in a Covaris microTube-50 (Woburn, MA, U.S.A.) on the Covaris 550 bp program. The resulting sample was run on a Fragment Analyzer 2538 (Advanced Analytical Technologies Ltd., Ankeny, IA, U.S.A.) to ensure sonication had succeeded, with a target peak of approximately 500-600 bp. Samples below 2 µg in 30 µL were concentrated with AMPureXP beads (Beckman Coulter, Brea, CA, U.S.A.) as per the manufacturer's instructions with the following modifications. An equal volume of AMPure beads were added rather than 1.8x volume, the reaction tubes were placed on the magnetic base for five minutes rather than the recommended two, 80% ethanol

(C<sub>2</sub>H<sub>5</sub>OH) was used for washes rather than 70%, the beads were left to air dry for three minutes at room temperature, 30 µL of elution buffer was added and left at room temperature for five minutes, the time placed on the magnetic base for final elution was increased to five minutes.

All samples were then analysed on a BluePippin (Sage Science, Beverly, MA, U.S.A.) for size selection purposes on the 550-broad program, which provided a size range of 450-650 bp. Quantification was then performed using the dsDNA BR (Broad Range) Assay Kit for the Qubit II Fluorometer (ThermoFisher Scientific, Waltham, MA, U.S.A.). A minimum of 100 ng, with 200 ng preferred, was required to continue with the library preparation. End repair was performed using the NEBNext End Repair Module (New England Biolabs, Ipswich, MA, U.S.A.) following the manufacturer's instructions. This was again cleaned using AMPureXP beads as above, with 26 µL of resuspension buffer used for the final step and again quantified using the Qubit. dA-tailing was then performed using the NEBNext dA-Tailing module (New England Biolabs, Ipswich, MA, U.S.A.) as follows: 3 µL of NEBNext dA-Tailing Reaction Buffer and 2 µL of Klenow Fragments were added to 25 µL of DNA. This mixture was then incubated at 37°C for 30 minutes.

LT Illumina adapters (San Diego, CA, U.S.A.) were then ligated to the samples. First, the 15 µM stock was diluted 1:1 in 10 mM Tris-HCl (pH 8.0) to provide a 7.5 µM stock. 0.3 µL of this was added per 100 ng of DNA and mixed by vortexing. An equal volume (approximately 31 µL) of NEB Blunt TA ligase master mix (New England Biolabs, Ipswich, MA, U.S.A.) was added and mixed by vortexing, before incubation at 22°C for 15 minutes. The ligated mixture was then cleaned using AMPureXP beads three times as described above. Firstly, an equal volume (approximately 61 µL) of beads and eluting in 55 µL of resuspension buffer; secondly using 55 µL of beads and 55 µL of resuspension buffer and finally using 55 µL of beads and 22 µL of resuspension buffer. Ligation was confirmed on the Fragment Analyzer 2538 and a library quantification kit (KAPA Biosystems, Wilmington, MA, U.S.A.) or a NEBNext kit (New England Biolabs, Ipswich, MA, U.S.A.) following the manufacturer's instructions, analysed on either a 7500 Real Time PCR System (Applied Biosystems, Foster

City, CA, U.S.A.) or a CFX96™ Real-Time System running CFX Manager version 1.3 (Bio-Rad, Watford, U.K.). Libraries were loaded onto V3 600 cycle kit and run on a MiSeq™ system following the manufacturer's instructions (Illumina, San Diego, CA, U.S.A.).

### **2.8.2 PacBio SMRT sequencing of DNA**

PacBio SMRT sequencing was performed at the Earlham Institute (Norwich, U.K.) on a PacBio RS II machine.

### **2.8.3 Oxford Nanopore Technologies sequencing**

Oxford Nanopore Technologies (ONT) sequencing was performed by Kieran J. Housley and Dr. Helen J. Bates using the SQK-LSK108 kit. Briefly, DNA was analysed on a TapeStation 4200 (Agilent Genomics, Santa Clara, CA, U.S.A.) to test for long length DNA fragments. Quantification was then performed using the dsDNA BR (Broad Range) Assay Kit for the Qubit II Fluorometer (ThermoFisher Scientific, Waltham, MA, U.S.A.). The DNA was then run on a BluePippin (Sage Science, Beverly, MA, U.S.A.) for size selection with the High-Pass filtering > 10 kB unstranded program. DNA was again quantified by Qubit. DNA repair was performed using the NEBNext FFPE DNA Repair Kit (New England Biolabs, Ipswich, MA, U.S.A.) for fifteen minutes at room temperature. DNA was again quantified by Qubit. The repaired DNA was re-eluted with a 45 minute incubation at 37°C. This was followed by end repair with the NEBNext Ultra™ II End Repair kit (New England Biolabs, Ipswich, MA, U.S.A.). DNA was again quantified by Qubit. Sequencing adapters were then ligated and the mixture was cleaned using AMPureXP beads at 0.6x volume. The library was then loaded onto the FAH69834 FLO-MIN106 flow cell and run on the GridION for approximately 28 hours.

#### **2.8.4 RNA sequencing**

RNA extraction procedures are described in the relevant results chapters. All RNA sequencing was performed by Novogene (Hong Kong, Special Administrative Region, China). Extracted RNA was prepared for sequencing by RT-PCR and run on an Illumina HiSeq 4000.

#### **2.9 Availability of code and data**

All computer code was uploaded to public GitHub repositories at the following locations: [https://github.com/harrisonlab/phytophthora\\_fragariae](https://github.com/harrisonlab/phytophthora_fragariae) and [https://github.com/harrisonlab/phytophthora\\_rubi](https://github.com/harrisonlab/phytophthora_rubi). All raw sequencing reads were uploaded to the sequencing read archive (SRA), for details of SRA codes, see the relevant results chapters. All assembled genomes with annotations were also uploaded to National Centre for Biotechnology Information (NCBI) GenBank in the BioProject PRJNA396163. Non-annotated genomes were also uploaded to NCBI GenBank in the BioProject PRJNA488213. These will be made publicly available upon publication of results in a peer reviewed journal.

#### **2.10 Computing resources**

Unless otherwise specified, all computer code was run on the NIAB EMR Beowulf cluster. This cluster system ran Linux Debian 7 (wheezy), kernel version 3.2.0-5. Unless otherwise detailed, R was version 3.2.1 (R core team, 2016) and Python was version 2.7.3 (Python Software Foundation, 2017).

## **Chapter 3: Characterisation of *Phytophthora fragariae* races and genome assembly of isolates of *P. fragariae* and *Phytophthora rubi***

### **3.1 Introduction**

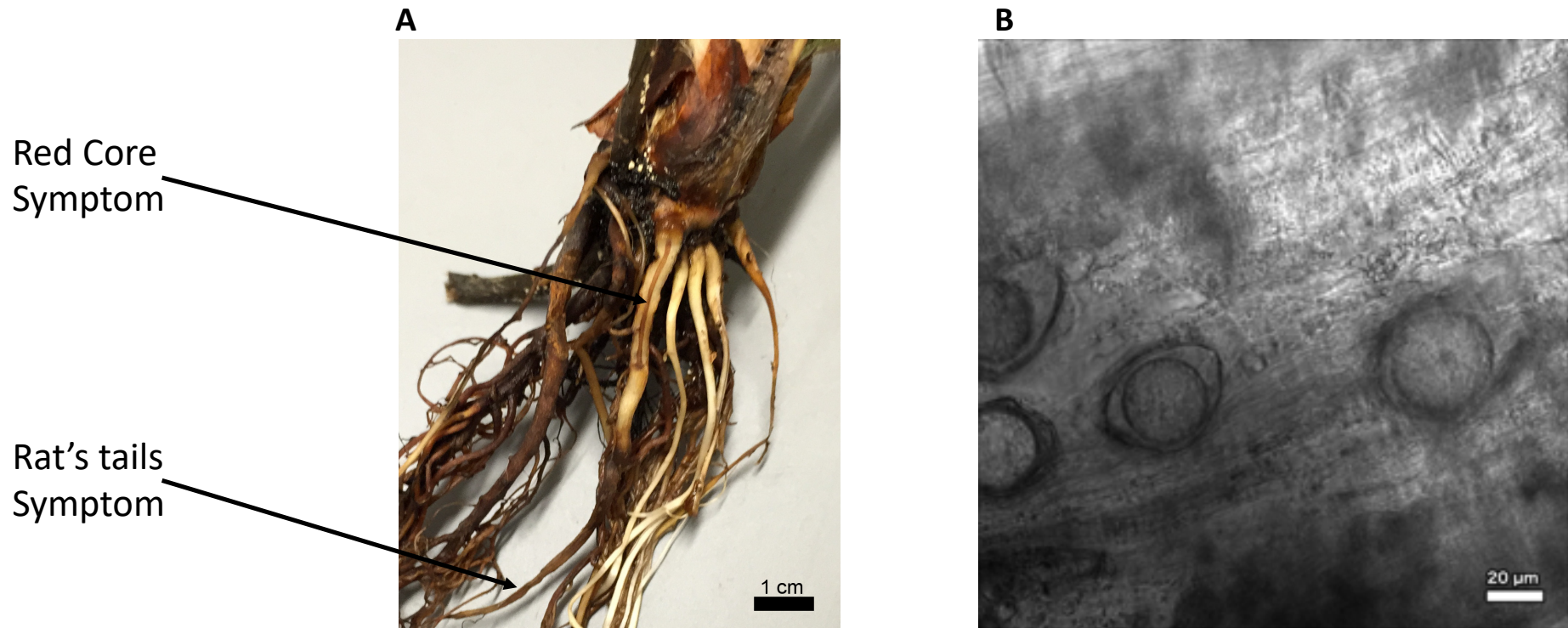
#### **3.1.1 The NIAB EMR collection of *Phytophthora fragariae* and *Phytophthora rubi***

NIAB EMR maintains a collection of isolates of the closely related plant pathogens *Phytophthora fragariae* and *Phytophthora rubi*, which infect strawberry (*Fragaria × ananassa*) and raspberry (*Rubus idaeus*), respectively. The majority of the isolates used in this study were acquired from collaborators (detailed in Chapter 2). These isolates represent a variety of different races of *P. fragariae* and *P. rubi* from different geographical locations, predominantly Canada, though some isolates were isolated in the U.S.A., the U.K. and Germany. This collection represents a valuable resource for investigating and understanding the genetic determination of variation in pathogenicity between the isolates. This investigation focuses on *P. fragariae* and pathogenicity on strawberry, with *P. rubi* acting as an outgroup.

#### **3.1.2 Assessment of the pathogenicity of *Phytophthora fragariae* and current knowledge of the pathogenicity races of the collection**

*P. fragariae* virulence on different *F. × ananassa* cultivars has previously been explained via a gene-for-gene model of resistance (van de Weg, 1997b; Chapter 1). Work conducted prior to the commencement of this study assigned a U.K. pathogenicity race to three of the isolates in the collection: BC-1 (U.K. race 1), BC-16 (U.K. race 2) and NOV-9 (U.K. race 3; C. Nellist, unpublished). These races were determined using a method based on that described by van de Weg *et al.* (1996), where plants were inoculated and destructively tested for below ground symptoms after six weeks. The key symptoms for this pathogen include: the rotting of the roots from the tips, known as 'rat's tails' and the discolouration

of the core tissue, known as the 'red core' symptom (Fig. 3.1A). Additionally, the method of van de Weg *et al.* (1996) uses the presence of oospores in infected tissue as a key diagnostic criterion (Fig. 3.1B). The races of other isolates were provided by collaborators (Chapter 2), however these were mostly in the Canadian race scheme with one in the U.S.A. race scheme and so would require further testing in order to convert to the equivalent U.K. race.



**Fig. 3.1: Below ground symptoms of *Fragaria x ananassa* plants infected with *Phytophthora fragariae*.** **A:** Photograph of the entire root system of a *Fragaria x ananassa* plant infected by *Phytophthora fragariae*. The red core symptoms and rat's tails symptoms are denoted by arrows (Image courtesy of C. F. Nellist). **B:** Light micrograph of root tissue from a *F. x ananassa* plant infected with *P. fragariae* displaying the presence of oospores of *P. fragariae* (Image courtesy of C. F. Nellist).



### 3.1.3 Genome assembly of *Phytophthora* species

Whole genome sequencing (WGS) is a powerful tool for improving the understanding of the genomic nature of a variety of organisms. Within the *Phytophthora* genus, this was first performed for the clade 7b species *Phytophthora sojae*, a pathogen of Soybean (*Glycine max*), and the clade 8c species *Phytophthora ramorum*, the causative agent of sudden oak death. For *P. sojae*, the authors produced a genome assembly of 78.0 Mb in 5,577 contigs and for *P. ramorum* the authors produced a genome assembly of 54.4 Mb in 7,588 contigs, despite predicted genome sizes of 95 Mbp for *P. sojae* and 65 Mbp for *P. ramorum* (Tyler *et al.*, 2006). Interestingly, the largest genome assembled within this genus is that of *Phytophthora infestans*, at approximately 240 Mb. This clade 1c species is notorious as the cause of late blight on potato (*Solanum tuberosum*) and as the disease involved in the Irish potato famine at the end of the 19<sup>th</sup> century. The genome assembly produced was 229 Mb in size in 18,288 contigs. Interestingly, the authors demonstrated a large proportion (approximately 74%) of the sequence was repetitive and they suggested that this indicated faster evolving regions of the genome (Haas *et al.*, 2009).

When the analyses in this thesis were conducted, there were two publicly available genome assemblies of *P. fragariae*. The oldest of these sequencing experiments resulted in 105x coverage and was assembled into a 73.68 Mb assembly in 1,616 scaffolds, however, an N50 value was not reported (Gao *et al.*, 2015). The more recent genome assembly was sequenced to 92x coverage and assembled into a 76 Mb assembly in 8,511 scaffolds with an N50 of 16,735 bp (Tabima *et al.*, 2017). Unfortunately, no information on the race of either of these sequenced isolates was reported.

More recently, long read technologies, such as Pacific Biosciences (PacBio) Single Molecule Real-Time (SMRT) sequencing, have been used to improve the contiguity of genome assemblies (Rhoads and Au, 2015). Recently, this has been used to sequence isolates of two races of the clade 1 species *Phytophthora nicotianae*. This produced genome assemblies of 80 Mb and 69 Mb despite the

predicted size of 90 Mb, in 6,142 and 6,116 contigs with N50 values of 23 kb and 22 kb respectively (Liu *et al.*, 2016). This demonstrates the value of long-read sequencing for producing improved reference genome assemblies of increased contiguity in *Phytophthora* spp..

### 3.2 Aims

Previously, the races of the *P. fragariae* isolates BC-1, BC-16 and NOV-9 were determined as being UK1, UK2 and UK3, respectively, based on pathogenicity experiments on strawberry (*F. × ananassa*) plants (C. F. Nellist, unpublished). The collection of isolates sequenced consisted of eleven isolates of *P. fragariae* and three isolates of *P. rubi*. The analyses conducted in this chapter aimed to assess the viability of converting from Canadian and U.S.A. races to U.K. races and to *de novo* assemble genomes of the isolates in the NIAB EMR collection.

The specific aims of this chapter were as follows:

1. Test all previously untested isolates listed as being race CA1, CA2, CA3 or US4 for their pathogenicity on host plants containing various different combinations of resistance genes.
2. Sequence and *de novo* assemble the genomes of each of the eleven *P. fragariae* isolates and three *P. rubi* isolates.
3. Create a new reference assembly through *de novo* assembly of long-read PacBio SMRT sequencing technology in an effort to produce an assembly of greater contiguity than the current publicly available assemblies.

### **3.3 Materials and Methods**

#### **3.3.1 Isolate preparation**

Isolates were grown out of medium-term storage and routinely subcultured as described in Chapter 2, section 2.3.

#### **3.3.2 Inoculation procedure**

Inoculation procedures were performed as described in van de Weg *et al.* (1996). Briefly, excised colonies of mycelium and the agar beneath were placed into a blender with ice H<sub>2</sub>O (1 g culture : 1 g ice H<sub>2</sub>O) and blended twice for two seconds. The resulting slurry was transferred to a cooled beaker which was kept on ice during the inoculation procedure. Prior to inoculation, each root system was washed in water, to remove sand and soil. Plants were then inoculated by dipping their roots into the slurry and then re-planted back into FP9 plastic pots (9.0 x 9.0 x 10.0 cm; Soparco, Chaingy, France) containing an autoclaved sand/soil mix (see Chapter 2, section 2.4).

#### **3.3.3 Experimental conditions**

Following inoculation, plants were placed in a walk-in growth chamber with air conditioning under 16 hours of fluorescent light per day at a constant 15°C. Inoculated plants stood in a shallow layer of water (2-7 mm) for the entire experiment and individual pots were watered from above twice a week.

#### **3.3.4 Disease assessment**

Inoculated plants were scored on their above ground symptoms on a weekly basis for the first five weeks on a three-point scale: a score of 1 represented no symptoms, 2 represented an intermediate state of the foliage beginning to collapse and 3 represented total foliar collapse. After six weeks, plants were removed from pots and the roots were rinsed to remove residual sand/soil mixture. Roots were then assessed for the characteristic 'rat's tails' phenotype

alongside red discolouration of the root core when cut using a scalpel (van de Weg *et al.*, 1996). Red core symptoms were considered as sufficient indication of successful infection. Samples for which infection was unclear were visualised under a light microscope for the presence of oospores. This was performed through squash-mounting of the root tissue, where a sample of root was excised and pressed between a microscope slide and cover slip. This sample was then examined at 40x magnification under high light intensity with a Leitz dialux 20 light microscope (Leica Microsystems, Wetzlar, Germany).

### **3.3.5 Preparation of raw data**

MiSeq reads were first assessed using the quality control program fastqc version 0.10.1 (Andrews, 2010) to provide statistics on several key metrics. These included: sequence quality, sequence content, percentage GC content, N content, sequence length distribution, sequence duplication, the presence of overrepresented sequences and the *K*-mer content. Errors and adaptor sequences were then trimmed using fastq-mcf (Aronesty, 2013), with 1,000,000 reads used for subsampling, Illumina PF filtering enabled, a skew percentage cut off of 20%, a 0.01% occurrence threshold before adapter clipping, a quality threshold of 30 and a 5% maximum adapter difference, before assessing the quality of the cleaned reads with fastqc. *K*-mer counting was also performed using *K*-mer counter (KMC) version 2.1.1 (Deorowicz *et al.*, 2013), with a *K*-mer length of 21 and only displaying *K*-mers with an abundance greater than 5, to produce a *K*-mer size distribution and to provide an estimate of genome coverage. Coverage was also assessed by counting the number of nucleotides contained in sequencing reads and dividing this value by the estimated genome size. Quality control and preparation of PacBio SMRT data was performed as part of the sequencing service by the Earlham Institute (Norwich, UK).

### **3.3.6 De novo assembly of genomes**

#### **3.3.6.1 Assembly of Illumina data**

MiSeq reads were assembled into draft genome assemblies using the program SPAdes version 3.11.0 (Bankevich *et al.*, 2012) with *K*-mer sizes of 21, 33, 55, 77, 99 and 127, a Phred quality offset of 33, the careful mode enabled and a read coverage cut off of 10. Contigs smaller than 500 bp were then removed (Fig. 3.2A). Assembly statistics were then collected using Quast version 3.0 with the scaffolds option (Gurevich *et al.*, 2013). Assembly completeness was assessed at this stage by running a CEGMA\_v2 (Core Eukaryotic Genes Mapping Approach) analysis to search for 1,788 genes designated as core eukaryotic genes (Parra *et al.*, 2007), and later a BUSCO (Benchmarking Universal Single-Copy Orthologs) analysis was run using BUSCO version 3.0.1 with the eukaryota\_odb9 database on the assembly to identify 303 genes in the database (Simão *et al.*, 2015). This change of approach also followed the recommendation of the team behind CEGMA following their withdrawal of support.

#### **3.3.6.2 Assembly of PacBio data**

For the PacBio SMRT data, the assembly program Canu version 1.4 (Koren *et al.*, 2017) was initially used with an estimated genome size of 95 Mb, provided from an average of the non-error size estimates by KMC version 2.1.1. Polishing was then performed using the Pilon version 1.17 program (Walker *et al.*, 2014), which first requires an alignment between the assembly and the MiSeq reads to be created using Bowtie 2 version 2.2.6 (Langmead and Salzberg, 2012) and Sequence Alignment/Map tools (SAMtools; Li *et al.*, 2009). Pilon version 1.17 was then run with the changes option enabled to correct single base call errors and erroneous small INDELS caused by the intrinsically higher error rate of SMRT sequencing in comparison to Illumina sequencing (Rhoads and Au, 2015). St. Petersburg genome assembler (SPAdes) version 3.11.0 was also used to perform an assembly of SMRT and Illumina data using *K*-mer sizes of: 21, 33, 55, 77, 99 and 127, a Phred quality offset of 33, the careful mode enabled

and a read coverage cut off of 50 (Bankevich *et al.*, 2012). A custom python script was then used to remove contigs smaller than 500 bp (Armitage, unpublished). Quast version 3.0 was then used with the scaffolds option to collect statistics about the assemblies (Gurevich *et al.*, 2013).

FALCON-Unzip was then used in an attempt to assemble a more contiguous genome than the Canu assembly (Chin *et al.*, 2016). This program was run on NIAB's Triticum high performance computing (HPC) system, a quad socket xeon server, running Debian version 8 (Jessie), kernel version 3.16.0.6. As FALCON-Unzip was still in an experimental state when it was run, the appropriate parameter values were unclear. As such, several iterations of the base FALCON assembler version 0.7+git.7a6ac0d8e8492c64733a997d72a9359e1275bb57 were run, optimising the following parameters to improve completeness and reduce duplication: `length_cutoff`, `length_cutoff_pr`, `falcon_sense_option`, `min_cov`, `max_n_read`, `overlap_filtering_setting`, `min_cov`, `max_cov` and `max_diff`. After every iteration, BUSCO genes were assessed using BUSCO version 3.0.1 with the eukaryota\_odb9 database (Simão *et al.*, 2015) and Quast version 3.0 was run to collect assembly statistics (Gurevich *et al.*, 2013). After assembly, the `bax2bam` function in Pitchfork version 0.0.2 was used to convert the PacBio `bax.h5` files to `bam` files required by the FALCON-Unzip stages (Pacific Biosciences, 2018). The FALCON-Unzip version 0.4.0 stages were then run, followed by the consensus-calling algorithm Quiver, installed via the FALCON-Unzip install (Chin *et al.*, 2013). These programs perform some error correction and create a phased assembly from the primary and accessory haplotigs generated by the initial FALCON assembly. Following assembly, Pilon version 1.17 (Walker *et al.*, 2014) was used to correct for the intrinsically higher error rate of SMRT sequencing in comparison to Illumina sequencing (Rhoads and Au, 2015) and statistics were collected by Quast version 3.0 (Gurevich *et al.*, 2013; Fig. 3.2B). Assembly completeness was then assessed using BUSCO version 3.0.1 with the eukaryota\_odb9 database (Simão *et al.*, 2015) as above.

Both the Canu and SPAdes assemblies and the FALCON-Unzip and Illumina only SPAdes assemblies were then merged using QuickMerge version 0.2, with an anchor contig length corresponding to the N50 of the most contiguous

assembly, an anchor overlap cutoff of 5.0, an extension overlap cutoff of 1.5 and a minimum alignment length of 5,000 (Chakraborty *et al.*, 2016). This merged assembly was again error corrected by Pilon version 1.17 (Walker *et al.*, 2014) and statistics were collected by Quast version 3.0 (Gurevich *et al.*, 2013). Assembly completeness was then assessed using BUSCO version 3.0.1 with the eukaryota\_odb9 database (Simão *et al.*, 2015) as above.

### 3.3.7 Removal of contaminant sequences from assemblies

DeconSeq version 0.4.3 (Schmieder and Edwards, 2011) was used to identify potential contaminant sequences. Two databases produced internally were used for this process, these were built from all available full genomes of *Bacillus* spp. and *Paenibacillus* spp. respectively. A database of all available complete *Phytophthora* spp. sequences was also developed. The *Bacillus* and *Paenibacillus* databases consisted of all available genomes on the National Centre for Biotechnology Information (NCBI) GenBank for these genera listed as being representative and complete genomes. The *Phytophthora* database consisted of all available genomes on NCBI GenBank (<https://www.ncbi.nlm.nih.gov/nucleotide/>, last accessed 03/01/2019) for this genus listed as being either a reference or representative genome (Armitage, unpublished). Contigs identified as potential contaminants were analysed by the Basic Local Alignment Search Tool (BLAST) BLASTn tool against the NCBI nucleotide collection (nr) database (NCBI Resource Coordinators, 2018) using their web server (Johnson *et al.*, 2008) to identify contaminant organisms. Those identified as genuine contaminants were removed either solely by DeconSeq for the *Paenibacillus* analysis, or manually for the *Bacillus* analysis due to a high false positive rate.

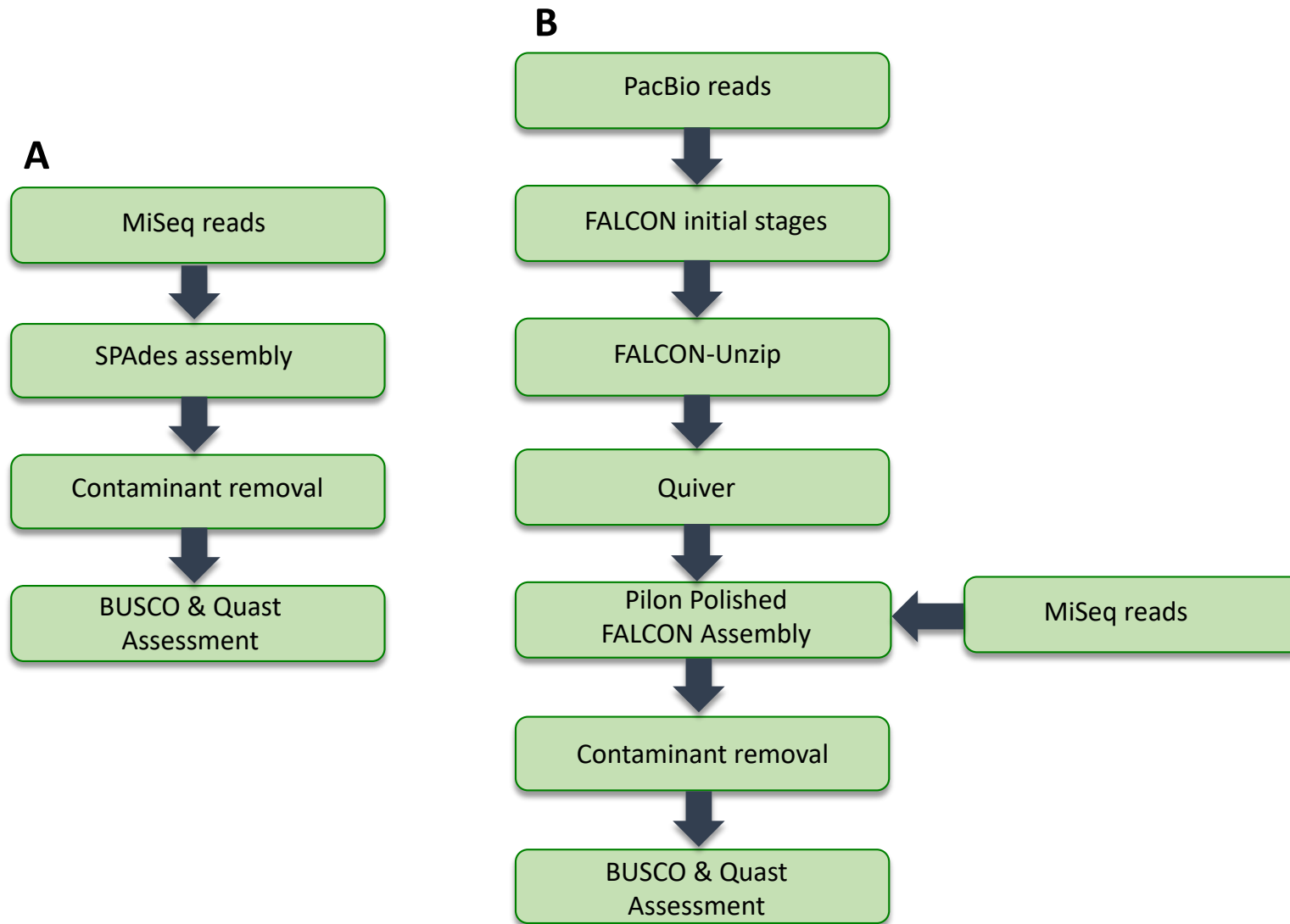
The assemblies were further checked for contamination by preliminarily uploading the fasta files of the assemblies to GenBank. As a part of this process, the GenBank team ran the assemblies through their own contamination screen. This produced a text file containing errors in the assemblies. This was downloaded and errors were removed.

Some assemblies were still found to contain bacterial contaminants, indicated by a suspiciously larger genome size and a few suspiciously long contigs, in one case with a lower GC content than the overall GC content of the assembly. These contigs were analysed by BLASTn against the NCBI nucleotide collection (nr) database (NCBI Resource Coordinators, 2018) using their web server (Johnson *et al.*, 2008) to confirm the identity of the sequences as bacterial contaminants. Following this, in the assembly where the contaminant had lower GC content, all contigs with a similar GC content were analysed by BLASTn against the NCBI nucleotide collection (nr) database using their web server. For the other organisms, a number of contaminant sequences were first identified by suspiciously high contig length. These were used to build a BLAST database in Geneious R10 (Kearse *et al.*, 2012). The remaining contigs in the genome were then analysed by BLASTn against this database in Geneious R10. Following this, any contigs that hit the database of contaminant sequences were selected. These were then analysed by BLASTn against the NCBI nucleotide collection (nr) database via Geneious R10. The contigs identified as contaminants by these methods were then removed from the assemblies using the previously mentioned python script. Statistics were then collected on these corrected assemblies using Quast (Gurevich *et al.*, 2013) and BUSCO (Simão *et al.*, 2015) as above.



**Fig. 3.2 (overleaf): An overview of assembly methods used in this study.**

Two assembly methods were used in this study depending on the sequencing method used. **A:** Illumina MiSeq reads were assembled using SPAdes (Bankevich *et al.*, 2012) to produce draft assemblies. Contaminant sequences were removed through a three-step process using: DeconSeq version 0.4.3 (Schmieder and Edwards, 2011), the NCBI Genbank contaminant screen and manual BLASTn searches against the NCBI nucleotide collection (nr) database (NCBI Resource Coordinators, 2018) implemented in Geneious R10 (Kearse *et al.*, 2012). Assembly statistics were then collected using Quast version 3.0 with the scaffolds option (Gurevich *et al.*, 2013). Assembly completeness was assessed by running a BUSCO (Benchmarking Universal Single-Copy Orthologs; Simão *et al.*, 2015) analysis was run using BUSCO version 3.0.1 with the eukaryota\_odb9 database on the assembly. **B:** PacBio single molecule real-time (SMRT) reads were assembled using FALCON-Unzip (Chin *et al.*, 2016), a diploid aware assembler developed by PacBio, followed by Quiver (Chin *et al.*, 2013). These programs produce a phased and error corrected assembly. Further error correction was performed using the MiSeq reads by Pilon (Walker *et al.*, 2014). Contaminant removal using DeconSeq version 0.4.3 (Schmieder and Edwards, 2011). Assembly statistics were then collected using Quast version 3.0 with the scaffolds option (Gurevich *et al.*, 2013). Assembly completeness was assessed by running a BUSCO (Benchmarking Universal Single-Copy Orthologs; Simão *et al.*, 2015) analysis was run using BUSCO version 3.0.1 with the eukaryota\_odb9 database on the assembly.



### **3.3.8 Masking of repetitive sequence and transposon sequences**

Masking of repetitive sequences was performed using RepeatMasker version open-4.0.5 (Smit *et al.*, 2013-2015) and RepeatModeler version 1.73 (Smit and Hubley, 2008-2015). These programs screen DNA sequences for interspersed repeats and low complexity DNA sequences and replaces them with Ns if hard masking was selected, or lowercase letters if soft masking was selected. Both hard and soft masked files were generated in this study. TransposonPSI release 22<sup>nd</sup> August 2010 (Haas, 2010) masks transposons using Position-Specific Iterative Basic Local Alignment Search Tool (PSI-BLAST) against a collection of transposon open reading frames (ORFs) to look for sequences with homology to known transposon sequences within the assembly and mask these with Ns. These programs also output a percentage of bases masked, however, genes can be contained within repetitive regions so caution was required when conducting further analyses.

### **3.3.9 Availability of sequencing data**

All Illumina and PacBio raw DNA sequencing reads have been submitted to the Sequencing Read Archive (SRA) maintained by NCBI. These will be available following the publication of these results in a peer reviewed journal (Table 3.1).

**Table 3.1: Details of the NCBI SRA codes for all sequencing reads uploaded for future release**

Species	Isolate	Sequencing Technology	SRA ID
<i>Phytophthora fragariae</i>	A4	Illumina MiSeq	SRR7668096
<i>Phytophthora fragariae</i>	BC-1	Illumina MiSeq	SRR7668095
<i>Phytophthora fragariae</i>	BC-16	Illumina MiSeq	SRR7668090
<i>Phytophthora fragariae</i>	BC-16	PacBio Single Molecule Real Time	SRR7668094
<i>Phytophthora fragariae</i>	BC-23	Illumina MiSeq	SRR7668088
<i>Phytophthora fragariae</i>	NOV-5	Illumina MiSeq	SRR7668091
<i>Phytophthora fragariae</i>	NOV-9	Illumina MiSeq	SRR7668100
<i>Phytophthora fragariae</i>	NOV-27	Illumina MiSeq	SRR7668092
<i>Phytophthora fragariae</i>	NOV-71	Illumina MiSeq	SRR7668086
<i>Phytophthora fragariae</i>	NOV-77	Illumina MiSeq	SRR7668087
<i>Phytophthora fragariae</i>	ONT-3	Illumina MiSeq	SRR7668089
<i>Phytophthora fragariae</i>	SCR245	Illumina MiSeq	SRR7668098
<i>Phytophthora rubi</i>	SCR249	Illumina MiSeq	SRR7668097
<i>Phytophthora rubi</i>	SCR324	Illumina MiSeq	SRR7668085
<i>Phytophthora rubi</i>	SCR333	Illumina MiSeq	SRR7668099

### 3.4 Results

#### 3.4.1 Characterisation of the race of a selection of *Phytophthora fragariae* isolates allowed conversion of three Canadian races and one U.S.A. race to the U.K. race scheme

As the isolates used in this study were received with either a Canadian race, a US race or no race assigned; those isolates assigned to the same Canadian race as BC-1, BC-16 or NOV-9 were assessed for their U.K. race in order to confirm that direct conversion is possible between the race schemes. These isolates were NOV-27 (race CA2), NOV-5 (race CA1) and NOV-71 (race CA2). Isolate A4 was thought to be UK2 and so was added to the experiment. Inoculations were performed on four cultivars of strawberry: 'Allstar' (containing *Rpf1*, *Rpf2* and *Rpf3*), 'Cambridge Vigour' (containing *Rpf2* and *Rpf3*), 'Hapil' (containing no resistance genes) and 'Redgauntlet' (containing *Rpf2*). Over the four weeks for which above ground symptoms were scored, very few above ground symptoms were observed, with only one 'Redgauntlet' plant inoculated with NOV-71 scoring at a value of two in the last two weeks of assessment.

After allowing two additional weeks for symptoms to develop in the roots, no change in above ground symptoms was observed. Following this, the roots of each plant were harvested, washed and investigated for rat's tails and red core symptoms, with identification of oospores used to judge whether an unclear set of symptoms for an individual was due to infection with *P. fragariae*. The presence of rat's tails was not deemed as sufficient for indicating *P. fragariae* infection, however both red core symptoms and oospores were deemed as sufficient for indicating a successful infection (Fig. 3.3; Table 3.2). These results confirmed the conversion between Canadian and U.K. race schemes for these three races, as well as confirming that A4 belongs to the UK2 race (Table 3.3).

**Fig. 3.3 (overleaf): Harvested roots from pathogenicity test.** Roots of *Fragaria* × *ananassa* harvested six weeks after inoculation with *Phytophthora fragariae* mycelial slurry. **A:** Compatible infection of the *P. fragariae* isolate NOV-27 (race CA2) on a susceptible 'Redgauntlet' plant (*Rpf2* only). **B:** Incompatible infection of the *P. fragariae* isolate A4 (US4) on a resistant 'Redgauntlet' plant (*Rpf2* only). **C:** Incompatible infection of the *P. fragariae* isolate NOV-27 (race CA2) on a resistant 'Allstar' plant (*Rpf1*, *Rpf2* and *Rpf3*). **D:** Compatible infection of the *P. fragariae* isolate A4 (race US3) on a susceptible 'Hapil' plant (no *Rpf* genes).



**Table 3.2: Results of pathogenicity test with deduced U.K. races**

Cultivar <sup>a</sup>	Isolate <sup>b</sup>				
	A4	NOV-27	NOV-5	NOV-71	Mock
Allstar <sub>1, 2, 3</sub>	0 <sup>3/4</sup>	0 <sup>2/3</sup>	0 <sup>3/3</sup>	0 <sup>4/5</sup>	0 <sup>1/1</sup>
Cambridge Vigour <sub>2, 3</sub>	0 <sup>5/5</sup>	0 <sup>4/5</sup>	+ <sup>5/5</sup>	0 <sup>4/5</sup>	0 <sup>1/1</sup>
Hapil	+ <sup>4/5</sup>	+ <sup>5/5</sup>	+ <sup>3/5</sup>	+ <sup>5/5</sup>	0 <sup>2/2</sup>
Redgauntlet <sub>2</sub>	0 <sup>5/5</sup>	+ <sup>5/5</sup>	+ <sup>5/5</sup>	+ <sup>5/5</sup>	0 <sup>2/2</sup>
Deduced UK race	Race 2	Race 3	Race 1	Race 3	N/A

<sup>a</sup>: Known resistance genes in the *Fragaria × ananassa* cultivars. <sup>b</sup>: Number of replicates showing the recorded phenotype compared to the number of replicates tested, with + representing a successful infection and 0 representing a resistance to the infection by the plant.

**Table 3.3 Summary of all pathogenicity races for all investigated *Phytophthora fragariae* isolates**

Isolate	UK race	CA race	US race
A4	2	3	4
BC-1	1	1	Unknown
BC-16	2	3	4
BC-23	Unknown	5	Unknown
NOV-5	1	1	Unknown
NOV-9	3	2	Unknown
NOV-27	3	2	Unknown
NOV-71	3	2	Unknown
NOV-77	Unknown	5	Unknown
ONT-3	Unknown	4	Unknown
SCR245	Unknown	Unknown	Unknown



### **3.4.2 Assessment of quality of sequencing data showed the data was sufficient for genome assembly**

Prior to genome assembly, the raw sequencing reads were assessed for read quality and average genome coverage. The sequencing reads were analysed both before and after trimming. These results showed that the reverse read libraries tended to perform worse than the corresponding forward read libraries, but the results overall suggested that the data was of good enough quality for draft *de novo* genome assembly. As expected, trimming of the reads improved their quality (Table 3.4 and 3.5).

KMC (Deorowicz *et al.*, 2013) provided an estimate of genome size and genome coverage, however, due to issues with filtering of error *K*-mers, occasionally the genome size was overestimated and hence the sequencing coverage was underestimated. From those isolates which did not show an excessively large estimated genome size from KMC (A4, BC-16, BC-23, NOV-27, NOV-5, NOV-77 and ONT-3), the average approximate genome size was determined to be 96,142,810 bp. This value was rounded to 96 Mb and used to more accurately determine the sequencing coverage by counting the number of nucleotides contained in the sequencing reads and dividing that value by the estimated genome size. A minimum coverage of 20 was required, otherwise additional sequencing was performed. A satisfactory coverage was generated for all isolates sequenced for this study.

**Table 3.4: Fastqc results from analysis of raw sequencing data**

Library Name	Basic Statistics	Per Base Sequence Quality	Per Sequence Quality Scores	Per Base Sequence Content	Per Base GC Content	Per Sequence GC Content	Per Base N Content	Sequence Length Distribution	Sequence Duplication Levels	Overrepresented Sequences	K-mer Content
A4_F	Pass	Fail	Pass	Fail	Fail	Fail	Pass	Warning	Warning	Warning	Warning
A4_R	Pass	Fail	Pass	Fail	Fail	Fail	Pass	Warning	Warning	Warning	Warning
BC-1_F.1	Pass	Fail	Pass	Fail	Fail	Fail	Pass	Warning	Warning	Warning	Warning
BC-1_F.2	Pass	Fail	Pass	Fail	Fail	Fail	Pass	Warning	Warning	Warning	Warning
BC-1_F.3	Pass	Fail	Pass	Fail	Fail	Fail	Pass	Warning	Warning	Warning	Warning
BC-1_R.1	Pass	Fail	Pass	Fail	Fail	Fail	Pass	Warning	Warning	Warning	Warning
BC-1_R.2	Pass	Fail	Pass	Fail	Fail	Fail	Pass	Warning	Warning	Warning	Warning
BC-1_R.3	Pass	Fail	Pass	Fail	Fail	Fail	Pass	Warning	Warning	Pass	Warning
BC-16_F.1	Pass	Fail	Pass	Fail	Warning	Fail	Pass	Warning	Warning	Pass	Warning
BC-16_F.2	Pass	Fail	Pass	Fail	Fail	Fail	Pass	Warning	Warning	Pass	Warning
BC-16_R.1	Pass	Fail	Pass	Fail	Fail	Fail	Pass	Warning	Warning	Pass	Fail
BC-16_R.2	Pass	Fail	Pass	Fail	Fail	Fail	Pass	Warning	Warning	Pass	Fail
BC-23_F	Pass	Fail	Pass	Fail	Fail	Fail	Pass	Warning	Warning	Pass	Warning
BC-23_R	Pass	Fail	Pass	Fail	Fail	Fail	Pass	Warning	Warning	Pass	Warning
NOV-5_F	Pass	Fail	Pass	Fail	Fail	Fail	Pass	Warning	Warning	Pass	Warning
NOV-5_R	Pass	Fail	Pass	Fail	Fail	Fail	Pass	Warning	Warning	Pass	Warning
NOV-9_F.1	Pass	Fail	Pass	Fail	Fail	Warning	Pass	Warning	Pass	Warning	Warning
NOV-9_F.2	Pass	Fail	Pass	Fail	Fail	Warning	Pass	Warning	Pass	Warning	Warning
NOV-9_F.3	Pass	Fail	Pass	Fail	Fail	Warning	Pass	Warning	Pass	Warning	Warning
NOV-9_R.1	Pass	Fail	Pass	Fail	Fail	Warning	Pass	Warning	Pass	Warning	Warning
NOV-9_R.2	Pass	Fail	Pass	Fail	Fail	Warning	Pass	Warning	Pass	Warning	Warning
NOV-9_R.3	Pass	Fail	Pass	Fail	Fail	Warning	Pass	Warning	Pass	Warning	Warning
NOV-27_F	Pass	Fail	Pass	Fail	Fail	Fail	Pass	Warning	Warning	Pass	Warning
NOV-27_R	Pass	Fail	Pass	Fail	Fail	Fail	Pass	Warning	Warning	Pass	Warning
NOV-71_F.1	Pass	Fail	Pass	Fail	Fail	Fail	Pass	Warning	Warning	Warning	Warning
NOV-71_F.2	Pass	Fail	Pass	Fail	Fail	Fail	Pass	Warning	Warning	Warning	Warning
NOV-71_R.1	Pass	Fail	Pass	Fail	Fail	Fail	Pass	Warning	Warning	Warning	Fail
NOV-71_R.2	Pass	Fail	Pass	Fail	Fail	Fail	Pass	Warning	Warning	Warning	Fail
NOV-77_F	Pass	Fail	Pass	Fail	Fail	Fail	Pass	Warning	Warning	Pass	Warning
NOV-77_R	Pass	Fail	Pass	Fail	Fail	Fail	Pass	Warning	Warning	Pass	Warning
ONT-3_F	Pass	Fail	Pass	Fail	Fail	Fail	Pass	Warning	Warning	Pass	Warning
ONT-3_R	Pass	Fail	Pass	Fail	Fail	Fail	Pass	Warning	Pass	Pass	Warning

SCR245_F	Pass	Fail	Pass	Fail	Warning	Fail	Pass	Warning	Warning	Pass	Warning
SCR245_R	Pass	Fail	Pass	Fail	Fail	Fail	Pass	Warning	Pass	Pass	Warning
SCR249_F	Pass	Fail	Pass	Fail	Warning	Fail	Pass	Warning	Warning	Pass	Warning
SCR249_R	Pass	Fail	Pass	Fail	Fail	Fail	Pass	Warning	Warning	Pass	Warning
SCR324_F	Pass	Fail	Pass	Fail	Warning	Fail	Pass	Warning	Warning	Pass	Warning
SCR324_R	Pass	Fail	Pass	Fail	Fail	Fail	Pass	Warning	Warning	Pass	Warning
SCR333_F	Pass	Fail	Pass	Fail	Fail	Warning	Pass	Warning	Warning	Pass	Warning
SCR333_R	Pass	Fail	Pass	Fail	Fail	Warning	Pass	Warning	Warning	Pass	Warning

---

**Table 3.5: Improved Fastqc results from analysis of trimmed sequencing data**

Library Name	Basic Statistics	Per Base Sequence Quality	Per Sequence Quality Scores	Per Base Sequence Content	Per Base GC Content	Per Sequence GC Content	Per Base N Content	Sequence length Distribution	Sequence Duplication Levels	Overrepresented Sequences	K-mer Content
A4_F	Pass	Warning	Pass	Pass	Warning	Fail	Pass	Warning	Warning	Warning	Warning
A4_R	Pass	Fail	Pass	Warning	Fail	Fail	Pass	Warning	Warning	Warning	Fail
BC-1_F.1	Pass	Warning	Pass	Pass	Pass	Fail	Pass	Warning	Warning	Pass	Warning
BC-1_F.2	Pass	Warning	Pass	Pass	Pass	Fail	Pass	Warning	Warning	Pass	Warning
BC-1_F.3	Pass	Warning	Pass	Pass	Pass	Fail	Pass	Warning	Warning	Pass	Warning
BC-1_R.1	Pass	Fail	Pass	Warning	Fail	Fail	Pass	Warning	Warning	Pass	Fail
BC-1_R.2	Pass	Fail	Pass	Pass	Pass	Fail	Pass	Warning	Pass	Pass	Fail
BC-1_R.3	Pass	Fail	Pass	Pass	Pass	Fail	Pass	Warning	Pass	Pass	Fail
BC-16_F.1	Pass	Warning	Pass	Pass	Pass	Fail	Pass	Warning	Warning	Pass	Warning
BC-16_F.2	Pass	Warning	Pass	Pass	Pass	Fail	Pass	Warning	Warning	Pass	Warning
BC-16_R.1	Pass	Fail	Pass	Fail	Fail	Fail	Pass	Warning	Warning	Pass	Fail
BC-16_R.2	Pass	Fail	Pass	Warning	Fail	Fail	Pass	Warning	Warning	Pass	Fail
BC-23_F	Pass	Warning	Pass	Pass	Warning	Fail	Pass	Warning	Warning	Pass	Warning
BC-23_R	Pass	Fail	Pass	Warning	Fail	Fail	Pass	Warning	Warning	Pass	Fail
NOV-5_F	Pass	Pass	Pass	Pass	Pass	Fail	Pass	Warning	Warning	Pass	Warning
NOV-5_R	Pass	Fail	Pass	Warning	Warning	Fail	Pass	Warning	Warning	Pass	Fail
NOV-9_F.1	Pass	Warning	Pass	Pass	Pass	Warning	Pass	Warning	Pass	Pass	Pass
NOV-9_F.2	Pass	Warning	Pass	Pass	Pass	Warning	Pass	Warning	Pass	Pass	Warning
NOV-9_F.3	Pass	Warning	Pass	Pass	Pass	Warning	Pass	Warning	Pass	Pass	Warning
NOV-9_R.1	Pass	Fail	Pass	Pass	Warning	Warning	Pass	Warning	Pass	Pass	Fail
NOV-9_R.2	Pass	Fail	Pass	Pass	Pass	Warning	Pass	Warning	Pass	Pass	Fail
NOV-9_R.3	Pass	Fail	Pass	Pass	Pass	Warning	Pass	Warning	Pass	Pass	Fail
NOV-27_F	Pass	Fail	Pass	Pass	Pass	Fail	Pass	Warning	Warning	Pass	Warning
NOV-27_R	Pass	Fail	Pass	Warning	Fail	Fail	Pass	Warning	Warning	Pass	Fail
NOV-71_F.1	Pass	Warning	Pass	Pass	Pass	Fail	Pass	Warning	Warning	Pass	Warning
NOV-71_F.2	Pass	Warning	Pass	Pass	Pass	Fail	Pass	Warning	Warning	Pass	Warning
NOV-71_R.1	Pass	Fail	Pass	Pass	Pass	Fail	Pass	Warning	Warning	Pass	Fail
NOV-71_R.2	Pass	Fail	Pass	Warning	Warning	Fail	Pass	Warning	Warning	Pass	Fail
NOV-77_F	Pass	Pass	Pass	Pass	Pass	Fail	Pass	Warning	Warning	Pass	Warning
NOV-77_R	Pass	Fail	Pass	Warning	Warning	Fail	Pass	Warning	Warning	Pass	Fail
ONT-3_F	Pass	Pass	Pass	Pass	Pass	Fail	Pass	Warning	Warning	Pass	Warning
ONT-3_R	Pass	Fail	Pass	Warning	Warning	Fail	Pass	Warning	Pass	Pass	Fail

SCR245_F	Pass	Warning	Pass	Warning	Warning	Fail	Pass	Warning	Warning	Pass	Warning
SCR245_R	Pass	Fail	Pass	Warning	Fail	Warning	Pass	Warning	Pass	Pass	Fail
SCR249_F	Pass	Pass	Pass	Pass	Pass	Fail	Pass	Warning	Warning	Pass	Warning
SCR249_R	Pass	Fail	Pass	Warning	Fail	Fail	Pass	Warning	Warning	Pass	Fail
SCR324_F	Pass	Pass	Pass	Pass	Pass	Fail	Pass	Warning	Warning	Pass	Warning
SCR324_R	Pass	Fail	Pass	Warning	Fail	Fail	Pass	Warning	Warning	Pass	Fail
SCR333_F	Pass	Warning	Pass	Pass	Pass	Warning	Pass	Warning	Warning	Pass	Warning
SCR333_R	Pass	Fail	Pass	Warning	Warning	Warning	Pass	Warning	Warning	Pass	Fail

### **3.4.3 *De novo* assembly of genomes**

#### **3.4.3.1 *De novo* assembly of genomes from Illumina data showed little variation at the genome level between the isolates**

Following assembly and filtering, the assemblies were analysed to provide information on the number of contigs, the assembly size, the N50, the L50, the percentage GC content and information on the number of mismatches (N's). N50 is a statistic which is used to assess the quality of an assembly. This is defined as the minimum contig length needed to cover 50% of the genome, specifically that half the genome sequence is in contigs larger than or equal to the value of the N50. As such it provides an indication of the level of fragmentation of the assembly. However, N50 values are not directly comparable unless the assembly size is equal. L50 is another diagnostic statistic that provides information on the fragmentation of the assembly. This is defined as the smallest number of contigs whose length sum produces the N50 value. Several isolates have a reported pre-assembly size greater than 100 Mb, this was likely due to poor filtering of error *K*-mers by KMC (Table 3.6).

**Table 3.6: Assembly statistics from pre-assembly and post-assembly**

Species	Isolate	Assembly Programme	Genome Size	Genome Coverage	Genome Size	Contig Number			Percentage	
			Pre-Assembly	Estimate	Post-Assembly	$\geq 500\text{bp}$	N50	L50	GC content	Number of N's
<i>Phytophthora fragariae</i>	A4	SPAdes	93,413,554	35.91	79,084,532	13,447	18,245	1,116	53.36	22,197
<i>Phytophthora fragariae</i>	BC-1	SPAdes	1,196,301,136	116.25	79,104,406	11,557	21,834	954	53.34	18,122
<i>Phytophthora fragariae</i>	BC-16	SPAdes	90,864,210	68.96	79,251,309	12,574	19,795	1,049	53.39	15,069
<i>Phytophthora fragariae</i>	BC-23	SPAdes	103,251,773	49.33	78,261,318	13,192	18,227	1,119	53.35	24,545
<i>Phytophthora fragariae</i>	NOV-5	SPAdes	95,350,039	40.21	78,998,334	13,529	17,887	1,134	53.38	25,934
<i>Phytophthora fragariae</i>	NOV-9	SPAdes	959,591,302	103.23	79,434,965	11,803	21,522	978	53.38	20,976
<i>Phytophthora fragariae</i>	NOV-27	SPAdes	93,851,233	52.27	78,754,876	12,490	19,406	1,046	53.38	24,931
<i>Phytophthora fragariae</i>	NOV-71	SPAdes	810,779,109	79.92	78,377,198	12,214	20,226	1,016	53.33	21,160
<i>Phytophthora fragariae</i>	NOV-77	SPAdes	92,399,813	46.29	78,809,760	13,322	18,909	1,102	53.39	19,838
<i>Phytophthora fragariae</i>	ONT-3	SPAdes	103,869,049	45.18	88,587,983	13,346	22,074	917	53.11	21,326
<i>Phytophthora fragariae</i>	SCR245	SPAdes	127,550,025	51.47	83,317,162	13,259	20,105	994	52.34	18,996
<i>Phytophthora rubi</i>	SCR249	SPAdes	105,406,182	51.00	77,788,370	14,023	16,620	1,232	53.45	26,455
<i>Phytophthora rubi</i>	SCR324	SPAdes	104,021,199	50.36	78,256,282	13,947	17,037	1,209	53.41	21,914
<i>Phytophthora rubi</i>	SCR333	SPAdes	102,239,327	49.79	77,817,047	13,679	16,946	1,210	53.43	15,429

The large discrepancies between pre-assembly estimates and post-assembly statistics was likely due to poor error *K*-mer filtering by KMC resulting in excessively large genome size predictions. This appeared to be the case for: BC-1, NOV-9, NOV-71 and SCR245. It could also be due to the inability of short reads to assemble repeat rich regions of the genome or the presence of contaminating sequences.

### **3.4.3.2 *De novo* assembly of a new reference assembly using PacBio sequencing data improved the contiguity of the genome assembly**

In order to produce a reference genome assembly of greater contiguity than the Illumina only assemblies, the BC-16 isolate (race UK2), was sequenced using PacBio SMRT sequencing technology. This produced reads of a much longer length than the Illumina technology, an average of approximately 15,000 bp as opposed to 250 bp, and hence aided in assembling difficult to assemble regions, such as repeat rich regions (Rhoads and Au *et al.*, 2015). Initially, two assembly methods were trialled. An assembly of just PacBio reads with Canu (Koren *et al.*, 2017) produced a far more contiguous assembly than a SPAdes hybrid approach utilising both PacBio and Illumina reads (Bankevich *et al.*, 2012; Table 3.7).

As this assembly was still relatively fragmented in 372 contigs, an assembler developed by PacBio, FALCON-Unzip (Chin *et al.*, 2016), was used. This assembler was designed to be diploid aware and was still in the final stages of development when used. However, following parameter optimisation to minimise potential spurious duplications, a more contiguous assembly of 180 contigs was generated (Table 3.7).

Attempts to further improve the assembly by merging the Canu and SPAdes assemblies or the FALCON-Unzip and the Illumina only SPAdes assemblies with QuickMerge (Chakraborty *et al.*, 2016) were unsuccessful in improving the contiguity with the number of contigs remaining unchanged at 372 for the Canu and SPAdes assemblies and 180 contigs for the FALCON-Unzip and SPAdes assemblies. As a result, the FALCON-Unzip only assembly was accepted as the draft assembly going forward. Part of the FALCON-Unzip assembly process involved phasing during the Unzip step (Chin *et al.*, 2016) and some error correction with Quiver (Chin *et al.*, 2013). Further correction was performed using ten iterations of Pilon (Walker *et al.*, 2014) with the Illumina reads for BC-16 (Table 3.7).



**Table 3.7: Assembly statistics for the *Phytophthora fragariae* isolate BC-16 utilising PacBio sequencing data**

Species	Isolate	Assembly Programme	Genome	Genome Size	Number of	N50	L50	Percentage GC content	Number of N's
			Coverage Estimate	Post-Assembly	Contigs ≥ 500bp				
<i>Phytophthora fragariae</i>	BC-16	SPAdes	61.29	70,028,085	10,900	47,885	405	52.97	0
<i>Phytophthora fragariae</i>	BC-16	Canu	61.29	96,985,215	372	622,936	47	53.57	0
<i>Phytophthora fragariae</i>	BC-16	FALCON	61.29	90,966,605	180	923,458	33	53.39	0
<i>Phytophthora fragariae</i>	BC-16	Quickmerge SPAdes & Canu	61.29	96,986,176	372	623,191	47	53.57	0
<i>Phytophthora fragariae</i>	BC-16	Quickmerge SPAdes & FALCON	61.29	90,966,605	180	923,458	33	53.39	0

No pre-assembly statistics were generated as KMC was not able to run with PacBio data.

### **3.4.4 Identification and removal of error and contaminant sequences identified bacterial contaminants and improved the assemblies**

A known problem during genome sequencing and assembly, is the inadvertent sequencing of non-target species, whether this is due to the contamination of the biological material from which DNA is extracted, or during library preparation (Merchant *et al.*, 2014). Initially, sequences were removed using DeconSeq (Schmieder and Edwards, 2011) with a database of *Paenibacillus* spp. as the database to mark a sequence for exclusion and a database of *Phytophthora* spp. as the database to mark a sequence for retention (Armitage, unpublished). This removed several contaminant contigs for ONT-3, SCRP245 and SCRP324.  $\phi$ X was removed from all assemblies except for BC-16. The DNA of this bacteriophage is used in Illumina sequencing as a quality control measure and so its presence was not a concern. The contaminated assemblies contained bacterial sequences which showed BLASTn hits to several genera, including: *Paenibacillus*, *Thermobacillus*, *Brevibacillus*, *Bacillus* and *Staphylococcus*. Following this, DeconSeq was again run using a larger database of *Bacillus* spp., however the contigs identified as contaminants frequently showed hits to *Phytophthora* sequences in the NCBI nucleotide collection (nr) database (NCBI Resource Coordinators, 2018) following BLASTn analysis via the web server (Johnson *et al.*, 2008). These sequences were therefore kept in the assemblies (Table 3.8).

Following DeconSeq corrections, the assemblies were submitted to NCBI GenBank for preliminary checks before further analyses were conducted. As a part of this process, the sequences were run through NCBI's contamination and error screen. This showed further sequences identified as contaminants or errors in the assemblies of: A4, BC-1, NOV-5, NOV-9, NOV-71, SCRP245 and SCRP324. Some non-contaminant errors were identified, namely: A4 still had an adaptor sequence present, NOV-5 still had five adaptor sequences present and NOV-71 had a duplicate contig present. Isolates BC-1, NOV-5, NOV-9, SCRP245 and SCRP324 contained contaminant sequences from several different possible contaminant organisms, these were identified by NCBI as:

*Paenibacillus* sp., *Bacillus infantis*, *Bacillus oceanisediminis*, Moloney murine leukemia virus, *Brassica napus* and *Penicillium solitum*. Some of these contaminants, such as the murine virus, seemed unlikely to be the genuine source, but as they did not hit *Phytophthora* sequences when analysed by BLASTn via their web server (Johnson *et al.*, 2008) against the NCBI nucleotide collection (nr) database (NCBI Resource Coordinators, 2018), the advice of the NCBI Genbank team was accepted and these sequences were removed (Table 3.9).

Following the correction of these errors and the removal of contaminant sequences, it was observed that the assemblies of isolates ONT-3, SCRP245 and SCRP324 still potentially contained contaminant sequences. This was due to a larger than expected largest contig length, some of which displayed a lower percentage GC content than that of the assembly as a whole. Additional candidate contaminant contigs were identified through a BLASTn of all the contigs in an assembly against the known contaminant contigs in an assembly. This set of candidate contaminant sequences were analysed by BLASTn on the NCBI web server (Johnson *et al.*, 2008) against the NCBI nucleotide collection (nr) database (NCBI Resource Coordinators, 2018) to identify whether the sequences hit *Phytophthora* sequences or potential contaminant organisms. In the case of poor-quality hits to both groups of organisms, sequences were kept. In the ONT-3 and SCRP324 assemblies, the likely contaminant organism was a *Paenibacillus* sp., though the identity of the species was unclear, possibly as it has not yet been sequenced. For SCRP245, the contaminant organism was a *Bacillus* sp., though again the identity of the species was unclear (Table 3.10).

**Table 3.8: Assembly statistics after contaminant removal by DeconSeq**

Species	Isolate	Assembly Size	Contig Number			Percentage	
			$\geq 500\text{bp}$	N50	L50	GC Content	Number of N's
<i>Phytophthora fragariae</i>	A4	79,079,019	13,446	18,245	1,116	53.36	22,197
<i>Phytophthora fragariae</i>	BC-1	79,098,893	11,556	21,842	953	53.34	18,122
<i>Phytophthora fragariae</i>	BC-16	90,966,605	180	923,458	33	53.39	0
<i>Phytophthora fragariae</i>	BC-23	78,255,805	13,191	18,227	1,119	53.36	24,545
<i>Phytophthora fragariae</i>	NOV-5	78,992,821	13,528	17,887	1,134	53.38	25,934
<i>Phytophthora fragariae</i>	NOV-9	79,429,452	11,802	21,522	978	53.38	20,976
<i>Phytophthora fragariae</i>	NOV-27	78,749,363	12,489	19,406	1,046	53.38	24,931
<i>Phytophthora fragariae</i>	NOV-71	78,371,685	12,213	20,226	1,016	53.33	21,160
<i>Phytophthora fragariae</i>	NOV-77	78,804,247	13,321	18,925	1,101	53.39	19,838
<i>Phytophthora fragariae</i>	ONT-3	84,322,417	13,292	20,565	988	53.22	20,496
<i>Phytophthora fragariae</i>	SCR245	83,127,441	13,249	20,056	995	52.37	18,612
<i>Phytophthora rubi</i>	SCR249	77,788,370	14,023	16,620	1,232	53.45	26,455
<i>Phytophthora rubi</i>	SCR324	78,256,282	13,947	17,037	1,209	53.41	21,914
<i>Phytophthora rubi</i>	SCR333	77,817,047	13,679	16,946	1,210	53.43	15,429

**Table 3.9: Assembly statistics after NCBI suggestions performed**

Species	Isolate	Assembly Size	Contig Number			Percentage	
			$\geq$ 500bp	N50	L50	GC Content	Number of N's
<i>Phytophthora fragariae</i>	A4	79,078,972	13,446	18,245	1,116	53.36	22,197
<i>Phytophthora fragariae</i>	BC-1	79,097,775	11,556	21,842	953	53.34	18,122
<i>Phytophthora fragariae</i>	NOV-5	78,992,522	13,531	17,887	1,134	53.38	25,934
<i>Phytophthora fragariae</i>	NOV-9	79,428,910	11,801	21,522	978	53.38	20,976
<i>Phytophthora fragariae</i>	NOV-71	78,371,082	12,212	20,226	1,016	53.33	21,160
<i>Phytophthora fragariae</i>	SCR245	83,123,062	13,251	20,056	995	52.37	18,612
<i>Phytophthora rubi</i>	SCR324	78,234,274	13,947	17,037	1,209	53.41	21,914

All isolates not listed in this table were accepted by NCBI with no errors

**Table 3.10: Assembly statistics following manual removal of contaminant sequences**

Species	Isolate	Assembly Size	Contig Number			Percentage	
			$\geq$ 500bp	N50	L50	GC Content	Number of N's
<i>Phytophthora fragariae</i>	ONT-3	79,064,189	13,265	18,830	1,104	53.37	20,496
<i>Phytophthora fragariae</i>	SCR245	77,843,592	13,231	18,154	1,120	53.37	18,612
<i>Phytophthora rubi</i>	SCR324	77,910,222	13,946	16,934	1,218	53.44	21,914

All isolates not listed in this table were not manually identified as containing further contaminant sequences

### **3.4.5 Assessment of genome completeness showed the assemblies represented the majority of the core genome of eukaryotic organisms**

Initially, genome completeness was assessed using CEGMA (Core Eukaryotic Genes Mapping Approach; Parra *et al.*, 2007). Later, a change was made to use BUSCO (Benchmarking Universal Single-Copy Orthologs) due to the ability of this method to assess not just completeness, but duplication as well (Simão *et al.*, 2015). The eukaryotic BUSCO database was used for this analysis as it represented the closest phylogenetic grouping available at the time. The Illumina only assemblies showed approximately 90% of BUSCO genes present as complete, single copy predictions, with only a small number (~5%) missing, which may have been due to these missing genes not being true eukaryotic BUSCOs. There was a slight decline in single copy, complete BUSCO gene predictions in the FALCON-Unzip assembly, though this was very slight and represented at most four percentage points of difference (88%; Table 3.11).

**Table 3.11: CEGMA and BUSCO statistics for genome assemblies**

Species	Isolate	Percentage Complete	Percentage Partial	Single Copy	Duplicated	Fragmented	Missing
		CEGMA genes	CEGMA genes	BUSCO genes	BUSCO genes	BUSCO genes	BUSCO genes
<i>Phytophthora fragariae</i>	A4	95.16%	97.98%	274 (90.43%)	6 (1.98%)	6 (1.98%)	17 (5.61%)
<i>Phytophthora fragariae</i>	BC-1	95.16%	97.58%	274 (90.43%)	6 (1.98%)	6 (1.98%)	17 (5.61%)
<i>Phytophthora fragariae</i>	BC-16	N/A	N/A	266 (87.79%)	9 (2.97%)	5 (1.65%)	23 (7.59%)
<i>Phytophthora fragariae</i>	BC-23	95.16%	97.58%	275 (90.76%)	5 (1.65%)	7 (2.31%)	16 (5.28%)
<i>Phytophthora fragariae</i>	NOV-5	94.76%	97.18%	273 (90.10%)	7 (2.31%)	6 (1.98%)	17 (5.61%)
<i>Phytophthora fragariae</i>	NOV-9	94.35%	97.18%	273 (90.10%)	6 (1.98%)	7 (2.31%)	17 (5.61%)
<i>Phytophthora fragariae</i>	NOV-27	94.76%	97.18%	273 (90.10%)	6 (1.98%)	7 (2.31%)	17 (5.61%)
<i>Phytophthora fragariae</i>	NOV-71	95.16%	97.98%	274 (90.43%)	6 (1.98%)	6 (1.98%)	17 (5.61%)
<i>Phytophthora fragariae</i>	NOV-77	94.76%	97.18%	272 (89.77%)	8 (2.64%)	6 (1.98%)	17 (5.61%)
<i>Phytophthora fragariae</i>	ONT-3	95.16%	97.18%	277 (91.42%)	7 (2.31%)	4 (1.32%)	15 (4.95%)
<i>Phytophthora fragariae</i>	SCR245	95.16%	97.18%	273 (90.10%)	7 (2.31%)	7 (2.31%)	16 (5.28%)
<i>Phytophthora rubi</i>	SCR249	93.95%	96.77%	273 (90.10%)	8 (2.64%)	3 (0.99%)	19 (6.27%)
<i>Phytophthora rubi</i>	SCR324	95.56%	97.98%	274 (90.43%)	8 (2.64%)	3 (0.99%)	18 (5.94%)
<i>Phytophthora rubi</i>	SCR333	94.35%	96.77%	275 (90.76%)	7 (2.31%)	3 (0.99%)	18 (5.94%)

CEGMA was not performed on the FALCON-Unzip assembly of BC-16 due to the adoption of BUSCO analyses, the total number of BUSCOs searched for was 303, the total number of CEGMA genes searches for was 1,788.



### **3.4.6 Identification and masking of regions of repetitive sequence showed a large proportion of the genome consisted of repeat rich regions**

Prior to any further analyses, repetitive sequences and putative transposon ORFs were identified. These statistics showed that in the Illumina only genomes, over 30% of the genome was masked, as opposed to 40% in the FALCON-Unzip assembly. This suggested that the reason for the BC-16 assembly showing increased contiguity and assembly size was due to an improved ability to assemble repeat rich regions. It was therefore decided that the FALCON-Unzip assembled genome of the BC-16 isolate represented a good reference assembly, despite showing marginally lower BUSCO completeness than the Illumina assemblies (Table 3.12).

**Table 3.12: Repeat masking and transposon masking statistics for the genome assemblies**

Species	Isolate	Number of bases masked			Assembly Size	Percentage of bases masked		
		RepeatMasker	TransposonPSI	Total		RepeatMasker	TransposonPSI	Total
<i>Phytophthora fragariae</i>	A4	24,225,474	6,237,528	25,918,245	79,078,972	30.63%	7.89%	32.78%
<i>Phytophthora fragariae</i>	BC-1	24,762,347	6,219,359	26,468,865	79,097,775	31.31%	7.86%	33.46%
<i>Phytophthora fragariae</i>	BC-16	34,585,513	8,335,020	36,352,455	90,966,605	38.02%	9.16%	39.96%
<i>Phytophthora fragariae</i>	BC-23	24,253,140	6,101,880	25,855,136	78,255,805	30.99%	7.80%	33.04%
<i>Phytophthora fragariae</i>	NOV-5	24,217,567	6,242,472	25,921,835	78,992,522	30.66%	7.90%	32.82%
<i>Phytophthora fragariae</i>	NOV-9	24,970,270	6,289,715	26,760,322	79,428,910	31.44%	7.92%	33.69%
<i>Phytophthora fragariae</i>	NOV-27	24,280,526	6,209,723	26,042,750	78,749,363	30.83%	7.89%	33.07%
<i>Phytophthora fragariae</i>	NOV-71	24,035,545	6,080,704	25,789,456	78,371,082	30.67%	7.76%	32.91%
<i>Phytophthora fragariae</i>	NOV-77	24,578,287	6,250,930	26,395,760	78,804,247	31.19%	7.93%	33.50%
<i>Phytophthora fragariae</i>	ONT-3	24,298,774	6,228,285	26,099,638	79,064,189	30.73%	7.88%	33.01%
<i>Phytophthora fragariae</i>	SCR245	24,077,643	6,029,983	25,831,280	77,843,592	30.93%	7.75%	33.18%
<i>Phytophthora rubi</i>	SCR249	23,906,929	5,953,026	25,659,011	77,788,370	30.73%	7.65%	32.99%
<i>Phytophthora rubi</i>	SCR324	23,969,622	5,940,402	25,657,326	77,910,222	30.77%	7.62%	32.93%
<i>Phytophthora rubi</i>	SCR333	23,126,516	5,961,557	25,001,991	77,817,047	29.72%	7.66%	32.13%

### **3.4.7 Comparison to published assemblies of *Phytophthora fragariae* demonstrated the improvement achieved by the PacBio assembly**

There were two publicly available genome assemblies of *P. fragariae* (Gao *et al.*, 2015; Tabima *et al.*, 2017) when this work was conducted. When these assemblies were compared to the assemblies generated in this study, the two assembly sizes were 20 Mb smaller than the FALCON-Unzip assembled genome.

However, the sizes of the Illumina-only sequenced genomes from this study were a similar size to the previously published assemblies. As both the Canu and FALCON-Unzip assemblies of the PacBio SMRT reads produced sizes of approximately 90 Mb, this appeared to be closer to the true genome size of *P. fragariae*. A lack of clear differences in the assemblies of isolates of differing races was also observed.

The phylogenetic identity of both available assemblies was first confirmed. The internal transcribed spacer (ITS) sequence was downloaded for the FJ172257 accession of *P. fragariae* from GenBank and for the KY785002.1 accession of *P. rubi* from GenBank (NCBI Resource Coordinators, 2018). When these were analysed using BLASTn in Geneious R10 (Kearse *et al.*, 2012) against the *P. fragariae* and *P. rubi* assemblies described above, identical partial portions of sequence were identified in both species. As an alternative, the  $\beta$ -tubulin gene sequence was downloaded for the AY564062.1 accession of *P. fragariae* from GenBank (NCBI Resource Coordinators, 2018). Using the BLASTn feature in Geneious R10,  $\beta$ -tubulin sequences were identified in all the assemblies generated in this study, as well as the two published genomes. At position 915 in the third exon of the  $\beta$ -tubulin gene there was a SNP from thymine (T) in the *P. fragariae* assemblies to cytosine (C) in the *P. rubi* assemblies. Both the published assemblies contained the *P. fragariae* type allele, confirming their identity as *P. fragariae*. The statistics for *P. sojae* appeared to show a similar assembly size to the assemblies generated in this study, especially given that the *P. sojae* assembly did not use long read sequencing data (Tyler *et al.*, 2006).

This also explained why the *P. sojae* assembly was still fairly fragmented compared to the FALCON-Unzip assembly. *P. infestans* showed a very large genome size, more than twice the size of the FALCON-Unzip assembly generated in this work (Haas *et al.*, 2009). This was likely not a concern for the assemblies produced here, as an inspection of the phylogenetic tree for this genus showed that these organisms are fairly distantly related within the diversity of the *Phytophthora* genus (Yang *et al.*, 2017; Table 3.13).

**Table 3.13: Comparison of statistics between BC-16 FALCON-Unzip assembly, the published *Phytophthora fragariae* assemblies and the assembly of the closely related soybean pathogen *Phytophthora sojae* and the late blight pathogen *Phytophthora infestans***

Source	Isolate	Species	Assembly Size (Mb)	Number Of Contigs	N50	L50
<i>de novo</i> FALCON Assembly This study	BC-16	<i>Phytophthora fragariae</i>	90.97	180	923,458	33
Gao <i>et al.</i> , 2015 GenBank GCA_000686205.4	309.62	<i>Phytophthora fragariae</i>	75.98	2,863	58,896	361
Tabima <i>et al.</i> , 2017 GenBank GCA_002025845.1	pd0101050015038	<i>Phytophthora fragariae</i>	76.97	10,440	18,987	1,071
Tyler <i>et al.</i> , 2006 GenBank GCA_000149755.2	P6497	<i>Phytophthora sojae</i>	82.60	862	385,992	61
Haas <i>et al.</i> , 2009 GenBank GCA_000142945.1	T30-4	<i>Phytophthora infestans</i>	228.54	18,288	44,484	899

## 3.5 Discussion

### 3.5.1 Pot based pathogenicity tests allowed assignment of additional isolates of *Phytophthora fragariae* to the U.K. race scheme

The results of the pathogenicity test allowed the assignment of U.K. races to four additional isolates, alongside the three already investigated prior to the commencing of this work (Table 3.3). Therefore, every isolate in the NIAB EMR collection which was initially assigned to races CA1, CA2 and CA3 has been assigned to a U.K. race. We can therefore conclude that race CA1 is identical to race UK1, race CA2 is identical to race UK3 and race CA3 is identical to race UK2. There was also a suggestion that A4 may be a race UK2 isolate, the results from the pathogenicity test have confirmed this, suggesting that race US4 is identical to race UK2. These results will aid with future work in mapping resistance genes in *F. × ananassa*, as well as aiding in further investigations in the following chapters.

### 3.5.2 *De novo* assembly of genomes of isolates of *Phytophthora fragariae* and *Phytophthora rubi* showed similar genome sizes and completeness and allowed the creation of a new reference assembly

Genomes of all eleven isolates of *P. fragariae* and all three isolates of *P. rubi* have been successfully assembled from Illumina MiSeq data, and the race UK2 isolate BC-16 has been assembled from PacBio SMRT data. The Illumina sequenced *P. fragariae* isolates assembled at an assembly size between 77 Mb and 80 Mb. Despite the lack of differences in assembly size, it was not possible to reject the possibility of the involvement of accessory chromosomes in pathogenicity differences, as has previously been shown in several pathogenic fungi, such as *Fusarium* spp. (Ma *et al.*, 2010; Croll and McDonald, 2012). However, there was equally no evidence to support the presence of accessory chromosomes. Following the assembly of PacBio reads for the BC-16 isolate, the true genome size was shown to likely be closer to 91 Mb, similar to the value of 96 Mb predicted from the raw Illumina MiSeq reads by KMC (Deorowicz *et*

*al.*, 2013). The slight size differences in assembly size do not appear to be related to a smaller amount of repetitive sequences, the number of masked bases was similar in all the Illumina sequenced isolates. In future, it would be interesting to assemble long read sequence data for other isolates to see whether the true genome size is indeed closer to 91Mb for all isolates and to provide assemblies of greater completeness for further investigations.

Long read sequencing is becoming a more common tool for genomic investigations, particularly within plant pathogens with complex genomes, such as the broad host range oomycete pathogen *Phytophthora nicotianae* (Liu *et al.*, 2016) and the fungal pathogen causing wheat yellow rust, *Puccinia striiformis* f. sp. *tritici* (Schwessinger *et al.*, 2018). In this study, the use of PacBio sequencing and the FALCON-Unzip assembler greatly improved the contiguity of the assembly of the BC-16 isolate of *P. fragariae*. However, it was a concern that there appeared to be around six extra BUSCO genes missing from the long-read assembly compared to those assembled solely from Illumina data. This was most likely due to the higher error rate inherent in PacBio sequencing compared to Illumina (Rhoads and Au, 2015), despite the fact that error correction was performed with both Quiver (Chin *et al.*, 2013) and ten iterations of Pilon (Walker *et al.*, 2014), rather than part of the genome no longer being assembled. Another possibility was that some BUSCO predictions may have been due to incorrect predictions in the Illumina only assemblies.

Interestingly, there were at least fifteen BUSCO genes from the eukaryotic dataset missing from all the assemblies generated in this study. This potentially suggested that these genes in the eukaryotic database may not in fact be universal to all eukaryotes, however, this was the closest phylogenetic grouping to *P. fragariae* available. Anecdotal evidence from other groups involved in the assembly of *Phytophthora* species suggests they observe a similar phenomenon (*Phytophthora* Sequencing Consortium, personal communication). The difference in BUSCO statistics between the Illumina only assemblies and the FALCON-Unzip assembly represented an approximate one percentage point decrease compared to the least complete *P. rubi* assembly

and an approximate two percentage point decrease compared to the least complete *P. fragariae* assembly. As this was very slight, it was decided that the FALCON-Unzip assembly was better to use as a new reference assembly. It is also possible that after gene annotation, differences in BUSCO predictions may change if some of these hits to the Illumina only assemblies were not genuine hits.

### **3.5.3 Comparison with other studied *Phytophthora* spp. showed similarities to a species in the same clade, but differences compared to a more distantly related species**

As the *Phytophthora* genus is well studied, it was of interest to compare the genome assemblies generated in this study with those generated for other species. In particular, assemblies of the most well studied species in this group, *P. infestans* and the best studied species in the same clade as *P. fragariae* and *P. rubi*, *P. sojae*. As *P. infestans* groups in clade 1c, whereas *P. fragariae*, *P. rubi* and *P. sojae* all group in clade 7, it is fairly distantly related to these species (Yang *et al.*, 2017). The current reference assembly of the T30-4 isolate of *P. infestans* had a total contig length of 190 Mb, but a predicted genome size of approximately 240 Mb (Haas, *et al.*, 2009). Therefore, this assembly was more than twice the size of the proposed reference assembly in this study. However, when comparing the proposed reference assembly to the more closely related species *P. sojae*, a genome size much closer to the size of our assembly was observed (Tyler *et al.*, 2006). This suggested that the long-read assembly was more representative of the true genome of *P. fragariae*. The proposed reference assembly was also compared to the two publicly available assemblies of *P. fragariae* and showed an increase in assembly size, contiguity and N50 length (Gao *et al.*, 2015 and Tabima *et al.*, 2017). Therefore, the assembly generated in this study will be useful as a reference genome for further investigations.



### 3.5.4 Conclusions

This study focused on confirming the race of several isolates of *P. fragariae* and carrying out the initial work required for further pathogenomic investigations. The results suggested that Canadian race 1 is equivalent to U.K. race 1, Canadian race 2 is equivalent to U.K. race 3 and Canadian race 3 is equivalent to U.K. race 2. This will aid in the transmission of knowledge about resistance genes between groups working with different race schemes, since the resistance genes are defined by which race of the pathogen they detect. This study also detailed the *de novo* assembly of a sizeable number of isolates of both *P. fragariae* and its closest relative, *P. rubi*. This resulted in high completeness assemblies of ten isolates of *P. fragariae* and three isolates of *P. rubi*, alongside a similarly complete assembly with increased contiguity of the BC-16 isolate of *P. fragariae*. This demonstrated the value of long read sequence data for producing reference genomes of difficult to assemble organisms. However, although the new reference assembly was of increased contiguity, it was still far from a chromosome level assembly, such as that announced recently for the fungal plant pathogen *Fusarium oxysporum* f. sp. *cepae* (Armitage *et al.*, 2018a). The assembly generated here was assembled into 180 contigs, whereas estimates of the diploid chromosome number for *P. fragariae* range between 10 and 12 (Brasier *et al.*, 1999). This was likely due to the high repeat content and large genome size of *P. fragariae*. The resources generated here will prove valuable for investigations discussed in further chapters, as well as for the scientific community as a whole.

## **Chapter 4: Genome annotation and initial investigations for the identification of race-specific effectors**

### **4.1 Introduction**

#### **4.1.1 Inoculation methods for *in vitro* plants with root infecting *Phytophthora* spp.**

Methods of inoculation of strawberry plants with the infectious zoospore life stage of *P. fragariae* have previously been described by Goode (1956), though these plants were returned to sand following inoculation. More recently, the use of an aeroponics system allowed for the propagation of *Eucalyptus marginata* roots in a soil free system. These roots were then inoculated with zoospores of *Phytophthora cinnamomi* and harvested four days after inoculation (Jackson *et al.*, 2000). A similar method was used for the investigation of *Phytophthora parasitica* infecting roots of *Lupinus angustifolius*, raised in Vermiculite and inoculated through immersion of the roots in a zoospore suspension (Blackman *et al.*, 2015).

Zoospore production in *Phytophthora* spp. is challenging, however the production has been performed through a variety of methods depending on the species being investigated. For *Phytophthora sojae* (formerly *Phytophthora megasperma* f. sp. *glycinea*), a plate of growing mycelium is simply flooded with distilled water, causing the growth of the zoospore containing sporangia (Ward *et al.*, 1989). A similar method is used for *Phytophthora infestans*, where the mycelium is grown on rye agar before flooding with distilled water to allow the development of sporangia and a cold shock is performed to liberate zoospores (Rohwer *et al.*, 1987). Whilst methods of zoospore suspension preparation for *P. fragariae* have been described (Hickman and English, 1951), this method relied on the submerging of mycelium growing on agar in pond water. Later work showed that sterilising the pond water by either autoclaving or filter sterilisation inhibited the development of zoospores and a large amount of variation was observed on the

ability of zoospores to develop depending on the source of the pond water (Goode, 1956).

#### **4.1.2 Molecular detection of *Phytophthora fragariae* infection**

Although *P. fragariae* causes clear symptoms on *F. × ananassa*, as first described by Hickman (1941), the red core symptom characteristic of this pathogen only develops later in the infection process. In order to detect *P. fragariae* infection earlier in the infection process, molecular methods have been developed to detect the presence of this pathogen. A species-specific nested Polymerase Chain Reaction (PCR) of the Internal Transcribed Spacer (ITS) region was described by Bonants *et al.* (1997). Enzyme-linked immunosorbent assay (ELISA) methods have also been developed for *P. fragariae*, however, these have lacked the specificity and sensitivity of the previously described PCR methods (Amouzou-Alladaye *et al.*, 1988; Olsson, 1995). This PCR method allows for the detection of *P. fragariae* infection long before symptom development.

#### **4.1.3 Methods of gene prediction and the identification of effector genes**

Gene prediction is a key bioinformatic process for the analysis of genomes that has advanced rapidly in recent years. The current standard for gene prediction process is to use RNA-Seq data as a guide to predict genes, including those that may not be included in the RNA-Seq data due to differences in expression levels in different conditions. One way in which RNA-Seq data is used is to train a model of hits for predicting genes, such as implemented in AUGUSTUS (Stanke *et al.*, 2008). AUGUSTUS is used within BRAKER1 as a method to predict gene models from aligned RNA-Seq data (Hoff *et al.*, 2015) with GeneMark-ET used to aid in training hints (Lomsadze *et al.*, 2014). An alternative method is to simulate a transcriptome assembly and use the assembled transcriptome as hints for predicting gene models, such as implemented in the CodingQuarry algorithm through the use of a Hidden Markov Model (HMM; Testa *et al.*, 2015). It is common in studies of *Phytophthora* spp. to supplement these RNA-Seq guided gene models with additional unguided models in order to ensure the entire gene

set is captured for analysis (Win *et al.*, 2006; Armitage *et al.*, 2018b). Prior to predicting effectors, as described in Chapter 1, it is common to identify genes with secretion signals in order to reduce the number of false positive effector identifications. Previous studies in *Phytophthora* spp. have used multiple versions of the SignalP software, namely versions 2.0, 3.0 and 4.1, in order to ensure all secreted proteins were captured (Nielsen *et al.*, 1997; Bendtsen *et al.*, 2004; Petersen *et al.*, 2011; Armitage *et al.*, 2018b). Phobius can also provide evidence of secretion signals (Käll *et al.*, 2007). However, these methods can incorrectly predict genes as secreted when they are in fact transmembrane bound, these proteins can be identified using programmes such as TMHMM (Krogh *et al.*, 2001), for identifying proteins with transmembrane helices, and GPI-SOM (Frankhauser and Mäser, 2005), for identifying proteins that may be GPI anchored.

#### **4.1.4 Changes in pathogen gene expression over time during infection by *Phytophthora* spp.**

Several studies in *Phytophthora* spp. have shown that the levels of expression of effector genes vary over time. Following the sequencing, genome assembly and annotation of *P. infestans* by Haas *et al.* (2009), a microarray was produced. Through the use of this microarray, it was shown that 494 genes increased their expression by at least a factor of two, this included 79 RxLR effectors, however, it was also shown that mycelia growing in the necrotrophic phase of infection showed a similar expression pattern to mycelium growing in plant extract media. The authors also showed that apoplastic effector genes were among the most upregulated genes and that although only a few Crinkler effectors (CRNs) were induced during infection, those which were induced were expressed at a high level. Additionally, metabolic enzymes were upregulated during infection and ~115 genes were down regulated, including elicitor-like genes (Haas *et al.*, 2009). Later, large scale RNA-Seq analyses in *Phytophthora capsici* during a number of *in vitro* life stages (mycelium, zoospores and germinating cysts with germ tubes) showed clear differences between the life stages and allowed for the analysis of a subset of effector genes known or thought to be involved in virulence, the majority of which were differentially expressed in the germinating cysts compared

to mycelium and zoospores. A further subset of these genes were analysed by Reverse Transcription Polymerase Chain Reaction (RT-PCR) in the same life stages, with additional samples from sporangia and seven time points following inoculation on *Nicotiana benthamiana* roots. All of these effector genes showed increased expression in infection stages compared with developmental stages (Chen *et al.*, 2013).

Interestingly, the levels of induction of RxLR effectors during infection has been shown to differ between isolates of *P. infestans* which displayed different infection phenotypes in terms of recognition by plant resistance genes and the length of the biotrophic phase. This work allowed for explaining how the key resistance gene, thought to be R7, in the 'Stirling' cultivar of potato (*Solanum tuberosum*) was broken in the field (Bradshaw *et al.*, 2004; Cooke *et al.*, 2012).

Additionally, a large-scale microarray experiment on *P. capsici* infecting tomato (*Solanum lycopersicum*) at three life stages (sporangia/zoospores, mycelium and germinating cysts) and six time points post inoculation showed dynamic expression level changes for RxLR effectors as the infection progressed, allowing the authors to define four classes of RxLR effectors. This suggested that transcriptional control is important for the infection process both during initiation of infection and the progression of the infection cycle (Jupe *et al.*, 2013). The same group additionally investigated CRNs in *P. capsici* again infecting tomato. Through analysis of the expression of CRNs at six time points post inoculation, the authors identified two classes of CRNs. One class of genes peaked at early time points, dropped in biotrophic stages and again rose in later stages. The other class of genes showed little expression early in infection, but peaked later (Stam *et al.*, 2013).

#### **4.1.5 Copy number variation in avirulence genes in *Phytophthora* spp.**

Previous work on *P. infestans* has described methods for detecting copy number variation (CNV) from Illumina sequencing data (Raffaele *et al.*, 2010; Cooke *et al.*, 2012; Pais *et al.*, 2018). This allowed for the inference of copy number, even when the assemblies were difficult to assemble. The method used by these

authors was based on a method described by Yoon *et al.* (2009). This method involves the calculation of average read depth and the adjustment of this value for average coverage and GC content, to control for sequencing biases. Whilst several avirulence genes in *P. infestans* displayed CNV among both avirulent and virulent isolates, investigations into these genes revealed that the CNV did not explain the differences by itself (Raffaele *et al.*, 2010; Cooke *et al.*, 2012; Pais *et al.*, 2018). This fits with the two-speed genome model in *P. infestans*, with effector genes being present in repetitive, transposon rich regions and therefore more prone to duplications (Dong *et al.*, 2015). Avirulence genes have also been identified which showed CNV in the more closely related clade 7b species *P. sojae*, however again the control of avirulence was not shown to be directly caused by the levels of CNV in different isolates (Dong *et al.*, 2009; Qutob *et al.*, 2009).

#### **4.1.6 Avirulence genes and presence/absence variation in *Phytophthora* spp.**

The simplest explanation for variation in the virulence of isolates of an organism on host plants is that of presence/absence variation in an avirulence gene. One example of this variation was shown in the Dothideomycete pathogen of Brassica crops, *Leptosphaeria maculans*. The avirulence gene *AvrLm1* was identified through map-based cloning as the gene causing avirulence on plants containing the *Rlm1* resistance gene. Virulent isolates of the pathogen did not show amplification by PCR and hence lacked this gene (Gout *et al.*, 2006). Within the *Phytophthora* genus, the *Avr1d* gene in *P. sojae* was identified as an RxLR effector resulting in recognition and resistance by the *Rps1d* gene in soybean (*G. max*). This gene was identified through the creation of several virulent and avirulent F<sub>2</sub> isolates. The gene was always present in avirulent isolates, but lacking in virulent isolates when assessed by PCR (Yin *et al.*, 2013). Within *P. infestans*, the avirulence gene *PiAvr4* was identified through a combined mapping, transcriptional profiling and bacterial artificial chromosome marker landing approach. This gene was identified as an RxLR effector that was always present in isolate avirulent to potato (*S. tuberosum*) plants containing *R4*. In virulent isolates, a frameshift mutation resulted in a truncated protein (van Poppel

*et al.*, 2008). Therefore, whilst this is not complete presence/absence variation, this variation may be identifiable through analysis of the gene complements of isolates varying in their virulence phenotypes. Recently, the detection of these variations has been parallelised in a method known as pathogen enrichment sequencing (PenSeq). This method was developed for *P. infestans* and *P. sojae*. In this method, extracted DNA is enriched for motifs with known involvement in pathogenicity, in this case RxLR effectors, which are subsequently next generation sequenced. This provides a cost-effective alternative to whole genome sequencing and simplifies the process of effector gene identification (Thilliez *et al.*, 2018).

## 4.2 Aims

In order to conduct pathogenomic investigations, the assembled isolates from Chapter 3 were *ab initio* annotated using RNA-Seq data generated from an inoculation time course experiment for *P. fragariae* and using reads sourced from the Grünwald Lab at Oregon State University for *Phytophthora rubi*. Following this, effector genes were predicted and investigations were carried out to test for any genetic explanation for the differences in pathogenicity between different races of *P. fragariae*.

The specific aims of this chapter were as follows:

1. Develop a method to inoculate micropropagated plants in order to generate RNA-Seq data.
2. Develop a method to detect the presence of *P. fragariae* via RT-PCR in extracted total RNA.
3. Perform *ab initio* gene prediction on the assembled genomes of *P. fragariae* and *P. rubi*, predicting effector genes as a part of this process.
4. Conduct a differential expression analysis to assess the efficacy of the time course experiment to capture different stages of the infection process.
5. Align raw sequencing reads to investigate any CNV.

6. Conduct an orthology analysis and investigate the resultant orthogroups for race specific genes that could be involved in the resistance response.



## **4.3 Materials and Methods**

### **4.3.1 Production of flooding solutions**

Four flooding solutions were used as a part of this investigation: compost extract, lake water, stream water and Petri's solution. Compost extract was produced as described in Nellist *et al.* (2018), briefly, 50 g of compost was placed in 2 L of dH<sub>2</sub>O and incubated at room temperature overnight in the dark. Following this, the solution was passed through 150 mm 113V wet strengthened Whatman filter paper (GE Healthcare Life Sciences, Little Chalfont, U.K.). Before usage, this solution was diluted at a 1:1 ratio with dH<sub>2</sub>O. Lake water was collected from an onsite lake at Bradbourne House (Latitude: 51.294107, Longitude: 0.442702, Plus code: 7CVV+J3 East Malling). Stream water was collected from an onsite stream at Bradbourne House (Latitude: 51.295221, Longitude: 0.444462, Plus code: 7CWV+3Q East Malling). Petri's solution was made as described in Judelson *et al.* (1993) to the following concentration: 1 mM KCl (Sigma Chemical Co., St. Louis, MO, U.S.A.), 2mM Ca(NO<sub>3</sub>)<sub>2</sub> (BDH laboratory reagents, Poole, U.K.), 1.2 mM MgSO<sub>4</sub> (Fisher Scientific, Loughborough, U.K.) and 1 mM KH<sub>2</sub>PO<sub>4</sub> (Fisher Scientific, Loughborough, U.K.). This solution was then autoclaved at 121°C for 20 minutes. All solutions were chilled to 4°C before use.

### **4.3.2 Assessment of sporangia growth and production of zoospores**

Isolates were subcultured as described in Chapter 2, section 2.3. A sterilised number 6 cork borer (10mm diameter) was then used to take plugs of mycelium, with the agar below, from the leading edge of the mycelium before being transferred to a fresh 90 mm triple vent petri dish (Thermo Scientific, Waltham, MA, U.S.A.) with a sterilised cocktail stick. The appropriate flooding solution (described above) was then added to the plate to cover the plugs and the plates were sealed with Parafilm® (Bemis, Oshkosh, WI, U.S.A.). The plates were then incubated in constant light for three days at 13°C, with the flooding solution replaced every 24 hours. Following this incubation period, sporangia growth was observed under an Olympus SZ-ST SZ11 dissecting microscope (Waltham, MA, U.S.A.).

Following the incubation period, the plates were transferred to a 4°C cold room in constant light for 1 hour and then to room temperature in constant light for 1 hour in an attempt to prompt the sporangia to burst. Plates were again observed under an Olympus SZ-ST SZ11 dissecting microscope (Waltham, MA, U.S.A.) to assess the relative proportion of burst sporangia. The plates were then filtered through 90 mm Q100 filter paper (Fisher Scientific, Loughborough, U.K.) using a Büchner funnel and Büchner flask attached to a 2016 VacuGene Vacuum Blotting Pump (LKB, Stockholm, Sweden). The filtrate was then examined on a haemocytometer under a Leitz transmission microscope (Leica Microsystems, Wetzlar, Germany) after staining with Victoria Blue.

#### **4.3.3 Methods of inoculation of micropropagated *Fragaria* × *ananassa* plants with *Phytophthora fragariae***

Two methods of inoculation were trialled in this chapter, these were named 'slurry' and 'plugs' for ease of reference. The BC-16 isolate was used to inoculate *F.* × *ananassa* plants of the universal susceptible cultivar 'Hapil' and the BC-1 isolate was used to inoculate *F.* × *ananassa* plants of the cultivar 'Redgauntlet'.

The plugs method began as described above for sporangia growth, following the results of trial experiments; KBA, Green Bean Agar (GBA) and Broad Bean Agar (BBA) plates were used with stream water as the flooding solution. Plates were then unsealed in a Tri-MAT class-II microbiological safety cabinet (Contained Air Solutions, Manchester, U.K.) and the roots of approximately five micropropagated plants, of cultivars 'Redgauntlet' and 'Hapil' maintained on *Arabidopsis thaliana* salts (ATS) plates (see Chapter 2, section 2.2 and 2.4), were submerged in the solution contained in the plates for one hour. After this time, plants were transferred back to the ATS plates, similar to the method described in Blackman *et al.* (2015).

The slurry method of inoculation was performed by blending excised colonies of mycelium and the agar beneath with ice H<sub>2</sub>O (1 g culture : 1 g ice H<sub>2</sub>O) twice for two seconds. The resulting slurry was transferred to a cooled beaker which was

kept on ice during the inoculation procedure. Plants of cultivars 'Redgauntlet' and 'Hapil' maintained on ATS plates (see Chapter 2, section 2.2) were inoculated by dipping their roots into the slurry and then replaced onto ATS plates, similar to the method used in van de Weg *et al.* (1996).

Following inoculation, plates were sealed with Sellotape (Winsford, U.K.) and transferred to a Panasonic MLR-325H controlled environment test chamber (Osaka, Japan) at 15°C, with a 16-hour day length, a relative humidity of 85% and lights at setting 4. Roots were then harvested immediately post inoculation for mock inoculated plants and after twenty-one days for inoculated plants. Prior to harvesting, plant roots were rinsed in a succession of three beakers of dH<sub>2</sub>O in order to remove any media from the roots. Plants were harvested by separating roots and leaves with sterile scissors and discarding the crown tissue, then placing the root tissue and the leaf tissue into separate 2 mL Eppendorf tubes (Stevenage, U.K.) and immediately flash-frozen in liquid nitrogen. Material was then stored at -80°C.

#### **4.3.4 DNA extraction and testing for presence of *Phytophthora fragariae* in inoculated micropropagated *Fragaria × ananassa* plants**

DNA was extracted from inoculated plant roots using the QIAGEN DNEasy Plant Mini kit as per the manufacturer's instructions (Venlo, The Netherlands), with the initial disruption performed using sterile magnetic ball bearings in a 2010 Geno/Grinder<sup>®</sup> (SPEX<sup>®</sup> SamplePrep, Metuchen, NJ, U.S.A.) as per the manufacturer's instructions, ensuring tissue remained frozen by immersing the tubes and holders in liquid nitrogen periodically. Following extraction, a previously published PCR method for detection of *P. fragariae* was used to assess the efficiency of the various inoculation methods trialled (Bonants *et al.*, 1997). This nested PCR was performed in a Veriti 96-well thermocycler (Applied Biosystems, Foster City, CA, U.S.A.) exactly as described in Bonants *et al.* (1997). Briefly, this procedure consisted of two PCR steps. The first reaction amplified the entire Internal Transcribed Spacer (ITS) region and the second reaction amplified a segment of the ITS sequence specific to *P. fragariae*. The results of this PCR were analysed by electrophoresis on a 1% w/v agarose

(Fisher Bioreagents, Fairlawns, NJ, U.S.A.) gel in Tris-Acetate-EDTA (TAE) buffer, consisting of 1 mM EDTA (Ambion, Foster City, CA, U.S.A.), 40 mM Tris (Fisher Bioreagents, Fairlawns, NJ, U.S.A.) and 20 mM glacial acetic acid (Fisher Scientific, Loughborough, U.K.). 0.003% by volume Gel-Red (Biotium, Fremont, CA, U.S.A.) was added before polymerisation of the gel. The gel was run at 100 V for 1 hour and visualised on a GelDoc XR+ (BioRad, Watford, U.K.). Products were sized using a 100 bp Plus Ladder (ThermoFisher Scientific, Waltham, MA, U.S.A.).

#### 4.3.5 Design of novel primers for testing for the presence of *Phytophthora fragariae* in inoculated micropropagated *Fragaria* × *ananassa* plants

The  $\beta$ -tubulin gene of the BC-16 isolate of *P. fragariae* was identified via a search with the Basic Local Alignment Search Tool (BLAST) algorithm BLASTn in Geneious R10 (Kearse *et al.*, 2012) against the gene sequence from the AY564062.1 accession of *P. fragariae* from GenBank (NCBI Resource Coordinators, 2018). Using the modified Primer3 version 2.3.7 implemented in Geneious R10 (Untergasser *et al.*, 2012) a number of candidate primer pairs were designed (Table 4.1).

**Table 4.1: Primer sequences used to assess the presence of *Phytophthora fragariae* through detection of  $\beta$ -tubulin**

Primer Name	Primer Sequence
Btub1_F	5'-AGGAGATGTTCAAGCGCGTG-3'
Btub1_R	5'-GGTCGTTTCATGTTGGACTCG-3'
Btub2_F	5'-GGATAACGAGGCCCTGTACG-3'
Btub3_F	5'-CACGGCCGCTATTTAACTGC-3'
Btub2_R	5'-TGTTGTTGGGGATCCACTCG-3'

These primers were tested against gDNA extracted from *in vitro* grown mycelium of BC-16 as a positive control, gDNA extracted from *F. × ananassa* plants of the 'Hapil' cultivar as a control for non-specific binding of the primers and dH<sub>2</sub>O as a negative control. The annealing temperature was also varied for each primer pair, the following temperatures were trialled: 50°C, 52°C, 55°C, 57°C, 60°C and

62°C. The PCR mix consisted of 2.5 units of *Taq* DNA polymerase in the buffer supplied at its recommended concentration (ThermoFisher Scientific, Waltham, MA, U.S.A.), 200 µM dNTPs (Fermentas, Waltham, MA, U.S.A.), 0.2 µM of each primer and DNA stocks in a 15% v/v solution with dH<sub>2</sub>O. 20 µL reactions were performed in a Veriti 96-well thermocycler (Applied Biosystems, Foster City, CA, U.S.A.) with an initial denaturation step at 95°C for 3 minutes, followed by 35 cycles of a denaturation step of 95°C for 30 seconds, an annealing temperature as described above for 30 seconds and an extension step of 72°C for 30 seconds. This was followed by a final extension step of 72°C for 5 minutes before being held at 10°C. The results of this PCR were analysed by electrophoresis on a 1% w/v agarose (Fisher Bioreagents, Fairlawns, NJ, U.S.A.) gel in Tris-Acetate-EDTA (TAE) buffer, consisting of 1 mM EDTA (Ambion, Foster City, CA, U.S.A.), 40 mM Tris (Fisher Bioreagents, Fairlawns, NJ, U.S.A.) and 20 mM glacial acetic acid (Fisher Scientific, Loughborough, U.K.). 0.0012% by volume Gel-Red (Biotium, Fremont, CA, U.S.A.) was added before polymerisation of the gel. The gel was run at 100 V for 1 hour and visualised on a GelDoc XR+ (BioRad, Watford, U.K.). Products were sized using a 100 bp plus ladder (ThermoFisher Scientific, Waltham, MA, U.S.A.).

The amplicons were excised from the gel using a razor blade and the DNA was extracted from the agarose using the Macherey-Nagel NucleoSpin® Gel and PCR Clean-up kit (Macherey-Nagel GmbH & Co. KG, Düren, Germany) following the manufacturer's instructions. 5 µL of the extracted DNA stock, with 5 µL of 5 µM stock of either the forward or reverse primer, was then analysed by Sanger Sequencing by GATC's (now Eurofins) LightRun service (Eurofins Genomics, Ebersberg, Germany).

#### **4.3.6 Preparation of inoculated micropropagated *Fragaria* × *ananassa* plant material for RNA extraction**

Prior to inoculation, micropropagated *F. × ananassa* plants of the 'Hapil' cultivar were subcultured on to fresh ATS plates approximately 24 hours before inoculation. These plants were then inoculated with the BC-16 of *P. fragariae* using the above described 'plugs' method. In order to provide a time course of

the infection process, root tissue was harvested from both uninoculated plants and from inoculated plants at: 24 hours post inoculation (hpi), 48 hpi, 96 hpi, 144 hpi, 192 hpi and 240 hpi with stream water used as the final flooding solution in the preparation of the inoculum. Additionally, roots were harvested at 24 hpi, 48 hpi, 96 hpi and 144 hpi with Petri's solution used as the final flooding solution in the preparation of the inoculum.

#### **4.3.7 RT-PCR analysis of extracted RNA to test for the presence of *Phytophthora fragariae***

Extracted RNA was then analysed for the presence of actively growing *P. fragariae* through an RT-PCR based method utilising the best performing pair of primers described above, Btub2\_F and Btub2\_R. Reverse transcription was performed using the SuperScript™ III Reverse Transcriptase kit (ThermoFisher Scientific, Waltham, MA, U.S.A.) following the manufacturer's instructions, with an equal amount of RNA template added and the reaction made up to 13 µL with dH<sub>2</sub>O. The resulting cDNA was then transferred to a PCR reaction, with 200 µM dNTPs (Fermentas, Waltham, MA, U.S.A.), 0.2 µM of each primer, 2 µL of cDNA template and 2.5 units of *Taq* DNA polymerase and the buffer supplied (ThermoFisher Scientific, Waltham, MA, U.S.A.) in a 20 µL reaction. This was then run in a Veriti 96-well thermocycler (Applied Biosystems, Foster City, CA, U.S.A.) with an initial denaturation step at 95°C for 30 seconds, followed by 35 cycles of a denaturation step at 95°C for 30 seconds, an annealing temperature of 60°C for 30 seconds and an extension step of 72°C for 30 seconds. This was followed by a final extension step of 72°C for 5 minutes before being held at 10°C.

The product of this PCR reaction was then analysed by electrophoresis on a 1% w/v agarose (Fisher Bioreagents, Fairlawns, NJ, U.S.A.) in Tris-Acetate-EDTA (TAE) buffer, consisting of 1 mM EDTA (Ambion, Foster City, CA, U.S.A.), 40 mM Tris (Fisher Bioreagents, Fairlawns, NJ, U.S.A.) and 20 mM glacial acetic acid (Fisher Scientific, Loughborough, U.K.). 0.003% by volume Gel-Red (Biotium, Fremont, CA, U.S.A.) was added before polymerisation of the gel. The gel was run at 80 V for 90 minutes and visualised on a GelDoc XR+ (BioRad,

Watford, U.K.). Products were sized using a 100 bp plus ladder (ThermoFisher Scientific, Waltham, MA, U.S.A.).

The results of the RT-PCR analyses were used to select which time points from the inoculation time course experiment to sequence. Three samples for each timepoint, as well as three from *in vitro* grown mycelium were selected for sequencing based on RNA integrity, RNA concentration and the presence of any contaminants based on the 260/230 and 260/280 ratios. These samples were then sequenced by Novogene (Hong Kong, Special Administrative Region, China).

#### **4.3.8 Initial assessment of RNA-Seq data and cleaning of raw data**

RNA-Seq reads from Novogene were downloaded to the NIAB EMR cluster (see Chapter 2, section 2.10). Following this, poor quality reads were removed, sequences were trimmed and Illumina adapters were removed using fastq-mcf (Aronesty, 2013). Following this, the data quality was visualised using fastqc version 0.10.1 (Andrews, 2010) to provide statistics on several key metrics. These included: sequence quality, sequence content, GC content, N content, sequence length distribution, sequence duplication, the presence of overrepresented sequences and the *K*-mer content.

#### **4.3.9 Alignment of RNA-Seq reads to genome assemblies**

*P. fragariae* reads from *in vitro* grown mycelium were aligned to all *P. fragariae* genome assemblies using Spliced Transcripts Alignment to a Reference (STAR) version 2.5.3a (Dobin *et al.*, 2013) with the following options specified to improve sensitivity: --winAnchorMultimapNmax 200 --seedSearchStartLmax 30. Reads from inoculated micropropagated *F. × ananassa* plant roots were first aligned to the *Fragaria vesca* version 1.1 genome (Shulaev *et al.*, 2011) with STAR version 2.5.3a running with the previously specified additional options, and a file of non-mapping reads was created in order to remove reads from the host plant material. These unmapping reads were then aligned to the *P. fragariae* genome

assemblies, again using STAR version 2.5.3a with the previously specified additional options.

RNA-Seq data of *P. rubi* was acquired from the Grünwald Lab at Oregon State University as part of the *Phytophthora* sequencing consortium. These reads were aligned to the *P. rubi* genome assemblies, again using STAR version 2.5.3a running with the same options specified for the alignments of *P. fragariae* reads to the *P. fragariae* assemblies.

Aligned data was then concatenated using the Sequence Alignment/Map tools (SAMtools) version 1.5 (Li *et al.*, 2009) merge function to produce a single file containing all aligned RNA-Seq data to be used as evidence for gene prediction.

#### **4.3.10 *Ab initio* annotation of *Phytophthora fragariae* and *Phytophthora rubi* genome assemblies**

##### **4.3.10.1 Initial gene prediction**

Gene models were created using both RNA-Seq guided and unguided methods. Firstly, gene models were created using BRAKER1 version 2.0.1 (Hoff *et al.*, 2015) with the fungus option selected, using the concatenated RNA-Seq alignments as a training set. The BRAKER1 program uses GeneMark-ET (Lomsadze *et al.*, 2014), installed via the GeneMark-ES suite version 4.33, to train hints for AUGUSTUS version 3.1 (Stanke *et al.*, 2008). Guided predictions were also generated using CodingQuarry version 2.0 (Testa *et al.*, 2015) using the pathogen option, which used a transcriptome assembly from Cufflinks version 2.2.1 (Trapnell *et al.*, 2010) as a guide. Transcripts from CodingQuarry were added to the BRAKER1 predicted genes when CodingQuarry genes were predicted in regions of the genome not containing BRAKER1 models, using the BEDTools intersect function (Quinlan and Hall, 2010), as described in Armitage *et al.* (2018b). A final check was then performed using a custom Python script to remove any duplicated transcripts.



An unguided set of gene predictions were produced for all isolates by translating sequences following all start codons in the genome, with a stop codon within 50-250 amino acids, as in Armitage *et al.* (2018b).

#### **4.3.10.2 Prediction of effector genes**

Predicted genes from both guided and unguided methods were initially assessed for evidence of secretion. These were predicted via SignalP-2.0 (Nielsen *et al.*, 1997), SignalP-3.0 (Bendtsen *et al.*, 2004), SignalP-4.1 (Petersen *et al.*, 2011) and Phobius version 1.1 (Käll *et al.*, 2004). Following this, three effector classes were predicted: RxLR effectors, CRNs and apoplastic effectors.

RxLR effectors were predicted by two complementary methods. Firstly, a motif search was performed, based on previous N-terminal RxLR identification methods (Torto *et al.*, 2003). The Python regular expression (R.LR.{,40}([ED][ED][KR])) was used to identify putative RxLR effectors, as in Armitage *et al.* (2018b). This search scored a protein as possibly being an RxLR effector if it matched the following criteria: the presence of an RxLR motif up to 100 aa downstream of the signal peptide cleavage point, the presence of an EER motif within 40 aa downstream of the RxLR motif, and the predicted signal peptide cleavage site being between the 10th and 40th amino acid with a HMM score greater than 0.9. A set of lower confidence RxLR effectors was also identified where the presence of an EER domain was not required. RxLR effectors were also identified using a previously published HMM (Whisson *et al.*, 2007). Proteins scoring positive for potentially being an RxLR effector by this model were identified with HMMER version 3.1b2 (<http://www.hmmer.org>, last accessed 03/01/2019).

Crinkler effectors were identified using HMMs specific to two domains characteristic of crinkler proteins. These models were described in Armitage *et al.* (2018b). Alignment of these sequences allowed models to be trained to two key domains: LFLAK and DWL. Proteins scoring positive by this model were identified with HMMER version 3.1b2 (<http://www.hmmer.org>, last accessed

03/01/2019). Gene models were called as putative CRNs if they scored positive in analysis with both HMMs.

Apoplastic effectors were identified by screening, with ApoplastP version 1.0, a subset of all proteins predicted as secreted by the above described method (Sperschneider *et al.*, 2018).

Effector identification and secreted protein prediction was carried out on both guided and unguided gene predictions. Only unguided gene models showing evidence of secretion by SignalP were used for effector prediction. As the unguided gene models prediction method produced a number of overlapping features, redundancy was removed by retaining the gene model with the highest SignalP HMM score. An exception was in the prediction of crinkler effectors, where the LFLAK HMM score was used for merging overlapping features. As Phobius predictions do not produce a HMM score, they were not used to inform the removal of overlapping unguided gene models, however the results were retained as another layer of evidence for verification of proteins predicted to be secreted.

#### **4.3.10.3 Creation of a final gene model set**

Prior to further analysis, intersects of the various effector classes were identified using the BEDTools intersect function (Quinlan and Hall, 2010). This allowed the merging of predicted unguided effector gene models sited in intergenic regions of the BRAKER1 and CodingQuarry predicted transcripts. The lower confidence set of RxLR effectors from unguided models was used for this purpose, to ensure all possible effector genes were captured. ApoplastP was found to have an issue leading to the prediction of duplicate genes, these were removed with a custom python script. Following this, additional intersects between effectors from unguided gene models and all guided gene models were identified using BEDTools (Quinlan and Hall, 2010) and the guided gene models were preferentially kept with a custom Python script. Following this, an additional check for duplicated transcripts was performed, this utilised the same script used after

merging BRAKER1 and CodingQuarry gene models. Genes were also renamed to National Centre for Biotechnology information (NCBI) standards at this point.

Following this, draft functional annotations were identified using InterProScan version 5.18-57.0 (Jones *et al.*, 2014) and through a BLASTp (Altschul *et al.*, 1990) search, with an e-value threshold of  $1 \times 10^{-100}$ , against the Swiss-Prot database as of March 2018 (The UniProt Consortium, 2017). Searches were also performed for the presence of transmembrane helices with TMHMM version 2.0 (Krogh *et al.*, 2001) to test whether proteins predicted as secreted may in fact be transmembrane proteins. The potential of a protein being membrane anchored was also assessed via the GPI-Som web-server (Frankhauser and Mäser, 2005).

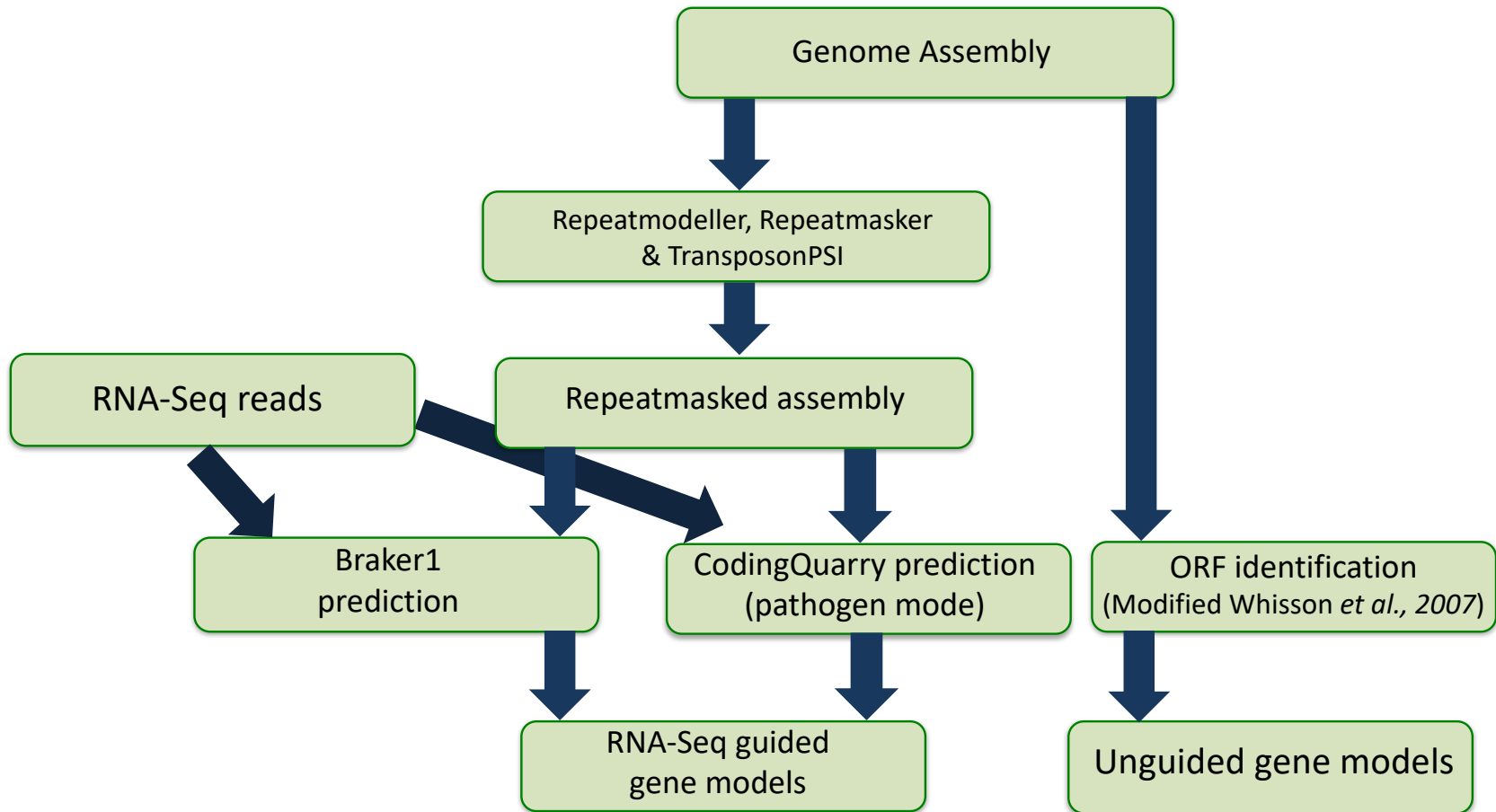
The completeness of the gene model set was then tested by identifying BUSCO genes using BUSCO version 3.0.1 with the eukaryota\_odb9 database (Simão *et al.*, 2015).

Differences in the amount of effector genes and secreted proteins, were assessed for significance using a Welch Two Sample t-test in R version 3.4.3 (R core team, 2017) running on a MacBook Pro, Early 2015, OS version 10.13.6.

#### **4.3.10.4 Submission of draft annotated assemblies to GenBank**

Following creation of a final gene model set and automated annotation, Annie version c1e848b was used to extract the automated functional annotations described above (Tate *et al.* 2014). These annotations were then used to build a genome file with Genome Annotation Generator (GAG) version 98da78e (Geib *et al.*, 2018). Tbl2asn was then used to generate an initial file for submission to NCBI (<https://www.ncbi.nlm.nih.gov/genbank/tbl2asn2/>, last accessed 03/01/2019). This produced an error log, which was used to further correct the formatting of the submission file for NCBI with a Python script (Armitage, unpublished). Finally, tbl2asn was again run to produce a final file for submitting to NCBI after a small number of further manual edits.

The gene prediction process is summarised in Fig. 4.1 below.



**Fig. 4.1: Summary of the gene prediction procedure conducted for all assembled isolates of *Phytophthora fragariae* and *Phytophthora rubi*.** Isolates were assembled as described in Chapter 3. RNA-Seq reads were either: generated from an inoculation time course of the BC-16 *P. fragariae* isolate for annotation of *P. fragariae* assemblies or acquired from the Grünwald Lab at Oregon State University for annotation of *P. rubi* assemblies.

#### **4.3.11 Analysis of differential expression between time points from the inoculation time course experiment and *in vitro* grown mycelium**

Following the alignment of the above described RNA-Seq data and the creation of a final gene model set, an analysis was conducted of the changes in expression level of the predicted genes during the inoculation time course compared to *in vitro* grown mycelium. Firstly, featureCounts version 1.5.2 was used to quantify the number of reads aligned to a gene model in each sample, with the fraction option enabled to improve handling of multi-mapping reads (Liao *et al.*, 2013). Following this, an analysis was conducted using the DESeq2 version 1.10.1 R package to produce normalised FPKM (Fragments Per Kilobase of transcript per Million mapped reads) counts and assess levels of differential expression between samples from the inoculation time course and the samples from *in vitro* grown mycelium (Love *et al.*, 2014).

#### **4.3.12 Assessment of copy number variation**

CNV was assessed in a method similar to that used in Raffaele *et al.* (2010), Cooke *et al.* (2012) and Pais *et al.* (2018). Only the BC-1, BC-16 and NOV-9 isolates of *P. fragariae* were investigated as representatives of the races UK1, UK2 and UK3 respectively. The Illumina reads of all three isolates were aligned to the reference assembly of the BC-16 isolate using Bowtie 2 version 2.2.6 (Langmead and Salzberg, 2012), followed by sorting and indexing with SAMtools version 1.5 (Li *et al.*, 2009). The read depth for each gene in the annotated BC-16 genome was then calculated using the SAMtools version 1.5 bedcov function. A Python script was then used to calculate the average read depth of each gene in each isolate, adjusted for GC content, and the copy number was calculated as described in Pais *et al.* (2018). If a gene had a change in copy number in one isolate relative to both other isolates, it was called as displaying CNV.

#### **4.3.13 Orthology analysis of all proteins predicted for all sequenced *Phytophthora fragariae* and *Phytophthora rubi* isolates**

Genes were assigned to orthology groups using OrthoFinder version 1.1.10 (Emms and Kelly, 2015). This programme relies on an BLASTp analysis (Altschul *et al.*, 1990) of all genes vs all genes to assign the genes to orthology groups. After the creation of orthology groups, groups containing effector genes of isolates in races UK1, UK2 and UK3 were identified for further analysis, alongside groups containing putative secreted proteins for these isolates. Orthogroups that contained only proteins from isolates of one race from the three of interest were called as unique groups and those where proteins from isolates of one race from the races of interest were present more than proteins from isolates of the other races of interest were called as expanded groups. These expanded and unique groups were manually investigated in Geneious R10 (Kearse *et al.*, 2012) to identify any putative avirulence genes for these three races. In a number of cases, BLASTn searches were performed within Geneious R10 to find orthologous sequence regions in isolates of other races. In one case, visualisation of the alignments generated for the investigation of CNV were visualised in Integrative Genomics Viewer (IGV) to check for assembly errors (Thorvaldsdóttir *et al.*, 2013).

#### **4.3.14 Availability of sequencing data and annotated assemblies**

All raw RNA sequencing reads have been submitted to the Sequencing Read Archive (SRA) maintained by NCBI. These will be available following the publication of these results in a peer reviewed journal (Table 4.2).

The annotated assemblies have been submitted to GenBank, maintained by NCBI as part of BioProject PRJNA396163 and will be publicly available following the publication of these results in a peer reviewed journal.

**Table 4.2: Details of the NCBI SRA codes for all sequencing reads uploaded for future release**

Species	Isolate	Sample Details	SRA code
<i>Phytophthora fragariae</i>	BC-16	24 hpi	SRR7764608
<i>Phytophthora fragariae</i>	BC-16	48 hpi	SRR7764609
<i>Phytophthora fragariae</i>	BC-16	96 hpi	SRR7764614
<i>Phytophthora fragariae</i>	BC-16	<i>in vitro</i> mycelium	SRR7764610

Hours post inoculation (hpi)

## 4.4 Results

### 4.4.1 Trial of methods to generate a zoospore suspension of *Phytophthora fragariae* were not sufficient for inoculation experiments

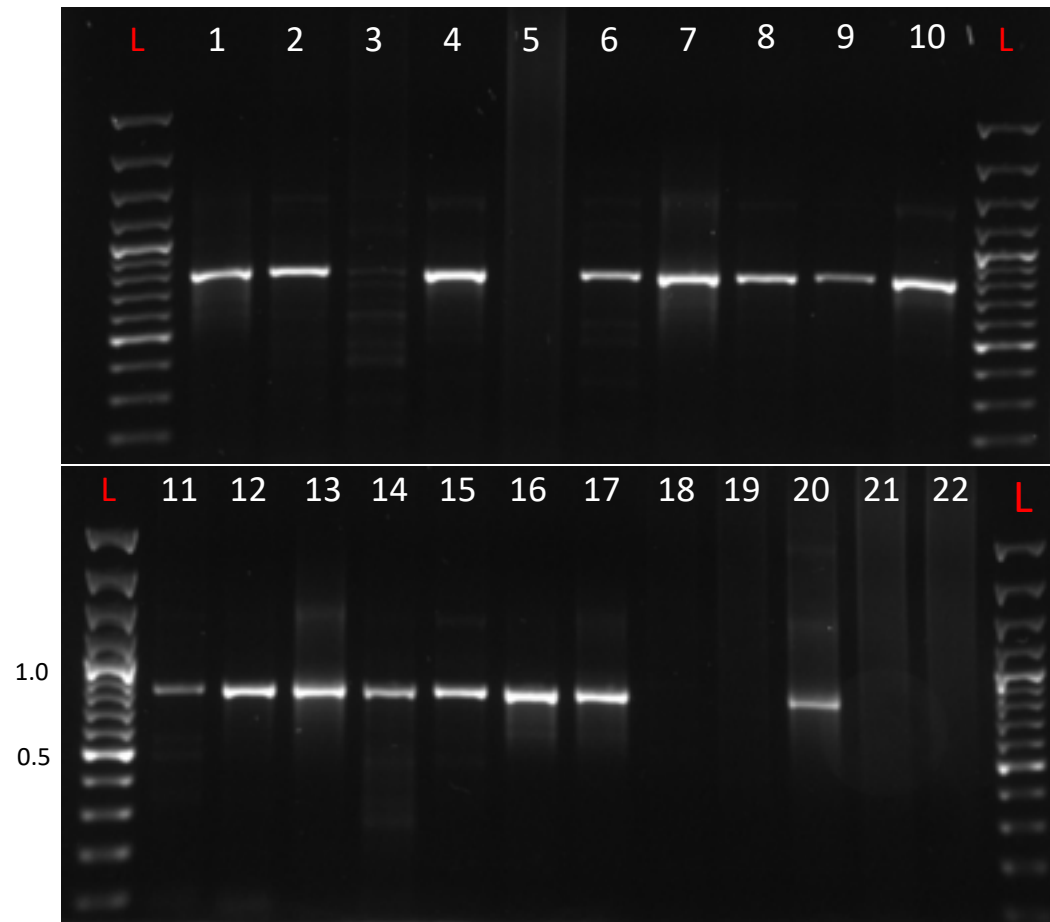
Unlike other members of the *Phytophthora* genus, the generation of a zoospore suspension of *P. fragariae* has proven difficult. Previously, several studies have used the method of Hickman and English (1951), where disks of mycelium growing on agar were submerged in unsterilised pond water for three days. However, attempts to replicate this method with the BC-16, BC-1 and NOV-9 isolates of *P. fragariae* were unsuccessful in producing a zoospore suspension, despite the successful formation of sporangia being observed. Several alternative methods were trialled for producing zoospores of the BC-16 isolate of *P. fragariae*, varying the agar mixture used as an initial growth substrate to be: KBA, Strawberry Material Agar (SMA), Strawberry Extract Agar (SEA), BBA, GBA, Soya Bean Agar (SBA), Edamame Bean Agar (EBA), Pea Broth Agar (PBA) and Mung Bean Shoot Agar (MBSA). Mycelial growth was not observed on either SEA or SMA, perhaps due to the presence of growth inhibitory compounds in the roots. Following flooding with compost extract, sporangia growth was not observed on SBA, EBA, MBSA and PBA. Despite sporangia growth being observed on GBA, BBA and KBA, attempts to trigger sporangia burst with temperature changes were unsuccessful at producing a zoospore suspension.

### 4.4.2 Development of an inoculation procedure for micropropagated *Fragaria* × *ananassa* plants with *Phytophthora fragariae* allowed the conducting of an *in vitro* pathogenicity time course experiment

Following the failure of attempts to generate a spore suspension, alternative methods were trialled for performing inoculation of micropropagated *F. × ananassa* plants with *P. fragariae*. Both the 'plugs' and the 'slurry' method were trialled with BBA, GBA and KBA. Following observations from the attempts to generate a spore suspension, it was determined that stream water was the best solution to use for sporangia growth in the 'plugs' method. This was replaced with



sterile Petri's solution for the final twenty-four hours of sporangia growth in an attempt to reduce the risk of contamination with organisms present in the stream water. Plants of the 'Redgauntlet' cultivar were tested with the U.K. race 1 isolate of *P. fragariae*, BC-1 and plants of the 'Hapil' cultivar were tested with the U.K. race 2 isolate of *P. fragariae*, BC-16. Plant roots were harvested at 21 days post-inoculation. Following DNA extraction from a subset of inoculated roots, a previously reported nested PCR method was used to detect the presence of *P. fragariae* in the inoculated plants (Bonants *et al.*, 1997; Fig. 4.2). These results showed that 'Redgauntlet' plants were successfully inoculated by all preparations except the BBA slurry and that 'Hapil' plants were successfully inoculated by all preparations. It also demonstrated negative results in mock inoculated plants with plugs, but not with slurry. It was therefore decided that KBA plugs was the best method for these inoculations.



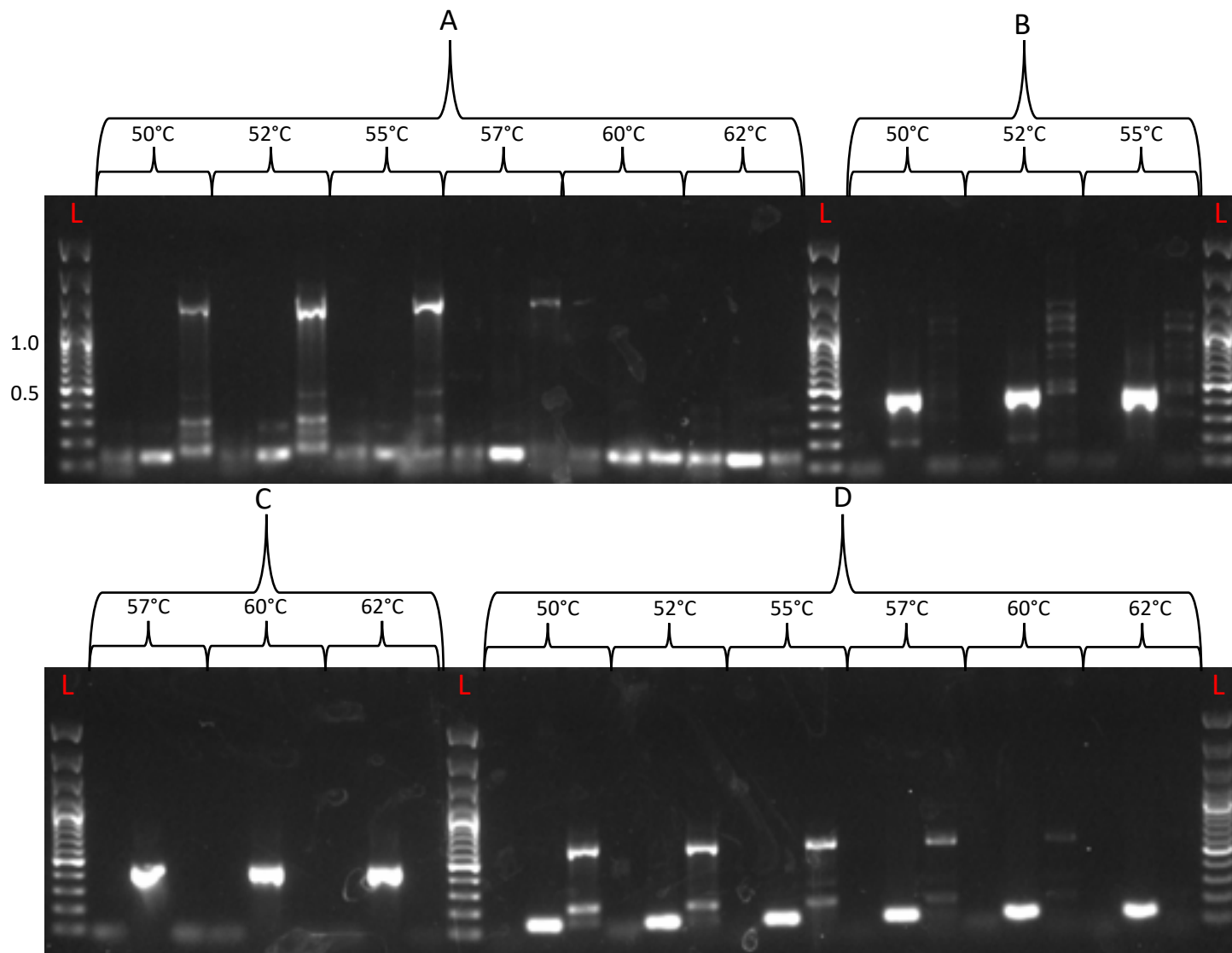
**Fig. 4.2: Agarose gel electrophoresis of a nested PCR to detect *Phytophthora fragariae* in inoculated *Fragaria × ananassa* micropropagated plants. L: 100 bp plus ladder (ThermoFisher Scientific, Waltham, MA, U.S.A.) sizes listed in kb. Additional lanes detailed in the additional table overleaf.**

Lane Number	<i>Fragaria</i> × <i>ananassa</i> cultivar	<i>Phytophthora fragariae</i> isolate	Media Type	Inoculation Method
1	'Hapil'	BC-16	KBA	Plugs
2	'Hapil'	BC-16	BBA	Plugs
3	'Hapil'	None	KBA	Plugs
4	'Redgauntlet'	BC-1	KBA	Slurry
5	'Redgauntlet'	BC-1	BBA	Slurry
6	'Redgauntlet'	BC-1	GBA	Slurry
7	'Redgauntlet'	BC-1	KBA	Plugs
8	'Redgauntlet'	BC-1	BBA	Plugs
9	'Redgauntlet'	BC-1	GBA	Plugs
10	'Hapil'	BC-16	KBA	Slurry
11	'Hapil'	BC-16	BBA	Slurry
12	'Hapil'	BC-16	GBA	Slurry
13	'Hapil'	BC-16	KBA	Plugs
14	'Hapil'	BC-16	BBA	Plugs
15	'Hapil'	BC-16	GBA	Plugs
16	'Redgauntlet'	None	KBA	Slurry
17	'Hapil'	None	KBA	Slurry
18	'Redgauntlet'	None	KBA	Plugs
19	'Hapil'	None	KBA	Plugs
20	None	BC-16	Pea Broth	<i>in vitro</i> mycelium growth
21	None	None	None	dH <sub>2</sub> O control
22	'Hapil'	None	None	Non-inoculated, from leaf tissue

#### **4.4.3 Design of primers for assessing presence of *Phytophthora fragariae* from RNA extracted from inoculated *Fragaria* × *ananassa* micropropagated plants resulted in the development of a suitable PCR based assay for detection of *P. fragariae***

Whilst the published nested PCR method for detection of *P. fragariae* had proven successful at identifying infected tissue following DNA extraction, it was not suitable for assessing the levels of infection from RNA samples, as the nested PCR target was the ITS region and so was not transcribed by the pathogen. Primers were designed for the housekeeping gene  $\beta$ -tubulin (Table 4.1). These primers were tested against dH<sub>2</sub>O as a no template control, BC-16 mycelial gDNA as a positive control and *F. × ananassa* cultivar 'Hapil' gDNA to assess the specificity of the amplification, PCR conditions were also optimised (Fig. 4.3). These results demonstrated that whilst all the primer pairs produced a distinct product in BC-16 gDNA, the product of Btub2\_F and Btub2\_R produced a larger target product at 440 bp that was more easily distinguished from primer-dimers formed when *P. fragariae* gDNA was not present. An annealing temperature of 60°C was selected as this showed the minimum primer dimer presence in the dH<sub>2</sub>O reaction and the minimum non-specific binding in the 'Hapil' gDNA reaction. Following this, the product of this PCR was gel-extracted and sanger sequenced. These sequencing reads were then aligned to the gene sequence and confirmed that the product was the *P. fragariae* isolate BC-16  $\beta$ -tubulin gene (Fig. 4.4).

**Fig. 4.3 (overleaf): Agarose gel electrophoresis of three primer pairs testing for the presence of *Phytophthora fragariae* isolate BC-16 by detection of  $\beta$ -tubulin. L:** 100 bp Plus Ladder (ThermoFisher Scientific, Waltham, MA, U.S.A.) sizes listed in kb. **A:** All lanes used the primer pair Btub1\_F and Btub1\_R, product size 129 bp. **B:** All lanes used the primer pair Btub2\_F and Btub2\_R, product size 440 bp. **C:** All lanes used the primer pair Btub2\_F and Btub2\_R, product size 440 bp. **D:** All lanes used the primer pair Btub3\_F and Btub2\_R, product size 127 bp. Each primer pair was trialled at six different annealing temperatures, displayed in the figure, and on three different templates. The templates were: **Left:** dH<sub>2</sub>O **Centre:** gDNA extracted from mycelium of the BC-16 isolate of *P. fragariae* grown on liquid media **Right:** gDNA extracted from the 'Hapil' cultivar of *Fragaria*  $\times$  *ananassa*. The temperatures trialled, from left to right in each set of three lanes were: 50°C, 52°C, 55°C, 57°C, 60°C and 62°C





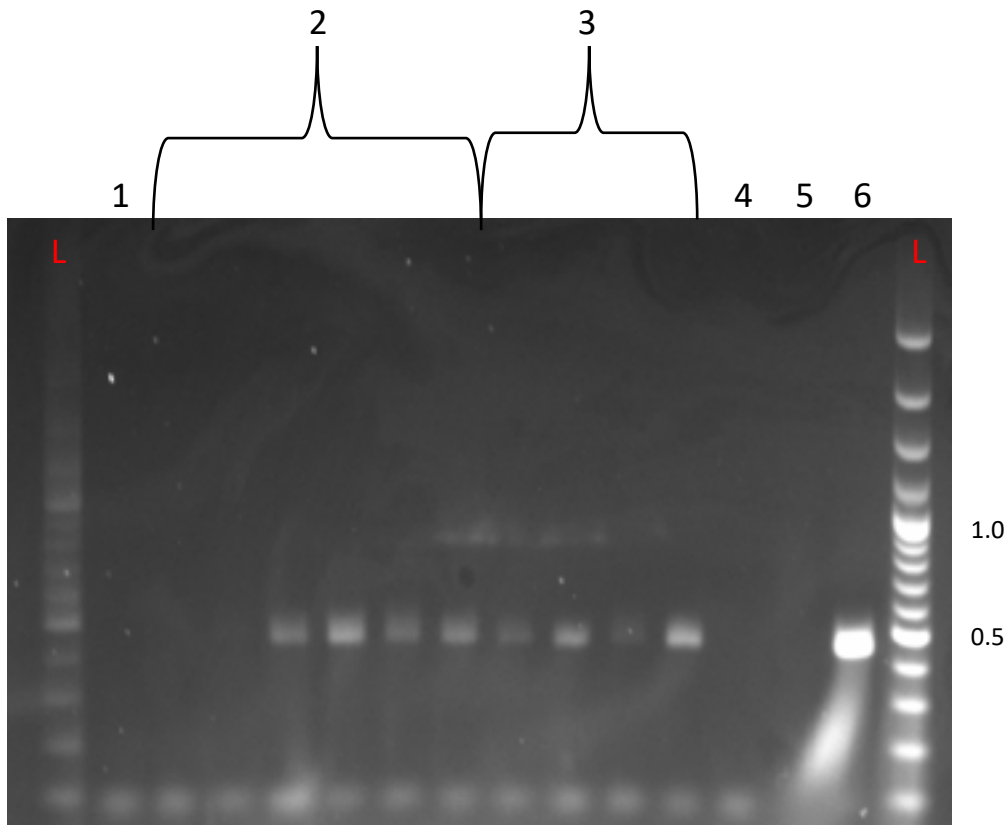
**Fig. 4.4: Alignment of Sanger sequencing results to the *Phytophthora fragariae*  $\beta$ -tubulin gene sequence.** The PCR product of Btub2\_F and Btub2\_R was sequenced and aligned to the predicted product, showing high identity and confirming that this was the correct sequence.

#### **4.4.4 Generation of RNA-Seq data for the BC-16 isolate of *Phytophthora fragariae* from an inoculation time course resulted in the sequencing of three time points post inoculation**

Using the above described 'plugs' method, samples were obtained of *F. × ananassa* cultivar 'Hapil' roots inoculated with the *P. fragariae* isolate BC-16, with either stream water or Petri's solution used as the final flooding solution. The following time points were harvested with stream water as the final flooding solution: 0, 24, 48, 96, 144, 192 and 240 hpi. Additionally, samples were taken at: 24, 48, 96 and 144 hpi with Petri's solution used as the final flooding solution. Total RNA was extracted from a single sample of each of these time points and an aliquot was subjected to RT-PCR to form total cDNA. This cDNA was then analysed using the previously described primers for *P. fragariae*  $\beta$ -tubulin in order to determine the earliest time point when the presence of *P. fragariae* could be confirmed (Fig. 4.5). These results showed that the pathogen was identified from 96 hpi when stream water was used as the final flooding solution, but from 24 hpi when Petri's solution was used as the final flooding solution. Due to these results, alongside the reduction in contamination risk when using Petri's solution over steam water, three samples from each of the following time points from the Petri's solution inoculation were sequenced: 24 hpi, 48 hpi and 96 hpi. Prior to sending samples for sequencing, several quality metrics were assessed to test for any risks of contamination with other biological molecules, assessing the concentration of the RNA and assessing the levels of degradation of the RNA (Table 4.3).

RNA was also extracted from mycelium of the BC-16 isolate of *P. fragariae* grown in liquid media and underwent the same quality control tests before being sent for sequencing (Table 4.4).





**Fig. 4.5: Agarose gel electrophoresis of an RT-PCR of a representative sample of each time point from an inoculation time course of the BC-16 isolate of *Phytophthora fragariae* on the 'Hapil' cultivar of the host plant *Fragaria × ananassa*. L: 100 bp Plus Ladder (ThermoFisher Scientific, Waltham, MA, U.S.A.) sizes listed in kb. **1:** Mock inoculated 'Hapil' plant. **2:** Time course of inoculated plants with stream water used as the flooding solution. Time points from left to right: 24 hpi, 48 hpi, 96 hpi, 144 hpi, 192 hpi and 240 hpi. **3:** Time course of inoculated plants with Petri's solution used as the flooding solution. Time points from left to right: 24 hpi, 48 hpi, 96 hpi and 144 hpi. **4:** dH<sub>2</sub>O control. **5:** 'Flamenco' cultivar gDNA control. **6:** BC-16 gDNA control.**

**Table 4.3: Quality control results for RNA samples extracted from *Fragaria* × *ananassa* cultivar 'Hapil' plants inoculated with the BC-16 isolate of *Phytophthora fragariae* before sequencing**

Sample ID	<i>Phytophthora fragariae</i>	<i>Fragaria</i> ×	Time point (hours post inoculation)	Concentration			
	Isolate	<i>ananassa</i> cultivar		260/280	260/230	(ng/μL)	RIN <sup>e</sup>
TA-07	BC-16	Hapil	24	2.03	2.20	158	6.6
TA-08	BC-16	Hapil	24	1.91	2.40	102	8.0
TA-09	BC-16	Hapil	24	1.93	2.42	133	6.6
TA-12	BC-16	Hapil	48	2.02	2.25	47.2	8.1
TA-13	BC-16	Hapil	48	2.08	2.16	63.4	7.0
TA-14	BC-16	Hapil	48	2.07	2.31	173	7.1
TA-18	BC-16	Hapil	96	2.05	2.30	94.6	7.9
TA-19	BC-16	Hapil	96	2.07	2.28	99.6	7.2
TA-20	BC-16	Hapil	96	2.04	2.32	67.2	8.2

**Table 4.4: Quality control statistics of RNA samples extracted from *in vitro* grown mycelium of the BC-16 isolate of *Phytophthora fragariae* before sequencing**

Sample ID	260/280	260/230	Concentration (ng/μL)	RIN <sup>e</sup>
TA-32	2.23	2.55	138	9.2
TA-34	2.19	2.29	124	9.8
TA-35	2.10	1.13	59.2	9.7

#### **4.4.5 Assessment of quality of sequencing data demonstrated a likely increase in the amount of *Phytophthora fragariae* RNA-Seq reads from the inoculation time course**

Prior to use of the sequencing data for downstream analysis, the raw reads were assessed to ensure they were of high enough quality for downstream analysis (Table 4.5). Several of these metrics were scored poorly, however, this was not unexpected for RNA-Seq data, which by its nature contains many more duplicate sequences than DNA-Seq data. It was also possible that since the pure mycelial RNA-Seq reads tend to perform better than the inoculated plant samples, that the presence of reads from two species has been marked as an indicator of poorer quality sequence by the program. It was also observed from these results that the percentage GC content of the reads increases as the time post inoculation increases, from 46% at 24 hpi, to 47.3% at 48 hpi and 49% at 96 hpi. This likely represents that a greater proportion of the reads in the later time points came from *P. fragariae* than in the earlier time points as *P. fragariae* has a higher percentage GC content than *F. × ananassa*, however, the values never reached the 60% GC content of pure mycelial RNA.

**Table 4.5: Fastqc results from analysis of trimmed RNA-Seq data**

Library Name	Timepoint or Sample Type	Basic Statistics	Per Base Sequence Quality	Per Sequence Quality Scores	Per Base Sequence Content	Per Base GC Content	Per Sequence GC Content	Per Base N Content	Sequence length Distribution	Sequence Duplication Levels	Overrepresented Sequences	K-mer Content
TA-07_F	24 hpi	Pass	Pass	Pass	Fail	Warning	Warning	Pass	Warning	Fail	Pass	Warning
TA-07_R	24 hpi	Pass	Pass	Pass	Fail	Warning	Warning	Pass	Warning	Fail	Pass	Warning
TA-08_F	24 hpi	Pass	Pass	Pass	Fail	Warning	Warning	Pass	Warning	Fail	Pass	Warning
TA-08_R	24 hpi	Pass	Pass	Pass	Fail	Warning	Warning	Pass	Warning	Fail	Pass	Warning
TA-09_F	24 hpi	Pass	Pass	Pass	Fail	Warning	Warning	Pass	Warning	Fail	Pass	Warning
TA-09_R	24 hpi	Pass	Pass	Pass	Fail	Warning	Warning	Pass	Warning	Fail	Pass	Warning
TA-12_F	48 hpi	Pass	Pass	Pass	Fail	Warning	Warning	Pass	Warning	Fail	Pass	Warning
TA-12_R	48 hpi	Pass	Pass	Pass	Fail	Warning	Warning	Pass	Warning	Fail	Pass	Warning
TA-13_F	48 hpi	Pass	Pass	Pass	Fail	Warning	Warning	Pass	Warning	Fail	Pass	Warning
TA-13_R	48 hpi	Pass	Pass	Pass	Fail	Warning	Warning	Pass	Warning	Fail	Pass	Warning
TA-14_F	48 hpi	Pass	Pass	Pass	Fail	Warning	Warning	Pass	Warning	Fail	Pass	Warning
TA-14_R	48 hpi	Pass	Pass	Pass	Fail	Warning	Warning	Pass	Warning	Fail	Pass	Warning
TA-18_F	96 hpi	Pass	Pass	Pass	Warning	Warning	Fail	Pass	Warning	Fail	Pass	Warning
TA-18_R	96 hpi	Pass	Pass	Pass	Warning	Warning	Fail	Pass	Warning	Fail	Pass	Warning
TA-19_F	96 hpi	Pass	Pass	Pass	Fail	Warning	Fail	Pass	Warning	Fail	Pass	Warning
TA-19_R	96 hpi	Pass	Pass	Pass	Warning	Warning	Warning	Pass	Warning	Fail	Pass	Warning
TA-20_F	96 hpi	Pass	Pass	Pass	Fail	Warning	Warning	Pass	Warning	Fail	Pass	Warning
TA-20_R	96 hpi	Pass	Pass	Pass	Fail	Warning	Warning	Pass	Warning	Fail	Pass	Warning
TA-32_F	<i>in vitro</i> mycelium	Pass	Pass	Pass	Warning	Pass	Warning	Pass	Warning	Fail	Pass	Warning
TA-32_R	<i>in vitro</i> mycelium	Pass	Pass	Pass	Warning	Pass	Pass	Pass	Warning	Fail	Pass	Warning
TA-34_F	<i>in vitro</i> mycelium	Pass	Pass	Pass	Warning	Pass	Warning	Pass	Warning	Fail	Pass	Warning
TA-34_R	<i>in vitro</i> mycelium	Pass	Pass	Pass	Warning	Pass	Warning	Pass	Warning	Fail	Pass	Warning
TA-35_F	<i>in vitro</i> mycelium	Pass	Pass	Pass	Warning	Pass	Warning	Pass	Warning	Fail	Pass	Warning
TA-35_R	<i>in vitro</i> mycelium	Pass	Pass	Pass	Warning	Pass	Warning	Pass	Warning	Fail	Pass	Warning

#### **4.4.6 Alignment of RNA-Seq data to assembled genomes of *Phytophthora fragariae* showed an increased proportion of *P. fragariae* reads as the inoculation time course progressed**

Initially, the RNA-Seq reads were aligned to the assembled genomes of *P. fragariae* and *P. rubi* (see Chapter 3). For reads from the inoculation time course, only those which did not map to the *F. vesca* genome were aligned to the *P. fragariae* assemblies and all mycelial reads were aligned to the assembled genomes (Table 4.6). These results showed a range of different mapping rates, with large variation within time points, for example, in the 24 hpi timepoint, one library averaged at 20.40% mapping rate, whereas the other samples averaged at 8.40% and 8.75% mapping rate. There was no clear cause as to this difference. The vast majority of unmapped reads were those rejected for being too short, this could possibly be due to degradation of the sample during library preparation at the external sequencing provider, though it was not possible to confirm this. However, despite the variation, an increase in the mapping rate was observed as the time course progresses, with an average of: 12.46% of reads mapping at 24 hpi, 29.46% of reads mapping at 48 hpi and 49.88% of reads mapping at 96 hpi. This suggested that despite attempts to remove reads from the host plant in all samples, that the proportion of reads from *P. fragariae* was increasing as the time course progressed. The reference BC-16 assembly also appeared to show marginally poorer mapping rates than the other *P. fragariae* isolates, this could potentially be due to sequencing errors that have not been corrected during the assembly process, or it may be that some of the alignments to the other isolates genomes were spurious.

For the *P. rubi* isolates, RNA-Seq data was obtained from the Grünwald Lab at Oregon State University as part of the *Phytophthora* Sequencing Consortium. These reads were mapped to the assemblies of *P. rubi* isolates (described in Chapter 3; Table 4.7). The mapping to SCRP249 appeared to be of poorer quality than that to SCRP324 and SCRP333, there were a larger number of unmapped reads when aligning to SCRP249, with the alignments rejected for being too short. It was possible that the SCRP249 isolate had sequence differences that resulted in these poorer statistics. It may also have been the case that since the

SCR249 assembly was more fragmented than the other two isolates that it resulted in shorter alignments, although this seemed unlikely as the N50 value was not dramatically different.

**Table 4.6: Mapping rates of RNA-Seq data from *Phytophthora fragariae* in vitro grown mycelium and time points post inoculation with the BC-16 isolate of *P. fragariae* on *Fragaria* × *ananassa* cultivar 'Hapil' plant roots against de novo assembled genomes of isolates of *P. fragariae***

Sample ID	Sample Source	Percentage Reads Mapped, unique sites and multiple sites										
		A4	BC-1	BC-16	BC-23	NOV-5	NOV-9	NOV-27	NOV-71	NOV-77	ONT-3	SCR245
TA-07	Inoculated roots, 24 hpi	20.58	20.26	19.51	20.58	20.52	20.33	20.62	20.25	20.54	20.57	20.62
TA-08	Inoculated roots, 24 hpi	8.38	8.23	8.00	8.38	8.37	8.26	8.39	8.24	8.37	8.39	8.40
TA-09	Inoculated roots, 24 hpi	8.74	8.60	8.41	8.73	8.72	8.64	8.75	8.61	8.73	8.75	8.75
TA-12	Inoculated roots, 48 hpi	19.14	18.64	18.11	19.12	19.07	18.73	19.18	18.65	19.14	19.14	19.15
TA-13	Inoculated roots, 48 hpi	44.22	43.37	42.12	44.17	44.04	43.56	44.31	43.35	44.18	44.24	44.25
TA-14	Inoculated roots, 48 hpi	25.98	25.32	24.45	25.96	25.86	25.42	26.05	25.31	25.98	25.98	26.01
TA-18	Inoculated roots, 96 hpi	66.43	64.98	62.89	66.36	66.12	65.26	66.61	64.95	66.44	66.41	66.38
TA-19	Inoculated roots, 96 hpi	62.92	61.16	59.34	62.83	62.55	61.51	63.09	61.15	62.90	62.90	62.86
TA-20	Inoculated roots, 96 hpi	22.10	21.47	20.83	22.06	21.98	21.59	22.17	21.46	22.10	22.11	22.09
TA-32	Liquid grown mycelium	95.80	93.49	91.24	95.80	95.51	94.30	96.12	93.85	95.63	95.84	95.77
TA-34	Liquid grown mycelium	96.08	93.78	91.53	96.13	95.77	94.63	96.41	94.17	95.99	96.09	96.06
TA-35	Liquid grown mycelium	96.02	93.50	91.52	96.00	95.66	94.43	96.31	93.93	95.85	96.08	95.95

Hours post inoculation (hpi)

**Table 4.7: Mapping rates of RNA-Seq data from *Phytophthora rubi* against de novo assembled genomes of isolate of *P. rubi***

Library Name	Percentage reads mapped, unique sites and multiple sites		
	SCR249	SCR324	SCR333
4671V8	47.91	97.59	97.78
Pr4671PB	47.92	97.71	97.82

#### **4.4.7 *Ab initio* gene prediction**

##### **4.4.7.1 Initial gene prediction showed similar numbers of genes in all isolates**

Following the alignment of RNA-Seq reads to the genome assemblies, genes were predicted by three methods: BRAKER1 (Hoff *et al.*, 2015), CodingQuarry (Testa *et al.*, 2015) and a non-guided approach using a modified version of the ORF finder script from Win *et al.* (2006). Non-guided predictions were not yet merged with the Braker and CodingQuarry predictions as stronger evidence was required to call these as predicted gene models. For building the final guided gene model set, individual CodingQuarry predictions were merged into the gene model set where they did not intersect with BRAKER1 predicted genes, this produced between 30,538 and 37,353 gene models (Table 4.8). These results showed a similar number of guided gene predictions in all the *P. fragariae* isolates, but with a much larger number in BC-16. Interestingly, this appeared to be due to a greatly increased number of CodingQuarry predictions in BC-16 compared to the other *P. fragariae* isolates as the number of BRAKER1 predictions was actually 2,149 lower than the average for the other *P. fragariae* and *P. rubi* isolates. The number of genes in the *P. rubi* isolates appeared to vary for both guided gene prediction methods and in the final number of guided genes predicted. As expected, the number of unguided gene predictions correlated with assembly size.



**Table 4.8: The number of genes present from both guided and unguided gene predictions in all sequenced isolates**

Species	Isolate	Number of			
		BRAKER1 predictions	CodingQuarry predictions	Guided gene predictions	Unguided ORF predictions
<i>Phytophthora fragariae</i>	A4	21,182	34,110	30,997	654,457
<i>Phytophthora fragariae</i>	BC-1	20,941	31,405	31,189	657,394
<i>Phytophthora fragariae</i>	BC-16	20,222	43,899	37,353	776,126
<i>Phytophthora fragariae</i>	BC-23	21,403	30,852	30,791	648,130
<i>Phytophthora fragariae</i>	NOV-5	21,004	34,007	31,072	654,072
<i>Phytophthora fragariae</i>	NOV-9	20,975	33,845	30,828	660,249
<i>Phytophthora fragariae</i>	NOV-27	21,202	31,478	31,432	653,798
<i>Phytophthora fragariae</i>	NOV-71	20,879	33,631	30,538	649,542
<i>Phytophthora fragariae</i>	NOV-77	21,163	33,884	30,932	653,276
<i>Phytophthora fragariae</i>	ONT-3	21,417	33,820	30,987	655,156
<i>Phytophthora fragariae</i>	SCR245	20,838	31,196	30,996	645,061
<i>Phytophthora rubi</i>	SCR249	29,583	28,270	31,808	645,852
<i>Phytophthora rubi</i>	SCR324	20,541	30,513	30,713	647,415
<i>Phytophthora rubi</i>	SCR333	29,696	31,450	32,946	646,159

#### **4.4.7.2 Prediction of putative effector genes showed similar numbers of possible effector genes in all isolates**

Following gene prediction, several key classes of effector gene were identified in both the guided and unguided gene prediction sets. For the guided gene models, all genes predicted as positive for secretion by any of the methods used were scored as coding for putatively secreted proteins. For the unguided gene models, it was common that several predictions overlapped and scored positive for a secretion signal. In these cases, the prediction with the highest SignalP score was accepted (Table 4.9 and 4.10). These results showed a similar number of secreted proteins from guided gene models in all *P. fragariae* and *P. rubi* isolates. However, there were a larger number of positive predictions in BC-16, though this was not unexpected given the larger number of total gene models in this isolate. This was similarly the case with the unguided gene models, although there were a smaller number of secreted proteins predicted in ONT-3 and SCRP245 than the other isolates. This did not reflect the number of unguided models predicted. As expected, there were a greater number of secreted proteins from unguided than guided gene models, though these may have been false positives.

Following the prediction of secretion signals, the subset of putatively secreted proteins was analysed for the presence of various classes of effector genes. Firstly, genes encoding RxLR effectors were searched for. Two methods were used for predicting RxLR effectors, a Hidden Markov Model (HMM; Whisson *et al.*, 2007) and a regular expression search (Regex) that searches for both high confidence effectors, containing an EER motif as well as an RxLR motif, and lower confidence effectors, containing just the RxLR motif (Armitage *et al.*, 2018b). These analyses were conducted in secreted protein sets from both guided and unguided gene models, with merging being carried out in the unguided set after RxLR prediction based on the SignalP score where models overlapped (Table 4.9 and 4.10). These results showed a similar number of putative RxLR effectors for each prediction method across both guided and unguided gene models. There was an increase in the number of predictions in BC-16, which likely reflected a larger number of total gene predictions in this

assembly. Interestingly, there was a decrease in the number of RxLR effectors predicted from unguided gene models in ONT-3 and SCRP245, which likely reflected the lower number of secreted proteins being carried through into the total RxLR number.

Additionally, CRNs were identified using HMM models specific to two key domains, LFLAK and DWL (Armitage *et al.*, 2018b). These analyses were conducted in both guided and unguided secreted protein sets, with merging carried out in the unguided after CRN prediction based on the LFLAK HMM score where gene models overlapped (Table 4.9 and 4.10). These results showed a similar number of predicted CRNs from guided gene models in all *P. fragariae* isolates. Within the *P. rubi* isolates, SCRP324 showed a reduced number of CRNs from guided gene models, though the value for this isolate was more in line with those for the *P. fragariae* isolates than the other *P. rubi* isolates. CRN predictions from unguided gene models also showed a similar number of CRNs in all *P. fragariae* isolates and a similar number in all *P. rubi* isolates, though there were slightly more predictions in the *P. rubi* isolates than the *P. fragariae* isolates. Interestingly, although there were more CRNs predicted from guided gene models in BC-16 than the other *P. fragariae* isolates, this was not the case for the predictions from unguided gene models, despite there being a larger number predicted by each individual HMM.

Finally, apoplastic effectors were predicted by ApoplastP (Sperschneider *et al.*, 2018). This analysis was conducted on secreted protein sets from both guided and unguided gene models, with merging being carried out on the ApoplastP predictions from unguided gene models based on the SignalP HMM score where gene models overlapped (Table 4.9 and 4.10). These results showed a similar number of predicted apoplastic effectors in both *P. fragariae* and *P. rubi* isolates, with a larger number in BC-16, likely due to a larger number of predicted secreted proteins in this isolate. This was also observed for predictions from unguided gene models, with again lower numbers of predicted apoplastic effectors in ONT-3 and SCRP245, which likely reflected the lower number of predicted secreted proteins in these isolates.

**Table 4.9: Predicted effector proteins from RNA-Seq guided gene models**

Species	Isolate	Secreted Proteins (Nielsen et al., 1997; Bendtsen et al., 2004; Petersen et al., 2011; Käll et al., 2007)		RxLR-EER	RxLR	Total RxLR Effectors	CRN	CRN	Total CRNs	Apoplastic Effectors (Sperschneider et al., 2018)
		RxLR HMM (Whisson et al., 2007)	Regex (Armitage et al., 2018b)	Regex (Armitage et al., 2018b)	LFLAK HMM (Armitage et al., 2018b)		DWL HMM (Armitage et al., 2018b)			
<i>Phytophthora fragariae</i>	A4	3,637	194	178	371	410	85	95	55	991
<i>Phytophthora fragariae</i>	BC-1	3,601	194	184	367	405	83	87	53	1,002
<i>Phytophthora fragariae</i>	BC-16	4,217	218	208	445	486	114	121	82	1,274
<i>Phytophthora fragariae</i>	BC-23	3,611	188	176	364	402	86	96	55	980
<i>Phytophthora fragariae</i>	NOV-5	3,626	196	186	370	408	84	91	53	986
<i>Phytophthora fragariae</i>	NOV-9	3,637	186	174	356	397	82	96	50	1,007
<i>Phytophthora fragariae</i>	NOV-27	3,690	183	172	369	405	87	93	57	1,011
<i>Phytophthora fragariae</i>	NOV-71	3,633	191	182	374	412	90	101	59	1,007
<i>Phytophthora fragariae</i>	NOV-77	3,620	166	150	341	378	86	83	62	1,010
<i>Phytophthora fragariae</i>	ONT-3	3,658	191	178	367	403	90	90	61	982
<i>Phytophthora fragariae</i>	SCR245	3,581	196	185	370	407	90	87	60	984
<i>Phytophthora rubi</i>	SCR249	3,697	197	188	363	407	156	139	128	1,017
<i>Phytophthora rubi</i>	SCR324	3,683	195	176	350	395	93	102	71	1,059
<i>Phytophthora rubi</i>	SCR333	3,832	207	188	359	410	154	142	128	1,090

Crinkler effectors (CRNs). Combining of RxLR effectors and CRNs was performed by merging a list of all genes predicted as being RxLR effectors or CRNs from all methods.

**Table 4.10: The number of different classes of effector proteins predicted via various methods on unguided gene predictions in *Phytophthora fragariae* and *Phytophthora rubi* isolates**

Species	Isolate	Secreted Proteins (Nielsen et al., 1997; Bendtsen et al., 2004; Petersen et al., 2011; Käll et al., 2007)		RxLR-EER Regex (Armitage <i>et al.</i> , 2018b)	RxLR Regex (Armitage <i>et al.</i> , 2018b)	Total RxLR Effectors	CRN LFLAK HMM (Armitage <i>et al.</i> , 2018b)	CRN DWL HMM (Armitage <i>et al.</i> , 2018b)	Total CRNs	Apoplastic Effectors (Sperschneider <i>et al.</i> , 2018)
<i>Phytophthora fragariae</i>	A4	40,784	199	269	2,176	2,187	243	380	106	10,322
<i>Phytophthora fragariae</i>	BC-1	41,080	199	270	2,189	2,200	243	380	106	10,414
<i>Phytophthora fragariae</i>	BC-16	48,478	218	296	2,448	2,460	276	425	118	12,431
<i>Phytophthora fragariae</i>	BC-23	40,484	204	273	2,172	2,188	244	367	106	10,157
<i>Phytophthora fragariae</i>	NOV-5	40,803	199	269	2,161	2,172	245	378	106	10,294
<i>Phytophthora fragariae</i>	NOV-9	41,177	200	268	2,203	2,214	243	377	105	10,489
<i>Phytophthora fragariae</i>	NOV-27	40,789	199	270	2,163	2,174	246	371	106	10,325
<i>Phytophthora fragariae</i>	NOV-71	40,711	199	269	2,159	2,170	247	377	105	10,320
<i>Phytophthora fragariae</i>	NOV-77	40,933	195	258	2,158	2,167	230	371	104	10,360
<i>Phytophthora fragariae</i>	ONT-3	38,272	205	269	2,031	2,041	241	361	103	9,670
<i>Phytophthora fragariae</i>	SCR245	37,714	202	270	2,024	2,034	239	363	105	9,602
<i>Phytophthora rubi</i>	SCR249	40,893	211	277	2,158	2,171	301	452	131	10,341
<i>Phytophthora rubi</i>	SCR324	40,842	212	270	2,156	2,170	294	440	126	10,280
<i>Phytophthora rubi</i>	SCR333	40,825	206	272	2,141	2,153	293	438	129	10,354

Crinkler effectors (CRNs). Combining of RxLR effectors and CRNs was performed by merging a list of all genes predicted as being RxLR effectors or CRNs from all methods.

#### 4.4.7.3 Creation of a final gene model set provided a set for further investigation to identify putative avirulence genes

Following the prediction of effectors, a final gene model set was created for all assemblies of *P. fragariae* and *P. rubi*. The prediction of an effector coding gene in the set of unguided gene models was considered evidence for merging the model with the guided gene models when the model resided in an intergenic region. Following this merging, several steps of correction of this merged file were performed including: removing duplicate predictions from ApoplastP and RxLR or CRN predictions, overlapping gene models were removed where unguided gene models intersected with RNA-Seq guided gene models and the genes were renamed to NCBI standards (Table 4.11).

These results showed that the number of predicted genes and proteins were similar in all examined *P. fragariae* and *P. rubi* isolates, with slightly more predicted genes and proteins in the BC-16 assembly, likely due to a larger assembly size. The number of final predicted secreted proteins was similar in the majority of *P. fragariae* isolates, with again a larger number predicted in the BC-16 assembly and also a slightly smaller number predicted in the SCRP245 assembly. However, for the *P. rubi* assemblies, the number of predicted secreted proteins was significantly ( $p < 0.05$ ) lower than in the assemblies of *P. fragariae* isolates ( $p = 0.03039$ ).

The number of RxLR effectors was similar for all isolates of *P. fragariae*, with marginally fewer predictions in the SCRP245 assembly and more predictions in the BC-16 assembly. Again, there were significantly ( $p < 0.05$ ) fewer predictions in the assemblies of the *P. rubi* isolates than in the assemblies of the *P. fragariae* isolates, likely due to a smaller number of input secreted proteins ( $p = 0.01256$ ).

The number of predicted CRNs was again similar in all assemblies of the *P. fragariae* isolates, with slightly more predictions in the BC-16 assemblies. Interestingly, there appeared to be more CRNs predicted in the assemblies of *P. rubi* isolates than in the assemblies of *P. fragariae* isolates, though the number of predictions in the SCRP324 assembly was closer to the number in the BC-16

assembly than the assemblies of other *P. rubi* isolates. This likely resulted in the non-significance ( $p < 0.05$ ) of any difference between the numbers of CRNs in these species ( $p = 0.07965$ ).

The number of putative apoplastic effectors appeared to be similar in the assemblies of all the *P. fragariae* isolates with again more predictions in the assembly of the BC-16 isolate of *P. fragariae* and marginally fewer predictions in the assembly of the SCRP245 isolate of *P. fragariae*. In the *P. rubi* assemblies, the number of putative apoplastic effectors was significantly ( $p < 0.05$ ) lower in the assemblies of isolates of *P. fragariae*, with again more predictions in the assembly of the SCRP324 isolate than the assemblies of other *P. rubi* isolates ( $p = 0.03890$ ).

Following the creation of a final gene model set for each isolate, testing for completeness was performed against the eukaryotic BUSCO database (Simão *et al.*, 2015; Table 4.12), alongside the changes with respect to the results of a BUSCO search on the assemblies prior to annotation being performed, as described in Chapter 3, sections 3.3.6.1 and 3.3.6.2. This process searches for 303 genes thought to be universal single copy orthologues in the specified database, in this case, all eukaryotes. These results showed a general decrease in the number of complete single copy BUSCO genes compared to the results of the BUSCO analysis on the assemblies alone for the majority of isolates, however, a decrease was also observed in the number of missing BUSCO genes in the majority of isolates. An increase was observed in the number of fragmented and duplicated BUSCO genes in the majority of assemblies. The magnitude of these changes was fairly minor and suggested that the annotations were likely to represent a high proportion of the true gene complement of these isolates. These values were comparable to those demonstrated for assemblies of several previously sequenced *Phytophthora* spp.: 272 were identified in *Phytophthora cactorum*, 271 were identified in *P. parasitica*, 257 were identified in *P. infestans*, 261 were identified in *P. capsici* and 262 were identified in *P. sojae*. This was compared to the identification of 271 - 274 in *P. fragariae* isolates, 263 in the reference BC-16 annotations and 265 - 270 in *P. rubi* isolates (Armitage *et al.*, 2018b).

**Table 4.11: Final numbers of genes, proteins and putative effector class genes in all sequenced isolates of *Phytophthora fragariae* and *Phytophthora rubi***

Species	Isolate	Gene Number	Protein Number	Secreted Proteins		CRN	
				(Nielsen et al., 1997; Bendtsen et al., 2004; Petersen et al., 2011; Käll et al., 2007)	RxLR Effectors (Whisson <i>et al.</i> , 2007; Armitage <i>et al.</i> , 2018b)	Effectors (Armitage <i>et al.</i> , 2018b)	Apoplasmic Effectors (Sperschneider <i>et al.</i> , 2018)
<i>Phytophthora fragariae</i>	A4	33,623	34,434	6,887	950	68	3,993
<i>Phytophthora fragariae</i>	BC-1	33,691	34,500	6,724	935	59	3,890
<i>Phytophthora fragariae</i>	BC-16	41,103	41,400	8,024	1,058	88	4,864
<i>Phytophthora fragariae</i>	BC-23	33,143	33,968	6,602	945	62	3,730
<i>Phytophthora fragariae</i>	NOV-5	33,699	34,515	6,884	932	67	4,003
<i>Phytophthora fragariae</i>	NOV-9	33,527	34,361	6,968	961	61	4,090
<i>Phytophthora fragariae</i>	NOV-27	33,797	34,624	6,696	908	63	3,793
<i>Phytophthora fragariae</i>	NOV-71	33,143	33,967	6,880	939	71	4,001
<i>Phytophthora fragariae</i>	NOV-77	33,580	34,408	6,901	945	75	4,022
<i>Phytophthora fragariae</i>	ONT-3	33,415	34,214	6,734	913	74	3,816
<i>Phytophthora fragariae</i>	SCR245	33,223	34,010	6,460	895	68	3,622
<i>Phytophthora rubi</i>	SCR249	34,139	34,484	6,262	839	133	3,343
<i>Phytophthora rubi</i>	SCR324	33,263	33,732	6,566	874	87	3,673
<i>Phytophthora rubi</i>	SCR333	35,223	35,539	6,309	827	135	3,340

Crinkler effectors (CRNs).



**Table 4.12: A comparison of BUSCO predictions from gene annotations against BUSCO predictions from unannotated assemblies for all sequenced *Phytophthora fragariae* and *Phytophthora rubi* isolates**

Species	Isolate	Single Copy BUSCO genes		Duplicated BUSCO genes		Fragmented BUSCO genes		Missing BUSCO genes	
		Predicted genes	Difference	Predicted genes	Difference	Predicted genes	Difference	Predicted genes	Difference
<i>Phytophthora fragariae</i>	A4	272 (89.77%)	-2 (-0.66%)	7 (2.31%)	+1 (+0.33%)	9 (2.97%)	+3 (+0.99%)	15 (4.95%)	-2 (-0.66%)
<i>Phytophthora fragariae</i>	BC-1	272 (89.77%)	-2 (-0.66%)	7 (2.31%)	+1 (+0.33%)	9 (2.97%)	+3 (+0.99%)	15 (4.95%)	-2 (-0.66%)
<i>Phytophthora fragariae</i>	BC-16	263 (86.80%)	-3 (-0.99%)	9 (2.97%)	0 (0.00%)	10 (3.30%)	+5 (+1.65%)	21 (6.93%)	-2 (-0.66%)
<i>Phytophthora fragariae</i>	BC-23	274 (90.43%)	-1 (-0.33%)	6 (1.98%)	+1 (+0.33%)	8 (2.64%)	+1 (+0.33%)	15 (4.95%)	-1 (-0.33%)
<i>Phytophthora fragariae</i>	NOV-5	273 (90.10%)	0 (0.00%)	7 (2.31%)	0 (0.00%)	8 (2.64%)	+2 (+0.66%)	15 (4.95%)	-2 (-0.66%)
<i>Phytophthora fragariae</i>	NOV-9	273 (90.10%)	-3 (-0.99%)	7 (2.31%)	0 (0.00%)	8 (2.64%)	+3 (+0.99%)	15 (4.95%)	0 (0.00%)
<i>Phytophthora fragariae</i>	NOV-27	273 (90.10%)	0 (0.00%)	7 (2.31%)	+1 (+0.33%)	8 (2.64%)	+1 (+0.33%)	15 (4.95%)	-2 (-0.66%)
<i>Phytophthora fragariae</i>	NOV-71	272 (89.77%)	-2 (-0.66%)	7 (2.31%)	+1 (+0.33%)	9 (2.97%)	+3 (+0.99%)	15 (4.95%)	-2 (-0.66%)
<i>Phytophthora fragariae</i>	NOV-77	271 (89.44%)	-1 (-0.33%)	7 (2.31%)	-1 (-0.33%)	9 (2.97%)	+3 (+0.99%)	16 (5.28%)	-1 (-0.33%)
<i>Phytophthora fragariae</i>	ONT-3	274 (90.43%)	-3 (-0.99%)	7 (2.31%)	0 (0.00%)	7 (2.31%)	+3 (+0.99%)	15 (4.95%)	0 (0.00%)
<i>Phytophthora fragariae</i>	SCR245	272 (89.77%)	-1 (-0.33%)	8 (2.64%)	+1 (+0.33%)	7 (2.31%)	0 (0.00%)	16 (5.28%)	0 (0.00%)
<i>Phytophthora rubi</i>	SCR249	269 (88.78%)	-4 (-1.32%)	9 (2.97%)	+1 (+0.33%)	8 (2.64%)	+5 (+1.65%)	17 (5.61%)	-2 (-0.66%)
<i>Phytophthora rubi</i>	SCR324	265 (87.46%)	-9 (-2.97%)	11 (3.63%)	+3 (+0.99%)	10 (3.30%)	+7 (+2.31%)	17 (5.61%)	-1 (-0.33%)
<i>Phytophthora rubi</i>	SCR333	270 (89.11%)	-5 (-1.65%)	10 (3.30%)	+3 (+0.99%)	7 (2.31%)	+4 (+1.32%)	16 (5.28%)	-2 (-0.66%)

All changes displayed were the result of subtracting the value for the annotations from the value for the assemblies. For example, a value of -2 represents two fewer positive hits in the annotations compared to the assemblies. All values were rounded to two decimal places where appropriate.

#### **4.4.7.4 Additional functional annotation of gene annotations provided additional evidence for possible involvement of genes in the pathogenicity process**

Following the creation of a final gene model set, additional functional annotations were performed to predict transmembrane helices and glycosylphosphatidylinositol (GPI) coupled proteins that were likely to be membrane bound. These features were predicted since a protein with these features would be unlikely to be involved in the pathogenicity process (Table 4.13). There did not appear to be any differences in the numbers of proteins with predicted transmembrane helices for all isolates of *P. fragariae* and *P. rubi*. The number of proteins with predicted GPI anchors was consistent within species, however, there appeared to be approximately 100 fewer in *P. rubi* than in *P. fragariae*.

**Table 4.13: Number of genes predicted to possess transmembrane helices or GPI anchors**

Species	Isolate	Number of proteins with TM Helicies (Krogh <i>et al.</i> , 2001)	Number of proteins with GPI anchors (Frankhauser and Mäser, 2005)
<i>Phytophthora fragariae</i>	A4	4,617 (13.41%)	1,105 (3.21%)
<i>Phytophthora fragariae</i>	BC-1	4,593 (13.31%)	1,039 (3.01%)
<i>Phytophthora fragariae</i>	BC-16	4,816 (11.63%)	1,181 (2.85%)
<i>Phytophthora fragariae</i>	BC-23	4,557 (13.75%)	1,030 (3.11%)
<i>Phytophthora fragariae</i>	NOV-5	4,641 (13.45%)	1,076 (3.12%)
<i>Phytophthora fragariae</i>	NOV-9	4,632 (13.48%)	1,086 (3.16%)
<i>Phytophthora fragariae</i>	NOV-27	4,587 (13.25%)	1,036 (2.99%)
<i>Phytophthora fragariae</i>	NOV-71	4,579 (13.48%)	1,074 (3.16%)
<i>Phytophthora fragariae</i>	NOV-77	4,580 (13.31%)	1,094 (3.18%)
<i>Phytophthora fragariae</i>	ONT-3	4,609 (13.47%)	1,041 (3.04%)
<i>Phytophthora fragariae</i>	SCR245	4,533 (13.33%)	1,004 (2.95%)
<i>Phytophthora rubi</i>	SCR249	4,546 (13.18%)	910 (2.64%)
<i>Phytophthora rubi</i>	SCR324	4,561 (13.52%)	972 (2.88%)
<i>Phytophthora rubi</i>	SCR333	4,577 (12.88%)	950 (2.67%)

#### **4.4.8 Analysis of expression in the BC-16 isolate of *Phytophthora fragariae* showed wide ranging transcriptional reprofiling occurring upon infection**

Following the creation of a final gene model set, it was of interest to examine whether the RNA-Seq dataset had indeed captured the expression dynamics during the early stages of the infection process. Following the determination of expression levels in a variety of samples, a principal component analysis (PCA) plot was produced for all the RNA-Seq samples with an rlog transformation (Fig. 4.6). The largest principal component, PC1, appeared to represent the differences between expression data from samples from the inoculation time course experiment and samples from *in vitro* grown mycelium and explained 71% of the variation. The smaller principal component, PC2, appeared to represent time since inoculation and explained 19% of the variation. These results suggested that although the largest difference was explained by the differences in growth conditions, there was a sizeable amount of variation over time, with no overlap between the time points. A total of 13,240 transcripts showed a FPKM value of five or higher in at least one sequenced sample, representing 31.98% of the total number of predicted transcripts in the BC-16 isolate of *P. fragariae*.

The number of differentially expressed transcripts at each time point post inoculation was plotted in a Venn diagram (Fig. 4.7). These results showed that a large number of differentially expressed transcripts were shared across all-time points post inoculation. It was also observed that when examining the intersections on the Venn diagram between two time points post inoculation, there was a dramatic decline in the 24 hpi time point and the 96 hpi time point intersection than the other two intersections (297 compared to 1,016 and 1,604). This suggested that the samples taken from the inoculation time course experiment had indeed captured the infection progressing from an early stage of the process to a later stage, with the transcript expression patterns changing to match this. Interestingly, although there were slightly more down regulated than up regulated transcripts in this dataset, the difference was only 598 transcripts (11.72% of the down regulated total). A total of 9,329 transcripts showed differential expression with a log<sub>2</sub> fold change (LFC) either greater than one or less than minus one, representing 22.53% of the predicted transcripts.

Additionally, 4,506 transcripts showed an LFC greater than one, representing up regulation. This consisted of 10.88% of predicted transcripts. Finally, 5,104 transcripts showed an LFC less than minus one, representing down regulation. This consisted of 12.33% of predicted transcripts. These results suggested that there was wide scale transcriptional change occurring during the infection of a host plant.

Analyses were also conducted on the levels of expression of predicted effector genes in order to assess possible errors or biases in the prediction methods. Firstly, it was of interest to assess the proportions of effectors that showed evidence of expression in these RNA-Seq datasets. For this analysis, an FPKM threshold of five was used to control for potential misalignments of the RNA-Seq reads to the assemblies. These results showed a smaller number of transcripts being expressed at early stages of the inoculation time course experiment when compared to later stages and *in vitro* grown mycelium, possibly due to a smaller number of pathogen reads being sequenced in earlier samples. However, the numbers expressed in the 96 hpi timepoint were similar to those from *in vitro* grown mycelium. This suggested that the lack of expressed transcripts in early time points was not solely due to the inoculation procedure. In total, just under a third of the predicted transcripts showed evidence of expression in these datasets. This could potentially have been due to errors in the gene prediction process, or it also could have been due to not all possible life stages of *P. fragariae* being sequenced. Most likely, it was due to a combination of these factors (Table 4.14).

Analysis of effectors showed that approximately 26% of predicted RxLR effectors showed evidence of expression and approximately 31% of CRNs showed evidence of expression. A far lower percentage, below 20%, of predicted apoplastic effectors showed expression. A similar proportion of secreted proteins and total transcripts showed evidence of expression, just under a third showed evidence of expression in at least one sample. These results suggested that predictions of RxLRs and apoplastic effectors were less precise than those for secreted proteins and CRNs, however they were all kept as putative gene models to ensure all potential avirulence genes were captured for analysis (Table 4.14).

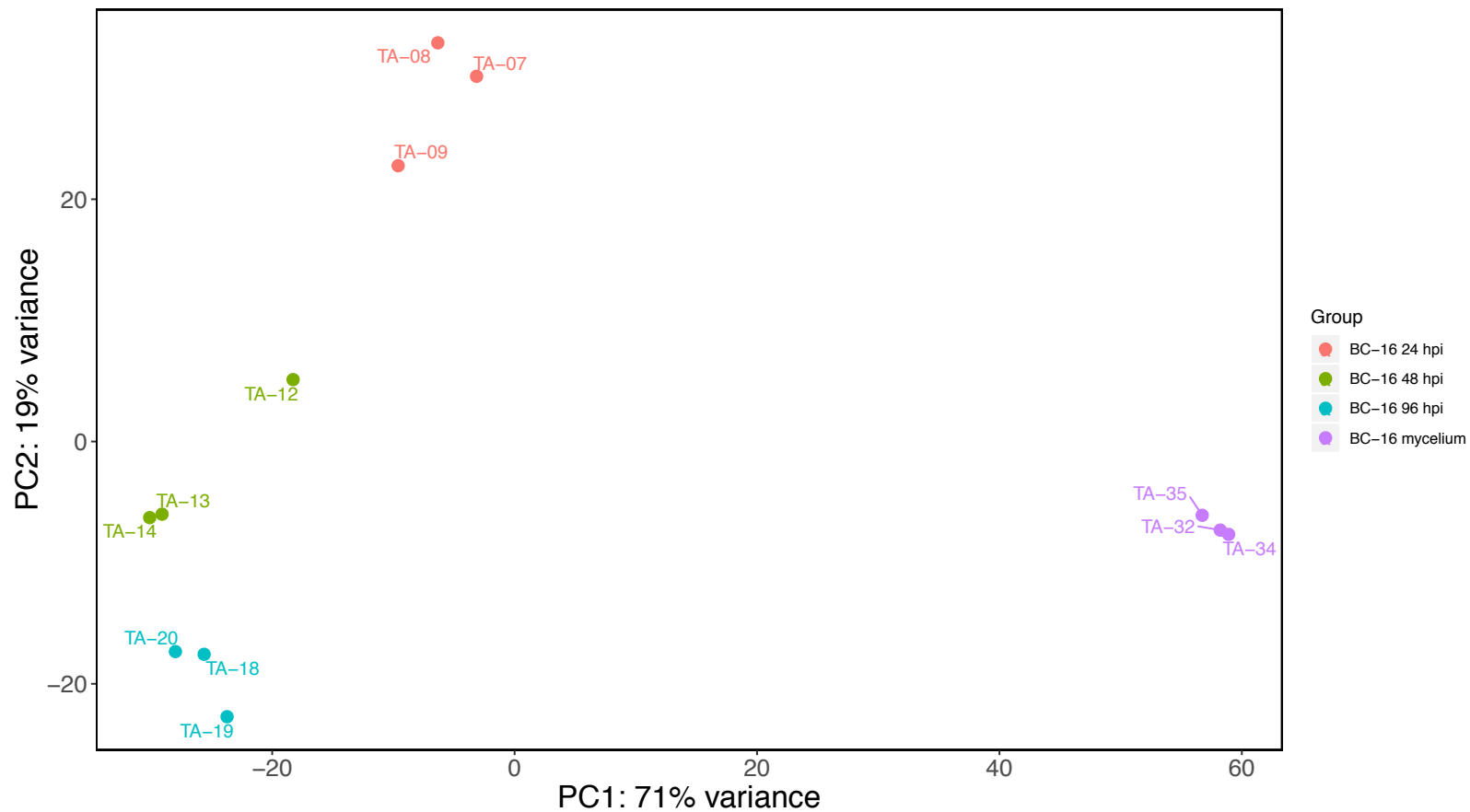
Differential expression analysis was also conducted for each predicted effector gene class. In total, approximately 23% of transcripts showed evidence of differential expression between samples from time points in the inoculation time course experiment and *in vitro* grown mycelium. This was approximately ten percentage points less than the number of transcripts that showed evidence of expression overall. This suggested that the differences in expression patterns were due to a specific transcriptional reprofiling, rather than solely due to expression patterns in *in vitro* grown mycelium.

For RxLR effectors, just under a quarter of predicted transcripts were differentially expressed between samples from time points in the inoculation time course experiment and *in vitro* grown mycelium, this was very similar to the number showing expression above the threshold previously defined. Only 21 predicted RxLR effectors were expressed, but did not show differential expression. A larger number of these predicted effectors showed up regulation during the infection process, however, interestingly a sizable number still showed down regulation, perhaps as an attempt to avoid defense responses or due to the clearing of transcriptional repression during growth on artificial media.

A similar proportion of predicted crinkler effectors were differentially expressed, however, in this case a greater proportion of predicted CRNs showed down regulation than up regulation during the infection process. This was expected as CRNs have been mostly shown to be late stage effectors (Stam *et al.*, 2013). Similar to the above expression results, a lower percentage of predicted apoplastic effectors showed differential expression, however, the values for differential expression and expression above a threshold were largely similar.

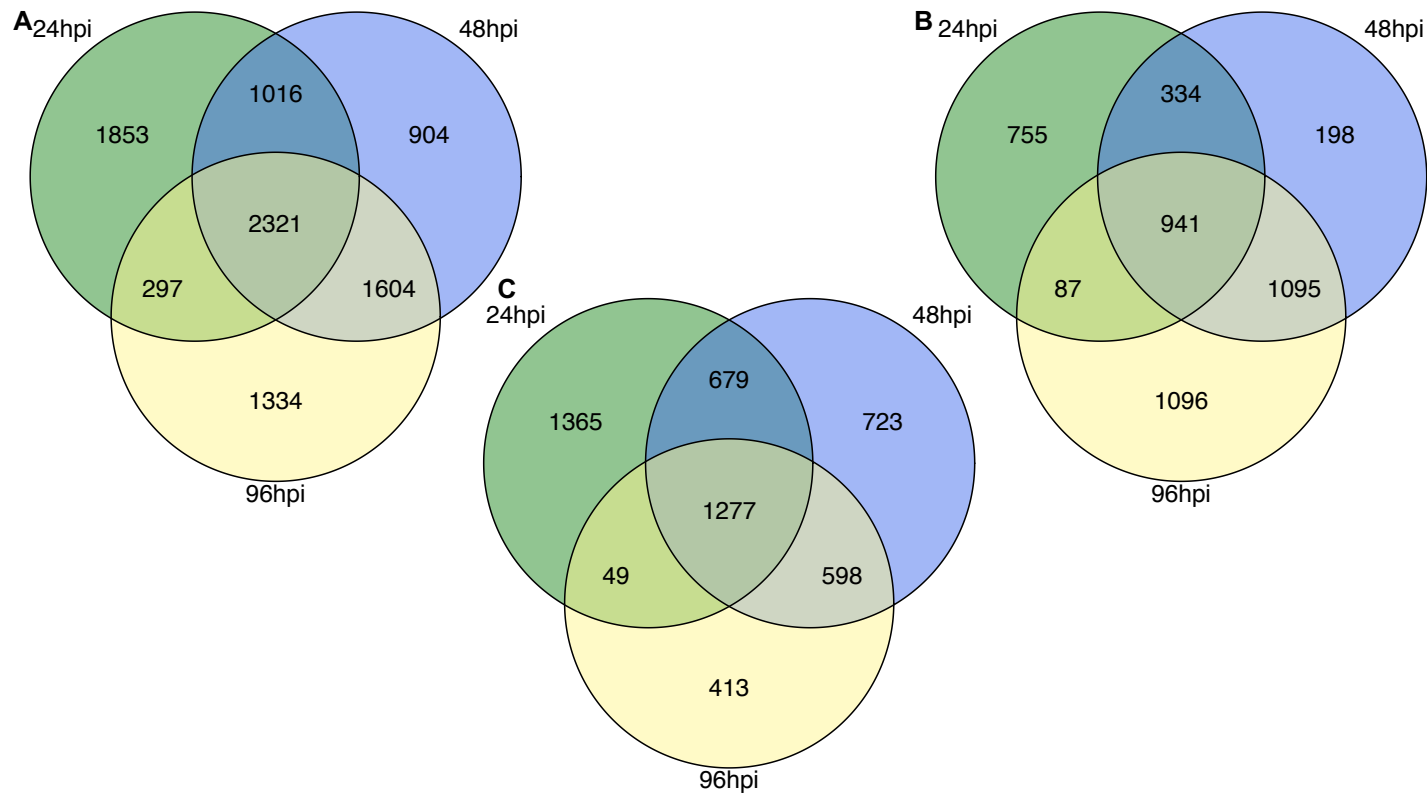
Slightly more predicted apoplastic effectors were predicted as being differentially expressed than those just expressed, likely due to genes not meeting the threshold for expression still being called as differentially expressed. Similar to predicted RxLR effectors, these showed a greater proportion being up regulated than down regulated during the infection process.

For secreted proteins, a larger proportion of the predictions were shown as differentially expressed than the proportion of predicted transcripts overall, with over 25% of secreted proteins showing this. Interestingly, unlike transcripts in general, a greater proportion of secreted proteins were up regulated than down regulated. This was similar to patterns observed for genes predicted as RxLR effectors and apoplastic effectors and suggested that a major transcriptional reprofiling was underway during infection (Fig. 4.8 - 4.11 and Table 4.15).

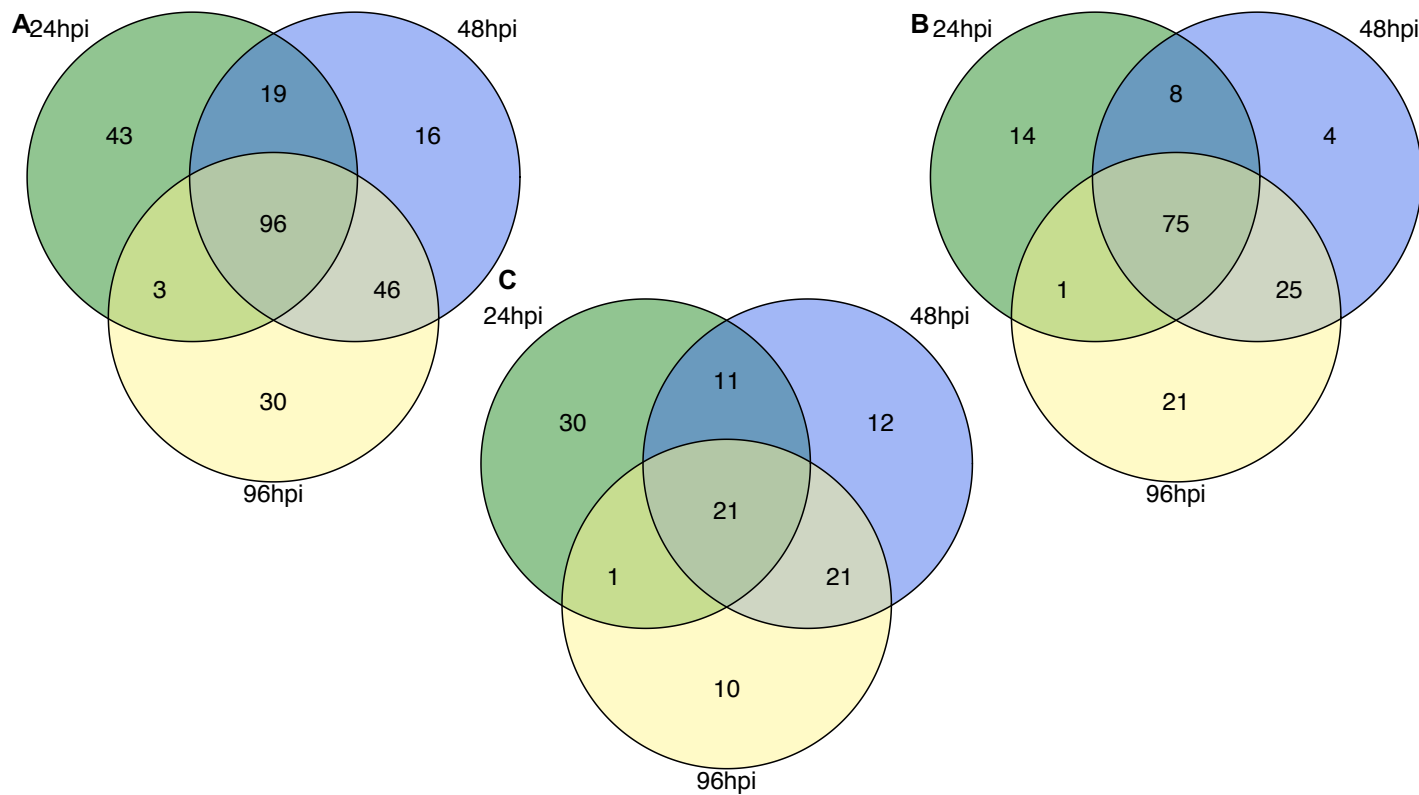


**Fig. 4.6: Principal component analysis following expression assessment in the BC-16 isolate of *Phytophthora fragariae*.** RNA-Seq reads were aligned to the assembly of the BC-16 isolate of *P. fragariae* using STAR version 2.5.3a (Dobin *et al.*, 2013). Predicted transcripts were then quantified with featureCounts version 1.5.2 (Liao *et al.*, 2014) and differential expression was identified with DESeq2 version 1.10.1 (Love *et al.*, 2014). Following this, an rlog transformation of the expression data was plotted as a PCA with R (R core team, 2016).

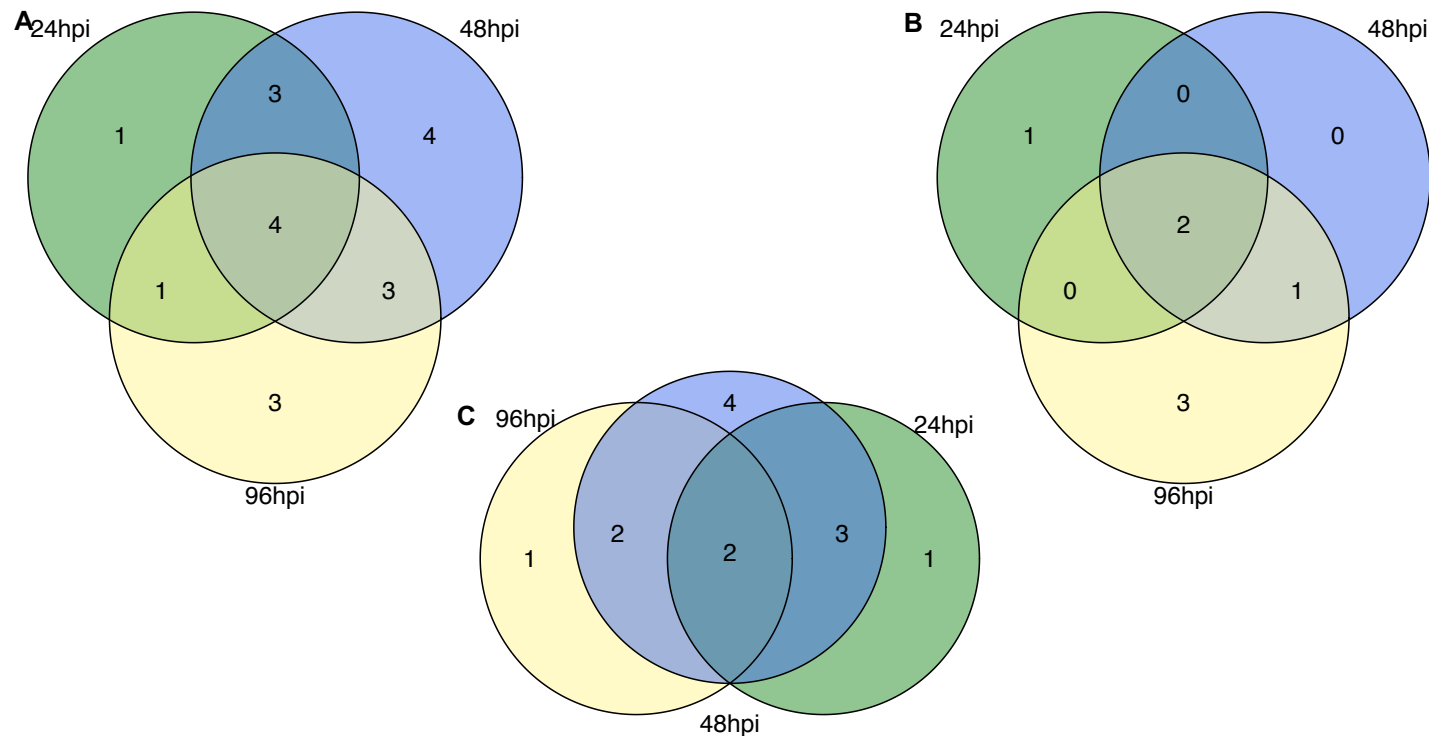




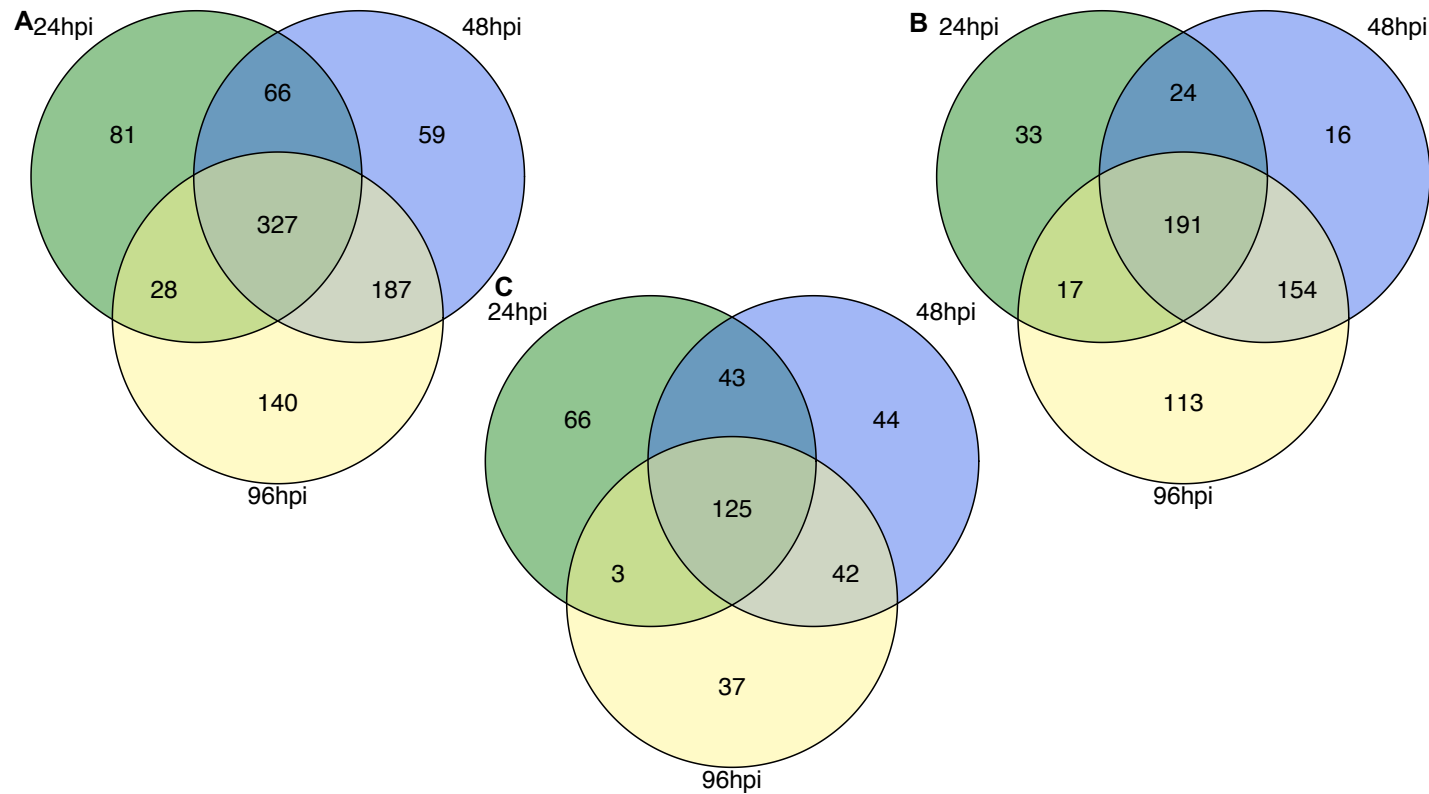
**Fig. 4.7: Venn diagram of all differentially expressed transcripts in the BC-16 isolate of *Phytophthora fragariae*.** RNA-Seq reads were aligned to the assembly of the BC-16 isolate of *P. fragariae* using STAR version 2.5.3a (Dobin *et al.*, 2013). Predicted transcripts were then quantified with featureCounts version 1.5.2 (Liao *et al.*, 2014) and differential expression was identified with DESeq2 version 1.10.1 (Love *et al.*, 2014). Following this venn diagrams were plotted using the VennDiagram R package version 1.6.20 (Chen and Boutros, 2011) in R (R core team, 2016). **A:** All differentially expressed transcripts. **B:** All differentially expressed transcripts displaying a log<sub>2</sub> fold change (LFC) greater than 1. **C:** All differentially expressed transcripts displaying an LFC less than -1.



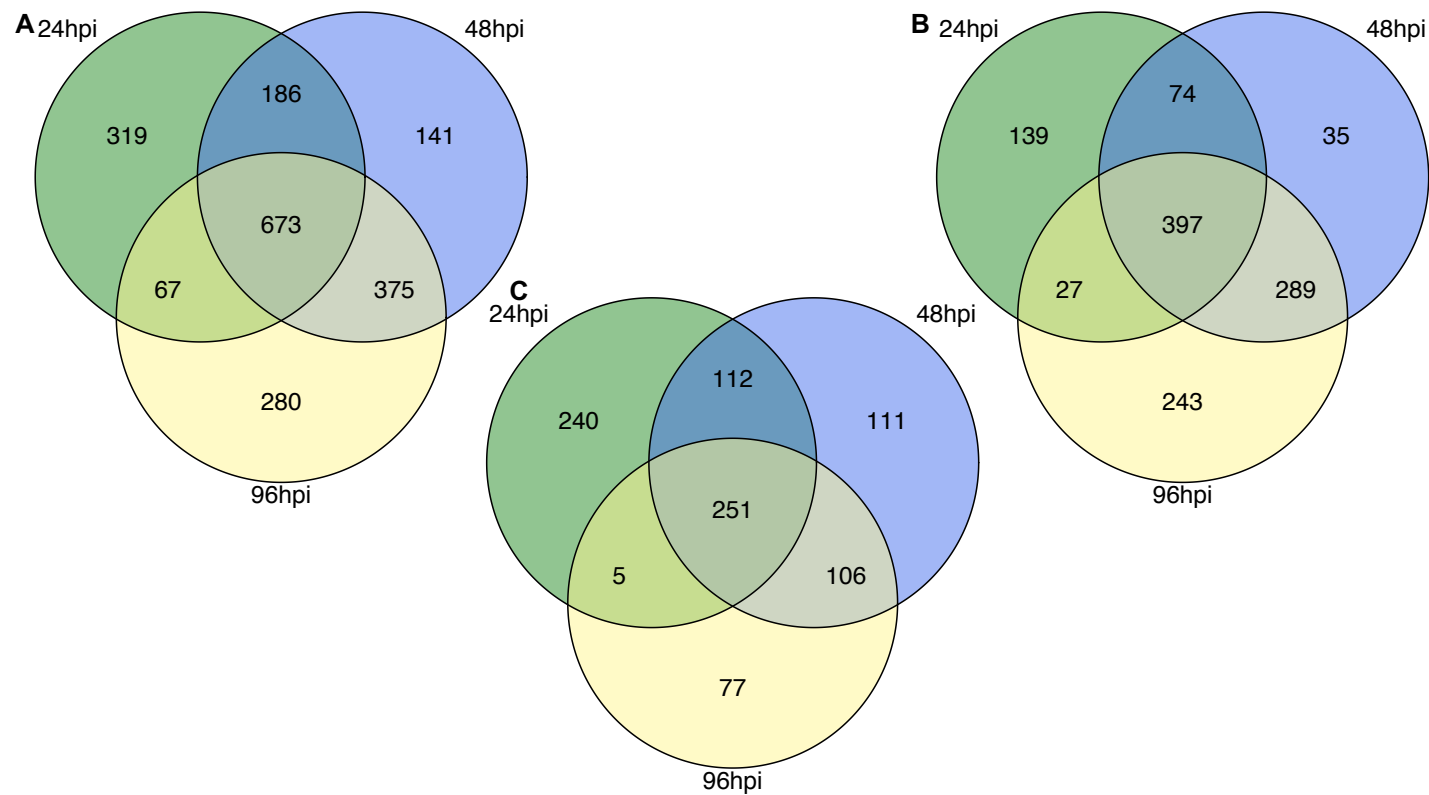
**Fig. 4.8: Venn diagram of all differentially expressed RxLR effectors in the BC-16 isolate of *Phytophthora fragariae*.** RNA-Seq reads were aligned to the assembly of the BC-16 isolate of *P. fragariae* using STAR version 2.5.3a (Dobin *et al.*, 2013). Predicted transcripts were then quantified with featureCounts version 1.5.2 (Liao *et al.*, 2014) and differential expression was identified with DESeq2 version 1.10.1 (Love *et al.*, 2014). Following this Venn diagrams were plotted using the VennDiagram R package version 1.6.20 (Chen and Boutros, 2011) in R (R core team, 2016). **A:** All differentially expressed RxLR effectors. **B:** All differentially expressed RxLR effectors displaying a log<sub>2</sub> fold change (LFC) greater than 1. **C:** All differentially expressed RxLR effectors displaying an LFC less than -1.



**Fig. 4.9: Venn diagram of all differentially expressed crinkler effectors in the BC-16 isolate of *Phytophthora fragariae*.** RNA-Seq reads were aligned to the assembly of the BC-16 isolate of *P. fragariae* using STAR version 2.5.3a (Dobin *et al.*, 2013). Predicted transcripts were then quantified with featureCounts version 1.5.2 (Liao *et al.*, 2014) and differential expression was identified with DESeq2 version 1.10.1 (Love *et al.*, 2014). Following this Venn diagrams were plotted using the VennDiagram R package version 1.6.20 (Chen and Boutros, 2011) in R (R core team, 2016). **A:** All differentially expressed crinklers effectors (CRNs). **B:** All differentially expressed CRNs displaying a log<sub>2</sub> fold change (LFC) greater than 1. **C:** All differentially expressed CRNs displaying an LFC less than -1.



**Fig. 4.10: Venn diagram of all differentially expressed apoplastic effectors in the BC-16 isolate of *Phytophthora fragariae*.** RNA-Seq reads were aligned to the assembly of the BC-16 isolate of *P. fragariae* using STAR version 2.5.3a (Dobin *et al.*, 2013). Predicted transcripts were then quantified with featureCounts version 1.5.2 (Liao *et al.*, 2014) and differential expression was identified with DESeq2 version 1.10.1 (Love *et al.*, 2014). Following this Venn diagrams were plotted using the VennDiagram R package version 1.6.20 (Chen and Boutros, 2011) in R (R core team, 2016). **A:** All differentially expressed apoplastic effectors. **B:** All differentially expressed apoplastic effectors displaying a log<sub>2</sub> fold change (LFC) greater than 1. **C:** All differentially expressed apoplastic effectors displaying an LFC less than -1.



**Fig. 4.11: Venn diagram of all differentially expressed secreted proteins in the BC-16 isolate of *Phytophthora fragariae*.** RNA-Seq reads were aligned to the assembly of the BC-16 isolate of *P. fragariae* using STAR version 2.5.3a (Dobin *et al.*, 2013). Predicted transcripts were then quantified with featureCounts version 1.5.2 (Liao *et al.*, 2014) and differential expression was identified with DESeq2 version 1.10.1 (Love *et al.*, 2014). Following this Venn diagrams were plotted using the VennDiagram R package version 1.6.20 (Chen and Boutros, 2011) in R (R core team, 2016). **A:** All differentially expressed secreted proteins. **B:** All differentially expressed secreted proteins displaying a log<sub>2</sub> fold change (LFC) greater than 1. **C:** All differentially expressed secreted proteins displaying an LFC less than -1.

**Table 4.14: The number of genes showing expression levels greater than the threshold FPKM value of 5 in the BC-16 isolate of *Phytophthora fragariae***

Transcript Class	Total Number	Number of transcripts showing expression				Total
		24hpi	48hpi	96hpi	Mycelium	
All	41,400	9,782 (23.63%)	10,603 (25.61%)	11,763 (28.41%)	11,723 (28.32%)	13,240 (31.98%)
RxLR Effectors	1,058	207 (19.57%)	224 (21.17%)	240 (22.68%)	210 (19.85%)	274 (25.90%)
CRNs	88	20 (22.73%)	21 (23.86%)	24 (27.27%)	26 (29.55%)	27 (30.68%)
Apoplatic Effectors	4,864	562 (11.55%)	642 (13.20%)	733 (15.07%)	649 (13.34%)	880 (18.09%)
Secreted Proteins	8,024	1,823 (22.72%)	2,028 (25.27%)	2,238 (27.89%)	2,063 (25.71%)	2,533 (31.57%)

Hours post inoculation (hpi).

**Table 4.15: The number of transcripts identified as being differentially expressed in any *in planta* time point post inoculation versus *in vitro* grown mycelium in the BC-16 isolate of *Phytophthora fragariae***

Transcript class	Total Number	Number of DETs	Number of Up Regulated transcripts	Number of Down Regulated Transcripts
All Genes	41,400	9,329 (22.53%)	4,506 (10.88%)	5,104 (12.33%)
RxLR Effectors	1,058	253 (23.91%)	148 (13.99%)	106 (10.02%)
Crinklers	88	19 (21.59%)	7 (7.95%)	13 (14.77%)
Apoplatic Effectors	4,864	888 (18.26%)	548 (11.27%)	360 (7.40%)
Secreted Proteins	8,024	2,041 (25.44%)	1,204 (15.00%)	902 (11.24%)

Differentially Expressed Transcript (DET). Up regulated transcripts were those showing a log<sub>2</sub> fold change (LFC) of greater than one. Down regulated transcripts were those showing an LFC of less than minus one.

#### **4.4.9 Assessment of gene copy number variation in Illumina-only genomes showed little evidence of copy number variation between isolates of three distinct pathogenicity races**

It has previously been shown that *de novo* assembly of only paired end reads potentially results in an underestimation of CNV of key effector genes, particularly within the well-studied, related pathogen *P. infestans* (Raffaele *et al.*, 2010; Cooke *et al.*, 2012; Pais *et al.*, 2018). This analysis was conducted on the *P. fragariae* isolates BC-1, BC-16 and NOV-9 representing the UK1, UK2 and UK3 races respectively. As the BC-16 assembly was the reference assembly generated in Chapter 3, Illumina reads for all investigated isolates were aligned to this assembly (Table 4.16). Only 123 genes showed CNV in at least one pairwise comparison of isolates. A similar phenomenon was observed in all analysed effector classes except CRNs, though a larger proportion of putative apoplastic effectors showed CNV. This suggested that the BC-16 assembly had effectively captured the majority of the gene space in the three races of *P. fragariae* analysed. There was a discrepancy between the total number of genes which showed CNV and the sum of the number of genes which displayed increased or decreased copy number in an isolate. This discrepancy was due to genes showing CNV for a single pairwise comparison, but not in the other two isolates. Following the identification of genes that were increased in copy numbers in these three isolates, these genes were checked for evidence of expression. All RxLR genes, apoplastic effectors and secreted proteins with evidence of CNV in any isolate did not show any evidence of expression in the BC-16 RNA-Seq reads, except g7404, a predicted apoplastic effector from unguided gene models. This gene showed an increased copy number in BC-1 and a decreased copy number in BC-16, with NOV-9 at an intermediate level. This gene's expression peaked with an FPKM value of 110 at 96 hpi and had no functional annotations from InterProScan and a search of the Swiss-Prot database. Though this level of expression was low, it was of interest for further analyses.

**Table 4.16: Assessment of copy number variation in three isolates of *Phytophthora fragariae*: BC-1, BC-16 and NOV-9**

Gene Class	Total Number	Number of genes showing copy number variation						Total
		BC-16 Increased	BC-16 Decreased	BC-1 Increased	BC-1 Decreased	NOV-9 Increased	NOV-9 Decreased	
All	41,400	12 (0.03%)	27 (0.07%)	20 (0.05%)	4 (0.01%)	5 (0.01%)	3 (0.01%)	105 (0.25%)
RxLR Effectors	1,058	1 (0.09%)	0 (0.00%)	0 (0.00%)	1 (0.09%)	0 (0.00%)	0 (0.00%)	3 (0.28%)
CRNs	88	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)
Apoplasmic Effectors	4,864	4 (0.08%)	12 (0.25%)	9 (0.19%)	3 (0.06%)	3 (0.06%)	1 (0.02%)	35 (0.72%)
Secreted Proteins	8,024	2 (0.02%)	9 (0.11%)	9 (0.11%)	3 (0.04%)	3 (0.04%)	0 (0.00%)	28 (0.35%)

Crinkler effectors (CRNs).



#### **4.4.10 Orthology analysis identified presence/absence variation between *Phytophthora fragariae* and *Phytophthora rubi* effectors with a single possible avirulence gene for race UK2**

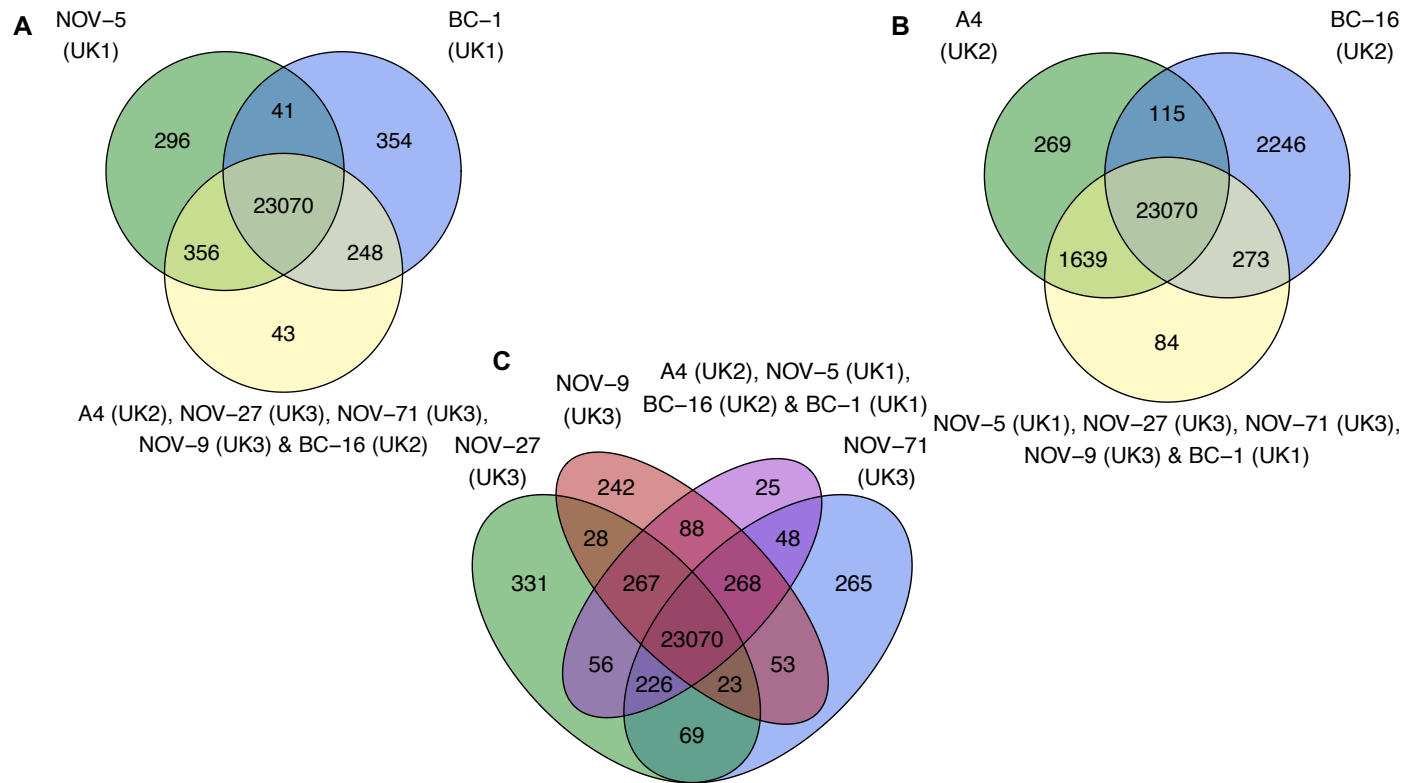
An orthology analysis successfully assigned 481,942 proteins, 98.7% of the total input proteins, into 38,891 orthogroups, excluding singleton orthogroups. Fifty percent of the input genes were in orthogroups with 14 or more proteins and were contained in the largest 12,375 orthogroups. The results also showed that 17,101 orthogroups contained all isolates sequenced and 13,132 of these consisted entirely of single-copy proteins.

Further analysis showed that 2,345 orthogroups were unique to *P. rubi* and 1,911 orthogroups were unique to *P. fragariae*. Within *P. fragariae*: 10 orthogroups were unique to isolates of race UK1, 26 orthogroups were unique to isolates of race UK2, 3 orthogroups were unique to isolates of race UK3, 302 orthogroups were unique to isolates of race CA4, 30 orthogroups were unique to isolates of race CA5 and 285 orthogroups were unique to isolates where the race was unknown. Further examination of isolates of races UK1, UK2 and UK3 showed 54 orthogroups contained only proteins from the seven isolates assigned to these races, whilst 23,070 groups contained proteins from the seven isolates assigned to these races, but also contained proteins from other isolates. Within this subset of isolates of races UK1, UK2 and UK3: 3,024 orthogroups were unique to these isolates, 41 orthogroups were unique to isolates of race UK1, 115 orthogroups were unique to isolates of race UK2 and 23 orthogroups were unique to isolates of race UK3. Venn diagrams were plotted focusing on isolates of each race (Fig. 4.12). This showed a large number (23,070) of orthogroups were shared between the seven isolates representing these three races. The results also showed large numbers of groups shared by some isolates of a race with isolates of other races but not with isolates of the same race, such as 248 proteins being shared between BC-1 and the non-UK1 isolates, but not with NOV-5.

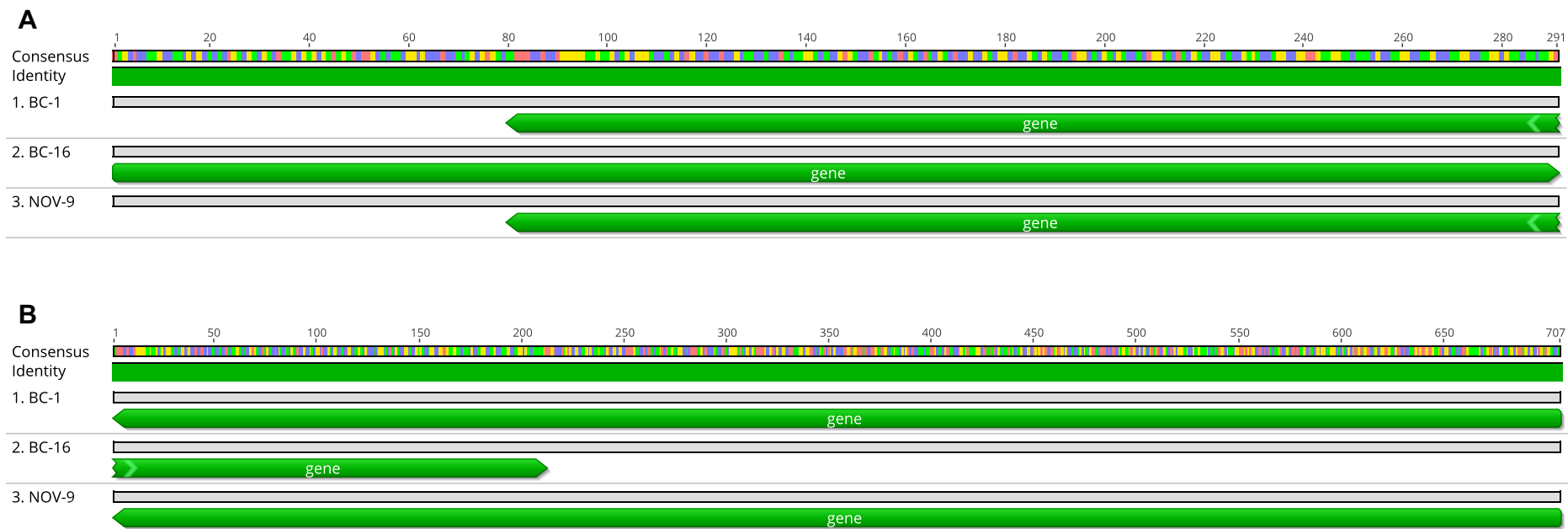
Following this, unique and expanded orthogroups were identified. In this case, unique orthogroups were orthogroups where only isolates of one of race UK1, UK2 or UK3 were present; though the presence of isolates of races CA4, CA5 or

unknown race was tolerated, as was the presence of *P. rubi* isolates. Expanded orthogroups were orthogroups where the number of proteins for isolates of one race of UK1, UK2 or UK3 was greater than the number of proteins for each isolate of the other two races of UK1, UK2 or UK3. Isolates of race CA4, CA5 or unknown race were not used for calling expanded groups, neither were *P. rubi* isolates. These orthogroups were then extracted and examined for the presence of previously identified effectors, along with secreted proteins for further analysis (Table 4.17). All candidate orthogroups were investigated manually to analyse whether they contained a candidate avirulence protein. There were many reasons for rejecting the proteins contained in candidate orthogroups, including: the protein sequence exactly matching that of other isolates, differences in the protein sequence within isolates of the race of interest, gene prediction errors and possible assembly errors. An example of a gene prediction error was in the orthogroup OG0036136, containing the BC-16 gene g36121 (Fig. 4.13). Here the prediction from ApoplastP on low confidence gene models in BC-16 was an exact sequence match to a prediction from BRAKER1 in BC-1 and NOV-9. In this particular case, there was also no evidence of expression in BC-16 for the apoplastic effector.

One candidate avirulence protein was identified in the UK2 race, g36121.t1 from BC-16 in OG0036240. This gene was an apoplastic effector from the lower confidence gene models. When the gene sequence was analysed by BLASTn in Geneious R10 (Kearse *et al.*, 2012) against the assemblies of BC-1 and NOV-9, only partial hits were obtained with a single deletion in the BC-16 sequence. Evidence of expression was then checked. This gene showed a peak FPKM of 259 at 48 hpi. As the BLASTn analysis showed only partial hits, the alignments generated for CNV identification were investigated visually in IGV (Thorvaldsdóttir *et al.*, 2013; Fig. 4.14 and 4.15). The Illumina reads of BC-1 and NOV-9 mapped with high accuracy and coverage to the gene region. This suggested that the lack of a BLASTn hit in the BC-1 and NOV-9 assemblies may be due to an assembly error, however it was investigated further.



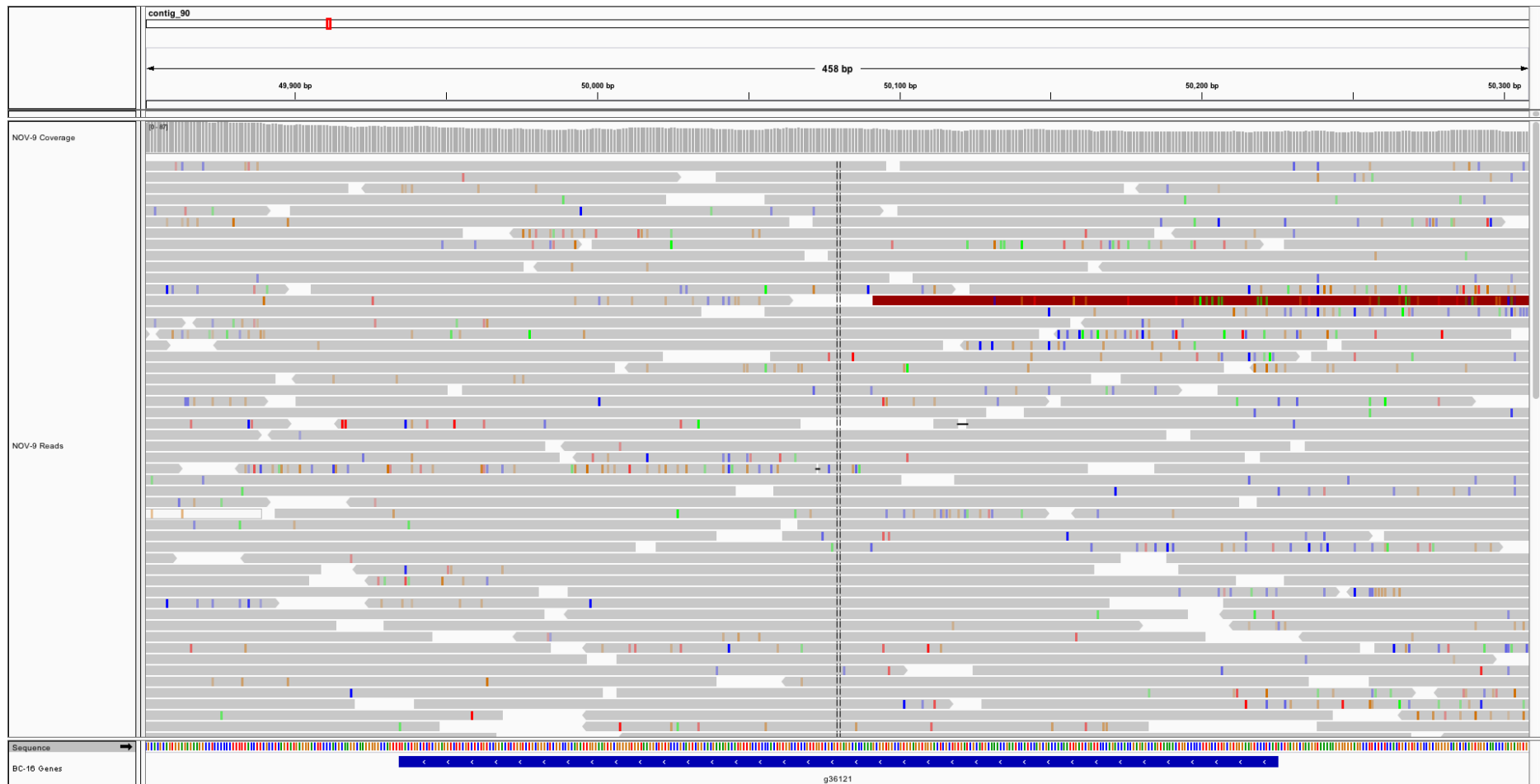
**Fig. 4.12: Venn diagram of the distribution of shared orthogroups within a subset of seven isolates of *Phytophthora fragariae* representing three different races.** Orthology groups were identified by OrthoFinder version 1.1.10 (Emms and Kelly, 2015) and Venn diagrams were plotted using the VennDiagram R package version 1.6.20 (Chen and Boutros, 2011) in R (R core team, 2016). **A:** Analysis focused on the *P. fragariae* isolates of race UK1: BC-1 and NOV-5 compared to isolates of races UK2 and UK3. **B:** Analysis focused on the *P. fragariae* isolates of race UK2: A4 and BC-16 compared to isolates of races UK1 and UK3. **C:** Analysis focused on the *P. fragariae* isolates of race UK3: NOV-5, NOV-27 and NOV-71 compared to isolates of races UK1 and UK2.



**Fig. 4.13: Alignments of BLASTn hits of a rejected candidate avirulence gene from BC-16 against BC-1 and NOV-9. A:** Alignment of the full-length prediction in BC-16 against the regions hit by BLASTn in BC-1 and NOV-9, showing 100% sequence identity despite different annotations. **B:** Alignment of the full-length prediction in BC-1 and NOV-9 against the region hit by BLASTn in BC-16, showing 100% sequence identity despite different annotations. Figures created using Geneious R10 (Kearse *et al.*, 2012).



**Fig. 4.14: Alignment of Illumina reads from BC-1 against the annotated BC-16 assembly.** g36121 was identified as a putative avirulence gene from manual searches of orthology groups. The BC-1 sequencing reads show up to 89x coverage of the gene, with high sequence identity. Figure created using IGV (Thorvaldsdóttir *et al.*, 2013).



**Fig. 4.15: Alignment of Illumina reads from NOV-9 against the annotated BC-16 assembly.** g36121 was identified as a putative avirulence gene from manual searches of orthology groups. The NOV-9 sequencing reads show up to 80x coverage of the gene, with high sequence identity. Figure created using IGV (Thorvaldssdóttir *et al.*, 2013).

**Table 4.17: Summary of expanded and unique orthogroup investigation for races UK1 (BC-1), UK2 (BC-16) and UK3 (NOV-9) of *Phytophthora fragariae***

Isolate	Expanded orthogroups	Unique orthogroups	Candidate expanded orthogroups	Candidate unique orthogroups	Putative avirulence genes
BC-1	38	41	9	10	0
BC-16	117	115	29	25	1 (Low confidence)
NOV-9	18	23	1	3	0

Candidate groups were identified as groups containing either: RxLR effectors, crinkler effectors, apoplastic effectors or secreted proteins. The low confidence avirulence protein in BC-16 may be due to an assembly error in BC-1 and NOV-9.

## 4.5 Discussion

### 4.5.1 Development of a novel method for the inoculation of micropropagated *Fragaria* × *ananassa* plants with *Phytophthora fragariae* allowed for the extraction of RNA from a time course experiment

The method of inoculation of *F.* × *ananassa* plants with *P. fragariae* as detailed in van de Weg *et al.* (1996) requires plants to be grown in soil. Therefore, it was of interest to develop an inoculation method that could be used on micropropagated plants, in order to perform RNA sequencing with a lower risk of contamination from other organisms in the soil. An initial concept was to produce a suspension of zoospores to pipette onto micropropagated plant roots to provide an inoculation with a known number of zoospores. Previously, methods of producing zoospore suspensions of *P. fragariae* have been described by Goode (1956). However, the author returned pots to sand following inoculation and attempts to replicate the production of a zoospore suspension using their methods were not successful. Additionally, attempts to generate a zoospore suspension by varying flooding solutions and growth media, in a method similar to that used for *P. sojae* (formerly *P. megasperma* f. sp. *glycinea*; Ward *et al.*, 1989), proved unsuccessful for *P. fragariae*. However, these attempts only failed at the stage where the sporangia were induced to burst in a synchronised manner to produce a zoospore suspension, meaning some of the trialled medias still had large numbers of intact sporangia present. During observations, a small number of motile zoospores were observed, with sporangia appearing to burst at a slow rate during incubation. Therefore, it was decided to test whether simply submerging the roots of micropropagated *F.* × *ananassa* plant roots in the flooded plates for a time would be sufficient for inoculation, similar to the method used in Blackman *et al.* (2015). It was found that being submerged for one hour resulted in successful infection of susceptible 'Hapil' plants by the BC-16 isolate of *P. fragariae*. The presence of the pathogen was confirmed by a nested PCR method (Bonants *et al.*, 1997). However, this nested PCR protocol amplified the ITS regions between ribosomal RNA (rRNA) coding genes. As a result, these sequences would not be present when total RNA was extracted, as was planned for the inoculation time course experiment. In order to allow detection of *P.*



*fragariae* from extracted RNA, a novel pair of primers for the  $\beta$ -tubulin gene of the BC-16 isolate were developed and optimised for an RT-PCR based method of detection.

Following a successful pilot experiment utilising both the novel inoculation and novel RT-PCR based detection, an inoculation time course was performed for the BC-16 isolate of *P. fragariae* inoculated on to the susceptible cultivar 'Hapil' of *F. × ananassa*. A number of time points post inoculation were sampled and tested for the presence of *P. fragariae*, leading to the selection of three samples from three time points for RNA-Seq, alongside three samples from *in vitro* grown mycelium.

#### **4.5.2 *Ab initio* annotation of assembled genomes and prediction of effector class genes produced a valuable resource for pathogenomic investigations**

The generation of RNA-Seq data, alongside the acquisition of *P. rubi* data from the Grünwald Lab at Oregon State University as part of the *Phytophthora* Sequencing Consortium, allowed for *ab initio* prediction of gene models in all the isolates assembled in Chapter 3. This resulted in approximately 30,000 RNA-Seq guided gene models in all the Illumina-only sequenced isolates and approximately 37,000 in the PacBio sequenced isolate of *P. fragariae*, BC-16. This complemented the results shown in Chapter 3, where the long-read sequencing produced a larger assembly size. This now showed that the extra assembled regions were likely not solely gene poor repeat rich regions and may contain valuable biological information.

Three classes of effector genes were also identified: RxLR effectors, CRNs and putative apoplastic effectors. In all Illumina-only sequenced isolates of *P. fragariae* and *P. rubi*, approximately 400 RxLR effectors were identified, approximately 50 CRNs were identified in *P. fragariae* and approximately 100 CRNs were identified in *P. rubi* and approximately 1,000 apoplastic effectors were identified in both species. These values were higher in the BC-16 assembly of *P. fragariae* with: 486 RxLR effectors, 82 CRNs and 1,274 apoplastic effectors. This increase was not unexpected as there were more genes in total in the BC-

16 assembly. However, it was unexpected to have twice the number of CRNs in *P. rubi* compared to *P. fragariae*, though this was not consistent for all isolates of *P. rubi* as SCRP324 had 50 less CRNs predicted than the other *P. rubi* isolates. These predictions represented approximately 50% more RxLR effectors and twice as many CRNs in the BC-16 genome than in the *P. sojae* genome (Tyler *et al.*, 2006), however, this may have been due to improvements in the ability to predict putative effectors as a similar number were of CRNs were predicted in *P. cactorum* (Armitage *et al.*, 2018b).

Effectors were also identified from unguided gene predictions; however, these predictions would likely contain a large number of false positive predictions. Between 2,100 - 2,200 RxLR effectors, approximately 100 CRNs in *P. fragariae* and approximately 130 CRNs in *P. rubi* and approximately 10,000 apoplastic effectors predicted in the Illumina-only assemblies. These values were again increased in the BC-16 assembly, with 2,448 RxLR effectors, 118 CRNs and 12,431 apoplastic effectors predicted. Again, the increase in predicted effectors in BC-16 was not unexpected, given the larger assembly size. However, it was again observed that there were an increased number of CRNs predicted in *P. rubi*, although this time the number predicted in SCRP324 was not significantly lower.

Low confidence effector genes were added into the gene model set when they resided in intergenic regions with respect to the guided gene models. This has resulted in approximately 33,000 predicted genes in the Illumina-only assemblies and 41,103 predicted genes in the BC-16 assembly. When effector genes were examined in this gene model set, the Illumina-only assemblies showed: between 800 and 900 RxLR effectors, between 60 and 70 CRNs for *P. fragariae* and 130 CRNs in two isolates of *P. rubi*, with 87 in SCRP324 and between 3,300 and 4,000 apoplastic effectors. Again, there were more CRNs in *P. rubi* than *P. fragariae*, this could potentially be a genuine biological difference, or it may be due to the CRNs in *P. rubi* being a better match for the HMM than in *P. fragariae*. This was similar to results from Armitage *et al.* (2018b) where between 35 and 265 CRNs were identified depending on the species investigated. Interestingly, there appeared to be fewer genes predicted as apoplastic effectors in *P. rubi* than

in *P. fragariae*, though again SCRP324 appeared to show a different number of predictions than the other *P. rubi* isolates, with approximately 300 more apoplastic effectors being predicted. This was similar to the lowest number of apoplastic effectors predicted in a *P. fragariae* isolate, so may not have been due to a biological difference. The number of predicted effectors were again increased in the BC-16 assembly, with 1,058 RxLR effectors, 88 CRNs and 4,864 apoplastic effectors being predicted.

Following the creation of a final set of draft annotations, the annotated genomes were parsed to the format required for submission to NCBI. These genomes were then uploaded to NCBI GenBank (NCBI Resource Coordinators, 2018) as part of Bioproject PRJNA396163 and will be made publicly available upon these assemblies appearing in a peer reviewed publication. These annotated assemblies will prove a useful tool for future research on both *P. fragariae* and the wider *Phytophthora* genus.

#### **4.5.3 Analysis of expression patterns during the inoculation time course and in relation to *in vitro* grown mycelium showed the time course experiment had captured numerous stages of the infection process**

The RNA-Seq time course data allowed for quantification of the expression levels of predicted transcripts at different times during the infection process. The results of the expression analysis showed over 30% of the predicted transcripts were expressed in mycelium or *in planta*. This was as expected as only growth on an artificial medium and in the early stages of the infection process were analysed. It could have been the case that other genes that were not showing evidence of expression may be expressed during different life stages of *P. fragariae*, similar to observations of different classes of expression patterns in RxLR effectors and CRNs in *P. capsici* (Jupe *et al.*, 2013; Stam *et al.*, 2013). Alternatively, they may only be expressed under certain environmental conditions that were not investigated as a part of this study. It was also possible that some of the low confidence effector genes added to the gene model set did not represent true genes. This was also suggested by the lower proportions of effector proteins of all classes showing expression than the number of transcripts overall. This was

especially true for putative apoplastic effectors, with only 18% showing evidence of expression in this analysis. However, these genes were identified by a greedy approach to ensure all effector genes were captured in order to allow for identification of candidate avirulence genes. This likely included several false positives, as suggested by the numbers of these genes identified being larger than for other *Phytophthora* spp. (Armitage *et al.*, 2018b). Interestingly, the proportion of secreted proteins showing expression was higher than that for transcripts overall, this suggested that the methods used to detect evidence of secretion may have had high accuracy. It was also possible that some of the predictions from the automated RNA-Seq guided methods may be erroneous, as during orthology analysis it was observed that regions with the exact same nucleotide sequence were annotated differently in the assemblies of different isolates.

Analysis of differential expression in each sequenced time point from the inoculation time course experiment against the expression in *in vitro* grown mycelium was also conducted. A large number (23%) of transcripts were differentially expressed in all time points from the inoculation time course experiment compared to *in vitro* grown mycelium, which was expected as the conditions growing in artificial media compared to growing within a host plant differed by a large amount. It was also possible that epigenetic changes may have occurred when the axenic culture was used to inoculate a host plant, as has been previously shown for the fungal pathogen of Brassica crops *L. maculans* (Soyer *et al.*, 2014). A slightly larger proportion of transcripts showed down regulation of expression than up regulation of expression, though the difference was very slight. This suggested a wide ranging transcriptional reprofiling was underway during infection. An investigation of transcripts showing differential expression in only two time points showed a larger number were shared between sequential time points than were shared between 24 hpi and 96 hpi only. This suggested that the inoculation time course experiment had captured the progression of the infection process. For RxLR effectors, apoplastic effectors and secreted proteins, a greater proportion of the differentially expressed transcripts showed evidence of up regulation in samples from the inoculation time course experiment than down regulation, whereas for CRNs a greater proportion were

down regulated during infection than were up regulated. This suggested that the transcriptional reprofiling during infection was specific, with genes potentially involved in pathogenicity being up regulated. It was interesting to observe a greater number of down regulated CRNs than up regulated. This may have been due to a clearance of silencing markers during axenic culture for CRNs, or that the changes in expression were specific and an attempt to avoid resistance. It may also be the case that the CRNs would increase in expression level later in the infection process.

#### **4.5.4 Copy number variation analysis showed the reference assembly had captured the majority of copy number variation in *Phytophthora fragariae*, with one gene being of possible future interest**

CNV within three isolates: BC-1 (race UK1), BC-16 (race UK2) and NOV-9 (race UK3) was investigated using the raw Illumina reads to assess average read depth (Raffaele *et al.*, 2010; Cooke *et al.*, 2012; Pais *et al.*, 2018). Only a very small number of genes showed any evidence of CNV, suggesting the long-read assembly had captured the majority of the gene space for these three races of *P. fragariae*. Investigation of the expression levels of all genes showing CNV showed that the majority of genes showing CNV showed no evidence of expression in the BC-16 isolate. There was one exception to this, g7404, an apoplastic effector from low confidence gene models, which showed an increased copy number in BC-1 and a decreased copy number in BC-16, with NOV-9 at an intermediate level. The gene showed a peak in expression at 96 hpi with an FPKM of 110 and had no identified functional domains. Therefore, this gene will be of interest for further investigations. Whilst previous work in *P. infestans* has identified CNV in avirulence genes (Raffaele *et al.*, 2010; Cooke *et al.*, 2012; Pais *et al.*, 2018), the control of the avirulence phenotype was not shown to be due to the CNV itself, but rather due to sequence polymorphisms or expression differences. Additional studies in *P. sojae* have demonstrated several avirulence genes, many of which showed CNV (Dong *et al.*, 2009; Qutob *et al.*, 2009). As with *P. infestans*, the control of the avirulence was not caused by the CNV but by sequence polymorphisms or expression differences.

#### **4.5.5 Orthology analysis identifies a candidate avirulence gene for race UK2**

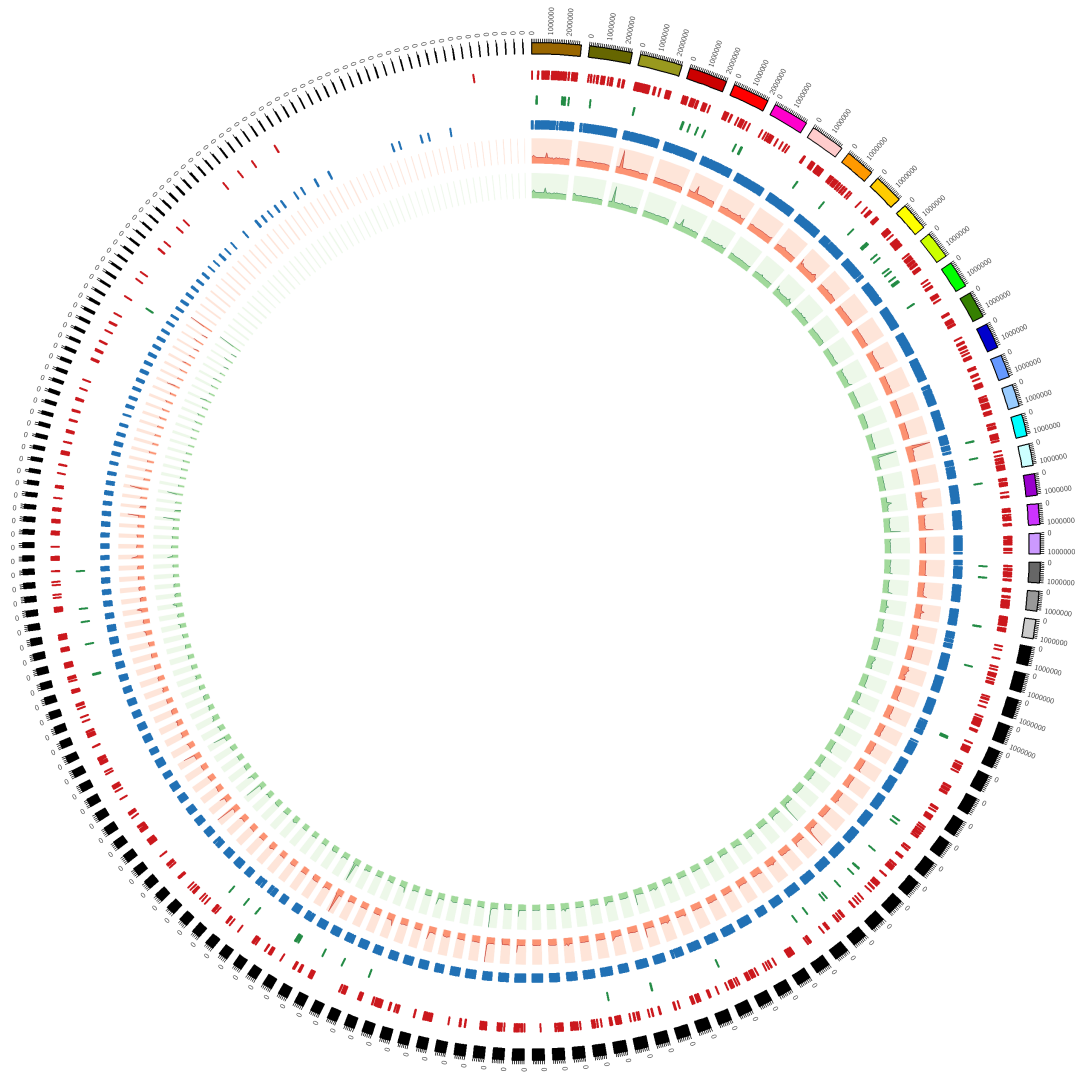
Orthology groups were created from all predicted proteins in all eleven sequenced isolates of *P. fragariae* and all three sequenced isolates of *P. rubi*. Analysis was focused on isolates of race UK1, UK2 and UK3. Only a small number of orthology groups were shown to be unique or expanded in each individual race, with over 23,000 groups being shared between these races. These groups were narrowed down to those showing the presence of effectors or secreted proteins. These orthogroups were investigated in order to identify orthogroups containing proteins that showed sequence differences between the races by manual investigation, reducing the number of candidate avirulence proteins yet further. Several cases were observed when a large, identical region of genome sequence would contain different predictions eg. a BRAKER1 gene in one isolate, but an apoplastic effector from low confidence models in another. This was likely due to errors in the automated RNA-Seq guided gene prediction processes and so these candidates were eliminated. Checks for expression in BC-16 allowed for the elimination of all but one candidate, g36121. However, as detailed in the results section of this chapter, this appeared to be due to an assembly error resulting in a portion of the gene failing to assemble in a number of the isolates as raw sequencing reads map completely. Despite these results, it remains a low confidence candidate for the avirulence gene in race UK2 and so the gene will be investigated further.

#### **4.5.6 Conclusions**

The development of a novel method of inoculating micropropagated *F. × ananassa* plants with *P. fragariae* allowed for the generation of RNA-Seq data. This will be a useful approach for work in Chapter 5 as well as for future investigations of *P. fragariae*, particularly as analyses of expression patterns at different time points post inoculation show changes in expression levels for predicted effectors. The *ab initio* gene predictions produced in this investigation represent a key step forward for pathogenomic analyses of this organism, both for further work in future chapters and for other researchers, following publication.

A key result from the work presented in this chapter is the annotation of a novel reference assembly of an isolate of *P. fragariae* from long read sequencing data (Fig. 4.16). Interestingly, the patterns of the coverage in BC-1 and NOV-9 showed similar patterns. This suggested high levels of similarity in the presence of regions in the genomes and potentially suggested a high level of sequence similarity. This will be examined in further work. Finally, two genes of interest have been identified as possibly being involved in the avirulence response. Firstly, g7404 in the BC-16 assembly, which showed CNV in BC-1 and NOV-9. It will be of interest to examine whether the expression patterns also show this difference. Secondly, g36121, which appeared to be unique in UK2 isolates through orthology analyses and showed expression in BC-16, leading it to be considered as a potential candidate for the avirulence gene of race UK2. However, it also appeared that the absence in other isolates may be due to assembly errors. Again, it will be of interest to see if there is evidence of expression in isolates of race UK1 and UK3.

**Fig. 4.16 (overleaf): Summary of the *de novo* assembly of the BC-16 isolate of *Phytophthora fragariae* from long read PacBio data, featuring the location of effector genes and mapping rates of raw short reads. Outside circle:** Every contig of the FALCON-Unzip assembly of the BC-16 isolate of *P. fragariae*. **Red ticks:** The location of RxLR effector genes. **Green ticks:** The location of crinkler effector genes. **Blue ticks:** The location of putative apoplastic effectors as predicted by ApoplastP (Sperschneider *et al.*, 2018). **Red plots:** The coverage of Illumina reads of the BC-1 isolate of *P. fragariae* mapped to the assembly with Bowtie 2 (Langmead and Salzberg, 2012). **Green plots:** The coverage of Illumina reads of the NOV-9 isolate of *P. fragariae* mapped to the assembly with Bowtie 2. Figure created with Circos (Krzywinski *et al.*, 2009).





## **Chapter 5: Population genetic investigations and identification of candidate avirulence genes in three pathogenicity races of *Phytophthora fragariae***

### **5.1 Introduction**

#### **5.1.1 Pathogenomic investigations beyond presence/absence variations**

As the orthology analysis in Chapter 4 failed to identify candidate avirulence genes for pathogenicity races UK1 and UK3, it was determined that the race structure must have been caused by variation beyond a simple presence/absence model. Whilst presence/absence variation can explain pathogenicity differences, factors such as differential expression and sequence polymorphisms can also be involved. An example of this was shown in *Phytophthora infestans* by Gilroy *et al.* (2011) with isolates showing both sequence polymorphisms and expression differences in the *PiAVR2* gene leading to the isolates regaining the virulence phenotype. However, the exact situation for pathogenicity on potato plants containing *R2* was shown to be highly complex, involving multiple genes in *P. infestans*.

Whilst the majority of pathogenomic work within the *Phytophthora* genus has been carried out on *P. infestans*, the race structure of *Phytophthora fragariae* was more similar to that of *Phytophthora sojae*. Investigations of the resistance gene *Rps1b* in soybean (*Glycine max*) allowed for the cloning of two avirulence genes: *Avr1b-1* and *Avr1b-2* from *P. sojae*. Critically, *Avr1b-1* was identified in both virulent and avirulent isolates of *P. sojae*, however, in some avirulent isolates several nucleotide substitutions were observed in the coding sequence of this gene. These substitutions led to changes in the amino acid sequence and the abolition of the avirulence phenotype (Shan *et al.*, 2003). A similar phenomenon was also shown during investigations of the resistance gene *Rps3c* in soybean. This resulted in the identification of the avirulence gene *Avr3c* in *P. sojae*. This gene was found to be present in both virulent and avirulent isolates of *P. sojae*, but in virulent isolates the gene showed amino acid polymorphisms,

which allowed the pathogen to evade effector recognition (Dong *et al.*, 2009). Interestingly, both these studies were performed for genes with copy number variation (CNV) between isolates of *P. sojae*.

A panel of single nucleotide polymorphisms (SNPs) in *P. fragariae* and *Phytophthora rubi* was recently identified by Tabima *et al.* (2018) with a genotyping by sequencing approach on 138 isolates of *P. rubi* and 45 isolates of *P. fragariae*. This study focused on *P. rubi* and called variants with reads aligned to a comparatively fragmented *P. rubi* reference assembly (Tabima *et al.*, 2017). As such it was of interest to identify new variant sites, using the assembly created in Chapter 3 of the BC-16 isolate of *P. fragariae* as a reference, despite the smaller sample size of this study.

Sequence variation was often not sufficient to explain all observed avirulence phenotypes in *P. sojae*. For instance, some of the isolates of *P. sojae* examined in Shan *et al.* (2003) showed virulence on soybean cultivars containing *Rps1b* despite the lack of substitution mutations in *Avr1b-1*. However, through a northern blotting analysis, these isolates were shown to lack mRNA of the avirulence gene. Investigation of *Avr1a* and *Avr3a* by Qutob *et al.* (2009) identified these two genes and demonstrated their CNV in *P. sojae* isolates. This work also showed that only avirulent isolates accumulated transcripts of these genes through reverse transcription polymerase chain reaction (RT-PCR), though they did not identify any cause for the abolition of expression of these genes (Qutob *et al.*, 2009). Later work by the same group on *Avr1c* in *P. sojae* identified the gene by sequencing pooled samples of virulent and avirulent isolates. One of the virulent isolates they examined lacked the avirulence gene, but the rest of the virulent isolates examined still had the gene present. Through the use of an RT-PCR analysis, all but one virulent isolate showed the abolition of expression of the transcript of this gene (Na *et al.*, 2014). More recently, investigations into the EC-1 clonal lineage of *P. infestans* revealed variation in pathogenicity of isolates on potato (*Solanum tuberosum*) lineages possessing the *Rpi-vnt1.1* resistance gene. The authors showed that in the absence of genetic mutations, differences in the expression levels of the effector gene

*Avrvnt1* were detected and correlated with virulence, perhaps due to differential epialleles (Pais *et al.*, 2018).

Recently, an investigation of seven *P. sojae* avirulence genes reanalysed several genes, including *Avr1a* and *Avr1c*. For *Avr1a*, the authors confirmed that complete deletion of the gene was not the sole explanation for the maintenance of virulence on soybean plants with the *Rps1a* gene. The authors identified new SNPs outside the gene coding region that correlated with virulence, though they were unable to determine the functional impact of these SNP sites it is possible they were influencing transcription factor binding or modifying promoter sequences. The authors also analysed *Avr1c*, where they suggested incorrect phenotyping was the cause of the results shown by Shan *et al.* (2003) and Na *et al.* (2014) described above. However, they did not rule out the possibility of epistatic effects causing the avirulence phenotype (Arsenault-Labrecque *et al.*, 2018). Therefore, it was of interest to investigate regions surrounding potential candidate avirulence genes in order to increase the understanding of the race structure of *P. fragariae*.

### **5.1.2 Population genetics insights in filamentous plant pathogens**

Population genetics analyses provide meaningful insights into the evolutionary state of pathogens and can be used to horizon scan and identify risks associated with pathogen evolution over time.

Most frequently in studies of *Phytophthora* spp., population genetics analyses are used to define the structure of subpopulations. In early work, this was performed using amplified fragment length polymorphisms (AFLPs) on *Phytophthora cactorum* from multiple hosts in California. This work calculated pairwise fixation index ( $F_{ST}$ ) values and demonstrated large degrees of population separation being driven by host preference (Bhat *et al.*, 2006). Additionally, work using simple sequence repeat (SSR) markers on *P. infestans* on different host species in Ecuador demonstrated substructuring of populations through the calculation of  $F_{ST}$  and showed evidence of co-evolution between populations of the pathogen and host species (Oliva *et al.*, 2007). Recently,

investigations focused on the closely related pathogen species *P. rubi* analysed populations of this pathogen using a panel of SNP sites. This work showed that clear separation between *P. fragariae* and *P. rubi* through the calculation of a genetic differentiation statistic,  $G_{ST}$ , a measure of genetic differentiation. The results also showed that the population of *P. rubi* within North America showed no evidence of population structure. The authors therefore suggested that *P. rubi* was mostly spread by human mediated plant movement in North America (Tabima *et al.*, 2018).

An analysis of microsatellites from the clade seven pathogen *Phytophthora alni* subsp. *uniformis* isolates for Europe and North America showed clear differentiation between the populations on each continent through the calculation of  $F_{ST}$  and an analogue of  $F_{ST}$  which assumes a stepwise mutation model,  $R_{ST}$ . Both populations were also shown to significantly deviate from Hardy Weinberg proportions, which the authors suggested indicated selfing as the predominant method of sexual reproduction in this species. However, they identified some signs of infrequent outcrossing in the North American population, such as incomplete linkage disequilibrium (LD) and gene diversity closer to that of other outcrossing species of oomycetes. From these results, the authors suggested *P. alni* subsp. *uniformis* was likely alien in Europe and indigenous in North America. However, they were not able to identify the centre of origin for this species (Aguayo *et al.*, 2012). Investigation of microsatellite markers and sequence data in *Phytophthora plurivora* showed low levels of recombination in this species, leading to a conclusion that this species reproduces mostly by selfing with occasional, rare outcrossing events. Analysis of the levels of polymorphism in this species suggested a centre of origin in Europe due to a higher level of diversity in this population compared to a U.S.A. population. The authors also investigated sequence diversity and found higher levels of diversity in regions which showed evidence of recombination, however they were unable to detect the action of selection through calculation of Tajima's D and Fu and Li's D (Schoebel *et al.*, 2014).

More recently, an analysis of 279 isolates of *P. infestans* of the EC-1 clonal lineage from central Colombia for twelve microsatellite markers provided

important information of the population of this key pathogen in an important agronomic area. Through the calculation of pairwise  $F_{ST}$  values, the authors demonstrated that neither the geographic location or host genotype caused significant population structure, despite pairwise isolation by distance measures showing a small amount of variation between isolates from Western and Northern regions of the studied area. Additionally, the authors identified a high level of diversity within this population, with the identification of 76 multi-locus genotypes among the single clonal population. The authors also showed a lack of sexual reproduction within this population, as multi-locus linkage disequilibrium was significantly greater than zero within each population for all comparisons of geographical location and host genotype. This lack of sexual reproduction was reinforced by the identification of all isolates as the A1 mating type (Chaves *et al.*, 2018).

More recently, work on the ash dieback fungus *Hymenoscyphus fraxineus* has investigated the history of this pathogen based on SNP site identification. This pathogen recently invaded Europe from Japan and this study examined a selection of European isolates and a selection of Japanese isolates. The authors showed that the two populations were divergent through calculation of  $F_{ST}$ . The authors also demonstrated that the European population appeared to have suffered a recent genetic bottleneck due to the founder effect. Interestingly, the authors suggested that the European population was founded by two genetically divergent individuals, possibly from the same ascocarp (fruiting body). They demonstrated this through comparison of levels of nucleotide diversity between the populations as well as the loss of both neutral and adaptive variation through calculation of Tajima's D. Additionally, they showed that the European population still undergoes sexual reproduction by analysis of LD decay. Interestingly, genomic regions with high diversity in the Japanese populations were less likely to be shared with the European populations and vice versa, further showing the separation of the populations. Despite the loss of variation due to the founder effect, calculation of the mean ratio of nonsynonymous sites to synonymous sites in genes showed that whilst the majority of genes in the European population showed neutrality, effector genes still showed a greater proportion of nonsynonymous sites, indicating diversifying selection (McMullan *et al.*, 2018).

## 5.2 Aims

Two genes of interest in *P. fragariae* were identified in Chapter 4, one (g36121.t1) as a candidate avirulence gene for race UK2 and one (g7404.t1) as a gene showing CNV polymorphisms in the BC-1 (UK1), BC-16 (UK2) and NOV-9 (UK3) isolates. The work in this chapter aims to further investigate these genes of interest and identify further candidate avirulence genes for races UK1, UK2 and UK3 using both existing and new sequencing data.

The specific aims of the chapter were as follows:

1. Identify a new set of variant sites and structural variants, using the FALCON assembly of BC-16, created in Chapter 3, acting as a reference.
2. Analyse individual pathogenicity races for private variants to identify further candidate avirulence genes.
3. Identify subpopulations and analyse population genetics statistics to investigate population dynamics and evolutionary history.
4. Perform additional RNA-Seq data of a single *in planta* time point of an isolate each of the races UK1 (BC-1) and UK3 (NOV-9), along with mycelial samples, in order to identify additional candidate avirulence genes. These results will be confirmed with Quantitative Reverse Transcription Polymerase Chain Reaction (RT-qPCR).
5. Investigate identified candidate avirulence genes for variants in *cis* and *trans* to improve the mechanistic understanding of the race structure of *P. fragariae*.

## 5.3 Materials and Methods

### 5.3.1 Alignment of raw Illumina sequencing reads to the FALCON assembly of BC-16

The Illumina reads of all three isolates were aligned to the reference assembly of the BC-16 isolate using Bowtie 2 version 2.2.6 (Langmead and Salzberg, 2012). Multi-mapping reads were removed by filtering out on the XS:i tag and reads with the “paired reads” and “properly paired reads” flags were kept using Sequence Alignment/Map tools (SAMtools) version 0.1.18 (Li *et al.*, 2009). Picard tools version 2.5.0 was then used to remove Polymerase Chain Reaction (PCR) and optical duplicates and prepare a genome reference index (<https://github.com/broadinstitute/picard>, last accessed 03/01/2019).

### 5.3.2 Variant calling process and annotation of variant sites

Variant sites were identified using the Genome Analysis Toolkit (GATK) version 3.6 HaplotypeCaller, with the ploidy argument set to two, specifying a diploid genome and allowing potentially misencoded quality scores (McKenna *et al.*, 2010). Following the identification of these sites, VariantsToAllelicPrimitives from GATK was used to convert multiple nucleotide polymorphisms (MNPs) to several SNPs. This produced a file containing all isolates. Another file of just *P. fragariae* isolates was created using vcfremovesamples from vcfliib. High quality, biallelic SNPs with no missing data were kept using vcfliib (Garrison, 2012) with a minimum quality of 40, minimum MQ of 30, minimum depth of 10, minimum GQ of 30. INDELs and sites with greater than 5% missing data were then removed and sites were re-coded using VCFTools (Danecek *et al.*, 2011).

Summary statistics were then calculated using vcf-stats from VCFTools (Danecek *et al.*, 2011). Following this, a custom Python script was used to calculate percentage similarity. These results were visualised as a heatmap and clustering dendrogram using the gplots R package version 3.0.1 (Magnusson *et al.*, 2016) in R (R core team, 2016). Monomorphic sites were removed using VCFTools with a minor allele count of 1 (Danecek *et al.*, 2011). A principal

component analysis (PCA) was also conducted and plotted using the SNPRelate (Zheng *et al.*, 2012), gdsfmt (Zheng *et al.*, 2012), ggplot2 version 2.2.1 (Wickman, 2016) and ggrepel version 0.6.5 (Slowikowski, 2018) R packages in R. A neighbour joining tree was also created from the SNP data using the ape R package version 5.1 (Paradis *et al.*, 2004) in R with 100 bootstrap replicates, after fasta alignments were produced with a Perl script (Bergey, 2012) and one allele was randomly picked at heterozygous sites to obtain haploid sequences with seqtk (Li, 2018b).

A custom SnpEff database was created for the BC-16 assembly and annotations to create the Bc16v1.0 database. This database was then used to annotate the variant sites with SnpEff (Cingolani *et al.*, 2012).

### **5.3.3 Assessment of population structure**

Prior to the assessment of population structure, Plink version 1.90 beta was used to convert the variant calls into the file type required for this analysis, with a window size of 100 kb, a step size variant ct of 1, an  $r^2$  threshold of 0.5, unrecognised chromosome codes allowed, FIDs set to 0, the make-bed option enabled and the recode option enabled (Purcell *et al.*, 2007). Population structure was then assessed for three progressively smaller groups of isolates: all isolates (eleven *P. fragariae* isolates and three *P. rubi* isolates), only *P. fragariae* isolates (eleven *P. fragariae* isolates) and isolates of races UK1, UK2 and UK3 (seven *P. fragariae* isolates) using fastSTRUCTURE with values of K (the number of populations to generate) ranging from one to five. Additionally, the analysis of the seven *P. fragariae* isolates of races UK1, UK2 and UK3 was analysed with the logistic-prior option enabled due to this showing increased sensitivity to subtle population structures (Raj *et al.*, 2014). Python scripts provided with the download of fastSTRUCTURE were used to generate advice on which K value was most informative and plot distruct plots (Rosenberg, 2004).



### **5.3.4 Identification of private and polarising variant sites in races UK1, UK2 and UK3**

In order to identify private, which occur only in a specified race, and polarising, which occur in both the race of interest and an outgroup. variant sites, the unfiltered files containing the race UK1, UK2 and UK3 isolates with and without the *P. rubi* isolates were created using `vcfremovesamples` from `vcflib` (Garrison, 2012) and were again filtered using `VCFTools` (Danecek *et al.*, 2011). A custom Python script was then used to identify variants that were present in either: isolates of a specific pathogenicity race to identify private variants, or isolates of a specific pathogenicity race as well as the isolates of *P. rubi* to identify polarising variants (Sobczyk, unpublished).

### **5.3.5 Identification of structural variants and larger INDELS**

In order to identify larger variants between the isolates, Illumina reads were aligned to the FALCON assembly with the Burrows-Wheeler Aligner (BWA) algorithm `BWA-mem` (Li, 2013). These alignments were used as the input for `SvABA`, to produce a file containing INDELS and a file containing structural variants (Wala *et al.* 2018). Private and polarising variants were then identified as in 5.3.4.

### **5.3.6 Assessments of linkage disequilibrium**

Prior to calculating LD, variant files were phased to assign alleles to specific chromosomal homologues using `Beagle` version 4.1 with a 1 kb window and an overlap of 100 (Browning and Browning, 2007). Following this, the  $r^2$  LD statistics was calculated using `VCFTools` with the `hap-r2` option, a minimum window size of 1 kb and a maximum window size of 100 kb (Danecek *et al.*, 2011). LD decay was calculated using a custom R script to fit the  $r^2$  decay to the Hill and Weir decay function (Hill and Weir, 1988) and plot the resulting curve with `ggplot2` R package version 2.2.1 (Wickman, 2016) in R (R core team, 2016) and calculate the half decay distance ( $LD_{50}$ ) and the distance where  $r^2$  first reaches 0.2. This

required a chromosome number, the estimate of Brasier *et al.* (1999) of  $n = 10 - 12$  was used to perform three separate runs.

LDHat was used to calculate LD statistics on the 28 contigs that were larger than 1 Mb (Auton and McVean *et al.*, 2007). Prior to running this analysis, the phased variants file was parsed to the format required for LDHat using VCFTools (Danecek *et al.*, 2011). The process of analysis with LDHat was as follows. Pairwise was run using the suggested value of  $\theta$ , maximum 4Ner of 100 and the number of points for the grid of 201. Following the generation of the lookup table, the default grid value for recombination was used, except when estimates were at the extreme of the grid, no sliding window analysis was performed, the full table was written, 4Ner was calculated by the moment method and the test for recombination was performed. Complete was then used to refine the lookup table, using the maximum 4Ner from pairwise, 201 points and the value of  $\theta$  from pairwise. Interval was then run to estimate the recombination rate with 10,000,000 iterations, a sample rate of 2,000 and a block penalty of 5. The interval results were then visualised using stat with a burn in of 50.

LDHot was then used to identify recombination hotspots with the output of LDHat with 1,000 simulations (Auton *et al.*, 2014). This was only possible on contig 8 due to the low number of SNPs on other contigs. The results of LDHat and LDHot were then plotted for each contig using the ggplot2 R package version 2.2.1 in R (Wickman, 2016).

### **5.3.7 Population genetics analyses with the Popgenome R package**

The Popgenome R package version 2.6.1 was used to calculate a variety of population genetics statistics for various selections of the analysed *P. fragariae* and *P. rubi* isolates (Pfeifer *et al.*, 2014). Prior to running these analyses, fasta alignments were created using a Python script (Sobczyk, unpublished). These files were used to calculate a variety of population genetics statistics. The four gamete test (Hudson and Kaplan, 1985) was calculated to assess evidence of historical recombination for both the population consisting of isolates of races UK1, UK2 and UK3 (hereafter termed the UK1-2-3 population) with the

population represented by the BC-23 and ONT-3 isolates and for all *P. fragariae* isolates with all *P. rubi* isolates. Pairwise  $F_{ST}$  per gene (Wright, 1943), Hudson's  $K_{ST}$  (Hudson *et al.*, 1992) and the average pairwise difference ( $D_{xy}$ ; Nei, 1972) were calculated per gene for the UK1-2-3 population compared to the population represented by the BC-23 and ONT-3 isolates. Nucleotide diversity ( $\pi$ ; Nei, 1987), Watterson's  $\theta$  (Watterson, 1975) and the frequency of segregating sites was calculated separately per gene for the UK1-2-3 population and the population represented by the BC-23 and ONT-3 isolates. The ratio of nucleotide diversity at synonymous and nonsynonymous sites ( $\pi_{ns}/\pi_s$ ; Nei, 1987) and Fu and Li's  $F^*$  and  $D^*$  (Fu and Li, 1993) were calculated just for the UK1-2-3 population. The McDonald-Kreitman test (MKT; McDonald and Kreitman, 1991) and Fay and Wu's  $H$  (Fay and Wu, 2000) were calculated with the UKR1-2-3 population and *P. rubi* isolates acting as an outgroup. The results of these analyses were plotted with the ggplot2 R package version 2.2.1 (Wickman, 2016) in R version 3.2.5 (R core team, 2016).

Genes showing high levels of population separation were identified using the Popgenome results. A high confidence gene showed a pairwise  $F_{ST}$  value greater than or equal to 0.5, a Hudson's  $K_{ST}$  value greater than or equal to 0.5 and a  $D_{xy}$  value greater than or equal to 0.05. If a gene met only one of these criteria, it was called a low confidence gene. Fisher's exact test (Fisher, 1922) contingency tables were created manually for effector classes and with a Python script for InterProScan annotations. Following this, Fisher's exact tests were conducted in R and the results were parsed and multi-test corrected using the Bonferroni and Benjamini-Hochberg methods in Python using the statsmodels module (Seabold and Perktold *et al.*, 2010).

Genes showing diversifying selection were identified using the Popgenome results. A gene was classed as showing diversifying selection if it showed a negative value for Fu and Li's  $F^*$  or  $D^*$ . Fisher's exact test contingency tables were created manually for effector classes and with a Python script for InterProScan annotations. Following this, Fisher's exact tests were conducted in R and the results were parsed and multi-test corrected using the Bonferroni and

Benjamini-Hochberg methods in Python using the statsmodels module (Seabold and Perktold *et al.*, 2010).

### **5.3.8 Preparation of RNA for sequencing**

#### **5.3.8.1 Testing of primers against additional isolates**

Prior to the preparation of RNA-Seq data, the Btub2\_F and Btub2\_R primers designed in Chapter 4 were tested for specificity against BC-1 and NOV-9. For this PCR, 2 µL of gDNA from BC-1, NOV-9, BC-16 and the 'Hapil' cultivar of *Fragaria × ananassa* were used as templates, with a water sample as a no template control in 20 µL reactions. The PCR reaction was performed as described in Chapter 4, section 4.3.5.

#### **5.3.8.2 Preparation of infection time course samples**

Micropropagated *F. × ananassa* plants of the 'Hapil' cultivar were inoculated using the 'plugs' method developed in Chapter 4, section 4.3.3. The BC-1, NOV-9 and A4 isolates were used for infections.

Time points were taken at: 24 hours post inoculation (hpi), 48 hpi, 72 hpi and 96 hpi. These were taken as described in Chapter 4, section 4.3.3.

#### **5.3.8.3 RT-PCR to assess which time points to sequence**

The presence of actively growing *P. fragariae* in the time point samples was assessed using the RT-PCR protocol described in Chapter 4, section 4.3.7.

The product of this PCR reaction was then analysed by electrophoresis as described in Chapter 4, section 4.3.7, except the gel was run at 80 V for 75 minutes rather than 90 minutes.

### **5.3.9 Initial assessment of RNA-Seq data, cleaning of raw data and aligning to assemblies**

Trimming and initial assessment of RNA-Seq reads was performed as described in Chapter 4, section 4.3.8. Alignments were additionally conducted as described in Chapter 4, section 4.3.9, however for this sequencing data reads were only aligned to the assemblies of the BC-1, BC-16 and NOV-9 isolates. Following this, analysis of expression and differential expression was carried out as described in Chapter 4, section 4.3.11.

### **5.3.10 Identifying candidate avirulence genes in the BC-1, BC-16 and NOV-9 isolates**

Candidate avirulence genes were identified through analysis of uniquely expressed genes and uniquely differentially expressed genes for each isolate. Uniquely expressed genes were identified as genes showing a Fragments Per Kilobase of transcript per Million mapped reads (FPKM) value greater than or equal to 5 in any *in planta* time point for only a single isolate. Uniquely differentially expressed genes were those with a minimum  $\log_2$  fold change (LFC) of 3 and a threshold *p*-value of 0.05 in any *in planta* time point for a single isolate. This analysis was performed separately using the assemblies of BC-1, BC-16 and NOV-9 as references. Following this, the results for all three analyses were combined to create lists of genes scored at different confidence levels. This was scored on a one - six scale, with the value increasing by one for each analysis the gene was scored as positive in. High confidence genes were those with a score of five or six, medium confidence genes had a score of three or four and low confidence genes had a score of one or two.

### **5.3.11 Confirmation of RNA-Seq results by RT-qPCR**

The expression levels of three effector genes were investigated in order to experimentally confirm the results from the RNA-Seq data. These effector genes were: g27513.t1 (an RxLR effector predicted to be a strong candidate avirulence gene in BC-16), g32018.t1 (an RxLR effector that was predicted to peak at the

24 hpi time point in BC-16) and g23965.t1 (an RxLR effector that was predicted to peak at the 48 hpi time point in BC-16). A shortlist of possible housekeeping genes was taken from Yan and Liou (2006) who analysed various housekeeping genes in the clade 1 species *Phytophthora parasitica*. The orthologous genes to those mentioned in the paper were identified by the BLASTn suite in Geneious R10 (Kearse *et al.*, 2012). The FPKM values of these genes were then investigated for stability in all sequenced samples, resulting in the selection of  $\beta$ -tubulin (g4288.t1) and WS41 (g28439.t1) for use as housekeeping genes. Primers for these five transcripts were designed using the modified primer3 version 2.3.7 implemented in Geneious R10 (Untergasser *et al.*, 2012). These primers were tested *in silico* for specificity to the BC-1, BC-16, NOV-9 and *F. vesca* v1.1 assemblies and only accepted if a single copy was amplified in the *P. fragariae* assemblies and did not amplify in the *F. vesca* assembly (Table 5.1).

Reverse transcription was performed on three biological replicates of the following RNA samples: BC-16 24 hpi, BC-16 48 hpi, BC-16 96 hpi, BC-16 *in vitro* mycelium, BC-1 24 hpi, BC-1 48 hpi, BC-1 72 hpi, BC-1 96 hpi, BC-1 *in vitro* mycelium, NOV-9 48 hpi, NOV-9 72 hpi, NOV-9 96 hpi, NOV-9 *in vitro* mycelium, A4 24hpi, A4 48hpi, A4 96 hpi and A4 *in vitro* mycelium. This was performed using the QuantiTect Reverse Transcription Kit (QIAGEN Inc., Venlo, The Netherlands) following the manufacturer's instructions. The melting temperature for the primers was then optimised on both gDNA and cDNA before performing Quantitative Polymerase Chain Reaction (qPCR).

qPCR was performed in a CFX96™ Real-Time PCR detection system running CFX manager version 1.3 (BioRad, Watford, U.K.). Reactions were run in 10  $\mu$ L volume, consisting of: 5  $\mu$ L of 2x qPCR BIO SyGreen Mix Lo-Rox (PCR Biosystems, London, U.K.), 2  $\mu$ L of a 1:3 dilution of the cDNA sample in dH<sub>2</sub>O and 400 nM of each primer. The qPCR reaction was run with the following conditions: 95°C for 3 minutes, 39 cycles of 95°C for 10 seconds, 62°C for 10 seconds and 72°C for 30 seconds. This was followed by 95°C for 10 seconds, and a 5 second step ranging from 65°C to 95°C by 0.5°C every cycle. At least two technical replicates for each sample were performed. Following completion of the reaction, the melt curve results were inspected and those which deviated

from a standard curve, consisting of a mix of all cDNA samples, were analysed by gel electrophoresis and repeated where necessary to ensure sufficient technical replication. Relative gene expression was calculated using the comparative cycle threshold ( $C_T$ ) method. Briefly, this is calculated as  $\Delta\Delta C_T = (C_{T \text{ target}} - C_{T \text{ reference}})_{\text{Time } x} - (C_{T \text{ target}} - C_{T \text{ reference}})_{\text{Time } 0}$  (Livak and Schmittgen, 2001). The target  $C_T$  value was calculated by taking the mean of all technical replicates. Where there was a difference of at least 1  $C_T$  value between the minimum and maximum results for technical replicates, further reactions were conducted. Outlier technical replicates were first identified as being outside 1.5 times the inter-quartile range. Following this, additional outliers were identified using the Grubb's test (Grubbs, 1950) and excluded from the analysis.  $C_{T \text{ reference}}$  was calculated as the geometric mean of the housekeeping genes. Expression values were calculated as the mean of three biological replicates and significance values were identified using one-way Analysis of Variances (ANOVA) and Tukey-Honestly Significant Different (Tukey-HSD) test in R. Graphs were plotted with the ggplot2 R package version 2.2.1 (Wickman, 2016) in R version 3.4.3 (R core team, 2017) running on a MacBook Pro, Early 2015, OS version 10.13.6.

**Table 5.1: Primer sequences used in the RT-qPCR experiment**

Species	Target	Gene Name	Gene ID	Primer Name	Primer Sequence	Length (bp)	Reference
<i>Phytophthora fragariae</i>	Housekeeping	WS41	g28439.t1	WS41_163F	5'-ATCGTGCTGTACCTGGGC-3'	156	This Study
				WS41_318R	5'-GATCTCGCTGGGCTTGAAGG-3'		
<i>Phytophthora fragariae</i>	Housekeeping	$\beta$ -tubulin	g4288.t1	Btub_44F	5'-CCGCGCCCGTACAGCAAC-3'	109	This Study
				Btub_152R	5'-TCGGAGATGACTTCCCAGAACTTG-3'		
<i>Phytophthora fragariae</i>	Gene of Interest	Candidate Avr	g27513.t1	cAvr_65F	5'-TGTCAAAGGCCGATCAGAGC-3'	180	This Study
				cAvr_244R	5'-CGAACAAACTATCCACACCAGC-3'		
<i>Phytophthora fragariae</i>	Gene of Interest	Early RxLR	g32018.t1	ERxLR_59F	5'-CTACCTCGACTACCAACGGC-3'	141	This Study
				ERxLR_199R	5'-TGATCTCCGCAGTGTCCACC-3'		
<i>Phytophthora fragariae</i>	Gene of Interest	Middle RxLR	g23965.t1	MRxLR_69F	5'-GGACCTCAGCCAAACCAAGC-3'	164	This Study
				MRxLR_187R	5'-CTCCTCGTCGTCTTGTCC-3'		
<i>Pseudomonas syringae</i> pv. <i>maculicola</i> 5422 (Arnold <i>et al.</i> , 1996)	Inter-plate calibrator	16S	N/A	U16SRT-F	5'-ACTCCTACGGGAGGCAGCAGT-3'	180	Clifford <i>et al.</i> , 2012
				U16SRT-R	5'-TATTACCGCGCTGCTGGC-3'		



### **5.3.12 Identification of putative promotor regions of a candidate avirulence gene**

#### **5.3.12.1 Building of a coexpression network**

In order to identify genes which displayed a similar expression pattern to a strong candidate avirulence gene in the BC-16 isolate, a coexpression analysis was performed using the weighted correlation network analysis (WGCNA) R package version 1.63 (Langfelder and Horvath, 2008, 2012) in R version 3.2.5 (R core team, 2016). Initially, input data was cleaned using the `goodSamplesGenes` function to remove genes with too many missing samples or zero variance. Further parameters were optimised, giving a soft threshold value of 22, a minimum module size of 30 and a merge threshold of 0.25.

#### **5.3.12.2 Identification of conserved regions upstream of genes of interest**

Following the identification of the coexpression module which contained the strong candidate avirulence gene in the BC-16 isolate, genes were manually inspected to produce four sets of genes for analysis: a set of high confidence genes, a set of low confidence genes, a set of highly expressed genes (showing an FPKM value greater than 9,000 in any BC-16 *in planta* sample) and the entirety of the module which contained the strong candidate avirulence gene in the BC-16 isolate prior to merging. High confidence genes showed a  $\log_2$  fold change of greater than one between the BC-16 24 hpi sample and the BC-16 mycelium, BC-1 48 hpi and NOV-9 72 hpi. The upstream regions were split into 100 bp sequences with `pyfasta` 0.5.2 (Pedersen, 2014) in Python. Non-target sequences were randomly selected from upstream regions of genes other than those in the set of interest. Discriminative Regular Expression Motif Elicitation (DREME), installed as part of the Multiple EM for Motif Elicitation (MEME) 4.11.2 suite, was then run with an e-value threshold of 0.5 and 10,000 regular expressions were examined (Bailey, 2011). This process was bootstrapped 100 times and any motifs identified in 90% of replicates were accepted.

### 5.3.13 Investigation of changes in *cis* of a candidate avirulence gene

#### 5.3.13.1 Assembly and repeat masking of Oxford Nanopore Technologies (ONT) reads

The UK3 isolate of *P. fragariae* NOV-9 was sequenced using Oxford Nanopore Technologies (ONT; see Chapter 2, section 2.8.3). The raw data was then base called with albacore version 2.2.7 (Oxford Nanopore Technologies, Oxford, UK) with the recursive option specified and 4,000 fastq reads per batch. Adapter sequences were then removed with Porechop version 0.2.0 (Wick, 2018). Coverage was then identified using a custom Perl script to count the number of nucleotides contained in sequencing reads and divide this value by the estimated genome size (Armitage, unpublished). Trimmed reads were then corrected using Canu version 1.4 with the correct flag enabled and the MinHash Alignment Process (MHAP) overlapper used, followed by Canu version 1.4 with the trim flag enabled and the MHAP overlapper used (Koren *et al.*, 2017). These trimmed reads were then assembled with SMARTdenovo version 1.0.0 (Ruan, 2018).

Following assembly, basic statistics were collected using Quast version 3.0 with the scaffolds option enabled (Gurevich *et al.*, 2013). A BUSCO (Benchmarking Universal Single-Copy Orthologs) analysis was also run using BUSCO version 3.0.1 with the eukaryota\_odb9 database on the assembly (Simão *et al.*, 2015).

Error correction was then first performed by aligning the ONT reads to the assembly using Minimap2 version 2.8-r711-dirty (Li, 2018a). This alignment was then used by Racon version 1.3.1 to correct sequencing errors (Vaser *et al.*, 2017). This process was run over ten iterations and the resulting assembly was filtered with a custom Python script to keep only contigs longer than 500 bp (Armitage, unpublished). Reads were then again mapped to the corrected assembly with Minimap2 version 2.8-r711-dirty (Li, 2018a). Nanopolish version 0.9.0 was then run on 50 kb fragments of the assembly, created with a Python script included in the Nanopolish install, to further correct sequencing errors with the Nanopolish variants command, with 100,000 max-haplotypes and a minimum candidate frequency of 0.2 (Simpson, 2018). These results were then merged

using a Python script included in the Nanopolish install. Finally, ten iterations of Pilon version 1.17 (Walker *et al.*, 2014) were used to correct further sequencing errors following alignment of the Illumina sequencing reads using Bowtie 2 version 2.2.6 (Langmead and Salzberg, 2012) and SAMtools version 1.5 (Li *et al.*, 2009).

After every stage of correction, basic statistics were collected using Quast version 3.0 with the scaffolds option enabled (Gurevich *et al.*, 2013). A BUSCO analysis was also run using BUSCO version 3.0.1 with the eukaryota\_odb9 database on the assembly (Simão *et al.*, 2015).

Masking of repetitive sequences was performed using RepeatMasker version open-4.0.5 (Smit *et al.*, 2013-2015) and RepeatModeler version 1.73 (Smit and Hubley, 2008-2015). These programs screen DNA sequences for interspersed repeats and low complexity DNA sequences and replaces them with Ns if hard masking was selected, or lowercase letters if soft masking was selected. Both hard and soft masked files were generated in this study. TransposonPSI release 22<sup>nd</sup> August 2010 (Haas, 2010) masked transposons using Position-Specific Iterative Basic Local Alignment Search Tool (PSI-BLAST) against a collection of transposon open reading frames (ORFs) to look for sequences with homology to known transposon sequences within the assembly and mask these with Ns. These programs also output a percentage of bases masked.

### **5.3.13.2 Identification of variant sites nearest to the candidate avirulence gene**

Following the assembly of the NOV-9 genome, the location of the orthologous sequence to the candidate avirulence gene identified in BC-16 was identified using the BLASTn suite implemented in Geneious R10 (Kearse *et al.*, 2012). The entire contigs were then extracted, the NOV-9 sequence was reversed due to differences in the assemblies and the contigs were trimmed to represent the same number of bases up and down stream of the gene. The upstream and downstream regions, plus the gene sequence, were then aligned using the

Multiple Alignment using Fast Fourier Transform (MAFFT) plugin in Geneious R10 (Katoh *et al.*, 2002; Katoh and Standley, 2013).

Variant sites were further investigated by aligning the BC-16 and NOV-9 Illumina reads to the ONT assembly with Bowtie 2 version 2.2.6 (Langmead and Salzberg, 2012). SAMtools version 1.5 was then used to create an index file (Li *et al.*, 2009) to allow for visualisation of the alignments in Integrative Genomics Viewer (IGV) (Thorvaldsdóttir, 2013).

#### **5.3.14 Investigation of changes in *trans* of a candidate avirulence gene**

A Hidden Markov Model (HMM) specific to transcription factors and transcriptional regulators in the stramenopiles was used to identify this class of genes in all the annotated genomes (Buitrago-Flórez *et al.*, 2014). Proteins scoring positive for potentially being either a transcription factor or transcriptional regulator by this model were identified with HMMER version 3.1b2 (<http://www.hmmer.org>, last accessed 03/01/2019).

These predicted transcription factors were then identified in Orthogroups (see Chapter 4, section 4.3.13) and assessed for any presence/absence variation. Orthogroups that contained only proteins from isolates of one race from the three of interest (UK1, UK2 and UK3) were called as unique groups and those where proteins from isolates of one race from the races of interest were present more than proteins from isolates of the other races of interest were called as expanded groups. These expanded and unique groups were manually investigated in Geneious R10 (Kearse *et al.*, 2012) to identify any putative transcription factors or transcriptional regulators that may have been causing the difference in expression levels of the candidate avirulence gene.

The genes nearest to variant sites, identified above, were investigated for the presence of positive hits from the HMM of transcription factors and transcriptional regulators.

Candidate calling was also performed using RNA-Seq data aligned to the BC-1, BC-16 and NOV-9 assemblies, as described in Section 5.3.10.

### **5.3.15 Investigation for a possible RxLR family**

The candidate avirulence gene in BC-16 was analysed against the rest of the predicted RxLRs in BC-16 by BLASTn from NCBI-BLAST version 2.2.3 (Altschul *et al.*, 1990) with an e-value threshold of 0.0000000001.

### **5.3.16 Assessment of synteny between the BC-16 and NOV-9 assemblies**

Satsuma version 3.0 was used to assess the levels of synteny between the long-read assemblies of BC-16 and NOV-9 (Grabherr *et al.*, 2010). These results were then visualised with Circos (Krzywinski *et al.*, 2009).

### **5.3.17 Availability of sequencing data and annotated assemblies**

All raw RNA sequencing reads and the ONT sequencing reads have been submitted to the Sequencing Read Archive (SRA) maintained by NCBI. These will be publicly available following the publication of these results in a peer reviewed journal (Table 5.2).

The NOV-9 assembly has been submitted to GenBank, maintained by NCBI as part of bioproject PRJNA488213 and will be publicly available following the publication of these results in a peer reviewed journal.

**Table 5.2: Details of the NCBI SRA codes for all sequencing reads uploaded for future release**

Speices	Isolate	Sequenced Molecule	Sample Details	SRA code
<i>Phytophthora fragariae</i>	BC-1	mRNA	48 hpi	SRR7764612
<i>Phytophthora fragariae</i>	BC-1	mRNA	<i>in vitro</i> mycelium	SRR7764615
<i>Phytophthora fragariae</i>	NOV-9	mRNA	72 hpi	SRR7764607
<i>Phytophthora fragariae</i>	NOV-9	mRNA	<i>in vitro</i> mycelium	SRR7764613
<i>Phytophthora fragariae</i>	NOV-9	gDNA	ONT reads	SRR7668093

Hours post inoculation (hpi); messenger Ribonucleic acid (mRNA); genomic deoxyribonucleic nucleic acid (gDNA); Oxford Nanopore Technologies (ONT).

## 5.4 Results

### 5.4.1 Alignment of all Illumina reads to the reference assembly of the BC-16 isolate of *Phytophthora fragariae* showed similar mapping rates for isolates of the same species

In order to investigate whether the race structure of *P. fragariae* was controlled by sequence level variation, the Illumina reads of all eleven sequenced isolates of *P. fragariae* and all three sequenced isolates of *P. rubi* were aligned to the reference assembly of the BC-16 isolate of *P. fragariae* described in Chapter 3 (Table 5.3). These results showed, in most cases, similar mapping rates for each species. The majority of *P. fragariae* isolates showed a mapping rate of above 90% and the majority of *P. rubi* isolates showed a mapping rate of approximately 75%. However, several isolates: BC-16, ONT-3 and SCRP245 all showed a reduction in mapping rates.

**Table 5.3: Mapping rates of DNA-Seq reads of all eleven sequenced isolates of *Phytophthora fragariae* and all three sequenced isolates of *Phytophthora rubi*, to the reference assembly of the BC-16 isolate of *P. fragariae***

Species	Isolate	Number of Paired Reads	Aligned Concordantly			Aligned Discordantly	Overall Alignment Rate
			0 times	1 time	> 1 time	1 time	
<i>Phytophthora fragariae</i>	A4	6,483,874	456,980 (7.05%)	3,204,351 (49.42%)	2,822,543 (43.53%)	1,680 (0.37%)	92.98%
<i>Phytophthora fragariae</i>	BC-1	21,412,244	1,804,325 (8.43%)	10,798,557 (50.43%)	8,809,362 (41.14%)	48,065 (2.66%)	91.80%
<i>Phytophthora fragariae</i>	BC-16	12,922,513	1,816,171 (14.05%)	6,086,410 (47.10%)	5,019,932 (38.85%)	17,404 (0.96%)	86.08%
<i>Phytophthora fragariae</i>	BC-23	9,034,592	940,674 (10.41%)	4,143,287 (45.86%)	3,950,631 (43.73%)	12,653 (1.35%)	89.73%
<i>Phytophthora fragariae</i>	NOV-5	7,088,293	515,248 (7.27%)	3,492,259 (49.27%)	3,080,786 (43.46%)	2,090 (0.41%)	92.76%
<i>Phytophthora fragariae</i>	NOV-9	19,084,039	1,411,478 (7.40%)	9,849,908 (51.61%)	7,822,653 (40.99%)	128,442 (9.10%)	93.28%
<i>Phytophthora fragariae</i>	NOV-27	10,053,852	726,988 (7.23%)	4,992,810 (49.66%)	4,334,054 (43.11%)	5,849 (0.80%)	92.83%
<i>Phytophthora fragariae</i>	NOV-71	14,432,530	1,168,460 (8.10%)	6,976,473 (48.34%)	6,287,597 (43.57%)	30,095 (2.58%)	92.11%
<i>Phytophthora fragariae</i>	NOV-77	8,174,990	608,271 (7.44%)	3,967,431 (48.53%)	3,599,288 (44.03%)	4,650 (0.76%)	92.62%
<i>Phytophthora fragariae</i>	ONT-3	8,002,167	1,139,950 (14.25%)	3,738,478 (46.72%)	3,123,739 (39.04%)	10,078 (0.88%)	85.88%
<i>Phytophthora fragariae</i>	SCR245	9,337,139	2,175,245 (23.30%)	3,733,421 (39.98%)	3,428,473 (36.72%)	3,816 (0.18%)	76.74%
<i>Phytophthora rubi</i>	SCR249	9,191,017	2,189,667 (23.82%)	3,574,372 (38.89%)	3,426,978 (37.29%)	16,899 (0.77%)	76.36%
<i>Phytophthora rubi</i>	SCR324	9,089,817	2,319,060 (25.51%)	3,432,777 (37.77%)	3,337,980 (36.72%)	22,219 (0.96%)	74.73%
<i>Phytophthora rubi</i>	SCR333	9,046,540	2,226,184 (24.61%)	3,636,246 (40.19%)	3,184,110 (35.20%)	21,992 (0.99%)	75.63%



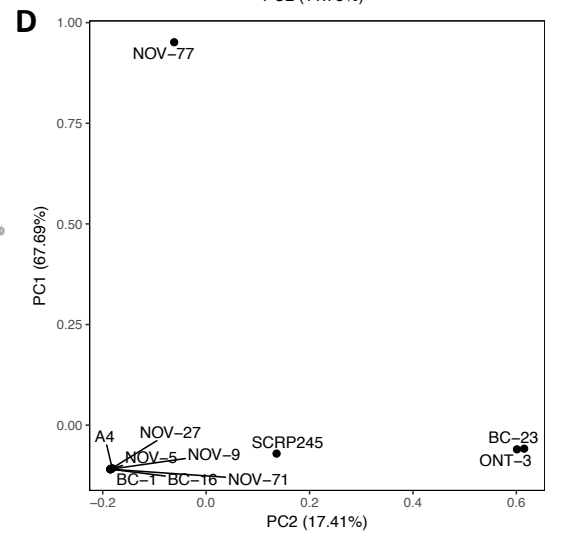
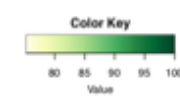
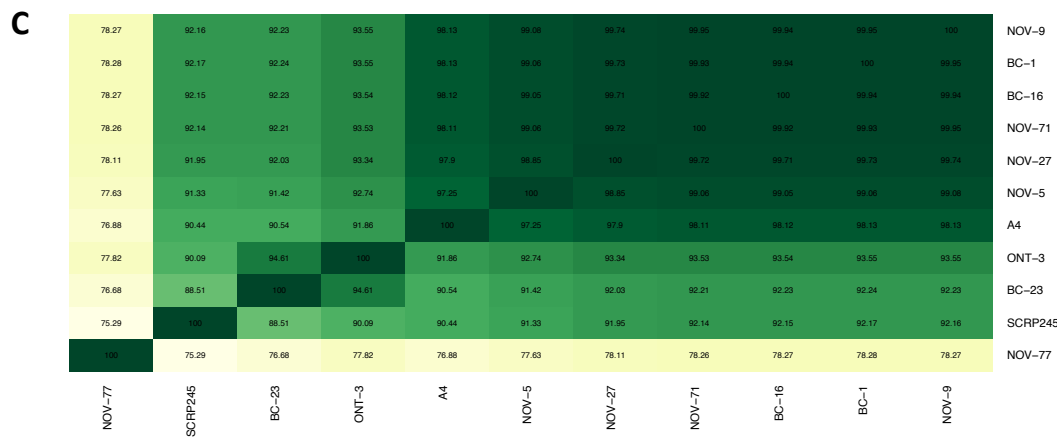
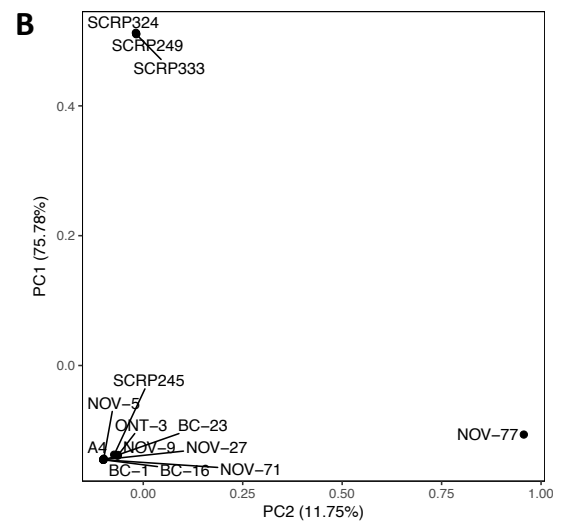
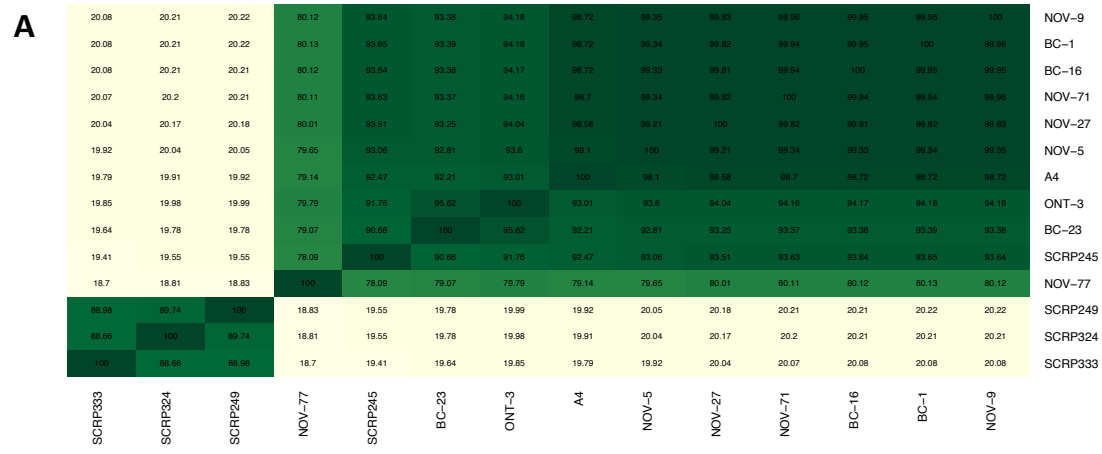
#### **5.4.2 Identification of variant sites allowed the construction of a phylogenetic tree of *Phytophthora fragariae* and demonstrated species separation between *P. fragariae* and *Phytophthora rubi***

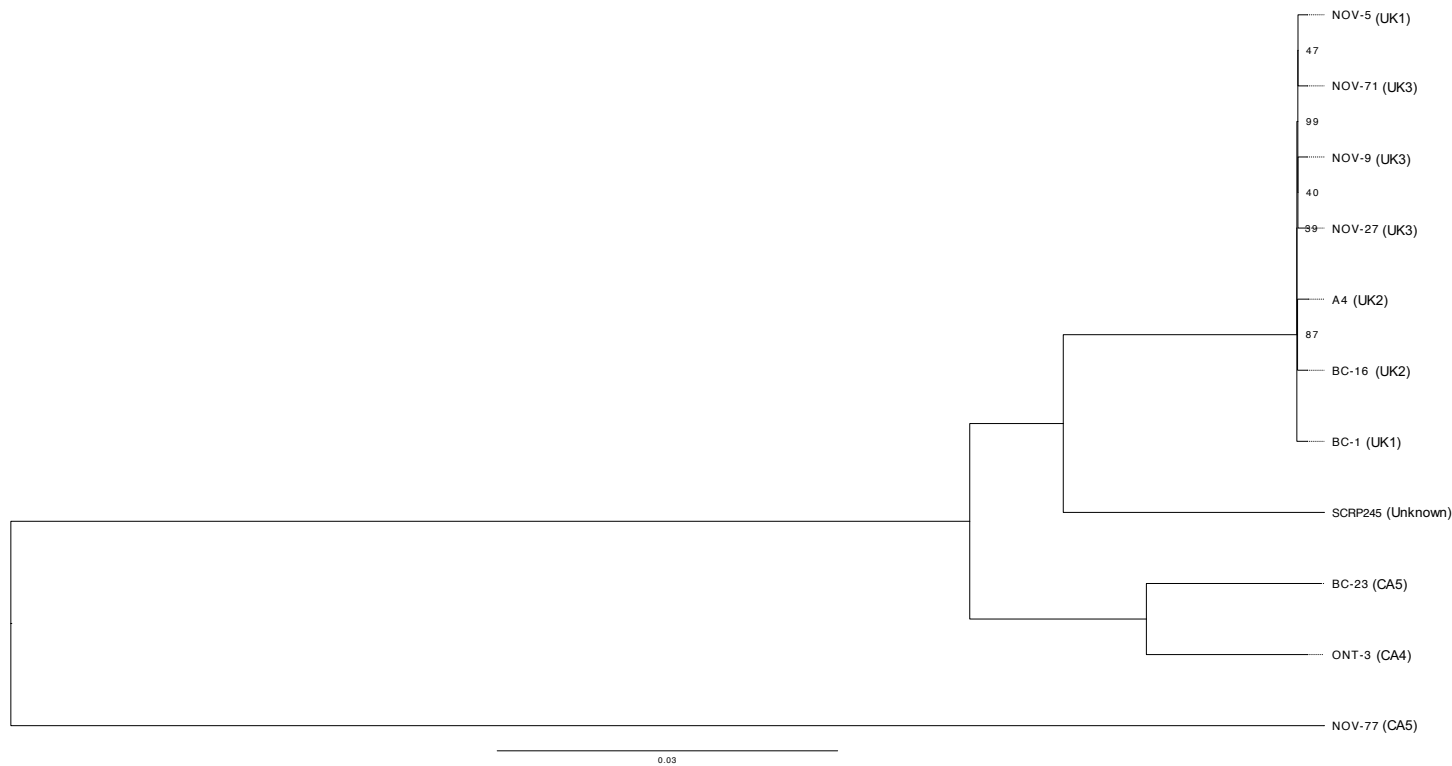
A total of 725,444 SNP sites were identified across all fourteen genomes, along with 95,478 insertions and deletions. Following filtering to retain only high-quality, biallelic SNPs for downstream analyses, 402,939 SNP sites were retained. These statistics were also assessed when the isolates of *P. rubi* were removed, with 545,365 SNP sites retained after filtering. This potentially still included monomorphic sites, following removal of these sites 401,009 SNPs were retained across all isolates and 138,672 SNPs were retained in only *P. fragariae* isolates. The percentage of shared biallelic SNP sites was then calculated both with and without *P. rubi* isolates and a heatmap of pairwise comparisons between each isolate was plotted (Fig. 5.1A and 5.1C). A PCA was also conducted both with and without *P. rubi* isolates (Fig. 5.1B and 5.1D). The plots with *P. rubi* isolates clearly showed a strong separation between the two species, as principal component one explained the majority of the variance (75.78%) between the isolates and only appeared to separate by species. The analysis also suggested that NOV-77 formed a separate group from the other *P. fragariae* isolates, as principal component two explained a small amount of the remaining variance (11.75%) and only appeared to separate NOV-77 from the other *P. fragariae* isolates. Interestingly, it appeared that the *P. rubi* isolates clustered away from NOV-77 along principal component two, though it was unclear what this might have been representing due to a small sample size for *P. rubi*.

There still appeared to be some variance within the remaining *P. fragariae* isolates and so the analysis was conducted again after removing the *P. rubi* isolates. Again NOV-77 grouped separately, this time along principal component one and explained the majority of the variance (67.69%) between the isolates. It was also observed that principal component two explained a small amount of the remaining variance (17.41%), with BC-23 and ONT-3 grouping at one extreme of this principal component, with SCR245 at an intermediate position and the remaining *P. fragariae* isolates clustered at the other extreme. These patterns can also be observed in the pairwise similarity heatmaps (Fig. 5.1A and 5.1C).

Using this same data, a tree was constructed of all the *P. fragariae* isolates, this again showed the seven isolates of races UK1, UK2 and UK3 grouping together with high similarity, BC-23 and ONT-3 forming a separate group with SCRP245 as an intermediate and NOV-77 at an extreme end of the variation (Fig. 5.2).

**Fig. 5.1 (overleaf): Pairwise comparison heatmaps and principal component analyses of percentage similarity of shared biallelic SNP sites for all sequenced isolates of *Phytophthora fragariae* and *Phytophthora rubi*.** Variant sites were identified by aligning Illumina reads of all the sequenced isolates to the reference assembly of the BC-16 isolate of *P. fragariae* (described in Chapter 3) with Bowtie 2 (Langmead and Salzberg, 2012) and analysis with the Genome Analysis Toolkit (GATK) haplotypcaller (McKenna *et al.*, 2010). Sites were filtered with VCFtools (Danecek *et al.*, 2011) and VCFlib (Garrison, 2012) to leave only high quality, biallelic SNP sites. Heatmaps were plotted using the gplots R package version 3.0.1 (Warnes *et al.*, 2016) and principal component analyses were plotted using the ggplot2 R package version 2.2.1 (Wickman, 2016) in R (R core team, 2016). **A:** Heatmap showing pairwise comparisons of the percentage of all high quality biallelic SNP sites with the same allele for eleven *P. fragariae* isolates and three *P. rubi* isolates. **B:** Principal component analysis of all high quality biallelic SNP sites for eleven *P. fragariae* and three *P. rubi* isolates. **C:** Heatmap showing pairwise comparisons of the percentage of all high quality biallelic SNP sites with the same allele for eleven *P. fragariae* isolates. **D:** Principal component analysis of all high quality biallelic SNP sites for eleven *P. fragariae* isolates.





**Fig. 5.2: Neighbour joining tree of all sequenced *Phytophthora fragariae* isolates based on high quality, biallelic SNP sites.**

Variant sites were identified by aligning Illumina reads of all the sequenced isolates to the reference assembly of the BC-16 isolate of *P. fragariae* (described in Chapter 3) with Bowtie 2 version 2.2.6 (Langmead and Salzberg, 2012) and analysis with the Genome Analysis Toolkit (GATK) haplotypcaller (McKenna et al., 2010). Sites were filtered with VCFtools (Danecek et al., 2011) and VCFlib (Garrison, 2012) to leave only high quality, biallelic SNP sites. Fasta alignments were produced with a Perl script (Bergey, 2012) and haploidised with seqtk (Li, 2018b) and a consensus tree was created from 100 bootstrap replicates with the ape R package version 5.1 (Paradis *et al.*, 2004) in R (R core team, 2016). Node labels represent the number of bootstrap replicates supporting this node, only values less than 100 are shown.

### 5.4.3 Annotation of SNP sites for their effects in the BC-16 isolate of *Phytophthora fragariae* showed less than half of SNP sites were in exonic regions

Following the identification of variant sites, the high quality, biallelic SNP sites were annotated for their effects in the BC-16 genome (Table 5.4). These results showed a greater proportion of intergenic variants than exonic variants. Of these exonic variants, more sites were silent than missense, with very few nonsense variants. A small proportion of these variants were intronic with small numbers in splice sites specifically, these variants could change the protein sequence formed from a transcript.

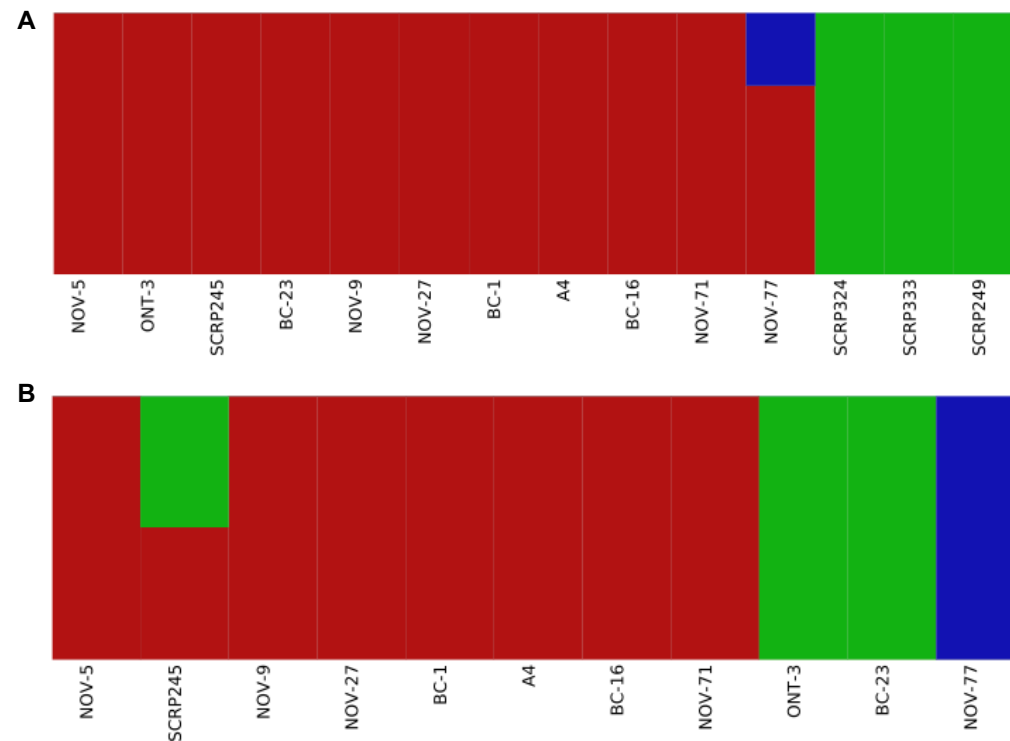
**Table 5.4: The classification of the effects of variant sites in the BC-16 genome assembly**

Classification of SNP	Number of sites	Percentage of sites
Total	401,009	100.000%
Exonic	177,968	44.074%
Missense	85,659	21.361%
Nonsense	1,168	0.291%
Silent	91,588	22.839%
Intergenic	204,222	50.576%
Intronic	18,890	4.678%
Splice Site Acceptor	100	0.025%
Splice Site Donor	103	0.026%
Splice Site Region	2,508	0.621%

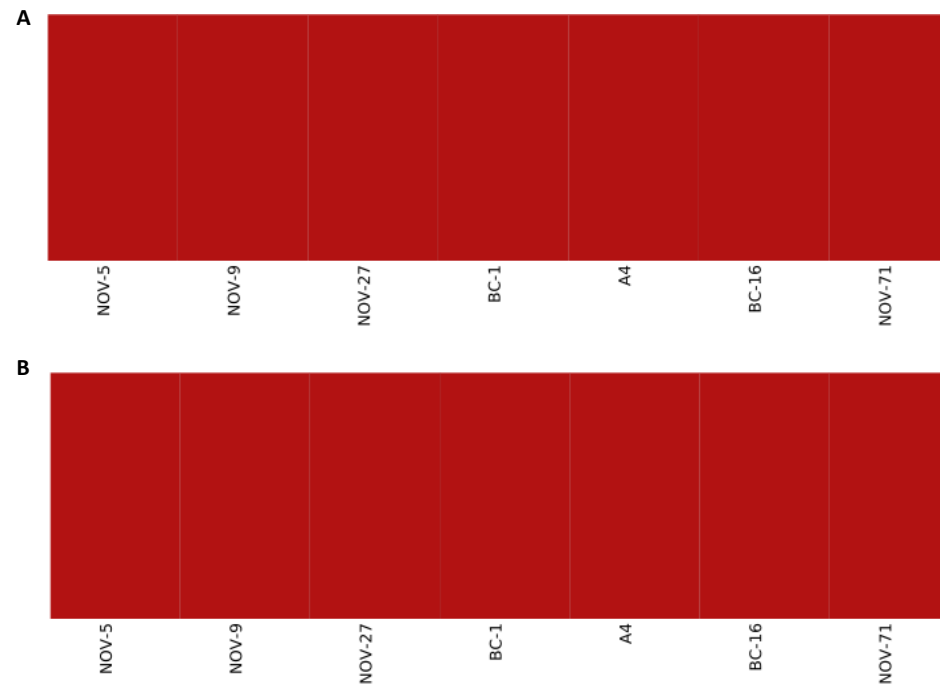
### 5.4.4 Investigation of a population structure within the sequenced isolates of *Phytophthora fragariae* and *Phytophthora rubi* showed all isolates of pathogenicity races UK1, UK2 and UK3 of *P. fragariae* grouped as a single population

Following the results of the SNP analysis, it appeared as though there may have been separation between the isolates of *P. fragariae* and *P. rubi* which were sequenced in this study. In order to investigate this, *distruct* (Rosenberg, 2004) plots were created, displaying a predicted population structure (Fig. 5.3 and 5.4).

The analysis of all *P. rubi* and *P. fragariae* isolates showed all three *P. rubi* isolates formed their own population, with all the *P. fragariae* isolates grouped into a separate population, with a small amount of evidence that NOV-77 may have been part of a separate population. When the analysis was conducted without the *P. rubi* isolates: NOV-77 formed a distinct population, BC-23 and ONT-3 formed another population and all the isolates of pathogenicity races UK1, UK2 and UK3 formed a further distinct population. Interestingly, SCRP245 appeared to either be part of an intermediate population without enough evidence to identify it as a separate population, or it may have been part of a hybrid population. Following this, analysis was conducted on just isolates of pathogenicity races UK1, UK2 and UK3 to detect any population structure that was not identified due to the variation between the other isolates preventing it from being detected. However, this did not show any population structure either with or without the logistic prior option, which is described as being able to detect more subtle variation than the default algorithm. This resulted in an identification of the UK1-2-3 population as a population of interest for further analysis.



**Fig. 5.3: Distruct plots of fastSTRUCTURE results conducted on high quality, biallelic SNP sites.** Variant sites were identified by aligning Illumina reads of all the sequenced isolates to the reference assembly of the BC-16 isolate of *Phytophthora fragariae* (described in Chapter 3) with Bowtie 2 (Langmead and Salzberg, 2012) and analysis with the Genome Analysis Toolkit (GATK) haplotypcaller (McKenna *et al.*, 2010). Sites were filtered with VCFtools (Danecek *et al.*, 2011) and VCFlib (Garrison, 2012) to leave only high quality, biallelic SNP sites. **A:** Distruct plot of fastSTRUCTURE (Raj *et al.*, 2014) results carried out on all sequenced isolates of *Phytophthora rubi* and *P. fragariae*. Each colour represents a different population. **B:** Distruct plot of fastSTRUCTURE (Raj *et al.*, 2014) results carried out on all sequenced isolates of *P. fragariae*. Each colour represents a different population.



**Fig. 5.4: Distruct plots of fastSTRUCTURE results conducted on high quality, biallelic SNP sites.** Variant sites were identified by aligning Illumina reads of all the sequenced isolates to the reference assembly of the BC-16 isolate of *Phytophthora fragariae* (described in Chapter 3) with Bowtie 2 (Langmead and Salzberg, 2012) and analysis with the Genome Analysis Toolkit (GATK) haplotypcaller (McKenna *et al.*, 2010). Sites were filtered with VCFtools (Danecek *et al.*, 2011) and VCFlib (Garrison, 2012) to leave only high quality, biallelic SNP sites. **A:** Distruct plot of fastSTRUCTURE (Raj *et al.*, 2014) results carried out on all sequenced isolates of *P. fragariae* of pathogenicity races UK1, UK2 and UK3. Each colour represents a different population. **B:** Distruct plot of fastSTRUCTURE (Raj *et al.*, 2014), with the logistic prior option enabled, results carried out on all sequenced isolates of *P. fragariae* of pathogenicity races UK1, UK2 and UK3. Each colour represents a different population.



#### **5.4.5 Investigation of polarising and private variant sites failed to explain the observed differences in pathogenicity race in the UKR1-2-3 population**

Following the identification of the UKR1-2-3 population as a population of interest, all variant sites within the UKR1-2-3 population were filtered and variant sites that were present in one race but not the other two in this population, termed a private variant, were identified. An investigation was also performed to identify any private variants that were also the same allele as in the *P. rubi* isolates, termed polarising variants. Private variant sites were only found in isolates of race UK2 (Table 5.5). Three variant sites were found to be within the coding sequence of genes, two of which were silent and lead to no change in the protein sequence. However, variant site 11 was shown to cause a change from a leucine to a proline in g27056. This gene encoded a putative GTPase with a PDZ domain, though it was not predicted as being secreted and so was not likely to be an avirulence gene. The expression of gene models nearest to the intergenic variant sites were checked for evidence of expression in BC-16. Only four gene models showed evidence of expression in at least one sequenced sample of the BC-16 isolate: g65, g66, g33391 and g33392. These genes were near variants 2 and 12, however, none of them showed evidence of secretion and so were not likely to be avirulence genes.

**Table 5.5: Details of all sites identified as private variants in the UKR1-2-3 population and their effects in BC-16**

Variant ID	Race	Allele in	Variant type		Position	Region type	Upstream		Downstream	
		<i>Phytophthora rubi</i>	Reference to Alternate	Contig			Nearest gene	Distance	Nearest gene	Distance
1	UK2	Identical	G to A SNP	14	964,434	Intergenic	g11265	978 bp	g11266	3,099 bp
2	UK2	Distinct	A to C SNP	1	217,370	Intergenic	g65	654 bp	g66	2,013 bp
3	UK2	Distinct	T to C SNP	2	1,219,725	Intergenic	g1731	7,004 bp	g1732	3,790 bp
4	UK2	Distinct	T to A SNP	13	673,883	CDS - Silent	g10482	Within CDS	g10482	Within CDS
5	UK2	Distinct	G to A SNP	14	964,434	Intergenic	g11265	978 bp	g11266	3,099 bp
6	UK2	Distinct	C to A SNP	14	964,446	Intergenic	g11265	990 bp	g11266	3,087 bp
7	UK2	Distinct	A to G SNP	14	1,308,196	Intergenic	g11448	739 bp	g11449	1,375 bp
8	UK2	Distinct	T to C SNP	19	961,636	CDS - Silent	g14158	Within CDS	g14158	Within CDS
9	UK2	Distinct	Multi-Allele InDel*	40	631,727	Intergenic	g23431	39 bp	g23432	3,247 bp
10	UK2	Distinct	T to C SNP	48	312,824	Intergenic	g25960	38 bp	g25961	1,997 bp
11	UK2	Distinct	A to G SNP	51	571,287	CDS - L to P substitution	g27056	Within CDS	g27056	Within CDS
12	UK2	Distinct	T to C SNP	76	248,619	Intergenic	g33391	4,412 bp	g33392	979 bp

\*There are four alternate alleles for this variant. The reference allele is AC. The alternate alleles are: A, ACC, AGC and AGCC.

#### **5.4.6 Structural variant site identification failed to explain the observed differences in pathogenicity race in the UKR1-2-3 population**

Following the identification of no private SNPs or small INDELS which could explain the race differences in the UKR1-2-3 population, larger structural variants were identified. A total of 80,388 INDELS were identified in the sequenced isolates, alongside 7,020 structural variants. Following the identification of these additional variants, sites were analysed for any that were private to a particular race of the UKR1-2-3 population. This analysis returned no private INDELS or structural variants for any race.

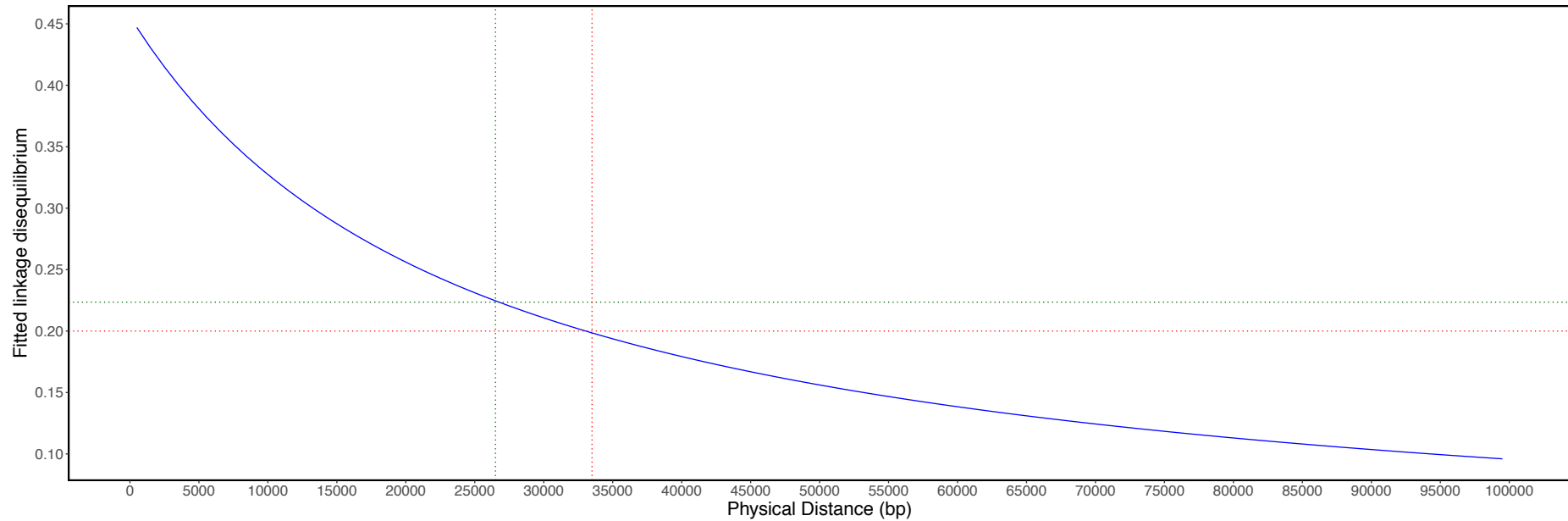
#### **5.4.7 Population Genetic Analyses**

##### **5.4.7.1 Investigation of linkage disequilibrium in the UKR1-2-3 population suggested a mixed reproductive strategy with clonal reproduction and occasional sexual reproduction**

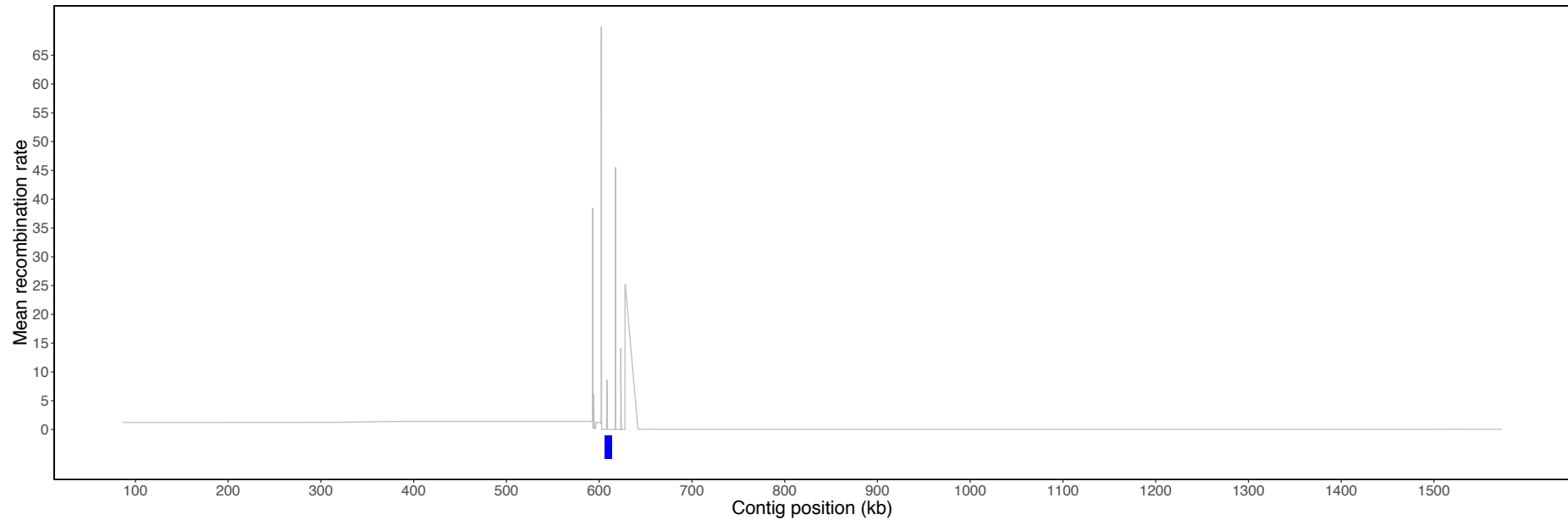
The previously identified high quality, biallelic SNP sites were used to assess linkage disequilibrium (LD) in the UKR1-2-3 population. Phased sites were used to calculate the squared Pearson coefficient of correlation  $r^2$  value. The decay of the  $r^2$  value with distance was additionally calculated by fitting the  $r^2$  values to the Hill and Weir decay function (Hill and Weir, 1988). The adjusted  $r^2$  values were then plotted by distance (Fig. 5.5). This revealed a half decay distance ( $LD_{50}$ ) of 26,500 bp for each possible chromosome number, as estimated by Brasier *et al.* (1999). This was similar to the value for *Schizosaccharomyces pombe* (Jeffares *et al.*, 2015) and so suggested a mixed reproductive strategy with clonal reproduction and occasional sexual reproduction (Nieuwenhuis and James, 2016).

Analysis was also performed on the twenty-eight contigs of the BC-16 reference assembly that were larger than 1Mb to produce estimates of the recombination rates along the contig. Contig 8 was the only contig with sufficient SNPs to test for recombination hotspots (Fig. 5.6). This showed the presence of a

recombination hotspot between position 605,923 and 613,302. This region contained eight predicted gene models (Table 5.6). None of these genes showed evidence of secretion and so were unlikely to be avirulence genes.



**Fig. 5.5: Assessment of linkage disequilibrium decay by distance in the BC-16 isolate of *Phytophthora fragariae*.** VCFtools (Danecek *et al.*, 2011) was used to calculate the  $r^2$  value of pairs of high quality, biallelic SNPs. These values were fitted to the Hill and Weir decay function (Hill and Weir, 1988) and plotted with the ggplot2 R package version 2.2.1 (Wickman, 2016) in R version 3.2.5 (R core team, 2016). The green dashed line displays the half decay distance (LD<sub>50</sub>). The red dashed line displays the first distance where the  $r^2$  value falls to 0.2.



**Fig. 5.6: Assessment of the recombination rate along contig 8 and depiction of a recombination hotspot.** Recombination rate was predicted for high quality, biallelic SNPs with LDHat (Auton and McVean *et al.*, 2007). Recombination hotspots were identified using the results from LDHat with LDHot (Auton *et al.*, 2014). These results were plotted with the ggplot2 R package version 2.2.1 (Wickman, 2016) in R version 3.2.5 (R core team, 2016). The blue rectangle below the plot represents the location of a recombination hotspot.

**Table 5.6: Details of gene models contained in the recombination hotspot on BC-16 contig 8**

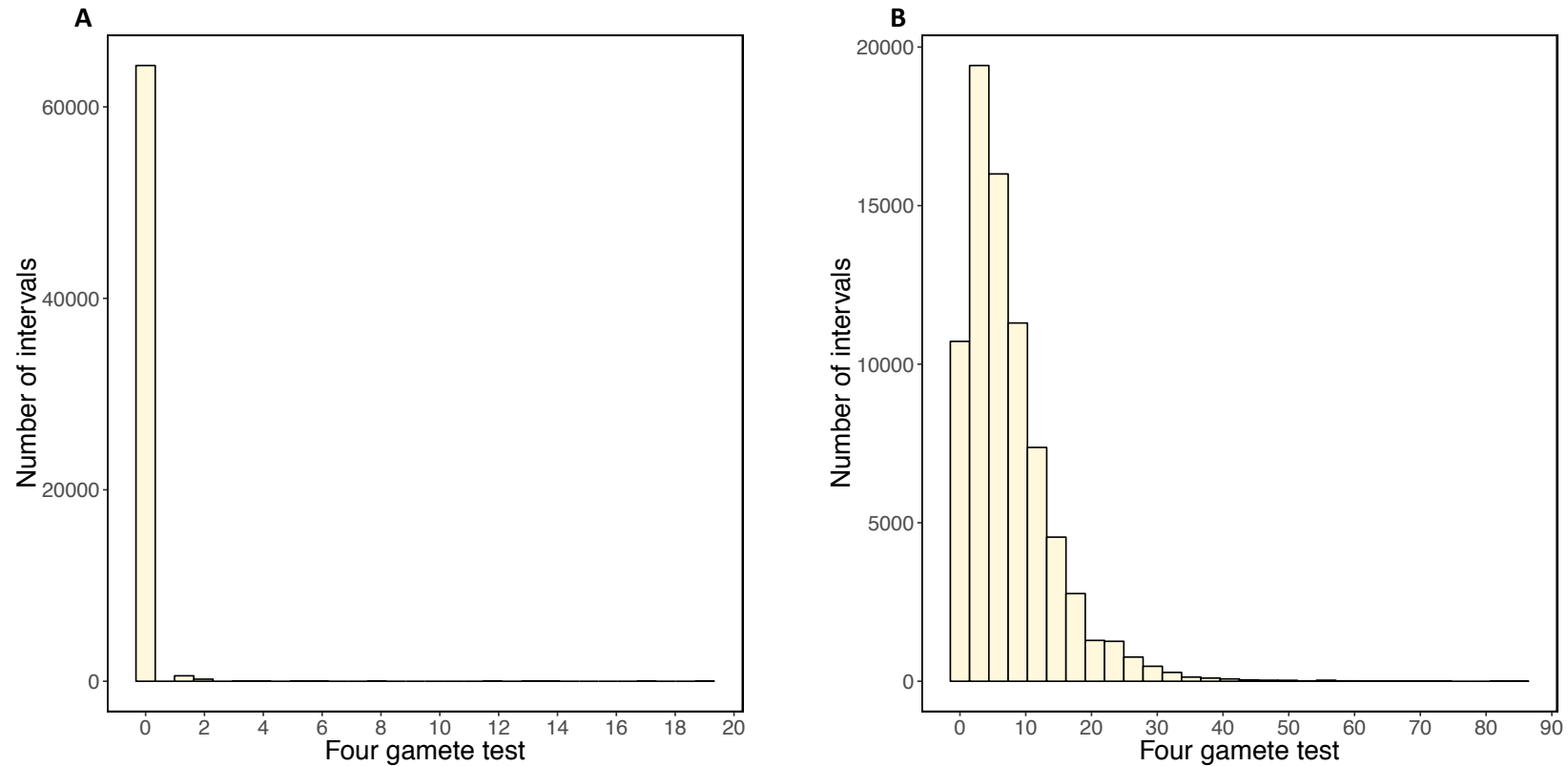
Gene ID	Orthogroup	Secretion Evidence	Functional Annotations	Expression Evidence
g6948	OG0000653	TMHMM		None
g6949	OG0031032	TMHMM		None
g6950	OG0000653			None
g6951	OG0001824	TMHMM		None
g6952	OG0000653			None
g6953	OG0002502	TMHMM		None
g6954	OG0000658		P-loop containing nucleoside triphosphate hydrolase, ABC transporter-like	None
g6955	OG0018628	TMHMM	ABC-2 type transporter	FPKM of 5 in BC-16 96 hpi

TMHMM is described in Krogh *et al.* (2001).

#### **5.4.7.2 Investigation of historical recombination within and between species showed greater levels of historical recombination between species than within species**

A four gamete test (Hudson and Kaplan, 1985) was performed using high quality, biallelic SNPs. Comparisons were performed between *P. fragariae* and *P. rubi* as well as between the UK1-2-3 population of *P. fragariae* isolates and the population represented by the BC-23 and ONT-3 isolates of *P. fragariae*. For the analysis conducted assessing evidence of historical recombination between the UK1-2-3 population of isolates and the BC-23 and ONT-3 isolates, in 26.53% of sliding windows, the statistic could not be calculated. For the windows where the statistic was calculated, 1.29% of these sliding windows showed non-zero results and hence some levels of historical recombination. This statistic peaked at 19 for 0.0077% of sliding windows (Fig. 5.7A). Interestingly, an analysis of high quality, biallelic SNPs in all sequenced isolates of *P. rubi* and *P. fragariae* showed much higher values, this suggested higher levels of historical recombination. This was likely due there being 505,622 fewer SNP sites used in the analysis of the UK1-2-3 population and the BC-23 and ONT-3 isolates, compared to the investigation of *P. rubi* and *P. fragariae* isolates. For the investigation of *P. fragariae* and *P. rubi* isolates, in 13.61% of sliding windows, the statistic could not be calculated. For the windows where the statistic was calculated, 91.82% of these sliding windows showed non-zero results and hence some levels of historical recombination. This statistic peaked at 85 for 0.0026% of sliding windows (Fig. 5.7B).





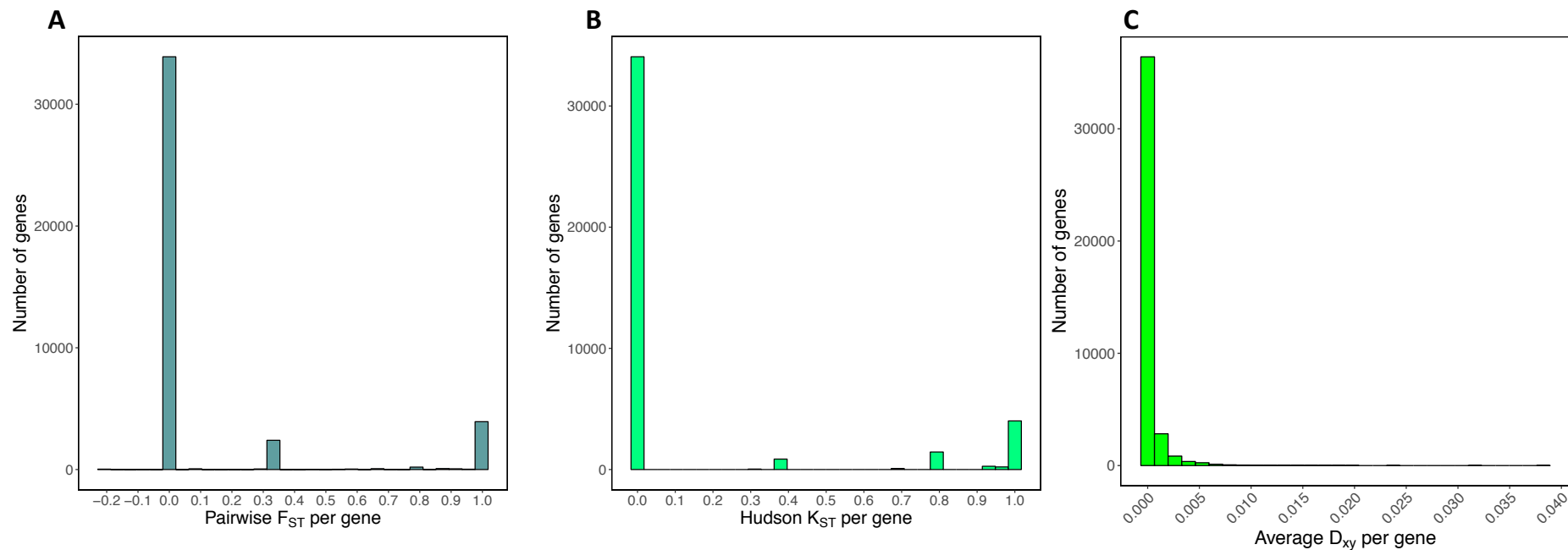
**Fig. 5.7: Results of a four gamete test conducted on different populations of *Phytophthora fragariae* and *Phytophthora rubi* isolates.** The Popgenome R package version 2.6.1 (Pfeifer *et al.*, 2014) was used to conduct a four gamete test (Hudson and Kaplan, 1985) for 10 kbp sliding windows with an interval size of 1 kbp. Figures were created using the ggplot2 R package version 2.2.1 (Wickman, 2016) in R version 3.2.5 (R core team, 2016). **A**: Plot of the results of a four gamete test conducted on the isolates of *P. fragariae* belonging to the UKR1-2-3 population and the population represented by the BC-23 and ONT-3 isolates of *P. fragariae*. **B**: Plot of the results of a four gamete test conducted on all sequenced *P. fragariae* and *P. rubi* isolates.

#### **5.4.7.3 Investigation of regions showing high levels of population separation showed variation across the genome with apoplastic effectors and a single functional term significantly overrepresented in genes which showed population separation**

As the population structure analysis results showed, seven sequenced isolates represented the UK1-2-3 population, and two isolates (BC-23 and ONT-3) represented a separate population. It was therefore of interest to investigate whether particular regions of the genome would explain the separation between these populations. This was investigated by calculating the pairwise  $F_{ST}$  (Wright, 1943), Hudson's  $K_{ST}$  (Hudson *et al.*, 1992) and average pairwise difference ( $D_{xy}$ ; Nei, 1972) statistics. The results of these calculations showed that for pairwise  $F_{ST}$  and Hudson's  $K_{ST}$ , the majority of genes clustered at a value of zero, this indicated that they did not show separation between the two populations. However, a sizeable number of genes also showed higher values, specifically the next highest peak was at one, which indicated complete separation by population structure. The average  $D_{xy}$  results also showed a peak of genes at zero, but additionally showed a number of genes with non-zero values. These were much lower than the values for  $F_{ST}$  and Hudson's  $K_{ST}$ , likely because this statistic is an absolute measure of genetic distance and takes no account of intrapopulation variation (Fig. 5.8).

Following the calculation of these statistics, the genes which showed the highest levels of population separation were extracted and analysed further. A total of 49 genes were identified as separated by population structure by all measures and were termed as high confidence and 4,554 genes were identified as separated by population structure in at least one measure and were termed low confidence. These produced 49 and 4,583 transcripts respectively. An investigation of effector gene and secreted protein enrichment was performed using Fisher's Exact test (Fisher, 1922) where the proportion of this gene class in the separated set was greater than the proportion in the genome. This was the case for apoplastic effectors and secreted proteins in both high confidence and low confidence sets, and crinkler effectors in the low confidence set only. The

Fisher's exact test showed significant ( $p < 0.05$ ) enrichment of apoplastic effectors in both sets, however, following multiple test correction, significance was only observed in the high confidence set. A similar process was also performed for InterProScan annotations. Only one term, IPR000612 (Proteolipid membrane potential modulator), was significant after multitest correction in the high confidence set and no terms were significant after multitest correction in the low confidence set. As apoplastic effectors are a fairly broad descriptor and potentially have a high false positive rate, it was unlikely that these genes were actually driving the population separation, though it could not be ruled out. The mechanism for how the significant InterProScan term could drive population separation was also not clear (Table 5.7 and A.1).



**Fig. 5.8: Results of pairwise  $F_{ST}$ , Hudson's  $K_{ST}$  and average  $D_{xy}$  analysis comparing the UK1-2-3 *Phytophthora fragariae* population to the population consisting of the isolates BC-23 and ONT-3 of *P. fragariae*.** The Popgenome R package version 2.6.1 (Pfeifer *et al.*, 2014) was used to calculate the values of pairwise  $F_{ST}$  (Wright, 1943), Hudson's  $K_{ST}$  (Hudson *et al.*, 1992) and average  $D_{xy}$  (Nei, 1972) for all genes in the annotations of the BC-16 isolate of *P. fragariae*. Figures were created using the ggplot2 R package version 2.2.1 (Wickman, 2016) in R version 3.2.5 (R core team, 2016). **A:** Pairwise  $F_{ST}$  plotted per gene. **B:** Hudson's  $K_{ST}$  plotted per gene. **C:** Average  $D_{xy}$  plotted per gene.

**Table 5.7: Investigation of the enrichment of classes of genes showing evidence of population separation between the *Phytophthora fragariae* UKR1-2-3 population and the population of the isolates BC-23 and ONT-3**

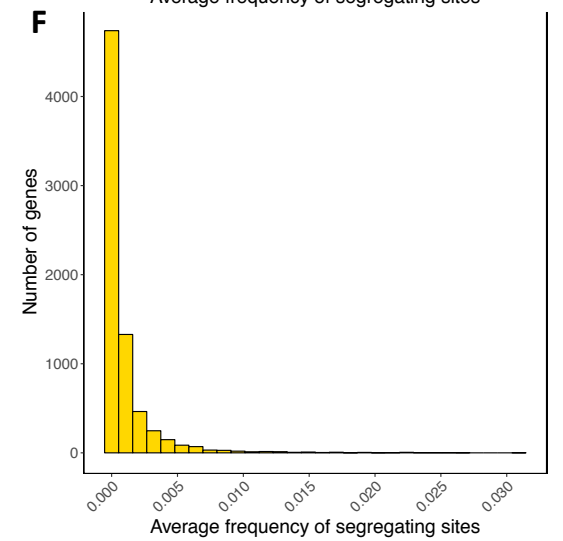
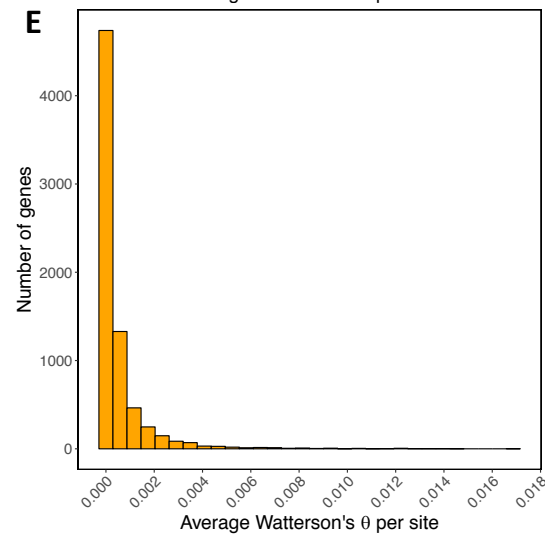
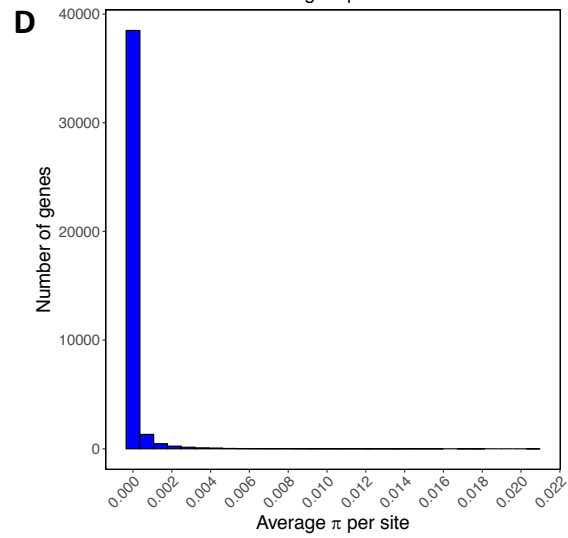
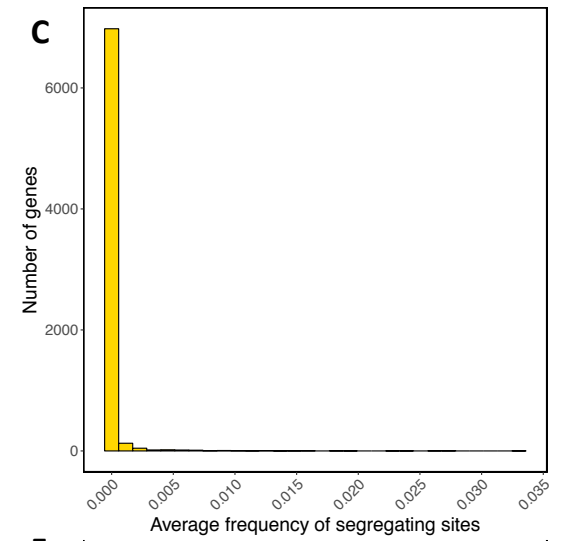
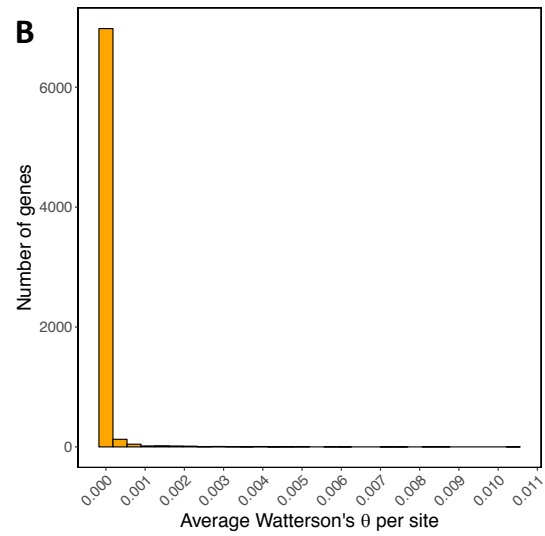
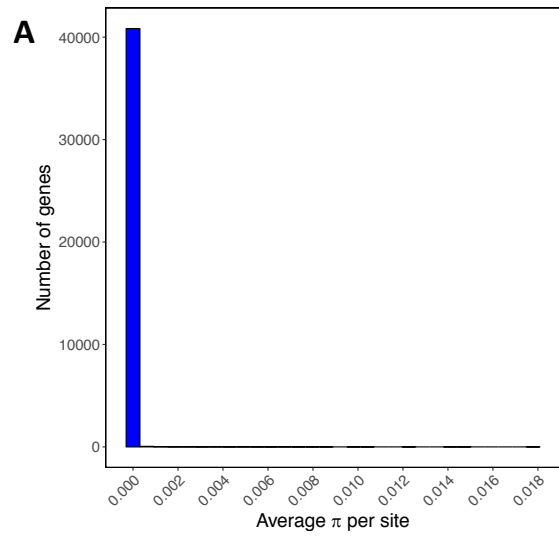
Transcript Classification	Confidence Level	Genomic Total	Separated Total	<i>p</i> -value adjusted			InterProScan Description
				<i>p</i> -value	Benjamini-Hochberg	Bonferroni	
<b>Apoplactic Effectors</b>	<b>High</b>	<b>4,864</b>	<b>11</b>	<b>0.0246</b>	<b>0.0491</b>	<b>0.0491</b>	N/A
Secreted Proteins	High	8,024	12	0.229	0.229	0.458	N/A
Crinkler Effectors	Low	88	11	0.383	0.383	1.00	N/A
Apoplactic Effectors	Low	4,864	575	0.0411	0.123	0.123	N/A
Secreted Proteins	Low	8,024	906	0.249	0.374	0.374	N/A
<b>IPR000612</b>	<b>High</b>	<b>2</b>	<b>1</b>	<b>0.00237</b>	<b>0.0331</b>	<b>0.0497</b>	<b>Proteolipid membrane potential modulator</b>

Genes with a *p*-value > 0.05 were accepted as statistically significant, these entries are displayed in bold. Only statistically significant InterProScan (Jones *et al.*, 2014) annotations are shown, the remainder are in Table A.1.

#### **5.4.7.4 Investigation of levels of variation in the UKR1-2-3 population and the population represented by the BC-23 and ONT-3 isolates showed lower variation in the UKR1-2-3 population despite the presence of a similar number of segregating sites in both populations**

In order to assess the levels of variation within the population, three statistics were calculated: nucleotide diversity ( $\pi$ ) per gene (Nei, 1987), Watterson's  $\theta$  per gene (Watterson, 1975) and the frequency of segregating sites per gene. These measures were calculated for both the UKR1-2-3 population of isolates of *P. fragariae* and the population represented by the BC-23 and ONT-3 isolates of *P. fragariae*. The  $\pi$  value peaked at 0.018 in the UKR1-2-3 population, compared to a value of 0.021 in the population represented by BC-23 and ONT-3. Despite these values being similar, plotting the results showed a larger tail of the distribution in the BC-23 and ONT-3 population. This suggested that although both populations showed low diversity, this population showed more diversity than the UKR1-2-3 population (Fig. 5.9A and 5.9D). Watterson's  $\theta$  is a similar measure to  $\pi$ , however it does not assess allelic frequencies as  $\pi$  does. Within the UKR1-2-3 population, Watterson's  $\theta$  peaked at 0.010 compared to a peak value of 0.017 in the population represented by BC-23 and ONT-3. Plotting the results again showed a larger tail of the distribution in the BC-23 and ONT-3 population. These results also suggested that although low levels of diversity were detected, there was more diversity in the BC-23 and ONT-3 population compared to the UKR1-2-3 population (Fig. 5.9B and 5.9E). Additionally, the average frequency of segregating sites was calculated for both populations. For the UKR1-2-3 population, this statistic peaked at 0.033 and for the BC-23 and ONT-3 population it peaked at 0.031. Although these values were similar, plotting of the results showed a larger tail of the distribution in the BC-23 and ONT-3 population compared to the UKR1-2-3 population. This, along with the results for Watterson's  $\theta$  suggested that the differences in levels of variation between these populations was influenced by differing numbers of segregating sites present in the populations, rather than being solely due to allelic differences (Fig. 5.9C and 5.9F).

**Fig. 5.9 (overleaf): Measures of variation in the UKR1-2-3 population and the BC-23 and ONT-3 population of *Phytophthora fragariae*.** The Popgenome R package version 2.6.1 (Pfeifer *et al.*, 2014) was used to calculate the values of average nucleotide diversity ( $\pi$ ) per site (Nei, 1987), average Watterson's  $\theta$  per site (Watterson, 1975) and average frequency of segregating sites per gene. These values were plotted for each gene in the annotations of the BC-16 isolate of *P. fragariae*. Figures were created using the ggplot2 R package version 2.2.1 (Wickman, 2016) in R version 3.2.5 (R core team, 2016). **A:** Average nucleotide diversity ( $\pi$ ) per site in the UKR1-2-3 population of *P. fragariae*. **B:** Average Watterson's  $\theta$  per site in the UKR1-2-3 population of *P. fragariae*. **C:** Average frequency of segregating sites per gene in the UKR1-2-3 population of *P. fragariae*. **D:** Average nucleotide diversity ( $\pi$ ) per site in the population represented by the BC-23 and ONT-3 isolates of *P. fragariae*. **E:** Average Watterson's  $\theta$  per site in the population represented by the BC-23 and ONT-3 isolates of *P. fragariae*. **F:** Average frequency of segregating sites per gene in the population represented by the BC-23 and ONT-3 isolates of *P. fragariae*.





#### **5.4.7.5 Investigation of levels and types of selection in the UKR1-2-3 population showed little selection compared to an outgroup and the action of diversifying and purifying selection within the UKR1-2-3 population**

The levels and direction of selection within the UKR1-2-3 population of *P. fragariae* were investigated through the use of five statistics: the McDonald-Kreitman test (MKT; McDonald and Kreitman, 1991), Fay and Wu's H (Fay and Wu, 2000), the ratio of nucleotide diversity ( $\pi$ ) at nonsynonymous sites to  $\pi$  at synonymous sites ( $\pi_{ns}/\pi_s$ ; Nei, 1987) and Fu and Li's  $F^*$  and  $D^*$  (Fu and Li, 1993). Both the MKT and Fay and Wu's H assessed selection through analysis with an outgroup, in this investigation the *P. rubi* isolates SCRP249, SCRP324 and SCRP333 were used. The MKT neutrality index showed the vast majority of genes scored at one, with a few genes showing a positive value, but less than one. This suggested that the vast majority of predicted genes fitted the neutrality null hypothesis and showed no evidence of selection away from the outgroup. However, for a minority of genes, there was some evidence of diversifying selection away from the outgroup (Fig. 5.10A). The peak of the Fay and Wu's H result was clustered around zero, though slightly shifted to the positive side, with a few genes at larger positive values (Fig. 5.10B). This suggested that for these genes there was evidence of low levels of purifying selection away from the outgroup. This could also be explained by a recent selective sweep or genetic bottleneck reducing the number of derived SNPs in the UKR1-2-3 population. There were also a few genes which showed negative values, this indicated diversifying selection away from the outgroup (Fig. 5.10B). Interestingly, the number of genes showing true neutral selection appeared to differ between the MKT and Fay and Wu's H analyses. This difference may be explained by the number of genes for which the statistic could be calculated: 18,355 in the MKT (of which 1,488 had a neutrality index  $< 1$ ) compared with 604 genes for the Fay and Wu's H analysis.

Fu and Li's  $F^*$  and  $D^*$  were calculated solely using the isolates from the UK1-2-3, population. These statistics showed similar results, with a peak at around -1.5 and a second peak between 0.5 and 0.75. However,  $F^*$  showed a small number

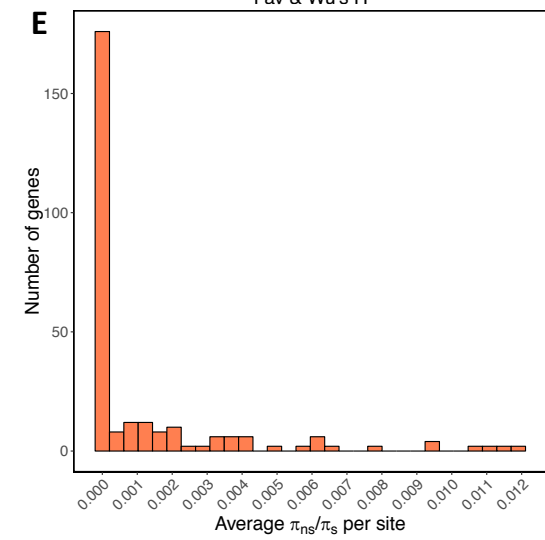
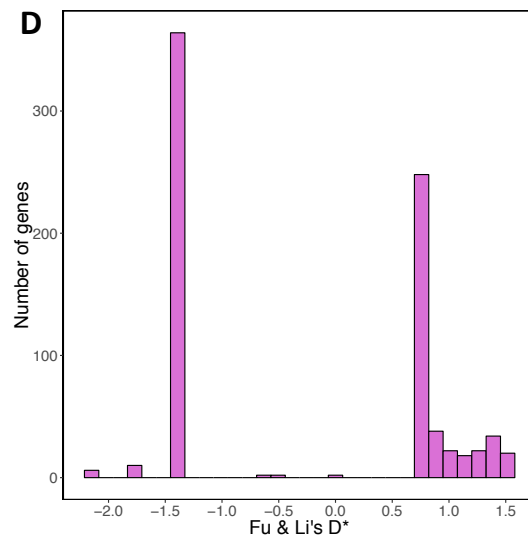
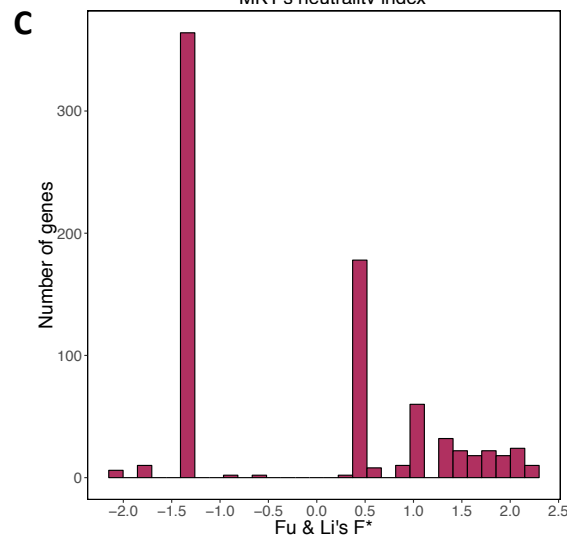
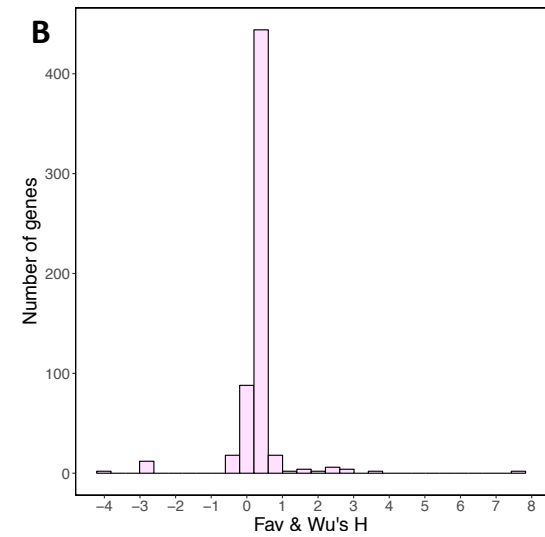
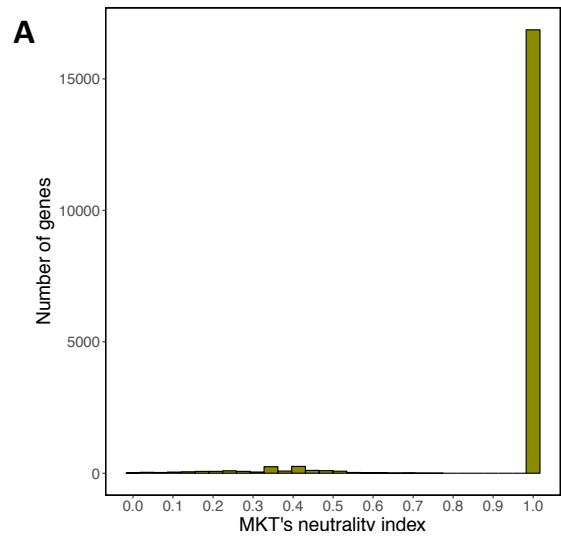
of genes at values between 1.5 and 2.5, which  $D^*$  did not show. Although the peak at -1.5 was larger than the positive peak in both tests, a greater number of genes showed positive values than negative values. Therefore, a large number of genes showed a lack of singleton variants and hence a positive value. This potentially indicated a recent selective sweep or genetic bottleneck in the population. However, there were a number of genes that showed negative values for these statistics. This indicated an excess of singleton variants. This potentially indicated the beginnings of differentiation following a recovery from a recent genetic bottleneck (Fig. 5.10C and 5.10D). This again showed a very small number of genes undergoing truly neutral selection, when compared to the results of the MKT analysis. Similar to the Fay and Wu's  $H$  results, this was likely explained by the smaller number of genes for which the statistic could be calculated: 788 genes for the Fu and Li's analyses, compared to 18,355 for the MKT analysis.

Selection was also investigated through the calculation of the ratio of the nucleotide diversity at nonsynonymous sites ( $\pi_{ns}$ ) and the nucleotide diversity at synonymous sites ( $\pi_s$ ). This statistic could only be calculated for 137 genes (0.33%), limiting the power of this test. The majority of these genes had values clustering at zero and no genes scored above 0.012. This suggested that the nucleotide diversity at synonymous sites was far greater than that at nonsynonymous sites; thus, suggesting that the majority of detected variation was neutral, as synonymous sites will not cause changes to the translated protein sequence (Fig. 5.10E).

Following the calculation of these statistics, the genes which showed diversifying selection from Fu and Li's  $F^*$  and  $D^*$  were extracted and analysed further. A total of 192 genes were identified as showing diversifying selection. These produced 193 transcripts. An investigation of effector gene and secreted protein enrichment was performed using Fisher's Exact test (Fisher, 1922), where the proportion of this gene class in the separated set was greater than the proportion in the genome. This was the case for all effector classes, however, none were

significant ( $p < 0.05$ ). A similar process was also performed for InterProScan annotations. No terms were significant after multitest correction (Table A.2 - A.3).

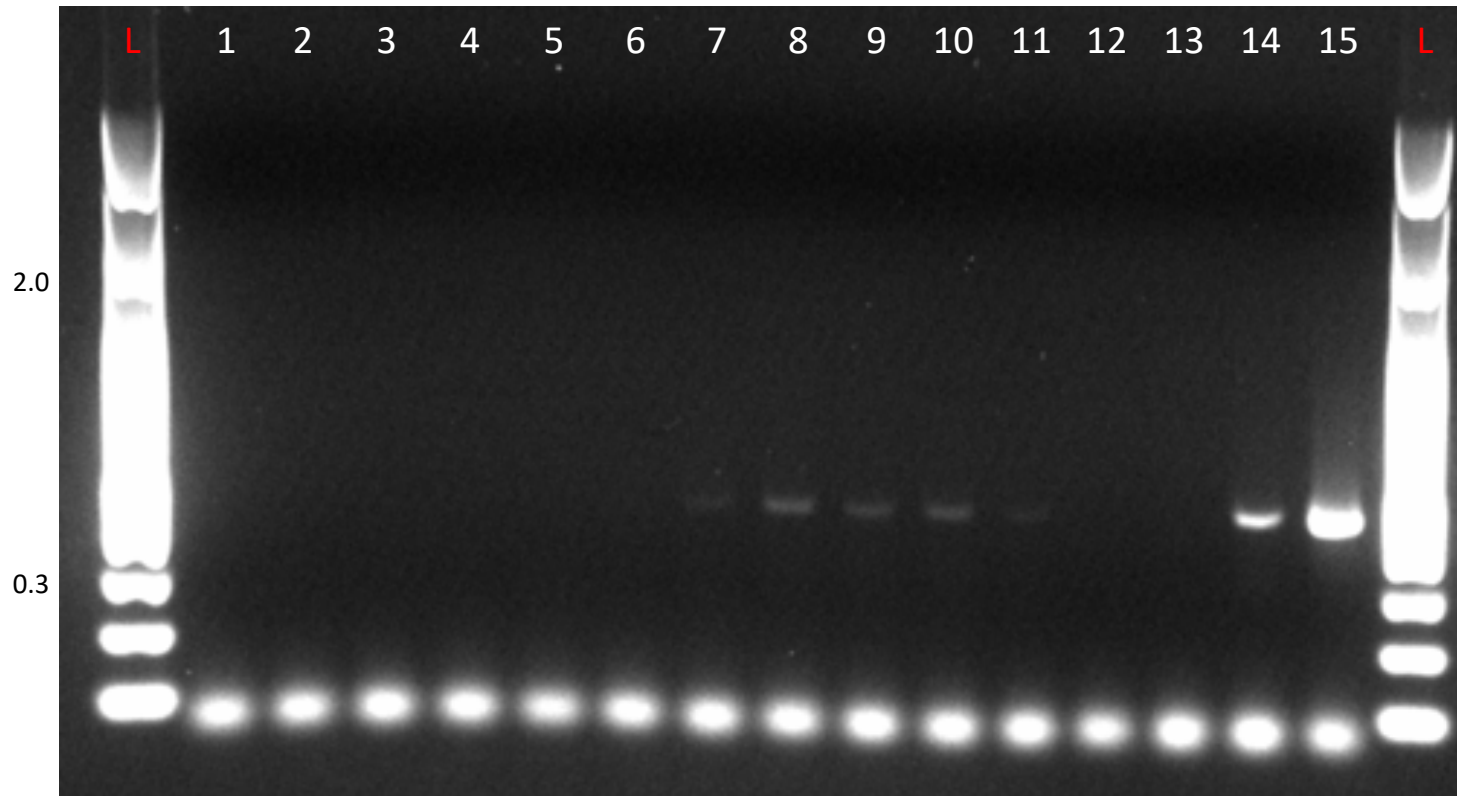
**Fig. 5.10 (overleaf): Assessment of selection levels and types in the UKR1-2-3 population of *Phytophthora fragariae*.** The Popgenome R package version 2.6.1 (Pfeifer *et al.*, 2014) was used to calculate the values of the McDonald-Kreitman test (MKT; McDonald and Kreitman, 1991), Fay and Wu's H (Fay and Wu, 2000), the ratio of nucleotide diversity ( $\pi$ ) at nonsynonymous sites to  $\pi$  at synonymous sites ( $\pi_{ns}/\pi_s$ ; Nei, 1987) and Fu and Li's  $F^*$  and  $D^*$  (Fu and Li, 1993). These values were plotted for each gene in the annotations of the BC-16 isolate of *P. fragariae*. Figures were created using the ggplot2 R package version 2.2.1 (Wickman, 2016) in R version 3.2.5 (R core team, 2016). **A:** The MKT neutrality index per gene. **B:** Fay and Wu's H per gene. **C:** Fu and Li's  $F^*$  per gene. **D:** Fu and Li's  $D^*$  per gene. **E:** Average  $\pi_{ns}/\pi_s$  per site, plotted for each gene.



#### **5.4.8 Sequencing of mRNA from additional *Phytophthora fragariae* isolates, BC-1 and NOV-9, for additional investigations into candidate avirulence genes**

Due to the lack of identified candidate avirulence genes for UK1 or UK3 and to further explore candidate genes identified for UK2, RNA-Seq was performed on the BC-1 (race UK1) and NOV-9 (race UK3) isolates. Prior to performing an infection time course with these isolates, the specificity of the  $\beta$ -tubulin primers were tested for specificity on BC-1 and NOV-9 gDNA. This analysis showed these primers functioned correctly in these isolates (Fig. A.1). Following a time course experiment, similar to that described in Chapter 4 for BC-16, RNA samples from each time point were analysed by RT-PCR for the presence of transcripts of  $\beta$ -tubulin. This showed that BC-1 was first detectable at 48 hpi and NOV-9 was first detectable at 72 hpi (Fig. 5.11). Therefore, these time points were sequenced, alongside mycelial RNA for each isolate. Prior to sending samples for external sequencing, several quality metrics were assessed to test for any risks of contamination, assessing the concentration of the RNA and assessing the levels of degradation of the RNA (Table 5.8).

Prior to use of the sequencing data for downstream analysis, the raw reads were assessed to ensure they were of high enough quality for downstream analysis. The sequencing reads were first trimmed to remove sequencing adaptors and poor-quality data. Similar to the results for the BC-16 reads described in Chapter 4, several metrics performed poorly. However, this was likely due to the nature of RNA-Seq compared to DNA-Seq. Mycelial reads also performed better than reads from *in planta* time points, similar to the results for BC-16 described in Chapter 4 (Table 5.9).



**Fig. 5.11: Agarose gel electrophoresis of the RT-PCR products of  $\beta$ -tubulin primers on samples from a time course of BC-1 and NOV-9 infection on *Fragaria*  $\times$  *ananassa* plants of the 'Hapil' cultivar. L: 100 bp plus ladder from New England Biolabs, sizes listed in kb. PCR templates: **1**: Uninoculated plant cDNA. **2**: NOV-9 12 hours post inoculation (hpi) cDNA. **3**: BC-1 12 hpi cDNA. **4**: NOV-9 24 hpi cDNA. **5**: BC-1 24 hpi cDNA. **6**: NOV-9 48 hpi cDNA. **7**: BC-1 48 hpi cDNA. **8**: NOV-9 72 hpi cDNA. **9**: BC-1 72 hpi cDNA. **10**: NOV-9 96 hpi cDNA. **11**: BC-1 96 hpi cDNA. **12**: dH<sub>2</sub>O. **13**: gDNA from a *Fragaria*  $\times$  *ananassa* cultivar 'Hapil' plant. **14**: gDNA from BC-16 mycelium. **15**: cDNA from BC-16 mycelium.**

**Table 5.8: Quality control results on extracted RNA prior to sequencing**

Sample ID	<i>Phytophthora fragariae</i>	<i>Fragaria</i> ×	Time Point (hours post inoculation)	Concentration			RIN <sup>e</sup>
	Isolate	<i>ananassa</i> cultivar		260/280	260/230	(ng/μL)	
TA_BC1_P1	BC-1	Hapil	48	2.06	1.69	35.4	7.2
TA_BC1_P2	BC-1	Hapil	48	1.94	2.09	55.4	7.7
TA_BC1_P3	BC-1	Hapil	48	1.80	1.71	20.6	6.6
TA_NOV9_P1	NOV-9	Hapil	72	2.09	2.13	55.2	8.4
TA_NOV9_P2	NOV-9	Hapil	72	2.07	2.23	88.0	8.1
TA_NOV9_P3	NOV-9	Hapil	72	1.98	2.20	47.2	8.7
TA_BC1_M1	BC-1	N/A	<i>in vitro</i> mycelium	2.24	0.15	33.0	9.5
TA_BC1_M2	BC-1	N/A	<i>in vitro</i> mycelium	2.43	0.11	20.4	9.3
TA_BC1_M3	BC-1	N/A	<i>in vitro</i> mycelium	2.12	0.69	27.5	9.0
TA_NOV9_M1	NOV-9	N/A	<i>in vitro</i> mycelium	2.18	0.94	113.2	9.7
TA_NOV9_M2	NOV-9	N/A	<i>in vitro</i> mycelium	2.27	0.25	53.2	9.8
TA_NOV9_M5	NOV-9	N/A	<i>in vitro</i> mycelium	2.15	1.26	25.2	8.4

**Table 5.9: Fastqc results from analysis of trimmed RNA-Seq data**

Library Name	Basic Statistics	Per Base Sequence Quality	Per Sequence Quality Scores	Per Base Sequence Content	Per Base GC Content	Per Sequence GC Content	Per Base N Content	Sequence length Distribution	Sequence Duplication Levels	Overrepresented Sequences	K-mer Content
TA_BC1_P1_F	Pass	Pass	Pass	Fail	Warning	Warning	Pass	Warning	Fail	Pass	Warning
TA_BC1_P1_R	Pass	Pass	Pass	Fail	Warning	Warning	Pass	Warning	Fail	Pass	Warning
TA_BC1_P2_F	Pass	Pass	Pass	Fail	Warning	Fail	Pass	Warning	Fail	Pass	Warning
TA_BC1_P2_R	Pass	Pass	Pass	Fail	Warning	Fail	Pass	Warning	Fail	Pass	Warning
TA_BC1_P3_F	Pass	Pass	Pass	Fail	Warning	Warning	Pass	Warning	Fail	Pass	Warning
TA_BC1_P3_R	Pass	Pass	Pass	Fail	Warning	Warning	Pass	Warning	Fail	Pass	Warning
TA_NOV9_P1_F	Pass	Pass	Pass	Fail	Warning	Warning	Pass	Warning	Fail	Pass	Warning
TA_NOV9_P1_R	Pass	Pass	Pass	Fail	Warning	Warning	Pass	Warning	Fail	Pass	Warning
TA_NOV9_P2_F	Pass	Pass	Pass	Fail	Warning	Warning	Pass	Warning	Fail	Pass	Warning
TA_NOV9_P2_R	Pass	Pass	Pass	Fail	Warning	Warning	Pass	Warning	Fail	Pass	Warning
TA_NOV9_P3_F	Pass	Pass	Pass	Fail	Warning	Warning	Pass	Warning	Fail	Pass	Warning
TA_NOV9_P3_R	Pass	Pass	Pass	Fail	Warning	Warning	Pass	Warning	Fail	Pass	Warning
TA_BC1_M1_F	Pass	Pass	Pass	Warning	Pass	Pass	Pass	Warning	Fail	Pass	Warning
TA_BC1_M1_R	Pass	Pass	Pass	Warning	Pass	Pass	Pass	Warning	Fail	Pass	Warning
TA_BC1_M2_F	Pass	Pass	Pass	Warning	Pass	Pass	Pass	Warning	Fail	Pass	Warning
TA_BC1_M2_R	Pass	Pass	Pass	Warning	Pass	Pass	Pass	Warning	Fail	Pass	Warning
TA_BC1_M3_F	Pass	Pass	Pass	Warning	Pass	Pass	Pass	Warning	Fail	Pass	Warning
TA_BC1_M3_R	Pass	Pass	Pass	Warning	Pass	Pass	Pass	Warning	Fail	Pass	Warning
TA_NOV9_M1_F	Pass	Pass	Pass	Warning	Pass	Pass	Pass	Warning	Fail	Pass	Warning
TA_NOV9_M1_R	Pass	Pass	Pass	Warning	Pass	Pass	Pass	Warning	Fail	Pass	Warning
TA_NOV9_M2_F	Pass	Pass	Pass	Warning	Pass	Pass	Pass	Warning	Fail	Pass	Warning
TA_NOV9_M2_R	Pass	Pass	Pass	Warning	Pass	Pass	Pass	Warning	Fail	Pass	Warning
TA_NOV9_M3_F	Pass	Pass	Pass	Warning	Pass	Pass	Pass	Warning	Fail	Pass	Warning
TA_NOV9_M3_R	Pass	Pass	Pass	Warning	Pass	Pass	Pass	Warning	Fail	Pass	Warning



## **5.4.9 Investigation of expression data to identify candidate avirulence genes for each race**

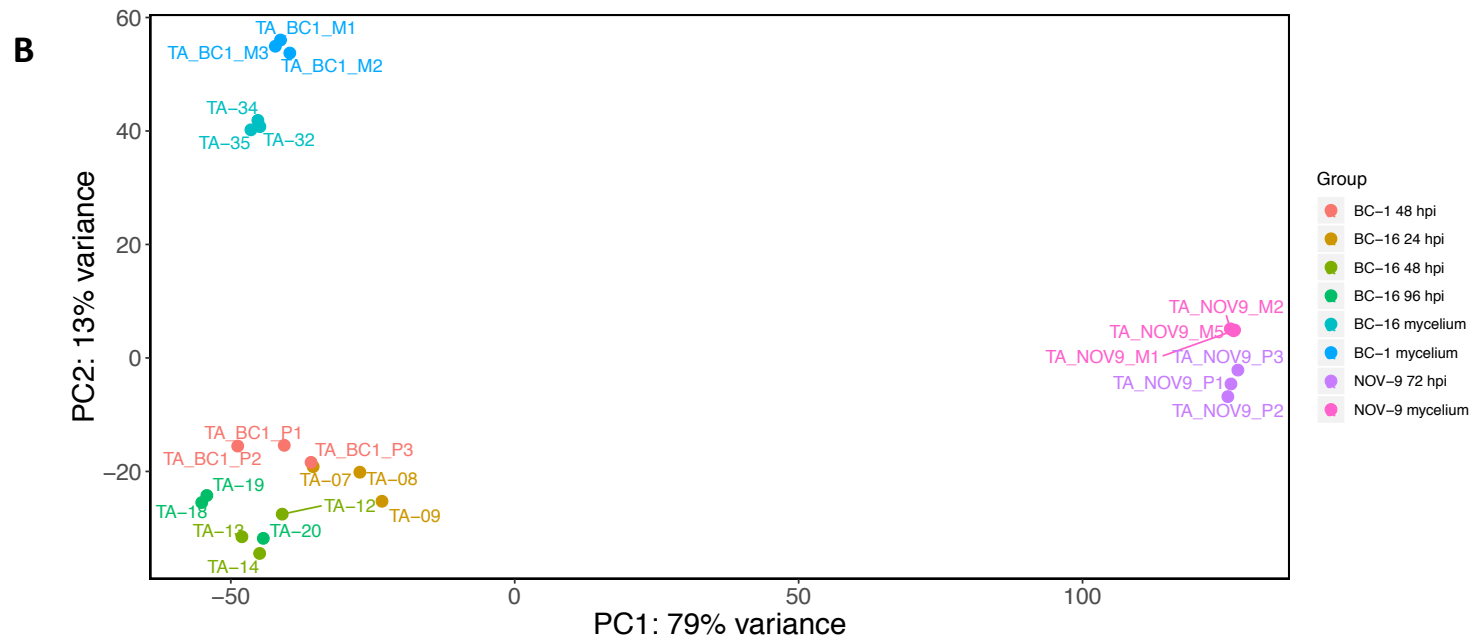
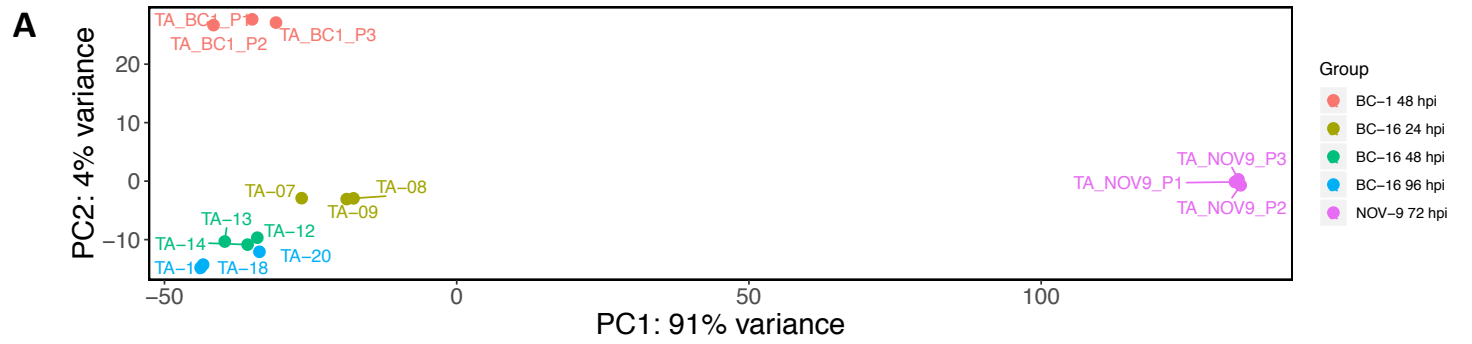
### **5.4.9.1 Alignment of RNA-Seq data and identification of differentially expressed genes from *in planta* time points compared to *in vitro* mycelium showed transcriptional reprofiling during the pathogenicity process**

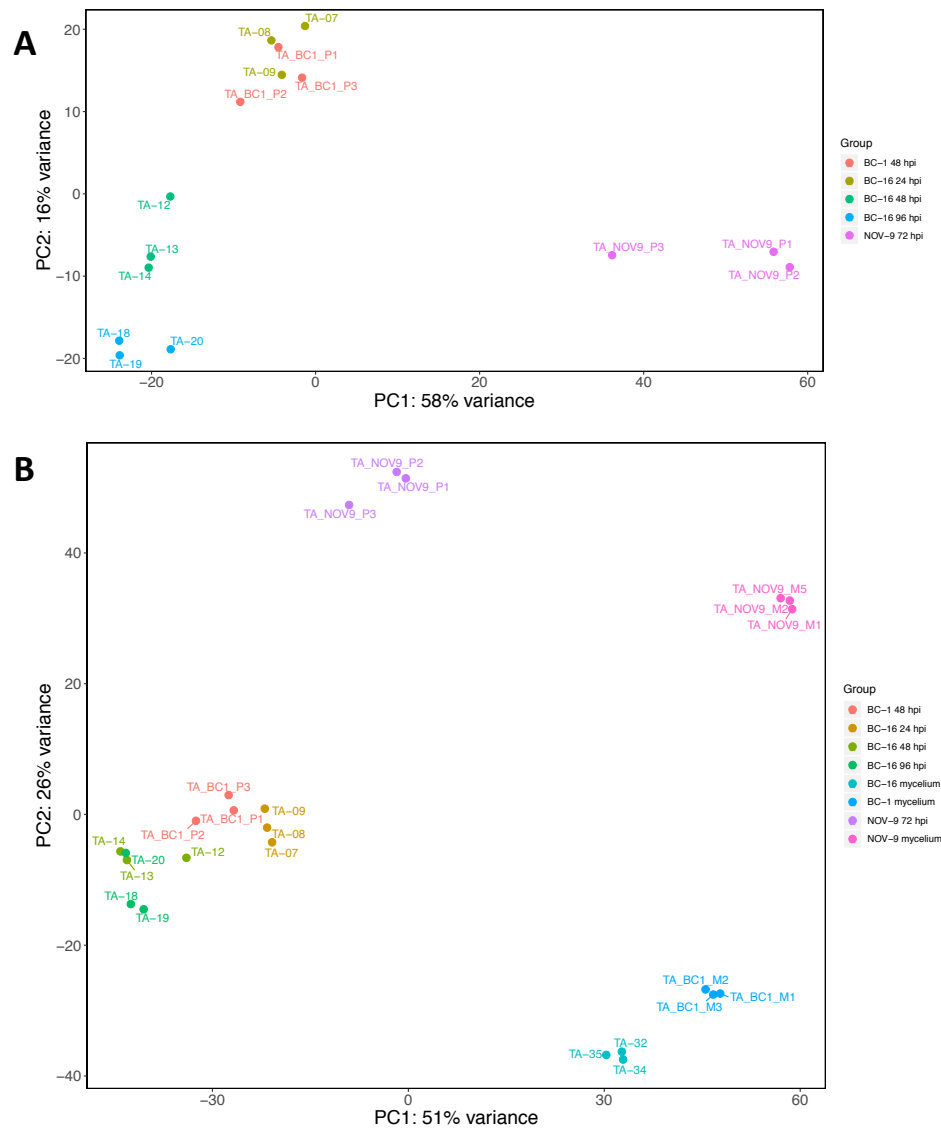
All RNA-Seq data was aligned to the assembled genomes of BC-1, BC-16 and NOV-9. Read mapping rates were similar to all assemblies, although aligning data to its source genome appeared to result in an improved mapping rate (Table A.4). PCA plots of the expression results showed that biological replicates for each timepoint clustered together, regardless of whether mycelial results were included or not. Where a principal component appeared to represent time since inoculation (eg. Principal component 2 in Fig. 5.13A) samples from BC-1 grouped with the earliest BC-16 timepoint. However, samples from NOV-9 appeared to group with the BC-16 samples from 48 hpi. Interestingly, the exact distribution of samples differed depending on which assembly was used as a reference for the analysis (Fig. 5.12 - 5.14).

Two genes of interest were identified in Chapter 4: g36121.t1 and g7404.t1. The gene g7404.t1 was identified as a gene of interest following the identification of CNV. This gene showed FPKM values of 110 at 96 hpi in BC-16, 41 at 48 hpi in BC-1 and 53 at 72 hpi in NOV-9 when the BC-16 assembly was used as a reference. Therefore, it was not identified as a candidate avirulence gene by this analysis as it is not differentially expressed between isolates of different races. The gene g36121.t1 was identified via orthology analyses, though as described in Chapter 4 it appeared to be an assembly error. This gene showed FPKM values of 259 at 48 hpi in BC-16, 176 at 48 hpi in BC-1 and 137 at 72 hpi in NOV-9 when the BC-16 assembly was used as a reference. Therefore, it was also not identified as a candidate avirulence gene by this analysis and further strengthened the conclusion of there being an assembly error in this region of the BC-1 and NOV-9 assemblies.

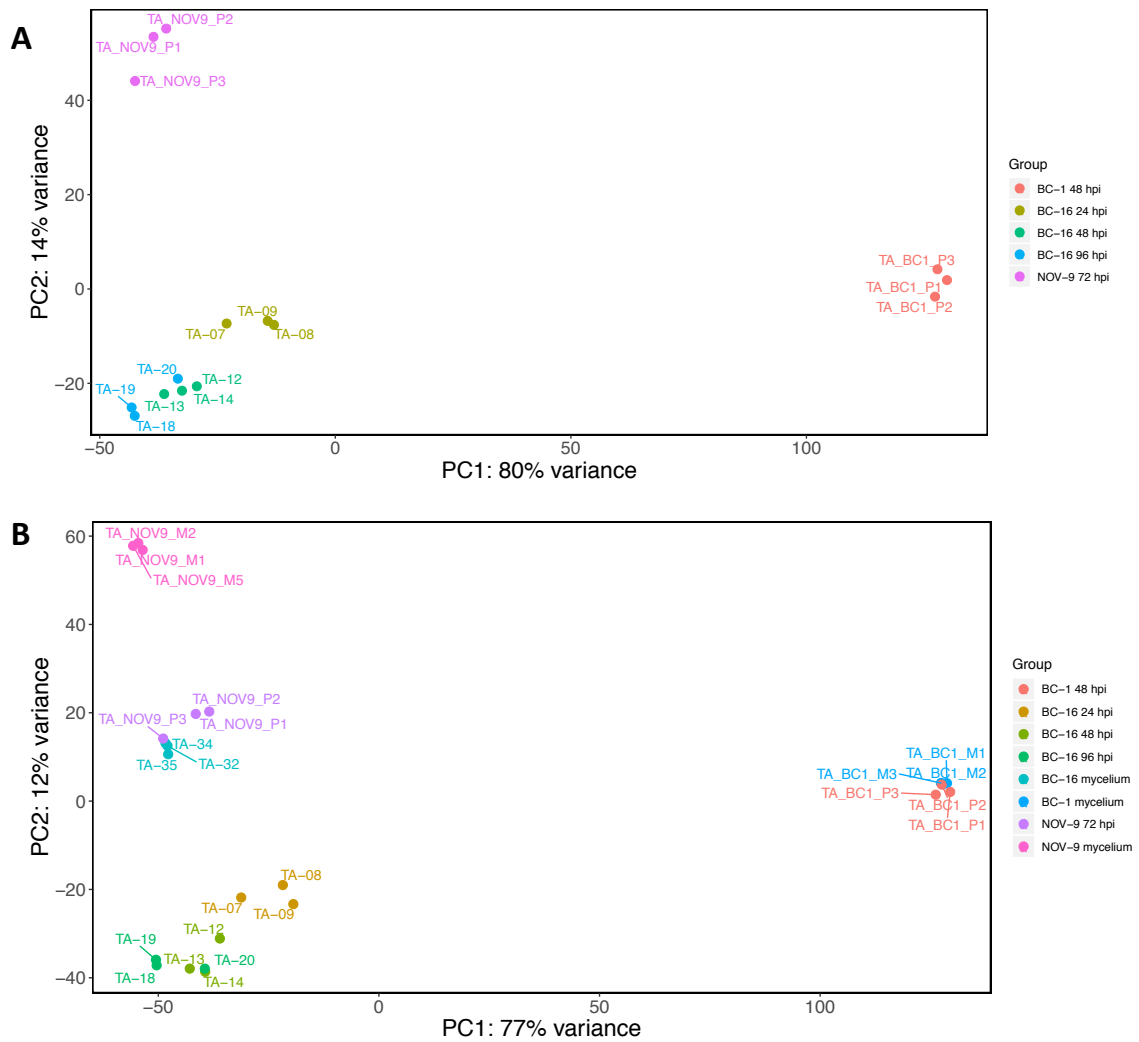
The number of genes showing evidence of expression (FPKM  $\geq$  5) were also investigated for each isolate's RNA-Seq aligned to each of the three reference assemblies used. Interestingly, BC-1 reads aligned to the NOV-9 genome and NOV-9 reads aligned to the BC-1 genome showed a far lower proportion of predicted transcripts being expressed than reads from the other isolates. For example, for all transcripts in BC-1, NOV-9 *in vitro* mycelial reads showed 4.90% of transcripts being expressed compared to 32.80% for BC-1 mycelial reads. Interestingly, for BC-16 predicted transcripts a similar proportion of transcripts showed evidence of expression in reads for all isolates. Despite this, it appeared that the proportions of each effector class that showed expression was similar across all three isolates (Table 5.10). Differential expression of transcripts from *in planta* samples and *in vitro* grown mycelium was also assessed for each isolate. Again, it was observed that analysing BC-1 with NOV-9 as the reference and vice versa resulted in a reduced number of predictions. However, for the BC-1 and NOV-9 data in the other assessments, twelve and fifteen percent of predicted genes were differentially expressed. This value was higher for BC-16, ranging between twenty-two and twenty-five percent, likely due to a larger amount of data being available (Table 5.11). This again represented a large scale transcriptional reprofiling upon infection, similar to that shown in Chapter 4.

**Fig. 5.12 (overleaf): Principal component analysis following expression assessment using RNA-Seq data from three isolates with BC-1 as a reference.** RNA-Seq reads were aligned to the assembly of the BC-1 isolate of *Phytophthora fragariae* using STAR version 2.5.3a (Dobin *et al.*, 2013). Predicted transcripts were then quantified with featureCounts version 1.5.2 (Liao *et al.*, 2014) and differential expression was identified with DESeq2 version 1.10.1 (Love *et al.*, 2014). Following this, an rlog transformation of the expression data was plotted as a PCA with R (R core team, 2016). **A:** Data plotted without RNA-Seq data from *in vitro* mycelium. **B:** Data plotted with RNA-Seq data from *in vitro* mycelium.





**Fig. 5.13: Principal component analysis following expression assessment using RNA-Seq data from three isolates with BC-16 as a reference.** RNA-Seq reads were aligned to the assembly of the BC-16 isolate of *Phytophthora fragariae* using STAR version 2.5.3a (Dobin *et al.*, 2013). Predicted transcripts were then quantified with featureCounts version 1.5.2 (Liao *et al.*, 2014) and differential expression was identified with DESeq2 version 1.10.1 (Love *et al.*, 2014). Following this, an rlog transformation of the expression data was plotted as a PCA with R (R core team, 2016). **A:** Data plotted without RNA-Seq data from *in vitro* mycelium. **B:** Data plotted with RNA-Seq data from *in vitro* mycelium.



**Fig. 5.14: Principal component analysis following expression assessment using RNA-Seq data from three isolates with NOV-9 as a reference.** RNA-Seq reads were aligned to the assembly of the NOV-9 isolate of *Phytophthora fragariae* using STAR version 2.5.3a (Dobin *et al.*, 2013). Predicted transcripts were then quantified with featureCounts version 1.5.2 (Liao *et al.*, 2014) and differential expression was identified with DESeq2 version 1.10.1 (Love *et al.*, 2014). Following this, an rlog transformation of the expression data was plotted as a PCA with R (R core team, 2016). **A**: Data plotted without RNA-Seq data from *in vitro* mycelium. **B**: Data plotted with RNA-Seq data from *in vitro* mycelium.

**Table 5.10: The number of transcripts showing expression levels greater than the threshold FPKM value of 5 in the three sequenced isolates of *Phytophthora fragariae*: BC-1, BC-16 and NOV-9**

Transcript Class	Total Predicted	Number of Transcripts Showing Expression								Total
		BC-1 48 hpi	BC-1 Mycelium	BC-16 24 hpi	BC-16 48 hpi	BC-16 96 hpi	BC-16 Mycelium	NOV-9 72 hpi	NOV-9 Mycelium	
BC-1 All Transcripts	34,500	10,184 (29.52%)	11,317 (32.80%)	9,730 (28.20%)	10,518 (30.49%)	11,635 (33.72%)	11,652 (33.77%)	1,642 (4.76%)	1,689 (4.90%)	14,264 (41.34%)
BC-1 RxLR Effectors	935	199 (21.28%)	172 (18.40%)	189 (20.21%)	206 (22.03%)	223 (23.85%)	199 (21.28%)	28 (2.99%)	32 (3.42%)	282 (30.16%)
BC-1 Crinkler Effectors	59	12 (20.34%)	18 (30.51%)	12 (20.34%)	13 (22.03%)	14 (23.73%)	17 (28.81%)	5 (8.47%)	5 (8.47%)	23 (38.98%)
BC-1 Apoplastic Effectors	3,890	566 (14.55%)	553 (14.22%)	508 (13.06%)	587 (15.09%)	670 (17.22%)	595 (15.30%)	151 (3.88%)	152 (3.91%)	955 (24.55%)
BC-1 Secreted Proteins	6,724	1,845 (27.44%)	1,875 (27.89%)	1,720 (25.58%)	1,908 (28.38%)	2,107 (31.34%)	1,959 (29.13%)	297 (4.42%)	307 (4.57%)	2,600 (38.67%)
BC-16 All Transcripts	41,400	10,296 (24.87%)	11,443 (27.64%)	9,782 (23.63%)	10,603 (25.61%)	11,763 (28.41%)	11,723 (28.32%)	10,063 (24.31%)	11,239 (27.15%)	13,988 (33.79%)
BC-16 RxLR Effectors	1,058	216 (20.42%)	185 (17.49%)	207 (19.57%)	224 (21.17%)	240 (22.68%)	210 (19.85%)	223 (21.08%)	190 (17.96%)	287 (27.13%)
BC-16 Crinkler Effectors	88	21 (23.86%)	27 (30.68%)	20 (22.73%)	21 (23.86%)	24 (27.7%)	26 (29.55%)	25 (28.41%)	23 (26.14%)	31 (35.23%)
BC-16 Apoplastic Effectors	4,864	622 (12.79%)	601 (12.36%)	562 (11.55%)	642 (13.20%)	733 (15.07%)	649 (13.34%)	635 (13.06%)	600 (12.34%)	948 (19.49%)
BC-16 Secreted Proteins	8,024	1,981 (24.69%)	1,984 (24.73%)	1,823 (22.72%)	2,028 (25.27%)	2,238 (27.89%)	2,063 (25.71%)	2,007 (25.01%)	1,984 (24.73%)	2,669 (33.26%)
NOV-9 All Transcripts	34,361	1,539 (4.48%)	1,661 (4.83%)	9,742 (28.35%)	10,490 (30.53%)	11,625 (33.83%)	11,608 (33.78%)	10,540 (30.67%)	11,055 (32.17%)	19,940 (58.03%)
NOV-9 RxLR Effectors	961	39 (4.06%)	41 (4.27%)	191 (19.88%)	207 (21.54%)	223 (23.20%)	199 (20.71%)	204 (21.23%)	178 (18.52%)	295 (30.70%)
NOV-9 Crinkler Effectors	61	5 (8.20%)	6 (9.84%)	11 (18.03%)	12 (19.67%)	13 (21.31%)	14 (22.95%)	14 (22.95%)	14 (22.95%)	21 (34.43%)
NOV-9 Apoplastic Effectors	4,090	138 (3.37%)	150 (3.72%)	505 (12.35%)	593 (14.50%)	684 (16.72%)	593 (14.50%)	581 (14.21%)	556 (13.59%)	999 (24.43%)
NOV-9 Secreted Proteins	6,968	279 (4.00%)	298 (4.28%)	1,733 (24.87%)	1,930 (27.70%)	2,136 (30.65%)	1,976 (28.36%)	1,897 (27.22%)	1,898 (27.24%)	2,688 (38.58%)

Hours post inoculation (hpi).

**Table 5.11: The number of transcripts showing differential expression in *in planta* time points compared to *in vitro* mycelium for BC-1, BC-16 and NOV-9, with the assemblies of these isolates used as references**

Reference Assembly	Transcript Number	DETs	BC-1		DETs	BC-16		DETs	NOV-9	
			Upregulated	Downregulated		Upregulated	Downregulated		Upregulated	Downregulated
BC-1	34,500	5,181 (15.02%)	2,740 (7.94%)	2,441 (7.08%)	8,937 (25.90%)	4,843 (14.04%)	4,359 (12.63%)	410 (1.19%)	279 (0.81%)	131 (0.38%)
BC-16	41,400	5,422 (13.10%)	2,455 (5.93%)	2,967 (7.17%)	9,191 (22.20%)	4,523 (10.93%)	4,912 (11.86%)	5,101 (12.32%)	3,319 (8.02%)	1,782 (4.30%)
NOV-9	34,361	383 (1.11%)	203 (0.59%)	180 (0.52%)	8,477 (24.67%)	4,862 (14.15%)	3,796 (11.05%)	4,767 (13.87%)	3,379 (9.83%)	1,388 (4.04%)

The threshold to be scored as a differentially expressed transcript (DET) was a log<sub>2</sub> fold change (LFC) greater than one (for up regulation) or less than minus one (for down regulation).

#### **5.4.9.2 Identification of uniquely expressed transcripts in the BC-1, BC-16 and NOV-9 isolates showed a large number of genes showing expression in only a single isolate**

In order to aid in the identification of candidate avirulence genes, transcripts that were only expressed in one isolate were identified. This was performed using the assemblies of BC-1, BC-16 and NOV-9 as references to account for the differences in the numbers of expressed and differentially expressed genes identified when using a different reference assembly. A transcript was called as uniquely expressed if it showed an FPKM value of five or above in any *in planta* time point for one isolate, but not in the other two isolates. Values were also calculated for each identified effector class, alongside secreted proteins. These results showed substantial variation between the numbers of uniquely expressed genes depending on which assembly was used as the reference. In the most extreme case, BC-1 showed 93 genes with unique expression with the BC-16 assembly used as the reference, but showed 820 genes with unique expression with the NOV-9 assembly used as the reference (Table 5.12). This could have been due to assembly or annotation errors, or due to spurious alignments of the RNA-Seq data. Therefore, all this information will be used for identifying candidate avirulence genes.



**Table 5.12: Identification of uniquely expressed genes in the BC-1, BC-16 and NOV-9 isolates using the assemblies of these three isolates as a reference**

Reference Assembly	BC-1					BC-16					NOV-9				
	All Genes	RxLR Effectors (Whisson <i>et al.</i> , 2007; Armitage <i>et al.</i> , 2018b)	Crinkler Effectors (Armitage <i>et al.</i> , 2018b)	Apoplastic Effectors (Sperschneider <i>et al.</i> , 2018)	Secreted Proteins (Nielsen <i>et al.</i> , 1997; Bendtsen <i>et al.</i> , 2004; Petersen <i>et al.</i> , 2011; Käll <i>et al.</i> , 2007)	All Genes	RxLR Effectors (Whisson <i>et al.</i> , 2007; Armitage <i>et al.</i> , 2018b)	Crinkler Effectors (Armitage <i>et al.</i> , 2018b)	Apoplastic Effectors (Sperschneider <i>et al.</i> , 2018)	Secreted Proteins (Nielsen <i>et al.</i> , 1997; Bendtsen <i>et al.</i> , 2004; Petersen <i>et al.</i> , 2011; Käll <i>et al.</i> , 2007)	All Genes	RxLR Effectors (Whisson <i>et al.</i> , 2007; Armitage <i>et al.</i> , 2018b)	Crinkler Effectors (Armitage <i>et al.</i> , 2018b)	Apoplastic Effectors (Sperschneider <i>et al.</i> , 2018)	Secreted Proteins (Nielsen <i>et al.</i> , 1997; Bendtsen <i>et al.</i> , 2004; Petersen <i>et al.</i> , 2011; Käll <i>et al.</i> , 2007)
BC-1	159	2	0	14	16	2,021	45	2	162	306	808	20	4	119	60
BC-16	93	4	0	10	12	1,588	33	2	148	19	476	7	4	44	49
NOV-9	820	27	3	116	58	2,083	49	0	210	391	470	6	2	55	44

#### **5.4.9.3 Identification of uniquely differentially expressed transcripts showed a large number of genes showing differential expression in only a single isolate**

Another line of evidence generated to aid the identification of candidate avirulence genes was the identification of uniquely differentially expressed transcripts (DETs). This was again performed using the assemblies of BC-1, BC-16 and NOV-9 as references to account for the differences in the numbers of expressed and differentially expressed genes identified when using a different reference assembly. A transcript was called uniquely differentially expressed if, for a specific isolate, a transcript showed a  $\log_2$  fold change (LFC) greater than or equal to 3, with a minimum  $p$ -value of 0.05, in at least one comparison of an *in planta* time point for the isolate of interest versus an *in planta* time point for the other isolates. Values were also calculated for each identified effector class, alongside secreted proteins. Similar to the results of unique expression, substantial variation was identified depending on which assembly was used as the reference. In the most extreme case, NOV-9 showed 397 uniquely differentially expressed transcripts when the BC-16 assembly was used as the reference, but 1,158 uniquely differentially expressed transcripts were identified when the NOV-9 assembly was used as the reference (Table 5.13). This could have been due to assembly or annotation errors, or due to spurious alignments of the RNA-Seq data. These results will aid in the identification of candidate avirulence genes by ensuring a gene is sufficiently different in its expression in the different isolates.

**Table 5.13: Identification of uniquely differentially expressed genes in the BC-1, BC-16 and NOV-9 isolates using the assemblies of these three isolates as a reference**

Reference Assembly	All Genes	BC-1				All Genes	BC-16				All Genes	NOV-9			
		RxLR Effectors (Whisson <i>et al.</i> , 2007; Armitage <i>et al.</i> , 2018b)	Crinkler Effectors (Armitage <i>et al.</i> , 2018b)	Apoplastic Effectors (Sperschneider <i>et al.</i> , 2018)	Secreted Proteins (Nielsen <i>et al.</i> , 1997; Bendtsen <i>et al.</i> , 2004; Petersen <i>et al.</i> , 2011; Käll <i>et al.</i> , 2007)		RxLR Effectors (Whisson <i>et al.</i> , 2007; Armitage <i>et al.</i> , 2018b)	Crinkler Effectors (Armitage <i>et al.</i> , 2018b)	Apoplastic Effectors (Sperschneider <i>et al.</i> , 2018)	Secreted Proteins (Nielsen <i>et al.</i> , 1997; Bendtsen <i>et al.</i> , 2004; Petersen <i>et al.</i> , 2011; Käll <i>et al.</i> , 2007)		RxLR Effectors (Whisson <i>et al.</i> , 2007; Armitage <i>et al.</i> , 2018b)	Crinkler Effectors (Armitage <i>et al.</i> , 2018b)	Apoplastic Effectors (Sperschneider <i>et al.</i> , 2018)	Secreted Proteins (Nielsen <i>et al.</i> , 1997; Bendtsen <i>et al.</i> , 2004; Petersen <i>et al.</i> , 2011; Käll <i>et al.</i> , 2007)
BC-1	421	11	0	61	95	381	12	0	37	69	808	20	4	108	67
BC-16	50	0	0	2	3	146	7	0	15	25	397	7	2	25	39
NOV-9	725	24	2	79	43	410	20	1	82	120	1,158	27	2	96	152

#### **5.4.9.4 Identification of candidate avirulence genes in the BC-1, BC-16 and NOV-9 isolates produced a small number of high to medium confidence candidate genes**

Using the results for uniquely expressed genes and uniquely differentially expressed genes, lists of candidate genes were identified with various confidence levels. Using the results of the orthology analysis (Chapter 4, section 4.4.10), the results of these analyses against different reference assemblies were compared. Transcripts were assigned a score of one - six, dependent on the presence of the transcript and the transcripts of its orthologues being scored as uniquely expressed or uniquely differentially expressed for the isolate of interest. High confidence avirulence genes were only identified in BC-16, likely due to the poor results of BC-1 data aligned to NOV-9 and vice versa. Only seventeen candidates were scored as high confidence in BC-16, four as medium confidence in BC-1 and forty-nine as medium confidence in NOV-9. These groups consisted of four RxLR effectors in BC-16, four in NOV-9 and none in BC-1; although one apoplastic effector was identified in BC-1 (Table 5.14; Table A.8). In order to identify possible candidates that may have been scored as false negatives by this process, the gene sets were investigated manually through inspection of FPKM values in each sequenced sample. Through this process, it was discovered that a medium confidence candidate RxLR effector, g27513.t1, was a very strong candidate for the avirulence gene in race UK2. This was due to extremely high levels of expression in BC-16 (peaking at an FPKM value of 9,392 at 24 hpi and reducing in subsequent time points). However, it was not scored as uniquely expressed with any reference genome, as the expression in BC-1 or NOV-9 exceeded the threshold of 5, despite the predicted expression being at 1,000 times lower in these isolates than in BC-16, at an FPKM value of 5 in BC-1 and 34 in NOV-9. This difference was greater than that for any of the high confidence candidates and so this gene was selected for further investigation.

**Table 5.14: Identified candidate avirulence genes in the BC-1, BC-16 and NOV-9 isolates of *Phytophthora fragariae***

Reference Assembly	All Genes	BC-1				All Genes	BC-16				All Genes	NOV-9			
		RxLR Effectors (Whisson <i>et al.</i> , 2007; Armitage <i>et al.</i> , 2018b)	Crinkler Effectors (Armitage <i>et al.</i> , 2018b)	Apoplastic Effectors (Sperschneider <i>et al.</i> , 2018)	Secreted Proteins (Nielsen <i>et al.</i> , 1997; Bendtsen <i>et al.</i> , 2004; Käll <i>et al.</i> , 2007)		RxLR Effectors (Whisson <i>et al.</i> , 2007; Armitage <i>et al.</i> , 2018b)	Crinkler Effectors (Armitage <i>et al.</i> , 2018b)	Apoplastic Effectors (Sperschneider <i>et al.</i> , 2018)	Secreted Proteins (Nielsen <i>et al.</i> , 1997; Bendtsen <i>et al.</i> , 2004; Käll <i>et al.</i> , 2007)		RxLR Effectors (Whisson <i>et al.</i> , 2007; Armitage <i>et al.</i> , 2018b)	Crinkler Effectors (Armitage <i>et al.</i> , 2018b)	Apoplastic Effectors (Sperschneider <i>et al.</i> , 2018)	Secreted Proteins (Nielsen <i>et al.</i> , 1997; Bendtsen <i>et al.</i> , 2004; Käll <i>et al.</i> , 2007)
High	0	0	0	0	0	17	4	0	8	15	0	0	0	0	0
Middle	4	0	0	1	2	206	24	0	75	175	49	4	2	24	26
Low	120	13	0	62	102	96	6	2	39	67	177	21	0	74	130
Total	124	13	0	63	104	319	34	2	122	257	226	25	2	98	156

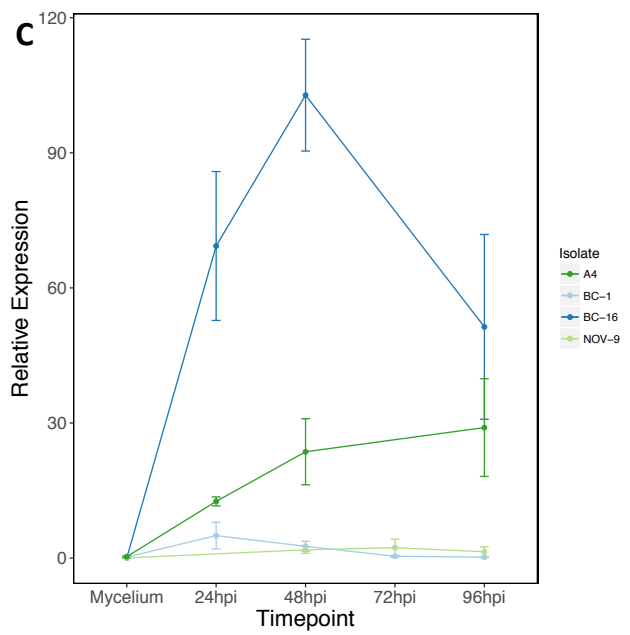
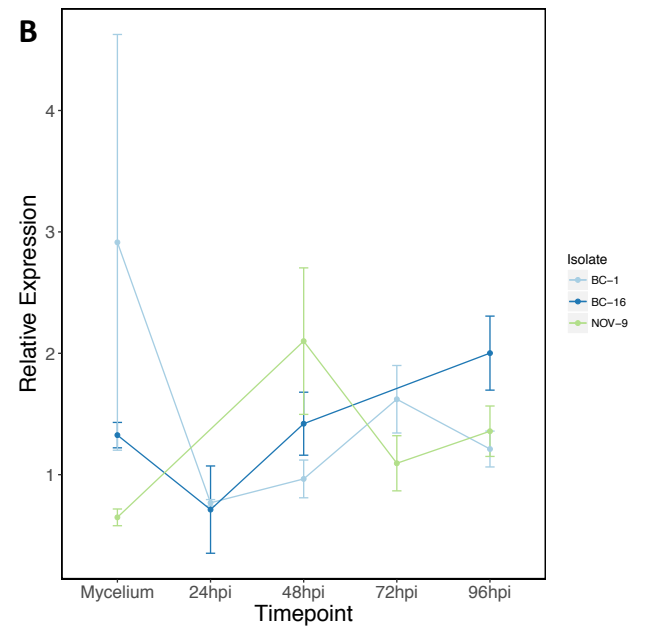
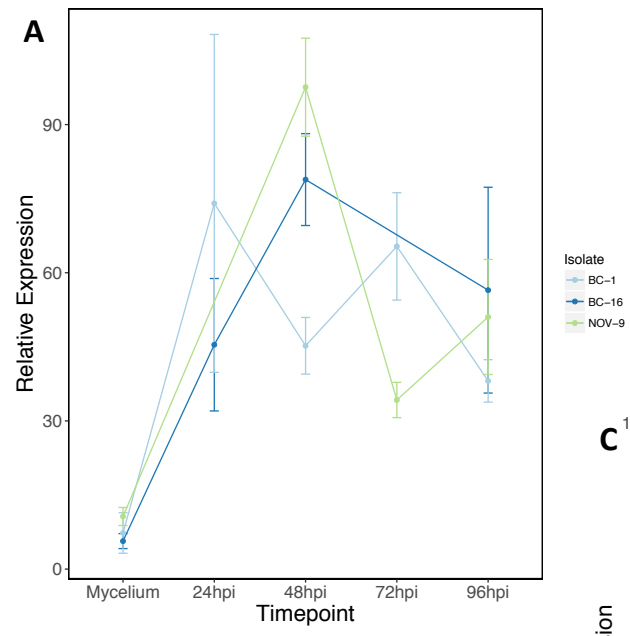
#### **5.4.9.5 Confirmation of RNA-Seq data using RT-qPCR reinforces the identification of the high quality avirulence gene candidate in BC-16**

In order to confirm the expression levels identified by the RNA-Seq analysis the following genes were investigated by RT-qPCR: the strong BC-16 candidate avirulence gene (g27513.t1), an RxLR effector that was predicted to peak in the 24 hpi time point (g32018.t1) and an RxLR effector that was predicted to peak in the 48 hpi time point (g23965.t1). Unfortunately, the expression patterns for the early and middle peaking RxLR could not be confirmed due to high levels of variation between biological replicates in the measurements. Despite these error levels, it was confirmed that the candidate avirulence gene was only expressed in the UK2 isolates A4 and BC-16 *in planta* samples, though the pattern of peaking at 24 hpi and then falling, as observed in the RNA-Seq results for BC-16, could not be confirmed. Interestingly, BC-16 appeared to show higher levels of expression than A4, this may have been due to slower growth of the A4 isolate compared to BC-16. The source of the statistical errors was determined to be due to variance between biological replicates rather than technical replicates. This was shown by removing outlier values using a threshold of 1.5 times the interquartile range, followed by the Grubb's test (Grubbs, 1950) for technical replicates. This resulted in the maximum value for the standard error of the mean (SEM) of 2.48, with only 4 biological samples having an SEM value above 1. Despite this, SEM values were still observed in excess of 10 in a large number of time points (Fig. 5.15).

Statistical significance of the differences between samples was also assessed. An analysis of variance (ANOVA) test was used to assess the difference between and within groups from the RT-qPCR results. For the candidate avirulence gene in BC-16 all investigated comparisons showed significant difference (Isolate, Timepoint and Isolate:Timepoint), for the 24 hpi peaking RxLR only Timepoint comparisons were significant and no comparisons were significant for the 48 hpi peaking RxLR. A Tukey honestly significant difference (HSD) test was also performed to assess comparisons between all time points and all isolates. For the candidate avirulence gene in BC-16, this showed *in planta* BC-16 time points

were significantly different to all BC-1 and NOV-9 samples, as well as all *in vitro* mycelial samples from BC-16 and the majority of A4 samples. A4 showed a significant difference when compared against all BC-1, BC-16 and NOV-9 data, but individual timepoint comparisons did not show significant differences, except for a rare few comparisons against BC-16 data. For the RxLR effector predicted to peak at 24 hpi, significance was only observed between samples from *in planta* time points and *in vitro* mycelium. Further investigation however showed not all these comparisons to be significant between isolates. No statistically significant differences were identified in the RxLR effector predicted to peak at 48 hpi (Table 5.15 and A.3 - A.5).

**Fig. 5.15 (overleaf): Results of RT-qPCR analysis of three genes of interest in the A4, BC-1, BC-16 and NOV-9 isolates of *Phytophthora fragariae*.** Plots created by the ggplot2 R package version 2.2.1 (Wickman, 2016) in R version 3.4.3 (R core team, 2017). **A:** Expression data of an RxLR effector predicted to peak in expression at 24 hpi (g32018.t1). **B:** Expression data of an RxLR effector predicted to peak in expression at 48 hpi (g23965.t1). **C:** Expression data of a strong candidate for the avirulence gene possessed by BC-16, but not BC-1 and NOV-9 (g27513.t1).





**Table 5.15: Results of statistical testing for all significant comparisons from RT-qPCR data**

Candidate UK2 Avr Gene			RxLR Effector Peaking at 24 hpi		
Comparison	<i>p</i> -value adjusted	Test Used	Comparison	<i>p</i> -value adjusted	Test Used
Isolate	0.00000000000676	ANOVA	Timepoint	0.0000457	ANOVA
Timepoint	0.0000384	ANOVA	Mycelium-24hpi	0.00174	Tukey HSD
Isolate:Timepoint	0.0000137	ANOVA	Mycelium-48hpi	0.0000193	Tukey HSD
BC-1-A4	0.0395	Tukey HSD	Mycelium-72hpi	0.0159	Tukey HSD
BC-16-A4	0.00000200	Tukey HSD	Mycelium-96hpi	0.00754	Tukey HSD
NOV-9-A4	0.0487	Tukey HSD	BC-1:Mycelium-BC-16:48hpi	0.0433	Tukey HSD
BC-16-BC-1	0.00	Tukey HSD	BC-16:Mycelium-BC-16:48hpi	0.0356	Tukey HSD
NOV-9-BC-16	0.00	Tukey HSD	BC-1:Mycelium-NOV-9:48hpi	0.00410	Tukey HSD
Mycelium-24hpi	0.00415	Tukey HSD	BC-16:Mycelium-NOV-9:48hpi	0.00330	Tukey HSD
Mycelium-48hpi	0.0000111	Tukey HSD	NOV-9:Mycelium-NOV-9:48hpi	0.00634	Tukey HSD
Mycelium-96hpi	0.00665	Tukey HSD			
BC-16:24hpi-A4:24hpi	0.00157	Tukey HSD			
BC-16:48hpi-A4:24hpi	0.000000300	Tukey HSD			
BC-16:24hpi-BC-1:24hpi	0.000218	Tukey HSD			
BC-16:48hpi-BC-1:24hpi	0.00	Tukey HSD			
BC-16:96hpi-BC-1:24hpi	0.0201	Tukey HSD			
A4:48hpi-BC-16:24hpi	0.0235	Tukey HSD			
BC-1:48hpi-BC-16:24hpi	0.000117	Tukey HSD			
NOV-9:48hpi-BC-16:24hpi	0.0000953	Tukey HSD			
BC-1:72hpi-BC-16:24hpi	0.0000655	Tukey HSD			
NOV-9:72hpi-BC-16:24hpi	0.000108	Tukey HSD			
BC-1:96hpi-BC-16:24hpi	0.0000623	Tukey HSD			
NOV-9:96hpi-BC-16:24hpi	0.0000849	Tukey HSD			
A4:Mycelium-BC-16:24hpi	0.0000638	Tukey HSD			
BC-1:Mycelium-BC-16:24hpi	0.0000624	Tukey HSD			
BC-16:Mycelium-BC-16:24hpi	0.0000591	Tukey HSD			
NOV-9:Mycelium-BC-16:24hpi	0.0000595	Tukey HSD			
BC-16:48hpi-A4:48hpi	0.00000450	Tukey HSD			
BC-16:48hpi-BC-1:48hpi	0.00	Tukey HSD			
BC-16:96hpi-BC-1:48hpi	0.0115	Tukey HSD			
NOV-9:48hpi-BC-16:48hpi	0.00	Tukey HSD			
BC-1:72hpi-BC-16:48hpi	0.00	Tukey HSD			
NOV-9:72hpi-BC-16:48hpi	0.00	Tukey HSD			
A4:96hpi-BC-16:48hpi	0.0000180	Tukey HSD			
BC-1:96hpi-BC-16:48hpi	0.00	Tukey HSD			
BC-16:96hpi-BC-16:48hpi	0.00589	Tukey HSD			
NOV-9:96hpi-BC-16:48hpi	0.00	Tukey HSD			
A4:Mycelium-BC-16:48hpi	0.00	Tukey HSD			
BC-1:Mycelium-BC-16:48hpi	0.00	Tukey HSD			
BC-16:Mycelium-BC-16:48hpi	0.00	Tukey HSD			
NOV-9:Mycelium-BC-16:48hpi	0.00	Tukey HSD			
BC-16:96hpi-NOV-9:48hpi	0.00950	Tukey HSD			
BC-16:96hpi-BC-1:72hpi	0.00669	Tukey HSD			
BC-16:96hpi-NOV-9:72hpi	0.0107	Tukey HSD			
BC-16:96hpi-BC-1:96hpi	0.0063854	Tukey HSD			
NOV-9:96hpi-BC-16:96hpi	0.00853	Tukey HSD			
A4:Mycelium-BC-16:96hpi	0.00653	Tukey HSD			
BC-1:Mycelium-BC-16:96hpi	0.00639	Tukey HSD			
BC-16:Mycelium-BC-16:96hpi	0.00608	Tukey HSD			
NOV-9:Mycelium-BC-16:96hpi	0.00611	Tukey HSD			

Significance threshold of  $p < 0.05$  used. Only significant results shown, all results are in Table A.5 - A.7.

## **5.4.10 Further investigation of candidate avirulence gene identified in BC-16**

### **5.4.10.1 Construction of a coexpression network allowed the identification of genes which showed similar expression patterns, although attempts to identify a promoter sequence failed**

In order to understand the expression pattern of the strong candidate avirulence gene, g27513, in BC-16, further investigations were performed. Firstly, it was of interest to identify possible promoter regions upstream of the gene. Genes with a similar expression pattern were identified through a coexpression analysis. Following the construction of coexpression modules, the gene of interest was found in the 'red' coexpression module, which consisted of 1,021 genes. However, when this module was manually investigated, it appeared to include genes that did not show a similar expression pattern to the strong candidate avirulence gene in BC-16. Therefore, this was narrowed down to twenty-nine high confidence genes and thirty-one lower confidence genes. An analysis was also performed using all six genes which showed an FPKM value above 9,000 in at least one BC-16 timepoint. Also analysed was the entire module which included the candidate avirulence gene before merging was performed, this was the 'saddlebrown' module and consisted of 197 genes. DREME (Bailey, 2011) was used to identify possible promoter sequences, however, when bootstrapping with 100 repetitions this process with a newly randomly generated negative set for each iteration, no motifs could be reliably identified in any dataset.

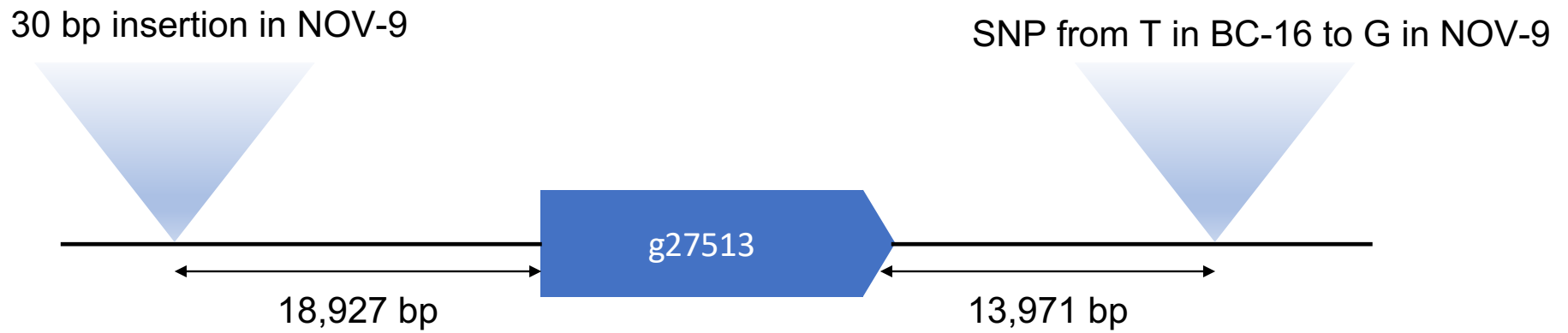
An investigation was also performed as to whether the candidate avirulence gene was a member of an effector family in BC-16. The gene was analysed by BLASTn against the predicted RxLR effectors (Altschul *et al.*, 1990). This identified two hits: g27504.t1 and g29093.t1, neither of which showed evidence of expression. Therefore, it appeared unlikely that this gene was part of a family sharing sequence identity.

#### **5.4.10.2 Investigation of changes in *cis* showed a SNP site 14 Kb downstream of the stop codon of the strong candidate avirulence gene in BC-16**

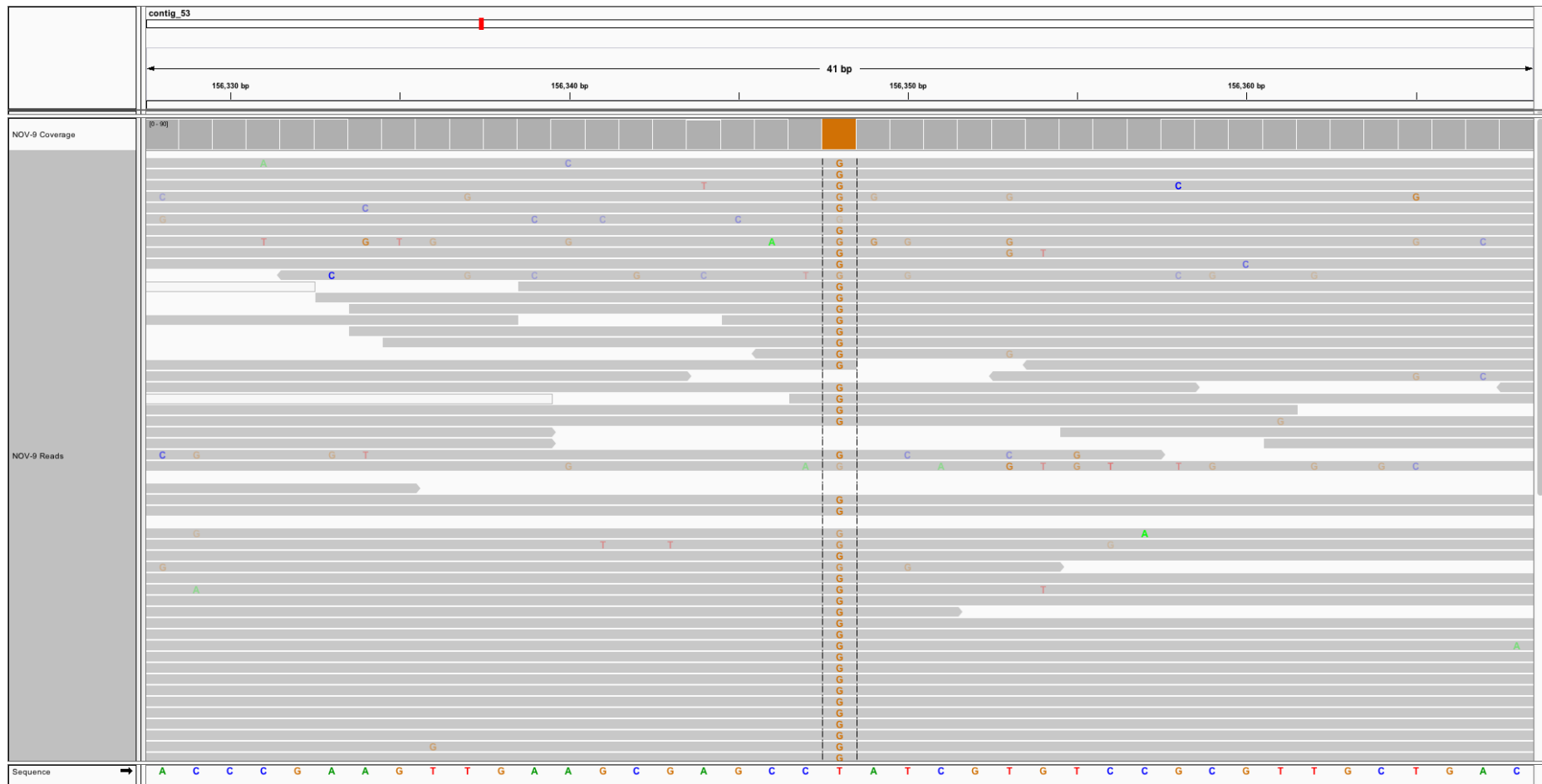
Further investigations were performed to investigate the cause of the difference in expression levels between the analysed isolates. As no variants were detected near this gene, additional long read sequencing was performed using nanopore technology developed by ONT, which provided 70.76x coverage. This was assembled and produced an assembly of greater contiguity than the FALCON-Unzip assembly of BC-16 (see Chapter 3, section 3.4.3.2), with a far larger N50 value. This assembly was also larger than the FALCON assembly, at 94 Mb compared to 91 Mb. Whilst more single copy BUSCO genes were identified, a greater number of fragmented and duplicated BUSCO genes were also identified. A larger amount of repetitive sequences were also identified (Table 5.16).

Through a BLASTn search, the location of the orthologous gene to the BC-16 candidate avirulence gene was identified in NOV-9 (Altschul *et al.*, 1990). These contigs were then trimmed to allow for the longest alignment to be produced, where the entirety of the sequence was covered in both assemblies. The gene coding sequences were shown to be identical (Fig. A.2). For the upstream region, the nearest variant was a 30 bp insertion in NOV-9, 18 Kb upstream of the start codon (Fig. 5.16). For the downstream region the nearest variant was a SNP from T in BC-16 to G in NOV-9 14 Kb downstream of the stop codon (Fig. 5.16). As both PacBio and ONT data are prone to sequencing errors, despite error correction being performed as part of the assembly process, the Illumina reads for these isolates were aligned to the assemblies. This showed that whilst the T to G SNP appeared to be reflected in the Illumina reads, there was very little evidence from these reads supporting the 30 bp insertion, though there was slightly more evidence for its presence in the NOV-9 reads than the BC-16 reads (Fig. 5.17 - Fig. 5.20 and Fig. A.3 - A.6). Further investigation showed this region to be in a predicted transposon ORF. Therefore, it appeared more likely to be a sequencing error in one of these isolates, that was not corrected due to poor quality alignments of the short reads.

The level of synteny between these two isolates was also assessed with Satsuma (Grabherr *et al.*, 2010), however, the assemblies proved too fragmented to show reliable results (Fig. 5.21).



**Fig. 5.16:** Graphical representation of sequence variants in the flanking regions of the candidate avirulence gene in BC-16 and its orthologous gene in NOV-9.



**Fig. 5.17: Alignment of the Illumina reads of NOV-9 to the BC-16 FALCON assembly at the identified SNP downstream of the strong candidate avirulence gene in BC-16.** Alignment was performed using Bowtie 2 (Langmead and Salzberg, 2012). Figure created in IGV (Thorvaldsdóttir *et al.*, 2013).



**Fig. 5.18: Alignment of the Illumina reads of NOV-9 to the BC-16 FALCON assembly at the identified INDEL upstream of the strong candidate avirulence gene in BC-16.** Alignment was performed using Bowtie 2 (Langmead and Salzberg, 2012). Figure created in IGV (Thorvaldsdóttir *et al.*, 2013).

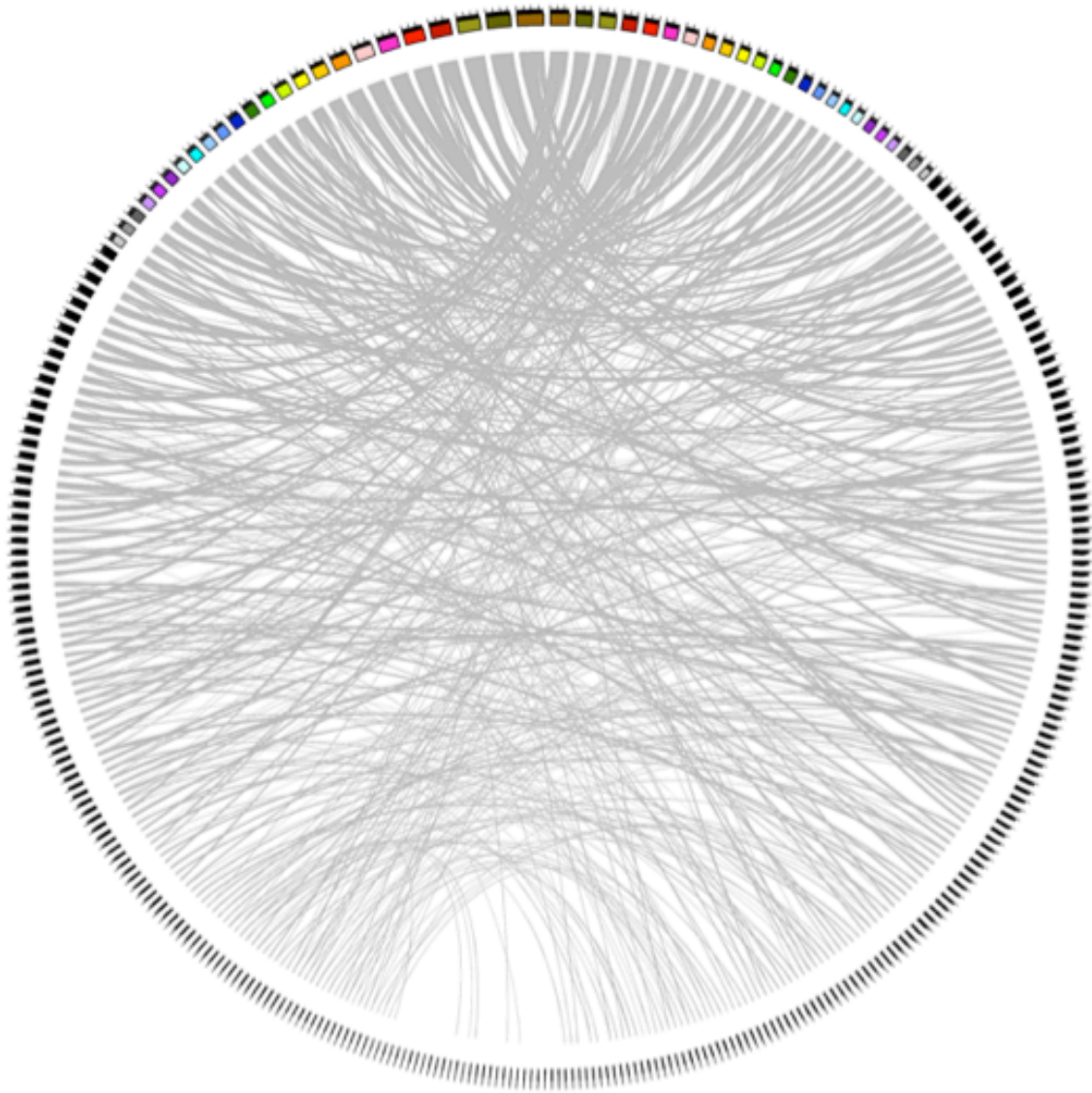


**Fig. 5.19: Alignment of the Illumina reads of BC-16 to the NOV-9 ONT assembly at the identified INDEL upstream of the orthologue of the strong candidate avirulence gene in BC-16.** Alignment was performed using Bowtie 2 (Langmead and Salzberg, 2012). Figure created in IGV (Thorvaldsdóttir *et al.*, 2013).





**Fig. 5.20: Alignment of the Illumina reads of NOV-9 to the NOV-9 ONT assembly at the identified INDEL upstream of the orthologue of the strong candidate avirulence gene in BC-16.** Alignment was performed using Bowtie 2 (Langmead and Salzberg, 2012). Figure created in IGV (Thorvaldsdóttir *et al.*, 2013).



**Fig. 5.21: Syntenic alignment of the assemblies of BC-16 and NOV-9.** Contigs of each isolate are arranged around the outside of this figure, with ribbons linking syntenic regions. Synteny alignments were created with Satsuma (Grabherr *et al.*, 2010). Figure created using Circos (Krzywinski *et al.*, 2009).

**Table 5.16: A comparison of assembly statistics between the FALCON assembly of BC-16 and the ONT assembly of NOV-9**

Isolate	Assembly Size	Contig Number (>= 500 bp)	N50	Largest Contig	% Repeat Masked	Single-Copy BUSCOs	Duplicated BUSCOs	Fragmented BUSCOs	Missing BUSCOs
BC-16	90.97 Mbp	180	93.46 Kb	2.70 Mbp	39.96	266 (87.79%)	9 (2.97%)	5 (1.65%)	23 (7.59%)
NOV-9	93.72 Mbp	124	1,260 Kb	3.69 Mbp	40.70	268 (88.45%)	12 (3.96%)	6 (1.98%)	17 (5.61%)

BUSCO genes were identified using BUSCO version 3.0.1 with the eukaryota\_odb9 database (Simão et al., 2015).

### 5.4.10.3 Investigation of changes in *trans* identified some candidate transcription factors and transcriptional regulators that may be involved in the differential expression of the strong candidate avirulence gene in BC-16

As no clear explanation for the expression differences was found from *cis* changes, investigations were performed to identify any changes in *trans*. Therefore, transcription factors and transcriptional regulators were identified in all genomes. This identified 269 genes as transcription factors or transcriptional regulators in BC-16, with between 236 and 249 genes identified in the other sequenced *P. fragariae* and *P. rubi* isolates. Further analyses of these genes showed no presence/absence variation in the orthology data and no variant sites at or near these genes. However, when these genes were included in the candidate calling process used for assessing avirulence genes, some were identified as possible candidates for explaining the expression differences. No genes were identified as high confidence in any isolate, although a small number were identified as medium confidence in each isolate. BC-16 had the most with 12 candidates (Table 5.17). Unfortunately, with the data available no further analysis was possible.

**Table 5.17: Numbers of candidate transcription factors identified in the BC-1, BC-16 and NOV-9 isolates**

Isolate	BC-1	BC-16	NOV-9
High Confidence	0	0	0
Medium Confidence	1	12	2
Low Confidence	6	7	10
Total	7	19	12

## 5.5 Discussion

### 5.5.1 Identification of variant sites allowed the resolution of a population structure with *Phytophthora fragariae*

Whilst previous work with a significantly larger sample size has been conducted for *P. rubi* and *P. fragariae*, the analysis focused on *P. rubi* (Tabima *et al.*, 2018). Additionally, Tabima *et al.* (2018) used a relatively fragmented reference assembly for variant calling. The results from this study called variant sites with respect to a novel, larger and more contiguous FALCON assembly of the BC-16 isolate of *P. fragariae* (see Chapter 3, section 3.4.3.2) with *P. rubi* acting as an outgroup species. Analysis of the levels of shared sites between isolates showed a clear population separation between *P. fragariae* and *P. rubi*, confirming the results of Tabima *et al.* (2018) and Man in't Veld (2007; Figure 5.1). Population structure was also assessed as a part of this study. Stepwise removal of isolates from outgroup populations showed a clear population consisting of isolates of the pathogenicity races: UK1, UK2 and UK3 (Fig. 5.3 and 5.4). Interestingly, the SCRP245 isolate was scored as an intermediate between the UK1-2-3 population and the population represented by the BC-23 and ONT-3 isolates. One possibility was that SCRP245 represented a rare hybrid of these two populations, however, due to the small sample size of isolates and the low number of SNP sites available for this analysis it appeared more likely that SCRP245 represented a separate population but the data available was unable to resolve this fully.

### 5.5.2 Investigation of variant sites within the UK1-2-3 population showed no sequence differences correlating with pathogenicity race

As no presence/absence variation was found to explain the differences in pathogenicity between isolates of races UK1, UK2 and UK3 through an orthology analysis (see Chapter 4, section 4.4.10), it was of interest to investigate whether there were small sequence differences that would result in a difference in protein sequence and possibly affect the ability of the host plant to detect this protein and trigger the resistance response, following the gene for gene model of resistance

proposed by van de Weg (1997b). Investigations of private sites between these isolates gave only private variants in isolates of race UK2 with eleven private sites identified. The lack of sites private to UK1 and UK3 was unexpected, though it may represent biological reality. The sequencing of additional isolates of races UK1 and UK3 may further elucidate this. However, these variant sites were not found to be either near or within genes involved in the pathogenicity process (Table 5.5).

### **5.5.3 Population genetic analyses showed the UK1-2-3 population was mostly clonally propagated with occasional sexual recombination and showed low levels of variation, likely due to a recent genetic bottleneck**

Since the population structure analysis had identified the UK1-2-3 population as a distinct group, it was of interest to investigate this population further to understand the population dynamics.

Given that the number of variant sites appeared to be low, it was of interest to investigate whether there was evidence of sexual recombination in the UK1-2-3 population. LD decay analysis showed an LD<sub>50</sub> value of 26,500 bp (Figure 5.5). This was slightly larger than the value for *S. pombe* (Jeffares *et al.*, 2015), which suggested a reproductive strategy consisting of clonality with some sexual reproduction (Nieuwenhuis and James, 2016). Only one hotspot of recombination was identified, though no genes within this region appeared to be involved in the pathogenicity process, due to the presence of transmembrane helices. Since there was evidence of some sexual reproduction within the population, it was of interest to investigate evidence for historical recombination between the UK1-2-3 population and the population represented by the BC-23 and ONT-3 isolates as well as between the *P. fragariae* and *P. rubi* species. There appeared only to be weak evidence for historical recombination between the UK1-2-3 population and the BC-23 and ONT-3 population, compared to the evidence of historical recombination between the *P. fragariae* and *P. rubi* species. However, this may be due to there being far fewer SNPs being available for the comparison of UK1-2-3 to BC-23 and ONT-3 than for the comparison between *P. fragariae* and *P.*

*rubri*. This analysis was conducted using the four gamete test (Hudson and Kaplan, 1985). This test relied on the infinite-sites assumption for calculation. As the recombination rate in UKR1-2-3 was very low, it was likely that the infinite-sites assumption was violated, casting doubt on the reliability of these results. Nevertheless, it did still show some evidence of historical recombination between the populations, though the exact age of the population separation remained unclear.

Population separation was found to vary across the genome, some regions showed no population separation and others showed extremely high population separation. Therefore, it was of interest to investigate whether regions which showed high separation contained genes with specific functions that might explain the population separation. Whilst this analysis has not previously been performed in *Phytophthora* spp., it has recently been performed on sunflowers (*Helianthus annuus*), though with a different threshold value used (Owens *et al.*, 2018). Only a small number of genes (49) were reliably identified as showing evidence of population separation. Despite this, apoplastic effectors and genes annotated with the “Proteolipid membrane potential modulator” InterProScan term were shown as statistically significant as being in regions of high population separation (Table 5.7). As previously discussed, apoplastic effector identification likely had a high false positive rate (see Chapter 4), therefore it appeared unlikely these genes were driving the population separation. It was also unclear how the “Proteolipid membrane potential modulator” InterProScan term would cause population separation and so this observation may not be causal.

In order to further investigate these two populations, levels of variation were assessed per gene. This showed that the BC-23 and ONT-3 population had higher levels of nucleotide diversity ( $\pi$ ) than the UKR1-2-3 population, despite a smaller sample size. However, there was a slightly larger distribution of both Watterson’s  $\theta$  and the raw frequency of segregating sites in the BC-23 and ONT-3 population than in the UKR1-2-3 population, potentially influencing the values of  $\pi$ . This potentially suggested that the UKR1-2-3 population had undergone a genetic bottleneck.

Investigations of selection in the UKR1-2-3 population showed the majority of the predicted genes in BC-16 had no evidence of selection and instead appeared to be neutrally evolving compared to the *P. rubi* outgroup. The McDonald-Kreitman test and Fay and Wu's H did show a small number undergoing diversifying selection and the Fay and Wu's H results indicated a low number of genes undergoing purifying selection. These results, combined with the Fu and Li's  $F^*$  and  $D^*$  results suggested the presence of a recent selective sweep or genetic bottleneck for this population, with the beginnings of a recovery being detected by Fu and Li's  $F^*$  and  $D^*$ . An analysis of genes showing negative values of Fu and Li's  $F^*$  and  $D^*$  failed to show any classification of gene being overrepresented in genes undergoing diversifying selection. However, this recovery did not appear to be causing diversification of predicted proteins, as evidenced by the low value for the ratio of nucleotide diversity at nonsynonymous sites ( $\pi_{ns}$ ) to nucleotide diversity at synonymous sites ( $\pi_s$ ),  $\pi_{ns}/\pi_s$ . Although the evidence for the historical bottleneck or sweep appeared robust, it was unclear what may have caused this. The apparent difference in the number of genes which showed true neutral selection between the MKT result and the Fay and Wu's H and the Fu and Li's  $F^*$  and  $D^*$  results was likely due to a stark decrease in the number of genes for which these statistics could be calculated, from 18,355 genes in the MKT analysis (with 1,488 genes with values  $< 1$ ), to 788 genes for the Fu and Li's statistics and 604 genes for the Fay and Wu's H analysis.

In summary, the UKR1-2-3 population showed a reproductive strategy of clonality with occasional sexual reproduction. However, there was no clear evidence of recombination hotspots. There was also no clear mechanism revealed for the population separation, this suggested that the separation could be due to factors outside of the genome, such as host preference or environmental factors. There was also evidence of a recent genetic bottleneck or clonal expansion driven by selection in the UKR1-2-3 population, though the cause for this was not clear. It may again have been due to an environmental factor or due to host preference, though this was unclear.



#### 5.5.4 Additional RNA-Seq data allowed for the identification of candidate avirulence genes

The expression levels of the two genes of interest from Chapter 4: g7404.t1 and g36121.t1 were investigated in order to assess whether they appeared to be candidate avirulence genes. Whilst the peak FPKM value in BC-16 *in planta* time points was always higher than that in BC-1 and NOV-9 *in planta* time points, it was not a large enough difference for either gene to be called as differentially expressed, nor were these genes present in the results of the candidate calling process. Hence, they were rejected as candidate avirulence genes.

A new set of candidate avirulence genes were identified through a process that analysed uniquely expressed genes and uniquely differentially expressed genes. The inclusion of uniquely differentially expressed genes reduced the number of false positive candidates by removing genes which showed FPKM values only slightly above the FPKM threshold for expression in one isolate but not in the other isolates. This resulted in the identification of seventeen high confidence avirulence genes in BC-16, three medium confidence avirulence genes in BC-1 and forty-seven medium confidence avirulence genes in NOV-9. These identified genes will be useful for future analyses. However, the low number of expressed genes identified from the alignment of BC-1 reads against NOV-9 and NOV-9 reads against BC-1 explained the lack of high confidence candidates (Table 5.10). This was due to the alignment of the reads to intergenic regions, though it raised questions about the accuracy of the alignments or gene prediction processes. The results of the RNA-Seq analyses were confirmed by RT-qPCR for a high-quality candidate avirulence gene in BC-16 (g27513.t1), an RxLR that appeared to peak at 24 hpi in BC-16 (g32018.t1) and an RxLR effector that appeared to peak at 48 hpi in BC-16 (g23965.t1). Whilst there was a large amount of variation between biological samples, the expression of the gene in BC-16 and A4 *in planta* timepoints, alongside the lack of expression of g27513.t1 in BC-1 and NOV-9 samples, as well as mycelial samples for BC-16 was confirmed.

Therefore, it appeared that the difference in pathogenicity phenotype for the UKR1-2-3 population was controlled at the transcriptional level. This was similar to recent results for two isolates of *P. infestans* from the same population (Pais *et al.*, 2018) and so provided additional evidence that transcriptional variation can be a key determinant of pathogenicity in *Phytophthora* spp..

#### **5.5.5 Investigations of a high-quality candidate avirulence gene in BC-16 failed to show a cause for the lack of expression in BC-1 and NOV-9**

Following the identification of the high quality avirulence gene candidate in BC-16, investigations were carried out to identify any variation in *cis* or *trans* of this gene that might explain the abolition of the expression of its orthologous genes in BC-1 and NOV-9. Additional sequencing of NOV-9 with nanopore sequencing, developed by ONT, allowed the identification of sequence variants, although an upstream INDEL appeared to be a sequencing error that had remained uncorrected due to the poor mapping of Illumina reads to a transposon ORF, a SNP from T in BC-16 to G in NOV-9 was identified 14 Kb downstream of the stop codon. Whilst this could be involved, perhaps though mutation of an enhancer binding site, it would be unlikely to result in the total abolition of expression of this gene.

An investigation was also conducted on transcription factors and transcriptional regulators to identify changes in *trans*. Whilst none of these genes were identified as candidates through orthology analyses and the investigation of variant sites, some were detected as candidates using the same process used to identify candidate avirulence genes. The reasons for this variation remained unclear. As attempts to identify a promoter sequence failed, it was not possible to identify which transcription factors and transcriptional regulators were involved in the expression of this gene with the current data.

It was possible that these changes in expression may have been due to the possession of different epialleles in these isolates. Recently, the addition of a methyl group to the sixth carbon on adenosines (A) in DNA, N6-methyladenine

(6mA) DNA methylation has been demonstrated in the clade 1 species *P. infestans*, whereas the addition of a methyl group to the fifth carbon on cytosines (C) in DNA, 5-methylcytosine (5mC) methylation has not been detected in this species (van West *et al.*, 2008; Chen *et al.*, 2017). It was also possible that histone modifications or chromatin remodelling may be involved in the variation in expression levels (van West *et al.*, 2008). However, this was not analysed in the present study.

### 5.5.6 Conclusions

The work in this chapter resulted in the development of a novel panel of variant sites for *P. fragariae* and *P. rubi*. Whilst the number of isolates used was lower than a recent study (Tabima *et al.*, 2018), this work has the advantage of using a more contiguous genome of higher completeness as a reference. Analysis of these variant sites showed no private variants that could explain the race structure. Interestingly, private variants were only identified for the UK2 race, though the cause of this was unclear. Population genetic analyses showed the isolates of races UK1, UK2 and UK3 formed a distinct population, called the UK1-2-3 population. Further population genetic analyses showed this population reproduce clonally with occasional sexual reproduction, similar to *S. pombe* (Jeffares *et al.*, 2015; Nieuwenhuis and James, 2016). This population was also shown to have low levels of variation and evidence of a recent genetic bottleneck, though the source of this was unclear. Some evidence of diversifying selection was identified, which suggested the beginnings of a recovery, though it did not appear to be acting on a specific gene class. The population separation between the UK1-2-3 population and the population represented by the BC-23 and ONT-3 isolates was also investigated, whilst apoplastic effectors and the 'Proteolipid membrane potential modulator' were significantly enriched in genes with high population separation, it was doubtful that these truly explained the separation. Therefore, the population separation may have been due to host preference or environmental factors. However, these results were possibly due to a lack of SNPs in the UK1-2-3 population, as the amount of biallelic SNP sites available for analysis was reduced to 2,552 from 401,009 for all sequenced *P. fragariae*

and *P. rubi* isolates. A set of candidate avirulence genes were then identified through the use of additional RNA-Seq data for BC-1 (race UK1) and NOV-9 (race UK3). This data resulted in the elimination of genes of interest from Chapter 4, but did identify additional candidates for each of the UK1, UK2 and UK3 pathogenicity races. A particularly high-quality candidate was identified for BC-16 (g27513.t1) and was confirmed by RT-qPCR. This candidate gene was further investigated and did not show any clear changes in *cis* or *trans* that may explain the expression pattern. Further investigation of this candidate gene, along with the other highly ranked candidates, would be of great interest to increase understanding of the race structure of *P. fragariae*.

## Chapter 6: General Discussion

### 6.1 Key findings

The studies contained in this thesis aimed to improve the understanding of the race structure explaining the observed variation in pathogenicity of different isolates of the *Phytophthora* clade 7a species *Phytophthora fragariae* against a variety of cultivars of strawberry (*Fragaria × ananassa*) plants hypothesised to contain different resistance (*R*) genes. Using the gene for gene model of resistance first proposed by Van de Weg (1997b), several isolates were classified into the U.K. pathogenicity race scheme, in order to correctly relate to hypothesised *R* genes. Taking advantage of the ever-reducing costs of genome sequencing technologies, genome assemblies of numerous isolates of *P. fragariae* and the closely related raspberry (*Rubus idaeus*) pathogen *Phytophthora rubi* were assembled and annotated. This included the development of a reference assembly of the BC-16 isolate of *P. fragariae* utilising long-read PacBio single molecule real time (SMRT) data. Annotation of these genomes was performed to provide predicted gene sets for each sequenced isolate. The annotation process required the development of a novel method of inoculation of micropropagated strawberry plants in order to allow for the generation of RNA-Seq data, due to the failure of zoospore suspension production methods used for other *Phytophthora* spp.. Effector genes were identified through a variety of methods in order to provide a set of genes of interest to be mined for candidate avirulence genes by several methods, which led to the identification of candidate avirulence genes at three confidence levels. Additionally, a novel panel of genome wide variant sites were identified using the BC-16 reference genome. Population genetic analyses were then conducted to investigate the population dynamics of these isolates. This work represents a significant advance in the bioinformatic exploration of this economically important plant pathogen species.

## **1. Phenotyping of isolates led to the conversion of three Canadian and one American race to the U.K. race scheme**

Although pathogenomic studies have become a key tool for the understanding of the molecular control of the resistance to plant pathogens, the phenotypic classification of examined isolates must first be confirmed. Although *P. fragariae* has been demonstrated to exhibit a gene for gene resistance model with *F. × ananassa* (Van de Weg, 1997b), there are several contradictory race schemes in existence (eg. CA2 is not necessarily equivalent to UK2). The *P. fragariae* isolates used in this study were mostly assigned using the Canadian race scheme, although A4 was assigned using the U.S.A. scheme. Prior to the commencement of this study, Dr. Charlotte F. Nellist characterised the pathogenicity of three isolates of Canadian races 1, 2 and 3 (BC-1, NOV-9 and BC-16) as U.K. races 1, 3 and 2 respectively. Following this, work described in Chapter 3 involved the characterisation of the remaining isolates received with Canadian race designations of 1, 2 and 3, along with the A4 isolate, which was anecdotally thought to be U.K. race 2. A pathogenicity test on rooted runners of plants of four cultivars with varying resistance gene complements allowed the assignment of these isolates to their respective U.K. races and supported the existence of a gene-for-gene model of resistance (Table 3.2). The results of this pathogenicity test will aid in investigations working to map the resistance genes present in *F. × ananassa*, as well as informing further investigations in this study by providing robust definitions of the pathogenicity race of a selection of the sequenced isolates.

## **2. Assembly of ten *P. fragariae* and three *P. rubi* isolates with Illumina sequencing data produced similar genome assemblies**

Prior to the commencement of this study, a single genome assembly of *P. fragariae* was publicly available (Gao *et al.*, 2015). During the course of this work, an additional assembly was announced and made publicly available (Tabima *et al.*, 2017). Unfortunately, neither author of the previous genome announcements provided any information about the pathogenicity race of the isolates they sequenced. The work in this thesis produced a large number of *de novo*

assemblies of isolates of both *P. fragariae* and *P. rubi*, providing a significant increase to the resources available for the study of these important pathogens. For the Illumina-only genome assemblies, the quality of the assemblies was comparable to the assemblies in the genome announcements of Gao *et al.* (2015) and Tabima *et al.* (2017). There was no clear differentiation between the assemblies of the isolates generated in this thesis, however it is not possible to dismiss the possibility that differing genomic elements may have been involved in the control of pathogenicity, as shown in a number of *Fusarium* spp. (Ma *et al.*, 2010; Croll and McDonald, 2012).

### **3. Construction of a reference assembly of the BC-16 isolate of *P. fragariae* with PacBio sequencing data resulted in improved contiguity**

Additionally, the BC-16 isolate of *P. fragariae* was sequenced using the long read PacBio single molecule real time (SMRT) technology. Long read sequencing technologies are becoming more common as a tool to aid in the assembly of difficult to assemble plant pathogens, such as the broad host range clade 1 species *Phytophthora nicotianae* (Liu *et al.*, 2016) and the fungal pathogen of wheat *Puccinia striiformis* f. sp. *tritici* (Schwessinger *et al.*, 2018). Through the use of an, at the time, experimental diploid-aware assembly programme FALCON-Unzip (Chin *et al.*, 2016), approximately 12 Mb of additional repeat rich regions were assembled into 180 contigs, increasing the assembly size from 79.3 Mb to 91.0 Mb. However, the contiguity was far short of the estimated chromosome number of 10 - 12 (Brasier *et al.*, 1999). This novel reference assembly represented a significant improvement in both assembly size and contiguity and approached the predicted genome size of 96 Mb as assessed from the raw Illumina sequencing reads. This improved assembly was similar in size to the well-studied clade 7b species *Phytophthora sojae* (Tyler *et al.*, 2006). There was a slight decline in assessments of BUSCO completeness in the PacBio assembly compared to the Illumina-only assembly, likely due to the higher inherent error rate of PacBio sequencing compared to Illumina, despite attempts at error correction (Rhoads and Au, 2015). Alternatively, it was possible that there may have been erroneous predictions of BUSCO genes in the Illumina assemblies. It could also have been due to differences in sequencing biases

between the different technologies, as it has previously been shown that the shorter read Illumina technologies underrepresents GC-rich/poor sequences, which PacBio sequencing avoids at the expense of a higher error rate and a bias towards larger DNA fragments (Ardui *et al.*, 2018).

#### **4. A novel *in vitro* inoculation method allowed for the generation of RNA-Seq data to annotate the genomes and predict putative effector genes**

In order to conduct investigations to identify candidate avirulence genes, it was required to predict gene models from these assemblies. In order to generate RNA-Seq data for use in gene predictions, attempts were made to generate a zoospore suspension of *P. fragariae* to conduct *in vitro* inoculations of strawberry (*F. × ananassa*) plants. Unfortunately, methods such as simply incubating mycelium on media under water, as with *P. sojae* (formerly *Phytophthora megasperma* f. sp. *glycinea*; Ward *et al.*, 1989), were insufficient to promote zoospore release in *P. fragariae*. Therefore, a method for the inoculation of micropropagated plants via uncontrolled zoospore release was developed, as described in Chapter 5. This inoculation method allowed the generation of RNA-Seq data from the period of the pathogen life cycle of interest to this study. The RNA-Seq data from the infection time course, alongside data from *in vitro* mycelium was used to predict gene models. Genes were predicted in *P. rubi* using data provided by the Grünwald Lab at Oregon State University as part of the *Phytophthora* Sequencing Consortium. Additional low confidence gene models were predicted as in Armitage *et al.* (2018b), though these were only included in the annotations if they were identified as putative effector genes. Key effector gene classes were then identified through the use of a variety of computational models. These effector gene classes were: RxLR effectors, Crinkler effectors and Apoplastic effectors. Effectors were also predicted from unguided gene models; however, these were treated with caution, due to statistical issues stemming from the nature of predicting functional classifications of genes, particularly for the machine learning method utilised by ApoplastP (Pritchard and Broadhurst, 2014; Sperschneider *et al.*, 2018) and the lack of prediction of the unguided gene models by methods making use of RNA-Seq data. However, all predicted effector genes were kept in the annotations in order



to ensure all possible effector genes were captured. Differential expression analyses showed a large scale transcriptional reprofiling upon infection of the host plant, with changes in different time points post-inoculation which demonstrated that distinct stages of the infection process had been captured. Therefore, a selection of novel annotated assemblies of both *P. fragariae* and *P. rubi* were produced, including a novel reference assembly of the BC-16 isolate (Fig. 4.16).

Recently, a method has been developed for *Phytophthora infestans* and *P. sojae* to identify effector genes involved in pathogenicity differences without performing whole genome sequencing. This method is known as pathogen enrichment sequencing (PenSeq) and involves the enrichment of genomic DNA for motifs associated with effector genes, in this case RxLR effectors (Thilliez *et al.*, 2018). Whilst this method would have allowed for a high throughput, cost effective determination of differences in RxLR effectors, it would potentially miss other classes of effector or variation due to genomic rearrangements.

## **5. Investigations of copy number variation, presence/absence variation and sequence variants allowed for the investigation of avirulence genes, yet failed to clearly identify candidate avirulence genes**

Previous studies in *P. sojae* have demonstrated several avirulence genes, many of which showed copy number variation (CNV; Dong *et al.*, 2009; Qutob *et al.*, 2009). In order to identify possible genes of interest showing CNV by assessment of average read depth, the raw reads of the U.K. race 1 isolate BC-1 and the U.K. race 3 isolate NOV-9 of *P. fragariae* were used in conjunction with the Illumina reads of the U.K. race 2 isolate BC-16 and the novel reference assembly generated in this study and normalised average read depth was assessed using a method which identified avirulence genes showing CNV in *P. infestans* (Raffaele *et al.*, 2010; Cooke *et al.*, 2012; Pais *et al.*, 2018). This identified a single gene of interest (g7404, a low confidence gene model predicted as an apoplastic effector) which showed increased copy number in BC-1, decreased copy number in BC-16 and at an intermediate value in NOV-9. In order to identify genes showing presence/absence variation between the races, an orthology

analysis was conducted on all predicted proteins from *P. fragariae* and *P. rubi*. Investigation of isolates of races UK1, UK2 and UK3 identified a single gene which potentially showed presence/absence variation, alongside expression evidence, which made it a candidate for the avirulence gene of race UK2 (g36121, a low confidence gene model predicted as an apoplastic effector). As discussed in Chapter 4, this appeared to be due to an assembly error in the U.K. race 1 and U.K. race 3 isolates as the Illumina reads of BC-1 and NOV-9 gave high levels of coverage for this gene. Variant sites including: single nucleotide polymorphisms (SNPs), small insertions and deletions (INDELs) and larger structural variants (SVs) were also identified through the alignment of sequencing reads to the BC-16 genome, however, there were only private variants in race UK2, and none were near or within genes that were potentially involved in the pathogenicity process, therefore the race typing phenotype was likely not controlled by nucleotide sequence variations.

## **6. Generation of additional RNA-Seq data for BC-1 and NOV-9 allowed the identification of candidate avirulence genes for race UK1, UK2 and UK3 along with a strong candidate for UK2.**

Following the generation of additional RNA-Seq data from BC-1 and NOV-9, these two genes of interest (g7404 and g36121) were no longer considered as candidate avirulence genes due to the level of expression being the same in all three sequenced isolates. As described in Chapter 5, genes which showed unique expression (Fragments Per Kilobase of transcript per Million mapped reads; FPKM > 5) and unique differential expression ( $\log_2$  fold change  $\geq 3$ ) in a single isolate were scored as candidate avirulence genes at a variety of confidence levels. This resulted in the identification of 17 high confidence candidates for the avirulence gene in race UK2, three medium confidence candidates for the avirulence gene in race UK1 and 47 medium confidence candidates for the avirulence gene in race UK3 (Table 5.14). As discussed in Chapter 5, issues with the alignment of BC-1 reads to the NOV-9 assembly and NOV-9 reads to BC-1 assembly likely explained the lack of high confidence candidates identified in BC-1 and NOV-9. A particularly strong candidate for the avirulence gene from race UK2, g27513.t1, was identified through manual

inspection of the candidates and confirmed by RT-qPCR. It therefore appeared that for the races UK1, UK2 and UK3, pathogenicity was controlled by transcriptional changes, similar to observations by Pais *et al.* (2018) on two isolates of differing pathogenicity phenotypes from the same population of *P. infestans*. These results therefore provided additional evidence that transcriptional variation can be a key determinant of pathogenicity in *Phytophthora* spp..

### **7. Investigation of the strong candidate for the avirulence gene in UK2 failed to identify clear *cis* or *trans* changes controlling the expression differences**

Further investigation of g27513, the strong candidate avirulence gene in BC-16 failed to show conclusive variation in *cis*. However, the presence of a SNP from T in BC-16 to G in NOV-9 was shown 14 kb downstream of the stop codon. This could potentially be involved in the expression polymorphism through the mutation of an enhancer site. Identification of a set of transcription factors and transcriptional regulators again did not show any clear variation that could have explained the expression level polymorphism, although some were identified as potentially involved via the same candidate calling process used for effector genes, as described in Chapter 5. However, as attempts to identify a promoter sequence were unsuccessful, it was not possible to identify whether any of these possible candidates may be involved. There was also no evidence for the presence of an effector family for g27513 in BC-16. It remained possible that the expression polymorphisms may have been due to epigenetic modification. However, little is known of this in *Phytophthora* spp., although recently the addition of a methyl group to the sixth carbon on adenosines (A) in DNA, N6-methyladenine (6mA) DNA methylation has been demonstrated in the clade 1 species *P. infestans*, whereas the addition of a methyl group to the fifth carbon on cytosines (C) in DNA, 5-methylcytosine (5mC) methylation has not been detected in *P. infestans* (van West *et al.*, 2008; Chen *et al.*, 2017). It was also possible that histone modifications or chromatin remodelling may be involved in the expression level variation (van West *et al.*, 2008).

**8. The identified variant sites were used to investigate a population of isolates consisting of isolates of race UK1, UK2 and UK3 with a mixed reproductive system and low levels of diversity which likely underwent a recent genetic bottleneck**

Following the identification of variant sites as described in Chapter 5, population genetics analyses were conducted. This allowed confirmation of the species separation of *P. fragariae* and *P. rubi* as shown by Man in't Veld (2007) and Tabima *et al.* (2018). The isolates of races UK1, UK2 and UK3 were also shown to form a single distinct population (named as the UK1-2-3 population), as did BC-23 and ONT-3, and a phylogenetic tree was constructed (Fig. 5.2). Population genetics analyses detailed in Chapter 5 examined the population dynamics of the UK1-2-3 population. Investigations into linkage disequilibrium (LD) decay showed a mixed reproductive strategy of clonality with some sexual reproduction, similar to *Schizosaccharomyces pombe* (Jeffares *et al.*, 2015; Nieuwenhuis and James, 2016). There was evidence of historical recombination between both the UK1-2-3 population and the BC-23 and ONT-3 population, as well as between *P. fragariae* and *P. rubi*, though this investigation utilised the four-gamete test which relies on the infinite-sites assumption which may well be violated due to the low recombination rate in this population (Hudson and Kaplan, 1985). There was evidence of population separation between the UK1-2-3 population and the BC-23 and ONT-3 population, potentially driven by apoplastic effectors and genes with the "Proteolipid membrane potential modulator" InterProScan term. Though the potential high error rate of ApoplastP due to inherent errors in machine learning models (Pritchard and Broadhurst, 2014; Sperschneider *et al.*, 2018) and an unclear mechanism for the InterProScan term to drive diversification cast doubt on these results. Investigation of selection and nucleotide diversity suggested that the population had recently undergone a genetic bottleneck or selective sweep due to alleles reaching fixation, though there was some amount of diversifying selection suggesting the beginning of a recovery. However, no genes were overrepresented in regions undergoing diversifying selection.

## Conclusion

In conclusion, the UKR1-2-3 population displayed clear separation from the other isolates of *P. fragariae* analysed in this study. This population appeared to have undergone a recent genetic bottleneck, with some evidence of recovery. The cause for this was, however, unclear. Within this population, the differences in pathogenicity appeared to be transcriptional in nature, similar to what has been shown in *P. infestans* by Pais *et al.* (2018). A set of candidate avirulence genes were identified for each of these three pathogenicity races, these will be useful for further investigations to aid in the breeding of resistance, following functional validation of these genes and the identification of the resistance genes, to this economically important pathogen. However, as control of virulence for these three races appeared to be transcriptional, it appeared doubtful that a field kit to identify which race is present in the field could be developed as part of a surveillance strategy, in order to aid in cultivar choice and breeding targets. The additional sequencing data generated as a part of this study will also contribute to wider work on the *Phytophthora* genus, leading to an increased understanding of a variety of plant pathogens that are highly destructive in both agricultural and wild systems.

## 6.2 Future directions

Improving the understanding of the genetic control of pathogenicity of *P. fragariae* will aid in the breeding of resistant cultivars of *F. × ananassa* through the identification of pairs of resistance genes and avirulence genes, followed by the pyramiding of the appropriate resistance genes. However, in order for these benefits to materialise, further study of the mechanisms driving the gene for gene model of resistance will be necessary. Additionally, it will be crucial to understand the distribution of pathogenicity races within different geographic areas and the ability of these pathogenicity races to move to new areas and controls that can be applied to reduce these risks.

The work in this thesis provided the first in depth investigation of the genomics of *P. fragariae* and resulted in novel insights into the control of virulence. Parallels

were drawn to other species of *Phytophthora*, though the UKR1-2-3 population did display a striking lack of diversification despite containing three different pathogenicity races and being isolated from different locations in Canada and the U.S.A.. Potential further work would first require the rationalisation of the divergent race typing systems, followed by the acquisition of additional sequencing data and isolates from a variety of geographic locations in order to generate hypotheses about the global state of this devastating plant pathogen and how best to reduce its impact in agricultural systems.

### **Determine the pathogenicity races of all remaining isolates in the NIAB EMR collection**

As Table 6.1 demonstrated, there are still many gaps in the conversion table between the three race typing schemes encountered in this study. Of particular interest would be race typing SCRP245, as this was isolated a lot earlier than the other isolates and could perhaps provide insight about the changes in pathogenicity races that have occurred throughout the recent evolutionary history of this pathogen following its identification in the 1920s.

### **Sampling of additional isolates in a rationalised scheme from specific geographical areas**

One of the major issues encountered in this study was the small sample size of isolates available. Whilst this pathogen is comparatively difficult to isolate, it would be of interest to sample the diversity of the population of this pathogen over time. It would also potentially be of interest to limit sampling to a distinct geographic area, for example the U.K.. However, assuming enough samples could be collected to ensure robust analyses, a wider scale investigation within Europe or on a global scale would provide greater insight.

### **Long read sequencing of additional isolates of *P. fragariae***

The assemblies described in this chapter which made use of long read sequencing data assembled larger amounts of the genome with greatly improved

contiguity. Whilst these regions were repeat rich, it was possible that effector genes may have resided in these regions, particularly as in *P. infestans* repeat rich regions are shown to have a greater amount of effector genes within them than the genome as a whole, described as the two-speed genome model (Dong *et al.*, 2015). Hence, assembly of these regions may provide additional candidate avirulence genes.

### **Additional sequencing with longer range technologies to further improve the reference assembly**

Whilst the reference genome presented in this thesis represented a significant improvement over the currently available reference assemblies of *P. fragariae*, it still fell short of a complete, chromosome level assembly. Whilst ONT data appeared to improve the assembly when roughly assembled with default parameters, it was unclear whether the improvement in contiguity would remain following parameter optimisation to avoid the erroneous duplication of repeat rich regions. Hence, the use of sequencing technologies such as BioNano Genomics optical mapping or Dovetail Genomics chromosome conformation capture data could further improve the contiguity of the assembly, as described by Jiao *et al.* (2017) for the plant species *Arabis alpina*, *Euclidium syriacum* and *Conringia planisiliqua*. However, the short lengths of the smaller contigs in the reference assembly presented here may limit the effectiveness of BioNano Genomics data due to technical constraints of this method. These technologies may also improve the accuracy of the assemblies by removing possible misassemblies from PacBio data.

### **Perform additional validation of gene models**

As the genomes presented here were annotated automatically, it was possible that some gene models may have been spurious predictions and that true genes may have been missed. This is a common problem with the annotation of non-model species and large-scale sequencing projects tend to employ a manual curation approach. This would improve the robustness of the orthology analyses presented in this study. However, with 478,569 gene models predicted to

produce 488,156 mRNA transcripts across the fourteen sequenced isolates, this represents a formidable task. Despite this, additional RNA-Seq data from a variety of life stages may aid in improving the quality of the annotations.

### **Generate additional RNA-Seq data from additional time points and isolates**

Although this work generated candidate avirulence genes for the UK1, UK2 and UK3 pathogenicity races, this determination relied upon the use of only a single time point post inoculation in the UK1 and UK3 races. The sequencing of additional time points would improve the robustness of this determination. The sequencing of additional isolates of these races would also improve robustness. Additionally, the RT-qPCR data emphasised the variation in biological replicates from the inoculation experiments conducted here. This was likely due to the imprecision inherent in the inoculation method and as such further refinement of the inoculation procedure would also improve the robustness of the identification of candidate avirulence genes.

### **Validation of candidate avirulence genes through molecular methods**

As the number of candidates generated at the highest confidence levels for each isolate was not excessively large (a maximum of forty-seven in NOV-9), it would be of interest to characterise these genes through molecular methods. However, this is hampered by the current knowledge and techniques available. As the resistance genes in *F. × ananassa* remain unidentified, a susceptible cultivar could not be transformed with an additional resistance gene and assessed for the acquisition of resistance. Additionally, no method of transient expression has been developed for *F. × ananassa*, so a cultivar with *Rpf2* could not be transformed with candidate avirulence genes to assess for a hypersensitivity response (HR). Neither could a resistant cultivar have the resistance gene knocked out and assessed for the acquisition of susceptibility, despite the recent demonstration of CRISPR/Cas9-mediated gene editing in *F. × ananassa* (Wilson *et al.*, 2018). There is also currently no available method for the transformation of *P. fragariae*. There has been some progress with the transformation of its close relative *P. rubi*, however this process is still undergoing development (E. Gilroy,



personal communication). Therefore, it is not yet possible to transform an isolate of a different race with a candidate avirulence gene and assess for the loss of virulence, or to mutate a candidate avirulence gene and assess for the acquisition of virulence.

### **Identify epigenetic markers in the region of the strong candidate avirulence gene in BC-16**

Previous work in *P. sojae* has shown that transgenerational silencing of the *Avr3a* effector allowed isolates of the pathogen to gain virulence on plants containing the *Rps3a* gene, though the precise epigenetic variant was not identified (Qutob *et al.*, 2013). As the investigation of variants in both *cis* and *trans* for the candidate avirulence gene for race two did not show a clear cause of the loss of expression, it would be of interest to investigate epigenetic markers to see if epialleles play a role. Whilst both PacBio and ONT sequencing can be used to identify methylation states, they have recently been shown to vary in which sites are called as methylated and how much methylation they detect, hence additional sequencing data would need to be generated for this analysis (McIntyre *et al.*, 2017). Additionally, as the material for sequencing was grown *in vitro*, it was possible that epigenetic markers may have been modified from those present during pathogenicity, as evidenced by the differences in expression in RNA-Seq reads from this condition. Another possibility would be to use Bisulfite sequencing to determine the methylation status of positions of the genome.

### **Identify transcription factors involved in the expression of the strong candidate avirulence gene in BC-16**

Another cause of the change in expression may be a difference in transcription factors driving the expression of the gene. Transcription factors could be identified *ab initio* from a method such as proteomics of isolated chromatin segments (PICh), though this has issues with a low signal-to-noise ratio (Déjardin and Kingston, 2009). However, it must be noted that the binding of a transcription factor does not necessarily mean the protein will be functional in inducing expression.

## References

- Aguayo, J., G. C. Adams, F. Halkett, M. Catal, C. Husson, Z. Á. Nagy, E. M. Hansen, B. Marçais and P. Frey (2012) Strong Genetic Differentiation Between North American and European Populations of *Phytophthora alni* subsp. *uniformis*. *Phytopathology*. **103**(2): 190-199
- Alcock, N. L. and D. V. Howells (1936) The Phytophthora disease of strawberry. *Scientific Horticulture*. **4**: 52-58
- Altschul, S. F., W. Gish, W. Miller, E. W. Myers and D. J. Lipman (1990) Basic local alignment search tool. *Journal of Molecular Biology*. **215**(3): 403-410
- Amouzou-Alladye, E., J. Dunez and M. Clerjeau (1988) Immunoenzymatic Detection of *Phytophthora fragariae* in Infected Strawberry Plants. *Phytopathology*. **78**: 1022-1026
- Andrews S. (2010) FastQC: A quality control tool for high throughput sequence data. Available online at: <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/> (Last accessed, 03/01/2019)
- Ardui, S., A. Ameer, J. R. Vermeesch and M. S. Hestand (2018) Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. *Nucleic Acids Research*. **46**(5): 2159-2168
- Armitage, A. D., A. Taylor, M. K. Sobczyk, L. Baxter, B. P. J. Greenfield, H. J. Bates, F. Wilson, A. C. Jackson, S. Ott, R. J. Harrison *et al.* (2018a) Characterisation of pathogen-specific regions and novel effector candidates in *Fusarium oxysporum* f. sp. *cepae*. *Scientific Reports*. **8**(1): 13530
- Armitage, A. D., E. Lysøe, C. F. Nellist, L. A. Lewis, L. M. Cano, R. J. Harrison and M. B. Brurberg (2018b) Bioinformatic characterisation of the effector repertoire of the strawberry pathogen *Phytophthora cactorum*. *PLoS ONE*. **13**(10): e0202305
- Arnold, D. L., A. Athey-Pollard, M. J. Gibbon, J. D. Taylor and A. Vivian (1996) Specific oligonucleotide primers for the identification of *Pseudomonas syringae* pv. *plasi* yield one of two possible DNA fragments by PCR amplification: evidence for phylogenetic divergence. *Physiological and Molecular Plant Pathology*. **49**(4)

- Aronesty, E. (2013) Comparison of sequencing utility programs. *The Open Bioinformatics Journal*. **7**(1): 1-8
- Arsenault-Labrecque, G., H. Sonah, A. Lebreton, C. Labbé, G. Marchand, A. Xue, F. Belzile, B. J. Knaus, N. J. Grünwald and R. R. Bélanger (2018) Stable predictive markers for *Phytophthora sojae* avirulence genes that impair infection of soybean uncovered by whole genome sequencing of 31 isolates. *BMC Biology*. **16**(1): 80
- Auton, A. and G. McVean (2007) Recombination rate estimation in the presence of hotspots. *Genome Research*. **17**(8): 1219-1227
- Auton, A., S. Myers, G. McVean (2014) Identifying recombination hotspots using population genetic data. *arXiv*. 1403.4264
- Bailey, T. L. (2011) DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics*. **27**(12): 1653-1659
- Baldauf, S. L., A. J. Roger, I. Wenk-Siefert and W. F. Doolittle (2000) A Kingdom-Level Phylogeny of Eukaryotes Based on Combined Protein Data. *Science*. **290**(5493): 972-977
- Bankevich, A., S. Nurk, D. Antipov, A. A. Gurevich, M. Dvorkin, A. S. Kulikov, V. M. Lesin, S. I. Nikolenko, S. Pham, A. D. Prjibelski *et al.* (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology*. **19**(5): 455-477
- Bendtsen, J. D., H. Nielsen, G. von Heijne and S. Brunak (2004) Improved Prediction of Signal Peptides: SignalP 3.0. *Journal of Molecular Biology*. **340**(4): 783-795
- Bergey, C. M. (2012) vcf-tab-to-fasta; <http://code.google.com/p/vcf-tab-to-fasta> (Last accessed, 03/01/2019)
- Bhat, R. G., P. M. Colowit, T. H. Tai, M. K. Aradhya and G. T. Browne (2006) Genetic and Pathogenic Variation in *Phytophthora cactorum* Affecting Fruit and Nut Crops in California. *Plant Disease*. **90**(2): 161-169
- Blackman, L. M., D. P. Cullerne, P. Torreña, J. Taylor and A. R. Hardham (2015) RNA-Seq Analysis of the Expression of Genes Encoding Cell Wall Degrading Enzymes during Infection of Lupin (*Lupinus angustifolius*) by *Phytophthora parasitica*. *PLoS ONE*. **10**(9): e0136899
- Bonants, B., M. Hagenaar-de Weerd, M. van Gent-Pelzer, I. Lacourt, D. Cooke and J. Duncan (1997) Detection and identification of *Phytophthora fragariae*

Hickman by the polymerase chain reaction. *European Journal of Plant Pathology*. **103**(4): 345-355

Bonardi, V., S. Tang, A. Stallmann, M. Roberts, K. Cherkis and J. L. Dangl (2011) Expanded functions for a family of plant intracellular immune receptors beyond specific recognition of pathogen effectors. *Proceedings of the National Academy of Sciences*. **108**(39): 16463-16468

Bos, J. I. B., M. R. Armstrong, E. M. Gilroy, P. C. Boevink, I. Hein, R. M. Taylor, T. Zhendong, S. Engelhardt, R. R. Vetukuri, B. Harrower *et al.* (2010) *Phytophthora infestans* effector AVR3a is essential for virulence and manipulates plant immunity by stabilizing host E3 ligase CMPG1. *Proceedings of the National Academy of Sciences*. **107**(21): 9909-9914

Boutemy, L. S., S. R. F. King, J. Win, R. K. Hughes, T. A. Clarke, T. M. A. Blumenschein, S. Kamoun and M. J. Banfield (2011) Structures of *Phytophthora* RXLR effector proteins: a conserved but adaptable fold underpins functional diversity. *Journal of Biological Chemistry*. **286**(41): 35834-35842

Bradshaw, J. E., B. Pande, G. J. Bryan, C. A. Hackett, K. McLean, H. E. Stewart and R. Waugh (2004) Interval Mapping of Quantitative Trait Loci for Resistance to Late Blight [*Phytophthora infestans* (Mont.) de Bary], Height and Maturity in a Tetraploid Population of Potato (*Solanum tuberosum* subsp. *tuberosum*). *Genetics*. **168**(2): 983-995

Brasier, C. M., D. E. L. Cooke and J. M. Duncan (1999) Origin of a new *Phytophthora* pathogen through interspecific hybridization. *Proceedings of the National Academy of Sciences*. **96**(10): 5878-5883

Browning, S. R. and B. L. Browning (2007) Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering. *The American Journal of Human Genetics*. **81**(5): 1084-1097

Brunner, F., S. Rosahl, J. Lee, J. J. Rudd, C. Geiler, S. Kauppinen, G. Rasmussen, D. Scheel, T. Nürnberger (2002) Pep-13, a plant defense-inducing pathogen-associated pattern from *Phytophthora* transglutaminases. *The EMBO Journal*. **21**(24): 6681-6688

Buitrago-Flórez, F. J., S. Restrepo and D. M. Riaño-Pachón (2014) Identification of transcription factor genes and their correlation with the high diversity of stramenopiles. *PLoS ONE*. **9**(11): e111841

Campbell, A. M., R. P. Moon, J. M. Duncan, S.-J. Gurr and J. R. Kinghorn (1989) Protoplast formation and regeneration from sporangia and encysted zoospores of *Phytophthora infestans*. *Physiological and Molecular Plant Pathology*. **34**(4): 299-307

Chakraborty, M., J. G. Baldwin-Brown, A. D. Long, J. J. Emerson (2016) Contiguous and accurate *de novo* assembly of metazoan genomes with modest long read coverage. *Nucleic Acids Research*. **44**(19): e147

Chaves, S. C., M. C. Rodríguez, M. F. Mideros, F. Lucca, C. E. Núñez and S. Restrepo (2018) Determining Whether Geographic Origin and Potato Genotypes Shape the Population Structure of *Phytophthora infestans* in the Central Region of Colombia. *Phytopathology*. PHYTO-05-18-0157-R

Chen, H. and P. C. Boutros (2011) VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R. *BMC Bioinformatics*. **12**(1): 35

Chen, H., H. Shu, L. Wang, F. Zhang, X. Li, S. Ochola, F. Mao, H. Ma, W. Ye, T. Gu *et al.* (2017) *Phytophthora* methylomes modulated by expanded 6mA methyltransferases are associated with adaptive genome regions. *bioRxiv*. 217646

Chen, X.-R., Y.-P. Xing, Y.-P. Li, Y.-H. Tong and J.-Y. Xu (2013) RNA-Seq Reveals Infection-Related Gene Expression Changes in *Phytophthora capsici*. *PLoS ONE*. **8**(9): e74588

Chin, C.-S., D. H. Alexander, P. Marks, A. A. Klammer, J. Drake, C. Heiner, A. Clum, A. Copeland, J. Huddleston, E. E. Eichler, *et al.* (2013) Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature Methods*. **10**(6): 563-569

Chin, C.-S., P. Peluso, F. J. Sedlazeck, M. Nattestad, G. T. Concepcion, A. Clum, C. Dunn, R. O'Malley, R. Figueroa-Balderas, A. Morales-Cruz, *et al.* (2016) Phased diploid genome assembly with single-molecule real-time sequencing. *Nature Methods*. **13**(12): 1050-1054

Cingolani, P., A. Platts, L. L. Wang, M. Coon, T. Nguyen, L. Wang, S. J. Land, X. Lu and D. M. Ruden (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly*. **6**(2): 80-92

Clarke, J., H.-C. Wu, L. Jayasinghe, A. Patel, S. Reid and H. Bayley (2009) Continuous base identification for single-molecule nanopore DNA sequencing. *Nature Nanotechnology*. **4**(4): 265

Clifford, R. J., M. Mililo, J. Prestwood, R. Quintero, D. V. Zurawski, Y. I. Kwak, P. E. Waterman, E. P. Lesho and P. McGann (2012) Detection of Bacterial 16S rRNA and Identification of Four Clinically Important Bacteria by Real-Time PCR. *PLoS ONE*. **7**(11): e48558

Cooke, D. E. L., L. M. Cano, S. Raffaele, R. A. Bain, L. R. Cooke, G. J. Etherington, K. L. Deahl, R. A. Farrer, E. M. Gilroy, E. M. Goss *et al.* (2012) Genome Analyses of an Aggressive and Invasive Lineage of the Irish Potato Famine Pathogen. *PLoS Pathogens*. **8**(10): e1002940

Couto, D. and C. Zipfel (2016) Regulation of pattern recognition receptor signalling in plants. *Nature Reviews Immunology*. **16**(9): 537

Croll, D. and B. A. McDonald (2012) The Accessory Genome as a Cradle for Adaptive Evolution in Pathogens. *PLoS Pathogens*. **8**(4): e1002608

Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry *et al.* (2011) The variant call format and VCFtools. *Bioinformatics*. **27**(15): 2156-2158

DEFRA (2017) *Horticultural statistics - 2016*. Retrieved from: [https://www.gov.uk/government/uploads/system/uploads/attachment\\_data/file/641111/hort-dataset-31aug17.xlsx](https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/641111/hort-dataset-31aug17.xlsx) (Last accessed, 03/01/2019).

Deorowicz, S., A. Debudaj-Grabysz and S. Grabowski (2013) Disk-based k-mer counting on a PC. *BMC Bioinformatics*. **14**(1): 160

Déjardin, J. and R. E. Kingston (2009) Purification of Proteins Associated with Specific Genomic Loci. *Cell*. **136**(1): 175-186

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

Dodds, P. N. and J. P. Rathjen (2010) Plant immunity: towards an integrated view of plant-pathogen interactions. *Nature Reviews Genetics*. **11**(8): 539

Dong, S., D. Qutob, J. Tedman-Jones, K. Kuflu, Y. Wang, B. M. Tyler and M. Gijzen (2009) The *Phytophthora sojae* Avirulence Locus *Avr3c* Encodes a Multi-Copy RxLR Effector with Sequence Polymorphisms among Pathogen Strains. *PLoS ONE*. **4**(5): e5556

- Dong, S., S. Raffaele and S. Kamoun (2015) The two-speed genomes of filamentous pathogens: waltz with plants. *Current Opinion in Genetics & Development*. **35**: 57-65
- EFSA Panel on Plant Health (PLH) (2014) Scientific opinion on the risks to plant health posed by *Phytophthora fragariae* Hickman var. *fragariae* in the EU territory, with the identification and evaluation of risk reduction options. *EFSA Journal*. **12**(1): 3539
- Emms, D. M. and S. Kelly (2015) OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology*. **16**(1): 157
- Emms, D. M. and S. Kelly (2017) STRIDE: Species Tree Root Inference from Gene Duplication Events. *Molecular Biology and Evolution*. **34**(12): 3267-378
- EPPO (2018) *EPPO A2 List of pests recommended for regulation as quarantine pests - version 2018-09*. Retrieved from: [https://www.eppo.int/ACTIVITIES/plant\\_quarantine/A2\\_list](https://www.eppo.int/ACTIVITIES/plant_quarantine/A2_list), (Last accessed, 03/01/2019)
- Erwin, D. C. and O. K. Ribeiro (1996) *Phytophthora diseases worldwide*. American Phytopathological Society (APS Press)
- Fankhauser, N. and P. Maeser (2005) Identification of GPI anchor attachment signals by a Kohonen self-organizing map. *Bioinformatics*. **21**(9): 1846-1852
- Fay, J. C. and C. I. Wu (2000) Hitchhiking under positive Darwinian selection. *Genetics*. **155**(3): 1405-1413
- Finn, C. E., J. B. Retamales, G. A. Lobos and J. F. Hancock (2013) The Chilean Strawberry (*Fragaria chiloensis*): Over 1000 Years of Domestication. *HortScience*. **48**(4): 418-421
- Fisher, R. A. (1922) On The Interpretation of  $\chi^2$  from Contingency Tables, and the calculation of P. *Journal of the Royal Statistical Society*. **85**(1): 87-94
- Flor, H. H. (1942) Inheritance of pathogenicity in *Melampsora lini*. *Phytopathology*. **32**: 653-669
- Franceschetti, M., A. Maqbool, M. J. Jiménez-Dalmaroni, H. G. Pennington, S. Kamoun and M. J. Banfield (2017) Effectors of Filamentous Plant Pathogens: Commonalities amid Diversity. *Microbiology and Molecular Biology Reviews*. **81**(2): e00066-16

Fu, Y.-X. and W.-H. Li (1993) Statistical Tests of Neutrality of Mutations. *Genetics*. **133**(3): 693-709

Gao, R., Y. Cheng, Y. Wang, L. Guo and G. Zhang (2015) Genome sequencing of *Phytophthora fragariae* var. *fragariae*, a quarantine plant-pathogenic fungus. *Genome Announcements*. **3**(2): e00034-15

Garrison, E. (2012) Vcflib. A C++ library for parsing and manipulating VCF files. Available online at <https://github.com/ekg/vcflib> (Last accessed, 03/01/2019)

Geib, S. M., B. Hall, T. Derego, F. T. Bremer, K. Cannoles and S. B. Sim (2018) Genome Annotation Generator: a simple tool for generating and correcting WGS annotation tables for NCBI submission. *GigaScience*. **7**(4): gij018

Gilroy, E. M., S. Breen, S. C. Whisson, J. Squires, I. Hein, M. Kaczmarek, D. Turnbull, P. C. Boevink, A. Lokossou, L. M. Cano *et al.* (2011) Presence/absence, differential expression and sequence polymorphisms between *PiAVR2* and *PiAVR2-like* in *Phytophthora infestans* determine virulence of *R2* plants. *New Phytologist*. **191**(3): 763-776

Goffeau, A., B. G. Barrell, H. Bussey, R. W. Davis, B. Dujon, H. Feldmann, F. Galibert, J. D. Hoheisel, C. Jacq, M. Johnston (1996) Life with 6000 Genes. *Science*. **274**(5287): 546-567

Goode, P. M. (1956) Infection of strawberry roots by zoospores of *Phytophthora fragariae*. *Transactions of the British Mycological Society*. **39**(3): 367-377

Goodwin, S., J. Gurtowski, S. Ethe-Sayers, P. Deshpande, M. C. Schatz and W. R. McCombie (2015) Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome. *Genome Research*. **25**(11): 1750-1756

Gout, L., I. Fudal, M.-L. Kuhn, F. Blaise, M. Eckert, L. Cattolico, M.-H. Balesdent and T. Rouxel (2006) Lost in the middle of nowhere: the *AvrLm1* avirulence gene of the Dothideomycete *Leptosphaeria maculans*. *Molecular Microbiology*. **60**(1): 67-80

Grabherr, M. G., P. Russell, M. Meyer, E. Mauceli, J. Alföldi, F. D. Palma and K. Lindblad-Toh (2010) Genome-wide synteny through highly sensitive sequence alignment: *Satsuma*. *Bioinformatics*. **26**(9): 1145-1151

Gremillion, K. J. and K. D. Sobolik (1996) Dietary Variability among Prehistoric Forager-Farmers of Eastern North America. *Current Anthropology*. **37**(3): 529-539



- Grubbs, F. E. (1950) Sample Criteria for Testing Outlying Observations. *The Annals of Mathematical Statistics*. **21**(1): 27-58
- Gurevich, A., V. Saveliev, N. Vyahhi and G. Tesler (2013) QUASt: quality assessment tool for genome assemblies. *Bioinformatics*. **29**(8): 1072-1075
- Haas, B. J., S. Kamoun, M. C. Zody, R. H. Y. Jiang, R. E. Handsaker, L. M. Cano, M. Grabherr, C. D. Kodira, S. Raffaele, T. Torto-Alalibo *et al.* (2009) Genome sequence and analysis of the Irish potato famine pathogen *Phytophthora infestans*. *Nature*. **461**(7262): 393-398
- Haas, B. (2010) TransposonPSI: an application of PSI-blast to mine (Retro-)transposon ORF homologies. Available online at <http://transposonpsi.sourceforge.net/> (Last accessed, 03/01/2019)
- Hancock, J. F. (1999) Strawberries crop production science in horticulture. *CABI Publishing, Oxford, UK*. 109-112
- Hayden, E. C. (2014) Technology: The \$1,000 genome. *Nature*. **507**(7492): 294-295
- Hein, I., E. M. Gilroy, M. R. Armstrong and P. R. J. Birch (2009) The zig-zag-zig in oomycete–plant interactions. *Molecular Plant Pathology*. **10**(4): 547-562
- Hickman, C. J. (1941) The Red Core Root Disease of The Strawberry Caused by *Phytophthora fragariae* n.sp.. *Journal of Pomology and Horticultural Science*. **18**(2): 89-118
- Hickman, C. J. (1962) Physiologic races of *Phytophthora fragariae*. *Annals of Applied Biology*. **50**(1): 95-103
- Hickman, C. J. and M. P. English (1951) Factors influencing the development of red core in strawberries. *Transactions of the British Mycological Society*. **34**(2): 223-236
- Hill, W. G. and B. S. Weir (1988) Variances and covariances of squared linkage disequilibria in finite populations. *Theoretical Population Biology*. **33**(1): 54-78
- Hirakawa, H., K. Shirasawa, S. Kosugi, K. Tashiro, S. Nakayama, M. Yamada, M. Kohara, A. Watanabe, Y. Kishida, T. Fujishiro *et al.* (2014) Dissection of the Octoploid Strawberry Genome by Deep Sequencing of the Genomes of *Fragariae* species. *DNA Research*. **21**(2): 169-181
- Hoff, K. J., S. Lange, A. Lomsadza, M. Borodovsky and M. Stanke (2015) BRAKER1: Unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics*. **32**(5): 767-769

- Houseknecht, J. L., S.-O. Suh and J. J. Zhou (2012) Viability of fastidious *Phytophthora* following different cryopreservation treatments. *Fungal Biology*. **116**(10): 1081-1089
- Hudson, R. R. and N. L. Kaplan (1985) Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics*. **111**(1): 147-164
- Hudson, R. R., D. D. Boos and N. L. Kaplan (1992) A statistical test for detecting geographic subdivision. *Molecular Biology and Evolution*. **9**(1): 138-151
- Jackson, T. J., T. Burgess, I. Colquhoun and G. E. St.J. Hardy (2000) Action of the fungicide phosphite on *Eucalyptus marginata* inoculated with *Phytophthora cinnamomi*. *Plant Pathology*. **49**(1): 147-154
- Jeffares, D. C., C. Rallis, A. Rieux, D. Speed, M. Převorovský, T. Mourier, F. X. Marsellach, Z. Iqbal, W. Lau, T. M. K. Cheng *et al.* (2015) The genomic and phenotypic diversity of *Schizosaccharomyces pombe*. *Nature Genetics*. **47**(3): 235-241
- Jiang, R. H. Y., S. Tripathy, F. Govers and B. M. Tyler (2008) RXLR effector reservoir in two *Phytophthora* species is dominated by a single rapidly evolving superfamily with more than 700 members. *Proceedings of the National Academy of Sciences*. **105**(12): 4874-4879
- Jiao, W.-B., G. G. Accinelli, B. Hartwig, C. Kiefer, D. Baker, E. Severing, E.-M. Willing, M. Piednoel, S. Woetzel, E. Madrid-Herrero *et al.* (2017) Improving and correcting the contiguity of long-read genome assemblies of three plant species using optical mapping and chromosome conformation capture data. *Genome Research*. **27**(5): 778-786
- Johnson, M., I. Zaretskaya, Y. Raytselis, Y. Merezuk, S. McGinnis and T. L. Madden (2008) NCBI BLAST: a better web interface. *Nucleic Acids Research*. **36**(suppl\_2): W5-W9
- Jones, J. D. G., R. E. Vance and J. L. Dangl (2016) Intracellular innate immune surveillance devices in plants and animals. *Science*. **354**(6316): aaf6395
- Jones, J. D. G. and J. L. Dangl (2006) The plant immune system. *Nature*. **444**(7117): 323
- Jones, P., D. Binns, H.-Y. Chang, M. Fraser, W. Li, C. McAnulla, H. McWilliam, J. Maslen, A. Mitchell, G. Nuka *et al.* (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics*. **30**(9): 1236-1240

Judelson, H. S., M. D. Coffey, F. R. Arredondo and B. M. Tyler (1993) Transformation of the oomycete pathogen *Phytophthora megasperma* f. sp. *glycinea* occurs by DNA integration into single or multiple chromosomes. *Current Genetics*. **23**(3): 211-218

Jupe, J., R. Stam, A. J. M. Howden, J. A. Morris, R. Zhang, P. E. Hedley and E. Huitema (2013) *Phytophthora capsici*-tomato interaction features dramatic shifts in gene expression associated with a hemi-biotrophic lifestyle. *Genome Biology*. **14**(6): R63

Käll L., A. Krogh and E. L. L. Sonnhammer (2004) A Combined Transmembrane Topology and Signal Peptide Prediction Method. *Journal of Molecular Biology*. **338**(5): 1027-1036

Kamoun, S. (2006) A Catalogue of the Effector Secretome of Plant Pathogenic Oomycetes. *Annual Reviews of Phytopathology*. **44**: 41-60

Katoh, K., K. Misawa, K. Kuma and T. Miyata (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*. **30**(14): 3059-3066

Katoh, K. and D. M. Standley (2013) MAFFT Multiple Sequence Alignment Software Version 7 Improvements in Performance and Usability. *Molecular Biology and Evolution*. **30**(4): 772-780

Kearse, M., R. Moir, A. Wilson, S. Stones-Havas, M. Cheung, S. Sturrock, S. Buxton, A. Cooper, S. Markowitz, C. Duran *et al.* (2012) Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*. **28**(12): 1647-1649

Koren S., B. P. Walenz, K. Berlin, J. R. Miller, N. H. Bergman and A. M. Phillippy (2017) Canu: scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. *Genome Research*. **27**(5): 722-736

Kourelis, J. and R. A. van der Hoorn (2018) Defended to the nines: 25 years of resistance gene cloning identifies nine mechanisms for R protein function. *The Plant Cell*. **30**(2): 285-299

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

- Krzywinski, M. I., J. E. Schein, I. Birol, J. Connors, R. Gascoyne, D. Horsman, S. J. Jones and M. A. Marra (2009) Circos: An information aesthetic for comparative genomics. *Genome Research*. **19**(9): 1639-1645
- Lai, Y. and T. Eulgem (2018) Transcript-level expression control of plant NLR genes. *Molecular Plant Pathology*. **19**(5): 1267-1281
- Langfelder, P. and S. Horvath (2008) WGCNA: an R package for weighted correlation and network analysis. *BMC Bioinformatics*. **9**(1): 559
- Langfelder, P. and S. Horvath (2012) Fast R Functions for Robust Correlations and Hierarchical Clustering. *Journal of Statistical Software*. **46**(11)
- Langmead, B. and S. L. Salzberg (2012) Fast gapped-read alignment with Bowtie 2. *Nature Methods*. **9**(4): 357-359
- Li, H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv*. 1303.3997v2
- Li, H. (2018a) Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*. **34**(18): 3094-3100
- Li, H. (2018b) seqtk: toolkit for processing sequences in FASTA/Q formats. Available online at <https://github.com/lh3/seqtk> (Last accessed, 03/01/2019)
- Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin and 1000 Genome Project Data Processing Subgroup (2009) The sequence alignment/map (SAM) format and SAMtools. *Bioinformatics*. **25**(16): 2078-2079
- Liao, Y., G. K. Smyth and W. Shi (2013) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*. **30**(7): 923-930
- Livak, K. J. and T. D. Schmittgen (2001) Analysis of Relative Gene Expression Data Using Real-Time Quantitative PCR and the  $2^{-\Delta\Delta CT}$  Method. *Methods*. **25**(4): 402-408
- Liu, H., X. Ma, H. Yu, D. Fang, Y. Li, X. Wang, W. Wang, Y. Dong and B. Xiao (2016) Genomes and virulence difference between two physiological races of *Phytophthora nicotianae*. *GigaScience*. **5**(1): 3
- Lomsadze, A., P. D. Burns and M. Borodovsky (2014) Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. *Nucleic Acids Research*. **42**(15): e119

Love, M. I., W. Huber and S. Anders (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*. **15**(12): 550

Ma, L.-J., H. Charlotte van der Does, K. A. Borkovich, J. J. Coleman, M.-J. Daboussi, A. Di Pietro, M. Dufresne, M. Freitag, M. Grabherr, B. Henrissat *et al.* (2010) Comparative genomics reveals mobile pathogenicity chromosomes in *Fusarium*. *Nature*. **464**(7287): 367

Ma, X., G. Xu, P. He and L. Shan (2016) SERKing Coreceptors for Receptors. *Trends in Plant Science*. **21**(12): 1017-1033

Maas, J. L. (1972) Growth and reproduction in culture of ten *Phytophthora fragariae* races. *Mycopathologia et Mycologia Applicata*. **48**(4): 323-334

Man in't Veld, W. A. (2007) Gene flow analysis demonstrates that *Phytophthora fragariae* var. *rubi* constitutes a distinct species, *Phytophthora rubi* comb. nov. *Mycologia*. **99**(2): 222-226

McDonald, J. H. and M. Kreitman (1991) Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature*. **351**(6328): 652

McIntyre, A. B. R., N. Alexander, A. S. Burton, S. Castro-Wallace, C. Y. Chiu, K. K. John, S. E. Stahl, S. Li and C. E. Mason (2017) Nanopore detection of bacterial DNA base modifications. *bioRxiv*. 127100

McKenna, A., M. Hanna, E. Bank, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Grimella, D. Altshuler, S. Gabriel, M. Daly *et al.* (2010) The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*. **20**(9): 1297-1303

McLellan, H., P. C. Boevink, M. R. Armstrong, L. Pritchard, S. Gomez, J. Morales, S. C. Whisson, J. L. Beynon and P. R. J. Birch (2013) An RxLR Effector from *Phytophthora infestans* Prevents Re-localisation of Two Plant NAC Transcription Factors from the Endoplasmic Reticulum to the Nucleus. *PLoS Pathogens*. **9**(10): e1003670

McMullan, M., M. Rafiqi, G. Kaithakottil, B. J. Clavijo, L. Bilham, E. Orton, L. Percival-Alwyn, B. J. Ward, A. Edwards, D. G. O. Saunders *et al.* (2018) The ash dieback invasion of Europe was founded by two genetically divergent individuals. *Nature Ecology and Evolution*. **2**(6): 1000

Merchant, S., D. E. Wood and S. L. Salzberg (2014) Unexpected cross-species contamination in genome sequencing projects. *PeerJ* **2**: e675

- Montgomerie, I. G. (1964) *Report of the Scottish horticultural research institute 1963-1964*. Pp. 78-83
- Montgomerie, I. G. (1967) Pathogenicity of British isolates of *Phytophthora fragariae* and their relationship with American and Canadian races. *Transactions of the British Mycological Society*. **50**(1): 57-67
- Na, R., D. Yu, B. P. Chapman, Y. Zhang, K. Kuflu, R. Austin, D. Qutob, J. Zhao, Y. Wang and M. Gijzen (2014) Genome Re-Sequencing and Functional Analysis Places the *Phytophthora sojae* Avirulence Genes *Avr1c* and *Avr1a* in a Tandem Repeat at a Single Locus. *PLoS ONE*. **9**(2): e89738
- NCBI Resource Coordinators (2018) Database resources of the National Centre for Biotechnology Information. *Nucleic Acids Research*. **46**(D1): D8-D13
- Nei, M. (1972) Genetic distance between population. *The American Naturalist*. **106**(949): 283-292
- Nei, M. (1987) *Molecular Evolutionary Genetics*. Columbia university press
- Nellist, C. F., R. Vickerstaff, M. K. Sobczyk, C. Marina-Montes, P. Brain, F. M. Wilson, D. W. Simpson, A. B. Whitehouse and R. Harrison (2018) Quantitative Trait Loci Controlling *Phytophthora cactorum* Resistance in the Cultivated Octoploid Strawberry *Fragaria × ananassa*. *bioRxiv*. 249573
- Nielsen, H., J. Engelbrecht, S. Brunak and G. von Heijne (1997) A neural network method for identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Engineering*. **10**(1): 1-6
- Nieuwenhuis, B. P. S. and T. Y. James (2016) The frequency of sex in fungi. *Philosophical Transactions of the Royal Society B*. **371**(1706): 20150540
- Njuguna, W., A. Liston, R. Cronn, T.-L. Ashman and N. Bassil (2013) Insights into phylogeny, sex function and age of *Fragaria* based on whole chloroplast genome sequencing. *Molecular Phylogenetics and Evolution*. **66**(1): 17-29
- Oliva, R. F., M. G. Chacón, D. E. L. Cooke, A. K. Lees and G. A. Forbes (2007) Is *Phytophthora infestans* a Good Taxonomist? Host Recognition and Co-evolution in the *Phytophthora/Solanum* Interaction. *Acta Horticulturae*. **745**: 465
- Olsson, C. H. B. (1995) Diagnosis of *Phytophthora* infections in raspberry and strawberry plants by ELISA tests. *Journal of Phytopathology*. **143**(5): 307-310
- Orsomando, G., L. Brunetti, K. Pucci, B. Ruggeri and S. Ruggieri (2011) Comparative structural and functional characterization of putative protein

effectors belonging to the *PcF toxin family* from *Phytophthora* spp. *Protein Science*. **20**(12): 2047-2059

Owens, G. L., G. J. Baute, S. Hubner and L. H. Rieseberg (2018) Genomic sequence and copy number evolution during hybrid crop development in sunflowers. *Evolutionary Applications*. 1-12

Pacific Biosciences (2018) Pitchfork: Prototyping pacbio build from github source. Available online at <https://github.com/PacificBiosciences/pitchfork> (Last accessed, 03/01/2019)

Pais M., K. Yoshida, A. Giannakopoulou, M. A. Pel, L. M. Cano, R. F. Oliva, K. Witek, H. Lindqvist-Kreuze, V. G. A. A. Vleeshouwers and S. Kamoun (2018) Gene expression polymorphism underpins evasion of host immunity in an asexual lineage of the Irish potato famine pathogen. *BMC Evolutionary Biology*. **18**(1): 93

Paradis, E., J. Claude and K. Strimmer (2004) APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics*. **20**(2): 289-290

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

Pauketat, T. R., L. S. Kelly, G. J. Fritz, N. H. Lopinot, S. Elias and E. Hargrave (2002) The Residues of Feasting and Public Ritual at Early Cahokia. *American Antiquity*. **67**(2): 257-279

Pedersen, B. (2014) pyfasta 0.5.2: fast, memory-efficient, pythonic (and command-line) access to fasta sequence files. Available online at: <https://pypi.org/project/pyfasta> (Last accessed, 03/01/2019)

Perry, M. and S. Raffle (2004) Strawberry red core. *Horticultural development council factsheet*.

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

Pfeifer, B., U. Wittelsbürger, S. E. Ramos-Onsins and M. J. Lercher (2014) PopGenome: An Efficient Swiss Army Knife for Population Genomic Analyses in R. *Molecular Biology and Evolution*. **31**(7): 1929-1936

Pritchard, L. and D. Broadhurst (2014) On the Statistics of Identifying Candidate Pathogen Effectors. In *Plant-Pathogen Interactions* (pp. 53-64). Humana Press, Totowa, NJ

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. W. de Bakker, M. J. Daly *et al.* (2007) PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics*. **81**(3): 559-575

Python Software Foundation (2017) Python programming language version 2.7. Available online at: <https://www.python.org> (Last accessed, 03/01/2019)

Quinlan, A. R. and I. M. Hall (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. **26**(6): 841-842

Qutob, D., J. Tedman-Jones, S. Dong, K. Kuflu, H. Pham, Y. Wang, D. Dou, S. D. Kale, F. D. Arredondo, B. M. Tyler *et al.* (2009) Copy Number Variation and Transcriptional Polymorphisms of *Phytophthora sojae* RXLR Effector Genes *Avr1a* and *Avr3a*. *PLoS ONE*. **4**(4): e5066

Qutob, D., B. P. Chapman and M. Gijzen (2013) Transgenerational gene silencing causes gain of virulence in a plant pathogen. *Nature Communications*. **4**: 1349

R Core Team (2016) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org> (Last accessed, 03/01/2019)

R Core Team (2017) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org> (Last accessed, 03/01/2019)

Raffaele, S., R. A. Farrer, L. M. Cano, D. J. Studholme, D. MacLean, M. Thines, R. H. Y. Jiang, M. C. Zody, S. G. Kunjeti, N. M. Donofrio *et al.* (2010) Genome Evolution Following Host Jumps in the Irish Potato Famine Pathogen Lineage. *Science*. **330**(6010): 1540-1543

Raj, A., M. Stephens and J. K. Pritchard (2014) fastSTRUCTURE: Variational Inference of Population Structure in Large SNP Data Sets. *Genetics*. **197**(2): 573-589

Reid, R. D. (1948) Strawberry breeding at Auchincruive. *Scottish Journal of Agriculture*. **27**: 1-7

Rhoads, A. and K. F. Au (2015) PacBio Sequencing and Its Applications. *Genomics, Proteomics & Bioinformatics*. **13**(5): 278-289



- Rizzo, D M., M. Garbelotto and E. M. Hansen (2005) *Phytophthora ramorum*: Integrative Research and Management of an Emerging Pathogen in California and Oregon Forests. *Annual Review of Phytopathology*. **43**: 309-335
- Rohwer, F., K.-H. Fritzemeier, D. Scheel and K. Hahlbrock (1987) Biochemical reactions of different tissues of potato (*Solanum tuberosum*) to zoospores or elicitors from *Phytophthora infestans*. *Planta*. **170**(4): 556-561
- Rosenberg, N. A. (2004) DISTRUCT: a program for the graphical display of population structure. *Molecular Ecology Notes*. **4**(1): 137-138
- Rousseau-Gueutin, M., E. Lerceteau-Köhler, L. Barrot, D. J. Sargent, A. Monfort, D. Simpson, P. Arús, G. Guérin and B. Denoyes-Rothan (2008) Comparative genetic mapping between octoploid and diploid *Fragaria* species reveals a high level of colinearity between their genomes and the essentially disomic behaviour of the cultivated octoploid strawberry. *Genetics*. **179**(4): 2045-2060
- Ruan, J. (2018) SMARTdenovo: Ultra-fast de novo assembler using long noisy reads. Available online at <https://github.com/ruanjue/smarddenovo> (Last accessed, 03/01/2019)
- Sanger, F., G. M. Air, B. G. Barrell, N. L. Brown, A. R. Coulson, J. C. Fiddes, C. A. Hutchison III, P. M. Slocombe and M. Smith (1977a) Nucleotide sequence of bacteriophage  $\phi$ X174 DNA. *Nature*. **265**(5596): 687
- Sanger, F., S. Nicklen and A. R. Coulson (1977b) DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*. **74**(12): 5463-5467
- Sargent, D. J., Y. Yang, N. Šurbanovski, L. Bianco, M. Buti, R. Velasco, L. Giongo and T. M. Davis (2016) HaploSNP affinities and linkage map positions illuminate subgenome composition in the octoploid, cultivated strawberry (*Fragaria × ananassa*). *Plant Science*. **242**: 140-150
- Schmeider, R. and R. Edwards (2011) Fast identification and removal of sequence contamination from genomic and metagenomic datasets. *PLoS ONE*. **6**(3): e17288
- Schoebel, C. N., J. Stewart, N. J. Grünwald, D. Rigling and S. Prospero (2014) Population History and Pathways of Spread of the Plant Pathogen *Phytophthora plurivora*. *PLoS ONE*. **9**(1): e85368
- Schwessinger, B., J. Sperschneider, W. S. Cuddy, D. P. Garnica, M. E. Miller, J. M. Taylor, P. N. Dodds, M. Figueroa, R. F. Park and J. P. Rathjen (2018) A Near-

Complete Haplotype-Phased Genome of the Dikaryotic Wheat Stripe Rust Fungus *Puccinia striiformis* f. sp. *tritici* Reveals High Interhaplotype Diversity. *mBio*. **9**(1): e02275-17

Scott, D. H., W. F. Jeffers, G. M. Darrow and D. P. Ink (1950) Occurrence of strains of the strawberry red stele fungus, *Phytophthora fragariae* Hickman, as shown by differential varietal response. *Phytopathology*. **40**: 194-198

Seabold, S. and J. Perktold (2010) Statsmodels: Econometric and statistical modeling with python. *Proceedings of the 9th Python in Science Conference*. (Vol. 57, p. 61). Austin: SciPy society.

Shan, W., M. Cao, D. Leung and B. M. Tyler (2003) The *Avr1b* Locus of *Phytophthora sojae* Encodes an Elicitor and a Regulator Required for Avirulence on Soybean Plants Carrying Resistance Gene *Rps1b*. *Molecular Plant-Microbe Interactions*. **17**(4): 394-403

Shulaev, V., D. J. Sargent, R. N. Crowhurst, T. C. Mockler, O. Folkerts, A. L. Delcher, P. Jaiswal, K. Mockaitis, A. Liston, S. P. Mane, *et al.* (2011) The genome of woodland strawberry (*Fragaria vesca*). *Nature Genetics*. **43**(2): 109-116

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

Simpson, J. (2018) Nanopolish: Signal-level algorithms for MinION data. Available online at <https://github.com/jts/nanopolish> (Last Accessed, 03/01/2019)

Slowikowski, K. (2018) ggrepel: Automatically Position Non-Overlapping Text Labels with 'ggplot2'. R package version 0.8.0. Available online at: <https://CRAN.R-project.org/package=ggrepel> (Last accessed, 03/01/2019)

Smakowska-Luzan, E., G. A. Mott, K. Parys, M. Stegmann, T. C. Howton, M. Layeghifard, J. Neuhold, A. Lehner, J. Kong, K. Grünwald *et al.* (2018) An extracellular network of *Arabidopsis* leucine-rich repeat receptor kinases. *Nature*. **553**(7688): 342

Smit, A. F. A. and R. Hubley (2008–2015) RepeatModeler Open-1.0. Available online at: <http://www.repeatmasker.org> (Last accessed, 03/01/2019)

Smit, A. F. A., R. Hubley and P. Green (2013–2015) RepeatMasker Open-4.0. Available online at: <http://www.repeatmasker.org> (Last accessed, 03/01/2019)

Soyer, J. L., M. El Ghalid, N. Glaser, B. Ollivier, J. Linglin, J. Grandaubert, M.-H. Balesdent, L. R. Connolly, M. Freitag, T. Rouxel *et al.* (2014) Epigenetic Control

of Effector Gene Expression in the Plant Pathogenic Fungus *Leptosphaeria maculans*. *PLoS Genetics*. **10**(3): e1004227

Sperschneider, J., P. N. Dodds, K. B. Singh and J. M. Taylor (2018) ApoplastP: prediction of effectors and plant proteins in the apoplast using machine learning. *New Phytologist*. **217**(4): 1764-1778

Stam, R., J. Jupe, A. J. M. Howden, J. A. Morris, P. C. Boevink, P. E. Hedley and E. Huitema (2013) Identification and Characterisation of CRN Effectors in *Phytophthora capsici* Show Modularity and Functional Diversity. *PLoS ONE*. **8**(3): e59517

Stanke, M., M. Diekhans, R. Baertsch and D. Haussler (2008) Using native and syntenically mapped cDNA alignments to improve *de novo* gene finding. *Bioinformatics*. **24**(5): 637-644

Stein, J. C., Y. Yu, D. Copetti, D. J. Zwickl, L. Zhang, C. Zhang, K. Chongule, D. Gao, A. Iwata, J. L. Goicoechea *et al.* (2018) Genomes of 13 domesticated and wild rice relatives highlight genetic conservation, turnover and innovation across the genus *Oryza*. *Nature Genetics*. **50**(2): 285

Tabima, J. F., B. A. Kronmiller, C. M. Press, B. M. Tyler, I. A. Zasada and N. J. Grünwald (2017) Whole genome sequences of the raspberry and strawberry pathogens *Phytophthora rubi* and *P. fragariae*. *Molecular Plant-Microbe Interactions*. **30**(10): 767-769

Tabima, J. F., M. D. Coffey, I. A. Zazada and N. J. Grünwald (2018) Populations of *Phytophthora rubi* show little differentiation and high rates of migration among states in the Western United States. *Molecular Plant-Microbe Interactions*. **31**(6): 614-622

Tate, R., B. Hall, T. DeRego and S. Geib (2014) Annie: the ANNotation Information Extractor (Version 1.0) Available from <http://genomeannotation.github.io/annie> (Last accessed, 03/01/2019)

Taylor, A., V. Vágány, A. C. Jackson, R. J. Harrison, A. Rainoni and J. P. Clarkson (2016) Identification of pathogenicity-related gene in *Fusarium oxysporum* f. sp. *cepae*. *Molecular Plant Pathology*. **17**(7): 1032-1047

Testa, A. C., J. K. Hane, S. R. Ellwood and R. P. Oliver (2015) CodingQuarry: highly accurate hidden Markov model gene prediction in fungal genomes using RNA-seq transcripts. *BMC Genomics*. **16**(1): 170

The UniProt Consortium (2017) UniProt: the universal protein knowledgebase. *Nucleic Acids Research*. **45**(D1): D158-D169

Thilliez, G. J. A., M. R. Armstrong, T.-Y. Lim, K. Baker, A. Jouet, B. Ward, C. van Oosterhout, J. D. G. Jones, E. Huitema, P. R. J. Birch *et al.* (2018) Pathogen enrichment sequencing (PenSeq) enables population genomic studies in oomycetes. *New Phytologist*. **221**(3): 1634-1648

Thomma, B. P. H. J., T. Nürnberger and M. H. A. J. Joosten (2011) Of PAMPs and Effectors: The Blurred PTI-ETI Dichotomy. *The Plant Cell*. **23**(1): 4-15

Thorvaldsdóttir, H., J. T. Robinson and J. P. Mesirov (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics*. **14**(2): 178-192

Tooley, P. W. (1988) Use of uncontrolled freezing for liquid nitrogen storage of *Phytophthora* species. *Plant Disease*. **72**(8): 680-682

Torto, T. A., S. Li, A. Styer, E. Huitema, A. Testa, N. A. R. Gow, P. van West and S. Kamoun (2003) EST Mining and Functional Expression Assays Identify Extracellular Effector Proteins From the Plant Pathogen *Phytophthora*. *Genome Research*. **13**(7): 1675-1685

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

Turner, R. S. (2005) After the famine: Plant pathology, *Phytophthora infestans*, and the late blight of potatoes, 1845—1960. *Historical Studies in the Natural Sciences*. **35**(2): 341-370

Tyler, B. M., S. Tripaty, X. Zhang, P. Dehal, R. H. Y. Jiang, A. Aerts, F. D. Arredondo, L. Baxter, D. Bensasson, J. L. Beynon, *et al.* (2006) *Phytophthora* genome sequences uncover evolutionary origins and mechanisms of pathogenesis. *Science*. **313**(5791): 1261-1266

Untergasser, A., I. Cutcutache, T Koressaar, J. Ye, B. C. Faircloth, M. Remm and S. G. Rozen (2012) Primer3 - new capabilities and interfaces. *Nucleic Acids Research*. **40**(15): e115

van de Weg, W. E. (1997a) Gene-for-gene relationships between strawberry and the causal agent of red stele root rot, *Phytophthora fragariae* var. *fragariae*. PhD thesis, University of Wageningen, The Netherlands, 93 p

- van de Weg, W. E. (1997b) A gene-for-gene model to explain interactions between cultivars of strawberry and races of *Phytophthora fragariae* var. *fragariae*. *Theoretical Applied Genetics*. **94**(3-4): 445-451
- van de Weg, W. E., S. Giezen, B. Henken and A. P. M. Den Nijs (1996) A quantitative classification method for assessing resistance to *Phytophthora fragariae* var. *fragariae* in strawberry. *Euphytica*. **91**(1): 119-125
- van de Weg, W. E., B. Henken and S. Giezen (1997) Assessment of the resistance to *Phytophthora fragariae* var. *fragariae* of the USA and Canadian differential series of strawberry genotypes. *Journal of Phytopathology*. **145**(1): 1-6
- van Dijk, T., G. Pagliarani, A. Pikunova, Y. Noordijk, H. Yilmaz-Temel, B. Meulenbroek, R. G. F. Visser and E. van de Weg (2014) Genomic rearrangements and signatures of breeding in the allo-octoploid strawberry as revealed through an allele dose based SSR linkage map. *BMC Plant Biology*. **14**(1): 55
- van Poppel, P. M. J. A., J. Guo, P. J. I. van de Vondervoort, M. W. M. Jung, P. R. J. Birch, S. C. Whisson and F. Govers (2008) The *Phytophthora infestans* Avirulence Gene *Avr4* Encodes an RXLR-dEER Effector. *Molecular Plant-Microbe Interactions*. **21**(11): 1460-1470
- van West, P., S. J. Shepherd, C. A. Walker, S. Li, A. A. Appiah, L. J. Grenville-Briggs, F. Govers and N. A. R. Gow (2008) Internuclear gene silencing in *Phytophthora infestans* is established through chromatin remodelling. *Microbiology*. **154**(5): 1482-1490
- Vaser, R., I. Sović, N. Nagarajan and M. Šikić (2017) Fast and accurate *de novo* genome assembly from long uncorrected reads. *Genome Research*. **27**: 1-10
- Wala, J. A., P. Bandopadhyay, N. Greenwald, R. O'Rourke, T. Sharpe, C. Stewart, S. Schumacher, Y. Li, J. Weischenfeldt, X. Yao, C. Nusbaum *et al.* (2018) SvABA: genome-wide detections of structural variations and indels by local assembly. *Genome Research*. **28**(4): 581-591
- Walker, B. J., T. Abeel, T. Shea, M. Priest, A. Abouelliel, S. Sakthikumar, C. A. Cuomo, Q. Zeng, J. Wortman, S. K. Young, *et al.* (2014) Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE*. **9**(11): e112963

Walshaw, S. (2009) *Fragaria* spp. Laboratory Guide To Archaeological Plant Remains From Eastern North America. <https://pages.wustl.edu/fritz/fragaria-spp>. (Last accessed, 03/01/2019)

Ward, E. W. B., D. M. Cahill and M. K. Bhattacharyya (1989) Early cytological differences between compatible and incompatible interactions of soybeans with *Phytophthora megasperma* f. sp. *glycinea*. *Physiological and Molecular Plant Pathology*. **34**(3): 267-283

Warnes, G. R., B. Bolker, L. Bonebakker, R. Gentleman, W. H. A. Liaw, T. Lumley, M. Maechler, A. Magnusson, S. Moeller, M. Schwartz *et al.* (2016) gplots: Various R Programming Tools for Plotting Data. R package version 3.0.1. <https://CRAN.R-project.org/package=gplots> (Last accessed, 03/01/2019)

Waterhouse, G. M. (1963) Key to the species of *Phytophthora* de Bary. *Mycological Papers*. **92**: 1-22

Watterson, G. A. (1975) On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*. **7**(2): 256-276

Wawra, S., F. Trusch, A. Matena, K. Apostolakis, U. Linne, I. Zhukov, J. Stanek, W. Koźmiński, I. Davidson, C. J. Secombes *et al.* (2017) The RxLR Motif of the Host Targeting Effector AVR3a of *Phytophthora infestans* Is Cleaved before Secretion. *The Plant Cell*. **29**(6): 1184-1195

Whisson, S. C., P. C. Boevink, L. Moleleki, A. O. Avrova, J. G. Morales, E. M. Gilroy, M. R. Armstrong, S. Grouffaud, P. van West, S. Chapman *et al.* (2007) A translocation signal for delivery of oomycete effector proteins into host plant cells. *Nature*. **450**(7166): 115-118

Wick, R. (2018) Porechop: adapter trimmer for Oxford Nanopore reads. Available online at <https://github.com/rrwick/Porechop> (Last accessed, 03/01/2019)

Wickman, H. (2016) *ggplot2: Elegant Graphics for Data Analysis*. Springer.

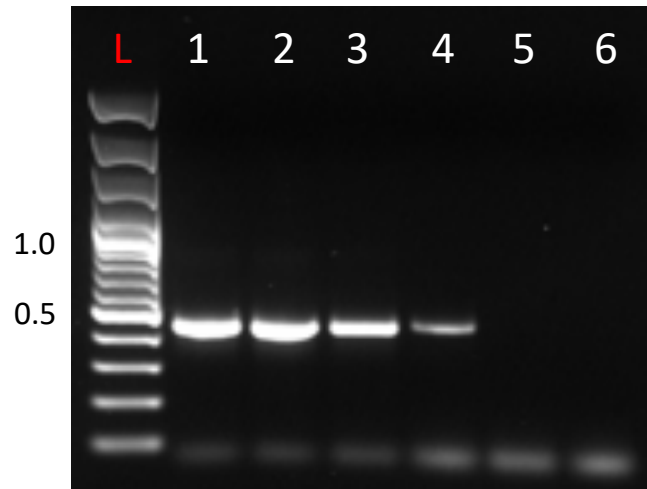
Wilcox, W. F., P. H. Scott, P. B. Hamm, D. M. Kennedy, J. M. Duncan, C. M. Brasier and E. M. Hansen (1993) Identity of a *Phytophthora* species attacking raspberry in Europe and North America. *Mycological Research*. **97**(7): 817-831

Wilhelm, S. and J. E. Sagen (1974) History of the strawberry. Division of Agricultural Sciences. *University of California, Berkeley, California, USA*.

Wilson, F., K. Harrison, A. D. Armitage, A. J. Simkin and R. J. Harrison (2018) CRISPR/Cas9-mediated mutagenesis of phytoene desaturase in diploid and octoploid strawberry. *bioRxiv*. 471680

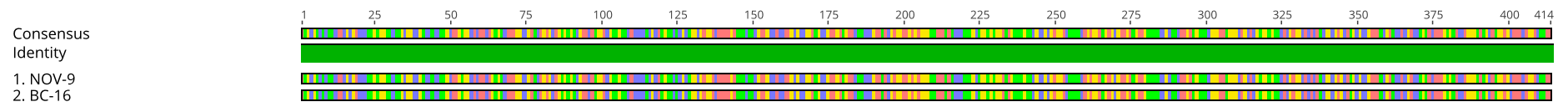
- Win, J., T.-D. Kanneganti, T. Torto-Alialibo and S. Kamoun (2006) Computational and comparative analyses of 150 full-length cDNA sequences from the oomycete plant pathogen *Phytophthora infestans*. *Fungal Genetics and Biology*. **43**(1): 20-23
- Wright, S. (1943) Isolation by distance. *Genetics*. **28**(2): 114
- Wu, C.-H., A. Abd-El-Haliem, T. O. Bozkurt, K. Belhaj, R. Terauchi, J. H. Vossen and S. Kamoun (2017) NLR network mediates immunity to diverse plant pathogens. *Proceedings of the National Academy of Sciences*. **114**(30): 8113-8118
- Wu, C.-H., L. Derevnina and S. Kamoun (2018) Receptor networks underpin plant immunity. *Science*. **360**(6395): 1300-1301
- Yan., H. Z. and R. F. Liou (2006) Selection of internal control genes for real-time quantitative RT-PCR assays in the oomycete plant pathogen *Phytophthora parasitica*. *Fungal Genetics and Biology*. **43**(6): 430-438
- Yang, X., B. M. Tyler and C. Hong (2017) An expanded phylogeny for the genus *Phytophthora*. *IMA Fungus*. **8**(2): 355-384
- Yin, W., S. Dong, L. Zhai, Y. Lin, X. Zheng and Y. Wang (2013) The *Phytophthora sojae Avr1d* Gene Encodes an RxLR-dEER Effector with Presence and Absence Polymorphisms Among Pathogen Strains. *Molecular Plant-Microbe Interactions*. **26**(8): 958-968
- Yoon, S., Z. Xuan, V. Makarov, K. Ye and J. Sebat (2009) Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Research*. **19**(9): 1586-1592
- Yu, D., H. Tan, Y. Zhang, Z. Du, H. Yu and Q. Chen (2012) Comparison and improvement of different methods of RNA isolation from strawberry (*Fragria × ananassa*) [sic]. *Journal of Agricultural Science*. **4**(7): 51-56
- Zhang, Y., R. Xia, H. Kuang and B. C. Meyers (2016) The Diversification of Plant NBS-LRR Defense Genes Directs the Evolution of MicroRNAs That Target Them. *Molecular biology and evolution*. **3**(10): 2692-2705
- Zheng, X., D. Levine, J. Shen, S. M. Gogarten, C. Laurie and B. S. Weir (2012) A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics*. **28**(24): 3326-3328

## Appendix

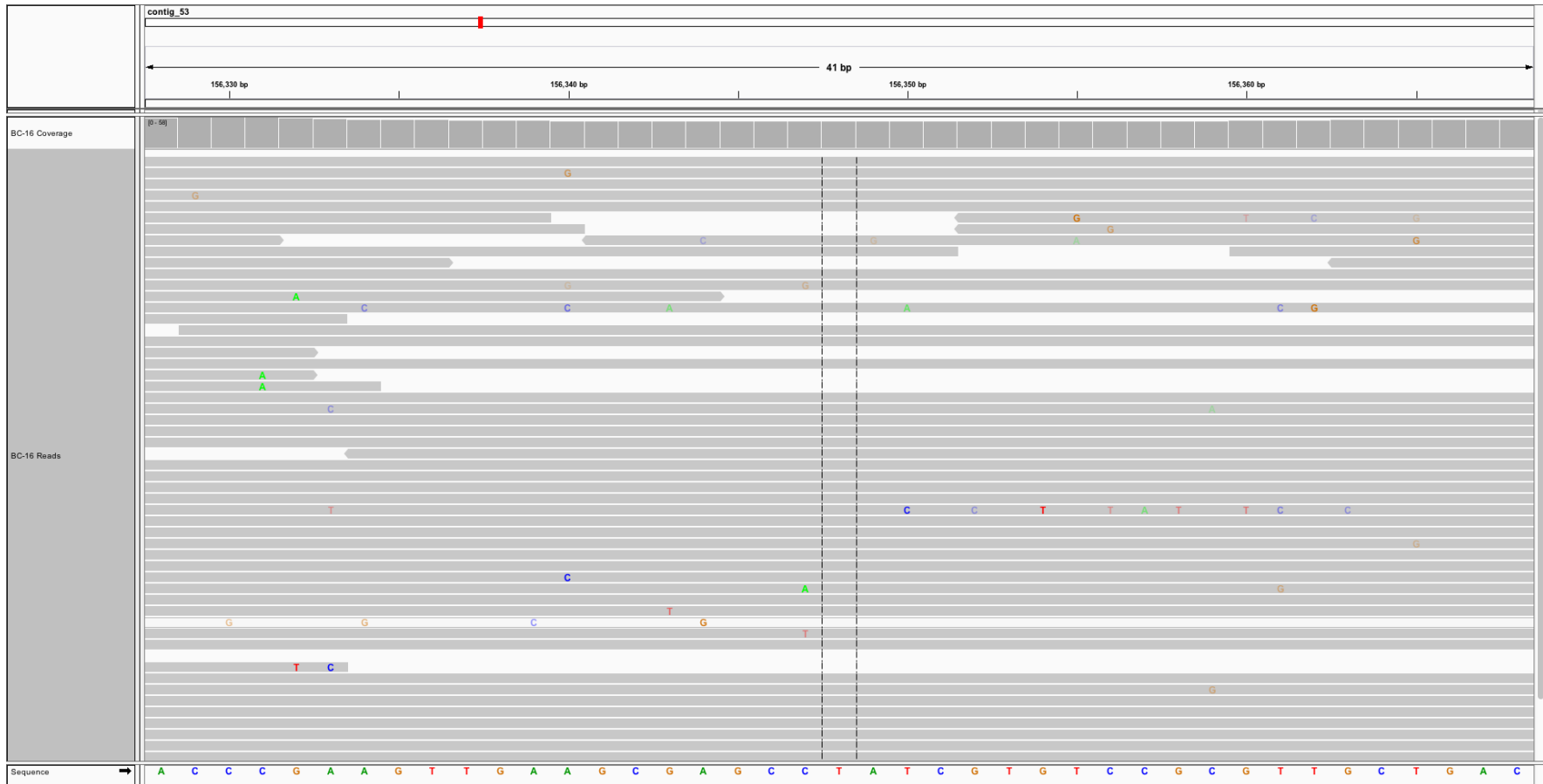


**Fig. A.1: Agarose gel electrophoresis of the PCR products of  $\beta$ -tubulin primers on BC-1 and NOV-9 gDNA. L:** 100 bp plus ladder from New England Biolabs, sizes listed in kb. PCR Templates: **1:** BC-1 gDNA. **2:** NOV-9 gDNA. **3 - 4:** BC-16 gDNA. **5:** gDNA from the 'Hapil' cultivar of *Fragaria × ananassa*. **6:** dH<sub>2</sub>O.





**Fig. A.2: Alignment of the strong candidate avirulence gene in BC-16 to the orthologous gene in NOV-9.** Figure created in Geneious R10 (Kearse *et al.*, 2012)



**Fig. A.3: Alignment of the Illumina reads of BC-16 to the BC-16 FALCON assembly at the identified SNP downstream of the strong candidate avirulence gene in BC-16.** Alignment was performed using Bowtie 2 (Langmead and Salzberg, 2012). Figure created in IGV (Thorvaldsdóttir *et al.*, 2013).



**Fig. A.4: Alignment of the Illumina reads of BC-16 to the NOV-9 ONT assembly at the identified SNP downstream of the orthologue of the strong candidate avirulence gene in BC-16.** Alignment was performed using Bowtie 2 (Langmead and Salzberg, 2012). Figure created in IGV (Thorvaldsdóttir *et al.*, 2013).



**Fig. A.5: Alignment of the Illumina reads of NOV-9 to the NOV-9 ONT assembly at the identified SNP downstream of the orthologue of the strong candidate avirulence gene in BC-16.** Alignment was performed using Bowtie 2 (Langmead and Salzberg, 2012). Figure created in IGV (Thorvaldsdóttir *et al.*, 2013).



**Fig. A.6: Alignment of the Illumina reads of BC-16 to the BC-16 FALCON assembly at the identified INDEL upstream of the strong candidate avirulence gene in BC-16.** Alignment was performed using Bowtie 2 (Langmead and Salzberg, 2012). Figure created in IGV (Thorvaldsdóttir *et al.*, 2013).

**Table A.1: Investigation of the enrichment of classes of genes showing evidence of population separation between the *Phytophthora fragariae* UKR1-2-3 population and the population of the isolates BC-23 and ONT-3**

Interproscan Annotation	Confidence Level	Genomic Total	Separated Total	<i>p</i> -value	<i>p</i> -value adjusted		InterProScan Description
					Benjamini-Hochberg	Bonferroni	
<b>IPR000612</b>	<b>High</b>	<b>2</b>	<b>1</b>	<b>0.00237</b>	<b>0.0331</b>	<b>0.0497</b>	<b>Proteolipid membrane potential modulator</b>
IPR000742	High	26	1	0.0303	0.0623	0.637	EGF-like domain
IPR001229	High	28	1	0.0326	0.0623	0.685	Jacalin-like lectin domain
IPR001926	High	16	1	0.0188	0.0563	0.394	Tryptophan synthase beta subunit-like PLP-dependent enzyme
IPR006626	High	32	1	0.0383	0.0671	0.805	Parallel beta-helix repeat
IPR013032	High	24	1	0.0280	0.0623	0.589	EGF-like, conserved site
IPR013111	High	9	1	0.0106	0.0412	0.223	EGF-like domain, extracellular
IPR013830	High	19	1	0.0223	0.0584	0.467	SGNH hydrolase-type esterase domain
IPR015889	High	10	1	0.0118	0.0412	0.247	Intradiol ring-cleavage dioxygenase, core
IPR027278	High	3	1	0.00355	0.0331	0.0745	1-aminocyclopropane-1-carboxylate deaminase/D-cysteine desulfhyrase
IPR030392	High	4	1	0.00473	0.0412	0.201	Intramolecular chaperone auto-processing domain
No Annotation	High	24,174	37	0.00956	0.041	0.201	No functional domains identified
IPR000300	Low	7	3	0.0337	0.272	1.00	Inositol polyphosphate-related phosphatase
IPR000306	Low	121	21	0.0247	0.272	1.00	FYVE zinc finger
IPR000408	Low	33	8	0.0247	0.272	1.00	Regulator of chromosome condensation, RCC1
IPR000718	Low	3	1	0.0340	0.272	1.00	Peptidase M13
IPR000851	Low	2	2	0.0123	0.272	1.00	Ribosomal protein S5
IPR001093	Low	3	1	0.0340	0.272	1.00	IMP dehydrogenase/GMP reductase
IPR001194	Low	10	4	0.0182	0.272	1.00	DENN domain
IPR001206	Low	7	3	0.0337	0.272	1.00	Diacylglycerol kinase, catalytic domain
IPR001245	Low	140	23	0.0346	0.272	1.00	Serine-threonine/tyrosine-protein kinase catalytic domain
IPR001314	Low	15	5	0.0192	0.272	1.00	Peptidase S1A, chymotrypsin family
IPR001461	Low	13	5	0.0099	0.272	1.00	Aspartic peptidase A1 family
IPR001678	Low	8	3	0.0495	0.272	1.00	SAM-dependent methyltransferase RsmB/NOP2-type
IPR001876	Low	34	8	0.0292	0.272	1.00	Zinc finger, RanBP2-type
IPR001972	Low	7	3	0.0337	0.272	1.00	Stomatin family
IPR001991	Low	13	4	0.0473	0.272	1.00	Sodium:dicarboxylate symporter
IPR002194	Low	8	3	0.0495	0.272	1.00	Chaperonin TCP-1, conserved site
IPR002423	Low	13	4	0.0473	0.272	1.00	Chaperonin Cpn60/TCP-1 family

IPR002591	Low	3	2	0.0340	0.272	1.00	Type I phosphodiesterase/nucleotide pyrophosphatase/phosphate transferase
IPR002594	Low	10	4	0.0182	0.272	1.00	Glycoside hydrolase family 12
IPR002634	Low	2	2	0.0123	0.272	1.00	BolA protein
IPR002880	Low	2	2	0.0123	0.272	1.00	Pyruvate flavodoxin/ferredoxin oxidoreductase, N-terminal
IPR002935	Low	3	2	0.0340	0.272	1.00	O-methyltransferase, family 3
IPR002939	Low	6	3	0.0209	0.272	1.00	Chaperone DnaJ, C-terminal
IPR003591	Low	46	11	0.0102	0.272	1.00	Leucine-rich repeat, typical subtype
IPR003846	Low	3	2	0.0340	0.272	1.00	Uncharacterised protein family UPF0061
IPR004328	Low	3	2	0.0340	0.272	1.00	BRO1 domain
IPR004567	Low	3	3	0.0014	0.272	1.00	Type II pantothenate kinase
IPR004821	Low	8	4	0.00727	0.272	1.00	Cytidyltransferase-like domain
IPR005052	Low	3	2	0.0340	0.272	1.00	Legume-like lectin
IPR005062	Low	3	2	0.0340	0.272	1.00	SAC3/GANP/THP3
IPR005113	Low	8	4	0.00727	0.272	1.00	uDENN domain
IPR005117	Low	2	2	0.0123	0.272	1.00	Nitrite/Sulfite reductase ferredoxin-like domain
IPR005324	Low	2	2	0.0123	0.272	1.00	Ribosomal protein S5, C-terminal
IPR005645	Low	8	3	0.0495	0.272	1.00	Serine hydrolase FSH
IPR005990	Low	2	2	0.0123	0.272	1.00	Inosine-5'-monophosphate dehydrogenase
IPR006066	Low	2	2	0.0123	0.272	1.00	Nitrite/sulphite reductase iron-sulphur/sirohaem-binding site
IPR006067	Low	2	2	0.0123	0.272	1.00	Nitrite/sulphite reductase 4Fe-4S domain
IPR006518	Low	3	2	0.0340	0.272	1.00	Trypanosome RHS
IPR006534	Low	4	3	0.00497	0.272	1.00	P-type ATPase, subfamily IIIA
IPR006620	Low	13	4	0.0473	0.272	1.00	Prolyl 4-hydroxylase, alpha subunit
IPR006845	Low	3	2	0.0340	0.272	1.00	Pex, N-terminal
IPR007070	Low	2	2	0.0123	0.272	1.00	GPI ethanolamine phosphate transferase 1
IPR007248	Low	11	6	0.000516	0.272	0.682	Mpv17/PMP22
IPR008269	Low	3	2	0.0340	0.272	1.00	Peptidase S16, Lon C-terminal
IPR008538	Low	7	3	0.0337	0.272	1.00	Domain of unknown function DUF820
IPR008701	Low	43	10	0.0170	0.272	1.00	Necrosis inducing protein
IPR008753	Low	3	2	0.0340	0.272	1.00	Peptidase M13, N-terminal domain
IPR008971	Low	6	3	0.0209	0.272	1.00	HSP40/DnaJ peptide-binding
IPR009003	Low	41	9	0.0326	0.272	1.00	Peptidase S1, PA clan
IPR009008	Low	6	3	0.0209	0.272	1.00	Valyl/Leucyl/Isoleucyl-tRNA synthetase, editing domain
IPR009091	Low	35	8	0.0343	0.272	1.00	Regulator of chromosome condensation 1/beta-lactamase-inhibitor protein II
IPR010921	Low	3	2	0.0340	0.272	1.00	Trp repressor/replication initiator

IPR011054	Low	6	3	0.0209	0.272	1.00	Rudiment single hybrid motif
IPR011761	Low	10	4	0.0182	0.272	1.00	ATP-grasp fold
IPR013122	Low	15	6	0.00378	0.272	1.00	Polycystin cation channel, PKD1/PKD2
IPR013319	Low	11	4	0.0261	0.272	1.00	Glycoside hydrolase family 11/12
IPR013810	Low	2	2	0.0123	0.272	1.00	Ribosomal protein S5, N-terminal
IPR013892	Low	13	4	0.0473	0.272	1.00	Cytochrome c oxidase biogenesis protein Cmc1-like
IPR014720	Low	11	5	0.00431	0.272	1.00	Double-stranded RNA-binding domain
IPR015915	Low	27	7	0.0245	0.272	1.00	Kelch-type beta propeller
IPR016064	Low	9	4	0.0119	0.272	1.00	NAD kinase/diacylglycerol kinase-like domain
IPR016160	Low	12	4	0.0358	0.272	1.00	Aldehyde dehydrogenase, cysteine active site
IPR017849	Low	28	8	0.00909	0.272	1.00	Alkaline phosphatase-like, alpha/beta/alpha
IPR017850	Low	28	8	0.00909	0.272	1.00	Alkaline-phosphatase-like, core domain
IPR017998	Low	8	3	0.0495	0.272	1.00	Chaperone tailless complex polypeptide 1 (TCP-1)
IPR018114	Low	6	4	0.00187	0.272	1.00	Serine proteases, trypsin family, histidine active site
IPR018247	Low	139	23	0.0322	0.272	1.00	EF-Hand 1, calcium-binding site
IPR018497	Low	3	2	0.0340	0.272	1.00	Peptidase M13, C-terminal domain
IPR018993	Low	2	2	0.0123	0.272	1.00	FGFR1 oncogene partner (FOP), N-terminal dimerisation domain
IPR019474	Low	2	2	0.0123	0.272	1.00	Ubiquitin conjugation factor E4, core
IPR021980	Low	2	2	0.0123	0.272	1.00	Transcription factor homeodomain, male germ-cell
IPR022742	Low	20	6	0.0180	0.272	1.00	Serine aminopeptidase, S33
IPR023267	Low	8	3	0.0495	0.272	1.00	RNA (C5-cytosine) methyltransferase
IPR023393	Low	119	20	0.0377	0.272	1.00	START-like domain
IPR025304	Low	2	2	0.0123	0.272	1.00	ALIX V-shaped domain
IPR025789	Low	18	5	0.0415	0.272	1.00	Histone-lysine N-methyltransferase DOT1 domain
IPR027409	Low	12	4	0.0358	0.272	1.00	GroEL-like apical domain
IPR027410	Low	9	4	0.0119	0.272	1.00	TCP-1-like chaperonin intermediate domain
IPR029045	Low	15	5	0.0192	0.272	1.00	ClpP/crotonase-like domain
IPR029061	Low	13	4	0.0473	0.272	1.00	Thiamin diphosphate-binding fold
IPR029063	Low	154	26	0.0163	0.272	1.00	S-adenosyl-L-methionine-dependent methyltransferase
IPR033116	Low	6	3	0.0209	0.272	1.00	Serine proteases, trypsin family, serine active site
IPR033121	Low	13	5	0.00994	0.272	1.00	Peptidase family A1 domain
IPR033380	Low	2	2	0.0123	0.272	1.00	Glutamate-aspartate symport protein GltP/GltT

Genes with a  $p$ -value < 0.05 were accepted as statistically significant, these entries are displayed in bold.



**Table A.2: Investigation of the enrichment of effector classes of genes showing evidence of diversifying selection in the *Phytophthora fragariae* UK1-2-3 population**

Transcript Classification	Genomic Total	Diversifying Total	$p$ -value	$p$ -value adjusted	
				Benjamini-Hochberg	Bonferroni
Apolastic Effectors	4,864	24	0.416	0.416	1.00
Crinkler Effectors	88	1	0.337	0.416	1.00
RxLR Effectors	1,057	3	0.416	0.416	1.00
Secreted Proteins	8,024	35	0.337	0.416	1.00

Genes with a  $p$ -value < 0.05 were accepted as statistically significant.

**Table A.3: Investigation of the enrichment of InterProScan terms of genes showing evidence of diversifying selection in the *Phytophthora fragariae* UK1-2-3 population**

Interproscan Annotation	Genomic Total	Diversifying Total	$p$ -value adjusted			InterProScan Description
			$p$ -value	Benjamini-Hochberg	Bonferroni	
IPR000959	1	1	0.00466	0.0398	0.676	POLO box duplicated domain
IPR002717	1	1	0.00466	0.0398	0.676	Histone acetyltransferase domain, MYST-type
IPR003135	1	1	0.00466	0.0398	0.676	ATP-grasp fold, ATP-dependent carboxylate-amine ligase-type
IPR004947	1	1	0.00466	0.0398	0.676	Deoxyribonuclease II
IPR005662	1	1	0.00466	0.0398	0.676	GTP-binding protein Era
IPR005875	1	1	0.00466	0.0398	0.676	Phosphoribosylaminoimidazole carboxylase, ATPase subunit
IPR006809	1	1	0.00466	0.0398	0.676	TAFII28-like protein
IPR010960	1	1	0.00466	0.0398	0.676	Flavocytochrome c
IPR013880	1	1	0.00466	0.0398	0.676	Yos1-like
IPR016301	1	1	0.00466	0.0398	0.676	Phosphoribosylaminoimidazole carboxylase
IPR017359	1	1	0.00466	0.0398	0.676	Uncharacterised conserved protein UCP038021, RWD
IPR018737	1	1	0.00466	0.0398	0.676	Protein LIN52
IPR019414	1	1	0.00466	0.0398	0.676	RNA polymerase II assembly factor Rtp1, C-terminal domain 2
IPR019451	1	1	0.00466	0.0398	0.676	RNA polymerase II assembly factor Rtp1, C-terminal
IPR021013	1	1	0.00466	0.0398	0.676	ATPase, vacuolar ER assembly factor, Vma12
IPR021967	1	1	0.00466	0.0398	0.676	Nuclear protein 96
IPR028933	1	1	0.00466	0.0398	0.676	Lebercilin domain
IPR000031	2	1	0.00930	0.0482	1.00	Phosphoribosylaminoimidazole carboxylase PurE domain
IPR000977	2	1	0.00930	0.0482	1.00	DNA ligase, ATP-dependent
IPR002306	2	1	0.00930	0.0482	1.00	Tryptophan-tRNA ligase
IPR004044	2	1	0.00930	0.0482	1.00	K Homology domain, type 2
IPR004241	2	1	0.00930	0.0482	1.00	Autophagy protein Atg8 ubiquitin-like
IPR007230	2	1	0.00930	0.0482	1.00	Peptidase S59, nucleoporin
IPR007918	2	1	0.00930	0.0482	1.00	Mitochondrial distribution/morphology family 35/apoptosis
IPR009019	2	1	0.00930	0.0482	1.00	K homology domain, prokaryotic type
IPR010541	2	1	0.00930	0.0482	1.00	Domain of unknown function DUF1115

IPR013167	2	1	0.00930	0.0482	1.00	Conserved oligomeric Golgi complex, subunit 4
IPR019758	2	1	0.00930	0.0482	1.00	Peptidase S26A, signal peptidase I, conserved site
IPR001584	1045	11	0.0103	0.0516	1.00	Integrase, catalytic core
IPR000223	3	1	0.0139	0.0531	1.00	Peptidase S26A, signal peptidase I
IPR000738	3	1	0.0139	0.0531	1.00	WHEP-TRS domain
IPR002305	3	1	0.0139	0.0531	1.00	Aminoacyl-tRNA synthetase, class Ic
IPR003953	3	1	0.0139	0.0531	1.00	FAD-dependent oxidoreductase 2, FAD binding domain
IPR012308	3	1	0.0139	0.0531	1.00	DNA ligase, ATP-dependent, N-terminal
IPR012309	3	1	0.0139	0.0531	1.00	DNA ligase, ATP-dependent, C-terminal
IPR016059	3	1	0.0139	0.0531	1.00	DNA ligase, ATP-dependent, conserved site
IPR027477	3	1	0.0139	0.0531	1.00	Succinate dehydrogenase/fumarate reductase flavoprotein, catalytic domain
IPR030616	3	1	0.0139	0.0531	1.00	Aurora kinase
IPR009072	44	2	0.0180	0.0597	1.00	Histone-fold
IPR012310	4	1	0.0185	0.0597	1.00	DNA ligase, ATP-dependent, central
IPR015927	4	1	0.0185	0.0597	1.00	Peptidase S24/S26A/S26B/S26C
IPR015946	4	1	0.0185	0.0597	1.00	K homology domain-like, alpha/beta
IPR019759	4	1	0.0185	0.0597	1.00	Peptidase S24/S26A/S26B
IPR025995	4	1	0.0185	0.0597	1.00	RNA binding activity-knot of a chromodomain
IPR028360	4	1	0.0185	0.0597	1.00	Peptidase S24/S26, beta-ribbon domain
IPR000627	5	1	0.0231	0.0657	1.00	Intradial ring-cleavage dioxygenase, C-terminal
IPR000971	5	1	0.0231	0.0657	1.00	Globin
IPR009050	5	1	0.0231	0.0657	1.00	Globin-like
IPR009068	5	1	0.0231	0.0657	1.00	S15/NS1, RNA-binding
IPR012292	5	1	0.0231	0.0657	1.00	Globin/Protoglobin
IPR023211	5	1	0.0231	0.0657	1.00	DNA polymerase, palm domain
IPR016197	481	6	0.0259	0.0721	1.00	Chromo domain-like
IPR006575	6	1	0.0276	0.0729	1.00	RWD domain
IPR006630	6	1	0.0276	0.0729	1.00	RNA-binding protein Lupus La
IPR011054	6	1	0.0276	0.0729	1.00	Rudiment single hybrid motif
IPR029071	57	2	0.0292	0.0756	1.00	Ubiquitin-related domain
IPR004127	7	1	0.0322	0.0791	1.00	Prefoldin alpha-like
IPR009027	7	1	0.0322	0.0791	1.00	Ribosomal protein L9/RNase H1, N-terminal
IPR011320	7	1	0.0322	0.0791	1.00	Ribonuclease H1, N-terminal

IPR000121	8	1	0.0367	0.0831	1.00	PEP-utilising enzyme, C-terminal
IPR001300	8	1	0.0367	0.0831	1.00	Peptidase C2, calpain, catalytic domain
IPR001678	8	1	0.0367	0.0831	1.00	SAM-dependent methyltransferase RsmB/NOP2-type
IPR005645	8	1	0.0367	0.0831	1.00	Serine hydrolase FSH
IPR023267	8	1	0.0367	0.0831	1.00	RNA (C5-cytosine) methyltransferase
IPR001164	9	1	0.0412	0.0905	1.00	Arf GTPase activating protein
IPR003958	9	1	0.0412	0.0905	1.00	Transcription factor CBF/NF-Y/archaeal histone domain
IPR011761	10	1	0.0457	0.0959	1.00	ATP-grasp fold
IPR013815	10	1	0.0457	0.0959	1.00	ATP-grasp fold, subdomain 1
IPR015889	10	1	0.0457	0.0959	1.00	Intradiol ring-cleavage dioxygenase, core
IPR013525	80	2	0.0539	0.110	1.00	ABC-2 type transporter
IPR009053	12	1	0.0545	0.110	1.00	Prefoldin
IPR025660	12	1	0.0545	0.110	1.00	Cysteine peptidase, histidine active site
IPR012337	1544	1	0.0587	0.116	1.00	Ribonuclease H-like domain
IPR001938	13	1	0.0589	0.116	1.00	Thaumatococcus
IPR000169	14	1	0.0633	0.119	1.00	Cysteine peptidase, cysteine active site
IPR001412	14	1	0.0633	0.119	1.00	Aminoacyl-tRNA synthetase, class I, conserved site
IPR016185	14	1	0.0633	0.119	1.00	Pre-ATP-grasp domain
IPR013780	15	1	0.0677	0.126	1.00	Glycosyl hydrolase, all-beta
IPR001139	17	1	0.0764	0.135	1.00	Glycoside hydrolase family 30
IPR001199	17	1	0.0764	0.135	1.00	Cytochrome b5-like heme/steroid binding domain
IPR001951	17	1	0.0764	0.135	1.00	Histone H4
IPR013816	17	1	0.0764	0.135	1.00	ATP-grasp fold, subdomain 2
IPR025733	18	1	0.0807	0.1410	1.00	Iron/zinc purple acid phosphatase-like C-terminal domain
IPR001594	19	1	0.0850	0.142	1.00	Zinc finger, DHHC-type, palmitoyltransferase
IPR008963	19	1	0.0850	0.142	1.00	Purple acid phosphatase-like, N-terminal
IPR015914	19	1	0.0850	0.142	1.00	Purple acid phosphatase, N-terminal
IPR027443	19	1	0.0850	0.142	1.00	Isopenicillin N synthase-like
IPR006073	20	1	0.0892	0.145	1.00	GTP binding domain
IPR017972	20	1	0.0892	0.145	1.00	Cytochrome P450, conserved site
IPR015813	21	1	0.0935	0.151	1.00	Pyruvate/Phosphoenolpyruvate kinase-like domain
IPR000668	22	1	0.0977	0.154	1.00	Peptidase C1A, papain C-terminal
IPR013128	22	1	0.0977	0.154	1.00	Peptidase C1A

IPR002350	26	1	0.114	0.178	1.00	Kazal domain
IPR013781	126	2	0.117	0.181	1.00	Glycoside hydrolase, catalytic domain
IPR001609	28	1	0.123	0.183	1.00	Myosin head, motor domain
IPR002492	28	1	0.123	0.183	1.00	Transposase, Tc1-like
IPR016135	28	1	0.123	0.183	1.00	Ubiquitin-conjugating enzyme/RWD-like
IPR000743	30	1	0.131	0.194	1.00	Glycoside hydrolase, family 28
IPR003439	139	2	0.137	0.198	1.00	ABC transporter-like
IPR002401	32	1	0.139	0.198	1.00	Cytochrome P450, E-class, group I
IPR024936	32	1	0.139	0.198	1.00	Cyclophilin-type peptidyl-prolyl cis-trans isomerase
IPR026892	32	1	0.139	0.198	1.00	Glycoside hydrolase family 3
IPR002130	33	1	0.143	0.199	1.00	Cyclophilin-type peptidyl-prolyl cis-trans isomerase domain
IPR006626	33	1	0.143	0.199	1.00	Parallel beta-helix repeat
IPR000953	432	4	0.144	0.199	1.00	Chromo/chromo shadow domain
IPR029000	34	1	0.147	0.201	1.00	Cyclophilin-like domain
IPR005123	37	1	0.159	0.213	1.00	Oxoglutarate/iron-dependent dioxygenase
IPR010929	37	1	0.159	0.213	1.00	CDR ABC transporter
IPR000626	38	1	0.163	0.215	1.00	Ubiquitin domain
IPR023780	298	3	0.163	0.215	1.00	Chromo domain
IPR001128	41	1	0.174	0.228	1.00	Cytochrome P450
IPR016181	42	1	0.178	0.231	1.00	Acyl-CoA N-acyltransferase
IPR003591	46	1	0.194	0.248	1.00	Leucine-rich repeat, typical subtype
IPR013242	47	1	0.197	0.251	1.00	Retroviral aspartyl protease
IPR004324	49	1	0.205	0.258	1.00	Folate-biopterin transporter
IPR014729	51	1	0.212	0.265	1.00	Rossmann-like alpha/beta/alpha sandwich fold
IPR017441	186	2	0.215	0.267	1.00	Protein kinase, ATP binding site
IPR004316	53	1	0.219	0.270	1.00	SWEET sugar transporter
IPR001995	57	1	0.234	0.285	1.00	Peptidase A2A, retrovirus, catalytic
IPR002200	58	1	0.238	0.287	1.00	Elicitin
IPR011010	59	1	0.241	0.289	1.00	DNA breaking-rejoining enzyme, catalytic core
IPR005225	60	1	0.245	0.291	1.00	Small GTP-binding protein domain
IPR000477	915	6	0.256	0.302	1.00	Reverse transcriptase domain
IPR004843	66	1	0.266	0.311	1.00	Calcineurin-like phosphoesterase domain, apaH type
IPR023753	73	1	0.289	0.336	1.00	FAD/NAD(P)-binding domain

IPR013762	74	1	0.293	0.337	1.00	Integrase-like, catalytic domain
IPR001478	80	1	0.312	0.356	1.00	PDZ domain
IPR012340	81	1	0.315	0.357	1.00	Nucleic acid-binding, OB-fold
IPR029052	87	1	0.334	0.376	1.00	Metallo-dependent phosphatase-like
IPR011050	88	1	0.337	0.376	1.00	Pectin lyase fold/virulence factor
IPR008271	268	2	0.356	0.394	1.00	Serine/threonine-protein kinase, active site
IPR023779	96	1	0.362	0.394	1.00	Chromo domain, conserved site
IPR025724	96	1	0.362	0.394	1.00	GAG-pre-integrase domain
IPR003653	97	1	0.365	0.395	1.00	Ulp1 protease family, C-terminal catalytic domain
IPR012334	101	1	0.377	0.404	1.00	Pectin lyase fold
IPR011991	103	1	0.382	0.408	1.00	Winged helix-turn-helix DNA-binding domain
IPR027806	112	1	0.408	0.431	1.00	Harbinger transposase-derived nuclease domain
IPR000048	113	1	0.411	0.431	1.00	IQ motif, EF-hand binding site
IPR001611	129	1	0.453	0.473	1.00	Leucine-rich repeat
IPR001683	131	1	0.458	0.475	1.00	Phox homologous domain
IPR017853	144	1	0.490	0.504	1.00	Glycoside hydrolase superfamily
IPR005162	574	3	0.502	0.513	1.00	Retrotransposon gag domain
IPR006600	152	1	0.509	0.513	1.00	HTH CenpB-type DNA-binding domain
IPR029063	152	1	0.509	0.513	1.00	S-adenosyl-L-methionine-dependent methyltransferase
IPR003593	197	1	0.603	0.603	1.00	AAA+ ATPase domain

Genes with a *p*-value < 0.05 were accepted as statistically significant.

**Table A.4: Mapping rates of RNA-Seq data from *Phytophthora fragariae* in vitro grown mycelium and time points post inoculation with the BC-1 and NOV-9 isolates of *P. fragariae* on *Fragaria* × *ananassa* cultivar ‘Hapil’ plant roots against *de novo* assembled genomes of the BC-1, BC-16 and NOV-9 isolates of *P. fragariae***

Sample ID	Sample Source	Percentage Reads Mapped, unique sites and multiple sites		
		BC-1	BC-16	NOV-9
TA_BC1_P1	Inoculated roots, 48 hpi	30.73	29.28	29.28
TA_BC1_P2	Inoculated roots, 48 hpi	49.69	48.54	48.54
TA_BC1_P3	Inoculated roots, 48 hpi	22.94	22.47	22.47
TA_NOV9_P1	Inoculated roots, 72 hpi	34.66	34.66	35.75
TA_NOV9_P2	Inoculated roots, 72 hpi	14.08	14.08	14.51
TA_NOV9_P3	Inoculated roots, 72 hpi	50.84	50.84	52.18
TA_BC1_M1	Liquid grown mycelium	93.54	91.30	91.30
TA_BC1_M2	Liquid grown mycelium	92.84	90.40	90.40
TA_BC1_M3	Liquid grown mycelium	94.10	91.17	91.17
TA_NOV9_M1	Liquid grown mycelium	90.74	90.74	92.81
TA_NOV9_M2	Liquid grown mycelium	91.14	91.14	93.21
TA_NOV9_M5	Liquid grown mycelium	91.72	91.72	94.29

Hours post inoculation (hpi).

**Table A.5: All ANOVA and Tukey-HSD results for the strong candidate avirulence gene in BC-16**

Comparison	<i>p</i> -value adjusted	Test Used
<b>Isolate</b>	<b>0.00000000000676</b>	<b>ANOVA</b>
<b>Timepoint</b>	<b>0.0000384</b>	<b>ANOVA</b>
<b>Isolate:Timepoint</b>	<b>0.0000137</b>	<b>ANOVA</b>
<b>BC-1-A4</b>	<b>0.0395</b>	<b>Tukey-HSD</b>
<b>BC-16-A4</b>	<b>0.000000200</b>	<b>Tukey-HSD</b>
<b>NOV-9-A4</b>	<b>0.0487</b>	<b>Tukey-HSD</b>
<b>BC-16-BC-1</b>	<b>0.00</b>	<b>Tukey-HSD</b>
NOV-9-BC-1	1.00	Tukey-HSD
<b>NOV-9-BC-16</b>	<b>0.00</b>	<b>Tukey-HSD</b>
48hpi-24hpi	0.505	Tukey-HSD
72hpi-24hpi	0.969	Tukey-HSD
96hpi-24hpi	0.991	Tukey-HSD
<b>Mycelium-24hpi</b>	<b>0.00415</b>	<b>Tukey-HSD</b>
72hpi-48hpi	0.252	Tukey-HSD
96hpi-48hpi	0.198	Tukey-HSD
<b>Mycelium-48hpi</b>	<b>0.0000111</b>	<b>Tukey-HSD</b>
96hpi-72hpi	0.999	Tukey-HSD
Mycelium-72hpi	0.0691	Tukey-HSD
<b>Mycelium-96hpi</b>	<b>0.00665</b>	<b>Tukey-HSD</b>
BC-1:24hpi-A4:24hpi	1.00	Tukey-HSD
<b>BC-16:24hpi-A4:24hpi</b>	<b>0.00157</b>	<b>Tukey-HSD</b>
A4:48hpi-A4:24hpi	1.00	Tukey-HSD
BC-1:48hpi-A4:24hpi	1.00	Tukey-HSD
<b>BC-16:48hpi-A4:24hpi</b>	<b>0.000000300</b>	<b>Tukey-HSD</b>
NOV-9:48hpi-A4:24hpi	1.00	Tukey-HSD
BC-1:72hpi-A4:24hpi	1.00	Tukey-HSD
NOV-9:72hpi-A4:24hpi	1.00	Tukey-HSD
A4:96hpi-A4:24hpi	0.992	Tukey-HSD
BC-1:96hpi-A4:24hpi	1.00	Tukey-HSD
BC-16:96hpi-A4:24hpi	0.105	Tukey-HSD
NOV-9:96hpi-A4:24hpi	1.00	Tukey-HSD
A4:Mycelium-A4:24hpi	1.00	Tukey-HSD
BC-1:Mycelium-A4:24hpi	1.00	Tukey-HSD
BC-16:Mycelium-A4:24hpi	1.00	Tukey-HSD
NOV-9:Mycelium-A4:24hpi	1.00	Tukey-HSD
<b>BC-16:24hpi-BC-1:24hpi</b>	<b>0.000218</b>	<b>Tukey-HSD</b>
A4:48hpi-BC-1:24hpi	0.970	Tukey-HSD
BC-1:48hpi-BC-1:24hpi	1.00	Tukey-HSD
<b>BC-16:48hpi-BC-1:24hpi</b>	<b>0.00</b>	<b>Tukey-HSD</b>



NOV-9:48hpi-BC-1:24hpi	1.00	Tukey-HSD
BC-1:72hpi-BC-1:24hpi	1.00	Tukey-HSD
NOV-9:72hpi-BC-1:24hpi	1.00	Tukey-HSD
A4:96hpi-BC-1:24hpi	0.797	Tukey-HSD
BC-1:96hpi-BC-1:24hpi	1.00	Tukey-HSD
<b>BC-16:96hpi-BC-1:24hpi</b>	<b>0.0201</b>	<b>Tukey-HSD</b>
NOV-9:96hpi-BC-1:24hpi	1.00	Tukey-HSD
A4:Mycelium-BC-1:24hpi	1.00	Tukey-HSD
BC-1:Mycelium-BC-1:24hpi	1.00	Tukey-HSD
BC-16:Mycelium-BC-1:24hpi	1.00	Tukey-HSD
NOV-9:Mycelium-BC-1:24hpi	1.00	Tukey-HSD
<b>A4:48hpi-BC-16:24hpi</b>	<b>0.0235</b>	<b>Tukey-HSD</b>
<b>BC-1:48hpi-BC-16:24hpi</b>	<b>0.000117</b>	<b>Tukey-HSD</b>
BC-16:48hpi-BC-16:24hpi	0.271	Tukey-HSD
<b>NOV-9:48hpi-BC-16:24hpi</b>	<b>0.0000953</b>	<b>Tukey-HSD</b>
<b>BC-1:72hpi-BC-16:24hpi</b>	<b>0.0000655</b>	<b>Tukey-HSD</b>
<b>NOV-9:72hpi-BC-16:24hpi</b>	<b>0.000108</b>	<b>Tukey-HSD</b>
A4:96hpi-BC-16:24hpi	0.0763	Tukey-HSD
<b>BC-1:96hpi-BC-16:24hpi</b>	<b>0.0000623</b>	<b>Tukey-HSD</b>
BC-16:96hpi-BC-16:24hpi	0.979	Tukey-HSD
<b>NOV-9:96hpi-BC-16:24hpi</b>	<b>0.0000849</b>	<b>Tukey-HSD</b>
<b>A4:Mycelium-BC-16:24hpi</b>	<b>0.0000638</b>	<b>Tukey-HSD</b>
<b>BC-1:Mycelium-BC-16:24hpi</b>	<b>0.0000624</b>	<b>Tukey-HSD</b>
<b>BC-16:Mycelium-BC-16:24hpi</b>	<b>0.0000591</b>	<b>Tukey-HSD</b>
<b>NOV-9:Mycelium-BC-16:24hpi</b>	<b>0.0000595</b>	<b>Tukey-HSD</b>
BC-1:48hpi-A4:48hpi	0.918	Tukey-HSD
<b>BC-16:48hpi-A4:48hpi</b>	<b>0.00000450</b>	<b>Tukey-HSD</b>
NOV-9:48hpi-A4:48hpi	0.892	Tukey-HSD
BC-1:72hpi-A4:48hpi	0.834	Tukey-HSD
NOV-9:72hpi-A4:48hpi	0.908	Tukey-HSD
A4:96hpi-A4:48hpi	1.00	Tukey-HSD
BC-1:96hpi-A4:48hpi	0.825	Tukey-HSD
BC-16:96hpi-A4:48hpi	0.584	Tukey-HSD
NOV-9:96hpi-A4:48hpi	0.876	Tukey-HSD
A4:Mycelium-A4:48hpi	0.830	Tukey-HSD
BC-1:Mycelium-A4:48hpi	0.826	Tukey-HSD
BC-16:Mycelium-A4:48hpi	0.816	Tukey-HSD
NOV-9:Mycelium-A4:48hpi	0.817	Tukey-HSD
<b>BC-16:48hpi-BC-1:48hpi</b>	<b>0.00</b>	<b>Tukey-HSD</b>
NOV-9:48hpi-BC-1:48hpi	1.00	Tukey-HSD
BC-1:72hpi-BC-1:48hpi	1.00	Tukey-HSD

NOV-9:72hpi-BC-1:48hpi	1.00	Tukey-HSD
A4:96hpi-BC-1:48hpi	0.666	Tukey-HSD
BC-1:96hpi-BC-1:48hpi	1.00	Tukey-HSD
<b>BC-16:96hpi-BC-1:48hpi</b>	<b>0.0115</b>	<b>Tukey-HSD</b>
NOV-9:96hpi-BC-1:48hpi	1.00	Tukey-HSD
A4:Mycelium-BC-1:48hpi	1.00	Tukey-HSD
BC-1:Mycelium-BC-1:48hpi	1.00	Tukey-HSD
BC-16:Mycelium-BC-1:48hpi	1.00	Tukey-HSD
NOV-9:Mycelium-BC-1:48hpi	1.00	Tukey-HSD
<b>NOV-9:48hpi-BC-16:48hpi</b>	<b>0.00</b>	<b>Tukey-HSD</b>
<b>BC-1:72hpi-BC-16:48hpi</b>	<b>0.00</b>	<b>Tukey-HSD</b>
<b>NOV-9:72hpi-BC-16:48hpi</b>	<b>0.00</b>	<b>Tukey-HSD</b>
A4:96hpi-BC-16:48hpi	0.0000180	Tukey-HSD
<b>BC-1:96hpi-BC-16:48hpi</b>	<b>0.00</b>	<b>Tukey-HSD</b>
<b>BC-16:96hpi-BC-16:48hpi</b>	<b>0.00589</b>	<b>Tukey-HSD</b>
NOV-9:96hpi-BC-16:48hpi	0.00	Tukey-HSD
A4:Mycelium-BC-16:48hpi	0.00	Tukey-HSD
<b>BC-1:Mycelium-BC-16:48hpi</b>	<b>0.00</b>	<b>Tukey-HSD</b>
<b>BC-16:Mycelium-BC-16:48hpi</b>	<b>0.00</b>	<b>Tukey-HSD</b>
<b>NOV-9:Mycelium-BC-16:48hpi</b>	<b>0.00</b>	<b>Tukey-HSD</b>
BC-1:72hpi-NOV-9:48hpi	1.00	Tukey-HSD
NOV-9:72hpi-NOV-9:48hpi	1.00	Tukey-HSD
A4:96hpi-NOV-9:48hpi	0.619	Tukey-HSD
BC-1:96hpi-NOV-9:48hpi	1.00	Tukey-HSD
<b>BC-16:96hpi-NOV-9:48hpi</b>	<b>0.00950</b>	<b>Tukey-HSD</b>
NOV-9:96hpi-NOV-9:48hpi	1.00	Tukey-HSD
A4:Mycelium-NOV-9:48hpi	1.00	Tukey-HSD
BC-1:Mycelium-NOV-9:48hpi	1.00	Tukey-HSD
BC-16:Mycelium-NOV-9:48hpi	1.00	Tukey-HSD
NOV-9:Mycelium-NOV-9:48hpi	1.00	Tukey-HSD
NOV-9:72hpi-BC-1:72hpi	1.00	Tukey-HSD
A4:96hpi-BC-1:72hpi	0.533	Tukey-HSD
BC-1:96hpi-BC-1:72hpi	1.00	Tukey-HSD
<b>BC-16:96hpi-BC-1:72hpi</b>	<b>0.00669</b>	<b>Tukey-HSD</b>
NOV-9:96hpi-BC-1:72hpi	1.00	Tukey-HSD
A4:Mycelium-BC-1:72hpi	1.00	Tukey-HSD
BC-1:Mycelium-BC-1:72hpi	1.00	Tukey-HSD
BC-16:Mycelium-BC-1:72hpi	1.00	Tukey-HSD
NOV-9:Mycelium-BC-1:72hpi	1.00	Tukey-HSD
A4:96hpi-NOV-9:72hpi	0.648	Tukey-HSD
BC-1:96hpi-NOV-9:72hpi	1.00	Tukey-HSD

<b>BC-16:96hpi-NOV-9:72hpi</b>	<b>0.0107</b>	<b>Tukey-HSD</b>
NOV-9:96hpi-NOV-9:72hpi	1.00	Tukey-HSD
A4:Mycelium-NOV-9:72hpi	1.00	Tukey-HSD
BC-1:Mycelium-NOV-9:72hpi	1.00	Tukey-HSD
BC-16:Mycelium-NOV-9:72hpi	1.00	Tukey-HSD
NOV-9:Mycelium-NOV-9:72hpi	1.00	Tukey-HSD
BC-1:96hpi-A4:96hpi	0.521	Tukey-HSD
BC-16:96hpi-A4:96hpi	0.870	Tukey-HSD
NOV-9:96hpi-A4:96hpi	0.593	Tukey-HSD
A4:Mycelium-A4:96hpi	0.527	Tukey-HSD
BC-1:Mycelium-A4:96hpi	0.522	Tukey-HSD
BC-16:Mycelium-A4:96hpi	0.509	Tukey-HSD
NOV-9:Mycelium-A4:96hpi	0.511	Tukey-HSD
<b>BC-16:96hpi-BC-1:96hpi</b>	<b>0.00639</b>	<b>Tukey-HSD</b>
NOV-9:96hpi-BC-1:96hpi	1.00	Tukey-HSD
A4:Mycelium-BC-1:96hpi	1.00	Tukey-HSD
BC-1:Mycelium-BC-1:96hpi	1.00	Tukey-HSD
BC-16:Mycelium-BC-1:96hpi	1.00	Tukey-HSD
NOV-9:Mycelium-BC-1:96hpi	1.00	Tukey-HSD
<b>NOV-9:96hpi-BC-16:96hpi</b>	<b>0.00853</b>	<b>Tukey-HSD</b>
<b>A4:Mycelium-BC-16:96hpi</b>	<b>0.00653</b>	<b>Tukey-HSD</b>
<b>BC-1:Mycelium-BC-16:96hpi</b>	<b>0.00639</b>	<b>Tukey-HSD</b>
<b>BC-16:Mycelium-BC-16:96hpi</b>	<b>0.00608</b>	<b>Tukey-HSD</b>
<b>NOV-9:Mycelium-BC-16:96hpi</b>	<b>0.00611</b>	<b>Tukey-HSD</b>
A4:Mycelium-NOV-9:96hpi	1.00	Tukey-HSD
BC-1:Mycelium-NOV-9:96hpi	1.00	Tukey-HSD
BC-16:Mycelium-NOV-9:96hpi	1.00	Tukey-HSD
NOV-9:Mycelium-NOV-9:96hpi	1.00	Tukey-HSD
BC-1:Mycelium-A4:Mycelium	1.00	Tukey-HSD
BC-16:Mycelium-A4:Mycelium	1.00	Tukey-HSD
NOV-9:Mycelium-A4:Mycelium	1.00	Tukey-HSD
BC-16:Mycelium-BC-1:Mycelium	1.00	Tukey-HSD
NOV-9:Mycelium-BC-1:Mycelium	1.00	Tukey-HSD
NOV-9:Mycelium-BC-16:Mycelium	1.00	Tukey-HSD

Significant tests in bold.

**Table A.6: All ANOVA and Tukey-HSD results for an RxLR effector predicted to peak at 24 hpi**

Comparison	<i>p</i> -value adjusted	Test Used
Isolate	0.964	ANOVA
<b>Timepoint</b>	<b>0.0000457</b>	<b>ANOVA</b>
Isolate:Timepoint	0.0679	ANOVA
BC-16-BC-1	0.997	Tukey HSD
NOV-9-BC-1	0.962	Tukey HSD
NOV-9-BC-16	0.981	Tukey HSD
48hpi-24pi	0.802	Tukey HSD
72hpi-24hpi	0.924	Tukey HSD
96hpi-24hpi	0.863	Tukey HSD
<b>Mycelium-24hpi</b>	<b>0.00174</b>	<b>Tukey HSD</b>
72hpi-48hpi	0.295	Tukey HSD
96hpi-48hpi	0.168	Tukey HSD
<b>Mycelium-48hpi</b>	<b>0.0000193</b>	<b>Tukey HSD</b>
96hpi-72hpi	1.00	Tukey HSD
<b>Mycelium-72hpi</b>	<b>0.0159</b>	<b>Tukey HSD</b>
<b>Mycelium-96hpi</b>	<b>0.00754</b>	<b>Tukey HSD</b>
BC-16:24hpi-BC-1:24hpi	0.962	Tukey HSD
NOV-9:24hpi-BC-1:24hpi	NA	Tukey HSD
BC-1:48hpi-BC-1:24hpi	0.960	Tukey HSD
BC-16:48hpi-BC-1:24hpi	1.00	Tukey HSD
NOV-9:48hpi-BC-1:24hpi	0.993	Tukey HSD
BC-1:72hpi-BC-1:24hpi	1.00	Tukey HSD
BC-16:72hpi-BC-1:24hpi	NA	Tukey HSD
NOV-9:72hpi-BC-1:24hpi	0.716	Tukey HSD
BC-1:96hpi-BC-1:24hpi	0.830	Tukey HSD
BC-16:96hpi-BC-1:24hpi	1.00	Tukey HSD
NOV-9:96hpi-BC-1:24hpi	0.994	Tukey HSD
BC-1:Mycelium-BC-1:24hpi	0.0756	Tukey HSD
BC-16:Mycelium-BC-1:24hpi	0.0627	Tukey HSD
NOV-9:Mycelium-BC-1:24hpi	0.109	Tukey HSD
NOV-9:24hpi-BC-16:24hpi	NA	Tukey HSD
BC-1:48hpi-BC-16:24hpi	1.00	Tukey HSD
BC-16:48hpi-BC-16:24hpi	0.888	Tukey HSD
NOV-9:48hpi-BC-16:24hpi	0.322	Tukey HSD
BC-1:72hpi-BC-16:24hpi	0.999	Tukey HSD
BC-16:72hpi-BC-16:24hpi	NA	Tukey HSD
NOV-9:72hpi-BC-16:24hpi	1.00	Tukey HSD
BC-1:96hpi-BC-16:24hpi	1.00	Tukey HSD
BC-16:96hpi-BC-16:24hpi	1.00	Tukey HSD

BC-16:Mycelium-BC-16:24hpi	0.718	Tukey HSD
NOV-9:Mycelium-BC-16:24hpi	0.859	Tukey HSD
BC-1:48hpi-NOV-9:24hpi	NA	Tukey HSD
BC-16:48hpi-NOV-9:24hpi	NA	Tukey HSD
NOV-9:48hpi-NOV-9:24hpi	NA	Tukey HSD
BC-1:72hpi-NOV-9:24hpi	NA	Tukey HSD
BC-16:72hpi-NOV-9:24hpi	NA	Tukey HSD
NOV-9:72hpi-NOV-9:24hpi	NA	Tukey HSD
BC-1:96hpi-NOV-9:24hpi	NA	Tukey HSD
BC-16:96hpi-NOV-9:24hpi	NA	Tukey HSD
NOV-9:96hpi-NOV-9:24hpi	NA	Tukey HSD
BC-1:Mycelium-NOV-9:24hpi	NA	Tukey HSD
BC-16:Mycelium-NOV-9:24hpi	NA	Tukey HSD
NOV-9:Mycelium-NOV-9:24hpi	NA	Tukey HSD
BC-16:48hpi-BC-1:48hpi	0.884	Tukey HSD
NOV-9:48hpi-BC-1:48hpi	0.316	Tukey HSD
BC-1:72hpi-BC-1:48hpi	0.998	Tukey HSD
BC-16:72hpi-BC-1:48hpi	NA	Tukey HSD
NOV-9:72hpi-BC-1:48hpi	1.00	Tukey HSD
BC-1:96hpi-BC-1:48hpi	1.00	Tukey HSD
BC-16:96hpi-BC-1:48hpi	1.00	Tukey HSD
NOV-9:96hpi-BC-1:48hpi	1.00	Tukey HSD
BC-1:Mycelium-BC-1:48hpi	0.775	Tukey HSD
BC-16:Mycelium-BC-1:48hpi	0.725	Tukey HSD
NOV-9:Mycelium-BC-1:48hpi	0.864	Tukey HSD
NOV-9:48hpi-BC-16:48hpi	0.999	Tukey HSD
BC-1:72hpi-BC-16:48hpi	1.00	Tukey HSD
BC-16:72hpi-BC-16:48hpi	NA	Tukey HSD
NOV-9:72hpi-BC-16:48hpi	0.555	Tukey HSD
BC-1:96hpi-BC-16:48hpi	0.686	Tukey HSD
BC-16:96hpi-BC-16:48hpi	0.995	Tukey HSD
NOV-9:96hpi-BC-16:48hpi	0.969	Tukey HSD
<b>BC-1:Mycelium-BC-16:48hpi</b>	<b>0.0433</b>	<b>Tukey HSD</b>
<b>BC-16:Mycelium-BC-16:48hpi</b>	<b>0.0356</b>	<b>Tukey HSD</b>
NOV-9:Mycelium-BC-16:48hpi	0.0639	Tukey HSD
BC-1:72hpi-NOV-9:48hpi	0.911	Tukey HSD
BC-16:72hpi-NOV-9:48hpi	NA	Tukey HSD
NOV-9:72hpi-NOV-9:48hpi	0.110	Tukey HSD
BC-1:96hpi-NOV-9:48hpi	0.164	Tukey HSD
BC-16:96hpi-NOV-9:48hpi	0.675	Tukey HSD
NOV-9:96hpi-NOV-9:48hpi	0.492	Tukey HSD
<b>BC-1:Mycelium-NOV-9:48hpi</b>	<b>0.00410</b>	<b>Tukey HSD</b>
<b>BC-16:Mycelium-NOV-9:48hpi</b>	<b>0.00330</b>	<b>Tukey HSD</b>

<b>NOV-9:Mycelium-NOV-9:48hpi</b>	<b>0.00634</b>	<b>Tukey HSD</b>
BC-16:72hpi-BC-1:72hpi	NA	Tukey HSD
NOV-9:72hpi-BC-1:72hpi	0.930	Tukey HSD
BC-1:96hpi-BC-1:72hpi	0.974	Tukey HSD
BC-16:96hpi-BC-1:72hpi	1.00	Tukey HSD
NOV-9:96hpi-BC-1:72hpi	1.00	Tukey HSD
BC-1:Mycelium-BC-1:72hpi	0.189	Tukey HSD
BC-16:Mycelium-BC-1:72hpi	0.161	Tukey HSD
NOV-9:Mycelium-BC-1:72hpi	0.258	Tukey HSD
NOV-9:72hpi-BC-16:72hpi	NA	Tukey HSD
BC-1:96hpi-BC-16:72hpi	NA	Tukey HSD
BC-16:96hpi-BC-16:72hpi	NA	Tukey HSD
NOV-9:96hpi-BC-16:72hpi	NA	Tukey HSD
BC-1:Mycelium-BC-16:72hpi	NA	Tukey HSD
BC-16:Mycelium-BC-16:72hpi	NA	Tukey HSD
NOV-9:Mycelium-BC-16:72hpi	NA	Tukey HSD
BC-1:96hpi-NOV-9:72hpi	1.00	Tukey HSD
BC-16:96hpi-NOV-9:72hpi	0.996	Tukey HSD
NOV-9:96hpi-NOV-9:72hpi	1.00	Tukey HSD
BC-1:Mycelium-NOV-9:72hpi	0.976	Tukey HSD
BC-16:Mycelium-NOV-9:72hpi	0.962	Tukey HSD
NOV-9:Mycelium-NOV-9:72hpi	0.992	Tukey HSD
BC-16:96hpi-BC-1:96hpi	0.999	Tukey HSD
NOV-9:96hpi-BC-1:96hpi	1.00	Tukey HSD
BC-1:Mycelium-BC-1:96hpi	0.935	Tukey HSD
BC-16:Mycelium-BC-1:96hpi	0.908	Tukey HSD
NOV-9:Mycelium-BC-1:96hpi	0.972	Tukey HSD
NOV-9:96hpi-BC-16:96hpi	1.00	Tukey HSD
BC-1:Mycelium-BC-16:96hpi	0.408	Tukey HSD
BC-16:Mycelium-BC-16:96hpi	0.359	Tukey HSD
NOV-9:Mycelium-BC-16:96hpi	0.515	Tukey HSD
BC-1:Mycelium-NOV-9:96hpi	0.586	Tukey HSD
BC-16:Mycelium-NOV-9:96hpi	0.530	Tukey HSD
NOV-9:Mycelium-NOV-9:96hpi	0.698	Tukey HSD
BC-16:Mycelium-BC-1:Mycelium	1.00	Tukey HSD
NOV-9:Mycelium-BC-1:Mycelium	1.00	Tukey HSD
<b>NOV-9:Mycelium-BC-16:Mycelium</b>	<b>1.00</b>	<b>Tukey HSD</b>

Significant tests in bold.

**Table A.7: All ANOVA and Tukey-HSD results for an RxLR effector predicted to peak at 48 hpi**

Comparison	<i>p</i> -value adjusted	Test Used
Isolate	0.858	ANOVA
Timepoint	0.374	ANOVA
Isolate:Timepoint	0.0991	ANOVA
BC-16-BC-1	0.931	Tukey HSD
NOV-9-BC-1	0.853	Tukey HSD
NOV-9-BC-16	0.984	Tukey HSD
48hpi-24hpi	0.503	Tukey HSD
72hpi-24hpi	0.752	Tukey HSD
96hpi-24hpi	0.469	Tukey HSD
Mycelium-24hpi	0.350	Tukey HSD
72hpi-48hpi	0.998	Tukey HSD
96hpi-48hpi	1.00	Tukey HSD
Mycelium-48hpi	0.998	Tukey HSD
96hpi-72hpi	0.996	Tukey HSD
Mycelium-72hpi	0.978	Tukey HSD
Mycelium-96hpi	0.999	Tukey HSD
BC-16:24hpi-BC-1:24hpi	1.00	Tukey HSD
NOV-9:24hpi-BC-1:24hpi	NA	Tukey HSD
BC-1:48hpi-BC-1:24hpi	1.00	Tukey HSD
BC-16:48hpi-BC-1:24hpi	1.00	Tukey HSD
NOV-9:48hpi-BC-1:24hpi	0.902	Tukey HSD
BC-1:72hpi-BC-1:24hpi	0.998	Tukey HSD
BC-16:72hpi-BC-1:24hpi	NA	Tukey HSD
NOV-9:72hpi-BC-1:24hpi	1.00	Tukey HSD
BC-1:96hpi-BC-1:24hpi	1.00	Tukey HSD
BC-16:96hpi-BC-1:24hpi	0.942	Tukey HSD
NOV-9:96hpi-BC-1:24hpi	1.00	Tukey HSD
BC-1:Mycelium-BC-1:24hpi	0.306	Tukey HSD
BC-16:Mycelium-BC-1:24hpi	1.00	Tukey HSD
NOV-9:Mycelium-BC-1:24hpi	1.00	Tukey HSD
NOV-9:24hpi-BC-16:24hpi	NA	Tukey HSD
BC-1:48hpi-BC-16:24hpi	1.00	Tukey HSD
BC-16:48hpi-BC-16:24hpi	1.00	Tukey HSD
NOV-9:48hpi-BC-16:24hpi	0.874	Tukey HSD
BC-1:72hpi-BC-16:24hpi	0.995	Tukey HSD
BC-16:72hpi-BC-16:24hpi	NA	Tukey HSD
NOV-9:72hpi-BC-16:24hpi	1.00	Tukey HSD
BC-1:96hpi-BC-16:24hpi	1.00	Tukey HSD
BC-16:96hpi-BC-16:24hpi	0.921	Tukey HSD

NOV-9:96hpi-BC-16:24hpi	1.00	Tukey HSD
BC-1:Mycelium-BC-16:24hpi	0.272	Tukey HSD
BC-16:Mycelium-BC-16:24hpi	1.00	Tukey HSD
NOV-9:Mycelium-BC-16:24hpi	1.00	Tukey HSD
BC-1:48hpi-NOV-9:24hpi	NA	Tukey HSD
BC-16:48hpi-NOV-9:24hpi	NA	Tukey HSD
NOV-9:48hpi-NOV-9:24hpi	NA	Tukey HSD
BC-1:72hpi-NOV-9:24hpi	NA	Tukey HSD
BC-16:72hpi-NOV-9:24hpi	NA	Tukey HSD
NOV-9:72hpi-NOV-9:24hpi	NA	Tukey HSD
BC-1:96hpi-NOV-9:24hpi	NA	Tukey HSD
BC-16:96hpi-NOV-9:24hpi	NA	Tukey HSD
NOV-9:96hpi-NOV-9:24hpi	NA	Tukey HSD
BC-1:Mycelium-NOV-9:24hpi	NA	Tukey HSD
BC-16:Mycelium-NOV-9:24hpi	NA	Tukey HSD
NOV-9:Mycelium-NOV-9:24hpi	NA	Tukey HSD
BC-16:48hpi-BC-1:48hpi	1.00	Tukey HSD
NOV-9:48hpi-BC-1:48hpi	0.968	Tukey HSD
BC-1:72hpi-BC-1:48hpi	1.00	Tukey HSD
BC-16:72hpi-BC-1:48hpi	NA	Tukey HSD
NOV-9:72hpi-BC-1:48hpi	1.00	Tukey HSD
BC-1:96hpi-BC-1:48hpi	1.00	Tukey HSD
BC-16:96hpi-BC-1:48hpi	0.985	Tukey HSD
NOV-9:96hpi-BC-1:48hpi	1.00	Tukey HSD
BC-1:Mycelium-BC-1:48hpi	0.447	Tukey HSD
BC-16:Mycelium-BC-1:48hpi	1.00	Tukey HSD
NOV-9:Mycelium-BC-1:48hpi	1.00	Tukey HSD
NOV-9:48hpi-BC-16:48hpi	1.00	Tukey HSD
BC-1:72hpi-BC-16:48hpi	1.00	Tukey HSD
BC-16:72hpi-BC-16:48hpi	NA	Tukey HSD
NOV-9:72hpi-BC-16:48hpi	1.00	Tukey HSD
BC-1:96hpi-BC-16:48hpi	1.00	Tukey HSD
BC-16:96hpi-BC-16:48hpi	1.00	Tukey HSD
NOV-9:96hpi-BC-16:48hpi	1.00	Tukey HSD
BC-1:Mycelium-BC-16:48hpi	0.809	Tukey HSD
BC-16:Mycelium-BC-16:48hpi	1.00	Tukey HSD
NOV-9:Mycelium-BC-16:48hpi	0.999	Tukey HSD
BC-1:72hpi-NOV-9:48hpi	1.00	Tukey HSD
BC-16:72hpi-NOV-9:48hpi	NA	Tukey HSD
NOV-9:72hpi-NOV-9:48hpi	0.988	Tukey HSD
BC-1:96hpi-NOV-9:48hpi	0.996	Tukey HSD
BC-16:96hpi-NOV-9:48hpi	1.00	Tukey HSD
NOV-9:96hpi-NOV-9:48hpi	0.999	Tukey HSD



BC-1:Mycelium-NOV-9:48hpi	0.998	Tukey HSD
BC-16:Mycelium-NOV-9:48hpi	0.999	Tukey HSD
NOV-9:Mycelium-NOV-9:48hpi	0.836	Tukey HSD
BC-16:72hpi-BC-1:72hpi	NA	Tukey HSD
NOV-9:72hpi-BC-1:72hpi	1.00	Tukey HSD
BC-1:96hpi-BC-1:72hpi	1.00	Tukey HSD
BC-16:96hpi-BC-1:72hpi	1.00	Tukey HSD
NOV-9:96hpi-BC-1:72hpi	1.00	Tukey HSD
BC-1:Mycelium-BC-1:72hpi	0.919	Tukey HSD
BC-16:Mycelium-BC-1:72hpi	1.00	Tukey HSD
NOV-9:Mycelium-BC-1:72hpi	0.991	Tukey HSD
NOV-9:72hpi-BC-16:72hpi	NA	Tukey HSD
BC-1:96hpi-BC-16:72hpi	NA	Tukey HSD
BC-16:96hpi-BC-16:72hpi	NA	Tukey HSD
NOV-9:96hpi-BC-16:72hpi	NA	Tukey HSD
BC-1:Mycelium-BC-16:72hpi	NA	Tukey HSD
BC-16:Mycelium-BC-16:72hpi	NA	Tukey HSD
NOV-9:Mycelium-BC-16:72hpi	NA	Tukey HSD
BC-1:96hpi-NOV-9:72hpi	1.00	Tukey HSD
BC-16:96hpi-NOV-9:72hpi	0.996	Tukey HSD
NOV-9:96hpi-NOV-9:72hpi	1.00	Tukey HSD
BC-1:Mycelium-NOV-9:72hpi	0.552	Tukey HSD
BC-16:Mycelium-NOV-9:72hpi	1.00	Tukey HSD
NOV-9:Mycelium-NOV-9:72hpi	1.00	Tukey HSD
BC-16:96hpi-BC-1:96hpi	0.999	Tukey HSD
NOV-9:96hpi-BC-1:96hpi	1.00	Tukey HSD
BC-1:Mycelium-BC-1:96hpi	0.649	Tukey HSD
BC-16:Mycelium-BC-1:96hpi	1.00	Tukey HSD
NOV-9:Mycelium-BC-1:96hpi	1.00	Tukey HSD
NOV-9:96hpi-BC-16:96hpi	1.00	Tukey HSD
BC-1:Mycelium-BC-16:96hpi	0.995	Tukey HSD
BC-16:Mycelium-BC-16:96hpi	1.00	Tukey HSD
NOV-9:Mycelium-BC-16:96hpi	0.892	Tukey HSD
BC-1:Mycelium-NOV-9:96hpi	0.765	Tukey HSD
BC-16:Mycelium-NOV-9:96hpi	1.00	Tukey HSD
NOV-9:Mycelium-NOV-9:96hpi	1.00	Tukey HSD
BC-16:Mycelium-BC-1:Mycelium	0.741	Tukey HSD
NOV-9:Mycelium-BC-1:Mycelium	0.236	Tukey HSD
<b>NOV-9:Mycelium-BC-16:Mycelium</b>	<b>1.00</b>	<b>Tukey HSD</b>

Significant tests in bold.

**Table A.8: Gene names of all genes identified as putative candidate avirulence genes**

High	BC-1		High	BC-16		High	NOV-9	
	Medium	Low		Medium	Low		Medium	Low
	g4972.t1	g1097.t1	g370.t1	g161.t1	g103.t1		g229.t1	g510.t1
	g5367.t1	g1250.t1	g6872.t1	g344.t1	g854.t1		g1306.t1	g633.t1
	g20984.t1	g1521.t1	g8968.t1	g415.t1	g1023.t1		g1759.t1	g925.t1
	g13347.t2	g1879.t1	g9383.t1	g565.t1	g1333.t1		g1867.t1	g1067.t1
		g1901.t1	g9855.t1	g615.t1	g1837.t1		g2507.t1	g1303.t1
		g2000.t1	g11145.t1	g829.t1	g2377.t1		g4097.t1	g1396.t1
		g2236.t1	g14271.t1	g850.t1	g2605.t1		g4330.t1	g1594.t1
		g2237.t1	g14290.t1	g1533.t1	g3089.t1		g4449.t1	g1601.t1
		g4406.t1	g17733.t1	g1668.t1	g3706.t1		g5545.t1	g1615.t1
		g4421.t1	g25743.t1	g1720.t1	g4033.t1		g6999.t1	g1623.t1
		g4507.t1	g27961.t1	g1932.t1	g4205.t1		g7047.t1	g1631.t1
		g4730.t1	g30505.t1	g2286.t1	g4209.t1		g7413.t1	g1758.t1
		g6448.t1	g33857.t1	g2531.t1	g4572.t1		g9940.t1	g1955.t1
		g6482.t1	g35474.t1	g2844.t1	g4714.t1		g11126.t1	g2025.t1
		g6653.t1	g35765.t1	g2941.t1	g5022.t1		g12024.t1	g2032.t1
		g7036.t1	g35940.t1	g3466.t1	g5657.t1		g12369.t1	g2235.t1
		g7166.t1	g39572.t1	g3615.t1	g6268.t1		g12447.t1	g3113.t1
		g7294.t1		g3666.t1	g6891.t1		g13975.t1	g3425.t1
		g8065.t1		g3747.t1	g7336.t1		g13977.t1	g3710.t1
		g8453.t1		g3781.t1	g7578.t1		g16217.t1	g3946.t1
		g8634.t1		g3877.t1	g7910.t1		g16995.t1	g4037.t1
		g9217.t1		g4005.t1	g8418.t1		g17295.t1	g4098.t1

g9223.t1	g4262.t1	g8479.t1	g17824.t1	g4197.t1
g9555.t1	g4559.t1	g8488.t1	g18495.t1	g4268.t1
g9765.t1	g4601.t1	g8612.t1	g18743.t1	g4281.t1
g10489.t1	g4609.t1	g9021.t1	g19312.t1	g4583.t1
g10754.t1	g4756.t1	g9302.t1	g19584.t1	g4921.t1
g11311.t1	g4911.t1	g9904.t1	g19586.t1	g5256.t1
g11817.t1	g4966.t1	g10114.t1	g19681.t1	g5791.t1
g11818.t1	g4988.t1	g10975.t1	g21394.t1	g6205.t1
g12474.t1	g5040.t1	g11147.t1	g21655.t1	g6249.t1
g12833.t1	g5169.t1	g11426.t1	g21662.t1	g6332.t1
g12880.t1	g5250.t1	g12759.t1	g22173.t1	g6386.t1
g12910.t1	g5573.t1	g13235.t1	g22652.t1	g6956.t1
g13157.t1	g5762.t1	g13528.t1	g22887.t1	g7048.t1
g13922.t1	g5893.t1	g15118.t1	g23285.t1	g7698.t1
g14044.t1	g6829.t1	g15205.t1	g23699.t1	g7949.t3
g14225.t1	g6984.t1	g15499.t1	g23735.t1	g7970.t1
g14740.t1	g6989.t1	g16398.t1	g23749.t1	g8060.t1
g14901.t1	g7325.t1	g16436.t1	g23756.t1	g8115.t1
g15056.t1	g7421.t1	g17334.t1	g23759.t1	g8314.t1
g15306.t1	g7590.t1	g17617.t1	g24190.t1	g8321.t1
g15622.t1	g7920.t1	g19159.t1	g24350.t1	g8322.t1
g15623.t1	g8061.t1	g20060.t1	g24741.t1	g8450.t1
g15968.t1	g8166.t1	g20576.t1	g29071.t1	g8451.t1
g16294.t1	g8350.t1	g20785.t1	g31899.t1	g8456.t1
g16403.t1	g8422.t1	g20821.t1	g31911.t1	g8457.t1
g16404.t1	g8467.t1	g21213.t1	g32962.t1	g9247.t1
g16865.t1	g8696.t1	g22711.t1	g33209.t1	g9652.t1

g17221.t1	g8774.t1	g22714.t1	g9691.t1
g17642.t1	g8851.t1	g23959.t1	g10033.t1
g17733.t1	g9159.t1	g24070.t1	g10365.t1
g17906.t1	g9423.t1	g24164.t1	g10696.t1
g17926.t1	g9470.t1	g24382.t1	g10701.t1
g17954.t1	g9599.t1	g25214.t1	g11219.t1
g17991.t1	g9606.t1	g25239.t1	g11405.t1
g18256.t1	g9609.t1	g26385.t1	g11550.t1
g18282.t1	g9681.t1	g26605.t1	g11675.t1
g18290.t1	g9931.t1	g26631.t1	g11676.t1
g18414.t1	g10452.t1	g27252.t1	g11898.t1
g18742.t1	g10466.t1	g27464.t1	g12025.t1
g18753.t1	g10472.t1	g28110.t1	g12227.t1
g18758.t1	g10683.t1	g28448.t1	g12233.t1
g19032.t1	g11102.t1	g28479.t1	g12608.t1
g19102.t1	g11114.t1	g29197.t1	g12790.t1
g19110.t1	g11814.t1	g29646.t1	g12839.t1
g19219.t1	g11881.t1	g29759.t1	g12844.t1
g19407.t1	g11882.t1	g29765.t1	g13119.t1
g19550.t1	g12013.t1	g30352.t1	g13156.t1
g19566.t1	g12066.t1	g30592.t1	g13178.t1
g19777.t1	g12217.t1	g31101.t1	g13184.t1
g19794.t1	g12265.t1	g31139.t1	g13802.t1
g19795.t1	g13068.t1	g31383.t1	g13970.t1
g19807.t1	g13533.t1	g31524.t1	g13973.t1
g19875.t1	g13936.t1	g31565.t1	g14450.t1
g20061.t1	g14208.t1	g31730.t1	g14533.t1

g20431.t1	g14280.t1	g32042.t1	g15015.t1
g20460.t1	g14309.t1	g32469.t1	g15080.t1
g21245.t1	g14400.t1	g32681.t1	g15098.t1
g21252.t1	g14493.t1	g32696.t1	g16205.t1
g21282.t1	g14554.t1	g32700.t1	g16754.t1
g21348.t1	g14578.t1	g34865.t1	g16987.t1
g21838.t1	g14706.t1	g34956.t1	g17262.t1
g22332.t1	g14855.t1	g35112.t1	g17436.t1
g22342.t1	g15109.t1	g36743.t1	g17508.t1
g22546.t1	g15203.t1	g37324.t1	g17668.t1
g22741.t1	g15377.t1	g37554.t1	g17771.t1
g22771.t1	g15378.t1	g37884.t1	g17773.t1
g22897.t1	g15427.t1	g38072.t1	g17798.t1
g23072.t1	g15431.t1	g38282.t1	g18082.t1
g23172.t1	g15449.t1	g38472.t1	g18162.t1
g23749.t1	g15462.t1	g38811.t1	g18389.t1
g24005.t1	g15660.t1	g38937.t1	g18424.t1
g24017.t1	g15663.t1	g39022.t1	g18490.t1
g24020.t1	g16428.t1	g39164.t1	g18710.t1
g24076.t1	g16522.t1	g39367.t1	g18730.t1
g24134.t1	g16580.t1		g19079.t1
g24398.t1	g16684.t1		g19116.t1
g24530.t1	g16749.t1		g19117.t1
g25017.t1	g16751.t1		g19191.t1
g25099.t1	g16853.t1		g19196.t1
g26294.t1	g17022.t1		g19200.t1
g26701.t1	g17032.t1		g19256.t1

g26875.t1	g17034.t1	g19288.t1
g27325.t1	g17379.t1	g19296.t1
g27405.t1	g17381.t1	g19421.t1
g27515.t1	g17768.t1	g19422.t1
g27533.t1	g17953.t1	g19520.t1
g28068.t1	g18573.t1	g19557.t1
g29021.t1	g18648.t1	g19940.t1
g29092.t1	g18918.t1	g20156.t1
g29631.t1	g19083.t1	g20157.t1
g30226.t1	g19389.t1	g20179.t1
g30319.t1	g19792.t1	g20584.t1
g30414.t1	g19803.t1	g20767.t1
g31019.t1	g20192.t1	g21417.t1
g31067.t1	g20308.t1	g21508.t1
g31440.t1	g20357.t1	g21551.t1
g33213.t1	g20394.t1	g21565.t1
g33351.t1	g20970.t1	g21722.t1
	g21125.t1	g21927.t1
	g21550.t1	g22175.t1
	g22174.t1	g22198.t1
	g22538.t1	g22734.t1
	g22548.t1	g22762.t1
	g22736.t1	g22858.t1
	g22739.t1	g23130.t1
	g22748.t1	g23132.t1
	g23314.t1	g23149.t1
	g23813.t1	g23322.t1

g24059.t1  
g24093.t1  
g24114.t1  
g24603.t1  
g24804.t1  
g25229.t1  
g25891.t1  
g26029.t1  
g26340.t1  
g26710.t1  
g26845.t1  
g26898.t1  
g26903.t1  
g27152.t1  
g27155.t1  
g27513.t1  
g27518.t1  
g27654.t1  
g27710.t1  
g27717.t1  
g28252.t1  
g28356.t1  
g28363.t1  
g28601.t1  
g29058.t1  
g29166.t1  
g29382.t1

g23347.t1  
g23513.t1  
g23758.t1  
g24113.t1  
g24248.t1  
g24261.t1  
g24331.t1  
g24337.t1  
g24496.t1  
g24630.t1  
g24872.t1  
g24886.t1  
g25332.t1  
g25370.t1  
g25414.t1  
g25710.t1  
g25906.t1  
g26030.t1  
g26276.t1  
g26293.t1  
g26402.t1  
g26741.t1  
g26751.t1  
g26787.t1  
g27314.t1  
g27623.t1  
g28086.t1

g29475.t1  
g29673.t1  
g30071.t1  
g30330.t1  
g30549.t1  
g31140.t1  
g31180.t1  
g31182.t1  
g31343.t1  
g31447.t1  
g31525.t1  
g31843.t1  
g31961.t1  
g31996.t1  
g32122.t1  
g32218.t1  
g32776.t1  
g32924.t1  
g32978.t1  
g33068.t1  
g33270.t1  
g33280.t1  
g33794.t1  
g33816.t1  
g33956.t1  
g34072.t1  
g34498.t1

g28096.t1  
g28114.t1  
g28345.t1  
g28837.t1  
g28979.t1  
g29138.t1  
g29297.t1  
g29344.t1  
g29942.t1  
g30064.t1  
g30487.t1  
g30644.t1  
g31050.t1  
g31803.t1  
g32373.t1  
g32805.t1  
g33045.t1  
g33351.t1  
g33427.t1  
g33499.t1



g34614.t1  
g34971.t1  
g35400.t1  
g35402.t1  
g35471.t1  
g35542.t1  
g35575.t1  
g36405.t1  
g36549.t1  
g37209.t1  
g37526.t1  
g37998.t1  
g38181.t1  
g38206.t1  
g38313.t1  
g38450.t1  
g38670.t1  
g38790.t1  
g39255.t1  
g39448.t1  
g39560.t1  
g40739.t1

---

Gene names are listed respective to the isolate they are called as candidate avirulence genes for.