# An Investigation into the Signatures of Evolution in Pathogen Effector Genes

**Naveed Ishaque**

Thesis submitted for the degree of Doctor of Philosophy

University of East Anglia

The Sainsbury Laboratory

January 2012

© This copy of the thesis has been supplied on the condition that anyone who consults it is understood to recognize that its copyright rests with the author and that no quotation from the thesis, nor any information derived therefrom, may be published without the author's prior written consent.

# Abstract

The oomycete *Hyaloperonospora arabidopsidis* (*Hpa*) is a pathogen of *Arabidopsis thaliana* and a model for dissection of *A. thaliana* pathogen response networks. *Hpa* suppresses plant immunity by secreting effector proteins into the host thereby interfering with the host defence response and facilitating its own growth. The host's resistance genes are able to recognise certain alleles of some effectors and trigger an immune response. This interaction exerts selection pressure on effectors and resistance genes and has been likened to an evolutionary arms race. It has been shown that some effectors of *Hpa* that increase virulence activity and have certain alleles recognised by the host are under positive selection.

I investigated sequence variation in *Hpa* using the Illumina second generation sequencing platform. I was involved in the *Hpa* genome sequencing project. Using Illumina sequenced reads I isolated and removed contaminations, identified and integrated 4 Mb of novel sequence and developed new methods to evaluate genome completeness. I then trained and used various gene prediction algorithms to predict gene models for *Hpa*. Annotation and analysis of the gene models revealed interesting aspects about *Hpa* biology, including incomplete nitrogen and sulphur assimilation pathways, a reduced complement of effectors compared to other similar pathogens and a significant increase of sequence variation in candidate effectors (Baxter et al., 2010). For ease of visualisation I implemented a genome browser, which displays the gene models, sequence variation, and expression data of *Hpa*.

I developed a novel pipeline that performs phylogenetic and evolutionary analysis to identify genes under selection. Comparative genomics analysis using this pipeline revealed that some effectors and under higher selective pressure compared to other genes. Analysis of the most highly evolving genes reveals a novel class of effectors, providing a valuable resource for further elucidating mechanisms of effector biology.

# Table of Content

Acknowledgements ........ i

Publications arising from this thesis ........ ii

Chapter 1 – General Introduction ........ 1

1.1 Introduction and rationale ........ 1

1.2 Evolution, variation and selection ........ 2

1.2.1 Variation ........ 2

1.2.2 Mutation ........ 2

1.2.3 Modelling variation ........ 3

1.2.3.1 Genetic drift ........ 3

1.2.3.2 Hardy-Weinberg equilibrium ........ 3

1.2.3.3 Coalescent theory ........ 4

1.2.4 Evolution ........ 4

1.2.4.1 Natural selection ........ 4

1.2.4.2 Neutral theory of molecular evolution ........ 5

1.2.5 Evolutionary tests ........ 5

1.2.5.1 ω ratio ........ 5

1.2.5.2 Evolutionary modelling with variable ω ........ 5

1.2.5.3 Tests of neutrality ........ 6

1.2.5.3.1 Tajima's D ........ 7

1.2.5.3.1 Fu and Li's D, D*, F and F*. ........ 7

1.2.5.3.1 Fu's Fs ........ 7

1.3 Plant-pathogen co-evolution ........ 7

1.3.1 Plant immunity ........ 8

1.3.2 Effector translocation and structure ........ 9

1.3.3 *Hpa* effector evolution ........ 11

1.4 Genomics and Sequencing ........ 14

1.4.1 Second generation sequencing ........ 14

1.4.1.1 454 Pyrosequencing ........ 15

1.4.1.2 Sequencing by synthesis ........ 16

1.4.2 Second generation sequencing applications and tools ........ 17

1.4.2.1 Genome assembly ........ 17

1.4.2.2 Alignment and variant calling ........ 18

1.4.2.3 Transcriptomics ........................................................................................ 19
1.5 Objectives of the project........................................................................................ 20
Chapter 2 – Materials and Methods ........................................................................ 21
2.1 Biological material and sequencing ....................................................................... 21
2.1.1 Sanger sequencing.............................................................................................. 21
2.1.2 Illumina Sequencing ........................................................................................... 21
2.1.2.1 DNA extraction. ............................................................................................... 21
2.1.2.2 Illumina DNA library preparation and sequencing. ......................................... 21
2.1.2.3 Quality checking the Illumina preparation and sequencing.......................... 22
2.1.3 Illumina cDNA sequencing.................................................................................. 23
2.1.4 454 Sequencing .................................................................................................. 23
2.2 Software and protocols.......................................................................................... 24
2.2.1 Assembly ............................................................................................................ 24
2.2.1.1 Assembly of Sanger reads ............................................................................... 24
2.2.1.2 Short read assembly........................................................................................ 25
2.2.1.3 Hybrid assembly .............................................................................................. 25
2.2.2 Alignment ........................................................................................................... 25
2.2.2.1 DNA alignment ................................................................................................ 25
2.2.2.2 cDNA alignment............................................................................................... 28
2.2.3 Gene predictions ................................................................................................ 29
2.2.4 Gene annotation................................................................................................. 29
2.2.5 Evolutionary analysis.......................................................................................... 30
Chapter 3 – *Use of sequencing by synthesis to evaluate and improve the Hyaloperonospora arabidopsidis genome assembly* ............................................................................... 31
3.1 Introduction ........................................................................................................... 31
3.2 Results and discussion ........................................................................................... 32
3.2.1 Establishing a short read *de-novo* assembly for *Hpa* Emoy2 ............................. 32
3.2.1.1 *Hpa* version 7 assembly.................................................................................... 32
3.2.1.2 Identifying 'uncloned' regions of the *Hpa* genome.......................................... 33
3.2.1.3 De-novo short read assembly of *Hpa* Emoy2 using Velvet ............................. 34
3.2.1.4 Evaluating the quality of the Velvet de-novo assemblies ............................... 37
3.2.1.5 Comparing the Velvet assembly to the v7 assembly ...................................... 39
3.2.1.6 Using CEGMA to compare the core eukaryotic genes in the Velvet and v7 assemblies ............................................................................................................ 40

3.2.1.7 Comparing representation of genomic sequence between the Velvet and v7 assemblies ... 42
3.2.1.8 Comparing representation expressed sequences between the Velvet and v7 assemblies ... 43
3.2.2 Improving the *Hpa* v7 Sanger assembly using Illumina sequenced short reads . 45
3.2.2.1 Merging of full length BAC sequences with the Sanger shotgun assembly . 46
3.2.2.2 Iterative correction of the Sanger assembled sequence ... 46
3.2.2.3 Integration of the Sanger and Illumina assemblies ... 49
3.2.2.4 Identifying and removing contamination using Illumina sequencing ... 50
3.2.3 Evaluating the v8.3 hybrid Sanger and Illumina assembly ... 53
3.2.3.1 Estimating genome size using read coverage ... 53
3.2.3.2 Comparing the Velvet and v8.3 assemblies ... 56
3.2.3.3 Identifying the number of CEGs in the v8.3 assembly ... 57
3.2.3.4 Representation of genomic sequence in the v8.3 assembly ... 58
3.2.3.5 Representation of expressed sequence in the v8.3 assembly ... 59
3.2.4 Heterozygosity in *Hpa* Emoy2 ... 60
3.2.4.1 Identifying heterozygosity in *Hpa* Emoy2 ... 60
3.2.4.2 Heterozygosity in genes and effectors ... 61
3.3 Summary ... 62
Chapter 4 – *Use of Illumina sequencing to evaluate and improve the Hpa gene models* ... 63
4.1 Introduction ... 63
4.2 Results and Discussion ... 64
4.2.1 Existing gene models ... 64
4.2.1.1 Determining the number of genes ... 64
4.2.1.2 Genezilla ... 64
4.2.1.3 Snap ... 66
4.2.1.4 CEGMA - Core Eukaryotic Genes Mapping Approach ... 67
4.2.1.5 Integration of gene models ... 67
4.2.2 Evaluating gene models ... 68
4.2.2.1 Evidence for expression ... 68
4.2.2.2 Other quality issues in the v1 gene models ... 70
4.2.3 Generating version 2 of the *Hpa* gene models ... 70
4.2.3.1 Geneid ... 71
4.2.3.2 Augustus ... 72
4.2.3.3 Integrating Augustus and Geneid gene predictions ... 76

4.2.3.4 Evaluating the *Hpa* Emoy2 v2 gene models ............................................. 77
4.2.4 Generating version 3 of the *Hpa* gene models.................................................. 80
4.2.4.1 Identifying missing genes ........................................................................ 81
4.2.4.1.1 Identifying missing conserved eukaryotic genes................................ 81
4.2.4.1.2 Identifying missing oomycete genes ................................................ 81
4.2.4.1.3 Identifying missing effector genes ................................................... 82
4.2.4.1.4 Identifying missing genes with PASA assembled EST support ............. 83
4.2.4.2 Integrating the additional genes into the v2 gene models ......................... 83
4.2.4.2.1 Integrating missing CEG and PCEG genes............................................ 84
4.2.4.2.2 Integrating missing effector genes..................................................... 85
4.2.4.2.3 Integrating missing PASA assembled genes ........................................ 86
4.2.4.2.4 Removing redundancy and overlap.................................................... 87
4.2.4.3 Evaluating the *Hpa* Emoy2 v3 gene models ............................................... 88
4.2.4.3.1 Evidence of expression...................................................................... 88
4.2.4.3.2 Average length + GC .......................................................................... 89
4.2.5 Annotation....................................................................................................... 91
4.2.5.1 InterproScan functional annotation ........................................................... 91
4.2.5.2 GO term annotation .................................................................................. 92
4.2.5.3 Localisation................................................................................................ 94
4.2.5.4 Secreted and transmembrane proteins ...................................................... 94
4.2.5.5 Metabolic Pathway Analysis........................................................................ 95
4.2.5.6 Lack of nitrogen and sulphur assimilation pathways .................................. 97
4.2.5.7 Virulence related genes.............................................................................. 99
4.3 *Hpa* genome browser........................................................................................ 101
4.4 Summary ........................................................................................................... 103
Chapter 5 – Use of Illumina sequencing to investigate signatures of evolution in *Hpa* .... 105
5.1 Introduction ....................................................................................................... 105
5.2 Method development ........................................................................................ 106
5.2.1 Pipeline .......................................................................................................... 106
5.2.1.1 Input ......................................................................................................... 106
5.2.1.2 Output ....................................................................................................... 107
5.2.1.2.1 Stage 1 ............................................................................................... 107
5.2.1.2.2 Stage 2 ............................................................................................... 116
5.2.1.2.3 Stage 3 ............................................................................................... 117
5.2.1.2.4 Output format for data comparison.................................................. 120

5.3 Results and discussion ........ 121
5.3.1 Alignment ........ 121
5.3.2 Variant calling ........ 124
5.3.2.1 SNP and INDEL calls ........ 124
5.3.2.2 Heterozygosity ........ 124
5.3.2.3 Preferential SNP mutation ........ 125
5.3.2.4 Distribution of INDEL sizes ........ 127
5.3.3 Resequencing analysis of *Hpa* genes for 8 races of *Hpa* ........ 128
5.3.3.1 Coverage ........ 128
5.3.3.1.1 Percentage covered ........ 128
5.3.3.1.2 Mean coverage of each gene for each race ........ 130
5.3.3.1.3 Copy number variation ........ 134
5.3.3.1.3.1 Hemizygosity ........ 135
5.3.3.2 SNPs ........ 136
5.3.3.2.1 Protein coding effects of SNPs ........ 137
5.3.3.3 INDELs ........ 139
5.3.3.3.1 Protein coding effects of INDELs ........ 140
5.3.4 DnaSP Analysis ........ 143
5.3.4.1 Analysis of sample size ........ 143
5.3.4.2 S, Eta and the number of haplotypes ........ 146
5.3.4.3 Tajima's D ........ 148
5.3.4.4 Fu & Li's D and Fu & Li's F ........ 149
5.3.4.5 Fu's Fs ........ 150
5.3.5 PAML analysis ........ 151
5.3.5.1 Tree construction ........ 151
5.3.5.2 dN/dS ........ 153
5.3.5.3 Evolutionary models ........ 155
5.3.6 Comparison of effectors ........ 157
5.3.6.1 S, Eta and the number of haplotypes ........ 157
5.3.6.2 Tajima's D ........ 159
5.3.6.3 Fu & Li's statistics and Fu's Fs ........ 159
5.3.6.4 dN/dS – yn00 and codeml ........ 162
5.3.6.5 PAML evolutionary models ........ 164
5.3.6.6 Genes evolving like effectors ........ 166
5.4 Summary ........ 169

Chapter 6 – General discussion and outlook ........ 171
6.1 Modern day genome assembly ........ 171
6.2 The nature of obligate biotrophy ........ 173
6.3 High throughput analysis of evolutionary signatures ........ 174
6.4 Effector characterisation and evolution ........ 175
Appendices ........ 178
Appendices for Chapter 2 ........ 178
Appendix table 2.1: Cluster density of paired end reads ........ 178
Appendix table 2.2: Reads Summary ........ 180
Appendix figure 2.1 FastQC Read statistics ........ 183
Appendices for Chapter 3 ........ 184
Appendix table 3.1: Genome version history ........ 184
Appendix table 3.2: *Hpa* Emoy2 Illumina reads ........ 185
Appendix table 3.3: List of genes used to evaluate assemblies ........ 186
Appendix table 3.4: Final v8.3 assembly scaffold (AGP) ........ 187
Appendix table 3.5: Bacterial contamination on contigs (top 50 contigs) ........ 193
Appendix table 3.6: Bacterial contaminants (top 50 contigs) ........ 194
Appendix table 3.7: *Arabidopsis thaliana* contamination ........ 195
Appendices for Chapter 4 ........ 196
Appendix table 4.1: Go Terms ........ 196
Appendices for Chapter 5 ........ 199
Appendix table 5.1: Distribution of percentage coverage (Perc cov) of genes ........ 199
Appendix table 5.2: Multi copy genes ........ 201
Appendix table 5.3: SNP gene tables (top 20) ........ 202
Appendix table 5.4: Heterozygous SNP gene table (top 20) ........ 203
Appendix table 5.5: Protein coding effect of SNPs (top 20) ........ 204
Appendix table 5.6: INDEL gene tables (top 20) ........ 205
Appendix table 5.7: Heterozygous INDEL gene tables (top 20) ........ 206
Appendix table 5.8: DnaSP tables (top and bottom 20 Fu's Fs) ........ 207
Appendix table 5.9: PAML tables (top 20 codeml dN/dS) ........ 208
Abbreviations ........ 209
References ........ 211

# List of figures

Figure 1.1: Likelihood ratio test model comparisons modelled by codeml ..........................6

Figure 1.2: Zig-zag-zig plant defence response mechanism, showing the amplitude of defence response elicited by various stages of infection ....................................................8

Figure 1.3: Principles of effector biology in oomycete, fungus and bacteria ..................... 10

Figure 1.4: Domain structure of oomycete apoplastic and cytoplasmic effectors ............ 11

Figure 1.5: Alignment of predicted ATR1 proteins from 8 *Hpa* races ............................... 12

Figure 1.6: Sliding window analysis of synonymous and non-synonymous substitutions across ATR1 in 8 *Hpa* races ........................................................................................ 13

Figure 1.7: Roche 454 GS FLX sequencing method .......................................................... 15

Figure 1.8: Illumina sequencing method ........................................................................... 16

Figure 1.9: Illustration of how a read is reassembled using consensus overlap and the De-Bruijn principles ............................................................................................................ 18

Figure 2.1: Variation true positive recall rate of various mapping techniques at Q10 ....... 27

Figure 2.2: Variation false positive rate of various mapping techniques at Q10 ................ 27

Figure 2.3: Sensitivity of various mapping techniques at Q10 .......................................... 28

Figure 3.1: Histogram plot of k-mer coverage (k=21) ...................................................... 36

Figure 3.2: Histogram plot of k-mer coverage (k=23) ...................................................... 36

Figure 3.3: Histogram plot of k-mer coverage (k=25) ...................................................... 36

Figure 3.4: Extract from BLAST alignment of ATR13 to Velvet assembly of *Hpa* Emoy2 .... 38

Figure 3.5: The effect of sequence coverage on number of CEGs identifiable by the CEGMA pipeline ........................................................................................................................ 41

Figure 3.6: Identification of single copy core eukaryotic orthologous genes (CEGs) by the CEGMA pipeline ............................................................................................................ 42

Figure 3.7: Four stage assembly improvement pipeline for incorporating BAC and Illumina sequencing data ............................................................................................................ 45

Figure 3.8: Genome size estimation from Illumina and Sanger read coverage ................. 55

Figure 3.9: Visual representation of an alignment of the *Hpa* Emoy2 v8.3 assembly against the Velvet assembly ....................................................................................................... 57

Figure 3.10: Percentage of full and partial CEGs identified by the CEGMA pipeline in the v8.3 assembly ................................................................................................................ 58

Figure 3.11: Rate of heterozygosity in the *Hpa* v8.3 genome, genes and RxLR effector candidates ............ 61
Figure 4.1: Genezillas state transition model ............ 65
Figure 4.2: The Snap transition state model ............ 66
Figure 4.3: Box plots of the 90th percentiles of gene lengths and GC percentage ............ 75
Figure 4.4: Box plots of the 90th percentiles of gene lengths and GC percentage for v1, v2 and v3 gene models ............ 90
Figure 4.5: The distribution of GO terms identified in the *Hpa* v3 gene models ............ 93
Figure 4.6: Venn diagram of genes with overlapping annotation as secreted, transmembrane and effector or effector homolog genes ............ 95
Figure 4.7: Metabolic pathways in *Hpa*, P. infestans, P. sojae and P. ramorum ............ 96
Figure 4.8: *Hpa* Emoy2 v8.3 genome browser ............ 102
Figure 5.1: QQ-plots of likelihood of belonging to Poisson/Normal distribution for 5x coverage ............ 110
Figure 5.2: QQ-plots of likelihood of belonging to Poisson/Normal distribution for 9x coverage ............ 111
Figure 5.3: QQ-plots of likelihood of belonging to Poisson/Normal distribution for 20x coverage ............ 112
Figure 5.4: QQ-plots of likelihood of belonging to Poisson/Normal distribution for 40x coverage ............ 113
Figure 5.5: Distribution of INDEL sizes ............ 127
Figure 5.6: Frequency distribution of difference between maximum and minimum relative gene coverage between 8 races of *Hpa* ............ 130
Figure 5.7: Frequency distribution of the Poisson CDF for coverage ............ 134
Figure 5.8: Frequency distribution of the INDEL length over coding regions of genes ............ 141
Figure 5.9: Percentage distribution of the INDEL length over coding regions of genes ............ 142
Figure 5.10: Effect of increasing the sample size on the percentage of genes for which S, Eta and the number of haplotypes can be analysed ............ 144
Figure 5.11: Frequency distribution of predicted number of haplotypes per gene with increasing sample population ............ 144
Figure 5.12: Frequency distribution of predicted number of segregating sites (S) per gene with increasing sample population ............ 145
Figure 5.13 Frequency distribution of predicted number of mutations (Eta) per gene with increasing sample population ............ 145

Figure 5.14: Frequency distribution of the number of segregating sites (S), the number of mutations (Eta) and the number of haplotypes ............................................................ 146
Figure 5.15: Frequency distribution of Tajima's D, Fu & Lis D* and F* ............................ 149
Figure 5.16: Frequency distribution of Fu's Fs ............................................................... 150
Figure 5.17: Phylogenetic tree of *Hpa* races based on sequence homologous to P. infestans mitochondria ......................................................................................................... 15
Figure 5.18: Phylogenetic tree of *Hpa* races based on 11570 segregating sites on the largest *Hpa* contig ........................................................................................... 152
Figure 5.19: Frequency distribution of dN/dS values calculated by yn00 and codeml ..... 153
Figure 5.20: Percentage distribution of dN/dS values calculated by yn00 and codeml .... 154
Figure 5.21: Frequency distribution of the difference in dN/dS calculation between the yn00 and codeml method ......................................................................................... 155
Figure 5.22: Frequency distribution of evolutionary model testing using PAML ............. 156
Figure 5.23: Percentage distribution of S, the number of segregating sites, for the 472 sampled effectors, transmembrane genes and KOGs .................................................... 157
Figure 5.24: Percentage distribution of Eta, the total number of mutations, for the 472 sampled effectors, transmembrane genes and KOGs .................................................... 158
Figure 5.25: Percentage distribution of the number of haplotypes per gene for the 472 sampled effectors, transmembrane genes and KOGs .................................................... 158
Figure 5.26: Percentage distribution of Tajima's D per gene for the 472 sampled effectors, transmembrane genes and KOGs ..................................................................... 159
Figure 5.27: Percentage distribution of Fu and Li's D per gene for the 472 sampled effectors, transmembrane genes and KOGs ................................................................. 160
Figure 5.28: Percentage distribution of Fu and Li's F per gene for the 472 sampled effectors, transmembrane genes and KOGs ..................................................................... 160
Figure 5.29: Percentage distribution of Fu's F per gene for the 472 sampled effectors, transmembrane genes and KOGs ..................................................................... 161
Figure 5.30: Percentage distribution of dN/dS calculated by codeml for the 472 sampled effectors, transmembrane genes and KOGs ................................................................. 162
Figure 5.31: Percentage distribution of dN/dS calculated by yn00 for the 472 sampled effectors, transmembrane genes and KOGs ................................................................. 163
Figure 5.32: Percentage distribution of 2 * ln (M2a -M1a) model comparison likelihoods for the 472 sampled effectors, transmembrane genes and KOGs ....................................... 164
Figure 5.33: Percentage distribution of 2 * ln (M3 -M0) model comparison likelihoods for the 472 sampled effectors, transmembrane genes and KOGs ....................................... 165

Figure 5.34: Percentage distribution of 2 * ln (M8 -M7) model comparison likelihoods for the 472 sampled effectors, transmembrane genes and KOGs ........................................ 165
Figure 5.35: Sequence alignment of three highly evolving secreted *Hpa* genes with a P. infestans effector ........................................................................................................ 169

# List of tables

Table 1.1: Polymorphisms in ATR13 and Ppat5 ........ 13

Table 3.1: velvetg run statistics (no scaffolding) ........ 34

Table 3.2: velvetg run statistics (with scaffolding) ........ 37

Table 3.3: DNAdiff results between the *Hpa* Emoy2 v7 Sanger assembly and the *Hpa* Emoy2 Velvet assembly ........ 40

Table 3.4: Number and percentage of reads from a single lane aligning to the v7 and Velvet assemblies ........ 43

Table 3.5: Number of ESTs aligning to the *Hpa* Emoy2 v7 and Velvet assemblies ........ 43

Table 3.6: Number of cDNA reads aligning to the *Hpa* Emoy2 v7 and Velvet assemblies .. 44

Table 3.7: Number of SNPs and INDELs after each iteration of correction ........ 47

Table 3.8: Number and percentage of reads from a single lane aligning to the corrected assembly after X rounds of genome correction ........ 48

Table 3.9: Number and percentage of reads from a single lane aligning to the corrected assembly after genome modifications ........ 50

Table 3.10: Best BLAST hits when extended regions of very low coverage were blasted against the NR database ........ 51

Table 3.11: DNAdiff results between the *Hpa* Emoy2 v8.3 hybrid assembly and the *Hpa* Emoy2 Velvet assembly ........ 56

Table 3.12: Number and percentage of reads from a single lane aligning to the v7, v8.3 and Velvet assemblies ........ 59

Table 3.13: Number of ESTs aligning to the *Hpa* Emoy2 v7 and Velvet assemblies ........ 59

Table 3.14: Number of cDNA reads aligning to the *Hpa* Emoy2 v7 and Velvet assemblies 60

Table 4.1: The number and percentage aligning to the *Hpa* v8.3 genome assemble and v1 gene models using BLAT ........ 68

Table 4.2: The number and percentage cDNA reads aligning to A. thaliana and *Hpa* gene models using MAQ ........ 69

Table 4.3: Number and percentage of ESTs aligning to gene models using BLAT ........ 74

Table 4.4: Number and percentage of filtered Illumina cDNA reads aligning to gene models using MAQ ........ 74

Table 4.5: Number and percentage ESTs aligning to the genome, v1 and v2 gene models using BLAT ........ 78

Table 4.6: Number and percentage cDNA aligning to the genome, v1 and v2 gene models using BLAT ........ 78
Table 4.7: List of genes missing from the *Hpa* v2 gene models ........ 80
Tables 4.8: Number of (A) genes added over each integration iteration to integrate the missing CEG, PCEG and PCEG homolog genes; (B) genes added from each set of gene predictions ........ 84
Tables 4.9: Number of (A) genes added over each integration iteration to integrate the missing Effector genes; (B) genes added from each set of gene predictions ........ 85
Tables 4.10: Number of genes (A) added over each integration iteration to integrate the missing CEG, PCEG and PCEG homolog genes; (B) genes added from each set of gene predictions ........ 86
Tables 4.11: Number of residual genes to add to the v2 assembly after removing redundant calls in each dataset ........ 87
Tables 4.12: Summary of (A) added genes; (B) genes removed due from v2 gene models; (C) the number of genes remaining to make the *Hpa* Emoy2 v8.3 genome v3 gene models .. 88
Table 4.13: The number and percentage of ESTs aligning to the genome, v1, v2 and the v3 gene models ........ 88
Table 4.14: The number and percentage of cDNA reads aligning to the genome, v1, v2, and v3 gene models ........ 89
Tables 4.15: Length and GC values for the v1, v2 and v3 gene models ........ 90
Table 4.16: Number and percentage of genes annotated using various programs ........ 92
Table 4.17: Breakdown of localisation predictions using WolfPsort on a fungal model ..... 94
Table 4.18: Gene IDs for nitrogen and sulphur assimilation enzymes in Phytophthora and *Hpa*. P. infestans, P. sojae and P. ramorum ........ 98
Table 4.19: Copy numbers of annotated *Hpa* genes implicated in pathogenesis. P. ramorum and P. sojae ........ 99
Table 5.1: Number and percentage of reads from 8 *Hpa* races aligning to the *Hpa* v8.3 assembly ........ 123
Table 5.2: Table of predicted SNPs and INDELs in the 8 sequenced races of *Hpa* ........ 125
Tables 5.3: Table of mutational spectrum of *Hpa* ........ 126
Table 5.4: Frequency distribution of percentage of nucleotides of genes covered ........ 128
Table 5.5: Table of genes displaying most variation in the percentage coverage ........ 129
Table 5.6: Genes with the most variation in the relative mean coverage ........ 131
Table 5.7: Genes with the most variation in the relative mean coverage, where the minimum mean coverage between the 8 races of *Hpa* is 10 ........ 132

Table 5.8: Genes with the most variation in the relative mean coverage, where they are single copy in the Emoy2 ........................................................................................ 133
Table 5.9: Number of coding and non-coding SNPs in *Hpa* races .................................... 136
Table 5.10: Number of coding and non-coding heterozygous (Het) SNPs in *Hpa* races ... 136
Table 5.11: Number of synonymous and non-synonymous SNPs in *Hpa* races ................ 138
Table 5.12: Number of coding and non-coding INDELs in *Hpa* races showing both homozygous and heterozygous INDELs .......................................................................... 139
Table 5.13: Number of coding and non-coding heterozygous INDELs in *Hpa* races ......... 139
Table 5.14: The 15 genes with the highest number of segregating sites (S) ..................... 147
Table 5.15: The 15 genes with the highest number of total mutations (Eta) ................... 148

# Acknowledgements

I would like to offer my humble gratitude, respect and huge thanks to my supervisor, Jonathan Jones, who gave me the opportunity to undertake my PhD under his supervision and guidance. I am grateful for the help and discussion from Eric Kemen, Dan MacLean and David Studholme who were always willing to help and point me in the right direction. I would also like to thank my PhD advisors, Eric Kemen, Dan MacLean, David Studholme and Richard Morris, for the support and advice. I am also grateful to BBSRC, Cogenics, and to Jonathan for funding my work.

I would like to thank the PPPP, Georgina Fabro, David Greenshields, Sophie Piquerez, Lennart Wirthmuller, Marie-Cecile Callaiud and Shuta Asia, who provided direction and useful discussion. I would also like to thank Brian Staskawicz for contributing Emwa1 sequence data. I am also very grateful for John McDowell, Brett Tyler, Jim Beynon, Laura Baxter and Sucheta Tripathy for their interaction and support whilst writing the *Hpa* genome paper.

I grateful to Anastasia Gardiner, Shyam Rallapalli, Michael Burrell and Neil Hastings for always being quick to respond to frequent pleas for help and support. I would also like to thank the rest of the JJ group for the friendship, support and useful discussion.

I would also like to thank my family, for their support despite the trials and tribulations over the last 4 years.

Finally, I would like to express my eternal gratitude to my best friend and wife, Erika Kuchen, without whom I would not have been able to complete my PhD. I would like to thank her for her optimistic outlook, encouragement and support throughout my PhD. To all the people I have mentioned and everyone that I forgot to mention, I thank you all once again.

# Publications arising from this thesis

Fabro G*, Steinbrenner J, Coates M, **Ishaque N**, Baxter L, Studholme DJ, Körner E, Allen RL, Piquerez SJ, Rougon-Cardoso A, Greenshields D, Lei R, Badel JL, Caillaud MC, Sohn KH, Van den Ackerveken G, Parker JE, Beynon J, Jones JD*. Multiple Candidate Effectors from the Oomycete Pathogen *Hyaloperonospora arabidopsidis* Suppress Host Plant Immunity. *PLoS Pathologens*, (11):e1002348, 2011

Caillaud MC*, Piquerez SJ, Fabro G, Steinbrenner J, **Ishaque N**, Beynon J, Jones JD*. Subcellular localization of the *Hpa* RxLR effector repertoire identifies a tonoplast-associated protein HaRxL17 that confers enhanced plant susceptibility. *Plant Journal* , doi: 10.1111/j.1365-313X.2011.04787.x, 2011

Baxter L*, Tripathy S*, **Ishaque N***, Boot N, Cabral A, Kemen E, Thines M, Ah-Fong A, Anderson R, Badejoko W, Bittner-Eddy P, Boore JL, Chibucos MC, Coates M, Dehal P, Delehaunty K, Dong S, Downton P, Dumas B, Fabro G, Fronick C, Fuerstenberg SI, Fulton L, Gaulin E, Govers F, Hughes L, Humphray S, Jiang RH, Judelson H, Kamoun S, Kyung K, Meijer H, Minx P, Morris P, Nelson J, Phuntumart V, Qutob D, Rehmany A, Rougon-Cardoso A, Ryden P, Torto-Alalibo T, Studholme D, Wang Y, Win J, Wood J, Clifton SW, Rogers J, Van den Ackerveken G, Jones JD*, McDowell JM*, Beynon J*, Tyler BM*. Signatures of adaptation to obligate biotrophy in the *Hyaloperonospora arabidopsidis* genome. *Science*, 330(6010):1549-51, 2010

**Ishaque N***, Piquerez SJ, Kemen E, Burrell M, Jones JD*. *Hyaloperonospora arabidopsidis* genome browser.
http://gbrowse2.tsl.ac.uk/cgi-bin/gb2/gbrowse/hpa_emoy2_publication/

# Chapter 1 – General Introduction

## 1.1 Introduction and rationale

Since the beginning of life on earth, approximately 3.4 billion years ago (Wacey et al., 2011) the key to survival has been the pursuit of nutrients. Successful organisms are able to gather nutrients from their surroundings, but not all organisms are able to do this independently and therefore need to interact with other organisms to acquire nutrients. This interaction "between two unlike organisms" is called symbiosis (de Bary, 1879). The symbiotic relationship between 2 organisms can be mutually beneficially to both organisms (mutualistic), beneficial to one of the organisms but neutral to the other (commensalistic) or beneficial to one at the expense of the other (parasitic) (Douglas, 2010). Depending on disease severity, a parasitic organism, which causes disease in its host, could also be considered a pathogen. These pathogens may require a living host in order to complete its lifecycle (biotrophy) or kill and feed on the host (necrotrophy). It is also possible for a pathogen to have a biotrophic phase followed by a necrotrophic phase (hemi-biotrophy). In order for pathogens to be successful, they must be able to colonise, survive and multiply on their hosts. Some pathogens are able to suppress host immunity by secreting so-called effector proteins into the host thereby interfering with the host defence response and facilitating their own growth. In order for a host to prevent disease, it must be able to stop the pathogen through either preformed barriers or by inducing a response to make the pathogen unable to colonise the host. In the case a pathogen is unable to colonise its target host, it must find a new host or overcome the defence barriers of its host to survive. When a pathogen is unable to parasitize multiple hosts, it becomes paramount that it specialises to overcome the host defence barriers. This continued co-evolutionary interaction between the host and pathogen would impose as a strong evolutionary selection on both the host and the pathogen.

In this dissertation I examined the signatures of evolution on *Hyaloperonospora arabidopsidis (Hpa)* using second generation sequencing and comparative genomics. *Hpa* is an obligate biotrophic pathogen that causes downy mildew of its host, the dicot plant *Arabidopsis thaliana.* Downy mildew pathogens are estimated to parasitize about 15% of all flowering plant families, and account for about 20% of the global fungicide market (The

Genome Institute, 2011). The downy mildews of sugar cane and maize are listed among seven plant pathogens considered to be major bioterror threats by the USA (Animal and Plant Health Inspection Service, 2002). Oomycete pathogens have also caused much devastation historically, such as the Great Irish Potato famine, caused by *Phytophthora infestans*, which caused 1 million deaths over 7 years. While *Hpa* is not a pathogen of a commercial crop plant, oomycete effector biology is an emerging field and any discoveries in the fundamental understanding of effector biology provide a platform to further study conserved mechanisms between systems, with the possibility of knowledge transfer to applied systems.

## 1.2 Evolution, variation and selection

### 1.2.1 Variation

Evolution is the cumulative change across successive generations in the characteristics of populations of biological organisms. Evolution leaves its signature on the genome as variation. Observable traits (phenotypes) of organisms result from their genetic constitutions (genotypes), as well as the environment. Natural selection acts on favourable phenotypes that are caused by variation at the DNA level, including genes encoding for proteins, non-coding RNA, non-coding DNA affecting expression levels and splice variation as well as epigenetic variation. The different types of variation in DNA sequence of individuals can be caused by mutations including single nucleotide polymorphisms (SNPs), insertions and deletions (INDELs), recombination, copy number variation (CNV), repeat number variation and change in ploidy.

### 1.2.2 Mutation

Mutations can be spontaneous or induced by a mutagen. Mutations can be classified as small scale mutations or large scale mutations. Small scale mutations are those affecting only a few nucleotides and include point mutations, insertions and deletions. Point mutations cause a change from one nucleotide to another and are often referred to as single nucleotide polymorphisms (SNPs) (Freese, 1959). SNPs can be classified as transitions (A ⇔ G, C ⇔ T) or transversions (T ⇔ G, C ⇔ G, T ⇔ A, C ⇔ A) (Freese, 1959). A SNP that occurs on a protein coding region of the genome can have 2 broad effects on the codon on which the mutation occurs: the new allele can code for the same amino acid (silent mutation/synonymous polymorphism) or it can code for another (non-synonymous polymorphism), which can either be a different amino acid (missense mutation) or a premature stop codon (nonsense mutation). Over half of all known human disease mutations are a result of non-synonymous mutations (Stenson et al., 2009). Insertions add one or more nucleotides to the DNA and deletions are the removal of one or more nucleotides from the DNA. Insertions and deletions are collectively referred to as INDELs. If an INDEL on a coding region of the genome is of a size not divisible by 3 (i.e. not a codon INDEL) they cause frameshifts, which can significantly modify the codons on the gene. Large scale mutations largely act at the level of the chromosomal structure. Examples of large scale mutations include gene duplications, deletions of large chromosomal sections, loss of heterozygosity, chromosomal inversions, interstitial deletions and chromosomal translocations.

### 1.2.3 Modelling variation

Variation in genes and genomic loci results in multiple forms of the gene (alleles) in the population. The number of distinct allele possibilities in a genomic locus (over a gene, genes or even a chromosome) that are transmitted together are referred to as the number of haplotypes. Haplotype variability encompasses the allelic variability and genetic recombination over a genomic locus in the population.

#### *1.2.3.1 Genetic drift*

The change in frequency of an allele in a population due to random sampling, rather than selective processes, is referred to as allelic drift (Joanna, 2011). Genetic drift is considered to be an evolutionary mechanism in small populations as it can be used to explain the loss of genetic variation in small populations due to sampling error (Zimmer, 2002).

#### *1.2.3.2 Hardy-Weinberg equilibrium*

Variability in genomes can be maintained through inheritance of haplotypes from parent to their offspring. Hardy and Weinberg (Hardy, 1908; Weinberg, 1908) described a principle that states that both allele and genotype frequencies in a sufficiently large population remain at equilibrium between generations unless disturbed by the influence of non-random mating, mutations, selection, limited population size, non-discrete generations, random genetic drift, gene flow and meiotic drive. This principle is commonly referred to as Hardy-Weinberg equilibrium, and describes an ideal state from which the extent of departure be measured. Another source of variation in diploid sexual populations includes the inheritance of one set of chromosomes from each parent. The interaction between the inherited parental haplotypes and how they characterise the resultant phenotypes is referred to a Mendelian inheritance (Mendel, 1865), which underlies much of the work carried out in the field of genetics.

#### *1.2.3.3 Coalescent theory*

Since all allelic states in the population are determined by the previous genealogical and mutational history of these genes, it is possible to identify ancestral forms of currently observed alleles. Coalescent theory models genetic drift backwards in time to attempt to identify a most recent common ancestor (MRCA) that provides the foundation for current allelic variation (Hudson, 1983; Kingman, 1982; Tajima, 1983). The coalescent provides insights into the probability of sample allelic configurations under stationary distribution of various population genetic models, and allows for maximum likelihood analysis of polymorphism data (Nordborg, 2007).

### 1.2.4 Evolution

#### *1.2.4.1 Natural selection*

A key mechanism of evolution is natural selection, which is the process whereby favourable genetic traits increase in abundance (segregate) within the population as a function of differential reproduction of the bearers of the traits (Darwin, 1859). Selection can be subcategorized as directional, stabilising, disruptive (diversifying), sexual and ecological. Directional selection is where a single phenotypic trait is favoured causing a shift in allele frequency towards this phenotypic trait. Stabilising selection is where a particular phenotypic trait is selected for in a population leading to decrease in genetic diversity. Diversifying selection is where 2 extreme phenotypic traits are preferred over an intermediate. Sexual selection is the process whereby phenotypic traits are favoured because, rather than improving survival fitness of the individual, they act to maximise reproductive success through mating characteristics. Finally, ecological selection is the favouring of phenotypic traits influenced only by the ecological processes without referencing mating characteristics (i.e. natural selection minus sexual selection).

### *1.2.4.2 Neutral theory of molecular evolution*

The neutral theory of molecular evolution states that the majority of evolutionary changes at the molecular level are caused by random drift of selectively neutral mutants (Kimura, 1983). This theory has recently been shown to be compatible with the theory of natural selection where adaptive change is modelled as a minority of DNA sequences changes (Fay, 2011).

## 1.2.5 Evolutionary tests

### *1.2.5.1 ω ratio*

There are a number of statistical tests for different types of evolution. One such test is the dN/dS (also referred to as Ka/Ks or ω) ratio, which is the ratio of non-synonymous substitutions per non-synonymous site (dN) to the number of synonymous substitutions per synonymous site (dS). With different ω values for genes, inferences about the evolutionary mechanisms acting on the gene can be made. If $\omega < 1$ it is assumed that purifying selection is acting on the amino acid changes to filter out deleterious mutations. If $\omega = 1$ it implies that the amino acid change is neutral. If $\omega > 1$ it implies that the amino acid change offers a selective advantage, providing convincing evidence for diversifying selection (Yang and Bielawski, 2000). Two implementations of this test have been described by (Goldman and Yang, 1994) and (Yang and Nielsen, 2000) amongst others (Ina, 1995; Nei and Gojobori, 1986). The Goldman and Yang's (1994) method implements a maximum likelihood codon substitution model, while the Yang and Nielsen (2000) method implements counting methods for estimating dN/dS. These methods have been implemented in the PAML suite (Yang, 2007) as yn00 (Yang and Nielsen, 2000) and codeml (Goldman and Yang, 1994).

### *1.2.5.2 Evolutionary modelling with variable ω*

Codeml also implements various site model tests, which treat the ω value for any codon in the gene as a variable, allowing for ω to vary over the gene (Nielsen and Yang, 1998; Yang and Bielawski, 2000). Studies have shown comparative likelihood ratio tests, comparing a model that does not allow for codons with $\omega > 1$ with another model that does, to be effective (Anisimova et al., 2001, 2002; Anisimova et al., 2003; Wong et al., 2004). The different models implemented in codeml are:

- M0 – one ratio (uniform selective pressure among sites)
- M3 – discrete (variable selective pressure among sites)
- M1a – nearly neutral (variable selective pressure, but no positive selection)
- M2a – positive selection (variable selective pressure, with positive selection)
- M7 – beta (beta distributed selective pressure)
- M8 – beta with ω (dN/dS or Ka/Ks) > 1 (beta plus positive selection)
- M8a – a special case of M8 testing for neutral selection

The model comparisons that can be used to infer positive selection include Model 0 (M0) - Model 3 (M3), Model 1a (M1a) – Model 2a (M2a) and Model 7 (M7) – Model 8 (M8) (fig 1.1).

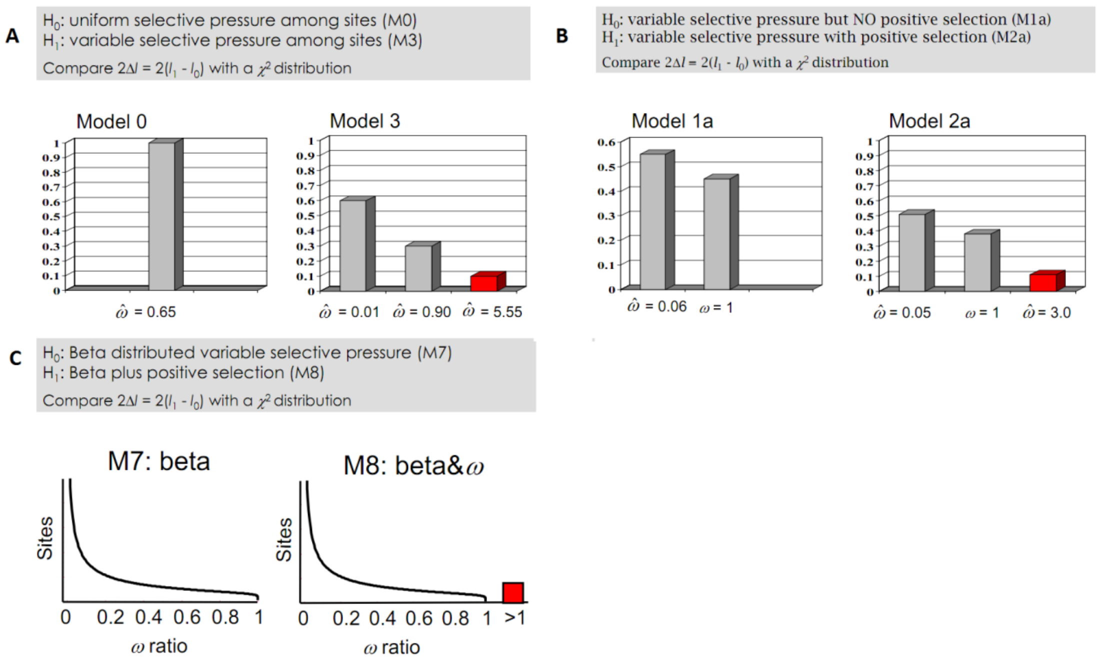



**Figure 1.1: Likelihood ratio test model comparisons modelled by codeml [reproduced from PAML: A program package by Ziheng Yang [http://abacus.gene.ucl.ac.uk/ziheng/data/pamlDEMO.pdf](http://abacus.gene.ucl.ac.uk/ziheng/data/pamlDEMO.pdf)]. Model comparisons include M0 – M3 (A), M1a – M2a (B) and M7 – M8a (C). An additional model comparison is used to differentiate model M8 from neutral drift (M8a – M8).**

### *1.2.5.3 Tests of neutrality*

Other commonly used statistical tests of neutrality are Tajima's D (Tajima, 1989), Fu & Li's D, D*, F & F* (1993) and Fu's Fs (Fu, 1997), which are all implemented in the program DnaSP (Librado and Rozas, 2009).

#### 1.2.5.3.1 Tajima's D

Tajima's D distinguishes between DNA sequence evolving neutrally and DNA sequence evolving under a non-neutral model. Tajima's test is based on the fact that estimates of the number of polymorphic (segregating) sites and of the average number of nucleotide differences are correlated under the neutral model of evolution. If the value of D is too large or too small, the neutral 'null' hypothesis is rejected. A negative value of Tajima's D indicates an excess of low frequency polymorphisms and may also signify purifying selection. A positive value indicates low levels of low and high frequency polymorphisms and may also signify balancing selection and heterozygote advantage. In general, values of Tajima's D above +2 and below -2 are likely to be significant, indicating selection (Tajima, 1989).

#### 1.2.5.3.1 Fu and Li's D, D*, F and F*.

Tajima did not base his test on coalescent. Fu and Li's tests are directly based on coalescent. The test statistics D and F require data from intraspecific polymorphism and sequence from a related outgroup species. The D* and F* do not require an outgroup species in the input. With 10 samples, values of D and D* less than -1.8 and greater than 1.4 are significant and values for F and F* less than -2 and greater than 1.55 are significant (Fu and Li, 1993).

#### 1.2.5.3.1 Fu's Fs

Fu's Fs test statistic is considerably more powerful than the previous tests, at rejecting the null hypothesis of neutrality of mutations in DNA samples under logistic population growth and genetic hitchhiking (where an allele experiences an increase in population frequency due to linkages with a gene positively selected for) (Fu, 1997).

## 1.3 Plant-pathogen co-evolution

*Hpa* has evolved only to be able to infect *A. thaliana* as a host (Goker et al., 2004). For this reason it has to overcome or evade triggering *A. thaliana* defence mechanisms and is under strong selection pressure to enhance its virulence mechanisms, while selecting against mechanisms that lead to recognition of the pathogen by the host. Likewise, the host, *A. thaliana,* is under similar selection pressures to enhance its resistance to *Hpa* infection.

### 1.3.1 Plant immunity

The plant's initial defence barrier is a general one, such as secretion of toxins and physical barriers to pathogen entry, which may be effective against a number of pathogens. Once a pathogen is able to overcome these barriers, more complex interactions between the host and pathogen become relevant. Jones and Dangl (2006) described a 'zig-zag-zig' model of defence response (fig 1.2). According to this model, the initial defence response after initial contact with the pathogen involves the recognition of pathogen associated molecular patterns (PAMPs). PAMPs are conserved molecular patterns in pathogens that are not easily lost by the pathogen such as bacterial flagellin, which contains a 22 amino acid conserved domain (flg22) (Chinchilla et al., 2006) and fungal chitin (Wan et al., 2008). PAMPs are bound by the host's receptor kinases. For instance, flg22 is recognised by the receptor kinase FLS2 of *Arabidopsis* (Chinchilla et al., 2006). Upon recognition, FLS2 is internalised and triggers a set of immune responses referred to as PAMP triggered immunity (PTI, which is also sometimes referred to as pattern triggered immunity).

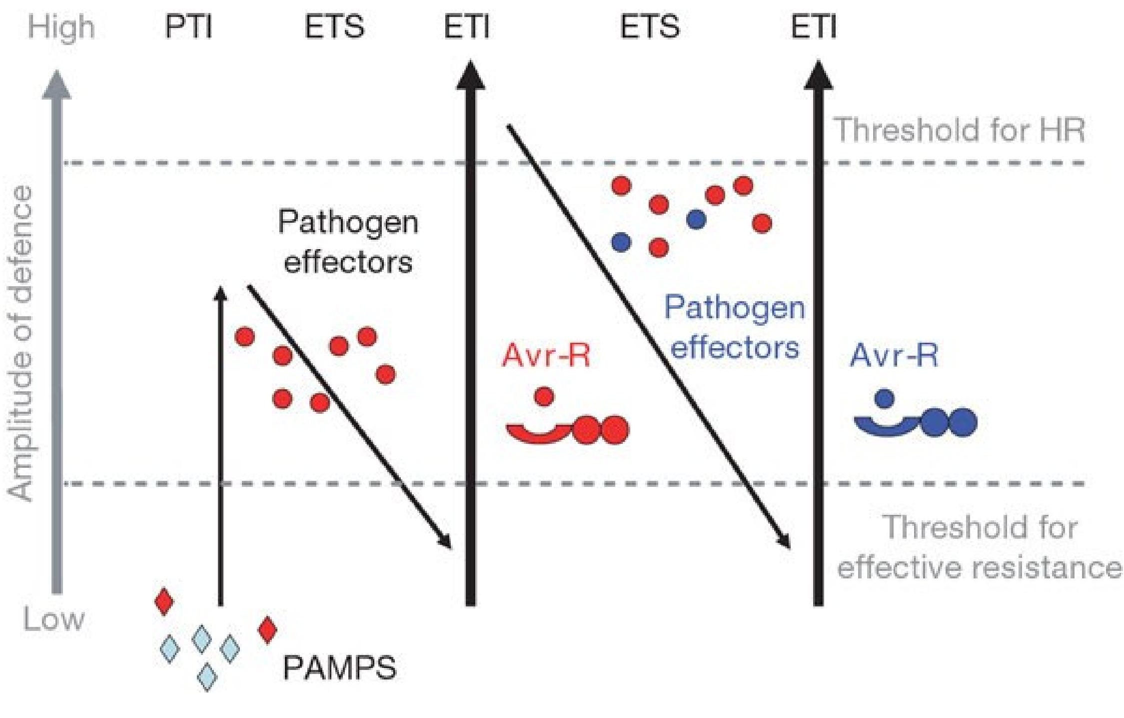



**Figure 1.2: Zig-zag-zig plant defence response mechanism, showing the amplitude of defence response elicited by various stages of infection [reproduced from (Jones and Dangl, 2006)]. ETS = effector-triggered susceptibility. Avr-R = interaction of the recognised effector (avirulence or avr gene) and the host resistance gene (the R-gene)**

In order to overcome this defence response the pathogen may have evolved the ability to secrete effectors, which are usually small secreted proteins that facilitate growth of the pathogen on the host by suppressing defence or manipulating host metabolic function for the benefit of the pathogen. A recent study has shown that pathogen effectors have evolved independently to target core genes in the plant immune network (Mukhtar et al., 2011).

In an evolutionary context, one would expect the fixation of a pathogen effector that is particularly effective in a small population. This is a hypothesis as to why plants have evolved the mechanisms to directly or indirectly recognise pathogen effectors through resistance genes, causing effector triggered immunity (ETI) (Scofield et al., 1996; Tang et al., 1996; Van der Biezen and Jones, 1998). To add to this intricate system of effector–R gene interaction, there have been reports that effectors also suppress ETI (Tsiamis et al., 2000) and there are plant resistance genes that recognise those effectors (Yucel et al., 1994).

### 1.3.2 Effector translocation and structure

Both bacterial and eukaryotic pathogens have been shown to secrete effectors. Effectors have either cytoplasmic (e.g. (Allen et al., 2004; Orbach et al., 2000; Rehmany et al., 2005; Win et al., 2006)) or apoplastic (e.g. (Rooney et al., 2005; Tian et al., 2007)) localisation in plants. They are translocated from the pathogen to the host via a secretory mechanism. In bacteria this is achieved by the type 3 secretion system for effectors to be delivered to the host cytoplasm and the type 2 secretion system for effectors to be delivered to the host apoplast. In oomycete pathogens the exact mechanism of translocation is not fully understood. Briefly, the effectors are translocated from the pathogen into the host apoplast. Here, apoplastic effectors interfere with apoplastic defence responses. Another translocation event occurs whereby cytoplasmic effectors are translocated into the plant cytosol where they interfere with plant cytoplasmic defence responses (fig 1.3).

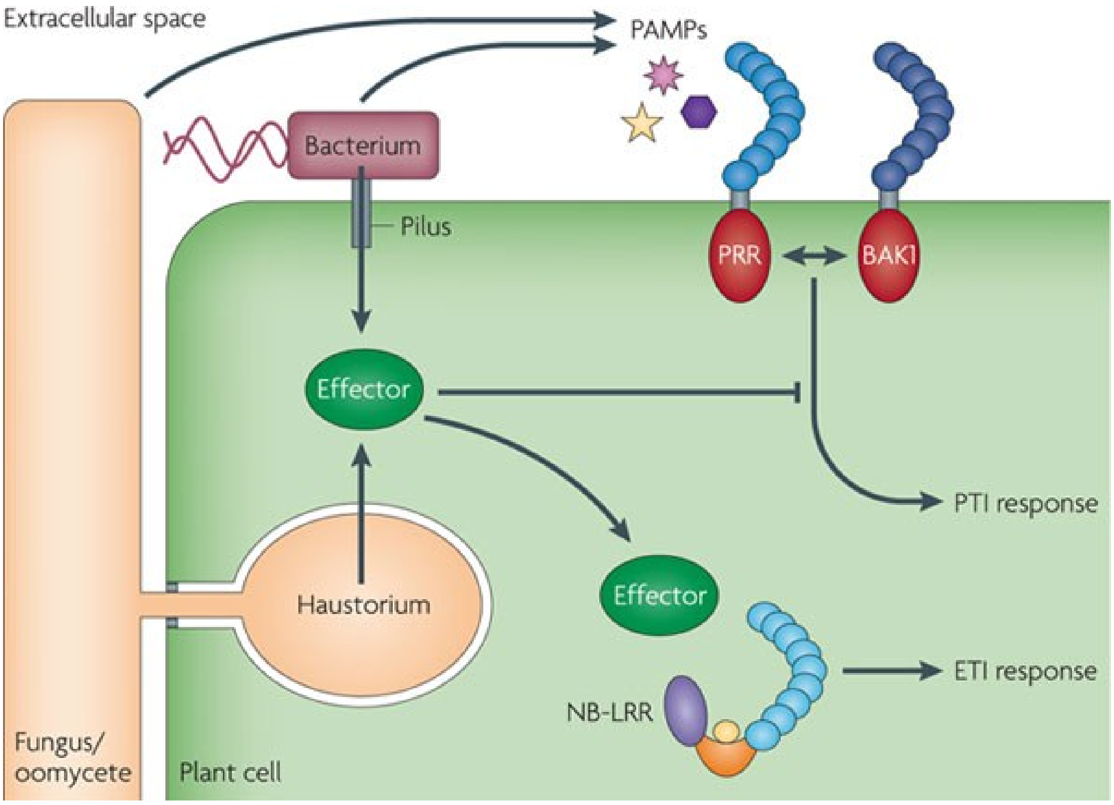



**Figure 1.3: Principles of effector biology in oomycete, fungus and bacteria [reproduced from (Dodds and Rathjen, 2010)]. PAMP = Pathogen Associated Molecular Pattern, PTI = Pathogen/Pattern Triggered Immunity, ETI = Effector Triggered Immunity, PRR = PAMP/Pattern Recognition Receptor, NB-LRR = Nucleotide Leucine Rich Repeat, BAK1 = Brassinosteroid Insensitive 1-Associated Kinase 1,**

Oomycete cytoplasmic effectors have a modular structure consisting of a secretion signal and a hypothetical secondary translocation domain (an RXLR motif in many oomycetes), followed by the C-terminal functional domain (commonly referred to as the effector domain) (fig 1.4). Until very recently, it was considered that effectors contained conserved RXLR motif after the signal peptide. This RXLR motif was considered to be involved in translocation of the effector into the host. However, recent studies have shown that within *Hpa*, which has RXLR effectors, there are effectors which do not have the RXLR motif (Bailey et al., 2011), and in the oomycete pathogen *Albugo laibachii* the majority of effector candidates carry a CHXC motif (Kemen et al., 2011).

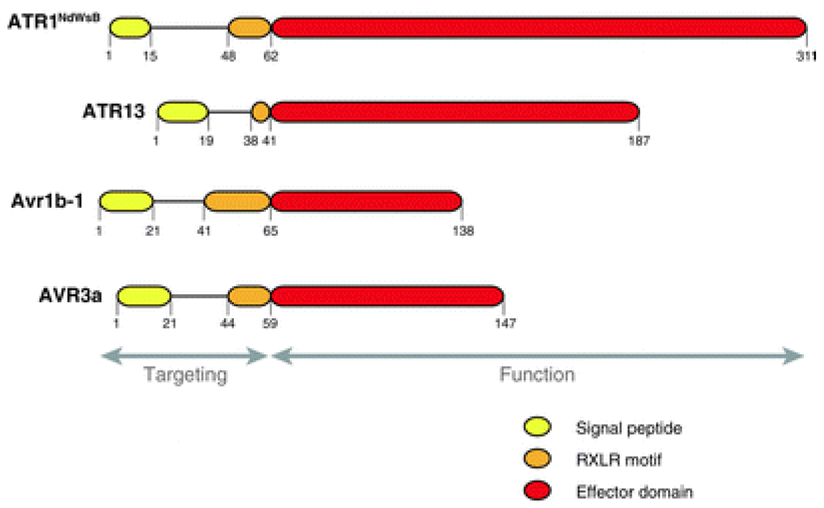



**Figure 1.4: Domain structure of oomycete apoplastic and cytoplasmic effectors [adapted from (Kamoun, 2006)].**

### 1.3.3 *Hpa* effector evolution

The importance of the pathogen having functional effectors that evade recognition, and the plant host having resistance genes that are able to recognise effectors, is paramount to the fitness of these organisms. Therefore, it is expected that their interaction will impose a very strong evolutionary selection on these organisms.

While there are more than 100 putative effector genes in *Hpa*, but only 3 have been confirmed to have avirulence activity due to recognition by an *A. thaliana* resistance gene: *ATR13* (Allen et al., 2004), *ATR1* (Rehmany et al., 2005) and *ATR5* (Bailey et al., 2011). Of these, the 2 best studied effectors in *Hpa* are *ATR1* and *ATR13*. *ATR1* was shown to have a high level of sequence polymorphism through comparative sequence analysis of 8 *Hpa* races (Rehmany et al., 2005) (fig 1.5), of which the majority of polymorphisms clustered towards the C-terminus of the gene, in the hypothesised functional domain. Rehmany et al. (2005) went on to show that the majority of the polymorphisms in the gene in this functional domain are non-synonymous mutations, while the few polymorphisms in the N-terminal domain (signal peptide and RXLR region) are synonymous mutations (fig 1.6).

```
       1                                                     50
Emoy2  MRVCYFVLVP SVALAVIATE SSETSGTIVH VFPLRDVADH RNDALINRAL
Hiks1  ·········· ·········· ·········· ·········· ··········
Waco5  ·········· ·········· ·········· ·········· ··········
Maks9  ·········· ·········· ·········· ·········· ··········
Emco5  ········L· ·········· ·····R···· ··T······· ··········
Noks1  ········L· ·········· ·····R···· ·········· ··········
Cala2  ········L· ·········· ·····R···· ·········· ··········
Emwa1  ········L· ·········· ·····R···· ·········· ··········
       51                                                    99
Emoy2  RAQTALDDDE ERWPFGPSAV EALIETIDRH GRVSLND-EA KMKKVVRTWK
Hiks1  ·········· ·········· ·········· ·······-·· ··········
Waco5  ·········· ·········· ·········· ·······-·· ··········
Maks9  ·········· ·········· ·········· ·······-·· ··E·······
Emco5  ···A······ ·······P·· ····K····· -······H·· ·····EQ···
Noks1  ···A······ ·········· K···R···S· --AG···-·· ···E··Q··N
Cala2  ···A······ ·········· K···R···S· --AG···-·· ···E··Q··N
Emwa1  ···A······ ·······L·· K···R···S· --AG···D·· ···E··Q··N
       100                                                  149
Emoy2  KLIERDDLIG EIGKHYFEAP GPLHDTYDEA LATRLVTTYS DRGVARAILH
Hiks1  ·········· ·········· ·········· ·········· ··········
Waco5  ·········· ·········· ·········· ·········· ··········
Maks9  ·········· ·········· ·········· ·········· ··········
Emco5  ······N··· ······S··· ···Y······ ·········· ··········
Noks1  ······N··· ······S··· ···Y······ ·········· ··········
Cala2  ······N··D ······SK·· ··VY·S···· ···M······ Y·········
Emwa1  ······N··D ······SK·· ··VY·S···· ···M······ Y·········
       150                                                  199
Emoy2  TRPSDPLSKK AGQAHRLEEA VASLWKGRGY TSDNVVSSIA TGHDVDFFAP
Hiks1  ·········· ·········· ·········· ·········· ··········
Waco5  ·········· ·········· ·········· ·········· ··········
Maks9  ·········· ·········· ······E··· ··H······· ·DD·······
Emco5  ·········· ·········· ·········· ·········· ··D·······
Noks1  ·········· ·········· ·········· ·········· ··D·······
Cala2  P·····I·N· ··R······· ······E··· ··H····N·· ·DD······S
Emwa1  P·····I·N· ··R······· ······E··· ··H····N·· ·DD······S
       200                                                  249
Emoy2  TAFTFLVKCV ESEDDANNAI FEYFGSNPSR YFSAVLHAME KPDADSRVLE
Hiks1  ·········· ·········· ·········· ·········· ··········
Waco5  ·········· ·········· ·········· ·········· ··········
Maks9  ·········· ·········· ·········· ·········· ··········
Emco5  ·········· ··KVE··Y·· ·K········ ·········· ··G··I·E·D
Noks1  ·········· ··KVE··Y·· ·K········ ·········· ··G··I·E·D
Cala2  ·········· ·········· ·········· ·········· ··G··I·K··
Emwa1  ·········· ·········· ·········· ·········· ··G··I·K··
       250                                                  297
Emoy2  SSKKWMFQCY AQKQ--FPTP VFERTLAAYQ SEDYAIRGAR NHYEKLSLSQ
Hiks1  ·········· ····--···· ·········· ·········· ··········
Waco5  ·········· ····--···· ·········· ·········· ··········
Maks9  ·········· ····--···· ·········· ·········· ··········
Emco5  N··N···RF· ·HA·EPLSST E··SM·PRV· ··E···H··Q ·D·····S··
Noks1  N··N···RF· ·HA·EPLSST E··SM·PRV· ··E···H··Q ·D·····S··
Cala2  N······RL· ··APEP·SP· D··WA··RF· D··H·FV··Q ·D·K···P··
Emwa1  N······RL· ··APEP·SP· G··WA··RF· D··H·FV··Q ·D·K···P··
       298        311
Emoy2  IEELVEEYSR IYSV.
Hiks1  ·········· ····.
Waco5  ·········· ····.
Maks9  ·········· ····.
Emco5  ·K···K···· ····.
Noks1  ·K···K···· ····.
Cala2  ··K······G ····TSRNFV GRASE.
Emwa1  ··K······G ····TSRNFV GRASE.
```

**Figure 1.5: Alignment of predicted *ATR1* proteins from 8 *Hpa* races [reproduced from (Rehmany et al., 2005)]. Emoy2, Hiks1, Waco9, Maks9, Emco5, Noks1, Cala2 and Emwa1 are different races of *Hpa*; dots in the sequence alignment indicate homology to the Emoy2 reference allele or ATR1.**

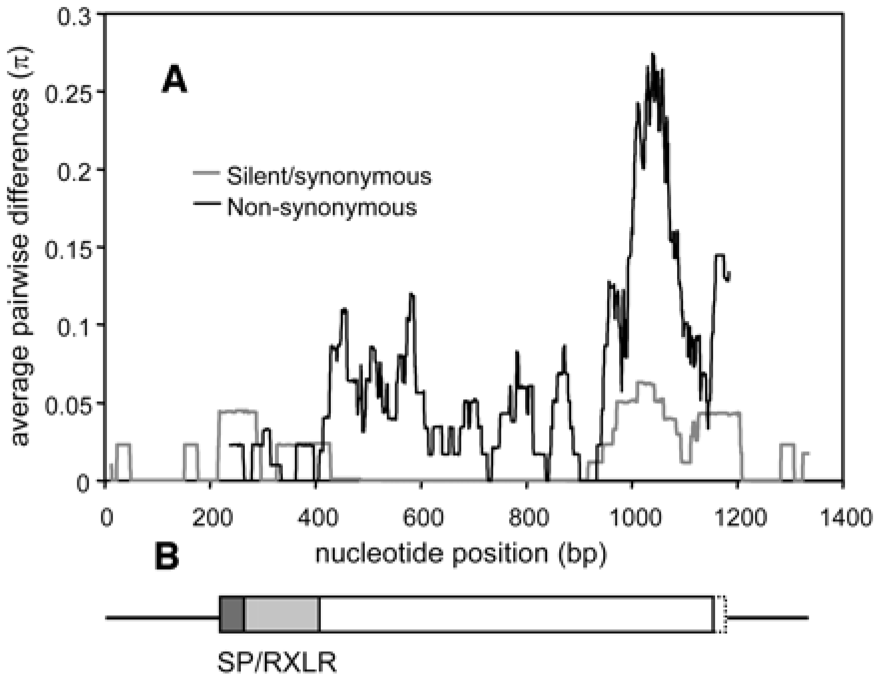



**Figure 1.6: Sliding window analysis of synonymous and non-synonymous substitutions across *ATR1* in 8 *Hpa* races [reproduced from (Rehmany et al., 2005)].**

This clustering of polymorphisms in the C-terminal region encoding for non-synonymous mutations is likely to be a signature of positive selection. A later study of *ATR13* analysed the observed polymorphisms of *Hpa* effector *ATR13* across 18 races of *Hpa*, where they put the polymorphisms into an evolutionary context (Allen et al., 2008). They showed that the dN/dS ratio of *ATR13* suggests that it is being subjected to positive selection (table 1.1).

**(a) Average pairwise differences per site (π)**

| Gene | Isolates | No. of alleles | π total | π syn* | π non† |
|---|---|---|---|---|---|
| *ATR13* | 18 | 15 | 0.042 | 0.015 | 0.05 |
| *ATR13*‡ | 17 | 14 | 0.033 | 0.014 | 0.039 |
| *Ppat5* | 16 | 11 | 0.002 | 0.010 | 0.00008 |

**(b) Chi-squared test comparing observed and expected polymorphism at *ATR13* and *Ppat5***

| Gene | Total no. of sites | Syn. obs.§ | Syn. exp.¶ | Non-obs.§ | Non-exp.¶ | $\chi^2$ | *P*-value | |
|---|---|---|---|---|---|---|---|---|
| *ATR13* | 528 | 7 | 21 | 84 | 70 | 12.04 | 0.0005 | Ka >> Ks |
| *ATR13*‡ | 531 | 7 | 14 | 55 | 48 | 4.80 | 0.0285 | Ka > Ks |
| *Ppat5* | 1983 | 12 | 3 | 1 | 10 | 4.80 | 0.0285 | Ka > Ks |

* Pairwise differences at synonymous sites.
† Pairwise differences at non-synonymous sites.
‡ Excluding Hind2 allele.
§ Number of polymorphisms observed at synonymous (or non-synonymous) sites.
¶ Expected number of polymorphisms at synonymous (or non-synonymous) sites assuming neutral evolution.

**Table 1.1: Polymorphisms in *ATR13* and *Ppat5* [reproduced from (Allen et al., 2008)].**

## 1.4 Genomics and Sequencing

The secrets of the evolutionary signature left on the genome can be unravelled by DNA sequencing, which has become an indispensable tool in many areas of biological research. About 15 years after the discovery of the double helix (Watson and Crick, 1953a, b), DNA sequencing began in 1968 by Wu and Kaiser (1968), and 3 years later they were able to report a 12 base sequence (Wu and Taylor, 1971). Since this initial publication of DNA sequencing, sequencing technology has improved dramatically.

The earliest rapid DNA sequencing technologies include Sanger's (chain termination/dideoxynucleotide) enzymatic method (Sanger et al., 1977) and Maxam and Gilbert's chemical method (Maxam and Gilbert, 1977).

### 1.4.1 Second generation sequencing

While Sanger sequencing was the pre-dominant method used for 30 years, it had various limitations which include its resource intensive library and template preparation, high running costs and relatively low throughput (Varshney et al., 2009). Advancements in microfluidics, biochemistry, nanotechnology and informatics have led to a number of new DNA sequencing technologies. These technologies were initially referred to as 'next generation sequencing' technologies, but since an even newer onset of DNA sequencing technologies, they are now more commonly referred to as 'second generation sequencing' technologies. The three most commonly used next generation sequencing technologies are pyrosequencing (employed by Roche, previously 454 Life Sciences, in the GS-FLX sequencer), sequencing by synthesis (usually referred to as Solexa sequencing, employed by Illumina, previously Solexa, in the Genome Analyser sequencer, HiSeq and MiSeq sequencers) and sequencing by ligation (employed by Applied Biosystems in the ABI SOLiD sequencer).

### *1.4.1.1 454 Pyrosequencing*

Developed by 454 Life Sciences (now owned by Roche), this method parallelised the sequencing by synthesis method (SBS) employed in pyrosequencing (Ronaghi et al., 1999). It was the first second generation sequencing technology to be made available. In the sequencing process DNA fragments anchored to beads and are amplified via emulsion PCR, which are then put into wells on a plate. dNTPs are washed over the wells in waves. As the nucleotides are incorporated into the new DNA strand, the intensity of the light given off is used as a measure of how many As, Ts, Cs or Gs have been incorporated (fig 1.8).

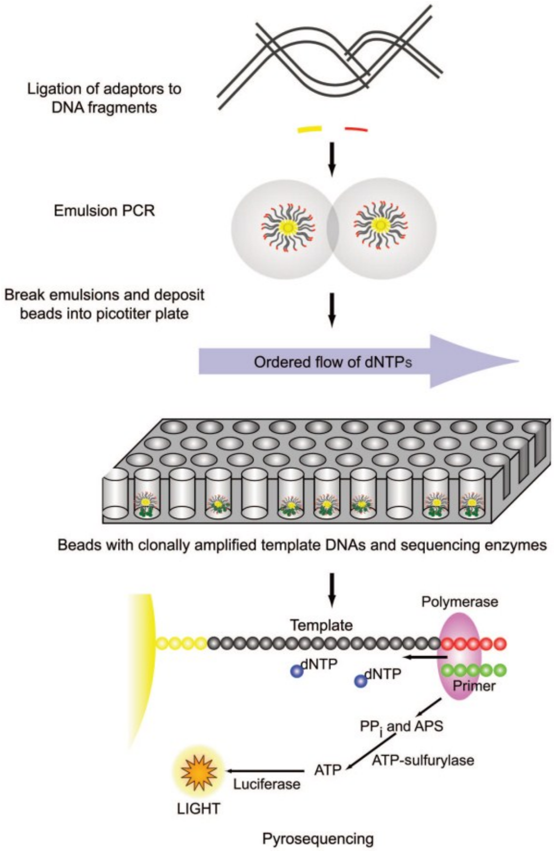



**Figure 1.7: Roche 454 GS FLX sequencing method [reproduced from (Voelkerding et al., 2009)]. $PP_i$ = pyrophosphate, APS = adenosine 5 –phosphosulfate.**

Previous reports showed that the 454 sequencing method is capable of sequencing 400-600 Mb of DNA over about 1 million reads of >=400 bp per 10-hour run (Voelkerding et al., 2009).

### *1.4.1.2 Sequencing by synthesis*

Sequencing by synthesis (SBS) (commonly referred to as Solexa sequencing) was developed by Solexa and was then acquired by Illumina and implemented in the Genome Analyser, HiSeq and MiSeq sequencers. This method uses a glass surface (flowcell) to capture molecules, which are subsequently bridge PCR amplified into clusters (fig 1.9). After this, dye labelled terminators are added and an image of the surface is taken, with information on fluorescence (correlation to bases) recorded. The dye is then cleaved and another layer of dye labelled terminators is added. This process is repeated until the whole fragment has been sequenced. The resultant images are processed to reveal the sequence of the DNA present at each cluster position.

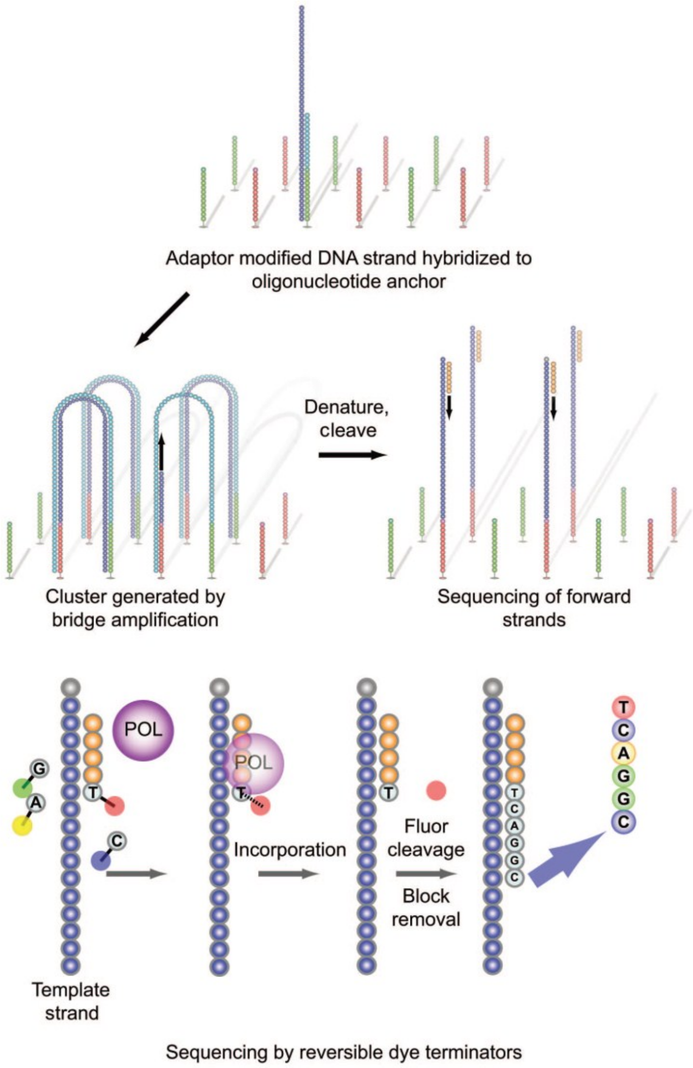



**Figure 1.8: Illumina sequencing method [reproduced from (Voelkerding et al., 2009)]. POL = polymerase.**

While the earlier protocols of Illumina sequencing yielded >1 Gb of 36 bp reads per run (Bentley et al., 2008), the newest iterations of the original machine (Illumina GAIIx) is capable of producing 100 Gb per run and 150 bp paired end read lengths (www.illumina.com).

### 1.4.2 Second generation sequencing applications and tools

While Sanger sequencing produces longer reads of higher quality than any second generation sequencing method, the high cost per base and lower throughput has led to the mass adoption of second generation sequencing for a myriad of tasks. I will discuss some of the applications and tools of second generation sequencing from the perspective of the Illumina platform.

#### *1.4.2.1 Genome assembly*

One would assume that genome assembly of eukaryotic organisms would be best left to sequencing technologies with longer reads, such as Sanger and 454. However, rapid advances in Illumina sequencing in throughput, read length and insert size have facilitated genome assemblies of, for example, *Albugo laibachii* (*A. thaliana* white rust) (Kemen et al., 2011) and the panda genomes (Li et al., 2010). Illumina sequencing is also the preferred platform for sequencing of prokaryotes, as prokaryotes have a less complex genome compared to eukaryotes, so the read length offered by Illumina sequencing is sufficient.

Recently, researchers have started to use second generation sequencing for metagenomics projects, such as for the sequencing of the human gut microbiome (Qin et al., 2010) and the human oral microbiome (Lazarevic et al., 2009).

Some of the first short read assemblers (Jeck et al., 2007; Warren et al., 2007) were based on the overlap consensus method. This method uses overlaps between sequences to create links between them, whereby a contig is formed when the links are followed as far as possible (fig 1.11). While this method was efficient when the throughput of sequencing machines was still comparatively low, it has the inherent problem that the memory requirements scale with the number of input reads. Therefore, it soon became unfeasible to use the overlap consensus method for genome assembly with the large amount of data produced by second generation sequencing technology.

The next phase of short read assemblers (Butler et al., 2008; Chaisson and Pevzner, 2008; Zerbino and Birney, 2008) were able to overcome the issue of memory requirement scaling directly with the number of reads by implementing a de-Bruijn graph based assembly method. This method uses a unique set of k-mers (all subsequences of length k within the read) and the reads are represented as a path between the k-mers (fig 1.11). By virtue of this method, the links between the reads are established as the data is read.

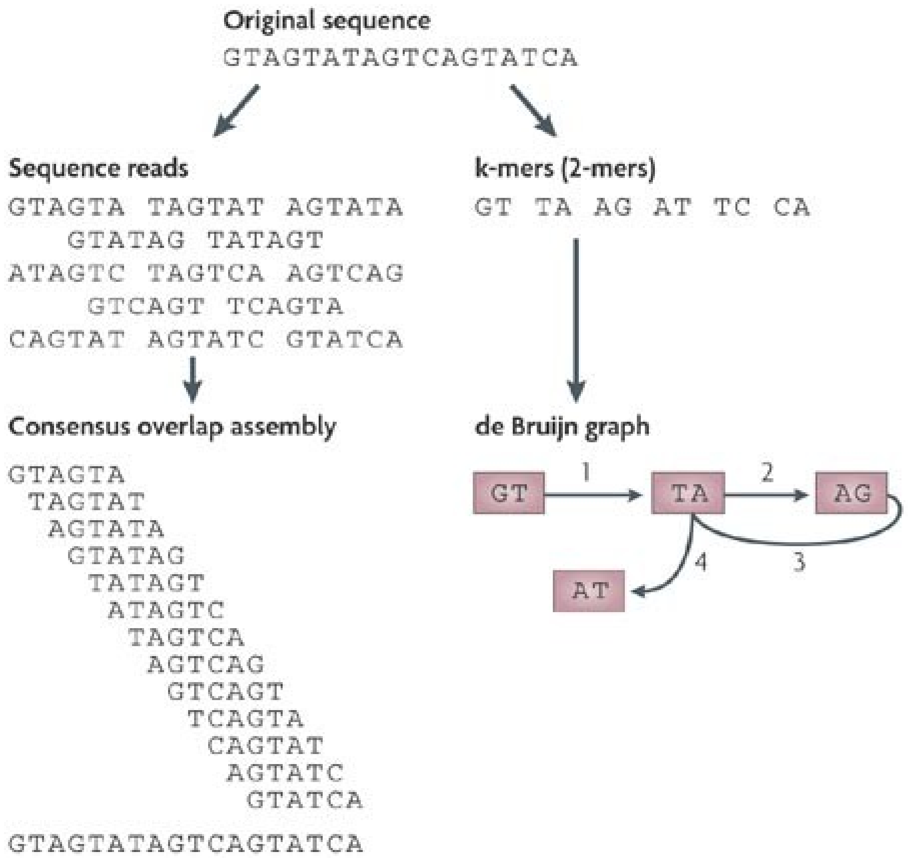



**Figure 1.9: Illustration of how a read is reassembled using consensus overlap and the De-Bruijn principles [reproduced from (MacLean et al., 2009)].**

### *1.4.2.2 Alignment and variant calling*

A major use of short read sequencing is to make inferences of variation between genomes by aligning short reads to a reference genome assembly. While there are many alignment algorithms, the early programs such as MAQ (Li et al., 2008a), SOAP (Li et al., 2008b) and SSAHA (Ning et al., 2001) have been superseded by aligners using the Burrow-Wheeler transform (Burrows and Wheeler, 1994) such as BWA (Li and Durbin, 2010), SOAP2 (Li et al., 2009b) and Bowtie (Langmead et al., 2009). Recently a number of alignment programs have started to employ the massive parallelisation offered by modern day graphics processing units (GPUs), such as MUMmerGPU (Schatz et al., 2007) and SOAP3.

After reads have been aligned to a reference genome, a number of inferences can be made by analysing differences between the reference sequence and the aligned reads:

- Nucleotide variation can be identified as SNPs and INDELs through programs such as SAMtools (Li et al., 2009a)
- Structural variation can be identified using programs such as BreakDancer (Chen et al., 2009)
- Copy number variation can be identified through programs such as CNV-seq (Xie and Tammi, 2009)
- Through bulk segregant mapping, gene mapping projects can be undertaken using programs such as SHOREmap (Schneeberger et al., 2009)
- ChIP-seq and Bis-seq experiments can be scored using programs such as PeakSeq (Rozowsky et al., 2009)

#### *1.4.2.3 Transcriptomics*

In addition to DNA sequencing, RNA, cDNA and expression tag sequencing can be performed using the second generation sequencing. There are a number of programs that are able to assemble RNA/cDNA sequencing data allowing discovery of unannotated transcripts, new isoforms (e.g. Cufflinks (Trapnell et al., 2010)) and splice site junctions (e.g. TopHat (Trapnell et al., 2009)).

Recently efforts have been made to make use of the very high throughput of Illumina sequencing for expression analysis. The benefit of using a sequencing based technique compared to microarrays is that the method is open and experiments can be designed more easily. Many of the methods make used of a protocol that digests the RNA/cDNA, ligation of barcoded adapters to allow for multiplexing and a number of repeats to infer statistical significance.

## 1.5 Objectives of the project

It is apparent that successful pathogens are able to colonise plant hosts, by delivering effectors that are able to suppress host defence and manipulate host function, to allow the pathogen to complete its lifecycle. In recent years, progress has been made in identifying and understanding oomycete effector biology, but there are many aspects of effector biology that remain unknown due to the lack of known effectors.

The main goal of this project is to use Illumina sequencing technology to better understand the nature of effectors from an evolutionary standpoint. During the start of my project the *Hpa* genome assembly project was in preliminary stages. The scope of my project was to use Illumina sequencing of the reference race, *Hpa* Emoy2, to assist with the genome assembly of *Hpa* (chapter 3). After genome assembly, expression data (Sanger sequenced ESTs and Illumina sequenced cDNA of *Hpa* Emoy2) were used to assist with gene model predictions, from which inferences about *Hpa* biology and better understanding of *Hpa*'s complement of virulence related genes can be made (chapter 4). After the foundations of a good reference sequence and good gene models with annotated effectors, I made use of Illumina reads of 8 natural *Hpa* races isolated from various geographical locations (Cala2, Emco5, Emoy2, Hind2, Maks9, Noco2 and Waco9 sequenced in house, and Emwa1 provided by Prof Brian Staskawicz, UC Berkeley) and performed comparative genomics analysis on *Hpa* genes using a custom made generalised analysis pipeline (chapter 5).

# Chapter 2 – Materials and Methods

## 2.1 Biological material and sequencing

The majority of the biological material sequenced used was that of *Hpa* Emoy2. The other races of *Hpa* sequenced by the lab using Illumina paired end sequencing include Noco2 from the Jones Lab (TSL, Norwich) and Cala2, Emco5, Hind2, Maks9, and Waco9 which were obtained from Prof Eric Holub at HRI, Warwick. Illumina paired end sequenced reads of *Hpa* Emwa1 were provided by Prof Brian Staskawicz, UC Berkeley.

The preparation and Sanger sequencing of the *Hpa* Emoy2 DNA, ESTs and BACs were coordinated by collaborators Dr Sucheta Tripathy (VBI, Virginia Tech) and Dr Laura Baxter (HRI, Warwick).

These material and methods are the same as those published in Baxter et al. (2010), with additional data used for the comparative genomics (chapter 5).

### 2.1.1 Sanger sequencing

The Sanger sequencing protocols for the DNA reads, BACs and ESTs are described in Baxter et al., 2010.

### 2.1.2 Illumina Sequencing

#### *2.1.2.1 DNA extraction.*

Genomic DNA was extracted from *Hpa* conidospores from infected *A. thaliana* Ws *eds1-1* plants using a Nucleon PhytoPure DNA extraction kit using the default protocol followed by a Phenol / Chloroform extraction and Isopropanol precipitation.

#### *2.1.2.2 Illumina DNA library preparation and sequencing.*

The non-paired end libraries were sequenced on the Illumina GA1 platform using 120bp inserts. The paired end libraries were sequenced on the Illumina GA2 platform using 400bp (+/- 10%) inserts. The protocol used was the same as the manufacturers protocol apart from the purification of the ligation of the Illumina adapters were performed on a 5% polyacrylamide gel and the library validation was performed a 6% polyacrylamide gel. The base calling was done on the Illumina GAP v1.0 pipeline for all runs before flowcell ID71 (appendix table 2.1; appendix table 2.2) after which the GAP v1.3 pipeline was used.

#### *2.1.2.3 Quality checking the Illumina preparation and sequencing.*

The libraries were sequenced on a single lane initially for quality checking after which the decision to sequence further lanes was made. For both the paired and non-paired sequencing runs, a PhiX control lane was also run to eliminate mechanical error. The raw reads generated from the Illumina Pipeline included errors in the form of PCR duplicate reads, adapter contamination and *Xanthomonas* contamination. Contamination was dealt with through post analysis filtering through sequence homology analysis. The reads were analysed for quality using FastQC (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/) (appendix figs 2.1). This analysis revealed certain quality issues:

- The per base quality drops drastically in the last third of the read for sequencing runs before the implementation of the GA pipeline v1.3
- The Emwa1 reads have high levels of Illumina paired end sequencing primer contamination
- The reads for each sequenced race have between 2% and 25% PCR duplication

Despite the per-base quality decrease in the last third of the read, the average read quality of the reads have a single peak around a Phred scaled quality score 30, which implies an overall error rate of 0.1%. Therefore, the reads were not filtered or trimmed before alignment, but instead, I modified the alignment parameters to soft-trim bad quality trailing bases and filtered the PCR duplicates post-alignment. The soft trimming was set between Phred scaled quality scores of Q10 (10% chance of the base call being incorrect) and Q20 (1% chance of the base call being incorrect) depending on the analysis. The specific values used are mentioned in each results section. PCR duplicates were removed after alignment to a reference genome sequence to avoid spurious variation calls.

### 2.1.3 Illumina cDNA sequencing.

*Hpa* RNA was extracted from infected leaves of 7 days post inoculation (d.p.i.) *A. thaliana* Ws *eds1-1* using TRI-REAGENT according to protocol (Sigma). RNA was resuspended in DEPC treated water. RNAse inhibitors (RNAseguard, Promega) was added and samples were DNAse treated (RNAse free, Roche). RNA was re-extracted with phenol/chloroform, EtOH precipitated and resuspended in DEPC treated water. First and second strand cDNA synthesis was performed using the default protocol from the Creator SMART cDNA Library Construction KitTM (Clontech). After the last amplification step cDNA was phenol/chloroform extracted followed by Isopropanol precipitation. The cDNA was then normalised using Duplex-specific nuclease (Evrogen) according to default protocol. The normalised cDNA was than prepared to be sequenced on the Illumina platform using 120bp inserts with a 35 bp read length.

### 2.1.4 454 Sequencing

*Hpa* RNA was extracted from infected leaves 3 d.p.i. of 3 week-old *A. thaliana* Ws eds1-1 using a protocol adapted from (White and Kaper, 1989). RNA was resuspended in DEPC treated water. RNAse inhibitors (RNAseguard, Promega) was added and samples were DNAse treated (RNAse free, Roche). RNA was re-extracted with phenol/chloroform, EtOH precipitated and resuspended in DEPC treated water. First and second strand cDNA synthesis was performed using the default protocol from the Creator SMART cDNA Library Construction KitTM (Clontech). After the last amplification step Proteinase K digestion was performed with the whole of the reaction and not just with half as in the Creator SMART protocol. cDNA was phenol/chloroform extracted and EtOH precipitated using 1.3 µg glycogen. For positive selection of *Hpa* cDNAs, 4 µg of genomic DNA, genomified using the GenomiPhi Kit (GE Healthcare), digested and biotinylated (Rougon-Cardaso, 2007) were mixed with the target (cDNA synthesized from RNA of infected leaves) and EtOH precipitated. The mixture was resuspended in 10 µl of sterile hybridization buffer after which the driver and target were denatured at 95$^{o}$C for 10 minutes and hybridized for 36 hours at 66$^{o}$C. The biotinylated DNA was captured by Streptavidin coated beads (Magnasphere Paramagnetic Beads (Promega)). Hybrids were recovered using a protocol adapted from Bashiardes et al. (2005). Positively selected cDNA was digested with Sfi I and Mse I restriction enzymes. Oligonucleotide fragments and salt were removed by spin-column chromatography through a Sephadex G-25 resin (Roche) after each digestion. The following primers were ligated to form adapters 454A and 454B: Biot-SfiAdaptor454Aoverhang, Biotin-AGCCTCCCTCGCGCCATCAGATTA; SfiAdapter454Acomp, $PO_4$-TCTGATGGCGCGAGGGAGGC; Mse-TOP, TACTGAGCGGG CTGGCAAGGC; Mse-BOT, GCCTTGCCAGCCCGCTCAG.

Four hundred ng of cDNA were ligated with 300 ng adapter 454A and 300 ng of adapter 454B. Biotinylated fragments were hybridized to 20 µl Magnasphere Paramagnetic Beads (Promega) pre-washed as specified by the manufacturer and pre-incubated with blocking agents. Beads with hybridized cDNA were washed 4 times with 0.1xSSC and captured with a magnet (Promega) and supernatant was discarded. After preparation of cDNA with 454 adaptors attached the sample was sent to 454 Life Sciences (Branford, Connecticut, USA) for further processing and sequencing with 454 GS-FLX technology.

The returned 454 sequenced reads were filtered for oomycete ribosomal genes and *A. thaliana* contamination.

## 2.2 Software and protocols

### 2.2.1 Assembly

#### *2.2.1.1 Assembly of Sanger reads*

The *Hpa* Emoy2 v7 assembly was sequenced to 9.5x phred Q20 redundancy (9.5X coverage) through 1,080,646 plasmid end reads and 25,516 fosmid end reads and 13,071 BAC end sequences. The combined sequence reads were assembled using the PCAP software (Huang et al., 2003). The 'bdocs' and 'bclean' commands of PCAP were then used to process the overlaps, and 'bcontig' to calculate the layout to generate the consensus sequence, using default parameters. Using this dataset, a round of automated sequence improvement was done. 23,855 of 32,122 pre-finishing reads were incorporated into the initial assembly.

The initial PCAP assembly, consisting of only plasmid end sequences, contained 1,053,419 reads, yielding more than 8 fold coverage for an estimated 70 Mb shotgun assembly. A total of 1,014,758 reads was assembled using PCAP. The final PCAP assembly included all plasmid sequences, 8346 BAC end sequences, 25,516 fosmid end sequences, and a round of automated pre-finishing. Additional filtering following assembly removed contigs less than 2kb, as well as *A. thaliana* and sequencing plasmid contaminants. 5354 contigs (including a large number of singletons) were removed by this process, with 5473 contigs and 1842 scaffolded contigs remaining. Approximately 99% of the 76 Mb shotgun assembly is covered.

#### *2.2.1.2 Short read assembly*

The Velvet algorithm (v0.7.55) (Zerbino and Birney, 2008) was used for short read assembly. Specific parameters used are described in chapter 3.

#### *2.2.1.3 Hybrid assembly*

The Hybrid assembly of the *Hpa* genome is described in chapter 3. It is based on the targeted assembly and re-integration method described by Ossowski et al. (2008), MAQ v0.7.1 (Li et al., 2008a), BLAT v34 (Kent, 2002), Velvet v0.7.55 (Zerbino and Birney, 2008) and custom scripts were used.

### 2.2.2 Alignment

#### *2.2.2.1 DNA alignment*

In chapter 3, Illumina sequenced DNA reads are aligned to the genome and variants were called using MAQ v0.7.1 (Li et al., 2008a). The parameters for each alignment are mentioned in the appropriate sections in chapter 3.

When I started the *Hpa* comparative genomics analysis a new selection of short read aligners were available, and MAQ was no longer supported. In chapter 5, BWA v0.5.8c (Li and Durbin, 2009) was used a primary aligner to its specificity and speed. A further round of alignment was done using Stampy v1.0v11 (Lunter and Goodson, 2011), which is slower but more sensitive than BWA. Variants were called using the SAMtools suite (Li et al., 2009a).

A simulation of SNP recall rate using various methods was performed. A total of 10,000 SNPs and 2000 INDELs were introduced in the *Hpa* Emoy2 genome in regions covered by reads, with average depth of coverage and at least 500 bp from existing and artificial variation. The SNPs were introduced as single SNPs, 2 SNPS that are 15-35 bp apart (simulating clustered SNPs), and 2 SNPs that are 2-15 bp apart (simulating clustered SNPs on the same seed). INDELs were introduced as mainly 1 bp INDELs, but also of length 2, 3, 4, 6, 9 and 12.

A number of pre-processing techniques were employed:

- Removing reads with any N's
- Removing reads with more than 1 N
- Read correction using HiTEC (Ilie et al., 2011) (using default parameters)
- Quality trimming reads from the end of the read so that all bases in a read have a minimum PHRED scaled quality of Q13 (Q13 ~ P(0.05) of error) using SolexaQA (Cox et al., 2010)

The alignment programs and variant call methods used were used were:

- MAQ for aligning and variant calling
- MAQ for aligning and SAMtools for variant calls
- BWA for aligning and SAMtools for variant calls (used for unfiltered and filtered reads)
- Stampy for aligning (with BWA pre-alignment) and SAMtools for variant calls

Three technical replicates of the alignments were performed and the variation recall rate and false positive rates were taken from the averages. The average variation recall rate, with a minimum quality score of Q10, showed that pre-filtering of reads had a very small negative effect on the number of variation recalled (fig 2.1). Results also showed that on average, Stampy was slightly more sensitive in recalling variation compared to BWA, and that MAQ was significantly worse than BWA and Stampy at recalling INDELs. It was interesting to note that the recall rate of deletions is on average 15% more than SNPs and 20% more than insertions.

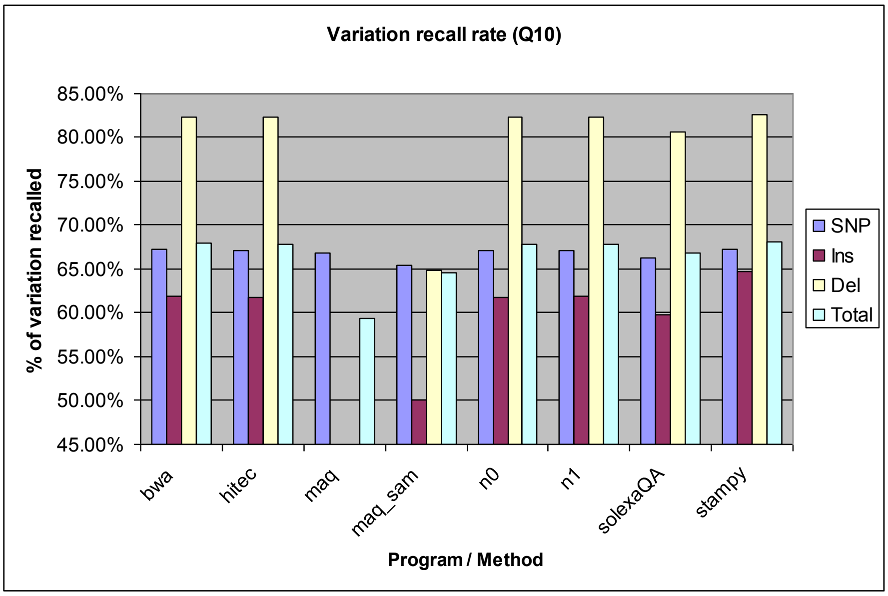



**Figure 2.1: Variation true positive recall rate of various mapping techniques at Q10.**

The average variation false positive rate (including call that were between 5-500 bp from inserted variation), with a minimum quality score of Q10 were 4.93% of the introduced variation (fig 2.2). For all methods there was a low false positives for INDEL predictions (averaging 2.01%), while the rate of false positive in SNPs was 5.32% on average. The variation between the false positives in the variation recall rate was 0.7%, so all methods were very similar.

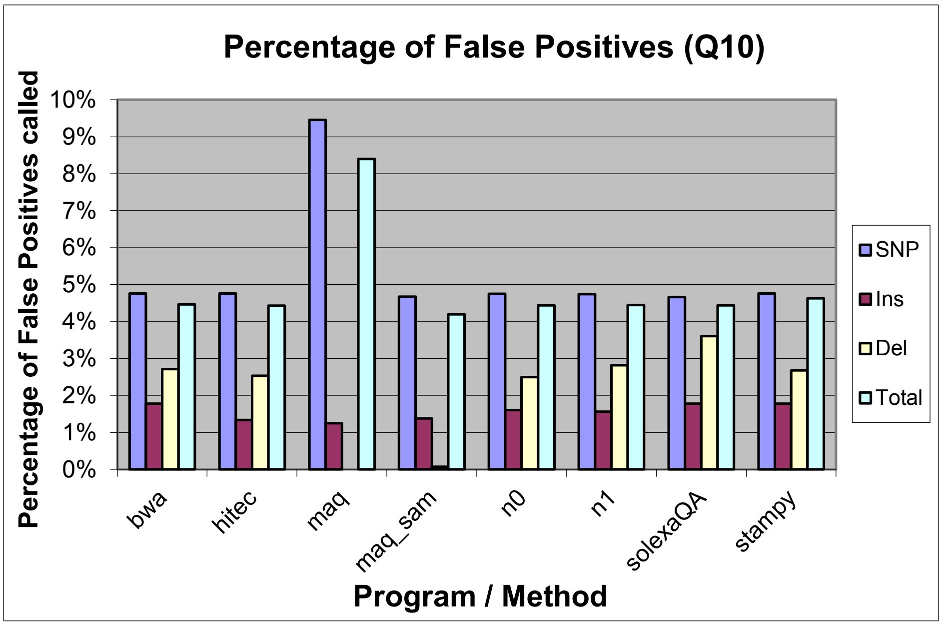



**Figure 2.2: Variation false positive rate of various mapping techniques at Q10**

The sensitivity (true positives / (true positives + false positive)) of each method was calculated (fig 2.3).

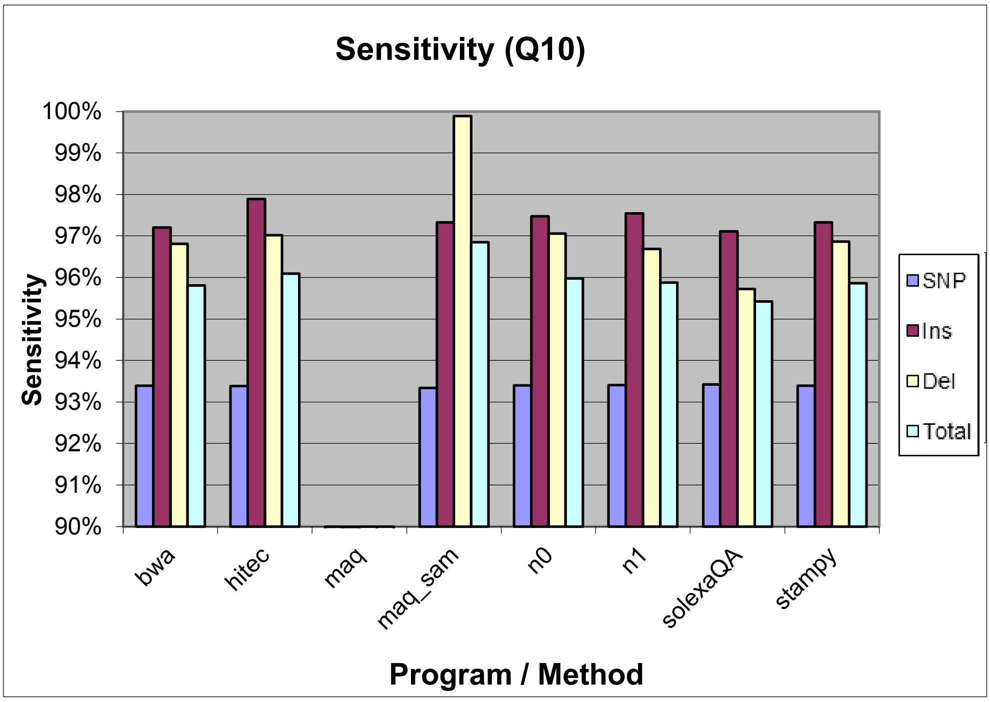



**Figure 2.3: Sensitivity of various mapping techniques at Q10**

While HiTEC read correction offered the best sensitivity, it was a very time consuming to compute for a single race. It was decided that using Stampy and BWA as a pre-aligner offered the best combination of true positive recall rate and sensitivity, and was used for the analysis performed in chapter 5.

### *2.2.2.2 cDNA alignment*

In chapter 3 and 4, Illumina sequenced cDNA reads are aligned to the genome using MAQ v0.7.1 (Li et al., 2008a) when evaluating the genome assembly and gene models. In order to train splice site predictions a combination of Bowtie (Langmead et al., 2009), Tophat (Trapnell et al., 2009) and Cufflinks (Trapnell et al., 2010) were used using default parameters.

For the genome browser the 5' SAGE tags were aligned using Novoalign from the Novocraft suite v 2.05.13 (www.novocraft.com) using default parameters but allowing for no mismatches. The reads had the sequencing primer, and barcodes removed using a custom script based on Hamming distances.

For the genome browser, the 454 ESTs were aligned using BLAT v34 (Kent, 2002).

### 2.2.3 Gene predictions

Gene prediction software used in chapter 4 are:

- Genezilla (Majoros et al., 2005)
- Snap (Korf, 2004)
- CEGMA (Parra et al., 2007)
- GeneID (Guigo, 1998)
- PASA (Haas et al., 2003)
- Augustus (Stanke et al., 2008)

Their usage is described in detail in chapter 4.

### 2.2.4 Gene annotation

The programs used for annotation of the gene models in chapter 4 are:

- ProDom (Bru et al., 2005) using BlastProDom (Blastall) (Zdobnov and Apweiler, 2001)
- PRINTS (Attwood et al., 2003) using FingerPRINTScan (Scordis et al., 1999)
- SMART (Letunic et al., 2002) using Hmmpfam (Finn et al., 2011)
- TIGRFAMs (Haft et al., 2003) using Hmmpfam (Finn et al., 2011)
- Pfam (Bateman et al., 2004) using Hmmpfam (Finn et al., 2011)
- PROSITE (Hulo et al., 2004) using ScanRegExp + ProfileScan (Thompson et al., 1994b)
- PIRSuperFamily (Wu et al., 2004) using Hmmpfam (Finn et al., 2011)
- SUPERFAMILY (Gough et al., 2001) using Hmmpfam (Finn et al., 2011)
- CATH (Pearl et al., 2000) using Hmmpfam (Finn et al., 2011)
- PANTHER (Thomas et al., 2003) using Hmmsearch (Finn et al., 2011)
- Transmembrane using TMHMM2.0 (Sonnhammer et al., 1998)

- Signal peptides using SignalPHMM (Bendtsen et al., 2004)
- Low complexity regions using SEG (Wootton and Federhen, 1993)
- 3D Structure using Gene3D
- Coiled coils using COILS (Lupas et al., 1991)
- WolfPsort (Horton et al., 2007)
- SignalP 3.0 HMM (Bendtsen et al., 2004)
- KAAS (Moriya et al., 2007)

There usages are described in more detail in chapter 4.

### 2.2.5 Evolutionary analysis

The evolutionary analysis of genes was performed using a combination of a customised pipeline, VariTale (described in chapter 5), PAML v4.0 (Yang, 2007) and DnaSP v5 (Librado and Rozas, 2009). The protocol used is described in chapter 5.

# Chapter 3 – *Use of sequencing by synthesis to evaluate and improve the Hyaloperonospora arabidopsidis genome assembly*

## 3.1 Introduction

At the start of this project, genome sequences for 3 oomycete pathogens *Phytophthora sojae*, *P. ramorum* (Tyler et al., 2006) and *P. infestans* (Haas et al., 2009) were published. Comparative genomics of *Hpa* with these *Phytophthora* species might enhance our understanding of conserved pathogenicity mechanisms in the Peronosporales and distinct mechanisms unique to the Peronosporacaea and Pythiacaea. In addition, the study of *Hpa* will improve our understanding of obligate biotrophy. It is also very important to characterise the evolutionary pressures being exerted in plant-pathogen interaction systems, and the *Hpa*-*Arabidopsis thaliana* interaction system provides a model system for studying a plant-pathogen interaction involving an oomycete obligate biotroph.

Effectors play an important part in pathogenicity. Before the publication of the *Hpa* genome sequences, only 2 effector genes had been characterised in *Hpa*, *ATR13* (Allen et al., 2004) and *ATR1* (Rehmany et al., 2005), which were identified through forward genetic approaches. The availability of the *Hpa* genome sequence provides the opportunity to identify and define the repertoire of effectors. This raises the potential for high throughput characterisation of effectors through targeted reverse genetic approaches to understand effector virulence and avirulence function.

In this chapter I discuss the development of the most recent version, version 8.3, of the *Hpa* Emoy2 genome assembly. This reference genome project was initially a capillary sequencing genome project. In the later stages of the project it became clear that Illumina short read sequences were able to contribute a lot more to the genome assembly than simply the identification of SNPs. In this chapter I describe how we developed novel methods to evaluate genome assemblies and improve the assembly using Illumina sequence data.

I also show that heterozygosity in diploid organisms is an important source of variation that is often overlooked, and can provide useful insights into signatures of selection pressure in organisms.

## 3.2 Results and discussion

### 3.2.1 Establishing a short read *de-novo* assembly for *Hpa* Emoy2

#### *3.2.1.1 Hpa version 7 assembly*

When I started work on the project, the *Hpa* Emoy2 version 7 (v7) (appendix table 3.1) was the most recent assembly making use of only Sanger sequenced reads. The assembly was performed by the Genome Sequencing Centre, University of Washington at St Louis, Missouri, in December 2007. The *Hpa* Emoy2 v7 assembly was sequenced to 9.5x phred Q20 redundancy (9.5X coverage) through 1,080,646 plasmid end reads, 25,516 fosmid end reads and 13,071 BAC end sequences. The combined sequence reads were assembled using the PCAP software (Huang et al., 2003). The 'bdocs' and 'bclean' commands of PCAP were then used to process the overlaps, and 'bcontig' to calculate the layout to generate the consensus sequence, using default parameters. Using this dataset, a round of automated sequence improvement was performed. 23,855 of 32,122 pre-finishing reads were incorporated into the initial assembly.

The initial PCAP assembly, consisting of only plasmid end sequences, contained 1,053,419 reads, yielding more than 8 fold coverage for an estimated 70 Mb shotgun assembly. A total of 1,014,758 reads was assembled using PCAP. The final PCAP assembly included all plasmid sequences, 8346 BAC end sequences, 25,516 fosmid end sequences, and a round of automated pre-finishing. Additional filtering following assembly removed contigs less than 2 kb, as well as *Arabidopsis thaliana* and sequencing plasmid contaminants. 5354 contigs (including a large number of singletons) were removed by this process, with 5473 contigs and 1842 scaffolded contigs remaining. Approximately 98-99% of the 76 Mb shotgun assembly is covered.

The 76 Mb assembly consisted of 1585 major scaffolds (larger than 2 kb) with an N50 scaffold number (the minimum number of scaffolded contigs to represent at least 50% of the genome assembly) of 68. The assembly consisted of 9 Mb of ‘N’s which represent unknown sequence between paired-end Sanger reads that were used for scaffolding.

The improvements of the *Hpa* Emoy2 v7 assembly over the previous version 6 (v6) (appendix table 3.1) are that *Arabidopsis thaliana* contamination and plasmid vector sequences were removed. Also, the *Hpa* Emoy2 v6 assembly used EST sequences in the assembly – this was a misuse of the ESTs as they may lead to assembly artefacts due to differences between the DNA and RNA due to splicing. This mistake of using ESTs was avoided in the v7 assembly.

#### *3.2.1.2 Identifying ‘uncloned’ regions of the Hpa genome*

*Hpa* may contain elements in its genome that cannot be cloned using the vectors used for the Sanger sequencing project of *Hpa* or which, due to the random sampling of shotgun sequencing, were not included in the clones.

The Illumina sequenced paired end short reads of *Hpa* Emoy2 were aligned to the 1,162,037 sequences of the *Hpa* Emoy2 Sanger shotgun reads, from the trace archives, using MAQ (Li et al., 2008a) (default parameters). 5,721,482 reads did not align against the Sanger shotgun reads. These reads were assembled using Velvet 0.7.18 (Zerbino and Birney, 2008) (k-mer length= 25, cov_cutoff=2). The assembly totalled to 1,061,433 bp over 1226 contigs with N50 length of 1157 bp. The longest contig length was 9010 bp.

To identify what was contained in the assembly of reads not sequenced by the Sanger shotgun sequencing method, a BLASTx (Altschul et al., 1990) search of this assembly versus the NCBI NR proteins database (July 2009) was conducted. Much of the assembly had DNA sequence similarity to *Hpa* sequences, BAC and to *Arabidopsis thaliana* sequences. This indicates that the stochastic nature of shotgun assemblies failed to sample known regions of the *Hpa* genome and that there is *Arabidopsi*s contamination in the Illumina reads of *Hpa*. After removing these sequences, what looked like bacterial and plant contamination was left. However, there were several sequences that showed about 80% DNA sequence identity with oomycete genes. These might plausibly be bona fide *Hpa* sequences that were not represented in the Sanger clones. Some notable hits included a delta-1-pyrroline-5-carboxylate reductase (*P. nicotianae*), a NADH dehydrogenase (*P. infestans*), and a putative nuclear LIM interactor-interacting protein (*NIF5*) gene (*P. sojae*).

A tBLASTx (Altschul et al., 1990) search against the NR nucleotide database (July 2009) found hits with similarity to *Phytophthora* species. Some notable hits included a necrosis and ethylene-inducing protein (*P. megakarya*), a hsf transcription factor (*P. sojae*), a reverse transcriptase (*P. parasitica*), and a rpL41-like protein (*P. sojae*).

These results suggested that even with an optimal assembly of all of the Sanger reads, the Illumina data contained novel sequence that can help improve the genome assembly of *Hpa* Emoy2.

### *3.2.1.3 De-novo short read assembly of Hpa Emoy2 using Velvet*

To evaluate the state of the *Hpa* Emoy2 assembly we decided to compare it to another existing assembly. At the beginning of the project the v6 assembly was the only comparative assembly using the same data. Unfortunately, the v6 assembly contained a significant amount of *A. thaliana* contamination, vector sequence and *Hpa* Emoy2 EST sequences, which made it less suitable for our comparison. Therefore, we compared the v7 assembly to an Illumina GA2 sequenced assembly, as this provided us with a comparative benchmark of the 2 technologies and insights into the limitations of both technologies.

Using the Velvet algorithm (v0.7.55) (Zerbino and Birney, 2008), we derived a strategy to assemble 8 lanes (56.7 Mb) of *Hpa* Emoy2 paired-end reads (appendix table 3.2). This strategy involved performing a parameter scan over the k-mer length and coverage cut-off.

The k-mer hash was constructed using the ‘velveth’ command with default parameters. The first assembly was performed using the ‘velvetg’ command with default parameters. The statistics for these assemblies is shown in table 3.1:

| k-mer length | Number of contigs | Mean contig length | Median contig length | Sum of contigs | Longest contig |
|---|---|---|---|---|---|
| 21 | 346,226 | 171 | 67 | 59,252,988 | 14,386 |
| 23 | 287,586 | 206 | 73 | 59,472,127 | 42,735 |
| 25 | 256,288 | 235 | 81 | 60,245,632 | 50,706 |

**Table 3.1: velvetg run statistics – default parameters**

These preliminary assemblies are very fragmented, as shown by the number of contigs. In order to improve the contiguity of the assembly we can perform a 'velvetg' assembly making use of the paired end information to connect contigs together. We also expect to see many singletons and short contigs in the assembly as a result of sequencing errors. To remove these we performed another 'velvetg' run with custom parameters for expected coverage, coverage cut-off and minimum contig length. The coverage cut-off and expected coverage were determined using the method described in the velvet manual (by plotting the k-mer coverage histogram and using the first minima for the cut-off and the maxima for the expected coverage) as shown in figures 3.1, 3.2 and 3.3:

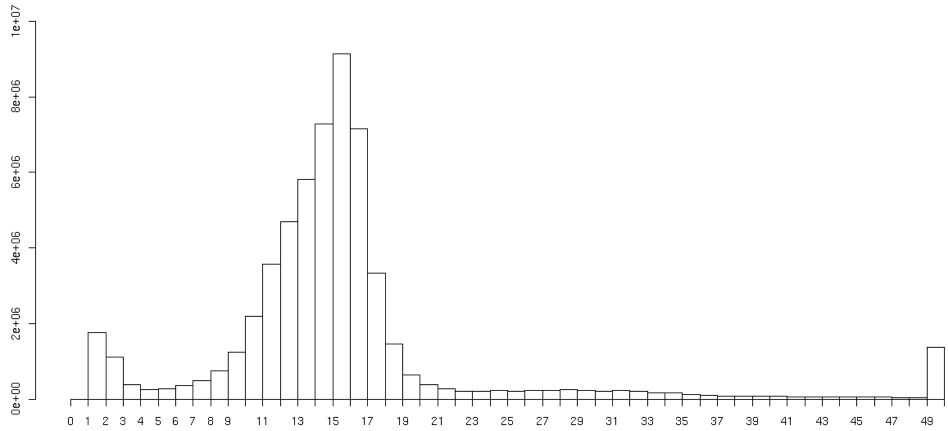



**Figure 3.1: Histogram plot of k-mer coverage in velvetg assembly with a k-mer size of 21. Expected k-mer coverage is 16, and estimated coverage cut-off is 5.**

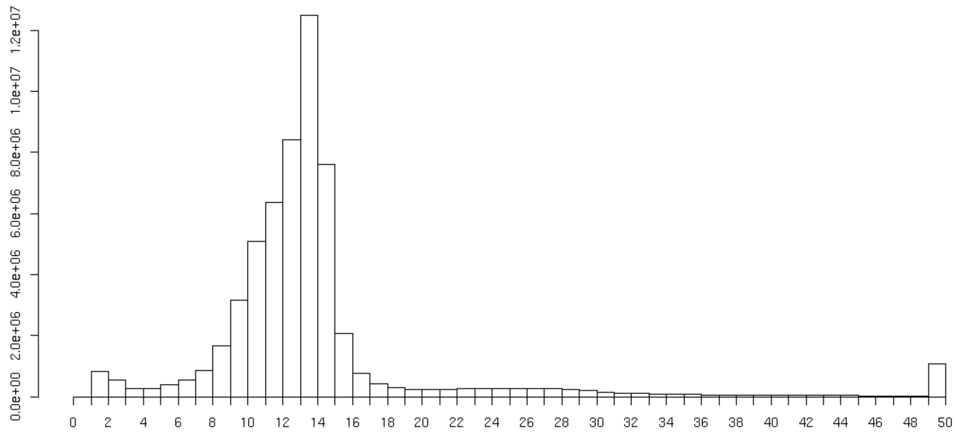



**Figure 3.2: Histogram plot of k-mer coverage in velvetg assembly with k-mer size 23. Expected k-mer coverage is 14, and estimated coverage cut-off is 4.**

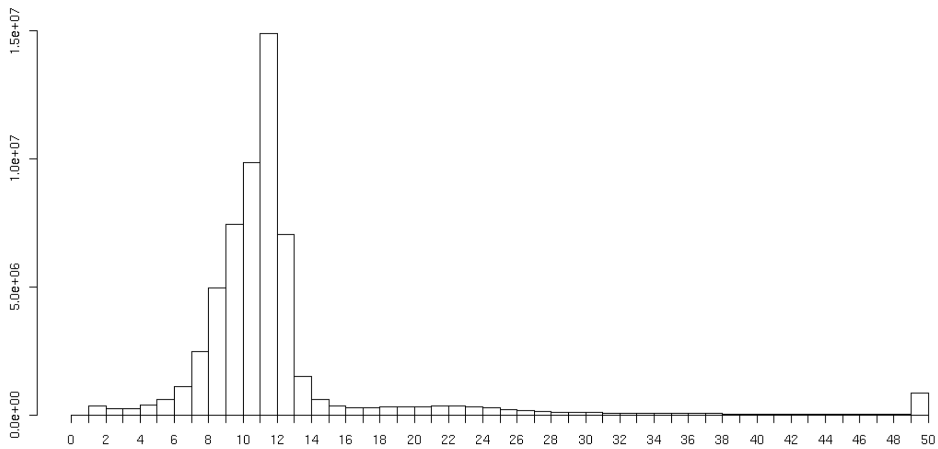



**Figure 3.3: Histogram plot of k-mer coverage in velvetg assembly with a k-mer size of 25. Expected k-mer coverage is 12, and estimated coverage cut-off is 3.**

We then performed a scaffolded assembly using 'velvetg'. We modified the parameters as follows:

- cov_cutoff 5/4/3 (this was based on a histogram of coverage per contig for each k-mer length 21,23,25)
- ins_length 410 (based on 2 x 36 bp reads, ~342 bp apart)
- ins_length_sd 20
- exp_cov 16/14/12 (this was based on a histogram of coverage per contig for each k-mer length 21,23,25)
- min_contig_length 100
- min_pair_count 4

The assembly statistics are presented in table 3.2.

| k-mer length | Number of contigs | Mean contig length | Median contig length | Sum of contigs | Longest contig |
|---|---|---|---|---|---|
| 21 | 19,104 | 2980 | 450 | 56,940,038 | 603,164 |
| 23 | 19,730 | 3028 | 336 | 59,742,995 | 684,975 |
| 25 | 20,744 | 2996 | 272 | 62,153,385 | 596,363 |

**Table 3.2: velvetg run statistics – scaffolding with modified parameters**

At this stage, with increasing k-mer length we see a:

- Reduction in the number of total contigs assembled
- Increase in mean and median contig size
- Increase in the sum of contig lengths
- Increase in size of longest contig

### *3.2.1.4 Evaluating the quality of the Velvet de-novo assemblies*

The assembly statistics of the various k-mer lengths in table 3.2 suggest that increasing the k-mer length for our dataset results in a more contiguous assembly, with more sequence assembled. Optimising the N50 is usually considered to be the optimal strategy to improve a genome assembly. However, these statistics do not reflect on the quality of the sequence assembled. To verify the accuracy of the sequence assembled we decided to see how well 2 known effector genes of *Hpa*, *ATR1* and *ATR13*, were assembled. In both assemblies with k-mer size 21 and 23, *ATR1* and *ATR13* were present at full length with no mismatches or gaps. With the assembly using a k-mer size of 25, *ATR1* assembled perfectly, but interestingly we saw an assembly artefact with the assembly of *ATR13* (fig 3.4).

```
…

Query: 421  ttaggaagataataaactcgcggaagcccatcgaaaccagttattcggctaaaggcatcc 480
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 1557 ttaggaagataataaactcgcggaagcccatcgaaaccagttattcggctaaaggcatcc 1616


Query: 481  acga---------gaagattataaaggcatacgatcgtcatgtcttcgaatctaagaagg 531
            ||||         |||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 1617 acgannnnnnnnngaagattataaaggcatacgatcgtcatgtcttcgaatctaagaagg 1676


Query: 532  cacacgatcgtcatgtctccaaatctaagaaggcacacggtcgtcatgtctccaaatcta 591
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 1677 cacacgatcgtcatgtctccaaatctaagaaggcacacggtcgtcatgtctccaaatcta 1736
```

**Figure 3.4: Extract from BLAST alignment of *ATR13* to Velvet assembly of *Hpa* Emoy2 with a k-mer size of 25. An assembly artefact was observed where a 9 bp insertion of n's is observed.**

We observed this assembly artefact with many preliminary assemblies utilising higher k-mer lengths with the data. This 'insertion artefact' was more frequent and pronounced the larger the k-mer length (data not shown). We believe this was due to a fault in the assembly algorithm during the time of assembly, but we have not re-assembled the data with a newer version of Velvet to verify this claim.

Observation of this artefact meant that we attempted no further assemblies with a k-mer length of 25.

To evaluate the quality of assembly of the Velvet k-mer size 21 and 23 assemblies, we downloaded a dataset of 52 amino acid sequences of known genes in the *Hyaloperonospora* genus from the Genbank database (taxonomic ID 184462, July 2009) (appendix table 3.3). 2 genes (a putative effector protein Avh341 and a MAP kinase) were found fully assembled in the assembly with k-mer length 21 but partially assembled in the assembly with k-mer length 23. There were no genes found that were fully assembled in the assembly with k-mer length 23, and partially assembled in the assembly with k-mer length 21. For this reason, we continued to compare the Velvet assembly of k-mer length 21 to the v7 Sanger assembly, and henceforth refer to the assembly as the 'Velvet assembly'.

### *3.2.1.5 Comparing the Velvet assembly to the v7 assembly*

The Velvet assembly adopted for the rest of this analysis (k-mer length 21) is 56.9 Mb (3.8 Mb N's) over 19,104 scaffolded contigs. The longest mean scaffolded contig length is 2980 bp and the largest scaffolded contig is 603,164 bp. The number of scaffolded contigs larger than 2 kb is 4229, and the N50 is 742.

Although at first glance, comparing the N50 values of the Velvet assembly (742) to the v7 assembly (68) (appendix table 3.1) suggests that the v7 assembly is much more contiguous than the Velvet assembly, it is remarkable that the Velvet assembly contained a residual 55.1 Mb of sequence (calculated by subtracting the N's from the scaffolded assembly) compared to residual 67.5 Mb of sequence in the v7 assembly. It was hypothesised that the 19% difference in the sequences was due the underlying De-Bruijn graphed structure used in the Velvet algorithm not being able to correctly resolve duplicate and repetitive regions.

We used DNAdiff from the EMBOSS package (Rice et al., 2000) to calculate the overlap of the v7 and Velvet assemblies (table 3.3). 94% of Velvet assembly contigs aligned to the v7 assembly with an average identity of 99%. The remaining 6% of the Velvet contigs that did not align to the v7 accounted for 9.5 Mb of sequence. Conversely, 68% of the v7 contigs aligned to the Velvet assembly with an average identity of 99%. The remaining 32% accounted for 33 Mb of sequence.

| | v7 | Velvet |
|---|---|---|
| **[Sequences]** | | |
| **Total** | 1842 | 19,104 |
| **Aligned** | 1260 (68.40%) | 17,959 (94.01%) |
| **Unaligned** | 582 (31.60%) | 1145 (5.99%) |
| | | |
| **[Bases]** | | |
| **Total** | 76,549,095 | 56,940,038 |
| **Aligned** | 43,682,374 (57.06%) | 47,463,168 (83.36%) |
| **Unaligned** | 32,866,721 (42.94%) | 9,476,870 (16.64%) |
| | | |
| **[Alignments]** | | |
| **1-to-1** | 21,338 | 21,338 |
| **Total Length** | 48,214,454 | 48,271,511 |
| **Average Length** | 2259.56 | 2262.23 |
| **Average Identity** | 99.14 | 99.14 |

**Table 3.3: DNAdiff results between the *Hpa* Emoy2 v7 Sanger assembly and the *Hpa* Emoy2 Velvet assembly.**

This data suggest that the majority of the Velvet assembled sequence is present in the v7 assembly. In addition, there may be ~6 Mb of novel sequence not present in the v7 assembly.

### *3.2.1.6 Using CEGMA to compare the core eukaryotic genes in the Velvet and v7 assemblies*

In order to evaluate the assembly of the gene space in each of the genome assemblies, we use the CEGMA pipeline (Parra et al., 2007) to identify the number of single copy core eukaryotic genes (CEGs). This method has been published as a useful metric to describe the assembled gene space (Parra et al., 2009). There are 248 CEGs, which are a subset of the 458 set of core eukaryotic genes (KOGs) conserved across 6 eukaryotic species (*A. thaliana*, *Homo sapiens, Drosophila melanogaster, Caenorhabditis elegans, Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*). We compared the number of predicted CEGs in the v7 and Velvet assemblies to predictions in *Phytophthora infestans, Phytophthora ramorum* and *Phytophthora sojae* (fig 3.6).

Stein et al. (2003) performed a simulation of the effect of sequence coverage in an assembly on percentage of CEGs identified in *Caenorhabditis briggsae*. They show that at

4x coverage, it is possible to identify more than 80% of the CEGs using the CEGMA pipeline. Therefore, given this data we would expect to identify approximately 95% of the CEGs in an assembly with 9.5x sequence coverage such as the *Hpa* Emoy2 v7 assembly (fig 3.5).

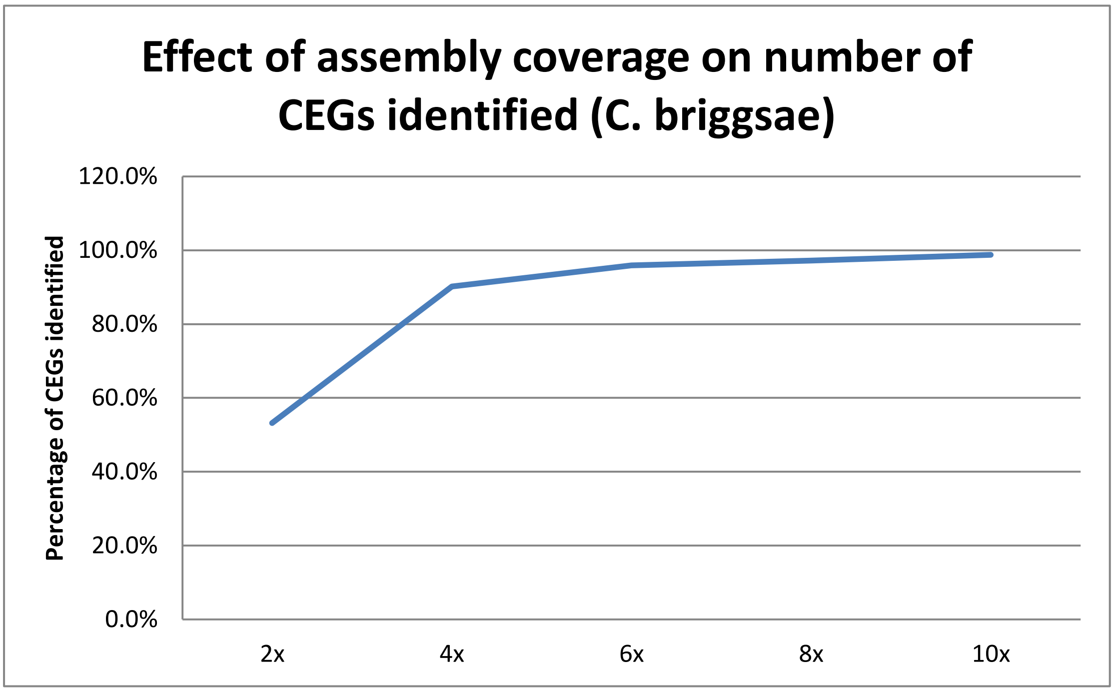



**Figure 3.5: The effect of sequence coverage and CEGs identifiable by the CEGMA pipeline**

The results of the CEGMA predictions are summarised in figure 3.6. 94.8% of the CEGs were identified in the Velvet assembly, which is what we expect based on simulations done by Stein et al. (2003). 89.9% of the CEGs were identifiable in the v7 assembly, which is a little less than expected. The CEGMA pipeline identified 12 more CEGs in the Velvet assembly (235) compared to the v7 assembly (223), which implies that assembly and identification of ~95% of the CEGs is achievable through short read assembly. This also suggests that some of the unique sequence in the Velvet assembly has additional genes not in the v7 assembly, and perhaps vice versa.

The average size of contigs, on which these CEGs were found, were approximately 10 fold higher in the v7 assembly (279,518) compared to the Velvet assembly (25,387), which suggests that although the Velvet assembly contains more of the CEG genes, the v7 assembly is much more contiguous.

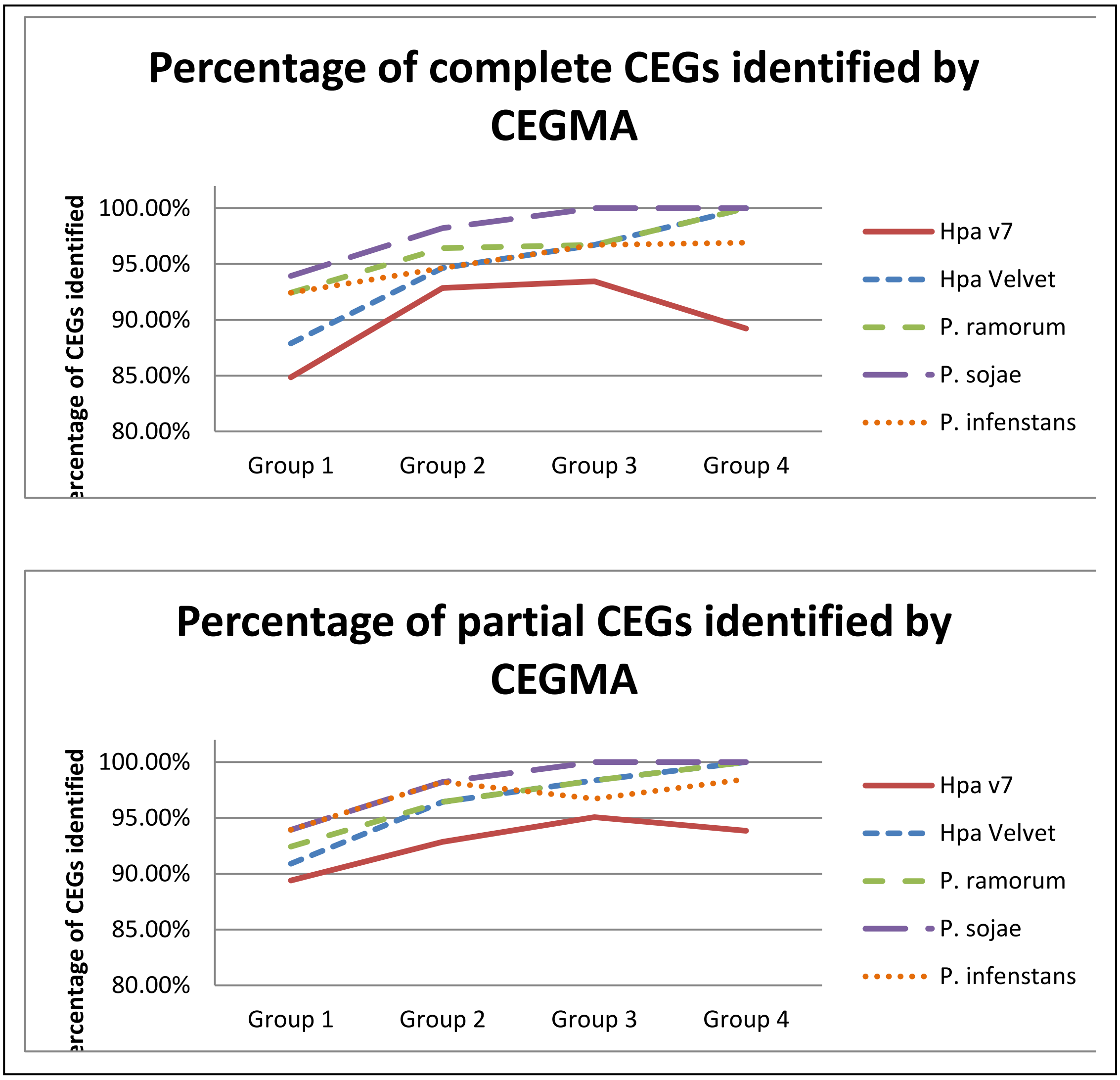



**Figure 3.6: Identification of single copy core eukaryotic orthologous genes (CEGs) by the CEGMA pipeline. Approximately 95% of the CEGs were identified in the *Hpa* Emoy2 Velvet assembly and ~90% in the *Hpa* Emoy2 v7 assembly. This is comparable to the number of CEGs identified in *P. infestans* (95%), *P. ramorum* (96%) and *P. sojae* (98%). The CEGs are split into 4 groups with Group 1 being the least conserved between organisms, and Group 4 being the most conserved.**

### *3.2.1.7 Comparing representation of genomic sequence between the Velvet and v7 assemblies*

In order to evaluate the representation of genomic sequence in each of the v7 and Velvet assemblies, we aligned a single lane of reads and identified the number of reads aligning to the genome (table 3.4). We observed that despite more CEGs being predicted in the Velvet assembly, 11.6% more reads aligned to the v7 assembly. We hypothesise that the extra reads aligning to the v7 assembly are not gene rich and mainly consist of highly repetitive and complex regions, as these are difficult to resolve using the de-Bruijn graph structure used by the Velvet assembly.

| Assembly | Number of reads aligned | Number of reads aligned as pair | Percentage of reads aligned | Percentage of reads aligned as pair |
|---|---|---|---|---|
| **Velvet** | 7,642,044 | 6,242,808 | 78.0% | 81.7% |
| **V7** | 8,881,216 | 8,134,393 | 90.4% | 91.6% |

**Table 3.4: Number and percentage of reads from a single lane (ID69 lane 39,794,370 reads) aligning to the v7 and Velvet assemblies. The reads were aligned using MAQ v0.7.1 with map parameters of n=1 e=60 a=650.**

### *3.2.1.8 Comparing representation expressed sequences between the Velvet and v7 assemblies*

In order to evaluate how much of the gene space is represented in each assembly we compare alignments of expressed sequence tags (ESTs) to each assembly (table 3.5). 31,759 EST sequences generated from *Hpa*. We aligned these to the Velvet and v7 assemblies using BLAT (Kent, 2002) (using parameters minIdentity=80 query=rna) and post filtering using Brian Haas' blat_top_hit.pl to find the best alignment for each EST (table 3.5). 2.6% more ESTs (or 844 ESTs) aligned to the Velvet assembly compared to the v7 assembly. This suggests that the Velvet assembly better represents the genes expressed during the spore stage of the lifecycle of *Hpa*.

| Genome Assembly | Number of ESTs aligned | Percentage of all ESTs |
|---|---|---|
| ***Hpa* Emoy2 v7** | 28,985 | 91.3% |
| ***Hpa* Emoy2 Velvet** | 29,829 | 93.9% |

**Table 3.5: Number of ESTs aligning to the *Hpa* Emoy2 v7 and Velvet assemblies. Total of 31,759 ESTs**

We also aligned 8,549,032 Illumina sequenced 36 bp cDNA reads of *A. thaliana* Ws eds1-1 plants 7 days post inoculation (d.p.i.) infected with *Hpa* Emoy2, to the assemblies using MAQ (Li et al., 2008a) (version 0.7.1, n=3, e=100) (table 3.6). 2.6% more cDNA reads (265,123 reads) aligned to the Velvet assembly compared to the v7 assembly. We come to a similar conclusion as with the ESTs, that the Velvet assembly better represents the genes that are expressed during the 7 d.p.i stage of infection of *Hpa*.

| Genome Assembly | Number of cDNA reads aligned | Percentage of all cDNA reads |
|---|---|---|
| ***Hpa* Emoy2 v7** | 2,145,339 | 25.1% |
| ***Hpa* Emoy2 Velvet** | 2,371,477 | 27.7% |

**Table 3.6: Number of cDNA reads aligning to the *Hpa* Emoy2 v7 and Velvet assemblies. Total of 8,549,486 reads. The reads were aligned using MAQ v0.7.1 with map parameters of n=3 e=100.**

Given that we saw approximately 3% more EST sequences and cDNA reads aligning to the Velvet assembly compared to the v7 assembly, we concluded that integrating the Illumina sequence data into the v7 assembly increases the representation of transcribed regions by approximately 3%. Given this rationale we proceeded with improving the v7 assembly making use of the Illumina reads.

### 3.2.2 Improving the *Hpa* v7 Sanger assembly using Illumina sequenced short reads

In the previous analysis I have shown that the v7 assembly can be improved by integrating additional data into the assembly from the Illumina reads. It also came to light that *Hpa* Emoy2 BAC sequences (provided by HRI, Warwick and sequenced by the Sanger Centre, Cambridge) were not integrated into the genome.

We developed a 4 stage iterative pipeline with which we improved the genome assembly. A flow chart describing the method is shown in figure 3.7

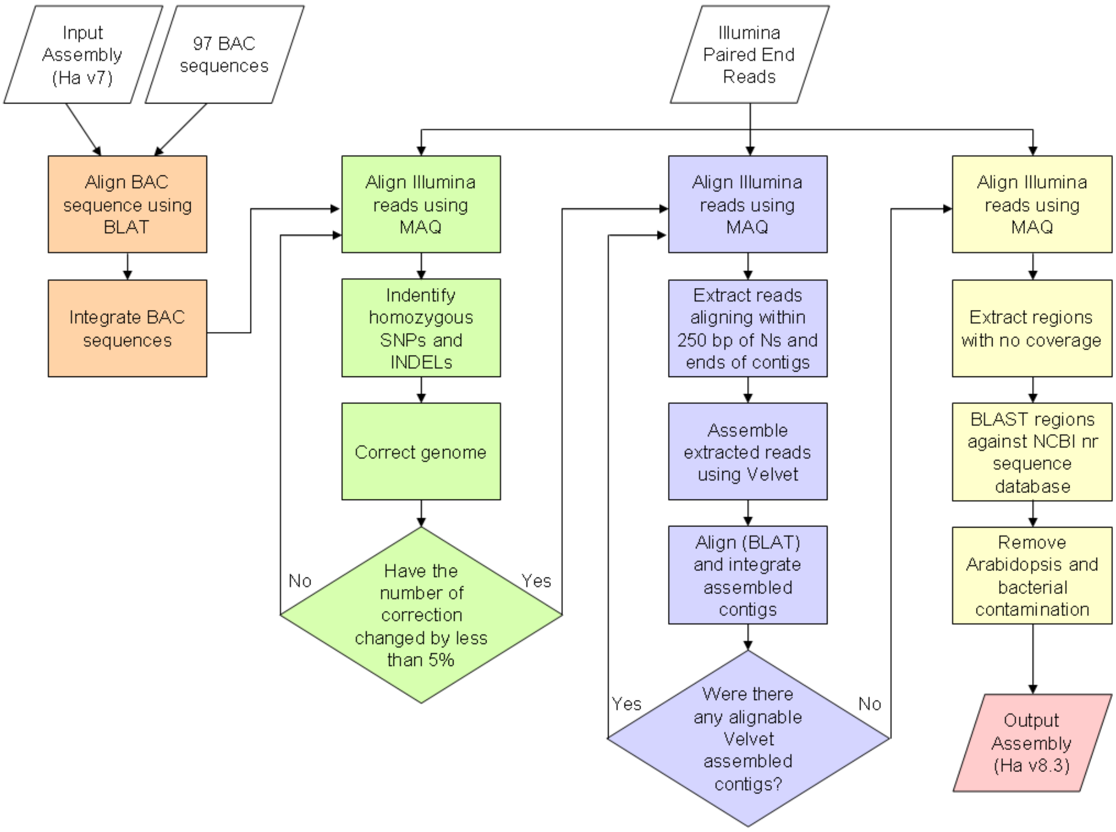



**Figure 3.7: Four stage assembly improvement pipeline for incorporating BAC and Illumina sequencing data. In stage 1 of the improvement pipeline (pale orange) 97 complete BAC sequences were integrated; in stage 2 (green) sequencing errors were corrected using the Illumina paired end reads; in stage 3 (blue) we used methods similar to those described by Ossowski et al. (2008) to integrate Velvet assembled short reads; in stage 4 (yellow) we identified and removed regions not covered by Illumina paired end reads that showed homology to possible contaminants. The resultant assembly was the v8.3 assembly.**

### *3.2.2.1 Merging of full length BAC sequences with the Sanger shotgun assembly*

The 95 BAC sequences were aligned to v7 using BLAT (Kent, 2002). Where the length of the BAC sequence differed from the length of the spanned assembly by less than 1%, the BAC sequence was automatically substituted for that region of the assembly. Where the BAC joined two contigs, the BAC sequence was automatically used to join the contigs if the replaced sequences differed no more than 1% from the BAC sequence. All other matches were reviewed manually. 57 BACs were integrated into the genome. Of these 57 BAC sequences, 28 integrated into existing contigs, 27 extended existing contigs and 2 merged existing contigs. Furthermore, 8 BAC sequences significantly overlapped each other and were merged into 3 larger contigs. 30 BACs were not easily integrated due to a very strict overlap agreement criteria used in order to minimise loss of sequence data through the integration of the BACs. These BAC sequences were appended to the assembly, and should be considered the authoritative assembly of the relevant regions. We calculated that this introduced 2.4 Mb of additional redundancy into the assembly. There were 58 short Sanger assembled contigs, totalling to 321 kb of sequence, entirely contained within full length BAC sequences. These 58 short contigs were removed from the assembly.

### *3.2.2.2 Iterative correction of the Sanger assembled sequence*

After the BAC sequences were integrated into the genome, I developed a novel method to iteratively correct a genome sequence. This method is based on the premise that a reference genome assembly and deep sequenced short read data of the same or similar species are available. Given that in most circumstances, the reference genome assembly has low sequence coverage, this method takes advantage of the deep coverage through, for example, second generation sequencing technologies such as the Illumina GA and Abi SOLiD. This would allow one to identify differences between the reference sequence and the deep sequenced species with reasonable accuracy. If the deep sequenced species is the same species and isolate as the reference sequence, any difference can be regarded as a mistake in the reference sequence. This observation forms the basis of our consensus based method to correct the Sanger genome sequence of *Hpa* Emoy2, as we also have *Hpa* Emoy2 short reads.

The strategy employed to correct sequencing errors consisted of the following steps:

1. Align short DNA reads to the genome using MAQ
    - Allowing for a maximum of 1 mismatch in the first 24 bp (n=1)
    - The maximum sum of qualities of mismatches over the entire read is less than 70 (e=70)
2. Predict the homozygous SNPs and INDELs in the *Hpa* Emoy2 genome (which indicate assembly errors in the *Hpa* Emoy2 genome assembly)
    - The read coverage does not deviate more than 50% from the expected coverage of 24x
    - The aligning bases are a minimum of 80% of a single consensus base for the SNP call
    - SNPs filtered using maq.pl command SNPfilter
        - Parameters d=12 D=36 Q=20 q=20 w=1 F=(predicted INDEL file)
    - INDELs predicted using indelpe command and filtered with minimum depth of 5 and 75% of reads agreeing with the INDEL prediction
    - SNPs do not overlap with predicted INDELs
3. Correct the reference sequence based on the above predicted variation
4. Repeat steps 1-3 until the number of corrections is less than 5%

The only corrections that were not considered were to the BAC sequences, and the location at which BAC sequences had been integrated into the genome.

| Round | SNPs identified | INDELs identified |
|---|---|---|
| **1** | 621 | 1489 |
| **2** | 467 | 253 |
| **3** | 976 | 132 |
| **4** | 388 | 99 |
| **5** | 230 | 87 |
| **6** | 147 | 75 |
| **7** | 399 | 70 |
| **8** | 189 | 62 |
| **9** | 103 | 60 |

**Table 3.7: Number of SNPs and INDELs after each iteration of correction. Corrections were made until the number of either SNPs or INDELs did not change by more than 5%.**

9 rounds of corrections were made before there was less than a 5% change to the number of predicted INDELs (table 3.7). In order to evaluate the effect of each of these rounds of corrections to the genome, we aligned a single lane of reads to the genome and identified the number of reads aligning to the genome. With each iteration the number of reads aligning to the modified genome increased until saturation, which suggests that the method works correctly. There is also a correlation between the number of SNPs predicted in each round and the number of reads aligning to the genome (an increase in the number of predicted SNPs correlates to fewer reads aligning, and a reduction in the number of predicted SNPs correlates with an increase in the number of reads aligning to the assembly). After these corrections the percentage of reads aligning from a single lane of Illumina data improved by 0.5% (table 3.8), which is remarkable given than no additional sequence was added.

| Round | Number of reads aligned | Number of reads aligned as pairs | Percentage of reads aligned | Percentage of reads aligned as pairs |
|---|---|---|---|---|
| 0 | 8,930,642 | 8,228,155 | 91.2% | 92.1% |
| 1 | 8,951,344 | 8,259,467 | 91.4% | 92.3% |
| 2 | 8,956,760 | 8,239,159 | 91.4% | 92.0% |
| 3 | 8,962,188 | 8,254,523 | 91.5% | 92.1% |
| 4 | 8,966,306 | 8,261,374 | 91.5% | 92.1% |
| 5 | 8,971,010 | 8,267,512 | 91.6% | 92.2% |
| 6 | 8,972,906 | 8,267,243 | 91.6% | 92.1% |
| 7 | 8,973,606 | 8,275,088 | 91.6% | 92.2% |
| 8 | 8,975,254 | 8,274,580 | 91.6% | 92.2% |
| 9 | 8,976,732 | 8,272,277 | 91.7% | 92.2% |

**Table 3.8: Number and percentage of reads from a single lane (ID69 lane 39,794,370 reads) aligning to the corrected assembly after X rounds of genome correction. Round 0 denotes the alignment statistics to the genome with the BACs integrated with no rounds of genome sequence correction.**

This method has further potential to be adapted to modify a reference sequence to represent the sequence of a closely related organism if SNP and INDEL prediction methods are modified, and additional resequencing artefacts are considered to include elevated coverage, represented possible copy number variation (CNV), or read pairs being aligned outside of the expected distribution of read pair insert sizes, indicating larger structural variation.

After the *Hpa* genome sequence was published and the method used to improve the genome sequence using the Illumina reads was described, a program working on the same principles of this method was published. This program is called 'iCorn' (iterative correction of reference nucleotides) and is described by Otto et al. (2010).

### *3.2.2.3 Integration of the Sanger and Illumina assemblies*

After the 9 rounds of iterative sequence correction, the next stage in the pipeline was to integrate novel sequence from the Illumina reads using a targeted assembly method. The method we used was adapted from Ossowski et al. (2008).

In order to integrate assembled Illumina reads into the genome, we first identified the reads that were novel and had potential to be integrated. These novel reads were identified by extracting all the reads that did not align to the reference assembly (using MAQ, default parameters). Since the assembly of these reads would result in novel sequence (based on how they were extracted), we had to develop a way of integrating the sequence into the genome. We hypothesised that the novel sequence would integrate into regions where there is unknown sequence in the reference assembly. The regions where the sequence would be unknown include regions of 'N's, sequence after the end of a contig and the sequence before the start of a contig. I therefore extracted reads, and their other pair, that aligned within 250 bp from regions of N's, and from the start and end of contigs. This additional inclusion of reads aligning to 250 bp from the fore mentioned regions will allow for overlap based integration of the subsequent generated contig assemblies.

A total of 14,148,204 reads were extracted, and assembled into 4234 contigs, with a mean contig length of 966 bp using Velvet (v 0.7.51, k-mer length 23). The assembly totalled to 4.1 Mb with the longest contig being 42 kb.

This assembly was then aligned back to the corrected Sanger assembly using BLAT to identify regions where the Velvet assembled sequence could be integrated into the genome sequence (minimum overlap of 40 bp). 313 Velvet contigs (totalling to ~500kb of sequence) were identified that showed significant matches to the reference sequence. Of the 313 contigs that were integrated, 148 integrated onto the ends of existing Sanger contigs. Of these, one Velvet contig connected 2 Sanger contigs. The remaining 165 integrated Velvet contigs integrated into the middle of contigs (over regions of Ns). The mean size of integrated contigs was 1.5kb, with the longest being 18.8kb.

The remaining Velvet sequence would need to be appended to the genome as additional sequence, but since the current Velvet assembled data contains reads from potential overlap locations a reassembly of reads that do not align to the genome is required. After these Velvet contigs were integrated into the assembly, we realigned (using the previous alignment protocol) all the Illumina paired end reads to the genome sequence and extracted the reads that did not align, and their pair. These reads were assembled using Velvet (k-mer length of 23) and 1468 contigs greater than 250 bp, that were unlikely to be contamination (determined by BLAST search against *A. thaliana*, the NCBI Bacterial genomes July 2009, and the human genome sequence) were appended to the *Hpa* genome sequence. The mean length of appended Velvet contigs was 1.4 kb, and the longest was 42.6kb.

The integration of the additional Illumina sequence improved the alignment of a single lane of Illumina reads to the assembled genome by 2.2%, which is the largest percentage increase compared to the previous modifications made (integrating the BAC sequences and correcting the sequencing assembly error) (table 3.9).

| **Genome** | **Number of reads aligned** | **Number of reads aligned as pair** | **Percentage of reads aligned** | **Percentage of reads aligned as pair** |
|---|---:|---:|---:|---:|
| **V7** | 8,881,216 | 8,134,394 | 90.7% | 91.6% |
| **Velvet** | 8,930,642 | 8,228,155 | 91.2% | 92.1% |
| **V7 + BACs** | 8,956,760 | 8,239,159 | 91.4% | 92.0% |
| **V7 + BACs (corrected)** | 8,976,732 | 8,272,277 | 91.7% | 92.2% |
| **V7 + BACS + Velvet** | 9,169,454 | 8,415,542 | 93.6% | 91.8% |
| **V7 + BACS + Velvet (filtered for contamination)** | 9,167,148 | 8,405,118 | 93.6% | 91.7% |

**Table 3.9: Number and percentage of reads from a single lane (ID69 lane 39,794,370 reads) aligning to the corrected assembly after genome modifications.**

### *3.2.2.4 Identifying and removing contamination using Illumina sequencing*

The final stage of the *Hpa* genome improvement pipeline was to identify and remove contamination in the genome. We initially identified regions of potential contamination as regions that were not covered by any Illumina reads when reads were aligned to the genome. The basis of the hypothesis that contamination in the genome assembly would not be covered by Illumina reads is that since the same organism was sequenced using Illumina and Sanger sequencing, all regions of the genome should be covered by Illumina reads. Any regions in the genome that were not covered by Illumina reads would either be due to contamination in the Sanger reads, or technical limitations of the protocol employed for the Illumina sequencing.

We aligned 6 lanes of Illumina paired end sequence against the *Hpa* genome and extracted 3360 regions with less than 3x coverage over 100 bp. We performed a preliminary BLAST search against the nucleotide NR database (August 2009, using blastn, no sequence filtering and a minimum e-value of 1 x 10-6). We observed that the majority of best hits of uncovered regions in the genome were to bacterial sequences (table 3.10). There were only 4 hits to *A. thaliana*, and no hits to cloning vectors as these were identified as contamination in the v6 assembly (by myself, using this same method) and rectified in the v7 assembly.

| **Best BLAST hit** | **Number of hits** |
|---|---|
| *Methylobacillus flagellatus* KT | 241 |
| *Xanthomonas campestris* pv. campestris | 122 |
| *Ricinus communis* | 78 |
| *Xanthomonas oryzae* pv. oryzicola BLS256 | 66 |
| *Xanthomonas axonopodis* pv. citri str. 306 | 58 |
| *Xanthomonas campestris* pv. vesicatoria str. 85-10 | 57 |
| *Xanthomonas oryzae* pv. oryzae PXO99A | 52 |
| *Flavobacterium johnsoniae* UW101 | 45 |
| *Sphingomonas* sp. SKA58 | 42 |
| *Chryseobacterium gleum* ATCC 35910 | 25 |
| *Xanthomonas campestris* pv. campestris str. 8004 | 23 |
| *Dechloromonas aromatica* RCB | 19 |
| *Acidovorax delafieldii* 2AN | 18 |
| *Xanthomonas oryzae* pv. oryzae KACC10331 | 18 |
| *Novosphingobium aromaticivorans* DSM 12444 | 17 |
| *Methylophilales bacterium* HTCC2181 | 16 |
| *beta proteobacterium* KB13 | 14 |
| *Cellvibrio japonicus* Ueda107 | 14 |
| *Janthinobacterium sp.* Marseille | 12 |
| *Pseudomonas fluorescens* Pf0-1 | 12 |
| *Chitinophaga pinensis* DSM 2588 | 10 |
| *Pseudomonas putida* W619 | 10 |

**Table 3.10: Best BLAST hits when extended regions of very low coverage were blasted against the NR database. Only results with at least 10 hits are shown. The majority of the best BLAST hits were from bacteria.**

We then proceeded to download the nucleotide databases from NCBI for human, *A. thaliana* and all bacteria (August 2009). We performed a BLAST search (using blastn and e-value cut-off of 1 x 10-10) against each of the sequence databases. Each hit was filtered to share at least 95% sequence identity over the match. Each potential contaminant region with a hit was then compared to the NR best BLAST hit to see if both hits are from the same phylogenetic kingdom.

We did not find any significant matches to human sequences. We found 261 contigs in 859 hits to the NCBI bacterial sequences from over 140 bacterial species. The total sum of bacterial contamination was 185 kb, with the main source of contamination (87 kb) being from 2 BAC sequences (Cu694975 and Cu694660, both of which originated from *Xanthamonas campestris* pv. campestris) (appendix table 3.5 & 3.6). We found 7 contigs in 13 hits to the *A. thaliana* TAIR 9 genome assembly. The total sum of *A. thaliana* contamination was 3 kb (appendix table 3.7).

After identification of potential known contamination, each of the regions of very low coverage with a hit to a contaminant was removed. A total of 85 contigs were removed due to having a >85% of the bases uncovered and having significant contamination, and a total of 119 contigs were split in order to remove contaminant regions.

This was the last stage of genome improvement utilising the Illumina reads. The resulting assembly was the v8.3 assembly.

### 3.2.3 Evaluating the v8.3 hybrid Sanger and Illumina assembly

The 82 Mb assembly consisted of 1783 major scaffolds (larger than 2 kb) with an N50 scaffold number of 75. The assembly consisted of 8.3 Mb of 'N's, which is 700 kb fewer than the v7 assembly. The reduction in the number of N's is due to the integration of the BAC sequences and Velvet assembled Illumina reads into regions of the v7 genome assembly, which had previously contained N's.

#### *3.2.3.1 Estimating genome size using read coverage*

The total length of the v8.3 assembly was 82 Mb, consisting of 73.7 Mb of "non-N" sequence. To independently estimate the total genome size, we conducted statistical analyses of the coverage provided by the Illumina reads and by the Sanger reads.

To estimate the genome size from the coverage provided by the Illumina reads, we used MAQ (v0.7.1 using default parameters) to align 2,393,125,128 bp of sequence from Illumina paired-end reads (66,475,698 total reads from six lanes of paired end sequencing) to the *Hpa* Emoy2 v7 Sanger read contig models (unscaffolded). The Illumina read coverage at each nucleotide position (67,509,127 positions) was calculated and the frequency of positions with each level of coverage was plotted (fig 3.8). A Gaussian curve was fitted to the main peak of the distribution by least squares and used to obtain the mean of the distribution (23.93).

To estimate the genome size (C) we used the following formula:

$$C = A * R / c$$

where $A$ = number of aligned reads

$R$ = read length

$c$ = average coverage per nucleotide

The genome size was estimated by dividing the total length of the Illumina reads by the mean coverage: 66475698 * 36 / 23.93 = 100.0 Mb.

To estimate the genome size from the unassembled Sanger read data, the coverage of every read was calculated from a blastn all-versus-all search of the trimmed random shotgun reads. For each read we counted the number of matches with > 95% identify over the length of the match, with a minimum overlap of 30 nt. The average coverage at the nucleotide positions defined by each read can then be obtained from the following formula:

$$C = 1 + A * R / [(R + L) - 2 * O]$$

where $A$ = number of aligned reads

$R$ = average length of all the trimmed trace files (720 nt)

$L$ = length of each individual query sequence

$O$ = minimum overlap required to call a match (30 nt)

The frequencies of reads with different coverages were then binned and plotted to identify a peak corresponding to the single copy sequences (fig 3.8). A Gaussian curve was fitted to the single copy peak by least squares and used to obtain the mean of the distribution (8.39). The genome size was estimated by dividing the total length of the Sanger reads by the mean coverage: 1140851 * 720 /8.39 = 97.9 Mb.

The close agreement of the two statistical estimates suggests that the actual genome size (mean of the two estimates = 99 Mb) is significantly larger than the assembled length of 82 Mb. An explanation for this discrepancy is suggested by the second prominent peak in figure 3.8. The presence of this peak suggests that there are a large number of sequences in the genome that are more than 95% identical and have an average copy number of around 3. Such sequences would most likely be assembled as single copy sequences by the assembly software. Plotting the sequence coverage provided by the Sanger reads against the assembled genome did not reveal any contigs or long segments of the assembly with elevated coverage (not shown), ruling out the presence of large triplicated regions or chromosomes. The reads with elevated copy number also did not correspond to contaminants such as bacteria or *Arabidopsis*. The reads with elevated copy number did not correspond to gene models from *Hpa*, suggesting that the repeats were largely confined to non-genic regions. The plot of the Illumina read coverage did not identify a sharp peak of triplicated sequences, but rather a long tail corresponding to high copy coverage. The different shapes of the two plots likely result from the fact that the Illumina reads were much shorter than the Sanger reads, and a perfect match for alignment was required for the Illumina reads, compared to a 95% match for the Sanger reads.

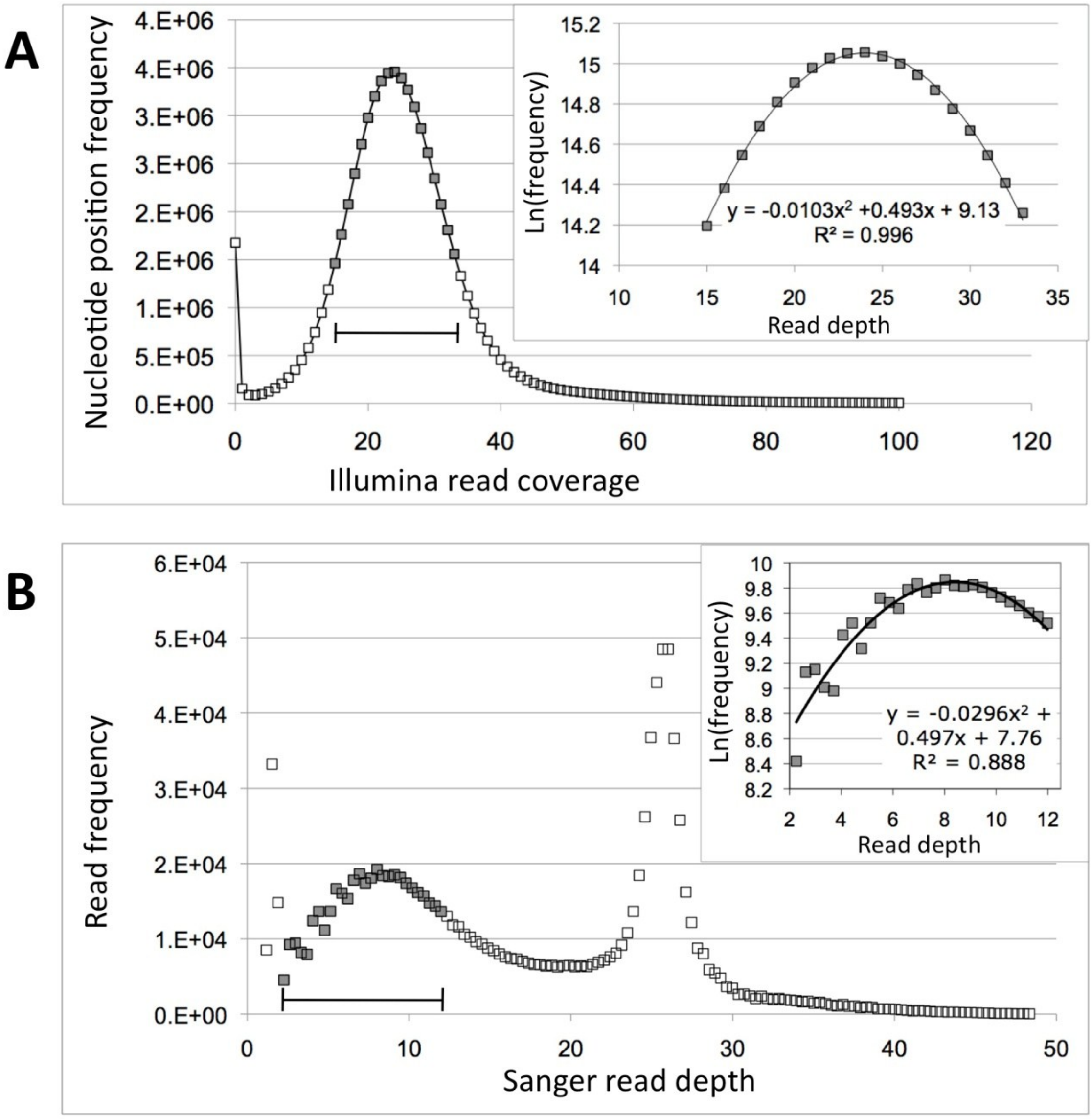



**Figure 3.8: Genome size estimation from Illumina and Sanger read coverage**

**(A) Frequency of nucleotide positions in the Sanger assembly with given Illumina read coverage. (B) Frequency of Sanger reads with given Sanger read coverage. In both (A) and (B), to obtain the mean coverage of the single copy sequences, a Gaussian curve was fitted to the main peak (indicated by shaded points and horizontal bar) by fitting a quadratic function to the natural-log-transformed frequency data (inset). From each fitted quadratic function $ax^2 + bx + c$, the mean of each Gaussian distribution was obtained as $-b/2a$.**

### *3.2.3.2 Comparing the Velvet and v8.3 assemblies*

We used DNAdiff (Kurtz et al., 2004) to compare the Velvet assembly to the v8.3. 98% of the Velvet assembly aligns to the current v8.3 assembly (table 3.11; fig 3.9). There are 14 scaffolds >2 kb that are unique to the Velvet assembly of which the largest is 6.2 kb. 70% of the v8.3 assembly aligns to the Velvet assembly. There are 422 scaffolds >2 kb that are unique to the v8.3 assembly, of which the longest is 12.6 kb. The difference between the size of the Velvet assembly and the v8.3 assembly is due to 27.2 Mb of sequence being collapsed into 5.9 Mb of scaffold in the Velvet assembly and 2 Mb of 'N's captured through the larger Sanger paired end reads.

| | **V8.3** | **Velvet** |
|---|---|---|
| **[Sequences]** | | |
| **Total** | 3138 | 19,104 |
| **Aligned** | 2182 (69.53%) | 18,642 (97.58%) |
| **Unaligned** | 956 (30.47%) | 462 (2.42%) |
| | | |
| **[Bases]** | | |
| **Total** | 82,051,642 | 56,940,038 |
| **Aligned** | 46,419,401 (56.57%) | 50,095,244 (87.98%) |
| **Unaligned** | 35,632,241 (43.43%) | 6,844,794 (12.02%) |
| | | |
| **[Alignments]** | | |
| **1-to-1** | 22,750 | 22,750 |
| **Total Length** | 50,990,271 | 51,040,070 |
| **Average Length** | 2241.33 | 2243.52 |
| **Average Identity** | 99.13 | 99.13 |

**Table 3.11: DNAdiff results between the *Hpa* Emoy2 v8.3 hybrid assembly and the *Hpa* Emoy2 Velvet assembly.**

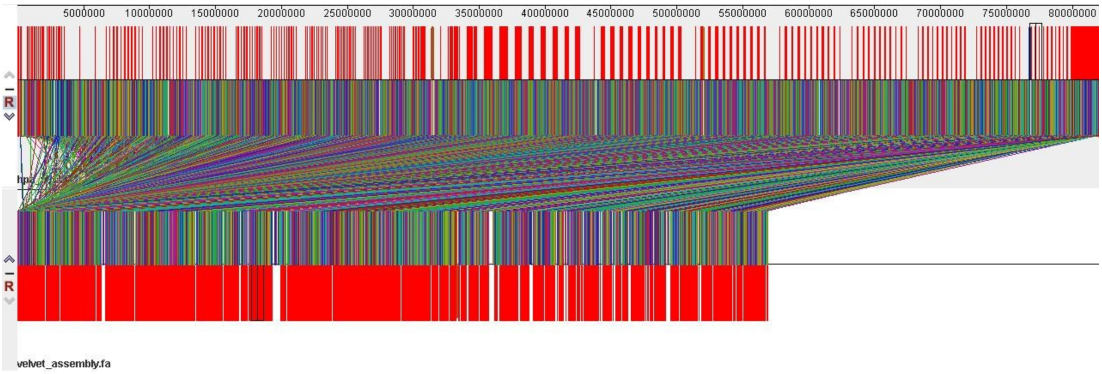



**Figure 3.9: Visual representation of an alignment of the *Hpa* Emoy2 v8.3 assembly against the Velvet assembly. Produced using Mauve (Darling et al., 2010).**

### *3.2.3.3 Identifying the number of CEGs in the v8.3 assembly*

236 CEGs were identified in the v8.3 assembly (fig 3.10). This represents a 6% increase in the number of CEGs identified compared to the v7 assembly in which 223 were predicted. This is an improvement over the v7 assembly. One CEG was predicted in the Velvet assembly that was not in the CEGMA predictions for the v8.3 assembly. The sequence of this CEG was extracted and aligned to the v8.3 assembly. The gene prediction was present in full length with no differences, but was omitted from the CEGMA predictions for unknown reasons.

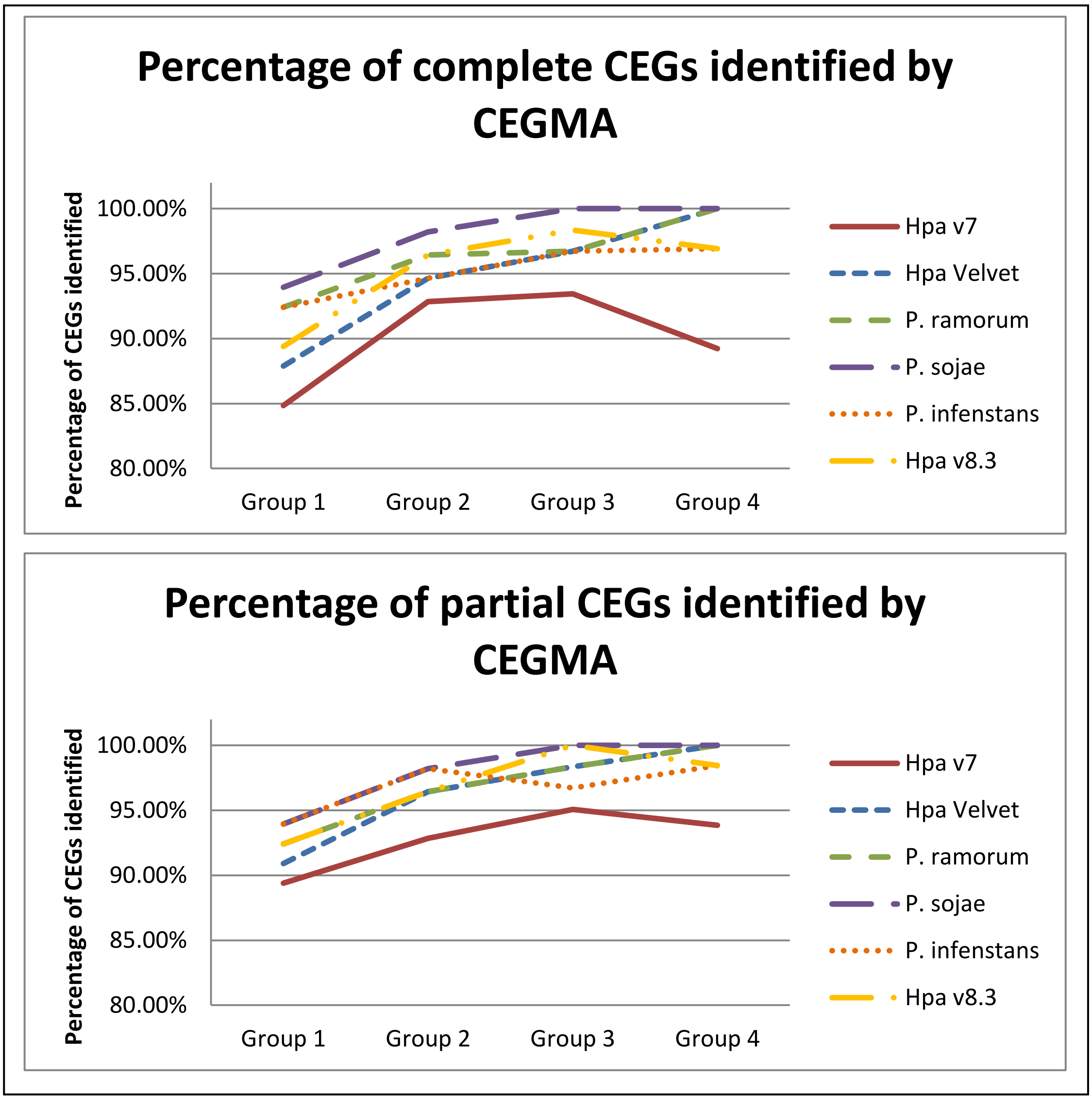



**Figure 3.10: Percentage of full and partial CEGs identified by the CEGMA pipeline in the v8.3 assembly. 95% of the CEGs were identified in the *Hpa* Emoy2 Velvet and v8.3 assemblies.**

### *3.2.3.4 Representation of genomic sequence in the v8.3 assembly*

In order to evaluate the representation of genomic sequence in each of the v8.3 assembly, we aligned a single lane of reads and identified the number of reads aligning to the genome and compared this to previous results of the v7 and Velvet assemblies (table 3.12). We observe a 3.2% increase in the reads that align to the v8.3 assembly compared to the v7 assembly. Given that we expect the genome size to be ~100 Mb, this increase in the number of reads aligning to the v8.3 assembly should correlate to an increase of 3.2 Mb of unique sequence introduced through integrating the BAC and Illumina sequences.

| Assembly | Number of reads aligned | Number of reads aligned as pair | Percentage of reads aligned | Percentage of reads aligned as pair |
|---|---|---|---|---|
| **Velvet** | 7,642,044 | 6,242,808 | 78.0% | 81.7% |
| **v7** | 8,881,216 | 8,134,393 | 90.4% | 91.6% |
| **v8.3** | 9,167,148 | 8,405,118 | 93.6% | 91.7% |

**Table 3.12: Number and percentage of reads from a single lane (ID69 lane 39,794,370 reads) aligning to the v7, v8.3 and Velvet assemblies. The reads were aligned using MAQ v0.7.1 using map parameters of n=1 e=60 a=650.**

### *3.2.3.5 Representation of expressed sequence in the v8.3 assembly*

We aligned the Sanger sequenced ESTs to the Velvet and v8.3 assemblies using BLAT (using parameters minIdentity=80 query=rna) and post filtering using Brian Haas' blat_top_hit.pl to find the best alignment for each EST. We observed a 3.6% improvement (851 ESTs) over the v7 assembly in the number of ESTs aligning to the genome assembly (table 3.13). Although the number of aligning ESTs did not differ significantly between the v8.3 and the Velvet assembly, we do capture the ESTs, and the genes to which they belong, in the v8.3 assembly in a much more contiguous genome space.

| Genome Assembly | Number of aligned ESTs | Percentage of all ESTs |
|---|---|---|
| ***Hpa* Emoy2 v7** | 28,985 | 91.3% |
| ***Hpa* Emoy2 Velvet** | 29,829 | 93.9% |
| ***Hpa* Emoy2 v8.3** | 29,836 | 93.9% |

**Table 3.13: Number of ESTs aligning to the *Hpa* Emoy2 v7 and Velvet assemblies based on a total of 31,759 ESTs.**

We also aligned the Illumina sequenced 36 bp cDNA reads to the v8.3 assembly using MAQ (table 3.14). Unlike the EST alignments, we saw increases in the number of cDNA reads aligning to the v8.3 assembly when compared to the v7 (3%) and the Velvet assembly (0.4%). We believe the difference in the increase of Illumina cDNA reads aligning to the v8.3 compared to the increase in ESTs aligning to the v8.3 assembly is because the ESTs were isolated from spores whereas the Illumina cDNA was isolated from infected plant tissue, and we would expect that different genes are expressed during these different developmental stages.

| Genome Assembly | Number of cDNA reads aligned | Percentage of all cDNA reads |
|---|---|---|
| ***Hpa* Emoy2 v7** | 2,145,339 | 25.1% |
| ***Hpa* Emoy2 Velvet** | 2,371,477 | 27.7% |
| ***Hpa* Emoy2 v8.3** | 2,397,839 | 28.1% |

**Table 3.14: Number of cDNA reads aligning to the *Hpa* Emoy2 v7 and Velvet assemblies based on a total of 8,549,486 reads. The reads were aligned using MAQ v0.7.1 using map parameters of n=3 e=100. The reason for the low percentage of cDNA read alignment to the *Hpa* assemblies is because approximately 60% of the cDNA originate from the host *A. thaliana* (this was determined through aligning the cDNA to the TAIR9 genome assembly), and MAQ is unable to align split reads over splice sites.**

### 3.2.4 Heterozygosity in *Hpa* Emoy2

The majority of genome sequencing project attempt to decipher the genome sequence of an organism. However, in diploid organisms that are not inbred differences between parental chromosomes may be a rich source of genome variation. This heterozygous variation is often not considered. I will describe how the Illumina reads were used to identify heterozygosity in *Hpa* Emoy2, and characterised the heterozygosity in all genes and effector genes (from a draft gene model prediction and annotation). The protocols used are described in Baxter et al., 2010.

#### *3.2.4.1 Identifying heterozygosity in Hpa Emoy2*

MAQ (Li et al., 2008) was used to align the paired-end Illumina reads to the v8.3 assembly. MAQ was used to predicted 59,358 high confidence SNPs (minimum of 10x nucleotide coverage by Illumina reads over the SNP call and a predicted SNP call accuracy of >99%). Of these, 8201 SNPs had a coverage of >80x. It is believe that these predicted SNPs are on regions of collapsed repeats, thus displaying higher than average coverage by Illumina reads, and are indistinguishable from real SNPs and mutations on different duplicated regions of the genome using this method. Furthermore, we observe that 99% of these predicted SNPs are heterozygous. The 1% of SNP calls that were homozygous is a combination of the SNP call error rate and errors in the genome sequence of *Hpa.*

### *3.2.4.2 Heterozygosity in genes and effectors*

It has been shown that *Hpa* effectors, ATR1 and ATR13, are highly polymorphic (Rehmany et al., 2005; Allen et al., 2004). We observe that the frequency of observation of a heterozygous SNP across candidate (1 per ~500bp) is five times higher than in other genes (1 per ~2500 bp) (fig 3.11). Under the neutral theory of evolution, one can explain that the rate of heterozygosity in genes being lower than the in the background rate observed in non coding regions as the majority of mutation are likely to be deleterious and would be selected against. However, the increased rate of heterozygosity in candidate effectors is near 5 times more than in genes, and more than twice as much as the genome, suggesting that there is selection for variation in candidate effectors.

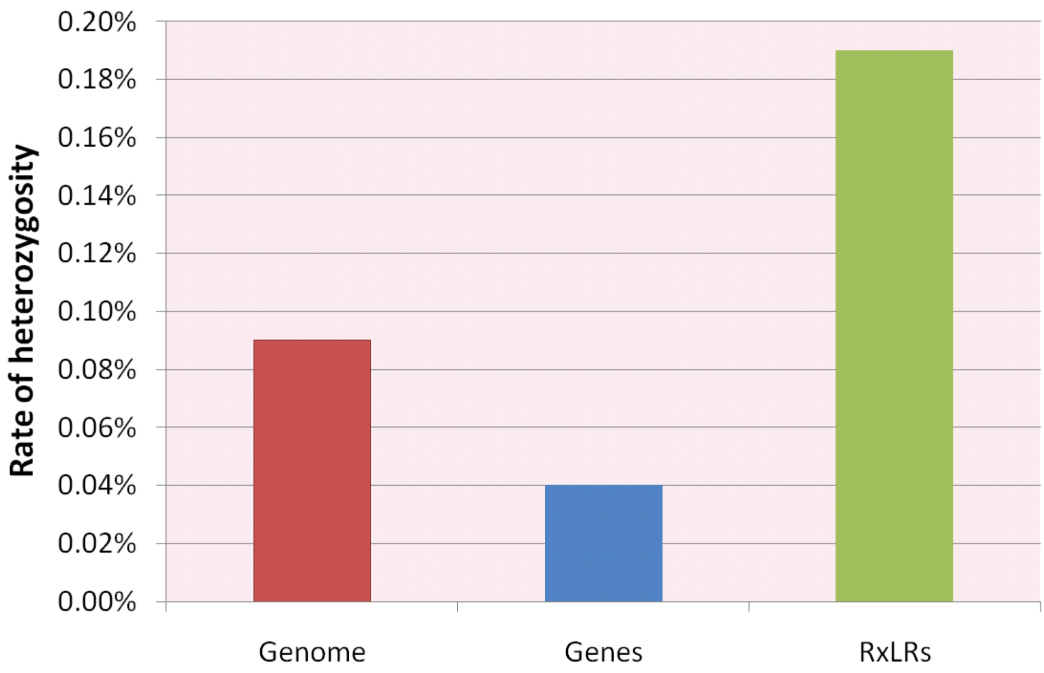



**Figure 3.11: Rate of heterozygosity in the *Hpa* genome, genes and RXLR effector candidates.**

## 3.3 Summary

We were able to show that integration of short read and long read sequences can lead to a better genome assembly by integrating Illumina short reads into the Sanger assembly of *Hpa* Emoy2. We developed several novel methods for evaluating genome completeness, genome correction, novel sequence integration and to identify contamination.

We improved the *Hpa* v7 assembly so that it contains an additional ~4 Mb (5%) of sequence, which should be relatively gene rich. We determined this expected gene richness by the increase in percentage of predicted CEGs (6%), percentage of Illumina sequence cDNA aligning (12%) and the percentage of ESTs aligning (3%) to the v8.3 hybrid assembly compared to the v8.3.

We also showed that Illumina short read assemblies of eukaryotic organisms with genomes of 100 Mb and less can be an efficient way of representing the gene encoding regions of an organism at the expensive of genome contiguity.

The resultant v8.3 assembly is a better representation of the true *Hpa* Emoy2 genome than either the v7 or Velvet assembly, and will thus support more reliable gene model predictions allowing us to better understand the biological functions of *Hpa* Emoy2.

We show that analysis of heterozygosity captures a subset of the variation in a larger population. The implication of this is that genomics experiments using second generation sequencing technologies can make use of heterozygosity to add to observed allelic variation. Analysis of heterozygous SNPs over the genome, genes and candidate effectors reveals that the rate of heterozygosity in effectors in almost 5 times higher than observed in other genes, suggesting that there is selection for maintaining variation in effectors.

# Chapter 4 – *Use of Illumina sequencing to evaluate and improve the Hpa gene models*

## 4.1 Introduction

In the previous chapter I described the establishment of the *Hpa* v8.3 genome assembly. The purpose of the genome assembly is to allow us to generate gene models from which we can identify gene families, predict gene function and ultimately make useful inferences about the underlying biology and chemistry of the organism of interest. In the case of *Hpa* it will increase our understanding of the genes involved in pathogenesis and may reveal clues to the obligate biotrophic nature of the pathogen.

I used a number of software packages to generate gene models and various data sets that provide evidence for gene expression:

- Sanger sequenced ESTs
- Illumina sequenced cDNAs
- 454 sequenced ESTs

I also annotated the resultant gene models and performed comparative analysis with other oomycete pathogens to further understand *Hpa* biology and pathogenicity.

## 4.2 Results and Discussion

### 4.2.1 Existing gene models

The existing gene models were compiled by the postdoctoral researchers Dr Sucheta Tripathy, (VBI, Virginia Tech, Virginia) and Dr Laura Baxter, (HRI, Warwick). I will describe briefly the rationale of design choices and methodologies employed to make the initial set of gene models.

#### *4.2.1.1 Determining the number of genes*

A whole genome BLAST search against the NR protein database was conducted (Oct 2009). The entire genome matched with 14,688 proteins from the NR database with >65% sequence identity over 100 bases. These matches were not inspected for evidence of pseudogenisation, which would have resulted in fewer true matches to existing genes. This analysis may not have shown all the *Hpa* specific genes, since no previous gene models for any *Hyaloperonospora* had been published. From this it is expected that the number of genes in *Hpa* will be between 14-15,000 genes.

#### *4.2.1.2 Genezilla*

Genezilla (Majoros et al., 2005), is an ab initio gene predictor that is based on a Generalised Hidden Markov Model (GHMM) (a GHMM models a continuous state space, as opposed to a HMM which models a discrete state space). Genezilla was chosen by Dr Sucheta Tripathy as the primary gene prediction method due to:

- Genezilla has a state transition model that enables it to consider different types of exons (i.e. initial, internal, final and single exons) using different content sensors (fig 4.1). This was one of the novel features of Genezilla compared to existing gene prediction algorithms
- independent training of exon and intron prediction
    - training using ESTs was done for prediction of exons and splice junctions
    - training using 16,999 'Illumina segments' (regions where 4 or more Illumina 36 bp cDNA reads align at locations at least 300 bp apart)

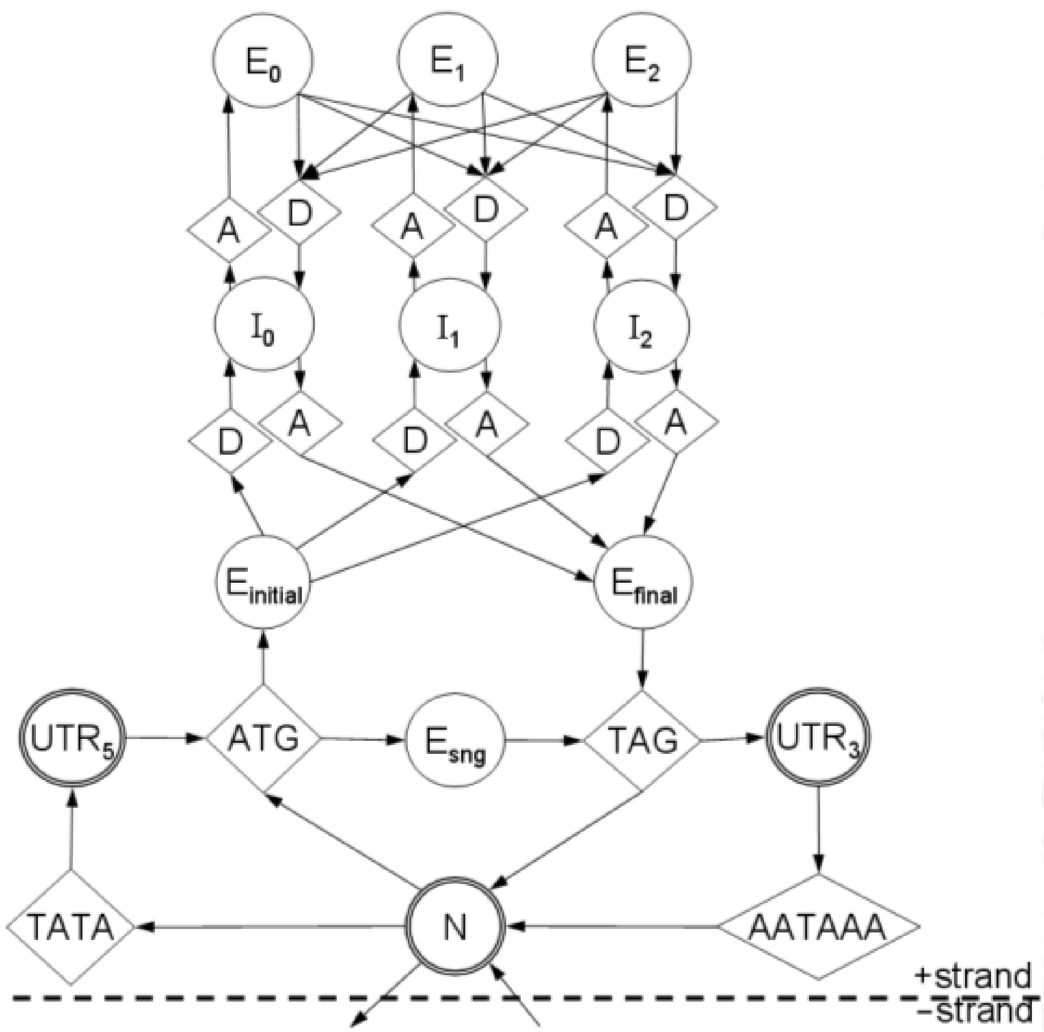



**Figure 4.1: Genezillas state transition model [reproduced from (Majoros et al., 2005)]. Each state of the HMM is represented by a shape and transitions between the states are represented by arrows. States include N: intergenic, $E_{sng}$: single-exon gene, $E_{initial}$: initial exon, $E_{final}$: terminal exon, $E_0$–$E_2$: exons in phase 0–2, $I_0$–$I_2$: introns in phases 0–2, A+D: acceptor and donor sites, TATA: transcription initiation site, AATAAA: transcription termination sites, $UTR_5$+$UTR_3$: 5' and 3' UTR regions, ATG+TAG: start and stop codon.**

3 different protocols were used for Genezilla in order to determine the best protocol. The mean intergenic length parameter was modified. It was reported that the mean intergenic length parameter need not match the actual mean length of intergenic regions in this organism; values quite different from the mean could give better prediction accuracy than the true mean, due to the dependencies between different parts of the underlying model. Increasing the mean intergenic length and the exon length parameters were reported to have also resulted in producing gene predictions with better EST support. A total of 16,166 genes predicted by Genezilla were used.

#### *4.2.1.3 Snap*

Snap (Korf, 2004), is a gene predictor that is based on a GHMM (fig 4.2). Snap was chosen to generate gene models to verify the Genezilla predictions and to provide alternative gene calls.

Snap's species-specific parameter estimation was performed using a training set derived from 100 manually curated *Hpa* genes with full-length EST support, and a HMM was built from these parameters. A total of 687 Snap gene predictions were used to complement the Genezilla models.

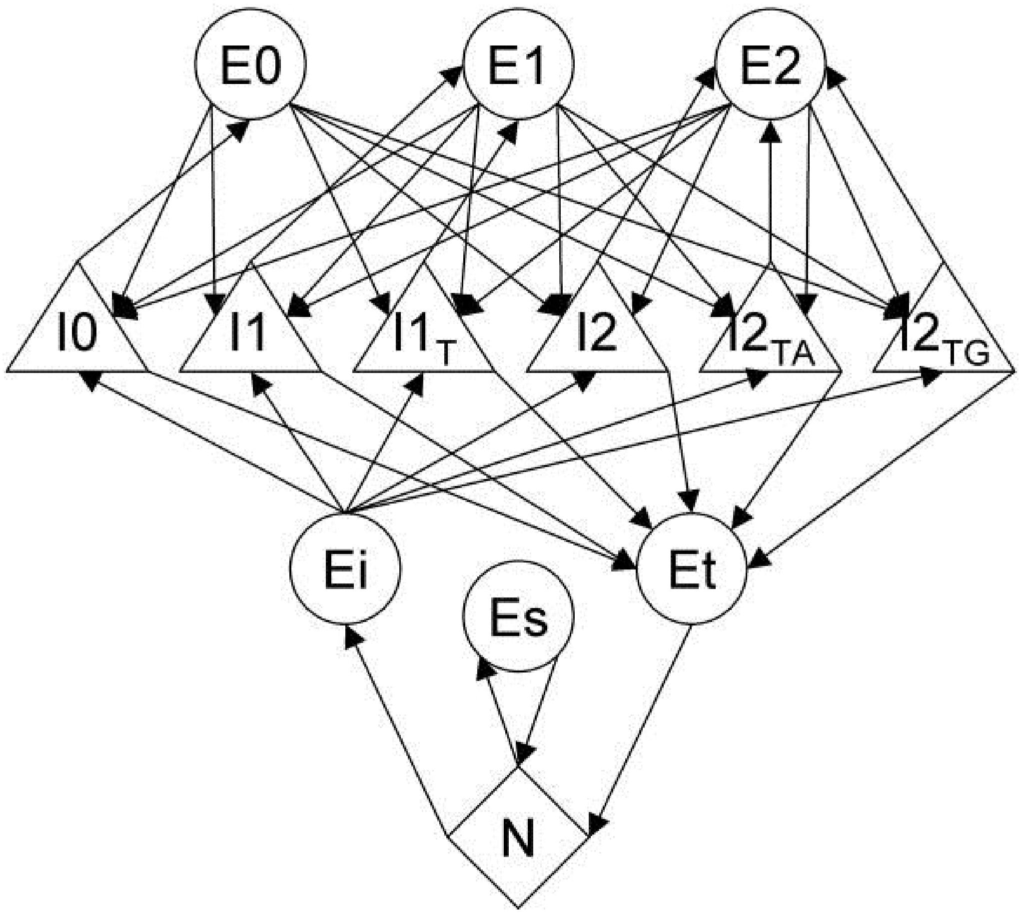



**Figure 4.2: The Snap transition state model [reproduced from (Korf, 2004)]. Each state of the HMM is represented by a shape and transitions between the states are represented by arrows. States include N: intergenic, $E_s$: single-exon gene, $E_i$: initial exon, Et terminal exon, $E_0$–$E_2$: exons in phase 0–2, $I_0$–$I_2$: introns in phases 0–2 (subscript of T, TA, or TG denotes the last bp or two bp of the intron – this is used to prevent in-frame stop codons across splice junctions).**

#### 4.2.1.4 CEGMA - Core Eukaryotic Genes Mapping Approach

The CEGMA pipeline (Parra et al., 2007) makes use of GeneID (Guigo, 1998), an ab-initio gene predictor, and Genewise (Birney et al., 2004), a homology based gene structure predictor, to predict 458 core eukaryotic proteins (KOGs) present in a wide range of taxa (*H. sapiens, D. melanogaster, C. elegans, A. thaliana, S. cerevisiae and S. pombe*). The pipeline identified 406 KOGs which were added to the gene models.

#### 4.2.1.5 Integration of gene models

Overlaps between all the predictions and the 'Illumina segments' were calculated based on GFF coordinates. Gene predictions from Genezilla, Snap and CEGMA that overlapped with the 'Illumina segments' (with 200 bp offset) were kept as an initial set of genes. 13,735 gene models from the different predictions resided within 200 bp of 12,368 'Illumina segments'. These genes were kept for the gene models (as set 1). Among the remaining genes that did not lie within 200 bp of 'Illumina segments', 158 genes had good BLAST homology with known genes in the NR database (set 2). The remaining gene models were filtered based on their length and other parameters. About 1021 genes were kept from this list (set 3). The CEGMA pipeline was used to build conserved eukaryotic gene models, resulting in 406 gene models predicted by CEGMA (set 4). The remaining 'Illumina segments' that did not overlap with any of the gene prediction programs were retrieved from the genome sequence with an addition 1 kb of flanking sequence on either side. A separate gene prediction was made on this data set. 1939 were gene models predicted from these extracted regions (set 5).

Sets 1-5 were manually integrated based on gene coordinate overlap. This resulted in 17,259 gene models being predicted for *Hpa*. 84% of the 'Illumina segments' lie within 200 bp of the gene models. We will call these gene models the version 1 (v1) gene models for the *Hpa* v8.3 assembly.

### 4.2.2 Evaluating gene models

#### *4.2.2.1 Evidence for expression*

As described in the previous chapter, we aligned both the ESTs and Illumina sequenced cDNA back to the reference genomes to compare how much of the transcription data is represented in the different reference genome assemblies. To identify how much of the transcribed sequence represented in the genome is described in the gene models, I aligned the ESTs to the gene models (using BLAT, minimum sequence identity of 80% and setting the query to RNA), and also aligned the Illumina cDNA reads to the gene models (using MAQ, mapping parameters of 3 mismatches in the 24 bp seed, and maximum sum of qualities of mismatching bases to 100).

We observed that 82.2% of the ESTs aligning to the v8.3 assembly also align to the v1 gene models presented by the VBI (table 4.1). From this analysis we would extrapolate that 80% of the genes of *Hpa* are represented by the gene models. This was considered to be reasonable for a draft genome assembly project. However, the 20% of the ESTs that aligned to the genome but not to the gene models, would suggest that 20% of the *Hpa* transcripts that are not annotated in the v1 gene models.

| | **Aligned EST** | **% of ESTs aligning** | **% of alignable ESTs** |
|---|---|---|---|
| ***Hpa* v8.3 genome assembly** | 29,836 | 93.9% | - |
| ***Hpa* v8.3 gene models (v1)** | 24,550 | 77.3% | 82.2% |

**Table 4.1: The number and percentage of the 31,759 ESTs aligning to *Hpa* v8.3 assembly and v1 gene models using BLAT (setting the query as RNA for the genome alignments, and setting the minimum sequence identity as 80% in both alignments). The percentage of alignable ESTs was calculated as the percentage of EST aligning to the genome that also aligned to the gene models.**

I also aligned the Illumina cDNA reads to the v8.3 genome assembly and the v1 gene models using MAQ (mapping parameters allowing for 4 mismatches in the 24 bp seed, and allowing for a maximum sum of qualities of mismatching bases to 100). 35.2% of the Illumina reads that aligned to the v8.3 assembly aligning to the v1 gene models (table 4.2). This is significantly less than the percentage of ESTs aligning to the v1 gene models. It was hypothesised that many Illumina cDNA reads would align to the untranslated regions (UTR) of genes and thus not align directly to the gene models as UTRs were not predicted.

The Illumina cDNA was obtained from *A. thaliana* Ws0-eds1 plants infected with *Hpa* Emoy2. In order to test the claim that much of the Illumina cDNA was aligning to the untranslated regions of genes, the cDNA was aligned to the:

- *A. thaliana* TAIR9 assembly (June 2009)
- *A. thaliana* TAIR9 gene models, including UTRs (June 2009)
- *A. thaliana* TAIR9 gene protein coding regions (i.e. not including UTRs) (June 2009)

Comparing the amount of cDNA aligning to the TAIR9 assembly and TAIR9 coding region provides a benchmark for comparing the *Hpa* genome assembly and gene models. Comparing the cDNA aligning to the TAIR9 gene models (including UTRs) and TAIR9 coding regions provides insight into how much of the cDNA is untranslated.

Aligning the cDNA using the previous protocol we found that 95.3% of the cDNA aligning to the TAIR9 genome assembly also aligned to the TAIR9 gene models including the UTR regions (table 4.2). It was surprising to observe that less than half of the aligning cDNA (45.8%) aligned to the protein coding regions of the genes. Since reads from *A. thaliana* Ws0-eds1 were aligned to the *A. thaliana* Col-0 genome and gene sequences, slightly more reads would be expected to align to a Ws-0 assembly. From this we hypothesise that with 'gold standard' gene models for *Hpa* we would expect at least 45% of the cDNA reads aligning to the genome to also align to the gene models, assuming a similar distribution of cDNA reads coming from UTRs and coding regions as we see in the *A. thaliana* Ws0-eds1 reads. Comparing this figure to the observed figure of 35.2% aligning to the *Hpa* v1 gene models, there is still an offset of 10% to our hypothetical optimum. We hypothesise that the difference is primarily due to missing gene models in the v1 set (table 4.2).

| | **Aligned cDNA** | **% of cDNA aligning** | **% of alignable cDNA** |
|---|---|---|---|
| ***A. thaliana* Tair9 assembly** | 3,074,292 | 36.0% | - |
| ***A. thaliana* Tair9 gene models** | 2,928,728 | 34.3% | 95.3% |
| ***A. thaliana* Tair9 gene CDS regions** | 1,410,598 | 16.5% | 45.8% |
| ***Hpa* v8.3 genome assembly** | 2,397,839 | 28.1% | - |
| ***Hpa* v8.3 gene models (v1)** | 844,524 | 9.9% | 35.2% |

**Table 4.2: The number and percentage of the 8,549,032 cDNA reads aligning to *A. thaliana* and *Hpa* gene models using MAQ (allowing for a 3 bp mismatch in the 24 bp seed, and a maximum sum of qualities of mismatching bases to 100). The percentage of alignable cDNA was calculated as the percentage of ESTs aligning to the genome that also aligned to the gene models.**

#### *4.2.2.2 Other quality issues in the v1 gene models*

I was dubious about the use of 'Illumina segments' to train Genezilla for gene predictions for exon content of genes, because the segments were constructed from reads separated by 400 bp. Since we are aware that the cDNA could be obtained from UTRs as well as exons, there is the possibility that these 'Illumina segments' cover introns that are shorter than 400 bp, and also possibly connect 2 genes that are separated by less than 400 bp. This would mean that Genezilla may have been trained using non-coding sequence, which may lead to incorrect gene predictions. This was an argument that was acknowledged but did not change the opinions of the post-doctoral researcher in charge of generating the gene models, with regards to the correctness of the method. While this remains a discussion point, it was decided that any modifications made to the gene models would be based on the Genezilla predictions.

Analysis of the gene lengths also revealed 2457 genes that were shorter than 50 amino acids. These were considered to the spurious gene calls.

TranposonPSI (Haas, 2010) was used to identify and analyse the *Hpa* v1 gene models for sequence similarity to known transposable elements using PSI-BLAST (Altschul et al., 1997). 1176 genes with high similarity to transposons over 75% of the length of the gene were identified. Both the genes less than 50 amino acids and the 1176 genes with homology to transposons are likely to be incorrect gene calls and should be removed.

### 4.2.3 Generating version 2 of the *Hpa* gene models

In the previous section problems with the current gene model predictions were identified, in that only approximately 80% of the expression data that aligns to the v8.3 assembly aligns to gene models, as well as other quality issues such as short gene models and failure to identify transposable elements. The aim for this part of the project was to provide a set of gene predictions to improve the current gene models. After each gene model prediction I evaluated how well they represent the transcribed sequence we have for *Hpa*. We aligned the ESTs using the previously described protocol (using BLAT with minimum sequence identity of 80%). We also evaluated gene models by aligning the Illumina cDNA, using a filtered subset of reads. Knowing that the Illumina cDNA reads contain *A. thaliana* transcribed sequence, we removed this to prevent a bias caused by reads aligning to both *A. thaliana* and *Hpa*, and belonging to transcribed sequence in both organisms. We performed the filtering using MAQ (v0.7.1) and extracted the reads that did not align to the *A. thaliana* TAIR9 genome assembly (using strict parameters allowing for 1 mismatch in the 24 bp seed, and a maximum of sum of mismatching bases to 40). This reduced the number of Illumina cDNA reads from 8,549,032 to 5,896,757. We aligned these filtered reads to the newer gene model predictions using MAQ (v0.7.1) allowing for 3 mismatch in the 24 bp seed, and a maximum of sum of mismatching bases to 100.

In the following section I will describe how additional gene prediction software was used to predict novel *Hpa* genes to identify unannotated *Hpa* transcripts. I will discuss how additional gene prediction software was used to generate additional gene models, and how they were integrated into the *Hpa* gene models.

#### *4.2.3.1 Geneid*

Geneid was chosen as one of the alternative gene prediction programs since it has already been pre-trained for predicting genes from another oomycete plant pathogen, *P. infestans*. This would then be ideal for recognising orthologous genes within the *Peronosporales*.
A default run with Geneid predicted 38,530 unfiltered gene models. Many of these predictions were very short (15,784 were under 50 amino acids) and 623 did not start with a Methionine. These spurious gene models were removed.

81.0% of the ESTs align to the Geneid gene models (table 4.3). This is an increase of 4.8% compared to the v1 gene models. However, we observed that 12.6% of the Illumina cDNA reads align to these gene models (table 4.4), which is 8.1% less reads aligning compared to the v1 gene models. This suggested that the Geneid models were able to predicted ~5% genes, compared to the v1 gene models, which are expressed in the zoospore stage of *Hpa* lifecycle. Similarly, Geneid predicted ~8% fewer genes, compared to the v1 gene models, which are expressed 7 days after infection.

The median length of the predicted genes (429 bp) (fig 4.3) is closer to the median length of gene predicted by Snap (448 bp), and less than the Genezilla (700) and CEGMA (1042) predictions. Despite the large number of genes predicted, the largest gene in the 90$^{th}$ percentile was 1803 bp.

It has been shown that genes have a higher GC content compared to that of the background (Pozzoli et al., 2008). The median GC percentage for the Geneid predictions is

52.6% which is similar to that of Genezilla. Despite the large number of genes predicted by Geneid, the distribution of GC percentages in the 90$^{th}$ percentile is not very large (fig 4.3), which is further evidence that the gene predictions by Geneid are fairly robust as the GC percentage across genes is consistent and higher that the background.

#### *4.2.3.2 Augustus*

We also decided to use Augustus (Stanke et al., 2008), which is a relatively new gene predictor that supports EST and Illumina cDNA training and was shown to performed well in the nGASP project (Coghlan et al., 2008).

In order to train Augustus with the EST sequences, they needed to be assembled using PASA (Haas et al., 2003), a program designed to align spliced alignments. PASA assembled the ESTs into 3601 genes, of which 1724 were randomly chosen as a training set for Augustus. I followed the Augustus manual for the training procedure, and used default parameters and did not predict UTRs. Using the training file and default parameters for Augustus, 12,678 gene models were predicted. Of these, 154 did not start with a Methionine and were removed. In addition, 3 genes were less than 50 amino acids long, but due to this low number were kept. For clarity I will refer to these as the “Augustus models”.

We aligned the ESTs to the Augustus gene models. 71.6% of the ESTs align to these gene models (table 4.3). This is a decrease of 7.4% compared to the v1 gene models. 13.5% of the Illumina cDNA reads align to these gene models (table 4.4), which is 1.5% less reads aligning compared to the v1 gene models.

We also observed the median length of the predicted genes (876 bp) (figure 4.3) to be between the median length of gene predicted by CEGMA (1042 bp) and Genezilla (700), and more than the Snap (448) predictions. The median GC percentage for the Augustus predictions is 53.1%, which is very close to median GC percentage observed in Geneid (52.6%) and Genezilla (53.2%). The distribution of GC percentage across the genes is very conservative, suggesting that the Augustus gene models are also robust and a good candidate set of gene predictions to complement the existing v1 gene models

An additional Augustus run was performed making use of the Illumina cDNA reads. We followed the Augustus protocol (BLAST alignment method) for generating “hints files” using Illumina cDNA data (http://augustus.gobics.de/binaries/readme.rnaseq.html). Using the additional "hints" provided by the Illumina cDNA reads as well as the EST training file yielded 34,028 gene models, of which 209 were shorter than 50 amino acids and 482 started without a Methionine. These genes were removed from the gene models. For clarity I will refer to these as the "Augustus hints models".

We aligned the ESTs to the Augustus hints gene models. 84.6% of the ESTs align to these gene models (table 4.3). This is an increase of 9.4% compared to the v1 gene models. 17.4% of the Illumina cDNA reads align to these gene models (table 4.4), which is 27.0% more reads aligning compared to the v1 gene models. This large increase in cDNA aligning is hypothesised to be due to the additional training of Augustus to make use of the Illumina cDNA. However, there is sufficient evidence in the increase in ESTs aligning to the Augustus hints model compared to the v1 gene models that there are a number of novel gene predictions with evidence of expression in the Augustus hints predictions.

We also observed the median length of the predicted genes (423 bp) (fig 4.3). This is very similar to the median observed in the Geneid predictions (429), although it did predict a number of much larger genes. The majority of the genes predicted were much smaller than the genes predicted by Augustus without the cDNA training. This could be due to the "Augustus hints models" being trained to predict partial genes in regions of high transcriptional activity identified by the Illumina cDNA reads.

The median GC percentage for the Augustus hints predictions is 50.4%. This was the lowest observed median GC percentage. We also saw a much larger distribution of different GC percentages compared to the other gene predictions (fig 4.3). This would suggest that the "Augustus hints model" are the least robust and should only be incorporated into the v1 gene models where there is direct evidence for expression. The large variation in GC percentage is also likely due to the large number of genes predicted, of which a large number are likely to be spurious,

| Gene Model | ESTs aligning | % of ESTs aligning | % change compared to v1 |
|---|---|---|---|
| **v8.3 gene models v1** | 24,550 | 77.3% | - |
| **Augustus** | 22,735 | 71.6% | -7.4% |
| **Augustus hints** | 26,863 | 84.6% | +9.4% |
| **Geneid** | 25,726 | 81.0% | +4.8% |

**Table 4.3: Number and percentage of ESTs aligning to gene model predictions using BLAT to align 31,759 ESTs with a minimum identity of 80%. The percentage change in alignment compared to the v8.3 gene models (v1) were calculated relative to the number of ESTs aligned to the v1 gene models.**

| Gene Model | cDNA aligning | % of cDNA aligning | % change compared to v1 |
|---|---|---|---|
| **v8.3 gene models v1** | 805,562 | 13.7% | - |
| **Augustus** | 795,443 | 13.5% | -1.5% |
| **Augustus hints** | 1,023,688 | 17.4% | +27.0% |
| **Geneid** | 740,541 | 12.6% | -8.1% |

**Table 4.4: Number and percentage of filtered Illumina cDNA reads aligning to gene model predictions using MAQ to align 5,896,757 reads (with mapping parameters allowing for 2 mismatches in the 24 bp seed and allowing for a maximum of 70 as the sum of qualities of mismatching bases). The percentage change in alignment compared to the v8.3 gene models (v1) were calculated relative to the number of cDNA reads aligned to the v1 gene models.**

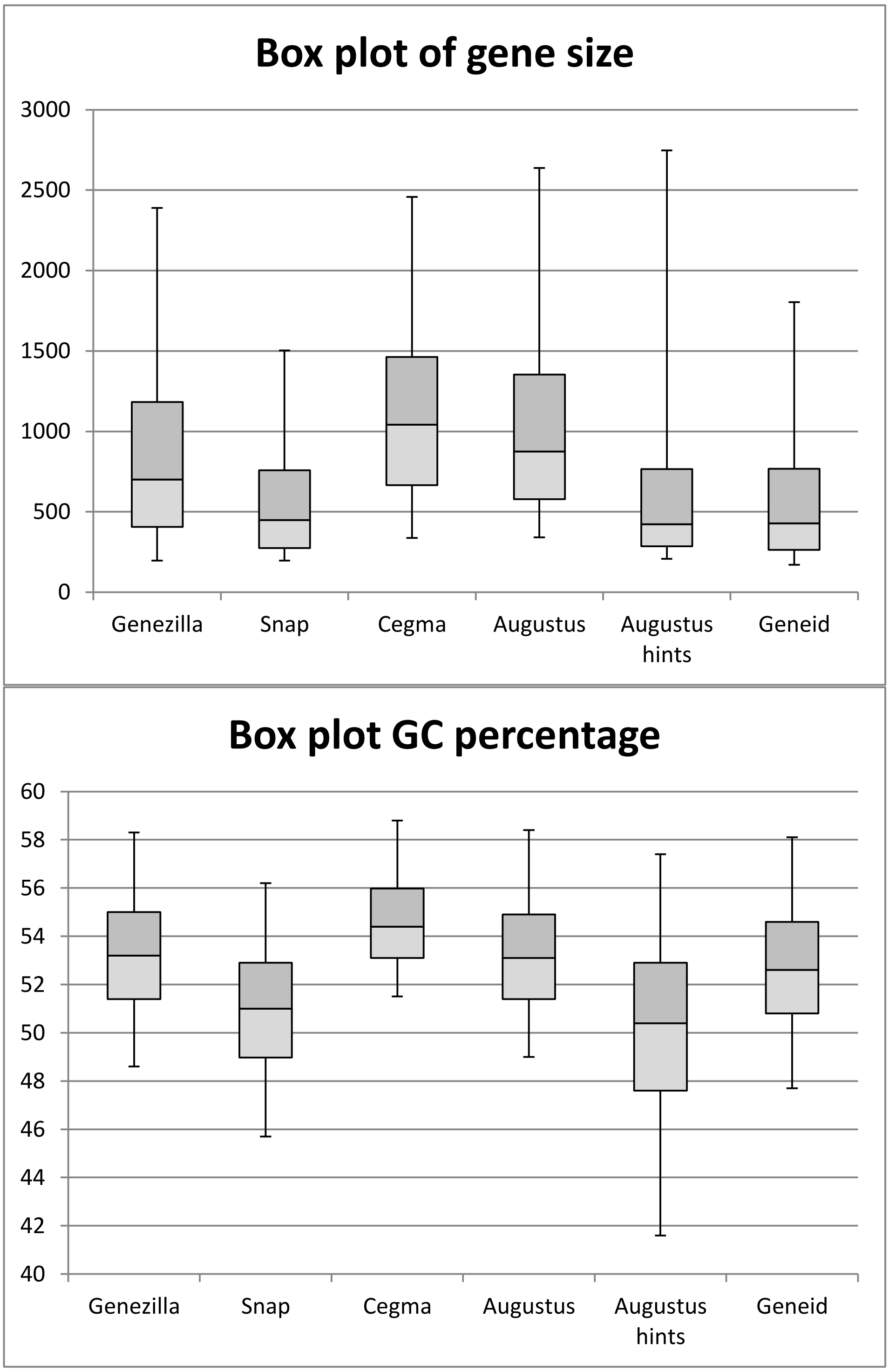



**Figure 4.3: Box plots of the 90th percentiles of gene lengths and GC percentage. Similar lower size distributions were observed between Geneid, Augustus hints and Snap, and similar higher size distributions between Genezilla, CEGMA and Augustus. The size of the distributions of GC percentage is similar in all predictions apart from the Augustus hints, which has a larger distribution.**

### *4.2.3.3 Integrating Augustus and Geneid gene predictions*

Our previous results show that there are a number of genes that should be removed from the v1 gene models and the evidence suggests that there are novel genes in the Geneid, Augustus and Augustus hints gene models with evidence of expression during zoospore and 7 d.p.i. stages. The strategy we adopted to identify gene predictions to complement the existing models was to identify genes predicted by Geneid and Augustus that aligned to expressed sequence that did not align to exisitng v1 gene predictions, and where the expressed sequences aligned better to the newer gene predictions compared to the v1 gene predicitons (either indicating longer gene predictions, or representing better prediction of intron exon boundaries).

All gene models from the Geneid, Augustus and Augustus hints models that shared the same coordinates as genes in the v1 gene models were removed. After this, we identified all gene models that shared the same coordinates in the new gene predictions and removed duplicate calls. These genes were then combined into a single file containing a redundant set of novel gene predictions. The genes were then filtered so that they did not contain transposable elements (identified by TranposonPSI, using the same protocol as described previously), were at least 50 amino acids in length and did not contain more than 25% interspersed repeats and low complexity DNA in the gene sequence (identified by RepeatMasker v3.1.6 (Smit et al., 1996-2010) using default parameters). Low complexity sequence is masked as this remove the majority of transposons from the analysis.

The ESTs were aligned to the list of v1 genes and new gene predictions using BLAT (minimum of 80% sequence identity). The ESTs aligned better to 993 genes from the new set of genes compared to the v1 gene models. These genes were then removed from the combined set of new genes and kept as potential new genes to add to the v1 gene models.

A set of “Illumina pseudo-ESTs” (cDNA that was assembled by alignment to the v8.3 genome) by extracting regions of the genome that had greater than 2x coverage of 50 bp when the filtered Illumina cDNA was aligned (using MAQ, mapping parameters n=3 e=100) were constructed. A total of 23,844 “Illumina pseudo-ESTs” were created. These “Illumina pseudo-ESTs” overcome the previously described quality issues of the “Illumina segments", as the "Illumina pseudo-ESTs" only described transcribed regions whereas the "Illumina segments" may also describe non transcribed regions. The "Illumina pseudo-ESTs" were aligned to the combined gene models and found that a number preferentially aligned to 1652 of the newer gene predictions compared to existing genes. These genes were extracted and added to the 993 genes chosen for their EST alignment to make a set of 2645 additional gene predictions to integrate into the gene models.

The following integration of these novel genes was performed by Dr Sucheta Tripathy. To prime the v1 gene models set for integrating the new genes, the genes less than 50 amino acids long (2457 genes) and genes with high identity to transposons over 75% of the gene (1176 genes) were removed. Checking the genomic co-ordinates of these gene predictions revealed 1609 new gene predictions overlapping with 2049 v1 gene predictions. The majority of the new overlapping predictions were in the same coding frames, had an extension of the 5' and the same stop locus. There were also a number of new overlapping gene predictions that were in different coding frames but had extensions of the 5' and 3' regions, for which both gene models were kept. The final integration replaced 1361 v1 gene predictions with 1321 new gene predictions, 288 new gene predictions with overlapping coordinates with v1 gene predictions but in different coding frames, and 1036 new gene predictions that did not overlap with existing v1 gene models .

To summarise the changes, the v1 gene models lost 2457 short genes (less than 50 amino acids), 1176 genes with high similarity to transposable elements and low complexity sequence. 1361 were replaced by newer gene predictions with better expression support. A total of 2645 genes from the new gene predictions were added, resulting in 14,910 gene predictions making the *Hpa* Emoy2 v8.3 assemblies "v2 gene models".

#### *4.2.3.4 Evaluating the Hpa Emoy2 v2 gene models*

In order to identify how much of the transcribed sequence represented in the genome is described in the v2 gene models, the ESTs were aligned to the gene models (using BLAT, minimum sequence identity of 80% and setting the query to RNA), and also aligned the Illumina cDNA reads to the gene models (using MAQ, mapping parameters of 3 mismatches in the 24 bp seed, and maximum sum of qualities of mismatches bases to 100).

90.6% of the ESTs aligning to the v8.3 assembly also align to the v2 gene models which is a 10% improvement compared to the v1 gene models presented by the VBI (table 4.5). The percentage of alignable ESTs in the v2 assemblies is greater than the ESTs alignable to the Geneid, Augustus and Augustus hints gene models. This is a substantial improvement over the previous gene models in representing the genes expressed during the zoospore stage.

| | Aligned EST | % of ESTs aligning | % of alignable ESTs |
|---|---|---|---|
| ***Hpa* v8.3 genome assembly** | 29,836 | 93.9% | - |
| ***Hpa* v8.3 gene models (v1)** | 24,550 | 77.3% | 82.2% |
| ***Hpa* v8.3 gene models (v2)** | 27,039 | 85.1% | 90.6% |

**Table 4.5: The number and percentage of the 31,759 ESTs aligning to the v8.3 genome, v1 and v2 gene models using BLAT (setting the query as RNA for the genome alignments, and setting the minimum sequence identity as 80% in both alignments). The percentage of alignable ESTs was calculated as the percentage of EST aligning to the genome that also aligned to the gene models.**

The filtered Illumina cDNA reads were aligned to the v8.3 genome assembly and the v2 gene models using MAQ (mapping parameters allowing for 4 mismatches in the 24 bp seed, and allowing for a maximum sum of qualities of mismatching bases to 100). We observed a 12.4% increase in the number of alignable cDNA to the v2 gene models compared to the v1 gene models (table 4.6). The percentage of alignable cDNA to the v2 gene models (45.3%) is comparable to the alignable cDNA of the 'gold standard' *A. thaliana* TAIR9 gene models (45.8%). However, we did observe that 13% more cDNA aligned to the Augustus hints model compared to the v2 gene models. From our protocol we deduced that the additional cDNA aligning to the Augustus hints gene models would be primarily singletons that were not considered when we chose additional new gene predictions to complement the v1 gene models.

| | Aligned cDNA | % of cDNA aligning | % of alignable cDNA |
|---|---|---|---|
| ***Hpa* v8.3 genome assembly** | 2,003,800 | 34.0% | - |
| ***Hpa* v8.3 gene models (v1)** | 805,562 | 13.7% | 35.2% |
| ***Hpa* v8.3 gene models (v2)** | 909,087 | 15.4% | 45.3% |

**Table 4.6: The number and percentage of the 5,896,757 filtered cDNA reads aligning to the v8.3 genome, v1 and v2 gene models using MAQ (allowing for a 3 bp mismatch in the 24 bp seed, and a maximum sum of qualities of mismatching bases to 100). The percentage of alignable cDNA was calculated as the percentage of ESTs aligning to the genome that also aligned to the gene models.**

10 genes were randomly selected from the additional genes that had no overlap with genes previously predicted in the v1 models. A BLASTX search against the NCBI NR

database (November 2009) was performed. The best BLAST results (and species) that were not conserved hypothetical proteins were:

- 2 ADP-ribosylation factor family (*P. infestans*)
- Amino acid/polyamine/organocation Family transporter protein (*P. infestans*)
- Conserved hypothetical protein (similar to ring finger protein) (*P. infestans*)

All the best hits, including the other 6 conserved hypothetical proteins, came from *P. infestans.* Although most of the hits were to conserved hypothetical proteins, these genes would otherwise not have been identified and thus not have been noted as potential genes conserved between *Hpa* and *P. infestans* or potentially even in higher orders of phylogeny. It was also noteworthy that 2 genes show homology to ADP-ribosylation factor (ARF) family proteins, as these have been shown to play a role in drug resistance and virulence in *Candida albicans* (Epp et al., 2010). It has also been shown that the Amino acid/polyamine/organocation Superfamily transporter, a transport system that existed before archaea and eukarya diverged from bacteria, have functions in amino acid and choline transport in eukaryotes (Jack et al., 2000) and may thus also pertain to roles in *Hpa* virulence.

To summarise, after shortfalls in the *Hpa* v1 gene models were identified, short genes and transposons were removed and alternative gene models were generated using Geneid and Augustus. Genes predicted by these methods that had evidence for expression but were not predicted in the v1 gene models were identified and added to the v1 gene models. The number of genes predicted remained fairly constant but the evidence for expression increased by ~10%. Using the alignment of Illumina cDNA as a benchmark, the new v2 gene models are comparable to the completeness of the TAIR9 gene models.

### 4.2.4 Generating version 3 of the *Hpa* gene models

Version 2 of the *Hpa* Emoy2 gene models based on the v8.3 assembly was the version included in the *Hpa* genome paper and is available online from the VBI Microbial Database website (http://vmd.vbi.vt.edu/download/index.php).

Post-doctoral researcher Dr Eric Kemen identified that there were still some shortfalls in the v2 gene models. There were a number of genes present in the sequenced oomycete gene models that were absent from *Hpa* (table 4.7). It is possible that these genes are not present in *Hpa.* However, since the majority of the sequenced oomycetes are *Phytophthora spp.*, which are closely related to *Hpa*, it is more likely that the genes are missing from the annotation, but are present in the *Hpa* genome assembly.

| **Gene** | **Function** | **Present in *Hpa* v8.3** |
|---|---|---|
| AlNc14C5G754 | Conserved hypothetical protein | N |
| AlNc14C51G4016 | Protein kinase putative | N |
| AlNc14C4G597 | Conserved hypothetical protein | Y |
| AlNc14C970G12675 | Eukaryotic translation initiation factor 3 subunit C putative | - |
| AlNc14C82G5339 | Guanylatebinding protein putative | Y |
| AlNc14C2G279 | Conserved hypothetical protein | Y |
| AlNc14C337G10741 | Hypothetical protein PITG_12566 | Y |
| AlNc14C139G7195 | Conserved hypothetical protein | Y |
| AlNc14C155G7625 | Flagellar associated protein putative | Y |
| AlNc14C114G6467 | Sporangia induced deflagellation inducible protein putative | N |
| AlNc14C8G1072 | Conserved hypothetical protein | Y |
| AlNc14C38G3287 | RNA helicase putative | Y |
| AlNc14C95G5839 | Annexin Family putative | Y |
| AlNc14C158G7697 | Conserved hypothetical protein | Y |
| AlNc14C48G3826 | Conserved hypothetical protein | Y |

**Table 4.7: List of genes (from the *Albugo laibachii* Nc14 gene models) identified by Dr Eric Kemen to be present in all published gene models of oomycete pathogens (March 2011) but not in the**

***Hpa* v2 gene models. Presence in the *Hpa* v8.3 assembly was determined by BLAST peptide homology.**

It was also noted by other post-doctoral researchers that the other major exclusion from the v2 genome was the lack of the predicted effector genes. The effector genes were predicted, analysed and annotated in the "VBI Oomycete Genomic Workshops" (2007, 2010) and were curated by Dr Rays Jiang (previously at the VBI, currently at the Broad Institute). However, these effector genes were never formally integrated into the v1 or v2 gene models.

Given the shortfalls of missing core oomycete genes and effectors in the gene models, I decided to perform a further round of gene integration. We also decided to check that all of the conserved eukaryotic genes (KOGs) were included as also taking another round to see if there were gene models removed in the integration process that removed genes with EST support.

### *4.2.4.1 Identifying missing genes*

We used the "IntersectBed" utility from the BEDTools v2.11.2 Suite (Quinlan and Hall, 2010) to identify differences between gene predictions GFF files. We compared the v2 gene models to genes we expected to see and identified missing genes that did not have a reciprocal 75% overlap with a gene prediction in the v8.3.

#### 4.2.4.1.1 Identifying missing conserved eukaryotic genes

So far we have mainly considered the inclusion of the 248 core eukaryotic genes present at single or low copy numbers (CEGs) into the genome. The CEGs are a subset of the full set of 458 core eukaryotic genes that are present at varying copy numbers. Comparing the v2 gene model overlap with the predicted coordinates of 406 predicted KOGs showed that 23 of the predicted KOGs were missing from the gene models.

#### 4.2.4.1.2 Identifying missing oomycete genes

As previously mentioned, Dr Eric Kemen identified a number of genes that were not present in the *Hpa* v2 gene models, but were present in gene models from sequenced *Phytophthoras* and *A. laibachii*. Under the assumption that gene conservation is more likely between more phylogenetically related organisms we decided to focus on genes

conserved between *P .infestans, P. ramorum, and P. sojae, and* modifying the CEGMA pipeline to search for conserved single copy *Phytophthora* genes (PCEGs).
The steps in the CEGMA pipeline that we would need to modify are the KOG FASTA files (which are used to identify regions were KOGs may lie), the KOG HMM models, (which are used to predict gene models by GeneWise (Birney et al., 2004)), and the cut off table for HMM searches (which is used to evaluate the GeneWise).

A list of 7113 conserved genes present as single copy in *P. infestans, P. ramorum,* and *P. sojae* were computed. For this we used clustering software, OrthoMCL (Chen et al., 2005), to cluster the gene models (*P. ramorum* and *P. sojae* gene models were downloaded from http://vmd.vbi.vt.edu/download/index.php, and the *P. infestans* gene models were provided by Prof Sophien Kamoun's group, The Sainsbury Laboratory, Norwich). Sequences for all genes in the gene clusters with 1 gene from each organism (i.e. single copy, conserved genes in the *Phytophthora* lineage) were extracted. Then for each gene cluster a HMM was constructed using hmmer3 (Finn et al., 2011) based on a ClustalW (Thompson et al., 1994a) alignment of the genes in each gene cluster using default parameters. The HMM cut-off score were set to 50% of the average score of each gene in the gene cluster when running a HMM search using the constructed HMM.

After running this modified version of CEGMA, 5755 of the 7113 PCEGs were identified in *Hpa*. We found that 1527 of these gene predictions did not overlap with genes in the v2 gene models. These genes are likely to be conserved oomycete genes that are not expressed during the zoospore stage, or in the later stages of infection.

The sequences of these PCEGs were used to perform a BLAST search against the v8.3 genome (using BLASTn, minimum e-value of $1 \times 10^{-20}$, and over 75% of the length of the gene), and compared the coordinates of these BLAST results against the v2 gene models. In addition to the 1527 missing PCEGs, we identified 91 regions that may contain PCEG homologues where there was no existing gene prediction in the v2 gene models.

#### 4.2.4.1.3 Identifying missing effector genes

The predicted effectors were aligned to the v8.3 gene models using BLAT. After comparing alignments of the best BLAT hits with the v2 gene models, we found 254 of 580 complete predicted effector were not in the v2 gene models. This included 81 of the 141 high confidence effector set, and 19 of the 22 predicted crinkler-like genes.

#### 4.2.4.1.4 Identifying missing genes with PASA assembled EST support

The coordinates of the PASA assembled ESTs, which were ESTs that aligned and assembled to the genome to encode for a full reading frame, were compared to the coordinates of the v1 gene more. 3369 PASA assembled ESTs that did not overlap with existing gene models were identified as additional gene candidates to integrate into *Hpa* gene models.

### *4.2.4.2 Integrating the additional genes into the v2 gene models*

Before we started to incorporate these missing genes we performed a Genemark-ES (Lomsadze et al., 2005) gene prediction (using default parameters) under advice from Dr Eric Kemen. GeneMark-ES is an ab-intio gene predictor based on a self-training algorithm for eukaryotes and was successfully used as the main gene predictor for the *A. laibachii* genome project (Kemen et al., 2011). This yielded 20,940 gene model predictions.

We developed an iterative method to incorporate the missing genes from the existing predictions in the Augustus, Augustus hits, Geneid and Genemark-ES gene predictions. The pipeline is summarised as follows:

1. Identify the overlapping genes between the missing gene coordinates and each set of the gene predictions (Augustus, Augustus hints, Geneid and Genemark-ES), using the IntersectBed utility with at least, e.g. reciprocal 80% overlap.

2. Extract the genes from the gene predictions yielding the highest number of overlapping genes, and remove the genes with which they overlap from the missing gene coordinates.

3. Repeat 1-2 for each of the other gene predictions.

4. Repeat 1-3 reducing the reciprocal overlap by, e.g 10% until a minimum of 60%.

#### 4.2.4.2.1 Integrating missing CEG and PCEG genes

We combined the missing CEG and PCEG genes into a single dataset and started to look for overlapping genes. After 8 iterations of gene integrations we identified 1145 genes that had 60%-100% reciprocal overlap with the missing CEG and PCEG genes (tables 4.8). The remaining 441 missing genes (5 CEGs, 379 PCEGs and 57 regions with homology to PCEGs) were added to make 1586 potential additional genes.

**A**

| Iteration | % reciprocal overlap | Gene prediction | Genes added |
|---|---|---|---|
| 1 | 80 | Genemark | 623 |
| 2 | 80 | Augustus Hints | 112 |
| 3 | 80 | Geneid | 55 |
| 4 | 80 | Augustus | 11 |
| 5 | 60 | Genemark | 234 |
| 6 | 60 | Augustus Hints | 67 |
| 7 | 60 | Geneid | 40 |
| 8 | 60 | Augustus | 3 |
| 9 | 100 | CEG | 5 |
| 10 | 100 | PCEG | 379 |
| 11 | 100 | PCEG homologs | 57 |
| | | **Total** | **1586** |

**B**

| **Sum of Genes added** | |
|---|---|
| **Gene prediction** | **Total** |
| Augustus Hints | 179 |
| CEG | 5 |
| PCEG | 379 |
| PCEG homologs | 57 |
| Geneid | 95 |
| Genemark | 857 |
| Augustus | 14 |
| **Grand Total** | **1586** |

**Tables 4.8: Number of (A) genes added over each integration iteration to integrate the missing CEG, PCEG and PCEG homolog genes; (B) genes added from each set of gene predictions.**

#### 4.2.4.2.2 Integrating missing effector genes

After 19 iterations of gene integrations we identified 300 genes that had 40%-100% reciprocal overlap with the missing CEG and PCEG genes (tables 4.9). The remaining 207 missing effectors were added manually to make 507 potential additional genes.

**A**

| Iteration | % reciprocal overlap | Gene prediction | Genes added |
|---|---|---|---|
| 1 | 99 | Augustus | 120 |
| 2 | 99 | Genemark | 28 |
| 3 | 99 | Augustus Hints | 14 |
| 4 | 99 | Geneid | 4 |
| 5 | 95 | Augustus | 25 |
| 6 | 95 | Genemark | 8 |
| 7 | 95 | Geneid | 3 |
| 8 | 95 | Augustus Hints | 1 |
| 9 | 90 | Augustus | 20 |
| 10 | 90 | Genemark | 5 |
| 11 | 90 | Augustus Hints | 2 |
| 11 | 100 | JJ | 28 |
| 12 | 100 | Jamboree Effector | 179 |
| 13 | 60 | Augustus Hints | 29 |
| 14 | 60 | Genemark | 9 |
| 16 | 60 | Geneid | 6 |
| 17 | 40 | Augustus Hints | 16 |
| 18 | 40 | Geneid | 8 |
| 19 | 40 | Augustus | 2 |
| | | **Total** | **507** |

**B**

| **Sum of Genes added** | |
|---|---|
| **Gene prediction** | **Total** |
| Augustus | 167 |
| Augustus Hints | 62 |
| Geneid | 21 |
| Genemark | 50 |
| Jamboree Effector | 179 |
| JJ | 28 |
| **Grand Total** | **507** |

**Tables 4.9: Number of (A) genes added over each integration iteration to integrate the missing Effector genes; (B) genes added from each set of gene predictions. Jamboree effectors are the effectors that were predicted in the 2010 Jamboree at the VBI; JJ denotes effectors predicted in the Jones Lab, TSL.**

#### 4.2.4.2.3 Integrating missing PASA assembled genes

After 12 iterations of gene integrations we identified 1145 genes that had 60%-100% reciprocal overlap with the missing CEG and PCEG genes (tables 4.10). The remaining 441 missing genes (5 CEGs, 379 PCEGs and 57 regions with homology to PCEGs) were added to make 1586 potential additional genes.

**A**

| Iteration | % reciprocal overlap | Gene prediction | Genes added |
|---|---|---|---|
| 1 | 90 | Augustus Hints | 52 |
| 2 | 90 | Geneid | 25 |
| 3 | 90 | Genemark | 15 |
| 4 | 90 | Augustus | 3 |
| 5 | 75 | Augustus Hints | 169 |
| 6 | 75 | Geneid | 105 |
| 7 | 75 | Genemark | 41 |
| 8 | 75 | Augustus | 16 |
| 9 | 60 | Geneid | 308 |
| 10 | 60 | Augustus Hints | 173 |
| 11 | 60 | Genemark | 56 |
| 12 | 60 | Augustus | 10 |
| | | **Total** | **973** |

**B**

| **Sum of Genes added** | |
|---|---|
| **Gene prediction** | **Total** |
| Augustus | 29 |
| Augustus Hints | 394 |
| Geneid | 438 |
| Genemark | 112 |
| **Grand Total** | **973** |

**Tables 4.10: Number of genes (A) added over each integration iteration to integrate the missing CEG, PCEG and PCEG homolog genes; (B) genes added from each set of gene predictions.**

#### 4.2.4.2.4 Removing redundancy and overlap

There was a possibility that our dataset of additional genes could contain duplicates due to the iterative nature of the pipeline. After collating all the genes to be added, 78 redundant genes, which is 2.5% of the entire set (tables 4.11).

**A**

| | Genes added | Unique |
|---|---|---|
| CEGs + PCEGs | 1586 | 1562 |
| Effectors | 507 | 499 |
| PASA | 973 | 927 |
| Total | **3066** | **2988** |

**B**

| | Genes added | Unique |
|---|---|---|
| Augustus | 210 | 187 |
| Augustus Hints | 635 | 621 |
| Geneid | 554 | 547 |
| Genemark | 1019 | 1004 |
| CEG | 5 | 5 |
| PCEG | 379 | 379 |
| PCEG homologs | 57 | 57 |
| Jamboree Effector | 179 | 160 |
| JJ | 28 | 28 |
| Total | **3066** | **2988** |

**Tables 4.11: Number of residual genes to add to the v2 assembly after removing redundant calls in each dataset with regards to (A) gene type and (B) gene predictor**

After removing the internal redundancy of the set, we concatenated the gene list to the v2 gene models, and did a further round of inspection to remove redundant genes based on gene coordinate overlap. Genes that were completely inside another larger gene call in the same reading frame were deleted. It was also noted that some of the genes added to the v2 gene models were removed. This is due to the nature of the pipeline used to preferentially identify gene calls closer in size, e.g. a smaller gene may have been chosen over a larger gene because it had a higher reciprocal overlap with an EST, but the EST may only be a fraction of the real gene in which case the longer gene call is preferred. We removed a total of 2341 genes from the v2 gene models. 850 of these were duplicate gene calls, and 1491 were due to being replaced by larger genes in the additional gene set just generated. 976 genes from the additional set were removed due to overlapping with already existing larger gene calls in the v2 gene models. This reduced the number of additional genes to 2012 genes. The total number of genes in the new set of gene models, which we will refer to as the v3 gene models, is 14,582 (tables 4.12).

A

| Start | Number of genes |
|---|---|
| v2 genes | 14911 |
| | |
| | |
| | |
| | |
| CEG + CEG | 1562 |
| Effectors | 499 |
| PASA | 927 |
| **Total** | **17899** |

B

| Removed from | Why/ overlap | Number of genes |
|---|---|---|
| v2 | Duplicate | 850 |
| v2 | CEG+PCEG | 1245 |
| v2 | EFF | 48 |
| v2 | PASA | 198 |
| | | |
| CEG PCEG | v2 genes | 374 |
| EFF | v2 genes | 240 |
| PASA | v2 genes | 362 |
| **Total** | | **3307** |

C

| Final | Number of genes |
|---|---|
| v2 genes | 12570 |
| | |
| | |
| | |
| | |
| CEG + PCEG | 1188 |
| Effectors | 259 |
| PASA | 565 |
| **Total** | **14582** |

**Tables 4.12: Summary of (A) added genes; (B) genes removed due to duplication in the existing v2 gene models, or v2 genes being replaced with longer predictions, or removing newer predictions due to overlapping with larger existing v2 genes; (C) the number of genes remaining to make the *Hpa* Emoy2 v8.3 genome v3 gene models.**

### *4.2.4.3 Evaluating the Hpa Emoy2 v3 gene models*

#### 4.2.4.3.1 Evidence of expression

In order to identify how much of the transcribed sequence represented in the genome is described in the v3 gene models, I aligned the ESTs to the gene models (using BLAT, minimum sequence identity of 80% and setting the query to RNA), and also aligned the Illumina cDNA reads to the gene models (using MAQ, mapping parameters of 3 mismatches in the 24 bp seed, and maximum sum of qualities of mismatching bases to 100).

91.8% of the ESTs align to the v8.3 assembly which is a 1% improvement over the v2 gene models (table 4.13).

| | Aligned EST | % of ESTs aligning | % of alignable ESTs |
|---|---|---|---|
| ***Hpa* v8.3 genome assembly** | 29,836 | 93.9% | - |
| ***Hpa* v8.3 gene models (v1)** | 24,550 | 77.3% | 82.2% |
| ***Hpa* v8.3 gene models (v2)** | 27,039 | 85.1% | 90.6% |
| ***Hpa* v8.3 gene models (v3)** | 27,415 | 86.3% | 91.8% |

**Table 4.13: The number and percentage of the 31,759 ESTs aligning to the genome, v1, v2, and v3 gene models using BLAT (setting the query as RNA for the genome alignments, and setting the**

**minimum sequence identity as 80% in both alignments). The percentage of alignable ESTs was calculated as the percentage of ESTs aligning to the genome that also aligned to the gene models.**

The filtered Illumina cDNA reads were aligned to the v8.3 genome assembly and the v3 gene models using MAQ (mapping parameters allowing for 4 mismatches in the 24 bp seed, and allowing for a maximum sum of qualities of mismatching bases to 100). We observed a 4.8% increase in the number of alignable cDNA to the v3 gene models compared to the v2 gene models (table 4.14). The percentage of alignable cDNA to the v2 gene models (47.5%) is higher than the alignable cDNA of the 'gold standard' *A. thaliana* TAIR9 gene models (45.8%).

| | Aligned cDNA | % of cDNA aligning | % of alignable cDNA |
|---|---|---|---|
| ***Hpa* v8.3 genome assembly** | 2,003,800 | 34.0% | - |
| ***Hpa* v8.3 gene models (v1)** | 805,562 | 13.7% | 35.2% |
| ***Hpa* v8.3 gene models (v2)** | 909,087 | 15.4% | 45.3% |
| ***Hpa* v8.3 gene models (v3)** | 951,743 | 16.4% | 47.5% |

**Table 4.14: The number and percentage of the 5,896,757 filtered cDNA reads aligning to the genome, v1, v2, and v3 gene models using MAQ (allowing for a 3 bp mismatch in the 24 bp seed, and a maximum sum of qualities of mismatching bases to 100). The percentage of alignable cDNA was calculated as the percentage of EST aligning to the genome that also aligned to the gene models.**

#### 4.2.4.3.2 Average length + GC

We noticed a shift in the median and distribution of gene lengths from the v1, to the v2, to the newest v3 gene models. The increase in median gene length from v2 (663 bp) to the v3 (711 bp) genes models was expected due to preferential choice of longer genes. There is very little change in the mean GC content of genes from the v2 to the v3 gene models, which suggest that the additional genes follow the expected GC content and the gene models are as robust. It was also noted that the distribution of the GC content in the v1 gene models is not as tight as that in the v2 and v3, which does suggest that the quality of the gene calls in the v2 and v3 gene models is significantly better. (fig 4.4, tables 4.15).

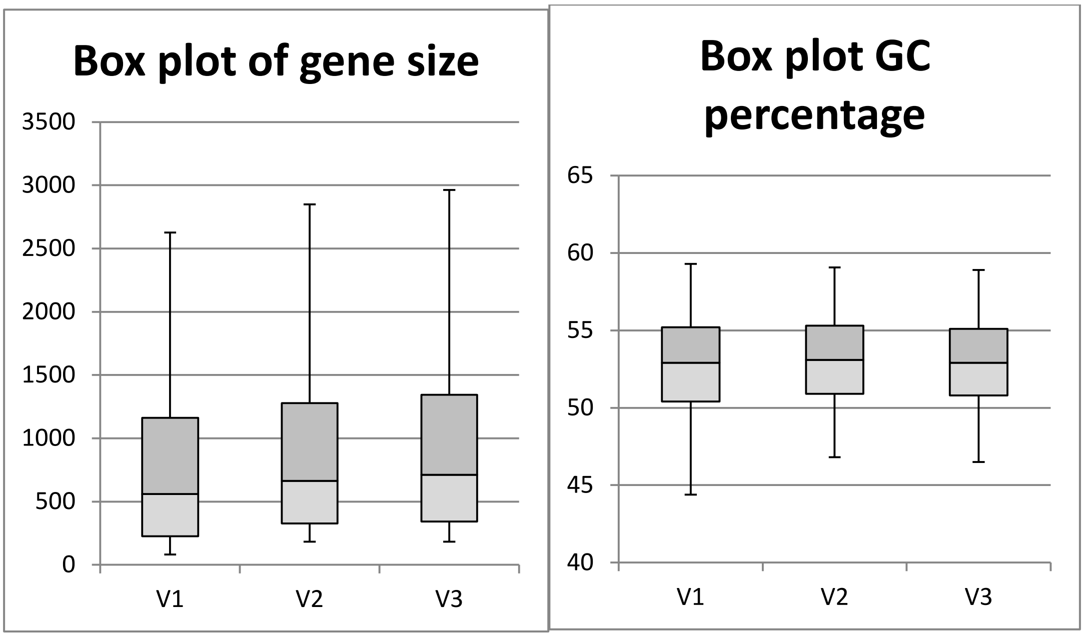



**Figure 4.4: Box plots of the 90th percentiles of gene lengths and GC percentage.**

| Length | v1 | v2 | v3 |
|---|---|---|---|
| 95 percentile | 2626.3 | 2848.65 | 2964 |
| 3rd quartile | 1162 | 1278 | 1344 |
| Median | 559 | 663 | 711 |
| 1st quartile | 226 | 327 | 342 |
| 5 percentile | 82 | 183 | 183 |

| GC | v1 | v2 | v3 |
|---|---|---|---|
| 95 percentile | 59.3 | 59.1 | 58.9 |
| 3rd quartile | 55.2 | 55.3 | 55.1 |
| Median | 52.9 | 53.1 | 52.9 |
| 1st quartile | 50.4 | 50.9 | 50.8 |
| 5 percentile | 44.39 | 46.8 | 46.5 |

**Table 4.15: Length and GC values for the v1, v2 and v3 gene models. The 5th percentile and 95th percentile were used as the lower and upper bounds for the box plots.**

### 4.2.5 Annotation

#### *4.2.5.1 InterproScan functional annotation*

We used the InterProScan v4.7 (Quevillon et al., 2005; Zdobnov and Apweiler, 2001), which annotates peptides against a number of databases:

- ProDom (Bru et al., 2005) using BlastProDom (Blastall) (Zdobnov and Apweiler, 2001)
- PRINTS (Attwood et al., 2003) using FingerPRINTScan (Scordis et al., 1999)
- SMART (Letunic et al., 2002) using Hmmpfam (Finn et al., 2011)
- TIGRFAMs (Haft et al., 2003) using Hmmpfam (Finn et al., 2011)
- Pfam (Bateman et al., 2004) using Hmmpfam (Finn et al., 2011)
- PROSITE (Hulo et al., 2004) using ScanRegExp + ProfileScan (Thompson et al., 1994b)
- PIRSuperFamily (Wu et al., 2004) using Hmmpfam (Finn et al., 2011)
- SUPERFAMILY (Gough et al., 2001) using Hmmpfam (Finn et al., 2011)
- CATH (Pearl et al., 2000) using Hmmpfam (Finn et al., 2011)
- PANTHER (Thomas et al., 2003) using Hmmsearch (Finn et al., 2011)
- Transmembrane using TMHMM2.0 (Sonnhammer et al., 1998)
- Signal peptides using SignalPHMM (Bendtsen et al., 2004)
- Low complexity regions using SEG (Wootton and Federhen, 1993)
- 3D Structure using Gene3D
- Coiled coils using COILS (Lupas et al., 1991)

In addition to these searches, we annotated the genes with Gene Ontology (GO) terms (Ashburner et al., 2000) using InterproScan.

78.5% of the v3 genes (11,451 genes) were functionally annotated using InterProScan. The breakdown of the number of genes annotated is shown in table 4.16.

| Annotation | Number of genes Annotated | % of all genes |
|---|---|---|
| **Coil** | 2020 | 13.9% |
| **GO** | 5431 | 37.2% |
| **HMMPfam** | 6143 | 42.1% |
| **HMMSmart** | 2329 | 16.0% |
| **InterPro** | 9734 | 66.8% |
| **ProfileScan** | 2271 | 15.6% |
| **Gene3D** | 4971 | 34.1% |
| **HMMPanther** | 5546 | 38.0% |
| **Seg** | 8683 | 59.5% |
| **Superfamily** | 5315 | 36.4% |
| **PatternScan** | 1480 | 10.1% |
| **SignalPHMM** | 2710 | 18.6% |
| **TMHMM** | 1967 | 13.5% |
| **HMMTigr** | 672 | 4.6% |
| **FPrintScan** | 979 | 6.7% |
| **HAMAP** | 165 | 1.1% |
| **All programs** | 11,451 | 78.5% |

**Table 4.16: Number and percentage of genes annotated using various programs.**

### *4.2.5.2 GO term annotation*

Analysing the distribution of GO terms, we see that the majority of genes annotated with a GO function pertain to molecular function, catalytic activity, biological process and metabolism (fig 4.5, appendix table 4.1). These GO terms account for more than 50% of all annotations. We also see a large subcomponent of GO annotation that may pertain to pathogenicity, for example, transferase activity, hydrolase activity and transport.

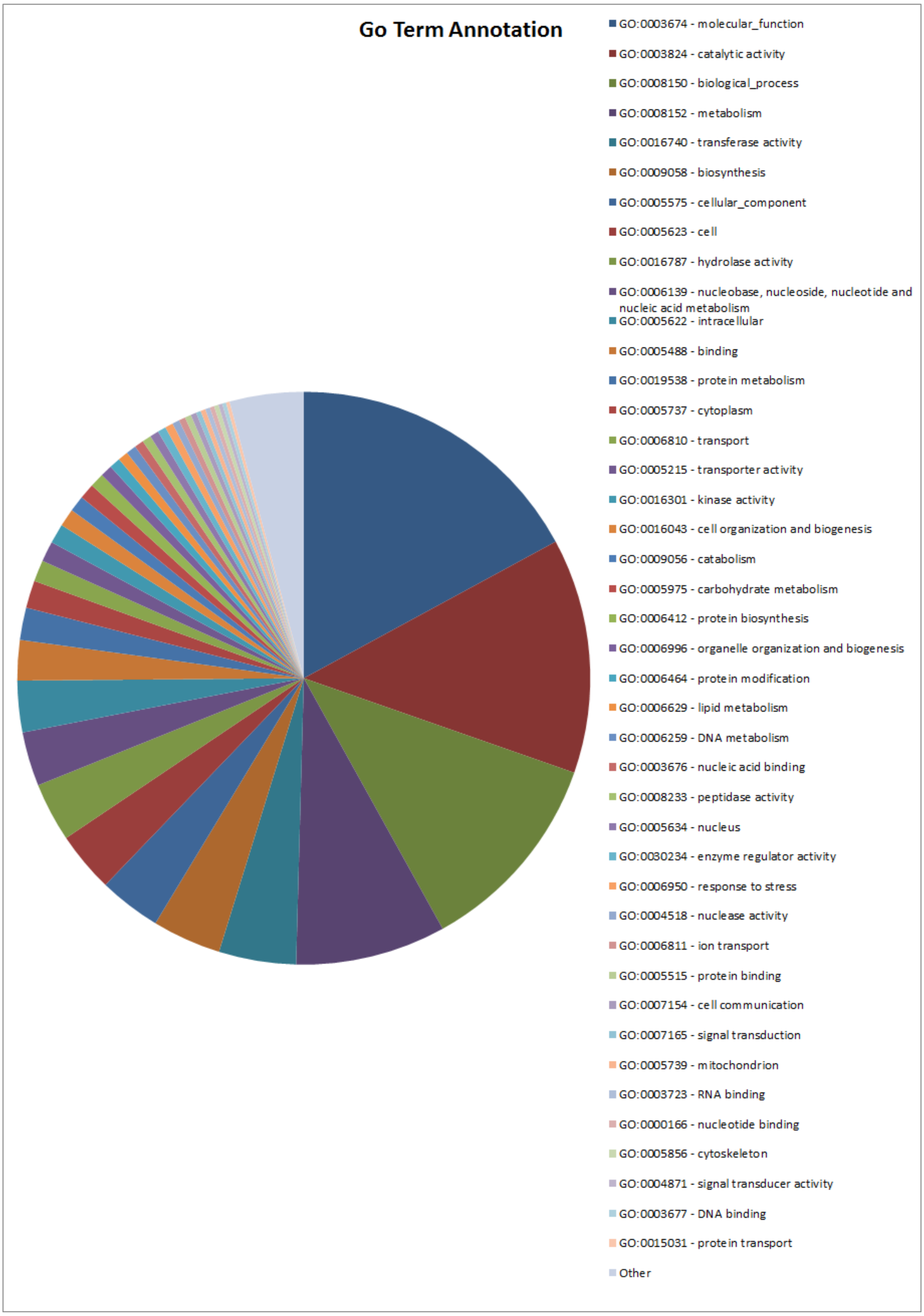



**Figure 4.5: The distribution of GO terms identified in the *Hpa* v3 gene models. The full list is described in appendix table 4.1.**

### *4.2.5.3 Localisation*

We predicted protein localisation using WolfPsort (Horton et al., 2007) on a fungal model (table 4.17).

| **Localisation** | **Number of genes** |
|---|---:|
| cytoskeleton | 439 |
| cytosol | 2000 |
| cytosol-mitochondria | 60 |
| cytosol-nuclear | 533 |
| cytosol-peroxisome | 6 |
| endoplasmic reticulum | 31 |
| extracellular | 1080 |
| Golgi apparatus | 9 |
| mitochondria | 4425 |
| mitochondria-nuclear | 42 |
| nuclear | 4679 |
| peroxisome | 13 |
| plasma membrane | 1240 |
| **Total** | **14,557** |

**Table 4.17: Breakdown of localisation predictions using WolfPsort on a fungal model.**

### *4.2.5.4 Secreted and transmembrane proteins*

We used SignalP 3.0 HMM eukaryotic model to predict the number of genes that are secreted. It was predicted that 2710 genes have a signal peptide and cleavage site using a cut-off of 90%. We found that 38.3% of the secreted proteins (1039 genes) also had predicted transmembrane domains, and 11.0% of secreted proteins (298 genes) are or are homologous to effectors.

Using the InterproScan search we identified 1967 proteins with predicted transmembrane helices. Of these proteins, 52.8% (1039 genes) had a predicted signal peptide and 2.5% (50 genes) are or are homologous to effectors. The overlap of the genes is show in figure 4.16.

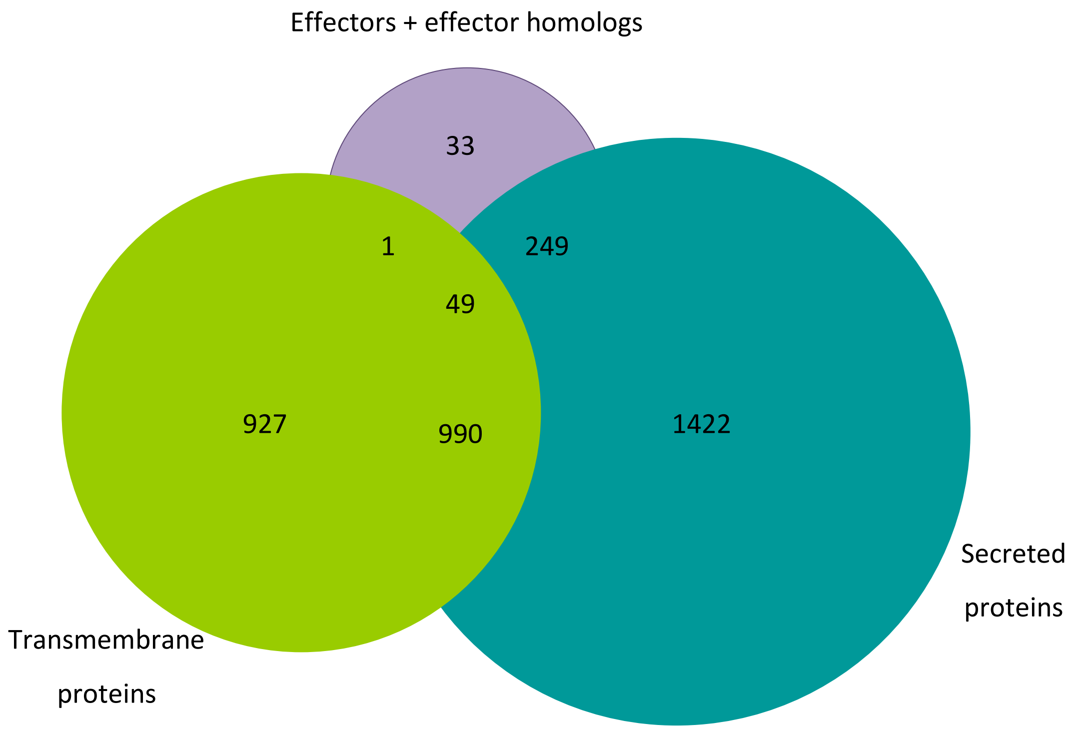



**Figure 4.6: Venn diagram of genes with overlapping annotation as secreted, transmembrane and effector or effector homolog genes.**

### *4.2.5.5 Metabolic Pathway Analysis.*

Pathway annotation for *Hpa* was done using KAAS (Moriya et al., 2007). The gene models were submitted to KAAS for assigning a KEGG Orthology (Ogata et al., 1999) identifier. The query sequences were blasted against the KEGG Genes reference database (containing genes from *Homo sapiens, Drosophila melanogaster, Caenorhabditis elegans, Arabidopsis thaliana, Saccharomyces cerevisiae, Cryptosporidium hominis, Escherichia coli K-12 MG1655, Neisseria meningitidis MC58, Helicobacter pylori 26695, Rickettsia prowazekii, Bacillus subtilis, Lactococcus lactis subsp. lactis IL1403, Clostridium acetobutylicum ATCC 824, Mycoplasma genitalium, Mycobacterium tuberculosis H37Rv, Chlamydia trachomatis D/UW-3/CX, Borrelia burgdorferi B31, Bacteroides thetaiotaomicron, Synechocystis sp. PCC6803, Deinococcus radiodurans, Aquifex aeolicus, Methanocaldococcus jannaschii* and *Aeropyrum pernix*), with homologs selected on reciprocal best blasts hits with a minimum sequence similarity threshold of 60%. These candidates were divided into KO groups and an assignment score was calculated. A pathway diagram was constructed using the non-organism specific option (fig 4.7).

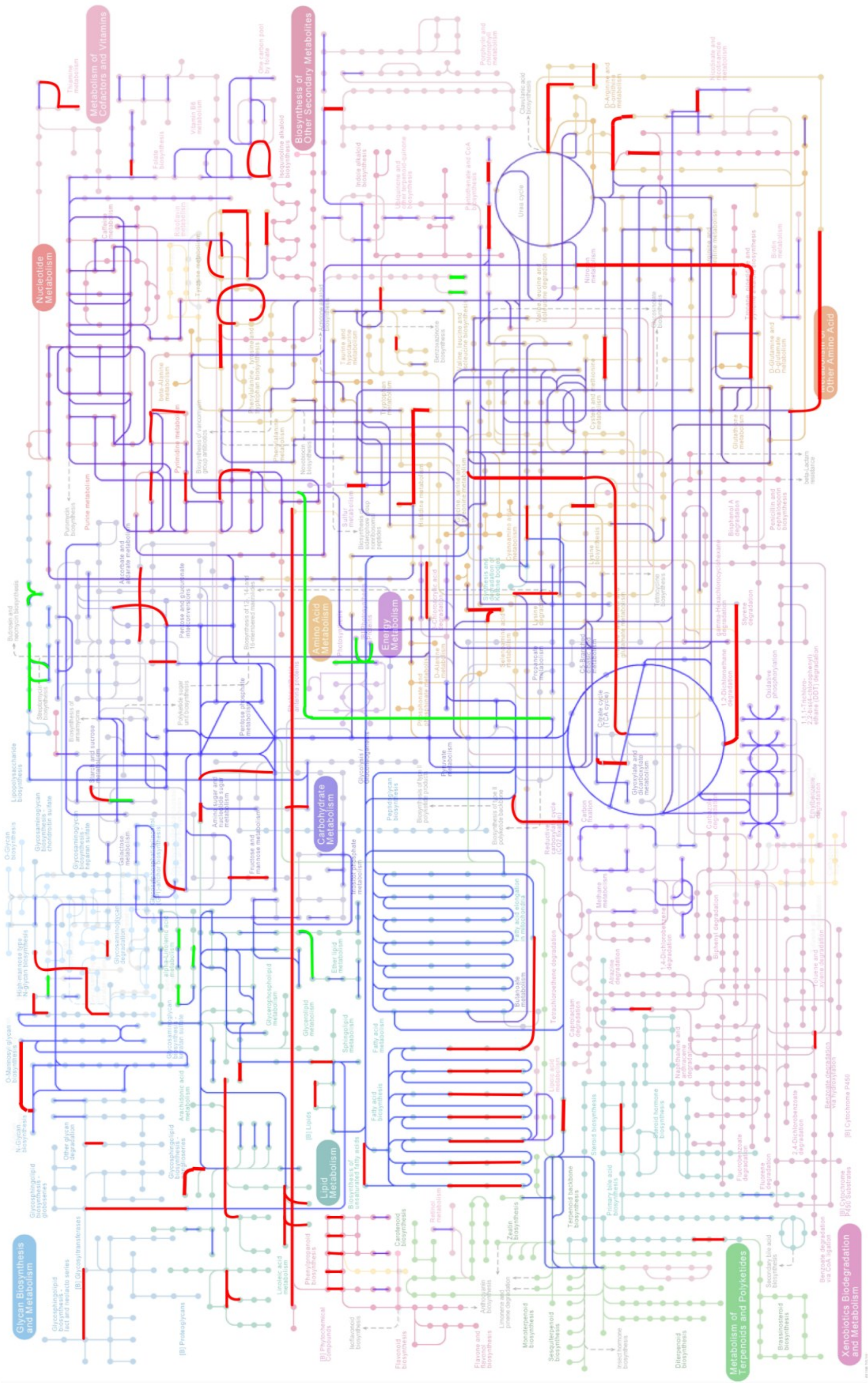



**Figure 4.7: Metabolic pathways in *Hpa*, *P. infestans*, *P. sojae* and *P. ramorum.* Components highlighted in blue are present in *Hpa* and at least one *Phytophthora* species. Components in red are absent in *Hpa* but present in at least 2 out of three *Phytophthora* species. Components in green are present in *Hpa* but are absent in the *Phytophthora* species. Dulled lines indicate components absent in all four species.**

### *4.2.5.6 Lack of nitrogen and sulphur assimilation pathways*

The genes for nitrate, nitrite and sulfite reductases could not be found (table 4.18) in the gene models, v8.3 assembly and Sanger ESTs and 454 ESTs. These genes are thus missing from the *Hpa* genome. In the case of the nitrate and nitrite reductases this conclusion is supported by the fact that those two genes are adjacent in the *Phytophthora* genome sequences whereas the syntenic region in the *Hpa* genome is simply missing the two genes, together with an adjacent nitrate transporter (Baxter et al., 2010). The loss of ability to assimilate nitrogen and sulphur was also observed in the biotrophic oomycete pathogen *A. laibachii* (Kemen et al., 2011). The same three nitrate assimilation genes are also present as a cluster in saprophytic fungi but are deleted in the obligate rust fungi *Melampsora populina-larici* and *Puccinia graminis f.sp. tritici* (Duplessis et al., 2011) and the obligate powdery mildew fungi *Blumeria graminis*, *Erysiphe pisi*, and *Golovinomyces orontii* (Spanu et al., 2010). The observed independent loss of genes in the nitrogen and sulphur assimilation pathways in various pathogens is a very important finding as it implicates reasons behind their obligate biotrophy.

| | *P. sojae* | *P. ramorum* | *P. infestans* | *H. arabidopsidis* |
|---|---|---|---|---|
| **Nitrate reductase *** | Ps140563 | Pr71442 | PITG_13012.1 | - |
| **Nitrite reductase** | Ps140562 | Pr76696 | PITG_13013.1 | - |
| **Glutamine synthetase** | Ps109140<br><br>Ps109139 | Pr72153<br><br>Pr72154 | PITG_14180.1<br><br>PITG_14179.1 | Ha802420 |
| **Glutamate synthase (NADH)** | Ps135530 | Pr72102 | PITG_07380.1 | Ha805196 |
| **Glutamate synthase (Ferridoxin)** | Ps130831 | Pr78125 | PITG_12037.1<br><br>PITG_16280.1 | Ha812981 |
| **Glutamate dehydrogenase** | Ps108919 | Pr71959 | PITG_07671.1 | Ha805610; Ha806617 |
| **Adenylsulfate kinase ATP sulfurylase Pyrophosphotase** | Ps112102 | Pr79353 | PITG_04010.1 | Ha813786 |
| **Phosphoadenosine reductase** | Pr74880 | Ps156997 | PITG_04601.1 | Ha809449 |
| **Sulfite Reductase** | Ps139493<br><br>Ps139488 | Pr71878<br><br>Pr81882 | PITG_19263.1<br><br>PITG_18187.1 | - |
| **Cysteine Synthetase** | Ps109172<br><br>Ps109175 | Pr71225<br><br>Pr71224 | PITG_12727.1<br><br>PITG_12725.1 | Ha814750 |

**Table 4.18: Gene IDs for nitrogen and sulphur assimilation enzymes in *Phytophthora* and *Hpa. P. infestans, P. sojae and P. ramorum* genes taken from (Baxter et al., 2010), and *Hpa* genes identified through reciprocal best BLAST. * Other enzymes in the nitrogen metabolism pathway (KEG m00910) that are present in the *Hpa* gene models include glycine synthase and carbonate hydrolase; L-Glutamate: ammonia ligase is also present, but this is involved in other metabolic pathways. The presence of these other enzymes may provide insight into the form in which nitrogen is taken up from the host.**

### *4.2.5.7 Virulence related genes*

We also observe a general trend of fewer genes implicated in virulence functions in *Hpa* compared to *P. sojae* and *P. ramorum* (table 4.19). There are less than half of the number of predicted effectors in *Hpa* compared to *P. sojae* and *P. ramorum,* although this observation may be biased towards the prediction method used in Tyler et al., 2006, whose method to predict RxLR encoding was primarily based on *Phytophthora* gene sequences and may thus not capture the full nature of effectors in *Hpa*. One example of a published effector gene that is not in the set of 141 high confidence effectors, but instead in the list of 272 less plausible effector genes, is *ATR5* (Bailey et al., 2011).

| **Gene product** | ***H. arabidopsidis*** | ***P. sojae*** | ***P. ramorum*** |
|---|---|---|---|
| Aspartyl proteases | 13 | 13** | 14** |
| Cysteine proteases | 18 | 29** | 35** |
| Glycosyl hydrolases | 91 | 125 | 114 |
| Endoglucanases (EGL12) | 8 | 10 | 8 |
| Polygalacturonases | 3 | 25 | 16 |
| Pectin methyl esterases | 3 | 19 | 15 |
| Cutinases | 2 | 16 | 4 |
| Chitinases | 3 | 5 | 2 |
| Phospholipases | 21 | 31** | 28** |
| Nonribosomal peptide synthetases | 1* | 4 | 4 |
| Polyketide synthases | 13 | 1 | 1 |
| Cytochrome P450's | 14 | 25 | 24 |
| ABC Transporters | 73 | 140 | 135 |
| NPP1 like | 21 | 39** | 59** |
| Elicitins | 16 | 40 | 57 |
| RxLR Effectors | 141 | 335** | 309** |
| Crinklers | 22 | 100 | 19 |

**Table 4.19: Copy numbers of annotated *Hpa* genes implicated in pathogenesis. *P. ramorum* and *P. sojae.* Figures from (Baxter et al., 2010), *Hpa* figures recalculated using InterProScan annotation: Cysteine proteases: Superfamily SSF50494; Glycosyl hydrolases: KEGG Ko01xxx; Endoglucanases : Superfamily SSF50685; Polygalacturonases: Pfam PF00295; Pectin methyl esterases: Pfam PF00295; Cutinases: Pfam PF01083; Chitinases: GO:0004568; Phospholipases : SuperFamily**

**SSF52151, SSF56024, SSF48537); polyketide synthases: SM00822, SM00823, SM00825, SM00827. SM00829; Cytochrome P450's: Pfam PF00067; ABC TRANSPORTERS: Pfam PF01061, PF00005; NPP1 like: Pfam PF05630; Elicitins: Pfam PF00964 .**Effectors for (Haas et al., 2009).**

## 4.3 *Hpa* genome browser

With the availability of the *Hpa* v8.3 assembly, various gene models and the wealth of expression data available, it is important to be able to access such information in a user friendly manner.

The Gbrowse generic genome browser system (Stein et al., 2002) framework was used to implement the *Hpa* genome browser (fig 4.8). The default protocol for uploading GFF was followed. The Illumina DNA coverage data was converted from MAQ pileup format to the WIG format to optimise loading times. The *Hpa* Emoy2 v8.3 Genome Browser (http://gbrowse2.tsl.ac.uk/cgi-bin/gb2/gbrowse/hpa_emoy2_publication/) visualises the v8.3 assembly and annotation. The tracks available for viewing are: global GC content (displays the GC content over the entire scaffold), predicted effectors (displaying the effectors predicted from the *Hpa* Jamboree held in Virginia Tech in 2007), 454 ESTs (extracted from 3 d.p.i infected *A. thaliana* WS eds1-1), *Hpa* Emoy2 Illumina cDNA (extracted from 7 d.p.i infected *A. thaliana* WS eds1-1), Sanger ESTs (extracted from spores), local 3-frame-forward, 3-frame-reverse and 6-frame translation, *Hpa* Emoy2 transcript models, *Hpa* DNA coverage by Illumina reads, polymorphisms (heterozygous INDELs and SNPs) and restriction sites. The genome browser has a search facility for genes, effectors and scaffolds, displays customisations and facilities the uploading of user defined GFF3 tracks.

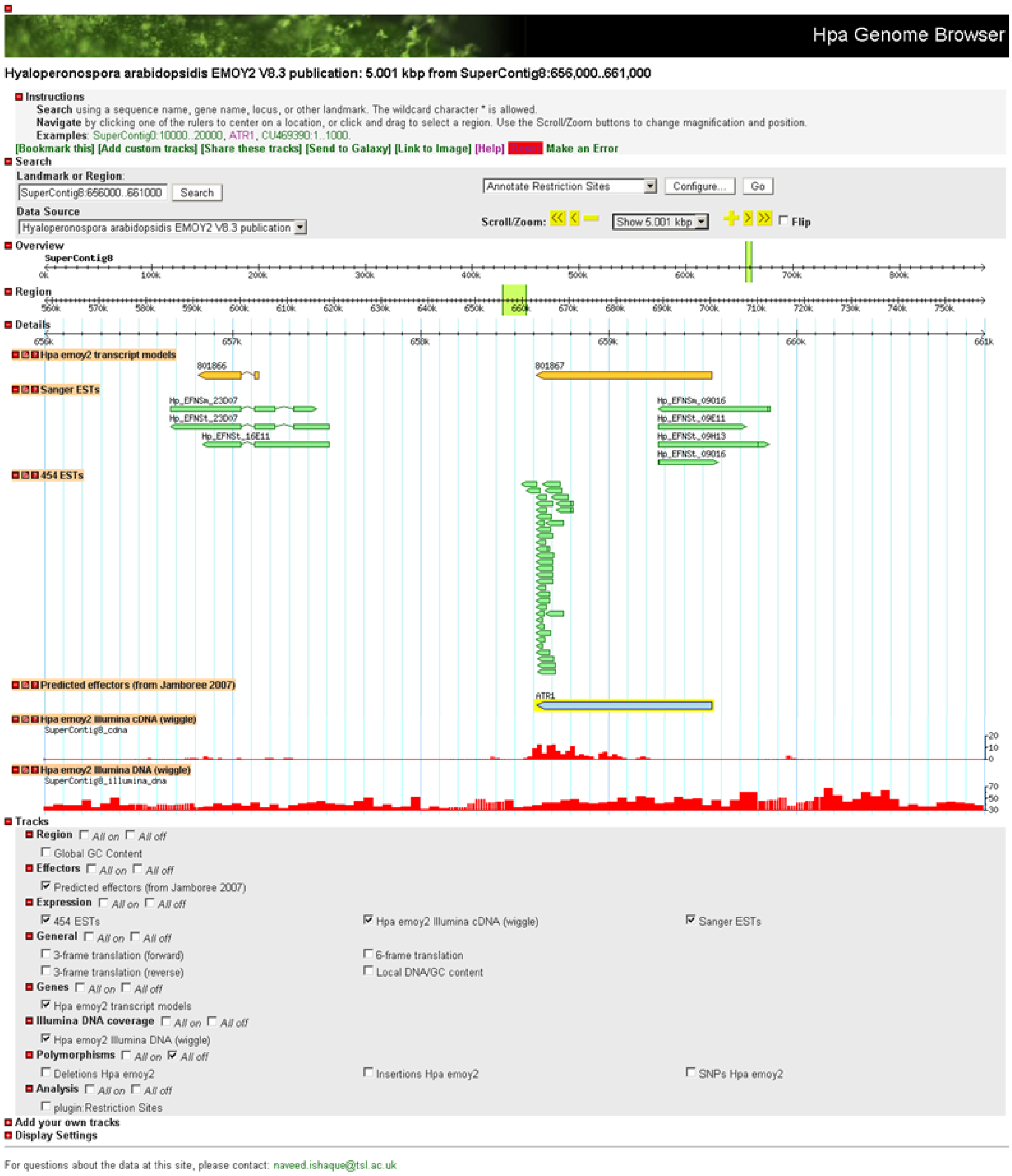



**Figure 4.8:** ***Hpa*** **Emoy2 v8.3 genome browser**

## 4.4 Summary

In this chapter we described the progressive improvement of the *Hpa* Emoy2 v8.3 assembly gene models from the v1 gene models containing many spurious genes with moderate evidence of expression, to the v3 gene models, which support 92% of the alignable ESTs, and have a higher percentage support of Illumina sequenced cDNA than we saw in the gold standard gene models of *A. thaliana.*

I successfully trained and used various gene prediction programs and integrated them into existing gene models using novel and robust methods. I also report, for the first time in a genome publication, methods used to evaluate gene model robustness using various sources of evidence.

The resulting annotations have resulted in interesting observations including the large proportion of the genome encoding secreted proteins, shared and unique metabolic pathways between *Hpa* and *Phytophthora,* incomplete nitrogen and sulphur assimilation pathways that may be the reason for obligate biotrophy, and the reduced number of genes encoding pathogen-related proteins compared to *P. ramorum* and *P. sojae*.

A genome browser was established to allow easy viewing of genomic regions, genes and expression data.

Other elements pertaining to the biology of *Hpa* that arose from the establishment of the *Hpa* gene models are described in Baxter et al. (2010).

# Chapter 5 – Use of Illumina sequencing to investigate signatures of evolution in *Hpa*

## 5.1 Introduction

In the previous chapter we made use of EST and Illumina expression data to improve the *Hpa* Emoy2 gene models. This improvement in gene models allows us to perform comparative genomics analysis using the Illumina sequence data of 7 other isolates (Cala2, Emco5, Hind2, Maks9, Noco2 and Waco9 sequenced at TSL and Emwa1 provided by Prof Brain Staskawicz).

Two *Hpa* effectors, *ATR1* and *ATR13*, have been shown to have a high level of nucleotide sequence variation between different races, leading to amino acid substitutions, and appear to be under positive selection (Allen et al., 2008; Rehmany et al., 2005). It is hypothesised that sequence variation is a result of selection pressures exerted by interaction with the plant immune system, i.e. recognition of these effectors by *A. thaliana* resistance genes *RPP1* and *RPP13* (Sohn et al., 2007).

In this chapter we will examine sequence polymorphism in the candidate *Hpa* effectors and other secreted proteins. We expect to see a high level of sequence polymorphism in genes of *Hpa* that are involved in interactions with the resistance genes of *A. thaliana*. We will further investigate the most polymorphic secreted protein families to detect potential effector candidates that do not carry the RXLR motif.

## 5.2 Method development

While there are a number of programs that are able to identify variation from second generation sequencing data, there are currently no computation tools that are able to make direct inferences about evolution from second generation sequencing data. It has been shown that *Hpa* effectors show signs of positive selection, so any method development comprises a significant contribution to the field of second generation sequencing analysis in the light of evolution. Firstly, I will describe the development of the algorithm, VariTale, which I use to make inferences about selection pressures acting on genes by processing variation predictions using second generation sequencing by performing tests of neutrality and selection.

### 5.2.1 Pipeline

The pipeline has been implemented as a set of Perl scripts. I decided that for this analysis the ability to create elaborate data structures, possible through object oriented programming languages, was not a necessity in this case. Scripts generated using high level programming languages such as Perl can be more readable and thus can be easily understood and modified. Perl has many existing libraries for data parsing and manipulation that facilitate quicker code generation and modularisation of the program. There are also well established Perl distributions for many platforms allowing for portability of code.

#### *5.2.1.1 Input*

The input data for the VariTale are:

- Reference sequence file in FASTA format
- Gene model file in GFF3 format
- Sequence alignment file in BAM format
- SNP and INDEL file in VCF-like format

The input file formats chosen are the *de facto* standard for each data type, apart from the SNP and INDEL data. In contrast to other file formats, there is still no widely adopted sequence variance call format. However, the variant call formats all share a similar layout, tab delimited, and have a minimum of 4 fields: chromosome/contig, position, reference base and variant call. These are the first 4 data elements of each line of the VCF format. I used these 4 data elements and describe the data parsing as VCF-like. The VCF-like parsing allows for input from very popular variant call formats including SAMtools VCF (Li et al., 2009a) and BCF (Danecek et al., 2011) and GATK (McKenna et al., 2010).

As with many data processing programs, the final output is highly reliant on the quality of input data. In previous chapters we have shown significant improvement in the *Hpa* Emoy2 genome assembly and gene models, which should translate to more reliable downstream analysis using VariTale.

### *5.2.1.2 Output*

VariTale is currently a 3 stage pipeline with distinct outputs at each stage. Stage 1 is the minimal pre-processing required for either stage 2 or 3.

#### 5.2.1.2.1 Stage 1

The first stage of the pipeline involves processing the input data (reference sequence, gene model co-ordinates, sequence alignments and variant calls) for each individual race. The data is parsed as follows:

- The FASTA reference sequence contigs are parsed as a Bioperl BIO::SeqIO object
- The GFF3 gene models are parsed using Bioperl Bio::Tool::GFF, and then stored in single level hash data structures for:
    - Gene direction
    - Exon start and end coordinates (stored as an array of elements)
    - UTRs (this is currently set to a fixed length of 250 nt as the UTRs for the gene models have not been predicted)
- The BAM alignment file is parsed using the Bioperl Bio::DB::Sam object
- The VCF-like variant calls split into SNPs and INDELs and are then parsed into a 2 level hash data objects in the format (using Perl nomenclature):

The output at this stage is a FASTA file containing sequences for the gene models (modified to incorporate the SNPs for each particular race) with additional statistics. The additional statistics are printed into the FASTA sequence identifier and are derived from the gene or the resequencing data. The statistics derived from the gene are the gene length in nucleotides, the GC content of protein coding regions, the orientation of the gene on the contig and the number of exons. The output derived from the resequencing information includes coverage statistics, number, effect and types of SNPs and number and effect of INDELs.

The mean read depth of coverage over coding regions and the percentage of the coding regions covered by reads (breadth of coverage) for each gene is extracted from the BAM alignments using the Bio::DB::Sam object. More important than the mean coverage, however, is to compare the observed average coverage for each gene with an expected coverage (calculated as the mean of all read coverage over all coding regions).

To determine whether the mean read depth of coverage per gene follows a normal or Poisson distribution over varying read depth, 1, 2, 4 and 8 lanes of Emoy2 sequence data were aligned to the v8.3 assembly. The average read coverage over the coding regions of all genes that are 100% covered by reads was extracted. For 1, 2, 4 and 8 lanes of sequence data we observed 5x, 9x, 20x and 40x read depth coverage. I compared the observed mean read depths to a set of read depths following normal and Poisson distributions, generated from the mean and variance observed in the set of mean read depths, using Quantile-Quantile-plots (QQ-plots). I then measured the goodness of fit using linear regression in the form of the Adjusted $R^2$ value. At low read depths, the QQ-plots of the observed mean read depths with the normal and Poisson generated sets were hard to distinguish by eye, but the adjusted $R^2$ value favoured the observed data following the Poisson distribution (fig 5.1; fig 5.2; fig 5.3; fig 5.4). At the highest read depth tested, the observed mean read depths followed the Poisson generated dataset much better than the normal generated dataset (fig 5.1; fig 5.2; fig 5.3; fig 5.4). I also compared the observed cumulative distribution function (CDF) to the normal and Poisson CDF. It was observed that with increasing read coverage that the data better fit the Poisson CDF compared to the expected normal CDF (fig 5.1; fig 5.2; fig 5.3; fig 5.4). Therefore, I decided to base calculations of deviance from expected read coverage on a Poisson distribution. In the QQ-plots and CDF comparisons for observed data, with average read depths of 20x and 40x, shoulders were observed in the graph at the points of half the average read depths (fig 5.3; fig 5.4). These shoulders support the idea that hemizygous genes can be identified from the read coverage.

Given that the observed mean read depth follows a Poisson distribution, the VariTale pipeline calculates the Poisson probability density function for each gene. With this probability density statistic, the user can infer whether the gene is present as a single copy (if the probability density function is within the e.g. 95% confidence interval), the gene underwent a duplication event (if the probability density function is above e.g. 97.5%), or a single or both parental haplotype copies of the gene are lost, truncated or pseudogenised (if the probability density function is below e.g. 2.5%). In the case of complete loss of both parental haplotypes, we would simply observe no read coverage. However it is less trivial to distinguish between the loss of a single parental haplotype and a gene truncation. For the loss of a single parental haplotype, we would expect the entire gene to be covered by reads but at a lower than expected mean coverage (and a lower probability density). In case of a truncation we would expect that the lower average read coverage is due to parts of the gene not being covered by reads.

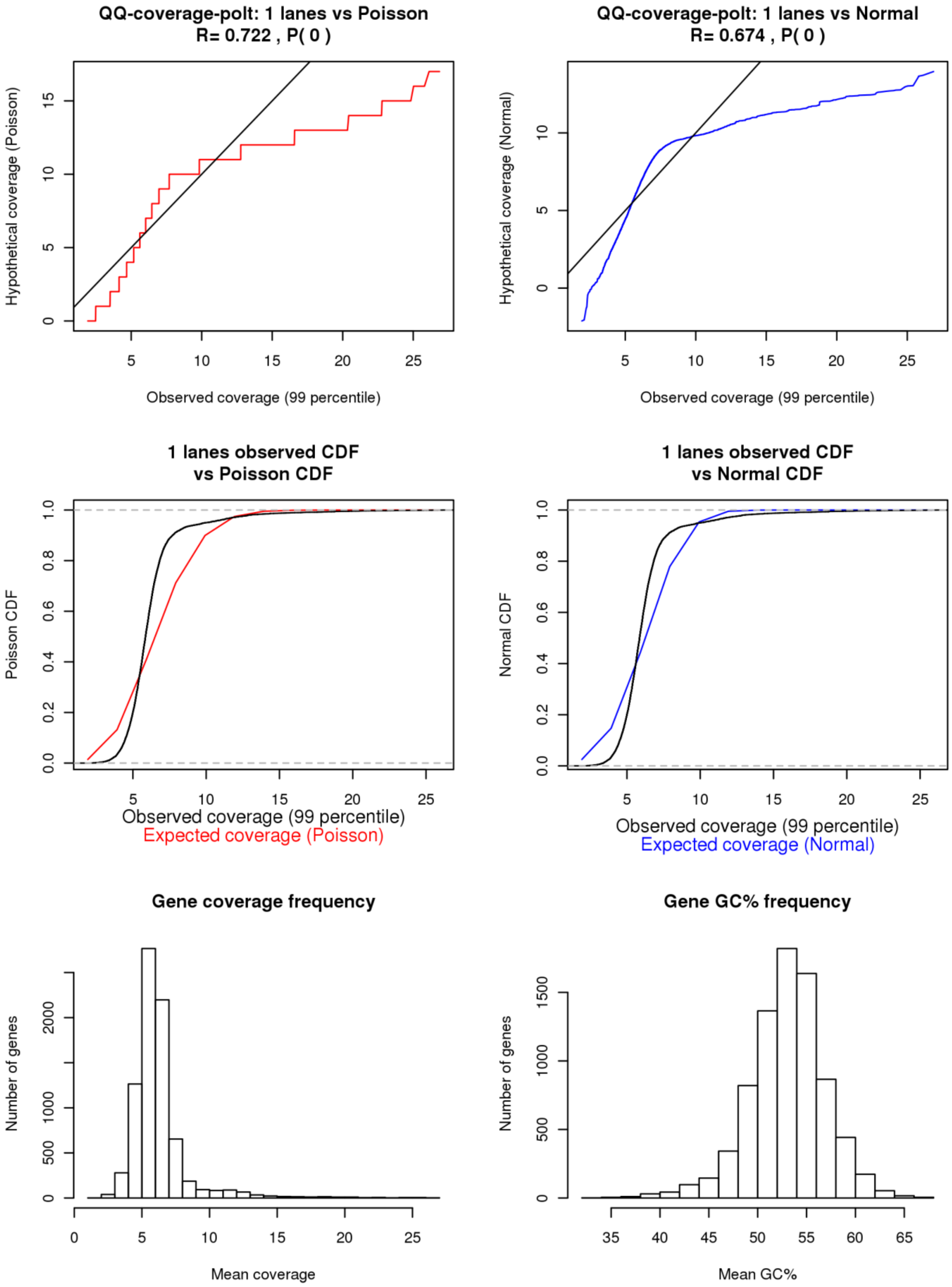



**Figure 5.1:** QQ-plots of likelihood of belonging to Poisson/Normal distribution for 5x coverage

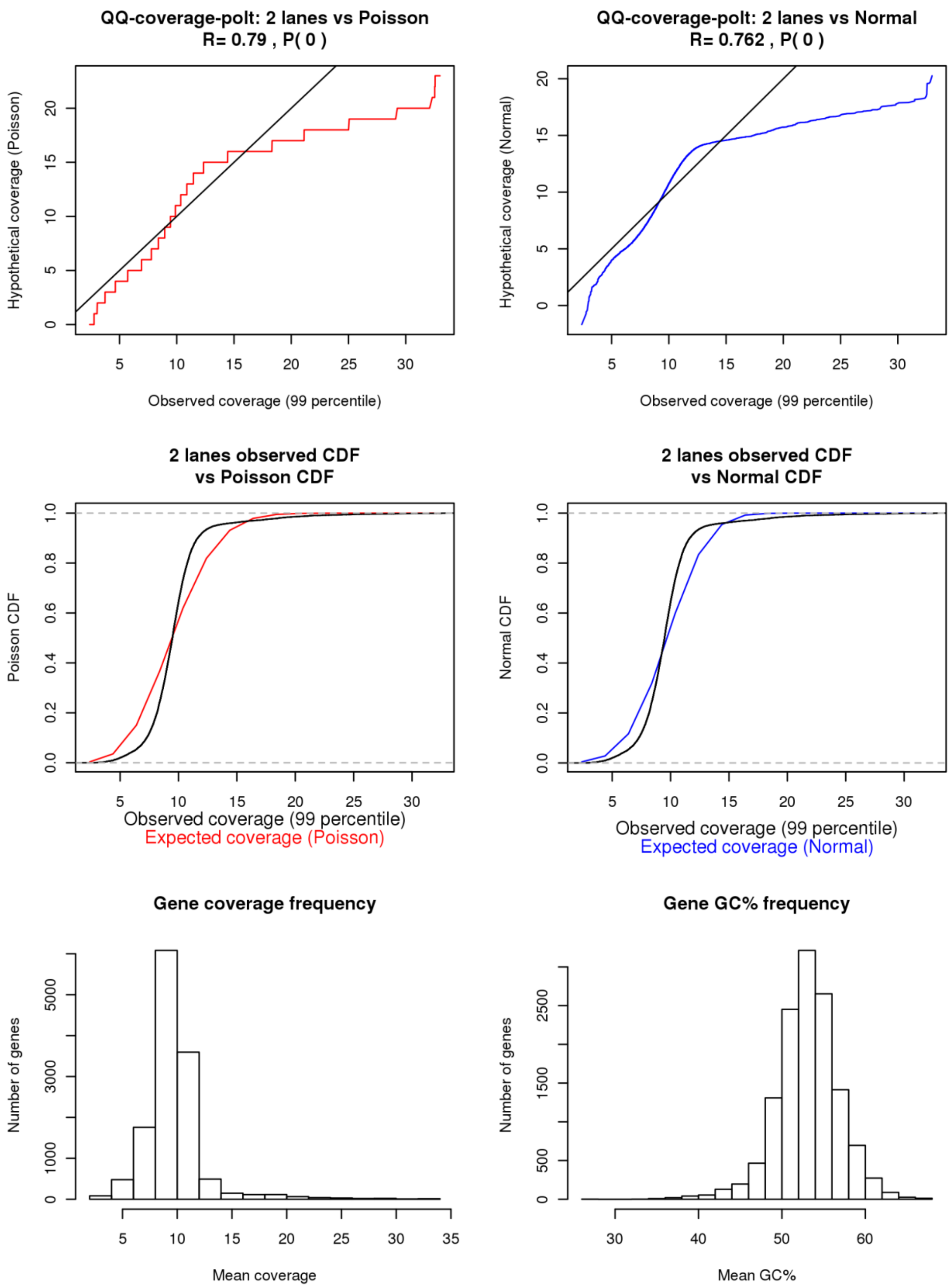



**Figure 5.2:** QQ-plots of likelihood of belonging to Poisson/Normal distribution for 9x coverage

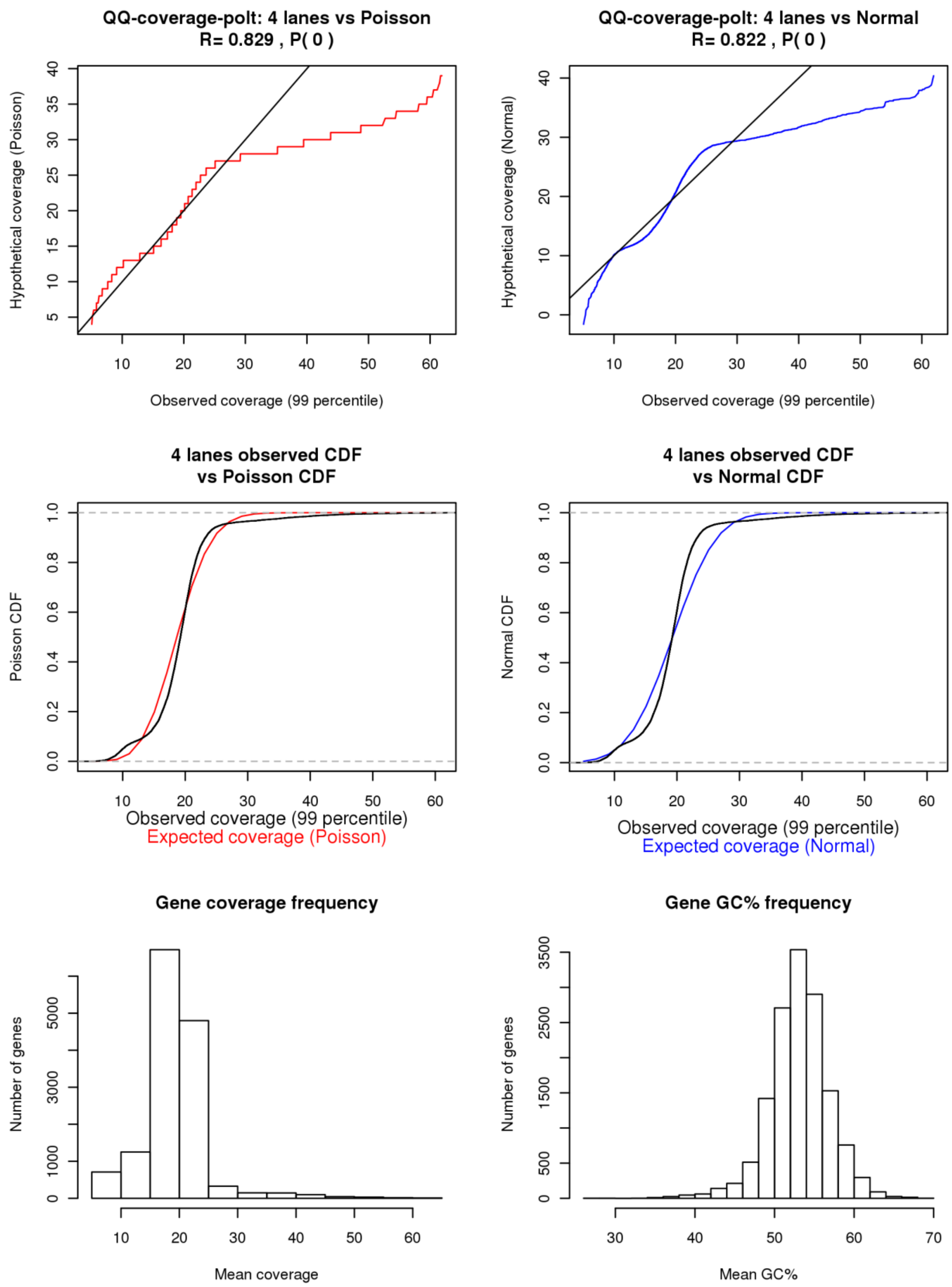



**Figure 5.3:** QQ-plots of likelihood of belonging to Poisson/Normal distribution for 20x coverage

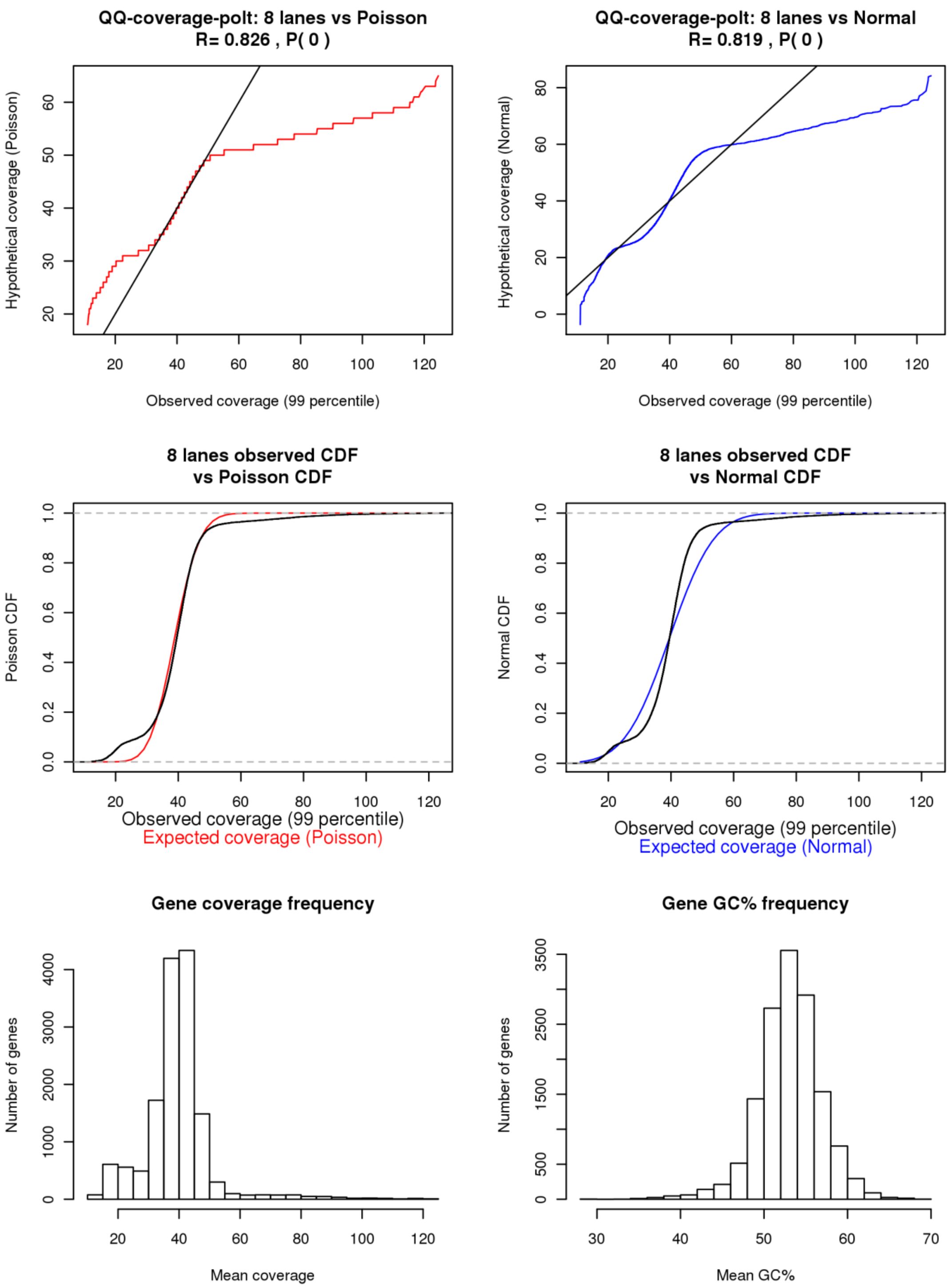



**Figure 5.4:** QQ-plots of likelihood of belonging to Poisson/Normal distribution for 40x coverage

VariTale also generates several SNP statistics. For each gene, the homozygous and heterozygous SNPs are reported on the 5' and 3' UTR (or for a fixed distance and downstream the gene if the UTRs are unknown), exons and introns. It is important to also investigate the up and downstream regions of genes as mutations may affect the transcriptional regulation of the gene rather than the gene sequence itself. Recording SNPs that are within the introns of genes are also important as this can support inferences of gene evolution, for example, a gene is likely to be under stabilising selection if a higher number of SNPs is observed in intronic regions compared to exonic regions. For each SNP on the exon, the codon, on which the SNP lies, is extracted and the amino acid encoded by this codon determined. For each gene, the number of SNPs causing synonymous and non-synonymous mutations is reported. In addition, heterozygous SNPs may lead to the same gene encoding 2 different protein sequences. These SNPs are reported and I will refer to them as 'heterozygous non-synonymous' SNPs. Some of the most drastic effects of SNPs on gene function are those affecting the start and end of the gene sequence. Any SNP that mutates the start codon or introduces a premature stop codon (and hence truncated protein) is also reported.

Up until now the majority of genome variation that has been studied using second generation sequencing include SNPs, CNVs and large scale chromosomal rearrangements. Although INDEL prediction has existed in the earliest of second generation sequence aligners, their implications have not been analysed as routinely as the above. I believe this is primarily due to the non-trivial nature of predicting INDELs and analysing their effect on genes, especially where INDELs result in frame shifts. For each gene, the homozygous and heterozygous INDELs are reported on the 5' and 3' UTR (or for a fixed distance and downstream of the gene if the UTRs are unknown), exons and introns.

As mentioned previously, it is interesting to investigate the effects of INDELs at the protein coding level. For example, a deletion of 2 nucleotides at the beginning of a gene will lead to a frame shift downstream of that deletion, dramatically changing the amino acid sequence encoded by the gene. However, while the frameshift has a dramatic effect on the protein coding sequence of the gene, it is possible that a downstream INDEL can correct the reading frame. For instance, a deletion of 2 nucleotides in one region of the gene may be complemented by an insertion of 2 nucleotides elsewhere in the sequence. In this case we would observe a frame shift in the region lying in between the deletion and the insertion, while the beginning and the end of the gene sequence is retained, including the start and stop codon. For this reason I also decided to report for each gene the 'net INDEL length', which is the sum of all the INDELs predicted in the coding region of the gene. When the 'net INDEL length' is exactly divisible by 3 we are likely to observe a conservation of the start and stop codons, while the internal sequence may have undergone a frame shift. The conservation of the start and stop codon may indicate selection pressures that are preserving the presence of the gene, while exerting selection pressure to modify the gene function (due to the effects of the internal frame shift caused by the INDELs). Thus, evolutionarily INDELs may be powerful tools to generate sequence diversity and should generally have much greater effects on protein functions than SNPs. We also report the numbers of INDELs that are exactly divisible by 3 over gene exons as this indicates exact codon loss and whether all the INDELs over the coding region are exactly divisible by 3.

To summarise, in stage 1, for each gene we display the following information:

- The gene nucleotide sequence, modified with the predicted SNPs
    - Heterozygous SNPs are displayed as IUPAC ambiguous codes
- Length
- GC content
- Direction
- Coverage
    - Mean coverage
    - Percentage covered
    - Cumulative distribution function of the observed mean coverage belonging to the expected Poisson distribution with mean equal to the observed mean over all genes
- SNPs
    - Number of homozygous and heterozygous SNPs
        - over 5', exons, introns, 3'
    - Number of synonymous SNPs
    - Number of non-synonymous SNPs
    - Number of heterozygous non-synonymous SNPs

- INDELs
    - Number of homozygous and heterozygous INDELs
        - over 5’, exons, introns, 3’
    - Net INDEL length
    - Number of INDELs exactly divisible by 3
    - If only INDELs exactly divisible by 3 are observed

**5.2.1.2.2 Stage 2**

The second stage of the pipeline performs population genetic analysis on the DNA sequences of genes obtained at the within species population level. The inputs are at least 3 different FASTA files produced by the first stage of the pipeline (i.e. 3 different races sequenced and analysed using the first stage of the analysis pipeline) and then performs several tests of neutrality using DnaSP v5 (Librado and Rozas, 2009). This approach allows variation at a population level and divergence from neutrality for each gene to be analysed. This is a novel approach to analysing effector genes, and may reveal further insights to effector biology and other biological mechanisms.

Before the tests of neutrality are performed the data have to be pre-processed. To recall, at the end of stage 1, the gene sequences with SNP modifications are printed. The heterozygous SNPs are displayed as IUPAC ambiguous codes. The DnaSP algorithm requires a nucleotide sequence for batch input processing and cannot currently disambiguate heterozygous calls to their parental haplotypes. Where the gene contains a single heterozygous SNP, the parental haplotypes can be determined easily.
When the gene contains 2 or more heterozygous SNPs, discerning the parental haplotypes is less trivial. The parental haplotypes can be reconstructed exactly in cases where the heterozygous SNPs are clustered over an area shorter than the length of the read. In these cases the parental haplotypes can be determined by observing SNP linkage over the reads with the heterozygous SNPs. The parental haplotypes can also be reconstructed exactly when the heterozygous SNPs are exactly the distance of the fragment apart, i.e. the SNP linkage can be observed by looking at the read pairs on which they occur. However the majority of the heterozygous calls do not lie within these 2 scenarios. To reconstruct the parental haplotypes from population data we use PHASE v2.1.1 (Stephens and Scheet, 2005). Although many algorithms exist for estimating haplotypes from genotype data, PHASE is one of the few that considers the decay of linkage equilibrium with distance and

the order and spacing of genotype markers. The SNPs and their positions are extracted and a PHASE run is performed on them (with parameters set at 100 iterations, a thinning interval of 1 and a burn-in of 100). From the output of PHASE, the parental haplotypes are extracted and printed in a FASTA file for use with DnaSP. DnaSP's batch processing mode is used to process the unphased haplotype FASTA gene sequences.

For each gene in each race with 100% sequence coverage and a coverage Poisson CDF of less than 97.5% (i.e. not within the 95% confidence interval of being a single copy gene based on a Poisson distribution),the statistics reported are:

- The number of segregating sites (S)
- The total number of mutations (Eta)
- The number of haplotypes
- Statistical test of neutral theory of molecular evolution (Kimura, 1983)
    - Tajima's D (Tajima, 1989)
    - Fu Li's D* (Fu and Li, 1993)
    - Fu Li's F* (Fu and Li, 1993)
    - Fu's Fs (Fu, 1997)

S, Eta and the number of haplotypes provide insight into the amount of variation seen in the gene and how varied alleles are in the sample population. The various neutrality tests will be able to report whether genes are evolving under neutrality, or if there are selective pressures being applied. While some of these tests may appear to be redundant, as high throughput analysis of this nature has not been performed previously we are currently unaware of the effectiveness and redundancy of each of these tests.

#### 5.2.1.2.3 Stage 3

The third stage of the pipeline performs phylogenetic and evolutionary analysis on the DNA sequences of genes obtained at the between species population level. The inputs are at least 3 different files produced by the first stage of the pipeline (i.e. 3 different races sequenced and analysed using the first stage of the analysis pipeline). PAML v4.0 (Yang, 2007) is then used to perform several likelihood ratio tests of the observed sequence variation following various evolutionary models. Although these tests are specifically for between species data where there is no gene flow between species, they can be used for within species data as suggestive evidence as used by Haas et al. (2009).

PAML requires 3 input files:

- Sequence alignment of the input
- A PAML control file, outlining the location of input and output files and processing to be carried out
- A phylogenetic tree of the organisms whose sequences are being analysed

For each gene, the sequence for each isolate is extracted for each organism and converted into a PHYLIP format (Felsenstein, 1989). Unlike stage 2, PAML is able to process sequences with ambiguous nucleotides, so there is no need for the pre-processing step in which the parental haplotypes are resolved. Only the full sequences are printed in the PHYLIP file as only genes that are 100% covered by reads and without INDELs are processed.

The control file generated declares the input file, output file and location of the tree file. It also defines the models to be run. 3 control files are produced to run:

- codeml (codon evolution) with evolutionary models
    - M0 – one ratio (uniform selective pressure among sites)
    - M3 – discrete (variable selective pressure among sites)
    - M1a – nearly neutral (variable selective pressure, but no positive selection)
    - M2 a– positive selection (variable selective pressure, with positive selection)
    - M7 – beta (beta distributed selective pressure)
    - M8 – beta with ω (dN/Ds or Ka/Ks) > 1 (beta plus positive selection)
    - pairwise comparison
- codeml with evolutionary model m8a (beta with ω = 1)
- yn00 (protein evolution)

The codeml evolutionary model analysis requires a phylogenetic tree file of all the organisms analysed with VariTale (super tree). The tree has to be in Newick (New Hampshire tree) format. PAML is unable to process trees that contain additional tips, so for each gene a 'pruned' tree is generated. The super tree is parsed as a Bio::Phylo::IO object, all tips of the tree corresponding to input sequences are kept, and unnecessary tips are removed. Any unbranched internal structures are removed producing a balanced tree that is a subtree of the original super tree.

This section of the processing is computationally very time consuming but has been optimised for running on a cluster managed by the LSF7 job management system. Alternatively, processing can very easily be modified to work on the PBS Torque job management system as well.

For each gene in each race with 100% sequence coverage and a coverage Poisson CDF of less 97.5% (i.e. not within the 95% confidence interval of being a single copy gene based on a Poisson distribution), the statistics reported are:

- ω as calculated by codeml pairwise comparisons
- ω as calculated by yn00
- Log likelihood difference between models (and their significance based on χ ):
    - M3 – M0 (testing for a variable ω among sites instead of a single ω for all sites)
    - M2a – M1a (testing for positive selection rather than nearly neutral evolution)
    - M8 –M7 (testing for the existence of sites with ω = 1 rather than ω < 1 for all sites)
    - M8 –M8a (testing for sites with ω >> 1 instead of ω = 1 as an indication of positive selection)

The codeml evolutionary model analysis provide insights into genes that have a sections of the gene under positive selection, while the dN/dS calculation try to evaluate if the gene as a whole is under positive selection. Previous analysis of a similar nature has been done before (Haas et al., 2009), but there method was limited to use of just yn00. Since systematic analysis of this high throughput nature has not been performed before, it is important to consider various, possibly redundant, models to evaluate their effectiveness.

#### 5.2.1.2.4 Output format for data comparison

Once all the processing stages have been completed, all the statistics generated are parsed into as simple tab delimited format:

[Gene]  [Program]  [Statistic]  [Value]  [Note/Significance]

Formatting the data in this way facilitates querying the data for different classes of genes, and allows for more efficient comparisons.

## 5.3 Results and discussion

The first stage of VariTale requires:

- The FASTA reference sequence contigs
- The GFF3 gene models
- The BAM alignment file
- The VCF-like variant

In previous chapters I have described the most recent *Hpa* Emoy2 genome assembly (v8.3) and the most recent gene models (v3). Here, I will describe how the alignment and variant files were produced.

### 5.3.1 Alignment

Illumina sequenced reads of 8 races of *Hpa* (Cala2, Emco5, Emoy2, Hind2, Maks9, Noco2 and Waco9 that were sequenced in house, and Emwa1 provided by Prof Brian Staskawicz), were analysed for quality using FastQC (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/) (appendix figures 2.1). This analysis revealed certain quality issues:

- The per base quality drops drastically in the last third of the read for sequencing runs on the before the implementation of the GA pipeline v1.3 (before ID71)
- The Emwa1 reads have high levels of Illumina paired end sequencing primer contamination
- The reads for each sequenced race have between 2% and 25% PCR duplication

Despite the per-base quality decrease in the last third of the read, the average read quality of the reads have a single peak around a Phred scaled quality score 30, which implies an overall error rate of 0.1%. Therefore, the reads were not filtered or trimmed before alignment, but instead, at a later point, I modified the alignment parameters to soft-trim bad quality trailing bases and filtered the PCR duplicates post-alignment, using the Samtools rmdup command.

The reads were converted from Solexa and Illumina to Sanger quality scores prior to alignment. The reads were aligned to the *Hpa* v8.3 assembly using BWA v0.5.8c (Li and Durbin, 2009). In addition, BWA's read trimming was used to clip trailing nucleotides with a quality score of less than 10. The aligned reads were converted to BAM files using BWA's 'sampe' command. I then extracted all the reads that did not align as pairs, and reads that did not align. After the initial BWA alignment there was a second round of alignment using a more sensitive aligner, Stampy v1.0v11 (Lunter and Goodson, 2011) using its 'sensitive' mode to align these extracted reads. Using Stampy on top of the BWA alignment increased the percentage of reads aligning by 3.72%, of which 1.31% mapped as pairs (table 5.1).
There was a notable difference in the percentage of reads aligning to the *Hpa* v8.3 assembly between races. More than 90% of the reads for Emoy2, Hind2, and Noco2 aligning to the *Hpa* v8.3 assembly when Stampy was used alongside BWA, which suggests that these are most similar to the reference race, Emoy2, out of the races studied (although it may be the case that Noco2 is contaminated with another *Hpa* race, as this level of similarity with Emoy2 was unexpected – if this is the case, the effect on downstream analysis would reduce the number of true positives, which is more desirable than increasing the number of false positives). Less than 50% of the Cala2 reads aligned to Emoy2 reference. We performed a Velvet assembly of the Cala2 reads which did not align to the reference sequence, and performed a BLAST search against the NR database, revealing significant *Xanthomonas* contamination and minor *Pseudomonas* contamination in the data. I believe this was due to a sample contamination rather than library prep or run contamination, because other libraries prepared at the same time were free of contamination, as were other samples sequenced on the same flowcell.

Once all the alignments were completed, they were merged, sorted and PCR duplicates removed using SAMtools v0.1.12a (Li et al., 2009a) rmdup command. It was noticed that as the total number or reads per race increased, so did the read duplicates. This could be due to stochastic error, or may imply that PCR duplicates are more likely to be observed with increasing read depth. It is possible that as the number of sequenced reads increases, the larger possibility of the observation of a sequenced fragment originating from same the genomic position as another sequenced fragment. This would, in fact, be a false positive PCR duplicate and an artefact caused by saturation of the system. This is a known issue of the Samtools rmdup command.

| | | | | All reads aligned | | | | | Reads aligned as pairs | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | BWA | | BWA + Stampy | | | BWA | | BWA + Stampy | | |
| Race | Number of reads | Duplicates | % Duplicates | Number of reads alinged | % of reads aligned | Number of reads alinged | % of reads aligned | % increase in readalignment | Number of reads alinged | % of reads aligned | Number of reads alinged | % of reads aligned | % increase in readalignment |
| **Cala2** | 102,102,496 | 7,702,096 | 7.54% | 41,145,830 | 40.30% | 43,985,611 | 43.08% | 6.90% | 38,106,640 | 37.32% | 38,354,846 | 37.57% | 0.65% |
| **Emco5** | 154,461,266 | 27,129,767 | 17.56% | 134,230,186 | 86.90% | 138,696,656 | 89.79% | 3.33% | 124,677,522 | 80.72% | 125,825,452 | 81.46% | 0.92% |
| **Emoy2** | 99,616,122 | 8,612,719 | 8.65% | 89,460,651 | 89.81% | 93,188,178 | 93.55% | 4.17% | 84,392,579 | 84.72% | 85,159,010 | 85.49% | 0.91% |
| **Emwa1** | 69,182,800 | 1,754,692 | 2.54% | 37,948,025 | 54.85% | 39,482,681 | 57.07% | 4.04% | 35,035,648 | 50.64% | 35,337,502 | 51.08% | 0.86% |
| **Hind2** | 117,257,018 | 22,530,277 | 19.21% | 109,017,009 | 92.97% | 111,987,100 | 95.51% | 2.72% | 101,136,710 | 86.25% | 102,775,426 | 87.65% | 1.62% |
| **Maks9** | 80,982,024 | 10,491,395 | 12.96% | 59,711,490 | 73.73% | 62,385,445 | 77.04% | 4.48% | 55,106,968 | 68.05% | 55,732,234 | 68.82% | 1.13% |
| **Noco2** | 116,719,072 | 27,523,234 | 23.58% | 104,879,442 | 89.86% | 107,737,917 | 92.31% | 2.73% | 97,604,628 | 83.62% | 99,265,106 | 85.05% | 1.70% |
| **Waco9** | 113,350,294 | 10,007,710 | 8.83% | 97,287,035 | 85.83% | 101,737,256 | 89.75% | 4.57% | 89,069,816 | 78.58% | 90,866,580 | 80.16% | 2.02% |
| **Average*** | - | - | 13.33% | - | 81.99% | - | 85.00% | 3.72% | - | 76.08% | - | 77.10% | 1.31% |

**Table 5.1: Number and percentage of reads from 8 *Hpa* races aligning to the *Hpa* v8.3 assembly. There is a large variation in the percentage of reads aligning to the v8.3 assembly for each race which is indicative of how similar the sequenced race is to the Emoy2 reference race. A very similar percentage of reads from Hind2 and Noco2 align to the reference as Emoy2, which indicates these may have very similar genomes. There is a very low percentage of reads aligning from Cala2, which is an artefact due to significant *Xanthamonas* contamination. There is an average of 3.72% of increase in reads aligning using Stampy on top of BWA, of which 1.31% align as pairs.**

*** Average is the calculated mean, excluding Cala2**

## 5.3.2 Variant calling

A list of the variant calls was generated using SAMtools. The variants were filtered using the varFilter script in the SAMtools package. The variant calls were filtered for:

- A minimum root mean squared (RMS) quality of 20 for SNPs [Q = 20]
- A minimum RMS quality of 20 for gaps [q = 20]
- A minimum read depth of 5 [d = 5]
- A maximum read depth of 80 [D = 80]
- A minimum SNP quality of 20 [S = 20]
- A minimum INDEL quality of 20 [i = 20]
- A window size of 3 for filtering dense SNPs [W = 3]

### *5.3.2.1 SNP and INDEL calls*

The 8 sequenced *Hpa* races have approximately 150,000-200,000 SNPs (table 5.2). Excluding the reference race Emoy2, the lowest number of SNPs predicted was for Noco2 (58,175), again suggesting that it is very similar to Emoy2. Hind2, the race with the highest percentage of reads aligning to the reference sequence, has about 3 times more SNPs than Noco2. This was also observed in the number of INDELs predicted in each race where Hind2 had a comparable number of INDELs to Cala2, Emco5 and Waco9, while Noco2 had about 1/3rd less INDELs. I also observed that the number of predicted insertions and deletions followed on average a 1:1 ratio. This indicates that there is no preference of insertions over deletions in *Hpa*.

### *5.3.2.2 Heterozygosity*

The average rate of observing a heterozygous SNP among SNP sites was found to be 25.25% (table 5.2; appendix table 5.4). This is nearly half of the rate of observing a heterozygous INDEL among INDEL sites, which is 44.72%. From this I concluded that ~35% of *Hpa* variation is heterozygous. It is also important to note that the high rates of heterozygous variants in Emoy2 are because of it being the reference strain. Any homozygous SNP would be due to a combination of errors in the reference sequence and error rate in the SNP calling. It was observed that the rate of heterozygous SNPs in Emoy2 (98.95%) was higher than the rate of heterozygous INDELs (87.79%). This is because the *Hpa* v8.3 assembly was corrected for SNP and INDEL errors with iterative variant calls using MAQ. While MAQ is very good at predicting SNPS, the INDEL prediction is not as accurate as the methods used in the BWA + Stampy alignment. Noco2 displayed a very high rate of heterozygosity in SNPs (77.01%) and INDELs (79.14%), making it the most heterozygous *Hpa* race in this study. It was also interesting to note that the rate of heterozygosity in the SNPs of Emwa1 was very low (0.24%), but extremely high in the INDELs (90.28%). The full list heterozygous INDELS on genes can be seen in appendix table 5.7.

| Race | SNPs | Het SNPs | % Het SNPs | INDELs | Het INDELs | % Het INDELs | %Het SNP: % Het INDELS | Insertions | Deletions | Insertions:Deletions | Total Variation |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Cala2** | 208,602 | 35,470 | 17.00% | 25,132 | 7083 | 28.18% | 0.60 | 12,295 | 12,837 | 0.96 | 233,734 |
| **Emco5** | 182,716 | 10,128 | 5.54% | 25,734 | 5672 | 22.04% | 0.25 | 12,859 | 12,875 | 1 | 208,450 |
| **Emoy2** | 54,283 | 53,626 | 98.79% | 9810 | 8612 | 87.79% | 1.13 | 4881 | 4929 | 0.99 | 64,093 |
| **Emwa1** | 146,530 | 350 | 0.24% | 53,259 | 48,084 | 90.28% | 0.00 | 30,916 | 22,343 | 1.38 | 199,789 |
| **Hind2** | 193,858 | 17,178 | 8.86% | 25,692 | 6529 | 25.41% | 0.35 | 12,816 | 12,876 | 1 | 219,550 |
| **Maks9** | 224,793 | 33,181 | 14.76% | 29,366 | 8386 | 28.56% | 0.52 | 14,512 | 14,854 | 0.98 | 254,159 |
| **Noco2** | 58,175 | 44,798 | 77.01% | 16,098 | 12,740 | 79.14% | 0.97 | 7674 | 8424 | 0.91 | 74273 |
| **Waco9** | 209,629 | 59,320 | 28.30% | 26,901 | 10,596 | 39.39% | 0.72 | 13,181 | 13,720 | 0.96 | 236,530 |
| **Average*** | | | 25.25% | | | 37.12% | 0.57 | | | 1.02 | |

**Table 5.2: Table of predicted SNPs and INDELs in the 8 sequenced races of *Hpa*. Het = heterozygous.**

*** The average was calculated as the mean without values from Emoy2, as it is the reference strain, and Emwa1 as this race had very few predicted heterozygous SNPs.**

### *5.3.2.3 Preferential SNP mutation*

For each race, the homozygous SNPs were extracted. A significant balance between reciprocal nucleotide changes was observed, which implies that the *Hpa* genome is under pressure to maintain its nucleotide compositions (tables 5.3). Preferential A/T ⇔ G/C transitional mutation were observed, which accounted for ~65% of all SNPs (outside of Emwa1). A/T ⇔ G/C transitional mutation has been described to be the most commonly observed point mutation and was also seen in 80 accessions of *A. thaliana* (Cao et al., 2011). A reciprocal balance of A/T →C/G with G/C→T:A and A/T→T/A with C/G→G/C was also observed.

| Cala2 | A | C | G | T |
|---|---|---|---|---|
| A | 0 | 9739 | 28,31 | 6027 |
| C | 9420 | 0 | 5007 | 27,803 |
| G | 28,07 | 4983 | 0 | 9546 |
| T | 6159 | 28,17 | 9866 | 0 |
| Total | | | | 173,11 |

| Balance | A | C | G | T |
|---|---|---|---|---|
| A | - | 1.0 | 1.0 | 0.9 |
| C | 0.9 | - | 1.0 | 0.9 |
| G | 0.9 | 1.0 | - | 0.9 |
| T | 1.0 | 1.0 | 1.0 | - |
| Std Dev | | | | 0.0 |

| Preferenc | A | C | G | T |
|---|---|---|---|---|
| A | - | 11.07 | 32.57 | 7.04% |
| C | - | - | 5.77% | 32.34% |
| G | - | - | - | 11.21% |
| T | - | - | - | - |
| Total | | | | 100.00 |

| Emco5 | A | C | G | T |
|---|---|---|---|---|
| A | 0 | 9818 | 28,00 | 6276 |
| C | 9615 | 0 | 4923 | 27,447 |
| G | 27,83 | 4970 | 0 | 9469 |
| T | 6370 | 28,04 | 9756 | 0 |
| Total | | | | 172,53 |

| Balance | A | C | G | T |
|---|---|---|---|---|
| A | - | 1.0 | 1.0 | 0.9 |
| C | 0.9 | - | 0.9 | 0.9 |
| G | 0.9 | 1.0 | - | 0.9 |
| T | 1.0 | 1.0 | 1.0 | - |
| Std Dev | | | | 0.0 |

| Preferenc | A | C | G | T |
|---|---|---|---|---|
| A | - | 11.26 | 32.37 | 7.33% |
| C | - | - | 5.73% | 32.16% |
| G | - | - | - | 11.14% |
| T | - | - | - | - |
| Total | | | | 100.00 |

| Emoy2 | A | C | G | T |
|---|---|---|---|---|
| A | 0 | 51 | 74 | 40 |
| C | 53 | 0 | 31 | 77 |
| G | 67 | 35 | 0 | 45 |
| T | 60 | 70 | 42 | 0 |
| Total | | | | 645 |

| Balance | A | C | G | T |
|---|---|---|---|---|
| A | - | 0.9 | 1.1 | 0.6 |
| C | 1.0 | - | 0.8 | 1.1 |
| G | 0.9 | 1.1 | - | 1.0 |
| T | 1.5 | 0.9 | 0.9 | - |
| Std Dev | | | | 0.1 |

| Preferenc | A | C | G | T |
|---|---|---|---|---|
| A | - | 16.12 | 21.86 | 15.50% |
| C | - | - | 10.23 | 22.79% |
| G | - | - | - | 13.49% |
| T | - | - | - | - |
| Total | | | | 100.00 |

| Emwa | A | C | G | T |
|---|---|---|---|---|
| A | 0 | 12,20 | 12,11 | 13,365 |
| C | 11,64 | 0 | 11,93 | 11,848 |
| G | 11,87 | 11,74 | 0 | 11,423 |
| T | 13,04 | 12,09 | 12,22 | 0 |
| Total | | | | 145,51 |

| Balance | A | C | G | T |
|---|---|---|---|---|
| A | - | 1.0 | 1.0 | 1.0 |
| C | 0.9 | - | 1.0 | 0.9 |
| G | 0.9 | 0.9 | - | 0.9 |
| T | 0.9 | 1.0 | 1.0 | - |
| Std Dev | | | | 0.0 |

| Preferenc | A | C | G | T |
|---|---|---|---|---|
| A | - | 16.39 | 16.49 | 18.15% |
| C | - | - | 16.27 | 16.46% |
| G | - | - | - | 16.25% |
| T | - | - | - | - |
| Total | | | | 100.00 |

| Hind2 | A | C | G | T |
|---|---|---|---|---|
| A | 0 | 9733 | 28,80 | 6399 |
| C | 9637 | 0 | 5184 | 28,715 |
| G | 28,45 | 5128 | 0 | 9601 |
| T | 6428 | 28,70 | 9858 | 0 |
| Total | | | | 176,64 |

| Balance | A | C | G | T |
|---|---|---|---|---|
| A | - | 1.0 | 1.0 | 1.0 |
| C | 0.9 | - | 1.0 | 1.0 |
| G | 0.9 | 0.9 | - | 0.9 |
| T | 1.0 | 1.0 | 1.0 | - |
| Std Dev | | | | 0.0 |

| Preferenc | A | C | G | T |
|---|---|---|---|---|
| A | - | 10.97 | 32.41 | 7.26% |
| C | - | - | 5.84% | 32.51% |
| G | - | - | - | 11.02% |
| T | - | - | - | - |
| Total | | | | 100.00 |

| Maks9 | A | C | G | T |
|---|---|---|---|---|
| A | 0 | 11,04 | 31,30 | 6686 |
| C | 10,69 | 0 | 5537 | 30,603 |
| G | 30,83 | 5497 | 0 | 10,449 |
| T | 6818 | 31,17 | 10,95 | 0 |
| Total | | | | 191,58 |

| Balance | A | C | G | T |
|---|---|---|---|---|
| A | - | 1.0 | 1.0 | 0.9 |
| C | 0.9 | - | 1.0 | 0.9 |
| G | 0.9 | 0.9 | - | 0.9 |
| T | 1.0 | 1.0 | 1.0 | - |
| Std Dev | | | | 0.0 |

| Preferenc | A | C | G | T |
|---|---|---|---|---|
| A | - | 11.34 | 32.43 | 7.05% |
| C | - | - | 5.76% | 32.24% |
| G | - | - | - | 11.17% |
| T | - | - | - | - |
| Total | | | | 100.00 |

| Noco2 | A | C | G | T |
|---|---|---|---|---|
| A | 0 | 801 | 2225 | 504 |
| C | 693 | 0 | 405 | 2031 |
| G | 2108 | 418 | 0 | 662 |
| T | 551 | 2145 | 793 | 0 |
| Total | | | | 13,336 |

| Balance | A | C | G | T |
|---|---|---|---|---|
| A | - | 1.1 | 1.0 | 0.9 |
| C | 0.8 | - | 0.9 | 0.9 |
| G | 0.9 | 1.0 | - | 0.8 |
| T | 1.0 | 1.0 | 1.2 | - |
| Std Dev | | | | 0.1 1 |

| Preferenc | A | C | G | T |
|---|---|---|---|---|
| A | - | 11.20 | 32.49 | 7.91% |
| C | - | - | 6.17% | 31.31% |
| G | - | - | - | 10.91% |
| T | - | - | - | - |
| Total | | | | 100.00 % |

| Waco9 | A | C | G | T |
|---|---|---|---|---|
| A | 0 | 8603 | 24,97 | 5360 |
| C | 8086 | 0 | 4315 | 23,782 |
| G | 23,84 | 4299 | 0 | 8077 |
| T | 5451 | 24,84 | 8650 | 0 |
| Total | | | | 150,28 9 |

| Balance | A | C | G | T |
|---|---|---|---|---|
| A | - | 1.0 | 1.0 | 0.9 |
| C | 0.9 | - | 1.0 | 0.9 |
| G | 0.9 | 1.0 | - | 0.9 |
| T | 1.0 | 1.0 | 1.0 | - |
| Std Dev | | | | 0.0 5 |

| Preferenc | A | C | G | T |
|---|---|---|---|---|
| A | - | 11.10 | 32.48 | 7.19% |
| C | - | - | 5.73% | 32.36% |
| G | - | - | - | 11.13% |
| T | - | - | - | - |
| Total | | | | 100.00 % |

| Averag | A | C | G | T |
|---|---|---|---|---|
| A | - | 1.0 | 1.0 | 0.9 |
| C | 0.9 | - | 1.0 | 0.9 |
| G | 0.9 | 1.0 | - | 0.9 |
| T | 1.0 | 1.0 | 1.0 | - |
| Std Dev | | | | 0.0 |

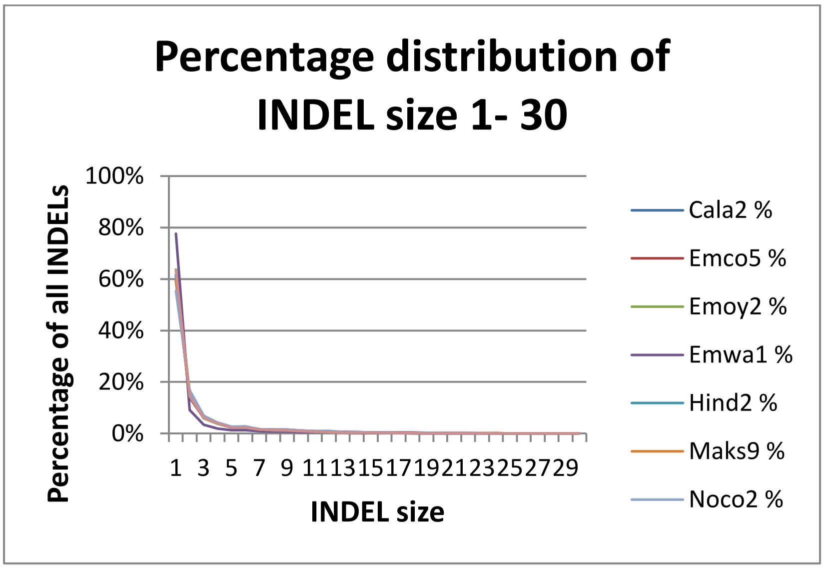
| Average | A | C | G | T |
|---|---|---|---|---|
| A | - | 11.16 | 32.46 | 7.30% |
| C | - | - | 5.83% | 32.15% |
| G | - | - | - | 11.10% |
| T | - | - | - | - |
| Total | | | | 100.00 |

**Tables 5.3: Table of mutational spectrum of *Hpa*. The 'Balance' tables display the ratio of reciprocal changes, e.g. from C→A and A→C, indicative of changes in nucleotide composition bias; they should be read across the diagonal from bottom left to top right. The 'Preferential' tables are calculated as the combined, e.g. A→C and T→G, mutations indicative of mutational preference. Std Dev = standard deviation.**

### *5.3.2.4 Distribution of INDEL sizes*

The distribution of INDEL sizes follows an exponential-like decay curve, with many small and very few large INDELs (fig 5.5 A). On closer inspection, many minor peaks can be seen in the decay, which aggregate on sizes that are exactly divisible by 3 (fig 5.5 B). These observed minor peaks are likely due to the retention of 'codon INDELs' on coding regions of the genome.

A




B

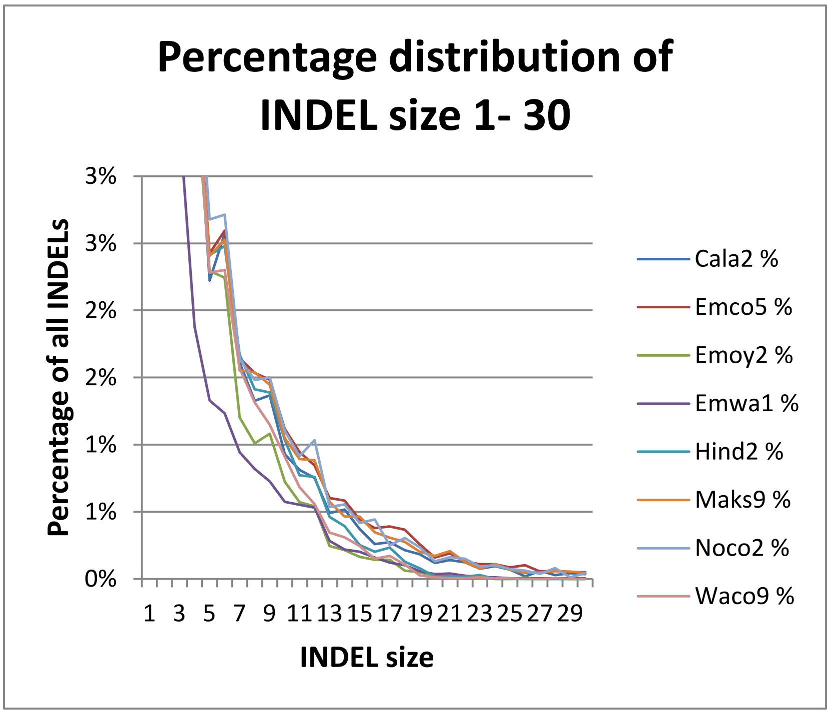



**Figure 5.5: Distribution of INDEL sizes. Only the INDELs between 1 and 30 bases are shown, and the percentage distribution is calculated as a percentage of the INDELs of size 1-30 (A). INDEL size distribution of INDEL sizes with axis set to 3% (B).**

### 5.3.3 Resequencing analysis of *Hpa* genes for 8 races of *Hpa*

#### *5.3.3.1 Coverage*

##### 5.3.3.1.1 Percentage covered

Analysing the percentage of each nucleotide of each gene covered by reads allows us to identify possible presence/absence polymorphisms and genes that are so divergent that the reads cannot align back to the reference sequence (appendix table 5.1). By comparing the number of genes that have a high percentage of their nucleotides covered by reads, we can postulate how similar the sequenced races are to the reference race, Emoy2. For Noco2 99.9% of genes are covered between 99%-100% (table 5.4) relative to Emoy2. This is the highest among the 8 races of *Hpa*. Cala2 has the lowest coverage with 98.3% of genes covered between 99%-100%. Analysing the genes that have high percentage coverage (90%-98%), but are not fully covered by reads, may give an indication of genes that are slightly divergent. Again we see a similar trend of Noco2 being most similar to Emoy2 and Cala2 being most divergent from Emoy2. This pattern is conserved for genes with 50%-89% coverage. Genes exhibiting less than 50% coverage are likely to be missing or are very highly diverged from Emoy2, and as the percentage coverage approaches 0%, the probability increases that the gene is not present in the race. Hind2 displays the most genes with less than 50% coverage (157), which is approximately twice as much as Cala2 (81), Emco5 (90), Emwa1 (58), Maks9 (82) and Noco2 (82) and 30% more than Waco9 (117).

| Percentage Covered | Cala2 | Emco5 | Emoy2 | Emwa1 | Hind2 | Maks9 | Noco2 | Waco9 |
|---|---|---|---|---|---|---|---|---|
| **99 to 100** | 14,031 | 14,202 | 14,271 | 14,227 | 14,053 | 14,170 | 14,257 | 14,118 |
| **90 to 98** | 288 | 160 | 125 | 161 | 238 | 200 | 130 | 199 |
| **50 to 89** | 178 | 127 | 102 | 130 | 127 | 129 | 105 | 144 |
| **11 to 49** | 72 | 78 | 65 | 39 | 117 | 72 | 65 | 92 |
| **0 to 10** | 9 | 12 | 11 | 19 | 40 | 10 | 17 | 25 |

**Table 5.4: Frequency distribution of percentage of nucleotides of genes covered in each *Hpa* race. The data was filtered to remove any genes predicted with 0% sequence coverage for Emoy2 as they are likely to be erroneous gene calls.**

The genes displaying the most variation in the percentage coverage in *Hpa* include 2 predicted effector genes (*HaRxL63* and *HaRxLL435*) and one gene with homology to an effector (table 5.5). Another 2 genes showed sequence homology to known genes. Gene *806362* showed homology to a 1,3-beta-glucanosyltransferase, which has been implicated in cell wall biosynthesis (Mouyna et al., 2000), and the other gene, *807641*, was homologous to a hypothetical effector as well as to a chromobox protein homolog 5, which encodes a highly conserved non-histone protein of the heterochromatin protein family. 3 out of the 10 genes with most variation in percentage coverage were predicted to be secreted and 1 had a predicted transmembrane domain. This is an overrepresentation of the secreted genes (with genes with a signal peptide accounting for 14.4% of all *Hpa* genes), while the expected number of transmembrane genes were observed (with transmembrane genes accounting for 13.5% of all *Hpa* genes).

| Gene | Cala2 | Emco5 | Emoy2 | Emwa1 | Hind2 | Maks9 | Noco2 | Waco9 | Std Dev | Effector | Best annotated BLAST hit | SP/TM |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ***806362*** | 21 | 0 | 100 | 100 | 0 | 100 | 100 | 13 | 49.38 | - | 1,3-beta-glucanosyltransferase | - |
| ***805120*** | 12 | 26 | 100 | 100 | 0 | 6 | 100 | 0 | 47.90 | - | - | - |
| ***HaRxL63*** | 0 | 100 | 100 | 100 | 0 | 100 | 100 | 100 | 46.29 | *HaRxL63* | - | SP |
| ***807641*** | 100 | 100 | 100 | 100 | 0 | 100 | 100 | 4 | 45.38 | *HaRxLL54* (partial) | chromobox protein homolog 5 | - |
| ***814385*** | 20 | 2 | 90 | 5 | 100 | 14 | 95 | 2 | 45.21 | - | - | - |
| ***805206*** | 21 | 100 | 100 | 8 | 1 | 84 | 26 | 100 | 44.78 | - | - | - |
| ***805119*** | 37 | 15 | 100 | 100 | 4 | 100 | 100 | 18 | 44.48 | - | - | SP |
| ***806374*** | 16 | 100 | 100 | 100 | 8 | 20 | 100 | 100 | 44.28 | - | - | - |
| ***810059*** | 10 | 100 | 100 | 100 | 0 | 100 | 100 | 100 | 44.06 | - | - | TM |
| ***eff_11049_g*** | 27 | 21 | 100 | 100 | 9 | 11 | 100 | 14 | 43.63 | *HaRxLL435* | - | SP |

**Table 5.5: Table of genes displaying most variation in the percentage coverage between 8 *Hpa* races. The best BLAST hits were identified through a tblastn search against NCBI NR; Signal peptides (SP) were predicted using SignalP 3 HMM (Bendtsen et al., 2004), and transmembrane helices (TM) were predicted using TMHMM (Sonnhammer et al., 1998).**

#### 5.3.3.1.2 Mean coverage of each gene for each race

Analysing the coverage of each gene relative to the observed mean coverage over all genes allows us to identify possible copy number variation (CNV). For each gene the relative mean coverage was calculated as the gene's mean coverage divided by the expected mean coverage. The majority of genes display a low variation in the maximum and minimum relative mean coverage and have very similar read coverage between races the 8 races of *Hpa* (fig 5.6). However, a secondary peak can be observed, suggesting that there is a subset of genes that are subject to CNV.

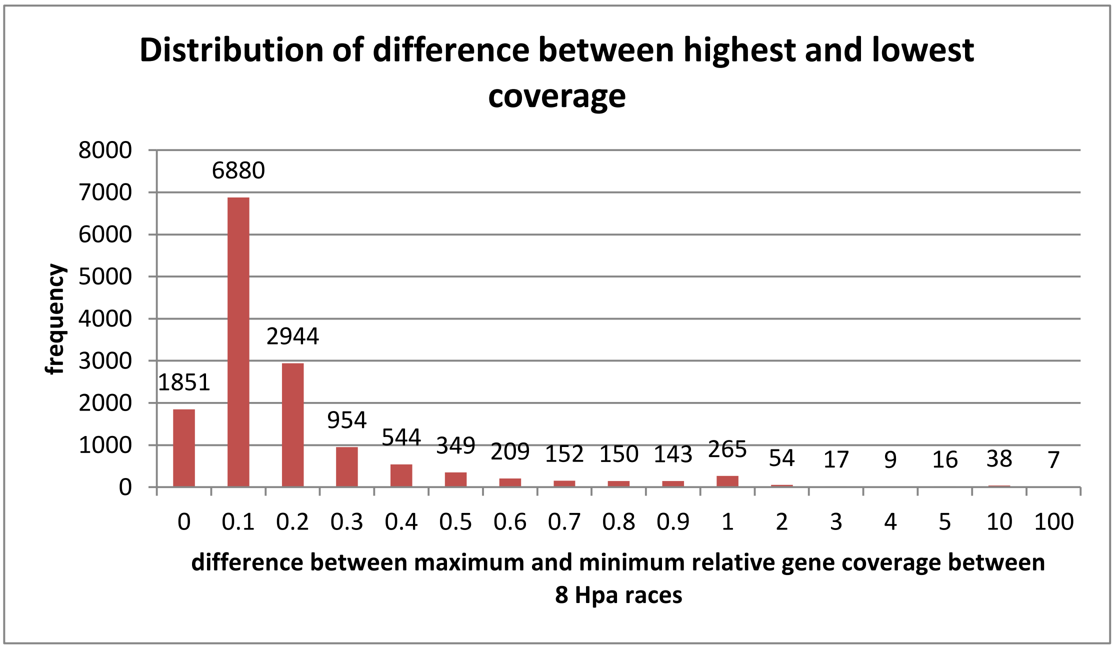



**Figure 5.6: Frequency distribution of difference between maximum and minimum relative gene coverage between 8 races of *Hpa*. The majority of genes are distributed around 0.1, suggesting most genes are of similar copy number. There is a secondary peak around 1, and a long tail indicating that there are a number of genes that have CNV polymorphisms.**

The genes displaying the most variation in relative mean coverage are also among the genes with the highest relative mean coverage (table 5.6). The 3 genes with the highest relative mean coverage are conserved in *A. laibachii*, *Plasmodium berghei* and *Caenorhabditis briggsae*. The other genes were not homologous to known proteins.

| Gene | Cala2 | Emco5 | Emoy2 | Emwa1 | Hind2 | Maks9 | Noco2 | Waco0 | Std Dev |
|---|---|---|---|---|---|---|---|---|---|
| *804802* | 464.5 | 331.1 | 410.7 | 432.2 | 333.9 | 313.4 | 424.0 | 200.1 | 86.12 |
| *800931* | 119.5 | 59.1 | 40.7 | 23.8 | 117.2 | 142.5 | 50.2 | 88.6 | 43.00 |
| *811772* | 34.1 | 50.7 | 109.4 | 128.1 | 44.0 | 23.9 | 121.5 | 87.6 | 41.68 |
| *800737* | 144.6 | 92.5 | 43.9 | 56.7 | 136.0 | 85.5 | 55.0 | 141.3 | 41.49 |
| *811584* | 1.1 | 83.3 | 73.8 | 75.9 | 1.1 | 1.1 | 83.7 | 19.0 | 39.93 |
| *814590* | 148.7 | 43.1 | 43.4 | 47.9 | 41.2 | 47.8 | 49.2 | 20.9 | 38.80 |
| *808660* | 165.6 | 59.7 | 139.9 | 156.4 | 116.8 | 123.5 | 159.7 | 178.4 | 37.79 |
| *814122* | 91.8 | 86.6 | 88.0 | 178.3 | 60.6 | 81.3 | 105.4 | 94.0 | 34.78 |
| *812642* | 98.0 | 113.9 | 131.4 | 145.1 | 108.4 | 189.6 | 154.6 | 133.2 | 29.27 |
| *814774* | 104.6 | 41.2 | 81.6 | 92.3 | 51.4 | 35.3 | 94.2 | 34.7 | 29.19 |

**Table 5.6: Genes with the most variation in the relative mean coverage. The 3 genes showing the most variation displayed homology to *A. laibachii*, *Plasmodium berghei* and *Caenorhabditis briggsae* (blastp against NCBI NR, e-value cut-off of 0.01). Std Dev = standard deviation.**

This analysis (identifying genes with the most variance in relative mean coverage) was repeated with genes that had an average minimum relative mean coverage of 10 between races to analyse the genes with the most variation among consistently high copy number genes (table 5.7). The most variable gene is homologous to an *A. laibachii* putative integrase, which is known to be a multicopy gene in oomycetes (Kemen et al., 2011). Given that this gene mean coverage is 73.8 times higher than expected in the reference isolate, it is likely that the assembly of this gene family has been collapsed at this genomics region.

| Gene | Cala2 | Emco5 | Emoy2 | Emwa1 | Hind2 | Maks9 | Noco2 | Waco0 | Std Dev |
|---|---|---|---|---|---|---|---|---|---|
| *811584* | 1.1 | 83.3 | 73.8 | 75.9 | 1.1 | 1.1 | 83.7 | 19.0 | 39.93 |
| *eff_g9604* | 70.4 | 25.7 | 14.4 | 19.9 | 15.4 | 16.6 | 13.5 | 9.7 | 19.67 |
| *813013* | 43.2 | 32.4 | 27.1 | 28.7 | 26.6 | 9.7 | 27.8 | 68.4 | 17.00 |
| *807483* | 20.6 | 29.4 | 13.8 | 25.1 | 20.7 | 53.8 | 6.3 | 15.8 | 14.23 |
| *803782* | 7.9 | 16.0 | 29.3 | 39.8 | 40.2 | 41.1 | 33.7 | 19.7 | 12.59 |
| *801561* | 8.3 | 7.0 | 20.6 | 25.9 | 20.7 | 2.6 | 22.6 | 39.4 | 11.97 |
| *814620* | 7.0 | 8.5 | 6.3 | 18.9 | 31.8 | 31.6 | 9.2 | 19.5 | 10.63 |
| *814857* | 6.7 | 3.5 | 13.6 | 5.2 | 1.8 | 9.5 | 12.9 | 28.7 | 8.57 |
| *810181* | 0.5 | 4.5 | 22.4 | 15.0 | 5.2 | 2.3 | 6.4 | 0.6 | 7.69 |
| *pasa_g19713* | 4.8 | 8.9 | 11.8 | 8.6 | 22.5 | 21.4 | 5.5 | 18.3 | 7.05 |

**Table 5.7: Genes with the most variation in the relative mean coverage, where the minimum mean coverage between the 8 races of *Hpa* is 10. The genes showing the most variation, gene 811584, displayed homology (blastp against NR, e-value cutoff of $1x10^{-6}$) to *A. laibachii* putative integrases, which are known to be multicopy genes. Std Dev = standard deviation.**

Many genes with a large variance in relative mean coverage and with the expected mean relative coverage (0.5-1.5) in Emoy2, are related to virulence function (table 5.8). For instance, the third most variable gene is the elongation factor TU, which has been described as a PAMP (Zipfel et al., 2006). It was also interesting to see the effector candidate HaRxL133, and 2 genes homologous to it, showing high variability in copy number. This could suggest that it may be recognised by some accession of *A. thaliana* so selection for loss of the gene would be advantages in *Hpa* where the host population consists of resistant ecotypes of *A.thaliana.*

| Gene | Cala2 | Emco5 | Emoy2 | Emwa1 | Hind2 | Maks9 | Noco2 | Waco9 | Std Dev | Homology |
|---|---|---|---|---|---|---|---|---|---|---|
| *802776* | 2.3 | 1.6 | 1.4 | 1.4 | 5.4 | 1.3 | 1.5 | 1.1 | 1.43 | - |
| *802778* | 2.4 | 1.5 | 1.5 | 2.0 | 5.2 | 1.6 | 1.6 | 1.1 | 1.30 | CENPB protein Homeodomain-like |
| *809897* | 4.2 | 1.0 | 0.9 | 2.9 | 0.9 | 1.6 | 1.1 | 0.8 | 1.23 | elongation factor Tu |
| *813537* | 1.0 | 4.0 | 0.8 | 0.8 | 0.9 | 1.7 | 0.9 | 2.2 | 1.13 | Fis family two component sigma 54 specific transcriptional regulator |
| *806770* | 1.5 | 1.5 | 0.7 | 0.3 | 0.1 | 3.4 | 0.9 | 2.4 | 1.10 | *HaRxL133* |
| *804929* | 3.5 | 1.0 | 0.9 | 1.8 | 0.8 | 1.1 | 0.9 | 0.9 | 0.93 | ATP synthase subunit beta |
| *814554* | 2.7 | 0.1 | 0.1 | 0.7 | 0.0 | 0.4 | 0.1 | 0.1 | 0.92 | Fis family two component sigma 54 specific transcriptional regulator |
| *803764* | 0.7 | 0.6 | 0.6 | 3.2 | 0.5 | 0.6 | 0.6 | 0.6 | 0.90 | Mitochondrial Carrier (MC) Family |
| *HaRxL133* | 1.1 | 1.1 | 0.7 | 0.5 | 0.4 | 3.2 | 0.8 | 1.4 | 0.90 | *HaRxL133* |
| *806769* | 0.7 | 0.8 | 0.7 | 0.7 | 0.7 | 3.0 | 0.8 | 0.3 | 0.85 | *HaRxL133* |

**Table 5.8: Genes with the most variation in the relative mean coverage, where they are single copy in the Emoy2. Std Dev = standard deviation.**

#### 5.3.3.1.3 Copy number variation

I have previously shown that read coverage follows a Poisson distribution. The distribution of coverage modelled as the Poisson CDF based on the observed mean over all genes, showed that a significant number of genes present as multiple copies (fig 5.7). It can also be seen that the distribution towards the lower end of the CDF spectrum is higher than expected. This is most likely due to a combination of:

- not having very high sequence coverage (a cleaner peak would be seen when read coverage is 100 fold);
- hemizygous regions of the genome having genuinely lower coverage;
- sequencing bias, where automated correction by the sequencing pipeline compensates for a genuine variance in genome nucleotide composition.

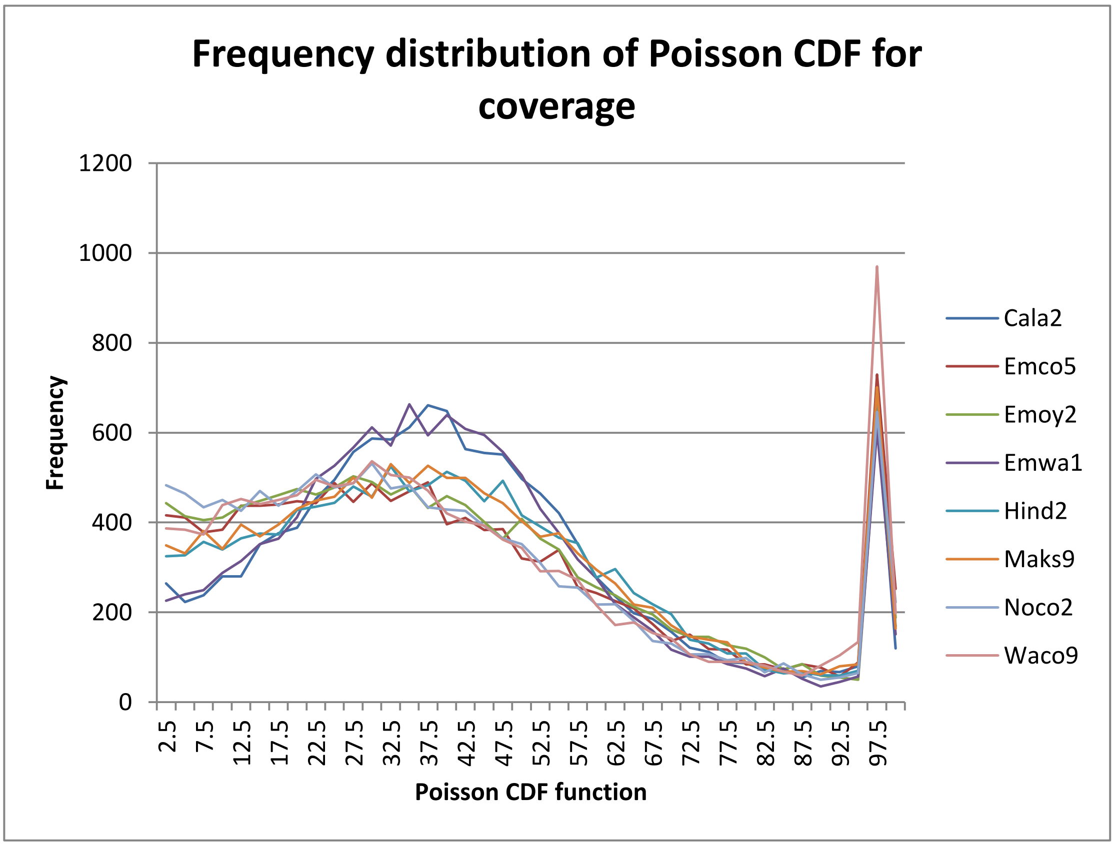



**Figure 5.7: Frequency distribution of the Poisson CDF for coverage. A general bell shaped curved can be observed, and is more pronounced for races Cala2 and Emwa1, with the other races having a higher frequency for the lower end of the CDF spectrum, possibly indicating more hemizygous genes.**

While observing a general picture of expected coverage distribution, it is also interesting to note that among the 10 genes showing most variance in Poisson CDF coverage, there is a hypothetical effector gene, *HaRxLL117* (appendix tables 5.2). Among the genes showing the most variance in expected copy number, while expected to be single copy genes in the reference race, Emoy2, we also observed the presence of another hypothetical effector gene, *HaRxL133* (appendix tables 5.2). This suggests that CNV and presence/absence polymorphisms may be a general trait of effectors.

***5.3.3.1.3.1 Hemizygosity***

Here, hemizygous regions of the genome are defined as regions with a low expected coverage and which are likely to belong to the distribution of expected single copy coverage. This can be defined as the genes/regions with a Poisson CDF for coverage of 0-1% (this falls outside the 98% confidence interval of being a single copy gene). Using this threshold, 1645 genes were identified to be hemizygous. To estimate the error rate of hemizygous calls, one can observe the number of heterozygous SNP calls made over the gene. Of the 1645 predicted hemizygous genes, 40 contained heterozygous SNPs on the coding region, and 69 contained SNPs over the gene +/-250 bp up and downstream of the gene. This equates to an accuracy of 96% for making true hemizygous calls. This 4% discrepancy is a combination of SNP calling error rate and possibility that these regions are caught in the tail of actual diploid distribution.

We observed 2 effector genes (*RXLR87* and *RXLR35*) that showed evidence of being hemizygous in some races of *Hpa* (appendix tables 5.2). Since hemizygosity is a by-product of presence/absence polymorphisms during sexual reproduction, this further supports our notion that effectors exhibit CNV polymorphisms.

### *5.3.3.2 SNPs*

The observed variation in the number of SNPs per race vaguely correlates the SNPs on exons in each race (table 5.9). A general ratio of 3:1 for Exon:Intron SNPs is observed as is a 1:1 Exon:Intron+UTRs. These ratios suggest that there is a selective balance in the number of SNPs in *Hpa* or that the majority of SNPs are accumulating through neutral drift. Alternatively, there may be a balanced selection pressure that prevents the acquisition of deleterious mutations exerted in protein coding regions (exons), regions affecting splice efficacy (introns) and expression (UTRs). However, this balance does not seem to be maintained in Emwa1, where the ratio of coding SNPs to non-coding SNPs is 1.63:1. The extent to which these ratios are conserved, despite the variation in the number of SNPs over coding and non-coding regions, is illustrated by the ration of heterozygous SNPs (table 5.10).

| SNP Position | Cala2 | Emco5 | Emoy2 | Emwa1 | Hind2 | Maks9 | Noco2 | Waco9 |
|---|---|---|---|---|---|---|---|---|
| 5' UTR | 9028 | 7777 | 2368 | 7715 | 8216 | 9286 | 2372 | 8970 |
| Exon | 26,709 | 21,713 | 7383 | 36,077 | 24,880 | 28,005 | 7166 | 27,000 |
| Intron | 7786 | 6502 | 2433 | 6744 | 7150 | 7933 | 2350 | 7489 |
| 3' UTR | 9379 | 7916 | 2536 | 7694 | 8817 | 9822 | 2523 | 9594 |
| Exon:Intron | 3.43 | 3.34 | 3.03 | 5.35 | 3.48 | 3.53 | 3.05 | 3.61 |
| Coding:Non-coding | 1.02 | 0.98 | 1.01 | 1.63 | 1.03 | 1.04 | 0.99 | 1.04 |

**Table 5.9: Number of coding and non-coding SNPs in *Hpa* races. Despite the variation in the number of total SNPs predicted on and near coding regions, the ration of coding:non-coding SNPs is maintained in most of the races (with the exception of Emwa1)**

| Het SNP Position | Cala2 | Emco5 | Emoy2 | Emwa1 | Hind2 | Maks9 | Noco2 | Waco9 |
|---|---|---|---|---|---|---|---|---|
| 5' UTR | 1256 | 304 | 2342 | 16 | 571 | 1043 | 1820 | 2393 |
| Exon | 3908 | 676 | 7341 | 67 | 1701 | 3453 | 5367 | 7431 |
| Intron | 889 | 205 | 2400 | 17 | 407 | 881 | 1866 | 1772 |
| 3' UTR | 1290 | 290 | 2472 | 23 | 575 | 1128 | 1909 | 2633 |
| Exon:Intron | 4.40 | 3.30 | 3.06 | 3.94 | 4.18 | 3.92 | 2.88 | 4.19 |
| Coding:Non-coding | 1.14 | 0.85 | 1.02 | 1.20 | 1.10 | 1.13 | 0.96 | 1.09 |

**Table 5.10: Number of coding and non-coding heterozygous (Het) SNPs in *Hpa* races. Despite the variation in the number of total SNPs predicted on and near coding regions, the ration of coding:non-coding SNPs is maintained.**

As mentioned previously, positive selection leaves its signature in form of sequence variation at the genome level. The accumulating SNPs is important for functional and evolutionary changes and identifying SNPs is one of the easier ways to make inferences about the selection pressure on genes. For example, *ATR1* is one such gene that displays a very high rate of SNPs over the coding region of the gene (on average 16.75 SNPs per race) (appendix tables 5.3). In addition, the variation in the number of SNPs observed between the races is an indication of the variation between the sampled populations. *ATR1* exhibits the highest variance in the number of SNPs on the coding region. However, *ATR1* also has a number of SNPs in the nearby non-coding region, which implies that there are certain genome dynamics leading to elevated variation in the *ATR1* region. This trend does not extend to heterozygous SNP positions.

#### 5.3.3.2.1 Protein coding effects of SNPs

While cataloguing sequence variation is important, it is the effect of SNPs on the protein code that leads to differentiation of function. The protein coding effect of each SNP over gene coding regions for each race was determined. A general ~4:1 ratio of non-synonymous : synonymous SNPs was observed (table 5.11). It was also interesting to note that ~60% of all heterozygous SNPs were 'heterozygous non-synonymous SNPs', where the heterozygous SNP leads to 2 different proteins being encoded). Conversely, this implies that ~40% of all heterozygous SNPs in *Hpa* do not alter the proteins encoded by the genes.

Other, more drastic, mutations caused by SNPs include mutated start codons and formation of premature stop codons. Apart from Emwa1 and Noco2 (which is hypothesised to be very similar to Emoy2), we observed approximately 60 mutated start codons and 200 premature stop codons per race due to SNPs.

| | Cala2 | Emco5 | Emoy2 | Emwa1 | Hind2 | Maks9 | Noco2 | Waco9 |
|---|---|---|---|---|---|---|---|---|
| Synonymous SNPs | 6006 | 4913 | 1600 | 18,416 | 5643 | 6197 | 1521 | 6140 |
| Non-synonymous SNPs | 20,618 | 16,700 | 5658 | 17,593 | 19,180 | 21,735 | 5565 | 20,740 |
| Heterozygous non-synonymous SNPs | 2429 | 470 | 3778 | 42 | 1125 | 2240 | 2942 | 4070 |
| Non-synonymous:Synonymous SNPs | 3.433 | 3.399 | 3.536 | 0.955 | 3.399 | 3.507 | 3.66 | 3.378 |
| All non-synonymous:Synonymous SNPs | 3.837 | 3.495 | 5.898 | 0.958 | 3.598 | 3.869 | 5.59 | 4.041 |
| Mutated start codon | 69 | 47 | 0 | 0 | 66 | 76 | 6 | 64 |
| Premature stop codon | 234 | 219 | 0 | 92 | 232 | 252 | 16 | 188 |

**Table 5.11: Number of synonymous and non-synonymous SNPs in *Hpa* races. An approximate 4:1 ratio of non-synonymous SNPs to synonymous SNPs can be observed, where all non-synonymous: synonymous SNPs refers to both the homozygous and heterozygous SNPs. On average, ~60% of all heterozygous SNPs lead to the encoding on 2 different proteins at the same locus. There is an observable trend of ~60 SNPs causing mutations in the start codon and ~200 SNPs causing premature stop codons in each race.**

We observe that *ATR1* has one of the highest non-synonymous:synonymous SNP ratios, and the highest variance in this category. These ratios provide further evidence that pathogen effectors, which have possible avirulence functions (recognised by the plant host), can leave signatures of accelerated evolution through elevated levels of SNPs with protein modifying effects (appendix tables 5.5). While *ATR1* exhibits a high level of variation, a small number of mutations is sufficient to modify the function of effectors as demonstrated by studies of the *Phytophthora infestans* effector Avr3a, where 2 modification are sufficient to switch from a resistant to susceptible allele of the effector (Armstrong et al., 2005).

Given the very low number of genes with modified start codons, I was interested to find an effector (*HaRxLL55*) to be among them (appendix tables 5.5). A total of 10 putative effectors were identified with premature stop codons introduced by SNPs (*HaRxLL447*, *HaRxLL105*, *HaRxLL181*, *HaRxLL14*, *HaRxLL53*, *HaRxLL100*, *HaRxLL115*, *HaRxLL133*, *HaRxLL89* and *HaRxLL176*). Given that 662 genes were predicted to be affected by SNPs causing premature stop codons (~4% of all genes), we would expect to see around 10 effectors to have premature stop codons (based on an estimate of 200-300 putative effectors), agreeing with the observations.

### *5.3.3.3 INDELs*

The observed variation in the number of INDELs per race is conserved in the INDELs on exons in each race (table 5.12). A general ratio of 1:1 for Exon:Intron INDELs is observed, as is a 1:3 coding:non-coding ratio. This suggests that there is a selective balance in the number of INDELs in *Hpa*, as was observed with the SNPs. This balance does not seem to be maintained in Emwa1, where the ratio of coding:non-coding INDELs is approximately 4 times more. Despite the variation in the number of INDELs over coding and non-coding regions between *Hpa* races, their ratios in each race are very similar. This is true to a lesser extent with heterozygous INDELs where a general ratio of 1.5:1 of Exon:Intron (with more variance than was observed with all INDEL positions), and 1:3 for coding:non-coding (table 5.13).

| All INDEL positions | Cala2 | Emco5 | Emoy2 | Emwa1 | Hind2 | Maks9 | Noco2 | Waco9 |
|---|---|---|---|---|---|---|---|---|
| Up | 1576 | 1539 | 547 | 2943 | 1529 | 1761 | 893 | 1652 |
| Exon | 1399 | 1290 | 617 | 9660 | 1270 | 1600 | 935 | 1354 |
| Intron | 1331 | 1238 | 523 | 2414 | 1230 | 1449 | 849 | 1283 |
| Down | 1410 | 1285 | 491 | 2604 | 1402 | 1551 | 796 | 1504 |
| Exon:Intron | 1.05 | 1.04 | 1.18 | 4.00 | 1.03 | 1.10 | 1.10 | 1.06 |
| Coding:Non-coding | 0.32 | 0.32 | 0.40 | 1.21 | 0.31 | 0.34 | 0.37 | 0.31 |

**Table 5.12: Number of coding and non-coding INDELs in *Hpa* races showing both homozygous and heterozygous INDELs. Non-coding regions are the 5' UTR, 3' UTR and Introns. Despite the variation in the number of total INDELs predicted on and near coding regions, the ratio of coding:non-coding INDELs is maintained in most of the races (with the exception of Emwa1).**

| Heterozygous INDEL positions | Cala2 | Emco5 | Emoy2 | Emwa1 | Hind2 | Maks9 | Noco2 | Waco9 |
|---|---|---|---|---|---|---|---|---|
| Up | 338 | 263 | 478 | 2682 | 309 | 381 | 706 | 604 |
| Exon | 545 | 344 | 528 | 8810 | 443 | 590 | 746 | 641 |
| Intron | 284 | 221 | 456 | 2199 | 255 | 330 | 656 | 422 |
| Down | 301 | 217 | 440 | 2365 | 259 | 336 | 651 | 540 |
| Exon:Intron | 1.90 | 1.55 | 1.16 | 4.01 | 1.74 | 1.79 | 1.14 | 1.52 |
| Coding:Non-coding | 0.59 | 0.49 | 0.38 | 1.22 | 0.54 | 0.56 | 0.37 | 0.41 |

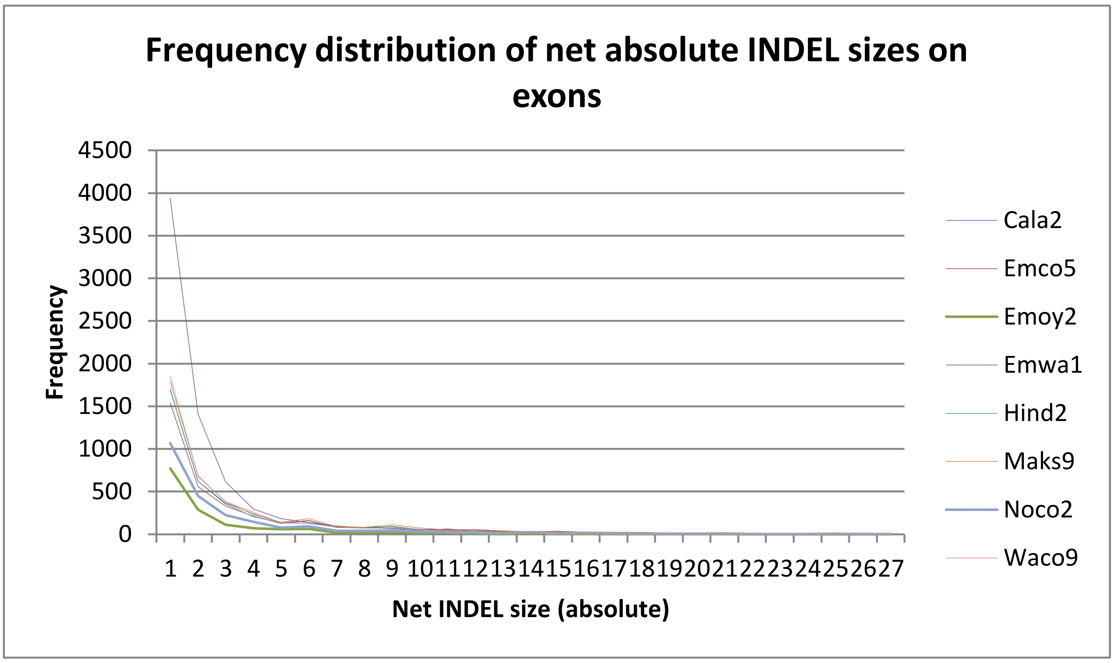
**Table 5.13: Number of coding and non-coding heterozygous INDELs in *Hpa* races. Non-coding regions are the 5' UTR, 3' UTR and Introns. Despite the variation in the number of total heterozygous INDELs predicted on and near coding regions, the ratio of coding:non-coding INDELs is maintained, with the exception of Emwa1.**

INDELs, especially if not a codon INDEL, are likely to have a much greater effect on the function of proteins than SNPs. *HaRxLCRN4*, a hypothetical crinkler protein, is one such gene that displays a very high rate of INDELs over its coding region (on average 2.25 INDELs per race) (appendix tables 5.6). This high level of variation is maintained when comparing the coding to non-coding INDEL ratio and when considering heterozygous INDELs (appendix tables 5.6).

#### 5.3.3.3.1 Protein coding effects of INDELs

While it is important to catalogue sequence variation, it is the effect of the INDEL on the protein code that allows for differentiation of gene function. INDELs may only have small effects on the amino acid sequence (e.g. loss of a single codon), but in general are likely to cause frameshift mutations, which often lead to a loss of the original gene function. There may also be a situation where 2 INDELs that lead to internal frameshift, with conserved start and stop codon, where the net INDEL length is exactly divisible by 3. It was observed that the majority of net INDEL lengths larger than 1 are exactly divisible by 3 (fig 5.8). Observing the frequency distribution of net INDEL lengths over coding regions only also revealed that the distribution of net INDEL lengths greater than 6 are highly similar, with the exception of Emoy2 (which is the reference strain) and Cala2. The percentage distribution of net INDEL lengths also showed that the net INDEL lengths are divisible by 3 and that the distribution of higher INDEL lengths is similar between *Hpa* races (fig 5.9).

A




B

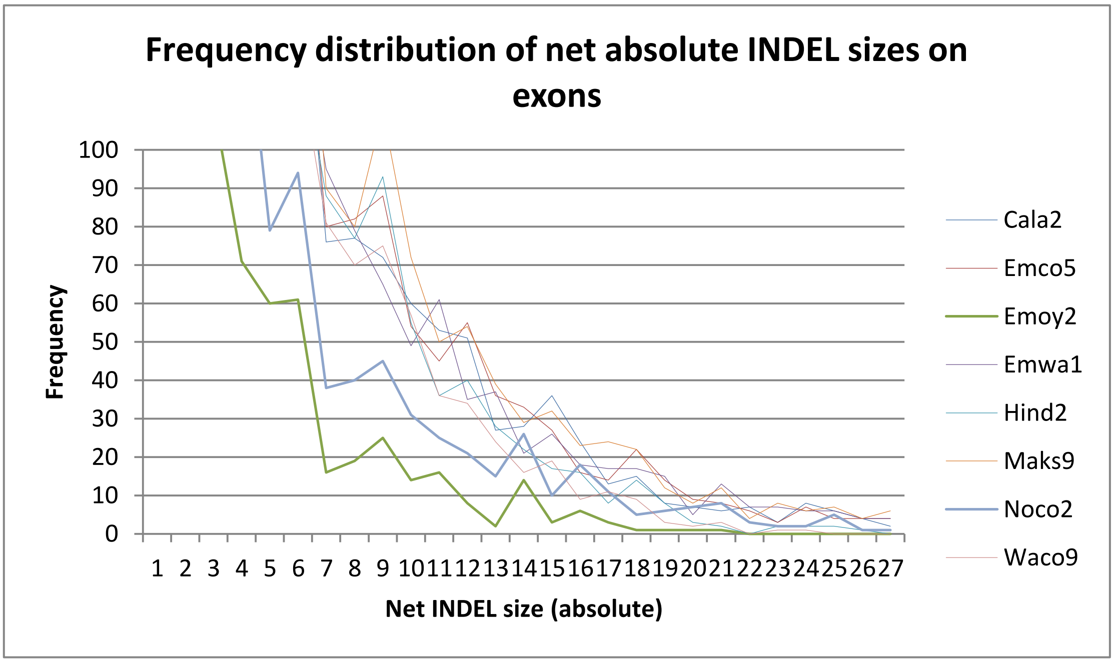
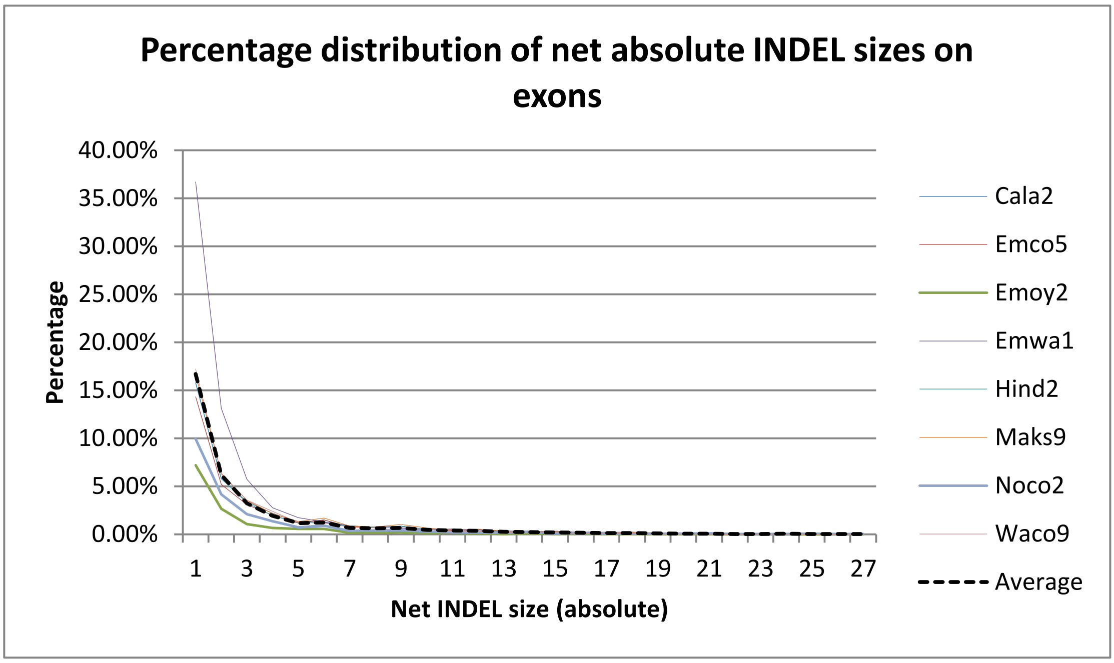



**Figure 5.8: Frequency distribution of the INDEL length over coding regions of genes with full frequency spectrum (A) and partial frequency spectrum (B). A general trend of a power law distribution can be observed, with additional peaks of net INDEL lengths divisible by 3, indicating net codon loss/gain and/or internal frameshifts. In addition, with an increase in the net INDEL length, the convergence in the variation in frequency between the *Hpa* races increases. It was also observed that in Emoy2 and Cala2, there are peaks at 14 and 16 and not at 15; I believe these are due to prediction errors,**

A




B

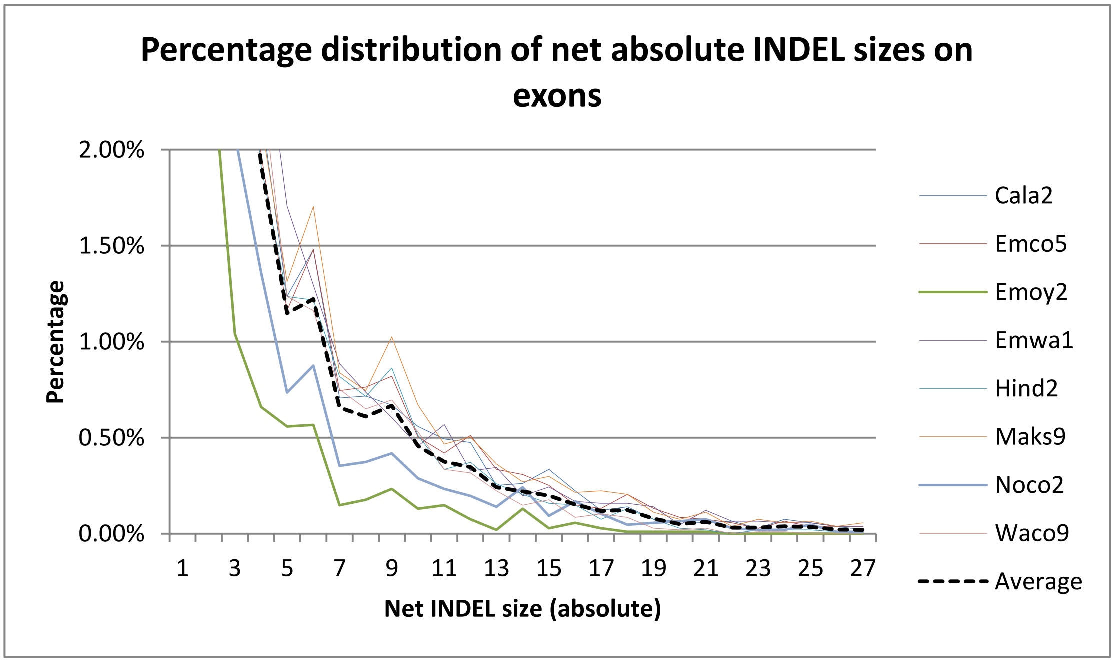



**Figure 5.9: Percentage distribution of the INDEL length over coding regions of genes with full percentage spectrum (A) and partial percentage spectrum (B). A general trend of a power law-like distribution can be observed, with additional peaks of net INDEL lengths divisible by 3, indicating net codon loss/gain and/or internal frameshifts. In addition, with an increase in the net INDEL length, the convergence in the variation in frequency between the *Hpa* races increases. The average peak is less pronounced at 15, due to possible prediction errors in Emoy2 and Cala2, which form peaks at net INDEL lengths of 14 and 16.**

### 5.3.4 DnaSP Analysis

DnaSP is a software package for the analysis of nucleotide polymorphisms from aligned DNA sequences (Librado and Rozas, 2009). Using the second stage of the VariTale pipeline, all the parental haplotype gene sequences, fully covered and lacking INDELs, over each race are generated. Since the analysis is based only on genes that are fully covered and lack INDELs, gene sequences will align without gaps. The calculations from this section of the analysis provide a base minimum of the true results, as we do not consider genes with INDELs, and parental haplotypes are predicted and not the actual parental haplotypes.

#### *5.3.4.1 Analysis of sample size*

In all comparative genomics analyses, the accuracy and significance of the described variation increases with the size of the sample population. To illustrate this point, I plotted the change in the percentage of genes that were analysed against the number of races analysed, where the number of races were selected randomly from the 8 *Hpa* races (fig 5.10). With just 2 races (equating to a maximum of 4 haplotypes) 98.8% of the genes had at least 1 parental haplotype identified. Using data from 5 races, every gene had at least 1 haplotype identified. The total number of segregating sites (S) and the total number of mutations (Eta) increase significantly with an increase in the percentage of genes analysed up to 5 races, after which subsequent increases in the number of analysed genes increase by less than 2% with each additional race. I therefore concluded that a minimum of 5 races (or 10 haplotypes) should be used for this type of comparative genomics analysis.

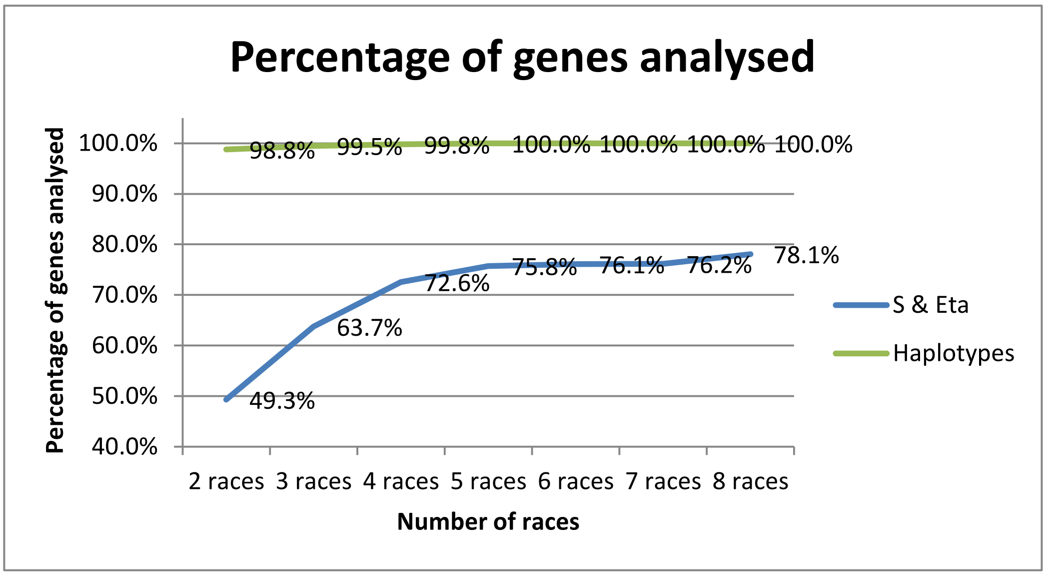



**Figure 5.10: Effect of increasing the sample size on the percentage of genes for which S, Eta and the number of haplotypes can be analysed. S is the number of segregating sites and Eta is the number of mutations.**

The frequency distribution of the predicted number of haplotypes per gene shows a similar trend, according to which changes in the distribution of predicted number of haplotypes per gene have a less significant effect at samples sizes above 5 races (fig 5.11).

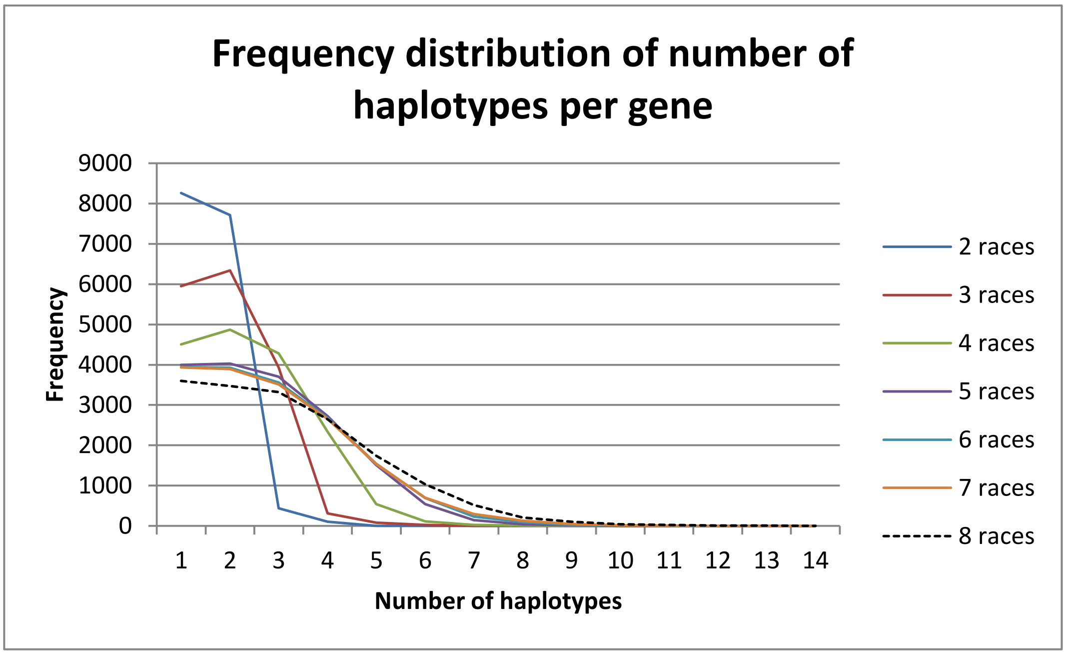



**Figure 5.11: Frequency distribution of predicted number of haplotypes per gene with increasing sample population. The change in the distribution becomes less significant after 5 races.**

Analysing the frequency distributions of S and Eta, shows that a minimum of 4-5 samples should be considered for this type of analysis (fig 5.12; fig 5.13). With this, I concluded that the analysis of the 8 races of *Hpa* should bring about meaningfull results for most of the genes in *Hpa*.

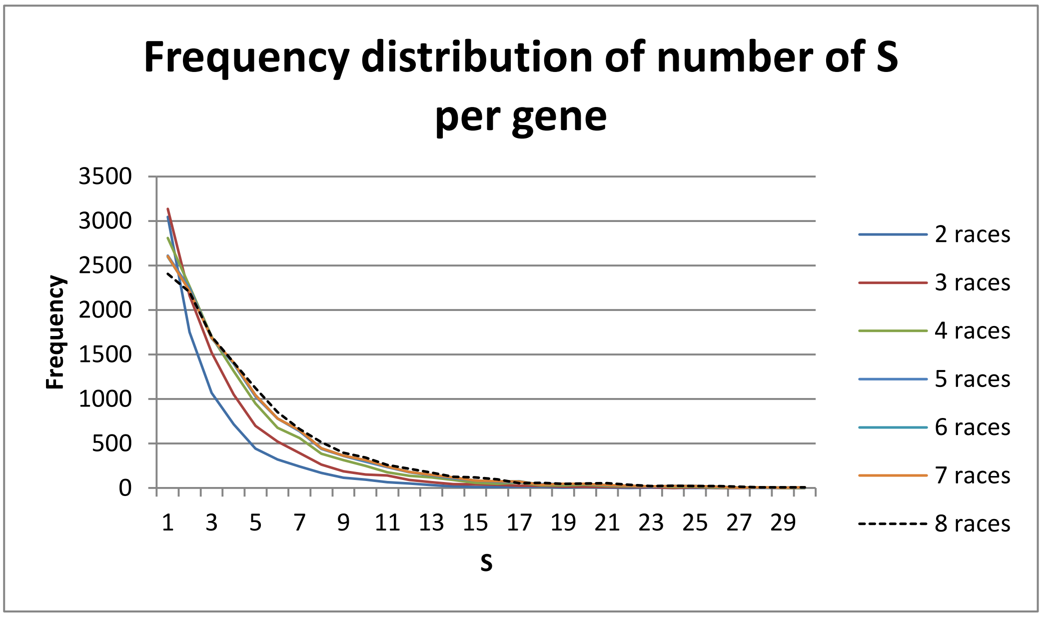



**Figure 5.12: Frequency distribution of predicted number of segregating sites (S) per gene with increasing sample population. The change in the distribution becomes less significant after 4-5 samples.**

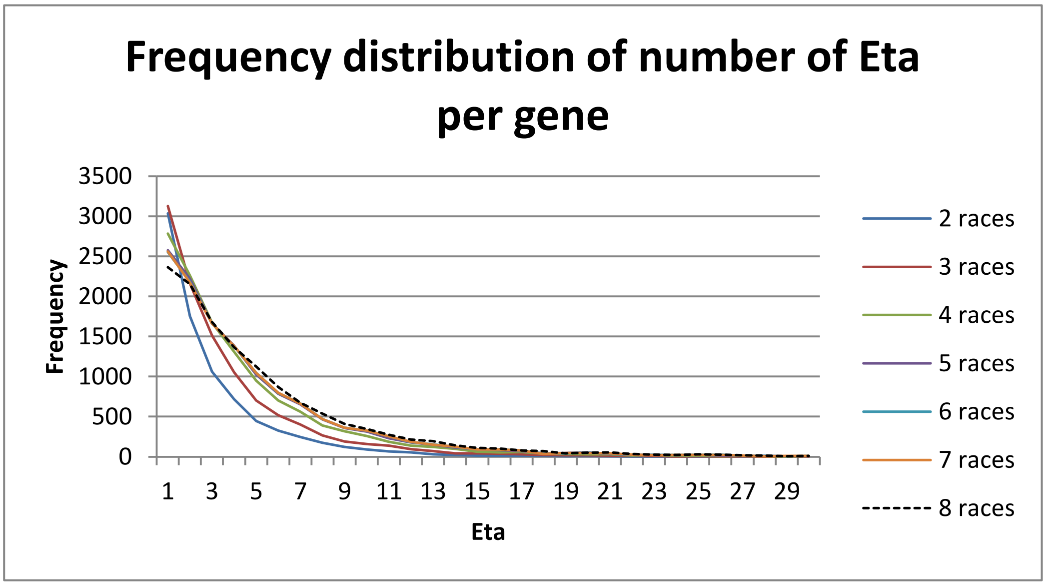



**Figure 5.13 Frequency distribution of predicted number of mutations (Eta) per gene with increasing sample population. The change in the distribution becomes less significant after 4-5 samples.**

#### *5.3.4.2 S, Eta and the number of haplotypes*

The frequency distribution of the number of predicted haplotypes per gene shows that most genes have a small number of haplotypes, with approximately 11% of all genes (1669) having 6 or more parental haplotypes between the 8 races of *Hpa* (i.e. from a set of 16 parental haplotypes) (fig 5.14). There was also a tight correlation between the distribution of S, the number of segregating sites, and Eta, the number of mutations, with both distributions following an exponential-like distribution.

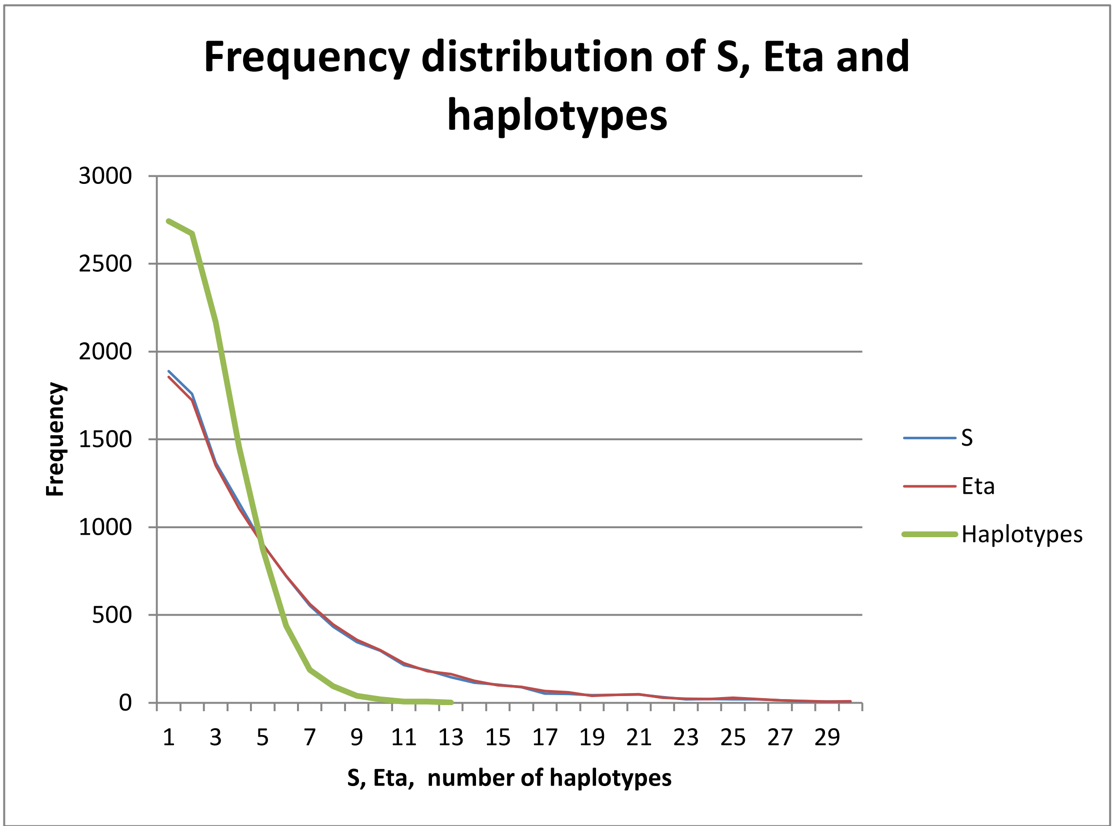



**Figure 5.14: Frequency distribution of the number of segregating sites (S), the number of mutations (Eta) and the number of haplotypes.**

Previously we have investigated the observed SNP and INDEL rate over genes. However, this would be a naïve method of analysing variation of a gene in a given population. For instance, if there are only 2 alleles in the population, but these alleles are very different from each other, a SNP/INDEL analysis may consider this gene to be variable due to the high average number of SNPs/INDELs and the large variation between the two alleles. Instead, considering the number of haplotypes, total segregating sites and the total number of mutations that are observed over the estimated parental haplotypes provides a more accurate measure of the variation in the sample population. Of the 15 genes with the highest number of segregating sites we find 5 effector genes, including *ATR1* (table 5.14). The gene with the highest number of segregating sites is also an effector candidate (*HaRxL19*). These 5 effector genes are also in the list of the 15 genes with the highest number of total mutations, which follows my previous suggestion that the number of segregating sites and the number of total mutations are tightly coupled (table 5.15).

| Gene | N | S | Eta | Hap | Tajima D | Sig D | FuLi D* | SigD | FuLi F* | SigF | FuFs | Effector |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *ceg_12014_g* | 16 | 67 | 67 | 7 | 1.1715 | n.s. | 1.6985 | ** | 1.79 | ** | 8.664 | *HaRxL19* |
| *803035* | 16 | 63 | 65 | 8 | 0.6639 | n.s. | 1.4529 | * | 1.42 | n.s. | 5.959 | |
| *ATR1_Emoy2* | 12 | 62 | 64 | 5 | 0.2156 | n.s. | 1.6512 | ** | 1.452 | # | 8.821 | *ATR1* |
| *pasa_gi_SuperContig27_149* | 10 | 61 | 61 | 10 | -0.0573 | n.s. | 0.7494 | n.s. | 0.6186 | n.s. | -1.683 | |
| *808490* | 8 | 61 | 64 | 4 | 1.1647 | n.s. | 1.6543 | ** | 1.7182 | * | 8.354 | *HaRxL128* |
| *eff_g11103* | 8 | 61 | 64 | 4 | 1.1647 | n.s. | 1.6543 | ** | 1.7182 | * | 8.354 | |
| *801867* | 12 | 60 | 62 | 5 | 0.1941 | n.s. | 1.6495 | ** | 1.4442 | n.s. | 8.611 | |
| *eff_g11210* | 12 | 60 | 62 | 5 | 0.1941 | n.s. | 1.6495 | ** | 1.4442 | n.s. | 8.611 | |
| *801132* | 16 | 56 | 56 | 6 | 1.1628 | n.s. | 1.6858 | ** | 1.7765 | ** | 9.612 | |
| *eff_g7948* | 16 | 56 | 56 | 6 | 1.1628 | n.s. | 1.6858 | ** | 1.7765 | ** | 9.612 | |
| *808092* | 16 | 50 | 56 | 10 | 0.2705 | n.s. | 1.5921 | ** | 1.4062 | n.s. | 2.189 | |
| *811590* | 16 | 50 | 50 | 9 | -0.5228 | n.s. | 1.2594 | n.s. | 0.8711 | n.s. | 2.185 | *HaRxL72* |
| *eff_5465_g* | 16 | 50 | 50 | 9 | -0.5228 | n.s. | 1.2594 | n.s. | 0.8711 | n.s. | 2.185 | |
| *807858* | 16 | 48 | 48 | 7 | 0.719 | n.s. | 1.5647 | ** | 1.5308 | # | 6.056 | |
| *805640* | 16 | 48 | 50 | 6 | -0.0403 | n.s. | 1.6767 | ** | 1.3746 | n.s. | 7.141 | *HaRxL73* |

**Table 5.14: The 15 genes with the highest number of segregating sites (S). Most of these genes have high values for the number of analysable haplotypes (Hap) and total number of mutations (Eta). There is less coupling between Eta and Hap compared to Eta and S. However this could be an artefact introduced by the method, which ignores genes with INDELs. Including genes with INDELs may increase the number of observed haplotypes. N = number of samples, n.s. = not significant, * = significant at p=0.05, ** = significant at p=0.01 and *** = significant at p=0.001.**

| Gene | N | S | Eta | Hap | Tajima D | Sig D | FuLi D* | Sig D | FuLi F* | Sig F | FuFs | Effector |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *ceg_12014_g* | 16 | 67 | 67 | 7 | 1.1715 | n.s. | 1.6985 | ** | 1.79 | ** | 8.664 | *HaRxL19* |
| *803035* | 16 | 63 | 65 | 8 | 0.6639 | n.s. | 1.4529 | * | 1.42 | n.s. | 5.959 | |
| *ATR1_Emoy2* | 12 | 62 | 64 | 5 | 0.2156 | n.s. | 1.6512 | ** | 1.452 | # | 8.821 | *ATR1* |
| *808490* | 8 | 61 | 64 | 4 | 1.1647 | n.s. | 1.6543 | ** | 1.7182 | * | 8.354 | *HaRxL128* |
| *eff_g11103* | 8 | 61 | 64 | 4 | 1.1647 | n.s. | 1.6543 | ** | 1.7182 | * | 8.354 | |
| *801867* | 12 | 60 | 62 | 5 | 0.1941 | n.s. | 1.6495 | ** | 1.4442 | n.s. | 8.611 | |
| *eff_g11210* | 12 | 60 | 62 | 5 | 0.1941 | n.s. | 1.6495 | ** | 1.4442 | n.s. | 8.611 | |
| *pasa_gi_SuperContig27_149* | 10 | 61 | 61 | 10 | -0.0573 | n.s. | 0.7494 | n.s. | 0.6186 | n.s. | -1.683 | |
| *801132* | 16 | 56 | 56 | 6 | 1.1628 | n.s. | 1.6858 | ** | 1.7765 | ** | 9.612 | |
| *eff_g7948* | 16 | 56 | 56 | 6 | 1.1628 | n.s. | 1.6858 | ** | 1.7765 | ** | 9.612 | |
| *808092* | 16 | 50 | 56 | 10 | 0.2705 | n.s. | 1.5921 | ** | 1.4062 | n.s. | 2.189 | |
| *811590* | 16 | 50 | 50 | 9 | -0.5228 | n.s. | 1.2594 | n.s. | 0.8711 | n.s. | 2.185 | *HaRxL72* |
| *eff_5465_g* | 16 | 50 | 50 | 9 | -0.5228 | n.s. | 1.2594 | n.s. | 0.8711 | n.s. | 2.185 | |
| *805640* | 16 | 48 | 50 | 6 | -0.0403 | n.s. | 1.6767 | ** | 1.3746 | n.s. | 7.141 | *HaRxL73* |
| *eff_g7740* | 16 | 48 | 50 | 6 | -0.0403 | n.s. | 1.6767 | ** | 1.3746 | n.s. | 7.141 | |

**Table 5.15 The 15 genes with the highest number of total mutations (Eta). Most of these genes have high values for the number of analysable haplotypes (Hap) and total segregating sites (S). There is less coupling between Eta and Hap compared to Eta and S. However this could be an artefact introduced by the method, which ignores genes with INDELs. Including genes with INDELs may increase the number of observed haplotypes. N = number of samples, n.s. = not significant, * = significant at p=0.05, ** = significant at p=0.01 and *** = significant at p=0.001.**

### *5.3.4.3 Tajima's D*

One way to infer positive selection on a gene is to show that the gene is not evolving neutrally. There are various hypothesis tests for neutrality, including Tajima's D, Fu and Li's D* and F* and Fu's Fs. Positive values for Tajima's D indicate low levels of both low and high frequency polymorphisms suggesting a decrease in population size and potentially balancing selection (Tajima, 1989). A negative value for Tajima's D indicates a large number of low frequency polymorphisms suggesting population size expansion and possibly purifying selection. While calculating p-values for this Tajima's D statistic is impossible for samples, it is generally accepted that values greater than +2 and values less than -2 are likely to be significant (Tajima, 1989). The observed frequency variation of the values calculated for Tajima's D follows a rough bell shaped curve with a median at around 0 (fig 5.15). The mean of 0.2 and variance of 0.8 are similar to the expected beta distribution around a mean of 0 with variance of 1 (Tajima, 1989). Only 10 genes show values of Tajima's D lower than -2 (an excess of low frequency polymorphisms) of which 2 are effectors (appendix table 5.8). 273 genes have values greater than 2 (indicating low levels of low and high frequency polymorphisms) of which 9 are effectors (appendix table 5.8). These results suggest that effectors may be among the most diverse and at the same time among the most conserved genes in *Hpa*, which supports the possibility of conserved core effectors and effectors that are highly diverse due to interactions with the host leading to a differential fitness.

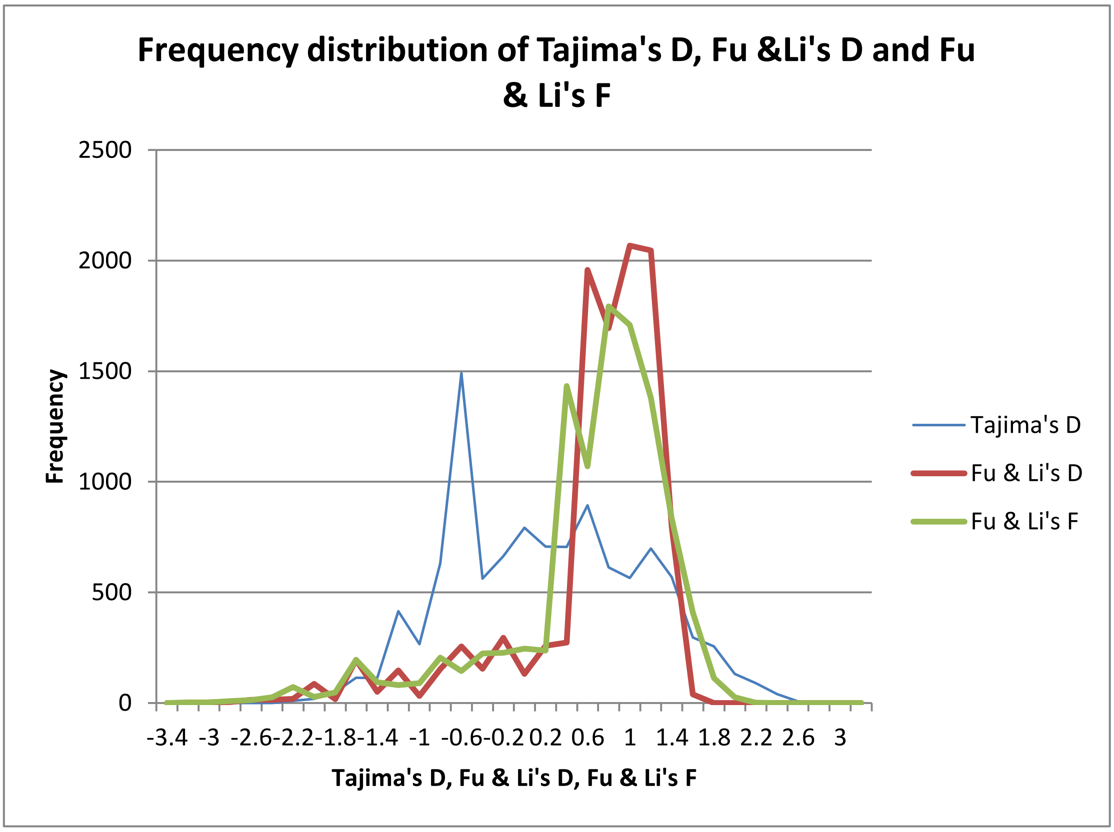



**Figure 5.15: Frequency distribution of Tajima's D, Fu & Lis D* and F*. Significance values for Tajima's D are +/- 2, and for Fu and Li's the are -1.8/+1.4 for D* and -2/+1.55 for F*.**

### *5.3.4.4 Fu & Li's D and Fu & Li's F*

While Fu and Li's statistics (Fu and Li, 1993) follow a similar principle to Tajima's, they also consider that some parts of the gene share a much longer ancestry than others. Applying Fu and Li's tests revealed a larger number of genes with test values of less than -1.8 (critical value for D*) and -1.4 (critical value for F*) (172 for Fu and Li's D*, and 478 for Fu and Li's F*, compared to only 10 for Tajima's D). For the Fu and Li D test these low valued genes include 5 effector genes (*HaRxLL27*, *HaRxLL27* homologue, *HaRxLL441*, *HaRxLL180* and *RxLRNEE3*) (appendix table 5.8) and for Fu and Li's F* 6 effectors (*HaRxLL27*, *HaRxLL27* homologue, *HaRxL89*, *HaRxLL441*, *RXLRNEE3* and *HaRxLL180*) (appendix table 5.8). It was also observed that 5 effectors are in the genes with the highest values for Fu and Li's D (*HaRxL123*, *HaRxL73*, *HaRxL128*, *ATR1* and *HaRxL19*) (appendix table 5.8), suggesting that a low value of Fu and Li's D may indicate effectors that are highly diverse. There was only 1 effector among the 20 highest values of Fu and Li's F (*HaRxL21*) (appendix table 5.8).

### *5.3.4.5 Fu's Fs*

Fu Fs (Fu, 1997) is a statistical test based on the infinite sites model of mutation, with a negative value being evidence for an excessive number of alleles expected from a recent population expansion and a positive value being evidence for a deficiency of alleles from a population bottleneck or over dominant selection (Fu, 1997). The distribution of Fu's Fs statistic seem normally distributed with a peak around 0 (fig 5.16). Among the genes with the 20 lowest values of Fu's Fs there are 5 effectors (*HaRxL51*, *HaRxLL163*, *HaRxLL133*, *HaRxLL15* and *HpRXLR104*) (appendix table 5.8). In the 20 highest scoring genes there are 3 effectors (*HaRxLL38*, *ATR1* and *HaRxL21*). This supports the idea of multiple selection pressures acting on effector genes.

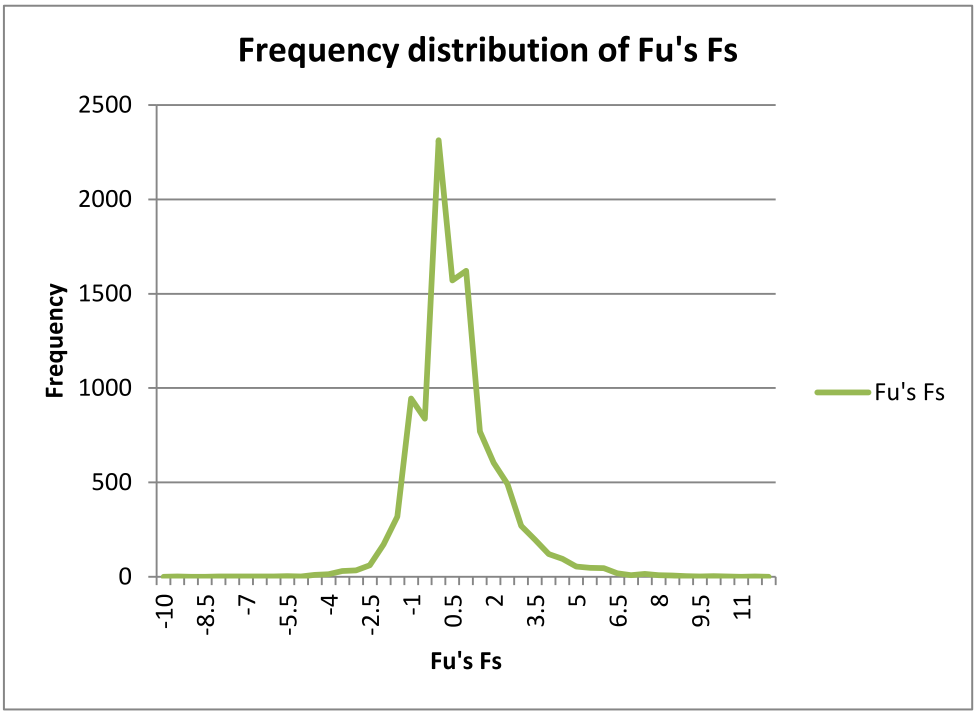



**Figure 5.16: Frequency distribution of Fu's Fs. The empirical significance cutoffs (P < 0.02) are -2.423 and 6.062**

### 5.3.5 PAML analysis

PAML consists of a suite of programs that are directly able to infer positive selection using direct dN/dS calculation, or modelling the gene using a number of sites to identify the likelihood of appositive selection acting on one of those sites. While these test are usually used to between species analysis and not interspecies analysis (like in this study), it may be possible to make inferences from the outcome as done so previously in Haas et al. (2009).

#### *5.3.5.1 Tree construction*

Some of the programs in the PAML suite require a phylogenetic tree of the input samples. Since there is currently no tree available for *Hpa* races, Mr Bayes v 3.1.2 (Ronquist and Huelsenbeck, 2003) was used to generate 2 phylogenetic trees. The first tree was constructed from the alignment of a region of *Hpa* that is homologous to the *Phytophthora infestans* mitochondrion over a Ribosomal L2 gene. The region had 30 segregating sites. Mr Bayes was run using the General Time Reversible models with gamma-shaped rate variation with a portion of invariable sites. The simulation was run over 1,000,000 generations with a sample frequency of 100 and diagnostics printed every 1000$^{th}$ generations. 9 heated chains are used in the Metropolis coupling to improve MCMC sampling of the dataset. A default of 25% was used for the burning (2500). The resultant tree clusters Emwa1, Emoy2 and Noco2 in one cluster and Waco9, Maks9, Emco5 and Hind2 in another cluster, while Cala2 clusters on its own (fig 5.17).

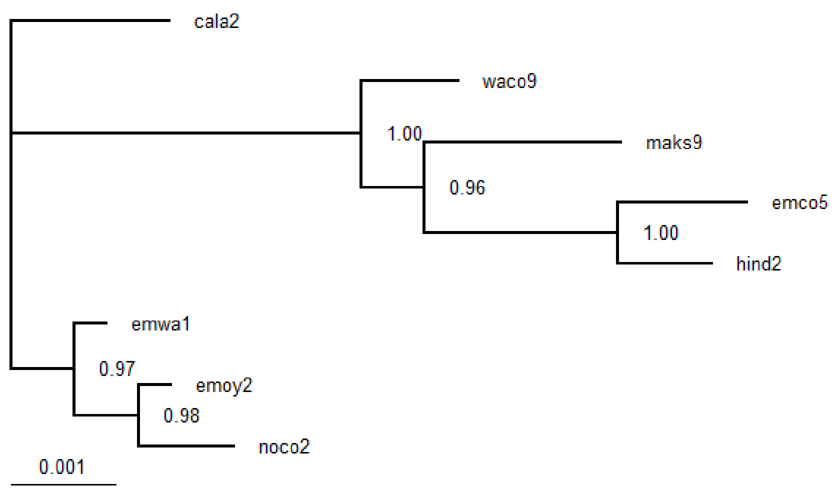



**Figure 5.17: Phylogenetic tree of *Hpa* races based on sequence homologous to *P. infestans* mitochondria. The tree was generated using Mr Bayes and bootstrap values are shown.**

A second tree was generated using all 11,570 segregating sites on the largest *Hpa* contig. Since only the segregating sites were used in the analysis, the Mr Bayes parameters were modified to model the variation rate as equal. In principle, the generated tree agreed with the previous tree, apart from Waco9, which did not cluster (compare fig 5.18 and fig 5.17 above). The bootstrap values were slightly higher than those of the previous tree.

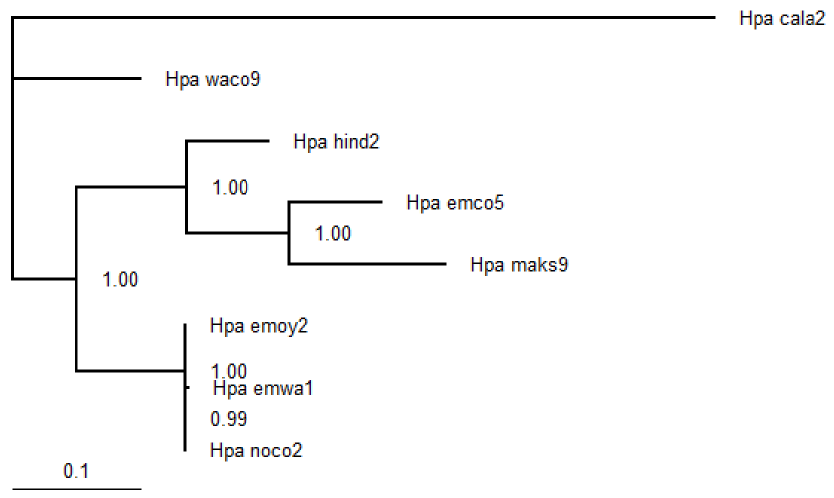



**Figure 5.18: Phylogenetic tree of *Hpa* races based on 11,570 segregating sites on the largest *Hpa* contig. The tree was generated using Mr Bayes and bootstrap values are shown.**

Both trees showed an overall good agreement. However, the tree generated using the 11,570 segregating sites on the largest *Hpa* contig generated slightly higher bootstrap values and was therefore used for further analysis.

### *5.3.5.2* dN/dS

PAML is able to calculate the dN/dS, the rate of non-synonymous substitutions per non-synonymous site divided by the number of synonymous substitutions per synonymous site, using numerous methods. In previous studies (Haas et al., 2009), the dN/dS values have been calculated using yn00. In this study, I also considered dN/dS calculations using codeml. The analysis is performed for each gene, which is present fully in at least 3 races without CNV or INDELs. The dN/dS values reported for each gene are presented as an average of dN/dS for each pairwise comparison. The results show that while the majority of dN/dS values lie between 0 and 1, a number of genes have dN/dS values of greater than 1 (fig 5.19).

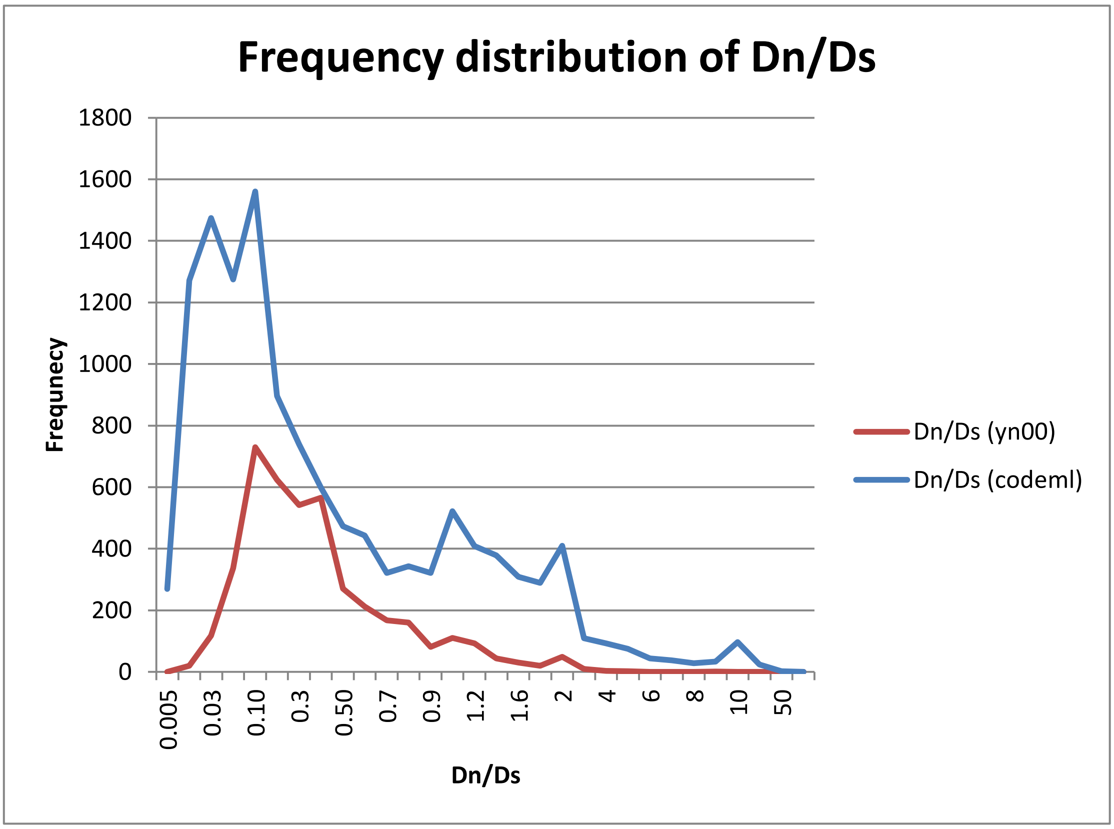



**Figure 5.19: Frequency distribution of dN/dS values calculated by yn00 and codeml. While the general trend of dN/dS distribution is the same for both methods, a clearer secondary peak of genes with dN/dS values higher than 1 are calculated using the codeml method. Yn00 predicted many more genes with a dN/dS of 0 (not shown on the graph and thus having lower area under the curve).**

The major difference between the yn00 and codeml calculation is that for each dN/dS range, more genes are identified to have a dN/dS value greater than 0 using codeml, and a clearer peak of dN/dS values greater than 1 (indicating positive selection) is present using codeml. This is also maintained when analysing the percentage distribution of the dN/dS values (fig 5.20).

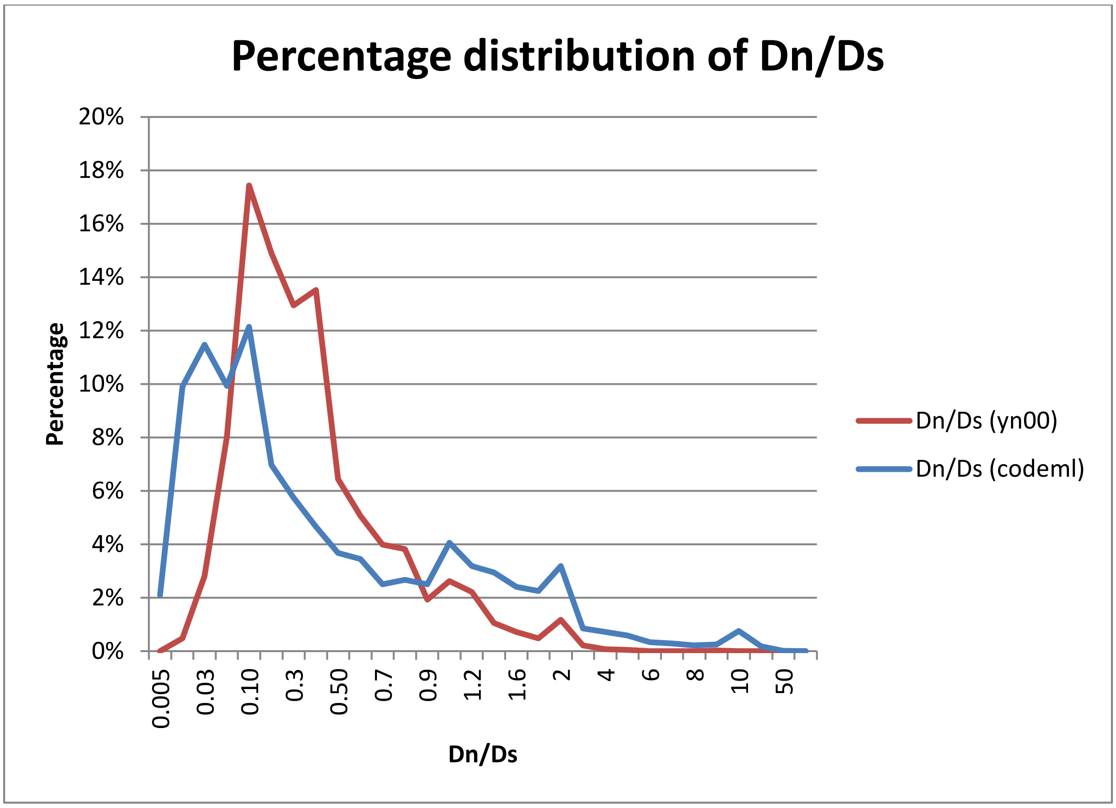



**Figure 5.20: Percentage distribution of dN/dS values calculated by yn00 and codeml. While the general trend of dN/dS distribution is the same by both methods a larger percentage of genes with dN/dS values higher than 1 are calculated using the codeml method.**

Plotting the frequency of the difference between the yn00 and codeml dN/dS calculations, it can be seen that for the majority of genes a similar dN/dS value is obtained with the two methods. However, for a number of genes the codeml calculated dN/dS is 1 to 3 units higher than that predicted by yn00 (fig 5.21). This is an important finding as it is possible that the yn00 method may incorrectly predict lower dN/dS values for genes compared to the codeml method and vice or versa.

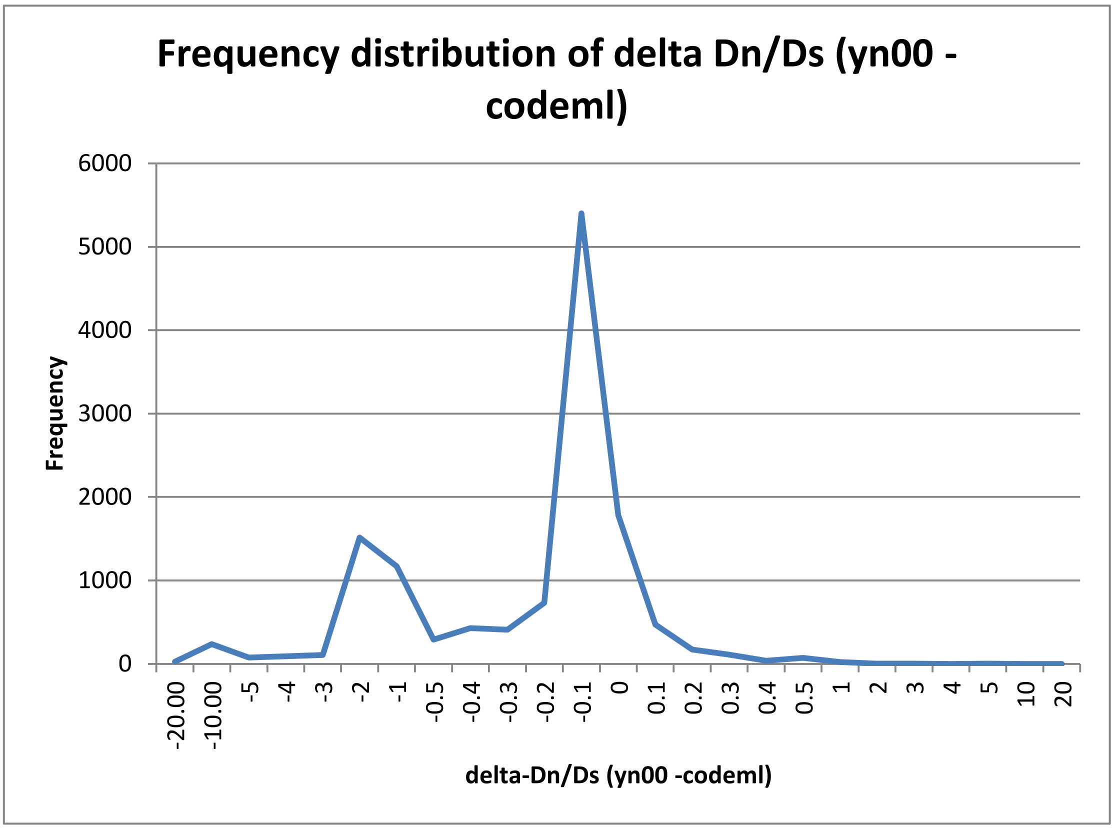


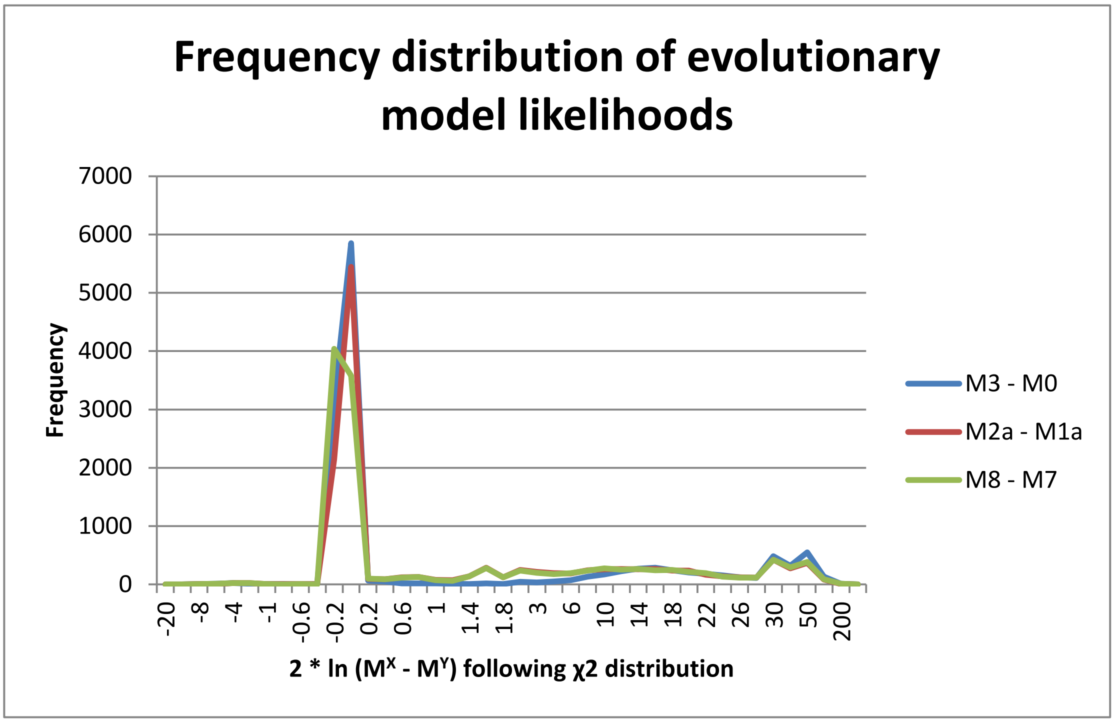

**Figure 5.21: Frequency distribution of the difference in dN/dS calculation between the yn00 and codeml method. There is little difference for most of the genes, but there is a large number of genes for which the dN/dS values calculated by codeml are 1-4 units higher than using the yn00 method.**

Among the genes with the 30 highest values for dN/dS calculated by yn00, there are 5 effector genes (appendix table 5.9). Out of the genes with the 30 highest values for dN/dS calculated by codeml, there are 3 effectors (appendix table 5.9). The effectors present in either list are mutually exclusive.

### *5.3.5.3 Evolutionary models*

The frequency distribution of each set of model comparisons (M3-M0, M2a-M1a and M8-M7) share similarities (fig 5.22). While for the majority of genes it is unlikely that there is a difference between the models, for each significance level of 95%, 99% and 99.9% (5.99, 9.21, 13.82) the difference between the number of genes between the M2a-M1a and M8-M7 comparison is very small (~25 genes more for M8-M7 at each significance level), while the M3-M0 comparison predicted ~300 more genes at each significance level. This is due to the M3-M0 comparison yielding many genes at a 99.9% level of significance. For 95% and 99% significance values, the number of genes predicted by each method does not vary more than 3%.

A




B

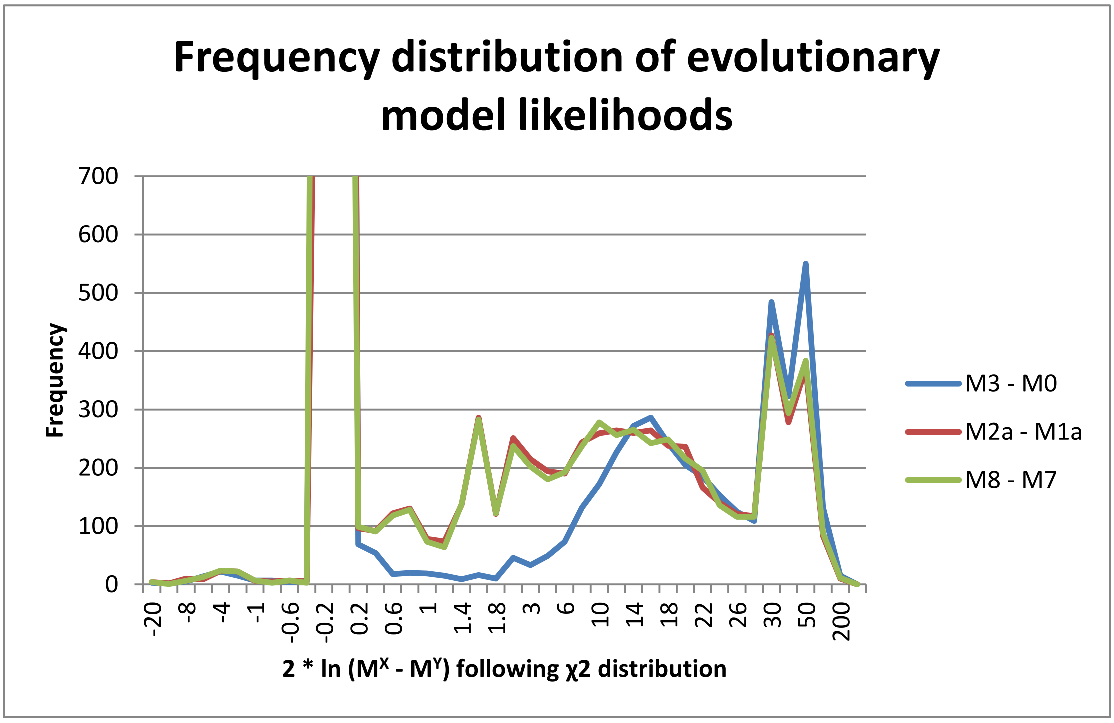



**Figure 5.22: Frequency distribution of evolutionary model testing using PAML showing full frequency spectrum (A) and partial frequency spectrum (B).**

### 5.3.6 Comparison of effectors

To better understand the evolutionary pressures acting on effectors, I compared the generated statistics between predicted effectors (including secreted proteins with homology to effectors), genes with predicted transmembrane domains and KOGs. 472 genes were selected randomly for each set for comparative analysis.

#### *5.3.6.1 S, Eta and the number of haplotypes*

In each sample effectors have a larger number of segregating sites compared to transmembrane genes and KOGs (fig 5.23). There are significantly more effectors with 13-30 and 48-63 segregating sites compared to transmembrane genes and KOGs. This trend was also observed with the total number of mutations per gene (fig 5.24), in agreement with the previous observation of correlation between the number of segregating sites and the number of total mutations per gene. The distribution of the number of haplotypes per gene show that effectors are more likely to have a higher number of haplotypes, with a secondary peak around 7-10 haplotypes (fig 5.25).

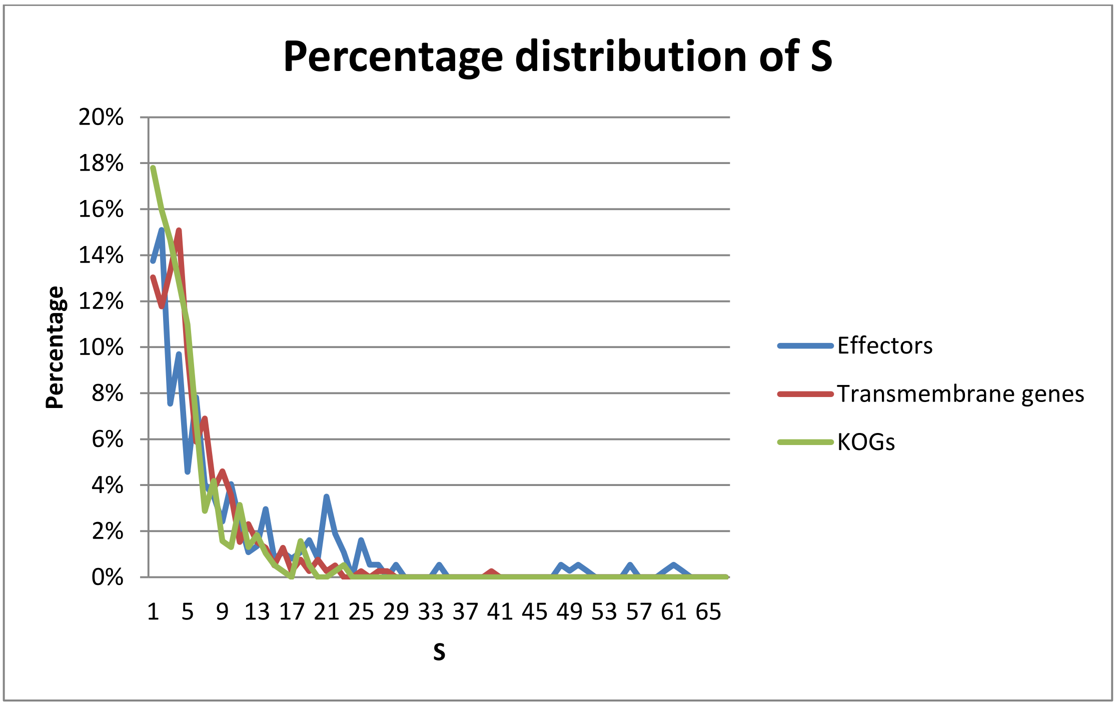



**Figure 5.23: Percentage distribution of S, the number of segregating sites, for the 472 sampled effectors, transmembrane genes and KOGs. The effectors have more genes with a higher S compared to transmembrane genes and KOGs.**

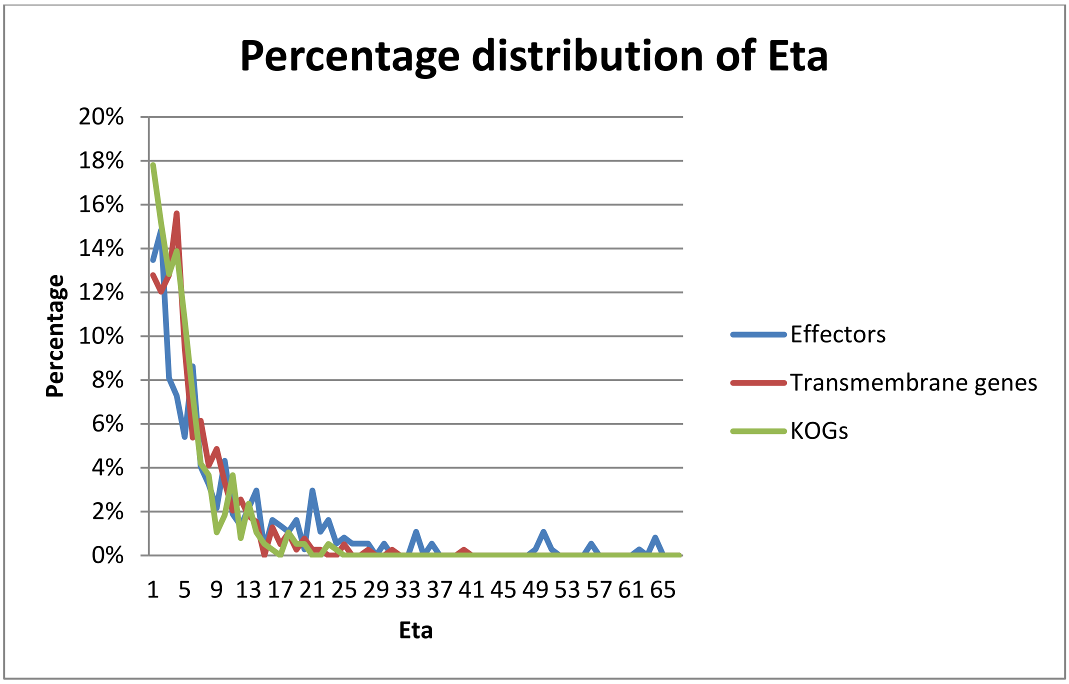



**Figure 5.24: Percentage distribution of Eta, the total number of mutations, for the 472 sampled effectors, transmembrane genes and KOGs. The effectors have more genes with a higher Eta compared to transmembrane genes and KOGs, correlating with observations for S.**

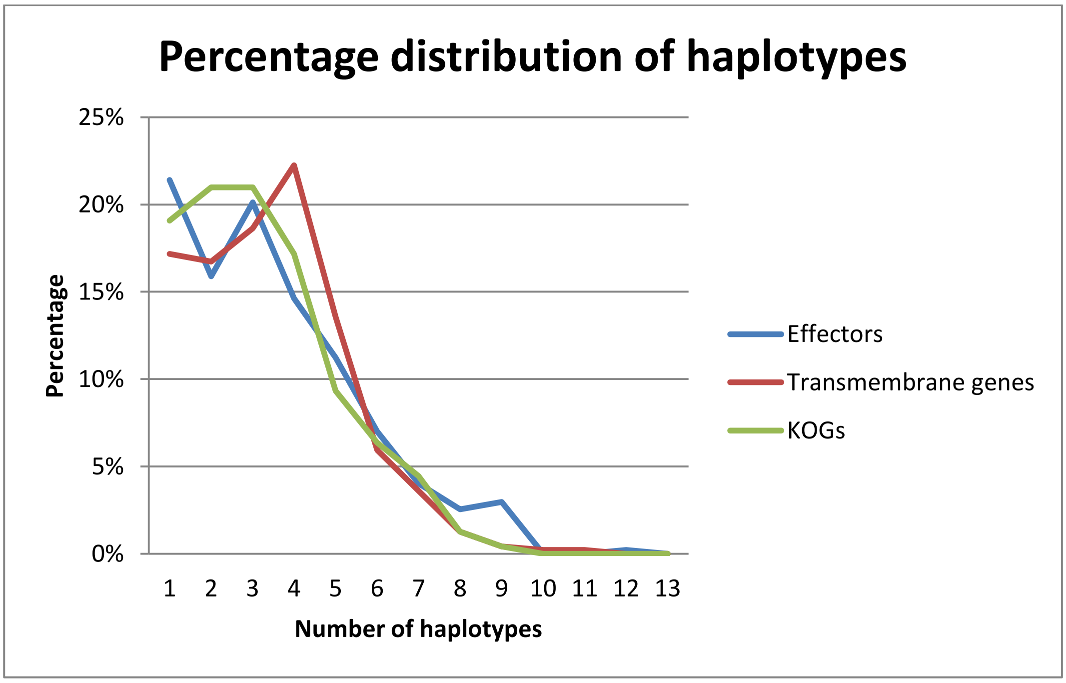



**Figure 5.25: Percentage distribution of the number of haplotypes per gene for the 472 sampled effectors, transmembrane genes and KOGs. The effectors show an elevated number of genes with higher number of haplotypes (7-10 haplotypes) compared to transmembrane genes and KOGs.**

### *5.3.6.2 Tajima's D*

No significant differences were detcted in the percentage distribution of Tajima's D for the effectors, transmembrane proteins and KOGs (fig 5.26). For lower values of Tajima's D (< -1.41), there are approximately twice as many effectors, but this represents only 2.5% of the entire set of effectors.

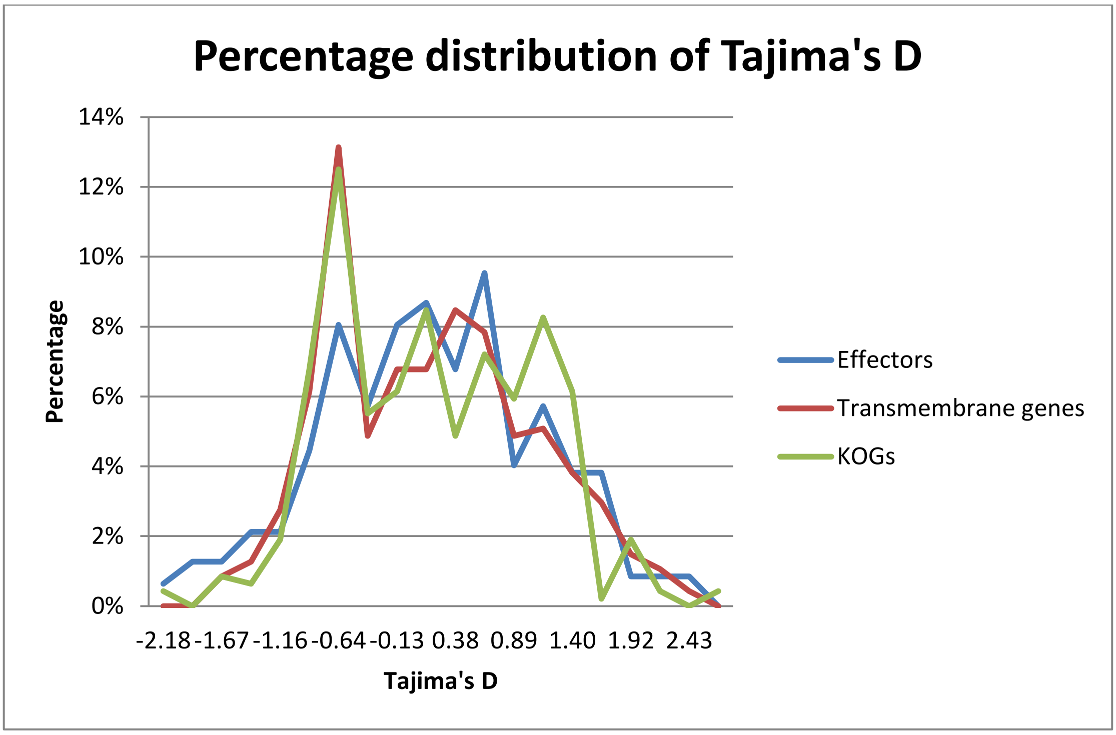



**Figure 5.26: Percentage distribution of Tajima's D per gene for the 472 sampled effectors, transmembrane genes and KOGs. The effectors show a slightly elevated number of genes with Tajima's D less than -1.41.**

### *5.3.6.3 Fu & Li's statistics and Fu's Fs*

There are no significant differences between the distribution of Fu and Li's D and for Fu and Li's F between the different sets of genes (fig 5.27; fig 5.28). The distributions of Fu's Fs is similar for the majority of the distribution of the different sets of genes, but the effectors have a large percentage of genes with Fu's Fs being larger than 4.4 with a secondary peak forming around 5.3 (fig 5.29).

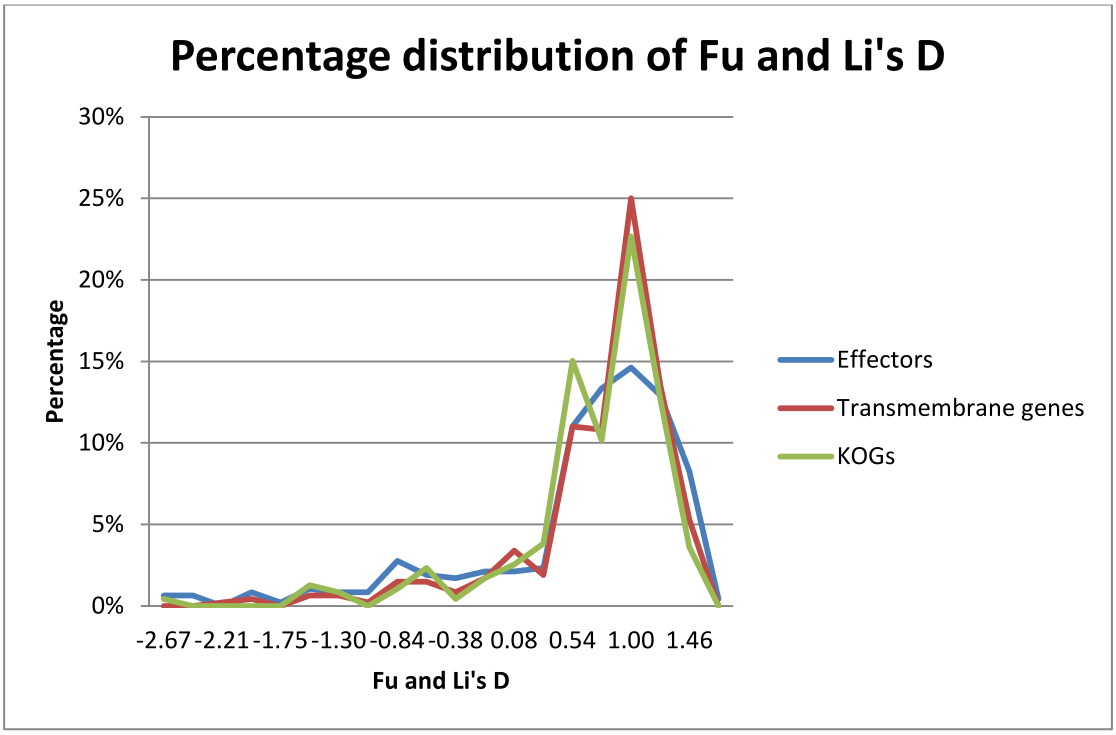



**Figure 5.27: Percentage distribution of Fu and Li’s D per gene for the 472 sampled effectors, transmembrane genes and KOGs. The effectors show a slightly elevated number of genes with Fu and Li’s D less than -0.84.**

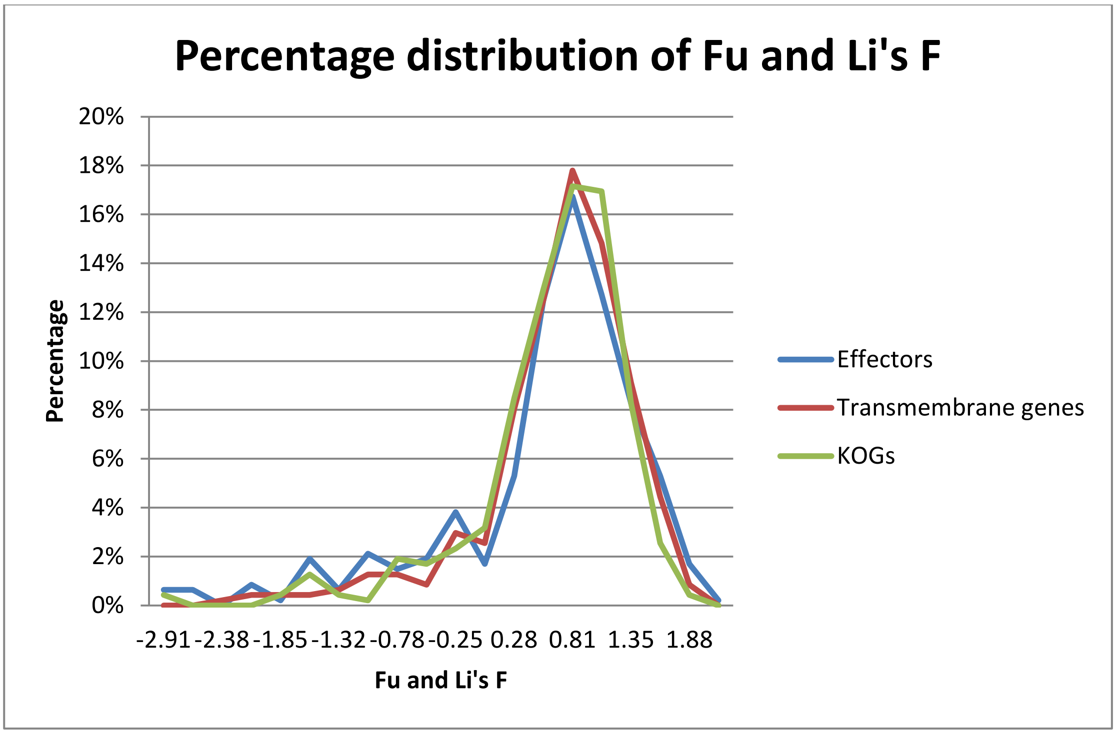



**Figure 5.28: Percentage distribution of Fu and Li’s F per gene for the 472 sampled effectors, transmembrane genes and KOGs. The effectors show a slightly elevated number of genes with Fu and Li’s F less than -1.05.**

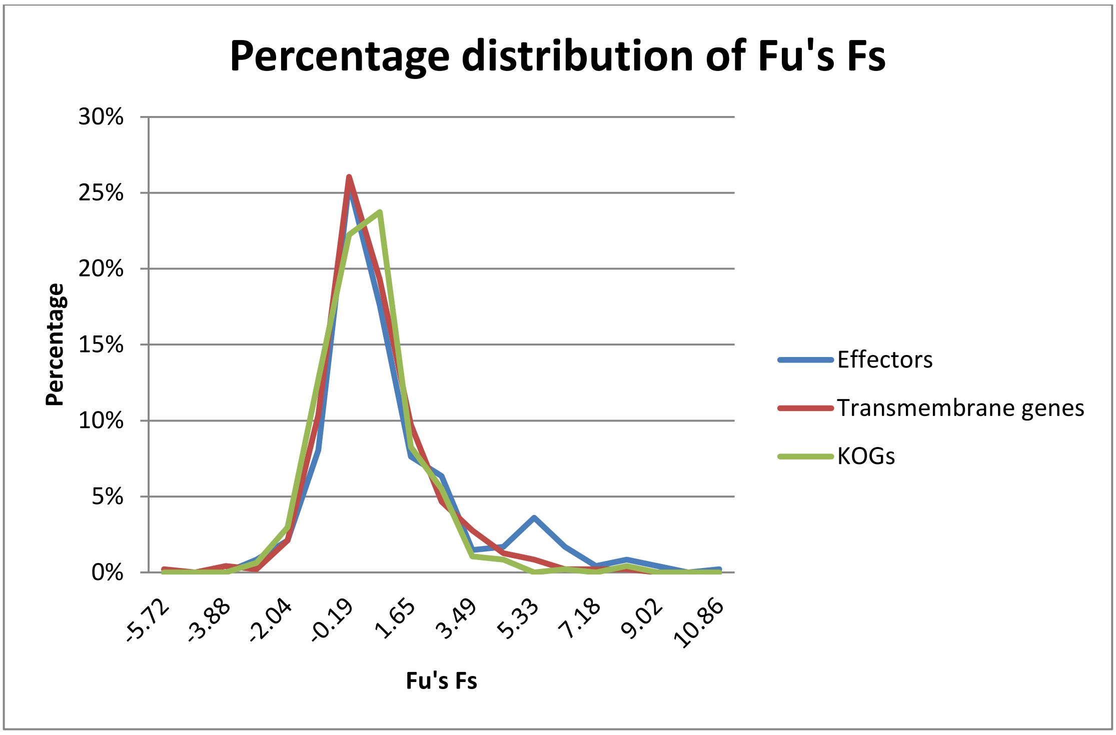



**Figure 5.29: Percentage distribution of Fu's F per gene for the 472 sampled effectors, transmembrane genes and KOGs. The effectors show an elevated number of genes with Fu's F greater than 4.41 and a secondary peak around 5.33.**

### *5.3.6.4* dN/dS – yn00 and codeml

The distribution of the dN/dS values as calculated by codeml show a clear secondary peak around 0.75 to 2.5 for effectors (fig 5.30). While this would be considered a modest value for dN/dS if inferring positive selection, the peak is clear and distinct from the other genes. The distribution of dN/dS values as calculate by yn00, does not have this same peak but instead has a few effectors with a dN/dS of greater than 1 (fig 5.31). This suggest that codeml is more suited to identify the selection acting on effectors, compared to yn00. This observation also suggest that the analysis done by Haas et al. (2009), which used yn00, may not be optimal.

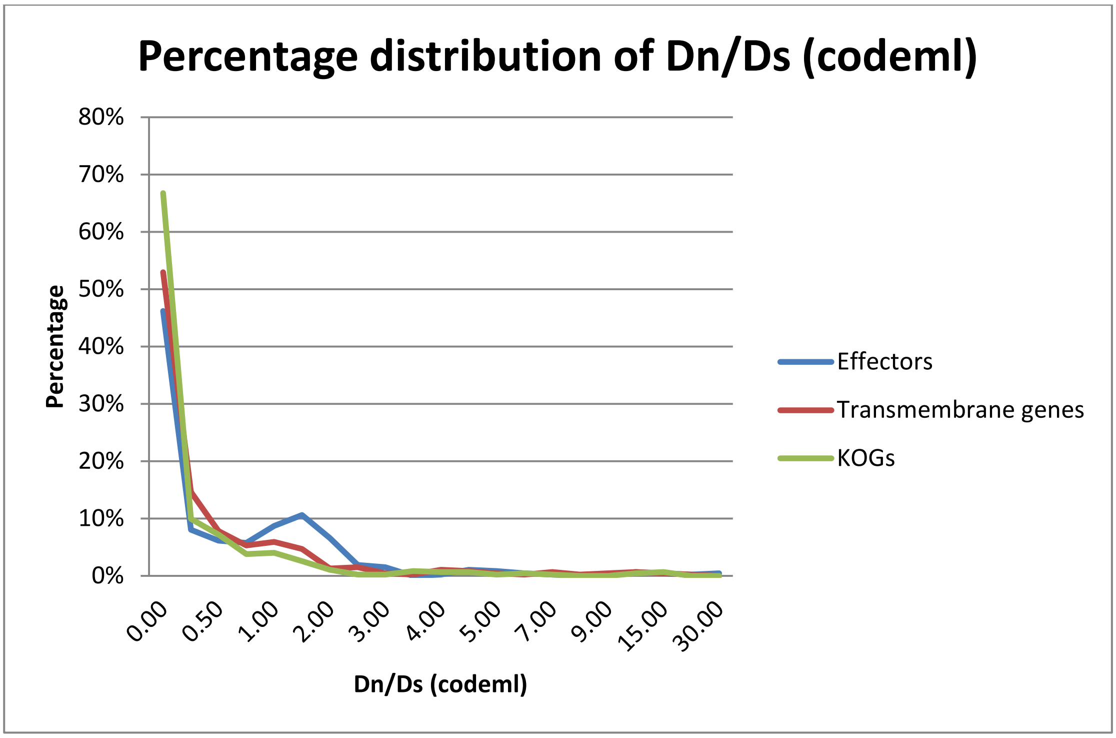
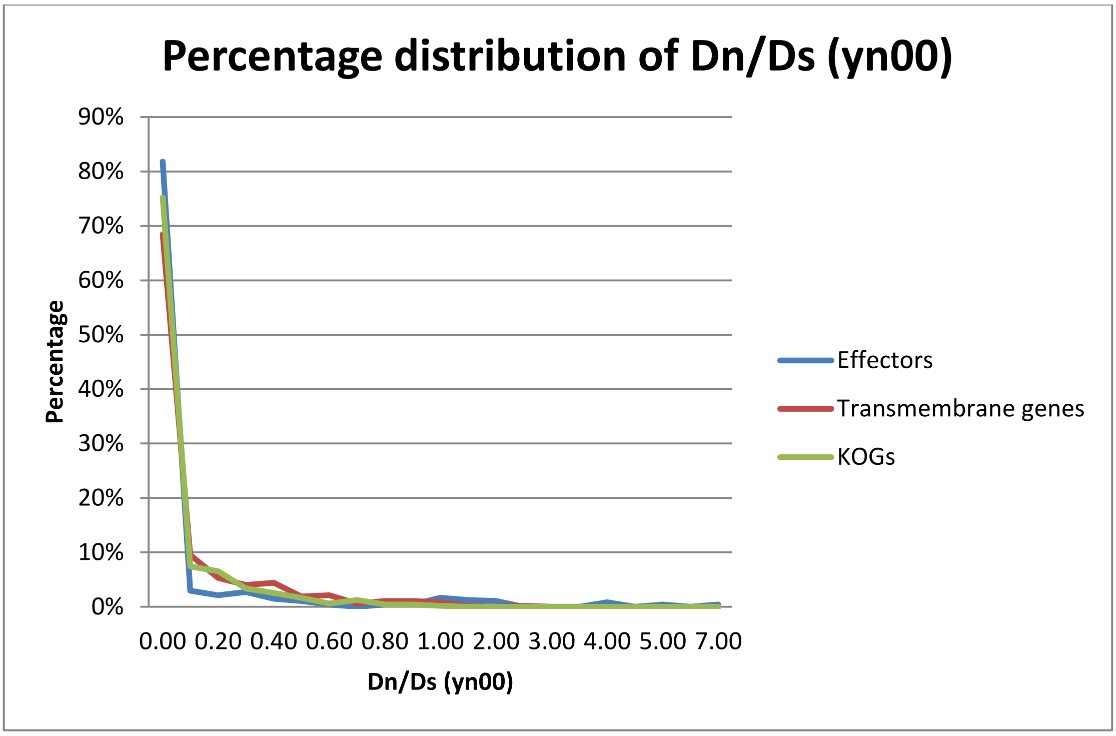



**Figure 5.30: Percentage distribution of dN/dS calculated by codeml for the 472 sampled effectors, transmembrane genes and KOGs. There is a clear peak between 0.75 and 2.5 for the effector genes, while the other genes follow exponential decay.**



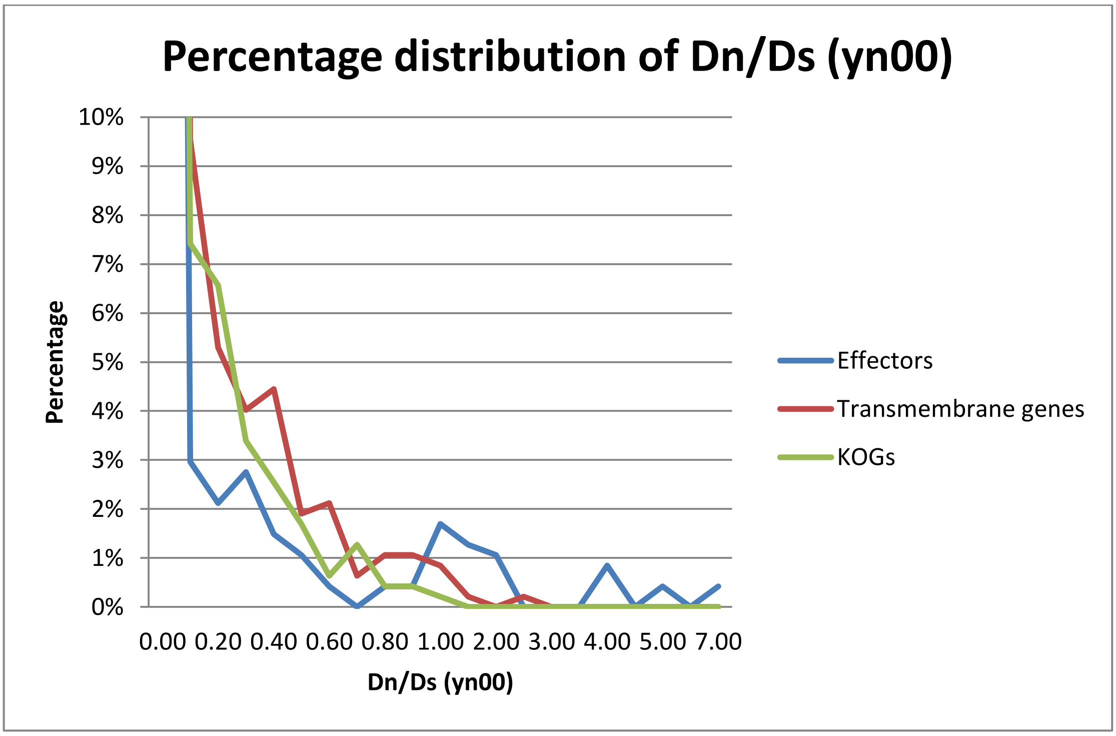




**Figure 5.31: Percentage distribution of dN/dS calculated by yn00 for the 472 sampled effectors, transmembrane genes and KOGs. There is a peak that was identified by codeml is not present, but on closer inspection it can be seen that a small number of effectors have a dN/dS value greater than 1.**

### *5.3.6.5 PAML evolutionary models*

The distributions of the likelihoods of the different sets of genes following the various PAML evolutionary models look very similar (fig 5.32; fig 5.33; fig 5.34). While the M3-M0 model comparison has the most similar distribution for the different genes, the M2a-M1a and M8-M7 model comparisons contain fewer effectors in the main peak (where positive selection is not implied), and slightly more genes distributing for higher likelihood values of being positively selected.

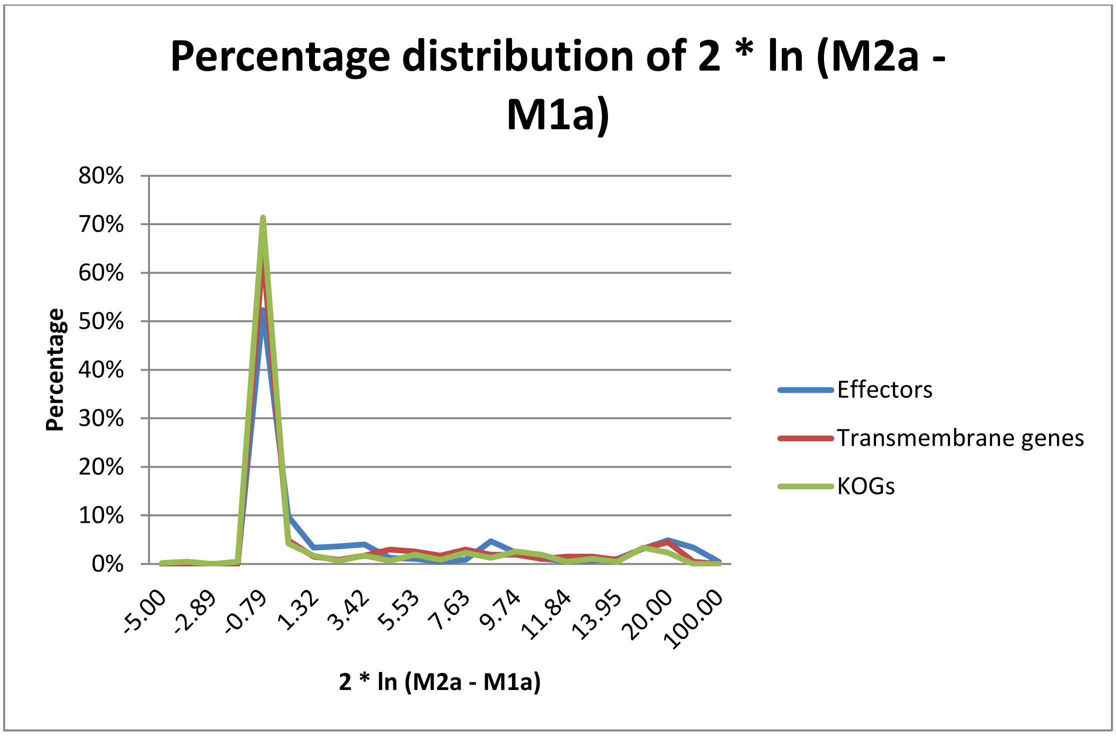



**Figure 5.32: Percentage distribution of 2 * ln (M2a –M1a) model comparison likelihoods for the 472 sampled effectors, transmembrane genes and KOGs.**

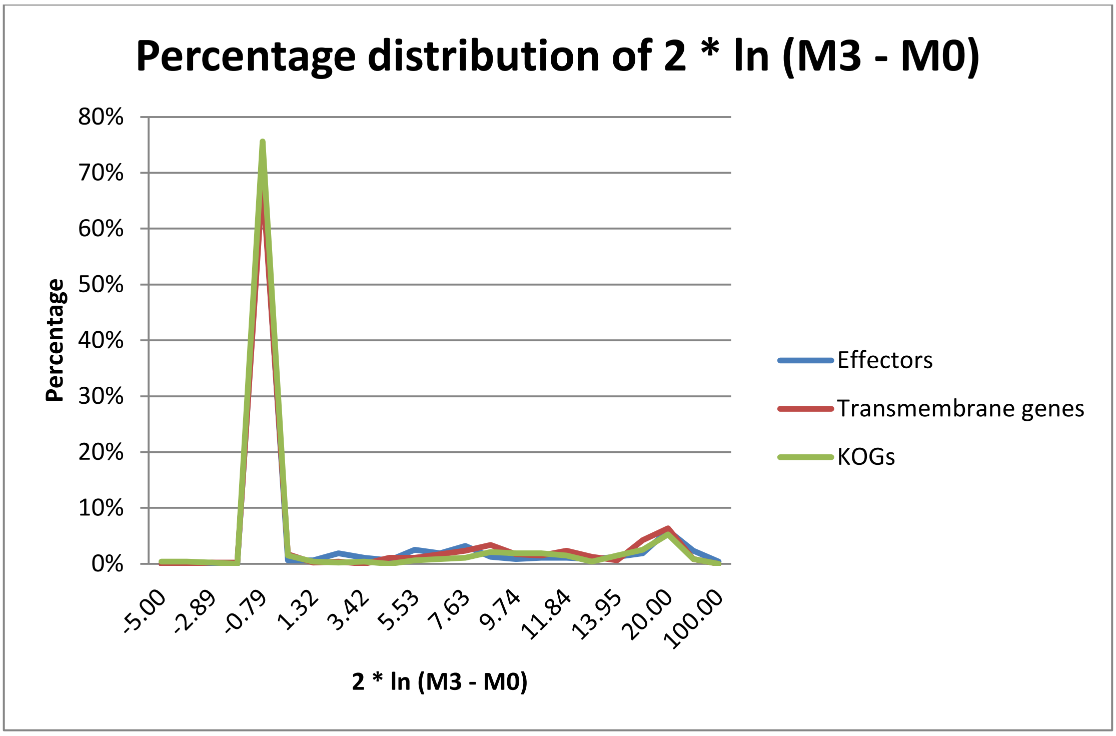



**Figure 5.33: Percentage distribution of 2 * ln (M3 –M0) model comparison likelihoods for the 472 sampled effectors, transmembrane genes and KOGs.**

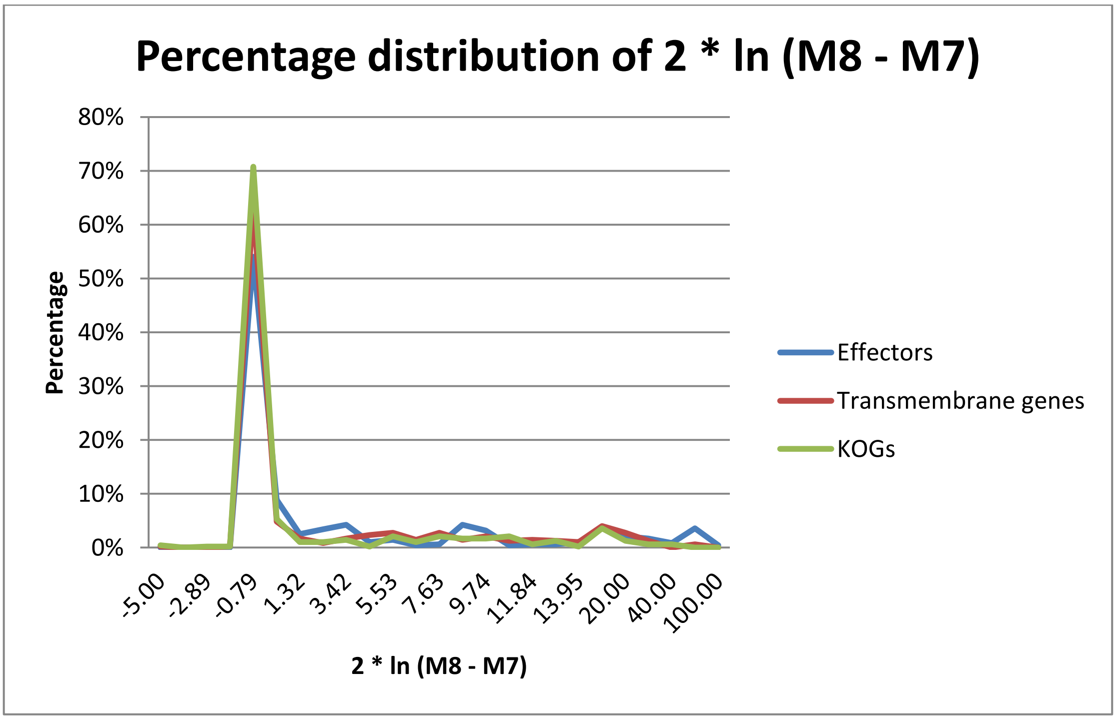



**Figure 5.34: Percentage distribution of 2 * ln (M8 –M7) model comparison likelihoods for the 472 sampled effectors, transmembrane genes and KOGs.**

### *5.3.6.6 Genes evolving like effectors*

In order to identify the genes evolving like effectors, one must understand information about known effectors. Since we are analysing within-species data and not between-species data, more robust inferences can be made from statistics generated from DNAsp rather than PAML. The DNAsp statistics for the 3 known *Hpa* effector genes, ATR1, ATR13 and ATR5 are shown in table 5.16.

| Effector? | n | S | Eta | Hap | TajimaD | Sig_TD | FuLiD* | Sig_FLD | FuLiF* | Sig_FLF | FuFs |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ATR1 | 12 | 62 | 64 | 5 | 0.2156 | n.s. | 1.6512 | ** | 1.452 | # | 8.821 |
| ATR13 | 10 | 15 | 15 | 3 | 0.3358 | n.s. | 1.5205 | ** | 1.3815 | n.s. | 5.4 |
| ATR5 | 16 | 21 | 25 | 8 | -1.2807 | n.s. | -0.5893 | n.s. | -0.9041 | n.s. | 0.144 |

**Table 5.16 – DNAsp neutrality statistics for ATR1, ATR13 and ATR5**

Analysing the 3 known effectors with DNAsp neutrality tests it can be seen that ATR5 has no significant test scores. ATR1 and ATR13 have high scores for Fu & Li's D* and F*, and Fu's Fs. They also have positive values for Tajima's D. It is interesting to note that the number of sequences sampled is 10 and 12 – this means that there were 5 and 6 races (of 8) for which sequences could be analysed. All the genes were filtered using the following criteria:

- Fu's Fs > 4
- Fu & Li's F* > 1
- Fu & Li's D* > 1.2
- Tajima's D > 0
- Number of samples (n) < 14

15 genes meet these criteria (table 5.17). 13 of the 15 genes are predicted to be secreted and 9 are highly similar to, or are predicted to be *Hpa* effectors. There are 3 groups of homologous genes. One group, containing genes 808594, 808490 (the most distant) and eff_g11324, are amongst the genes with the highest values of Fu's Fs. They also show homology to *Phytophthora* effectors, but have lost the RXLR motif as shown in an alignment with PiTG_09732, and RXLR effector gene from *P. infestans* (fig 5.35). This could suggest that there is strong selection against the effector so it lost the hypothetical RXLR translocation mechanism, or that the RXLR motif is not required for translocation but

instead a more general conserved region is required for translocation. The second group of homologous secreted proteins, containing genes803332, 802865, pasa_gi_SuperContig14_289 and eff_g17825, are all annotated effectors or effectors homologs. The third set of homologous effectors was ATR1 and ATR13 which shared distant homology over the first 100 amino acids. There were 3 genes that were not predicted to be secreted. One of the non-secreted genes, 811161, showed homology to *Phytophthora* choline/Carnitine O-acyltransferase, which was also confirmed by InterproScan predictions. Another non-secreted gene, 903729, was homologous to *Phytophthora* tRNA nucleotidyltransferase, again confirmed by InterproScan predictions. The last of the non-secreted genes, ceg_gi_SuperContig67_104, was homologous to myosin-like proteins. Myosins are a family of ATP-dependent motor proteins and are responsible for actin-based motility (Wessells et al., 1971). Interestingly, InterproScan results revealed that the gene contained a Phox homologous domain which a molecular function of phosphoinositide binding. It has recently been suggested that oomycete RXLR motifs enable binding to the phospholipid, phosphatidylinositol-3-phosphate (Kale et al., 2010). This could suggest that ceg_gi_SuperContig67_104 could also be involved in the virulence mechanism.

| Gene | Effector? | Secreted | n | S | Eta | Haplotypes | TajimaD | FuLiD* | FuLiF* | FuFs | BLAST | Homologous? |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **808594** | HaRxLL38_like | S | 14 | 36 | 36 | 4 | 1.2 | 1.6 | 1.7 | 10.3 | Pitg_09732 (RXLR) - Lost RxLR | + |
| **ATR1_Emoy2** | ATR1 | S | 12 | 62 | 64 | 5 | 0.2 | 1.7 | 1.5 | 8.8 | ATR1 | $ |
| **808490** | | S | 8 | 61 | 64 | 4 | 1.2 | 1.7 | 1.7 | 8.4 | Pitg_09732 (RXLR) - Lost RxLR | + |
| **eff_g11324** | | S | 14 | 34 | 34 | 5 | 1.4 | 1.6 | 1.8 | 7.7 | Avh153a1 (Ps); PITG 15110 (RxLR); Pitg_09732 (RXLR) - Lost RxLR | + |
| **809253** | HaRxL95 | S | 14 | 18 | 18 | 4 | 1.4 | 1.5 | 1.7 | 6.2 | Avh347 (Ps) | |
| **807906** | HaRxLL434 | S | 12 | 16 | 16 | 3 | 0.2 | 1.5 | 1.3 | 6.1 | HaRxL89 like | |
| **803332** | Emoy2cDNA_HpRXLR91 | S | 14 | 26 | 26 | 5 | 1.0 | 1.6 | 1.6 | 5.7 | Emoy2cDNA_HpRXLR91 | * |
| **802865** | HaRxL53_like | S | 12 | 17 | 17 | 4 | 1.6 | 1.5 | 1.8 | 5.5 | HaRxL56 like | * |
| **811161** | | | 14 | 14 | 14 | 4 | 1.9 | 1.5 | 1.8 | 5.4 | Phytophthora choline/Carnitine O-acyltransferase | |
| **813534** | ATR13 | S | 10 | 15 | 15 | 3 | 0.3 | 1.5 | 1.4 | 5.4 | ATR13 | $ |
| **803729** | | | 14 | 16 | 20 | 5 | 1.4 | 1.6 | 1.7 | 4.8 | tRNA nucleotidyltransferase (sojae and infestans) | |
| **pasa_gi_SuperContig14_289** | HaRxL53 | S | 12 | 14 | 14 | 4 | 1.5 | 1.5 | 1.7 | 4.5 | HaRxL53 | * |
| **eff_g17825** | HaRxL78_like | S | 10 | 12 | 12 | 3 | 0.0 | 1.5 | 1.3 | 4.3 | HaRxL78 like | * |
| **802512** | | S | 14 | 21 | 21 | 5 | 0.4 | 1.6 | 1.4 | 4.2 | Conserved hypothetical protein | |
| **ceg_gi_SuperContig67_104** | | | 8 | 11 | 11 | 3 | 0.9 | 1.5 | 1.5 | 4.0 | myosin-like protein | |

**Table 5.17: List of genes showing similar values to those of ATR1 and ATR13 from DNAsp analysis. N = number of samples; S = number of segregation sites; Eta = total number of mutations; BLAST = best blast results; Homologous? = shows homologous groups.**

```
Pitg_09732      MRLYSIAFLVAVALLA--KVDRATPS--ELTTADY----PTLTHSITGRQNDVTPRLLRR
eff_g11324      MRLHAIELWAALTLVSIEGDSAATAS--ELPAAKSRPPPPSSAYNSIDFQSEITSRKAES
808594          MRLHAIELWAALTLVSIEGDSAATAS--ELPAAKSRPPPPSSAYNSIDFQSEITSRKAES
808490          MRVHCLVFLASSALSA-RGDGVLEPADPDVAAPYSSSVARSLTENDNDNVPATMSFKSEG
                **::.: : .: :* :     .   .:  ::.:.        : : .  .       .  .

Pitg_09732      MEEDDEERGIGGTAISDLAAKLKSRTSSLVDKAVNLKAREMKAARAMRFGEIDDTLASSN
eff_g11324      PAAISEER---GTDRVSFAASWIRQAAEVEDTAWTTRA-------------VDKLQLEDN
808594          PAAISEER---GTDRVSFAASWIRQAAEVEDTAWTTRA-------------VDKLQLEDN
808490          PVNDDERI---MTGVSSLVGKVQERIHTAVQ--------------------GEKQKTFNF
                    .*.     *   .:...   :     :                    :.    .

Pitg_09732      INNLITKLKEINDKNRLVK--------V-------SLLGTLTTKYGDDAVAAALVTAKRS
eff_g11324      ISAAVLKLKSPTPLNKVLRIGNEKKPGV-------PLNEDSTERYALQNAVDDLVHASQS
808594          ISAAVLKLKSPTPLNKVLRIGNEKKPGV-------PLNEDSTERYALQNAVDDLVHASQS
808490          LAQSRTESNTQTFIENALKTKEWKALSVRYERSGGSMITPLLARYDCAEVARVLAPALKF
                :     : :  .  :. ::         *       .:       :*     ..  *. * :

Pitg_09732      ADS------PSVAQQIQKLQTEQLMK-WKDSGKSLGSVSKLLNFR---------YNRGLG
eff_g11324      TSE------KE--KRIADMWWKQLCALFVRYDQPIAGVAELLNVD---------GLKGLT
808594          TSE------KE--KRIADMWWKQLCALFVRYDQPIAGVAELLNVD---------GLKGLT
808490          NKDGVLQLPKKPASVKEKLAVDMLKH-WGEEKRPIRTLFTDLELNVQTETSPLYYNLG-G
                 ..        .  .   .:  . *   :    :.:  :    *:.              *

Pitg_09732      QKFQVLDEYAKLVKQS-DDTLLTTLIKSVGGEDNLGGVLYGARTNSAATKNKATKLENIL
eff_g11324      EKLEMLKYYIGVSPTSGPNTFLETMTANLGSEKDLVSFIGRAKLNDDVER-RATELEGLQ
808594          EKLEMLKYYIGVSPTSGPNTFLETMTANLGSEKDLVSFIGRAKLNDDVER-RATELEGLQ
808490          SRMRVLEAYRQYAKID-DQSYLDAVKTGFGGKTIFLEHLGNAKTFWNTAD-KADELEKIA
                .::.:*. *        .   :: * ::   ..*.:   :     :   *:     .    :* :** :

Pitg_09732      IERWTRGEQLPANVFQWLRLSGDVDDAFTASNLNRFMKYVDDFNAKNPGQKKPVLKLYTQ
eff_g11324      IAKWQAENKDPVELLNSMQMDKSM-DALISPALYTVMKYIAEHNLEHPDKKFSVLTPVRE
808594          IAKWQAENKDPVELLNSMQMDKSM-DALISPALYTVMKYIAEHNLEHPDKKFSVLTPVRE
808490          L---NTPSLNPLDVLKKTKPNSLRELIFDTDTFRLVSKYVE----ANPARKTTVLKVMMD
                :      .  * ::::  : .       : :  :  . **:       :* :* .**.   :

Pitg_09732      AFG-DAPVMRKLLSAMDDSTTNVA---AKKLLVE--RGVQKDNQSLGSMLRALNIDINQP
eff_g11324      RLG-DTHVMGALVAARSNGQSKVTEGFANELLSEMKKLWKGEGKSKREFLAA--------
808594          RLG-DTHVMGALVAARSNGQSKVTEGFANELLSEMKKLWKGEGKSKREFLAA--------
808490          SLGGEAHLLQALADFR----GKPTDAWLSCRLALFNQWRTTDKVTSTSGLQ---------
                 :* :: ::  *          : :    .  *    :     :   :   . *

Pitg_09732      TSIVNQKIDVLEQLAEVKEVRQVFIKAMSTQVGGNKMLAKILEGAEAATLQKKQFATWIG
eff_g11324      -----------SGLLDGQERKK----------------KK----------KKKK------
808594          -----------SGLLDGQERRNGIPGSASQRVTGCTVFSS----------TKRKF---VP
808490          --------DRLNSLSDAKQKRR--------------------------------------
                           . * : :: :.

Pitg_09732      EGVTPENFWKMIYKTETASNPVEEKIMAKFTAFYQSQKPGN------------------
eff_g11324      --------------------------------------NWAKHR--AKHELKRN----
808594          EGDVPME--TVMYRAAEATR-VEVKPCRELGLFFRGSAHPKLGDHRFICTDDMQRGQSAS
808490          ------------------------ITRMYTAFFEGHQ*--------------------

Pitg_09732      ------------------
eff_g11324      ------------------
808594          FARRRPFPVTYSFCATL*
808490          ------------------
```

**Figure 5.35: Sequence alignment of 3 highly evolving secreted *Hpa* genes with a *P. infestans* effector. The RXLR-EER amino acid motif region is highlighted in yellow.**

The gene models were searched to identify genes with a similar 5' regions using BLASTp, but no genes were found. The family of 3 novel effector candidates identified using this method are similar to ATR5 in that they have lost the RxLR motif but maintain the EE motif. ATR5 was shown to contain the W-Y motif. I performed a hidden Markov model search using the HMM from Boutemy et al., (2011) which reveals that 2 of the homologous genes (eff_g11324 and 808594) also contain the W-Y motif. This strongly suggests that these 2 genes are likely to be real effector candidates with possible.

## 5.4 Summary

In this chapter I presented the VariTale pipeline. While there are a number of pipelines that are able to identify synonymous and non-synonymous mutations in coding regions (Cingolani, 2011; Schneeberger et al., 2009), there are no pipelines that are able to make evolutionary inferences from resequencing data. The method builds on pervious manual methods used by Haas et al. (2009) and extends these methods to additionally consider:

- Breadth of coverage
- INDELs
- CNV
- Haplotype information
- Divergence from neutral evolution
- dN/dS calculations using codeml
- PAML evolutionary model analysis

I have shown that the Poisson distribution can be used to model coverage at varying read depths in *Hpa*, confirming previous findings (Xie and Tammi, 2009).

I also showed that the use of Stampy to align discordant read pairs improves the number of reads mapped to the genome by 3.72%, allowing for better variant calling. The alignment of 8 races of *Hpa* to the reference genome has revealed regions of the genome that have variable breadth and relative depths of coverage indicating duplicated, missing and hemizygous regions in the genome. The alignment has also revealed a number of caveats of the *Hpa* Emoy2 v8.3 genome assembly, in regions that are not covered by reads and therefore may be contamination, and other regions that have been collapsed.

I have shown that the number of variant calls differs between *Hpa* races, while they show consistency in their difference from the reference strain for the variant types. I have reported the homozygous and heterozygous SNPs, INDELs and their protein coding effects. The analysis revealed a 1:1 ratio insertions to deletions within each race which suggests an equal rate of insertion and deletion accumulation. A nucleotide composition equilibrium was revealed by analysing the rate of nucleotide mutations. We found that A/T ⇔ C/G mutations account for ~65% of all mutations. I also observed that the ratio of SNPs on exons to introns was 3:1, and the ratio of SNPs on coding regions to non-coding regions was 1:1 for most *Hpa* races.

I presented the protein coding effects of the variation on the genes. An elevated level of non-synonymous to synonymous SNPs (3.5:1) was observed, suggesting that there is evolutionary pressure acting to change gene functionality. I observed an overrepresentation of INDELs that lead to codon insertions and deletions, and the same overrepresentation is maintained for net INDEL length over coding regions of genes.

I also characterised the effect of SNPs on *Hpa* genes in an evolutionary framework, analysing divergence from neutrality and directional selection using DnaSP and PAML. I showed that at least 10 haploid samples are required for reliable analysis of this nature. I also revealed differences in the dN/dS calculations between yn00 and codeml, which may reveal additional information about selection on genes that is not always analysed.

This analysis has allowed for high-throughput characterisation of effectors, revealing that effector genes are likely to have a high number of segregating sites and total number of mutations, a higher Fu's F statistic and exhibit higher dN/dS values compared to transmembrane genes and KOGs. I also show that amongst the 15 genes showing signs of accelerated evolution sharing traits with ATR1 and ATR13, there are 12 secreted genes of which 11 are homologous to known effectors, providing further evidence that effectors are amongst the fastest evolving group of effectors. In these 12 effector like genes, there were 3 genes that show homology to *Phytophthora* effectors and share the classical characteristics of an effector but do not have an RXLR motif. This suggests that these may have been selected against as being effectors in *Hpa*, or that the RXLR motif is a specialisation of a more general effector motif. If it is the case that these genes are real effectors, they may contribute to understanding the effector translocation mechanism in *Hpa* and other oomycetes.

# Chapter 6 – General discussion and outlook

## 6.1 Modern day genome assembly

Genome assembly projects were expensive and time consuming. The onset of second generation sequencing has enabled the transition into a new era of genomics. Traditionally genome assemblies have employed Sanger sequencing technologies, but the trend is moving towards the usage of short read assemblies for small genomes (Farrer et al., 2009) and complex eukaryotic genomes alike (Li et al., 2010; Kemen et al., 2011). While the accuracy and length of Sanger sequenced reads are far superior to all second generation sequencing technologies, the cost and throughput of second generation sequencing techniques have enabled it to become the main contributor of genome sequences in data repositories.

In chapter 3 I described a pseudo-hybrid genome assembly method that was employed for the *Hpa* genome assembly. This hybrid approach makes use of the advantages of 2 different technologies: the read length and accuracy of Sanger reads, to resolve complex regions and optimise contiguity, and the depth of Illumina sequencing, to correct 3520 sequencing errors and provide an additional 4 Mb of *Hpa* sequence. It would also have been possible to combine two different second generation sequencing technologies, utilising the longer reads of 454 sequencing in combination with the high throughput of Illumina. By combining 2 sequencing technologies I have demonstrated that their strength can be combined and the weaknesses of the individual technologies can be overcome. The weaknesses in the technologies does not only include the previously mentioned read length, accuracy, throughput and cost but also other factors, which are not often considered, such as limitations in library preparation, bias in the sequencing methods and contamination. I have shown that an additional 4 Mb of sequence in the *Hpa* genome assembly was discovered using Illumina reads and Velvet assembly. Within these 4 Mb, there are genes that do not have homology to any other known sequence – these could be genomic regions truly unique to *Hpa* or could be highlighting a limitation of the previous sequencing technologies. While in-depth analysis was not performed on genomic regions that were present in the Sanger sequence but not the Illumina sequence, it is possible that during fragment size selection certain genomic regions did not fractionate to the selected size due to the physical properties of the DNA. By using this hybrid method I also showed that contamination in the reads can be identified.

Genome sequencing projects thus far do not usually consider heterozygous variation within the organism. I have shown that analysing heterozygosity in a single organism can reveal insights into interesting aspect of biology and provide naïve estimates of selective pressure. In chapter 1 I discovered that effector genes were 5 times more heterozygous than other genes, from which I hypothesised that effector genes are accumulating mutations due to selection pressures, which I prove later in chapter 5.

The method described in chapter 3 was not what I consider a true hybrid assembly, since both sequencing technologies were not used from the start of the assembly, but rather the Illumina sequencing was used to improve the existing Sanger assembly. Although the method did add 4 Mb of new sequence, it had the limitation that it was based on an existing assembly and may not be able to resolve complex errors. One such error is collapsed regions in the genome. These regions can be identified by a higher than expected depth of coverage of Illumina reads over a genomic region, but they cannot be resolved easily using this technology. I have been able to identify a number of genomic loci where gene paralogues have been collapsed into a single region due to high sequence similarity, including the *HaRxL79* gene family (data not shown). This gene family was assembled correctly in previous versions of the genome, but was collapsed in the v7 assembly, on which the v8.3 assembly was based. I have attempted to assemble the *Hpa* genome *de novo* using a hybrid assembly program, Mira3 (Chevreux et al., 2004). With this method I was able to correctly assemble the *HaRxL79* gene family, but at the expense of genome contiguity (results not shown). There is also additional complexity in that *Hpa* Emoy2 is a heterozygous wild organism with variation between the parental haploid genomes. Since resolving collapsed repeats using Illumina sequencing is not a trivial computational problem and programs do not exist to resolve this issue, I did not include this problem in the scope of my project.

DNA sequencing techniques are improving at an ever faster rate. At the start of my project the 3 major second generation sequencing technologies were Solexa, 454 and SOLiD. During the project, the evolution of the sequencing machine saw transitions from the Solexa to the Illumina Genome Analyser (GA), to the Genome Analyser 2 (GA2), to the Genome Analyser 2x (GA2x). Illumina now also has 4 other sequencing machines in addition to the GA2x – the MiSeq, HiSeq1000, HiSeq2000 and HiScanSQ. There has also been an increase in the number of sequencing technologies: single molecule sequencing, nanopore sequencing, ion semiconductor sequencing and DNA nanoball sequencing. While the rate of emergence and availability of all these sequencing technologies is further empowering the genomics era, enabling projects such as the 1000 genomes and 1001 genomes projects, it does make for a very dynamic current best practice.

## 6.2 The nature of obligate biotrophy

In chapter 4 I used Sanger sequenced ESTs and Illumina sequenced cDNA to assist with gene model prediction and evaluation. With the semi-automated combining of gene predictions from 6 methods, over 3 iterations of improvement, we establish the current *Hpa* gene models. While the gene models are dependent on the caveats of the genome sequence (e.g. not correctly predicting genes in the collapsed regions of the genome), they are the most complete to date, based on the expression data available to us.

With these gene models we were able to make inferences about obligate biotrophy. The current dogma dictates that obligate biotrophy is caused by a combination of a loss of biosynthetic pathway elements and the acquisition of mechanisms that facilitate growth on a host by evading recognition, suppressing defence and/or reprogramming nutrient trafficking (Kemen et al., 2011). With the *Hpa* gene models we were able to identify impaired nitrogen and sulphur assimilation pathways, which explains the obligate biotrophic nature of *Hpa* (Baxter et al., 2010). Similar observations of impaired nitrogen and sulphur metabolism were made by Duplessis et al. (2011) in the obligate biotrophic rust fungi, *Melamspora larici-populina*, and Kemen et al. (2011) in the obligate biotrophic white rust oomycete, *A. laibachii,* which are remarkable observation of independent convergent evolution towards obligate biotrophy. While many families of virulence related proteins that are present in *Phytophthoras* are also in *Hpa*, we see a dramatic reduction (or an expansion in *Phytophthora* species) in the number of genes encoding for RxLR effectors and other classes of virulence related proteins. There is further evidence that *Hpa* evolves toward a biotrophic lifestyle in the loss of pectin meythyl esterases, which have been implicated in host cell wall modification (Pelloux et al., 2007), and may be involved in triggering host defence. It is also interesting to note that while some of the more closely related species to *Hpa* such as *Phytophthora* and *Pythium* are able to induce host necrosis, *Hpa* shares many traits with biotrophic fungi that have lost degradative enzymes and nitrogen and sulphur assimilation pathways (Kamper et al., 2006; Martin et al., 2008; Spanu et al., 2010). This suggests independent but convergent evolution toward obligate biotrophy in both fungal and oomycete lineages.

The onset of the genomics era has shed new light on obligate biotrophy in recent years, and with more genome sequences becoming available there is the scope to investigate further the relationship between saprotrophy, necrotrophy, hemi-biotrophy, biotrophy, commensalism and mutualistic symbiosis.

## 6.3 High throughput analysis of evolutionary signatures

The signature of evolution appears as variation in the genome, so identifying and analysing sequence variation allows us to elucidate evolutionary mechanisms acting on the genome, genomic regions and genes of interest. Newer sequencing technologies have led to unprecedented levels of sequence data generation. Tools to analyse this type of data have progressed from being able to identify SNPs to presence/absence polymorphisms, INDELs, recombination, CNV, and most recently the ability to infer the protein coding effect of SNPs. In chapter 5, I took this a step further and put the predicted sequence variation into an evolutionary context in an automated pipeline termed VariTale. While studies exist where dN/dS is calculated using yn00 from the PAML package (Haas et al., 2009) the VariTale pipeline considers more tests including tests to resolve parental haplotypes, neutral evolutionary models and nucleotide divergence using DnaSP, and PAML selection-based evolutionary models.

Among existing programs used to test for selection (Tamura et al., 2011), DnaSP has the most functionality for testing for neutrality and nucleotide divergence and PAML implements the most sophisticated evolutionary selection models. The inclusion of DnaSP in the pipeline is not ideal, due to its inability to resolve phased data in the batch processing mode, the issue of cross platform compatibility (DnaSP runs only on Microsoft Windows, while PAML runs on all platforms), and most importantly its semi-automated running procedure require manual intervention making it impossible to script the entire pipeline. Ideally I would implement my own methods to perform the neutrality and divergence tests, but this would require significantly more time investment.

While these types of hypothesis testing are insightful, it is important to have a sufficiently large sample size to generate robust and meaningful results. My analysis revealed that robust results are generated at a minimum of 5 diploid samples or, by extension, 10 haploid samples, while there is little increase in robustness at more than 5 samples. Another unique feature of the VariTale pipeline is its ability to resolve parental haplotypes. This is a novel feature that has not previously been considered in second generation sequencing analysis. It is important that future resequencing analysis pays more attention to heterozygosity and treating diploid non-inbred wild organisms as 2 haploid samples to maximise the added value of high throughout sequencing in effectively providing twice as much population data. While I have implemented this method successfully for SNPs, I have yet to do so for INDELs. This was not attempted during the project as the INDEL prediction accuracy is significantly lower than the SNP prediction accuracy (data not shown). To improve INDEL prediction a better INDEL caller should be used. In this analysis I used the SAMtools pipeline. Using a program such as DINDEL (Albers et al., 2011), which reconstructs the genomic region around the potential INDEL to confirm whether it is real, would lead to more accurate INDEL calling and hence more reliable evolutionary inferences based on genes with predicted INDELs. Extending this further, it may be possible to reconstruct genes which have highly dissimilar regions, but identifying the region of interest, and then attempt a local assembly of the region, as was implemented in my genome improvement pipeline in chapter 3 inspired by Ossowski et al. (2008). This would improve the significance of the results as it enables comparison of genes from races that would otherwise have been left out of the analysis.

## 6.4 Effector characterisation and evolution

Despite the onset of the genomics era and the access to numerous genome assemblies of pathogens, it is not trivial to characterise effectors. In the oomycetes it was expected that effectors would be short secreted proteins carrying the RxLR domain (Birch et al., 2006). This was soon shown not to be the case for all oomycetes after it was found that the oomycete *A. laibachii* had CHxC effector genes (Kemen et al., 2011), suggesting that the RxLR motif is limited to the *Peronosporales*. However, the genome sequence of *Pythium* revealed that it also had genes encoding effector-like genes carrying the RxLR motif, suggesting that the *Albugonales* lost the ability to translocate effectors via the RxLR mechanism, or the RxLR translocation mechanism was gained in other oomycetes. The comparative analysis of effectors both within as well as between species is complex. Two characterised effectors of *Hpa*, *ATR1* and *ATR13* carry the expected RxLR motif. However the most recently characterised *Hpa* effector, *ATR5* (Bailey et al., 2011) carries a GRVR-EE domain, instead of the expected RXLR-EE. On closer analysis *ATR5* carries the C-terminal W and Y domains (Jiang et al., 2008), which suggests that the RxLR motif may not be the main element involved in the translocation of effectors to the host. *ATR5* despite not having sufficient evidence of divergence from neutrality did have 21 polymorphic sites, which is more than ATR13. It was also observed that ATR5 was present in full in all ecotypes tests. This could be because of (i) random choice of *Hpa* isolates sequenced (ii) ATR5 being an essential effector that *Hpa* cannot lose, or (iii) that there is sufficient diversity in ATR5 to usually avoid recognition by the host thus not presenting significant negative selection pressure for loss of the gene.

Despite only having identified 3 effectors in *Hpa* with avirulence function, we have been able to predict 141 'good' candidate effectors among a list of 647 potential effector candidates. For the analysis performed in chapter 5 I increased the sample set of effectors to 472 candidates showing no negative traits of being an effector, as by doing so I was able to include *ATR5*, and its 2 homologs. Comparative analysis of the effectors, transmembrane genes and KOGs revealed that the majority of the genes in each set fall into similar distributions for the majority of evolutionary test statistics. However, there were significant subsets of effector genes that were different from the other genes, in number of segregating sites, total number of mutations and predicted number of haplotypes per gene, Fu's Fs (Fu, 1997) and dN/dS as calculated by codeml. This is an improvement from describing effectors as polymorphic. Equally enlightening is that we observe many effector candidates that nevertheless do not have evidence for positive selection. This was observed repeatedly in the different evolutionary models tested. It is possible that different effectors are under different selection pressures – a subset of effectors may be functional but is not recognised by the host in which case they are selected for conservation while other effectors may be functional but do trigger an immune response in the host. This can be illustrated with examples from Fabro et al., (2011) who show that the candidate effectors contributing most to virulence include HaRxL66, for which Fu's Fs is 5.6 (higher than ATR13, and suggesting positive selection), and HaRxL44, for which Fu's Fs is 0.03 (no evidence for positive selection). In the co-evolutionary system, the effectors that are recognised have additional selection pressures acting on them due to the epispastic interaction with the host. One would expect to see these effectors in two forms in the *Hpa* population. One of those types of effectors includes those that are present in nearly all races due the effector being a key contributor to *Hpa* fitness. In this case it would be beneficial to keep the functionality of the effector but also have high levels of polymorphisms to avoid recognition in the host. The second form of effectors would be one with many presence/absence polymorphisms due to the effector being recognised in some accessions of the host. In this scenario it is possible that resistance gene alleles recognising this effector are in low frequency in the host population, so that there is significant enough negative selection to remove the effector from the pathogen population, with perturbation in the resistant gene frequency determining the frequency of the effector. In the extreme case where the majority of hosts are able to recognise the majority of alleles of the effector there may be enough negative selection pressure for complete loss of the effector. This type of analysis can only be performed when more avirulent effectors of *Hpa* have been identified.

With this improved understanding of effector evolution, I analysed a set of genes sharing similar levels of high selection pressure with ATR1 and ATR13. 15 genes were identified, of which 12 were secreted and 11 shared homology to known effectors. Among the genes showing homology to effectors, there are 3 genes that also show homology to *Phytophthora* RXLR effector gene candidates, but do not carry the RXLR motif. This suggests that not all translocated effectors carry an exact RXLR motif, an interpretation supported by three other pieces of evidence. ATR5 does not carry the RXLR motif (Bailey et al., 2011). The RXLR motif is not present in the major class of candidate effectors in *A. laibachii* (Kemen et al., 2011), and in a recent study by Yaeno et al. (2011) the phosphatidylinositol monophosphate mediated translocation system suggested by Kale et al. (2010) is not dependent on the RxLR motif but instead on a positive charged region found downstream of the RxLR corresponding to the W and Y motif regions, that is also present in ATR1, ATR5 and 2 of the 3 homologs of the *Phytophthora* effector genes (808594 and eff_g11324). This strongly suggests that genes 808594 and eff_g11324, and their *P. infestans* homologs, are real effectors and they provide an exciting opportunity to investigate a novel class of effector genes and help unravel the nature of effector translocation into the host.

# Appendices

## Appendices for Chapter 2

### Appendix table 2.1: Cluster density of paired end reads

| FlowCell ID | Lane | *Hpa* race | pM | Clusters | Clusters/pM |
|---|---|---|---|---|---|
| ID62 | Lane 1 | Hpa Waco9 | 6 | 107 | 17.8 |
| ID62 | Lane 2 | Hpa Hind2 | 6 | 119 | 19.8 |
| ID62 | Lane 3 | Hpa Emco5 | 6 | 116 | 19.3 |
| ID64 | Lane 1 | Hpa Hind2 | 6 | 87 | 14.5 |
| ID64 | Lane 2 | Hpa Hind2 | 6 | 77 | 12.8 |
| ID64 | Lane 3 | Hpa Hind2 | 6 | 89 | 14.8 |
| ID64 | Lane 5 | Hpa Emco5 | 6 | 83 | 13.8 |
| ID64 | Lane 6 | Hpa Emco5 | 6 | 83 | 13.8 |
| ID64 | Lane 7 | Hpa Emco5 | 6 | 82 | 13.7 |
| ID64 | Lane 8 | Hpa Emoy2 | 6 | 69 | 11.5 |
| ID66 | Lane 3 | Hpa Emoy2 | 6 | 74 | 12.3 |
| ID69 | Lane 1 | Hpa Emoy2 | 6 | 82 | 13.7 |
| ID69 | Lane 2 | Hpa Emoy2 | 6 | 71 | 11.8 |
| ID69 | Lane 3 | Hpa Emoy2 | 6 | 66 | 11.0 |
| ID69 | Lane 5 | Hpa Emoy2 | 6 | 85 | 14.2 |
| ID69 | Lane 6 | Hpa Cala2 | 6 | 99 | 16.5 |
| ID69 | Lane 7 | Hpa Cala2 | 6 | 100 | 16.7 |
| ID69 | Lane 8 | Hpa Cala2 | 6 | 91 | 15.2 |
| ID71 | Lane 1 | Hpa Cala2 | 6 | 93 | 15.5 |
| ID71 | Lane 2 | Hpa Maks9 | 6 | 79 | 13.2 |
| ID71 | Lane 3 | Hpa Maks9 | 6 | 87 | 14.5 |
| ID71 | Lane 5 | Hpa Maks9 | 6 | 86 | 14.3 |
| ID74 | Lane 1 | Hpa Noco2 | 6 | ? | ? |
| ID74 | Lane 2 | Hpa Noco2 | 6 | ? | ? |
| ID74 | Lane 3 | Hpa Noco2 | 6 | ? | ? |
| ID74 | Lane 5 | Hpa Noco2 | 6 | ? | ? |
| ID74 | Lane 8 | Hpa Maks9 | 6 | ? | ? |
| ID75 | Lane 1 | Hpa Waco9 | 6 | 91 | 15.2 |
| ID75 | Lane 2 | Hpa Waco9 | 6 | 94 | 15.7 |
| ID75 | Lane 3 | Hpa Waco9 | 6 | 91 | 15.2 |
| ID75 | Lane 5 | Hpa Waco9 | 6 | 96 | 16.0 |
| ID75 | Lane 6 | Hpa Waco9 | 6 | 94 | 15.7 |
| ID79 | Lane 2 | Hpa Emoy2 | 6 | 69 | 11.5 |
| ID79 | Lane 3 | Hpa Emoy2 | 6 | 70 | 11.7 |
| ID80 | Lane 1 | Hpa Noco2 | 6 | 81 | 13.5 |
| ID80 | Lane 2 | Hpa Noco2 | 6 | 76 | 12.7 |
| ID80 | Lane 3 | Hpa Cala2 | 6 | 61 | 10.2 |
| ID87 | Lane 3 | Hpa Waco9 | 8 | 78 | 9.8 |
| ID87 | Lane 4 | Hpa Waco9 | 8 | 74 | 9.3 |
| ID87 | Lane 5 | Hpa Hind2 | 8 | 67 | 8.4 |
| ID87 | Lane 6 | Hpa Hind2 | 8 | 69 | 8.6 |
| ID87 | Lane 7 | Hpa Hind2 | 8 | 70 | 8.8 |
| ID88 | Lane 1 | Hpa Emco5 | 9 | 99 | 11.0 |
| ID88 | Lane 2 | Hpa Emco5 | 9 | 101 | 11.2 |

| | | | | | |
|---|---|---|---|---|---|
| ID88 | Lane 3 | Hpa Emco5 | 9 | 101 | 11.2 |
| ID88 | Lane 4 | Hpa Emco5 | 9 | 106 | 11.8 |
| ID88 | Lane 5 | Hpa Hind2 | 9 | 106 | 11.8 |

## Appendix table 2.2: Reads Summary

| Date | ID | Lane | PE/SE | Insert Size | Race | Type | Concentration | Length | Number of Reads | Total bases sequenced |
|---|---|---|---|---|---|---|---|---|---|---|
| 2007-09-19 | ID19 | 6 | SE | | emoy2 | cdna | 4.5pM | 35 | 939,355 | 32,877,425 |
| 2007-09-19 | ID19 | 7 | SE | | emoy2 | cdna | 4.5pM | 35 | 1,153,202 | 40,362,070 |
| 2007-10-02 | ID21 | 7 | SE | | emoy2 | cdna | 5pM | 35 | 429,007 | 15,015,245 |
| 2007-10-02 | ID21 | 8 | SE | | emoy2 | cdna | 5pM | 35 | 503,955 | 17,638,425 |
| 2007-10-11 | ID23 | 7 | SE | | emoy2 | cdna | 5pM | 35 | 1,029,001 | 36,015,035 |
| 2007-10-11 | ID23 | 8 | SE | | emoy2 | cdna | 5pM | 35 | 1,946,103 | 68,113,605 |
| 2008-01-23 | ID32 | 8 | SE | | emoy2 | cdna | 5pM | 35 | 2,548,409 | 89,194,315 |
| 2008-10-22 | ID62 | 1 | PE | 334.305772 +/- 45.584579 | waco9 | dna | 6pM | 36 | 16,685,148 | 600,665,328 |
| 2008-10-22 | ID62 | 2 | PE | 352.198133 +/- 43.512337 | hind2 | dna | 6pM | 36 | 18,790,428 | 676,455,408 |
| 2008-10-22 | ID62 | 3 | PE | 359.373364 +/- 48.609409 | emco5 | dna | 6pM | 36 | 18,497,806 | 665,921,016 |
| 2008-11-03 | ID64 | 1 | PE | 348.531576 +/- 46.663678 | hind2 | dna | 6pM | 36 | 13,339,720 | 480,229,920 |
| 2008-11-03 | ID64 | 2 | PE | 348.368804 +/- 47.207299 | hind2 | dna | 6pM | 36 | 11,995,426 | 431,835,336 |
| 2008-11-03 | ID64 | 3 | PE | 348.436872 +/- 47.151158 | hind2 | dna | 6pM | 36 | 14,346,168 | 516,462,048 |
| 2008-11-03 | ID64 | 5 | PE | 355.161022 +/- 53.034060 | emco5 | dna | 6pM | 36 | 13,642,658 | 491,135,688 |
| 2008-11-03 | ID64 | 6 | PE | 355.096081 +/- 53.164785 | emco5 | dna | 6pM | 36 | 13,705,996 | 493,415,856 |
| 2008-11-03 | ID64 | 7 | PE | 355.052176 +/- 53.120832 | emco5 | dna | 6pM | 36 | 13,377,394 | 481,586,184 |
| 2008-11-03 | ID64 | 8 | PE | 339.052529 +/- 43.744176 | emoy2 | dna | 6pM | 36 | **11,044,778** | 397,428,480 |
| 2008-11-14 | ID66 | 3 | PE | 338.813991 +/- 43.742254 | emoy2 | dna | 6pM | 36 | **12,340,546** | 443,439,000 |
| 2008-12-02 | ID69 | 1 | PE | 338.685350 +/- 43.706918 | emoy2 | dna | 6pM | 36 | **12,997,822** | 458,173,584 |
| 2008-12-02 | ID69 | 2 | PE | 338.568591 +/- 44.134278 | emoy2 | dna | 6pM | 36 | **11,107,194** | 399,083,112 |
| 2008-12-02 | ID69 | 3 | PE | 338.675867 +/- 43.822840 | emoy2 | dna | 6pM | 36 | **10,424,806** | 352,597,320 |
| 2008-12-02 | ID69 | 5 | PE | 338.877557 +/- 43.778193 | emoy2 | dna | 6pM | 36 | **13,791,734** | 495,940,896 |
| 2008-12-02 | ID69 | 6 | PE | 327.371724 +/- 46.945084 | cala2 | dna | 6pM | 36 | 14,797,458 | 532,708,488 |
| 2008-12-02 | ID69 | 7 | PE | 327.443553 +/- 46.651466 | cala2 | dna | 6pM | 36 | 14,879,458 | 535,660,488 |
| 2008-12-02 | ID69 | 8 | PE | 327.378922 +/- 46.903094 | cala2 | dna | 6pM | 36 | 12,956,184 | 466,422,624 |

## Appendix table 2.2: Reads Summary

| Date | ID | Lane | PE/SE | Insert Size | Race | Type | Concentration | Length | Number of Reads | Total bases sequenced |
|---|---|---|---|---|---|---|---|---|---|---|
| 2008-12-12 | ID71 | 1 | PE | 327.043908 +/- 46.746053 | cala2 | dna | 6pM | 36 | 12,591,470 | 453,292,920 |
| 2008-12-12 | ID71 | 2 | PE | 333.627462 +/- 39.240630 | maks9 | dna | 6pM | 36 | 11,398,246 | 410,336,856 |
| 2008-12-12 | ID71 | 3 | PE | 333.914869 +/- 38.790438 | maks9 | dna | 6pM | 36 | 12,808,634 | 461,110,824 |
| 2008-12-12 | ID71 | 5 | PE | 334.095602 +/- 38.827157 | maks9 | dna | 6pM | 36 | 12,943,886 | 465,979,896 |
| 2009-01-05 | ID74 | 1 | PE | 336.995617 +/- 34.091977 | noco2 | dna | 6pM | 36 | 12,692,042 | 456,913,512 |
| 2009-01-05 | ID74 | 2 | PE | 337.123654 +/- 34.208484 | noco2 | dna | 6pM | 36 | 13,959,594 | 502,545,384 |
| 2009-01-05 | ID74 | 3 | PE | 337.087019 +/- 34.268842 | noco2 | dna | 6pM | 36 | 14,078,192 | 506,814,912 |
| 2009-01-05 | ID74 | 5 | PE | 337.310454 +/- 34.094262 | noco2 | dna | 6pM | 36 | 14,282,786 | 514,180,296 |
| 2009-01-05 | ID74 | 8 | PE | 334.466367 +/- 37.941812 | maks9 | dna | 6pM | 36 | 13,410,342 | 482,772,312 |
| 2009-01-12 | ID75 | 1 | PE | 330.821535 +/- 50.870072 | waco9 | dna | 6pM | 36 | 13,330,010 | 479,880,360 |
| 2009-01-12 | ID75 | 2 | PE | 330.629955 +/- 51.417488 | waco9 | dna | 6pM | 36 | 14,006,684 | 504,240,624 |
| 2009-01-12 | ID75 | 3 | PE | 329.992848 +/- 52.163552 | waco9 | dna | 6pM | 36 | 9,837,618 | 354,154,248 |
| 2009-01-12 | ID75 | 5 | PE | 331.021138 +/- 50.880626 | waco9 | dna | 6pM | 36 | 13,552,462 | 487,888,632 |
| 2009-01-12 | ID75 | 6 | PE | 331.167094 +/- 51.013012 | waco9 | dna | 6pM | 36 | 14,329,742 | 515,870,712 |
| 2009-02-16 | ID79 | 2 | PE | 339.291135 +/- 42.760396 | emoy2 | dna | 6pM | 36 | 13,849,592 | 498,585,312 |
| 2009-02-16 | ID79 | 3 | PE | 339.348849 +/- 42.578734 | emoy2 | dna | 6pM | 36 | 14,059,650 | 506,147,400 |
| 2009-02-17 | ID80 | 1 | PE | 337.236591 +/- 33.780184 | noco2 | dna | 6pM | 36 | 16,274,756 | 585,891,216 |
| 2009-02-17 | ID80 | 2 | PE | 337.119928 +/- 33.892590 | noco2 | dna | 6pM | 36 | 14,117,014 | 508,212,504 |
| 2009-02-17 | ID80 | 3 | PE | 327.009732 +/- 46.690504 | cala2 | dna | 6pM | 36 | 9,715,162 | 349,745,832 |
| 2009-04-15 | ID87 | 3 | PE | 332.110200 +/- 50.398237 | waco9 | dna | 8pM | 36 | 14,153,536 | 509,527,296 |
| 2009-04-15 | ID87 | 4 | PE | 331.922077 +/- 50.975731 | waco9 | dna | 8pM | 36 | 13,976,824 | 503,165,664 |
| 2009-04-15 | ID87 | 5 | PE | 349.398150 +/- 46.992682 | hind2 | dna | 8pM | 36 | 13,518,100 | 486,651,600 |
| 2009-04-15 | ID87 | 6 | PE | 349.466094 +/- 46.938720 | hind2 | dna | 8pM | 36 | 13,838,710 | 498,193,560 |
| 2009-04-15 | ID87 | 7 | PE | 349.527070 +/- 46.807906 | hind2 | dna | 8pM | 36 | 14,034,950 | 505,258,200 |
| 2009-04-21 | ID88 | 1 | PE | 355.199558 +/- 53.455711 | emco5 | dna | 9pM | 36 | 16,440,814 | 591,869,304 |
| 2009-04-21 | ID88 | 2 | PE | 355.669794 +/- 52.755553 | emco5 | dna | 9pM | 36 | 16,900,940 | 608,433,840 |
| 2009-04-21 | ID88 | 3 | PE | 355.876003 +/- 52.436018 | emco5 | dna | 9pM | 36 | 16,038,510 | 577,386,360 |

## Appendix table 2.2: Reads Summary

| Date | ID | Lane | PE/SE | Insert Size | Race | Type | Concentration | Length | Number of Reads | Total bases sequenced |
|---|---|---|---|---|---|---|---|---|---|---|
| 2009-04-21 | ID88 | 4 | PE | 355.857802 +/- 52.516396 | emco5 | dna | 9pM | 36 | 15,969,604 | 574,905,744 |
| 2009-04-21 | ID88 | 5 | PE | 349.195154 +/- 46.825088 | hind2 | dna | 9pM | 36 | 16,978,272 | 611,217,792 |
| 2009-11-30 | ID106 | 4 | PE | 356.988865 +/- 48.284072 | emco5 | dna | 12pM | 76 | 29,750,734 | 2,261,055,784 |
| 2009-11-30 | ID106 | 5 | PE | 328.725489 +/- 43.721368 | cala2 | dna | 12pM | 76 | 36,247,384 | 2,754,801,184 |
| 2009-12-16 | ID108 | 2 | PE | 338.662604 +/- 32.772048 | noco2 | dna | 12pM | 76 | 30,669,042 | 2,330,847,192 |
| 2009-12-16 | ID108 | 3 | PE | 335.986928 +/- 36.052768 | maks9 | dna | 12pM | 76 | 29,857,228 | 2,269,149,328 |
| | 42009 | 5 | PE | 291.462540 +/- 23.736702 | emwa1 | dna | ? | 76 | 18,662,002 | 1,418,312,152 |
| | 42009 | 6 | PE | 484.961356 +/- 48.770805 | emwa1 | dna | ? | 76 | 20,105,712 | 1,528,034,112 |
| | 42009 | 7 | PE | 484.938850 +/- 48.799756 | emwa1 | dna | ? | 76 | 13,729,394 | 1,043,433,944 |
| | 1E+05 | 6 | PE | 259.937447 +/- 38.686330 | emwa1 | dna | ? | 76 | 16,685,692 | 1,268,112,592 |

## Appendix figure 2.1 FastQC Read statistics

The FastQC read statistics are available in a separate Word and PDF document on the accompanying CD.

## Appendices for Chapter 3

**Appendix table 3.1: Genome version history. Release history and assembly statistics for the *H. arabidopsidis* genome. * includes 35.5x nucleotide coverage by Illumina paired end reads**

| Release version | Release Date | WGS plasmid reads | WGS fosmid reads | BAC ends | Illumina paired end reads | Total Input Reads | Size (Mb) | Coverage | Major scaffolds (>2 kb) | N50 scaffolds number |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | April 2006 | 1,053,419 | - | - | - | 1,053,419 | 70 | 8.0x | | |
| 2 | April 2006 | 1,053,419 | - | - | - | 1,053,419 | 70 | 8.0x | | |
| 3 | July 2007 | 1,055,973 | 18,814 | - | - | 1,074,767 | 75 | 9.0x | 2140 | 171 |
| 4 | November 2007 | 1,080,646 | 25,516 | - | - | 1,106,162 | 75 | 9.2x | 2073 | 174 |
| 5 | November 2007 | 1,080,646 | 25,516 | - | - | 1,106,162 | 75 | 9.2x | 2073 | 174 |
| 6 | December 2007 | 1,080,646 | 25,516 | 13,071 | - | 1,151,387 | 77 | 9.2x | 1739 | 71 |
| 7 | August 2008 | 1,080,646 | 25,516 | 13,071 | - | 1,119,233 | 76.5 | 9.5x | 1585 | 68 |
| Velvet | February 2009 | - | - | - | 56,727,498 | 56,727,498 | 56.9 | 35.5x* | 4429 | 742 |
| v8.3 | September 2009 | 1,080,646 | 25,516 | 13,071 | 56,727,498 | 57,846,731 | 82 | 45x* | 1783 | 75 |

- 

## Appendix table 3.2: Hpa Emoy2 Illumina reads

| Flowcell | Lane | Paired End | Inner distance between reads | Concentration | Length | Number of reads | Reads mapped to v8.3 | % of reads mapped to v8.3 | reads mapped as pair | % of reads aligning mapped as pair |
|---|---|---|---|---|---|---|---|---|---|---|
| ID16 | 7 | N | | 4pM | 35 | 1,011,420 | | | | |
| ID16 | 8 | N | | 4pM | 35 | 499,363 | | | | |
| ID45 | 1 | N | | 5pM | 35 | 4,667,054 | 3,781,889 | 81.0% | | |
| ID45 | 2 | N | | 5pM | 35 | 4,849,873 | 3,900,266 | 80.4% | | |
| ID45 | 3 | N | | 5pM | 35 | 4,907,559 | 3,924,460 | 80.0% | | |
| ID45 | 5 | N | | 5pM | 35 | 4,927,486 | 3,929,169 | 79.7% | | |
| ID45 | 6 | N | | 5pM | 35 | 4,806,600 | 3,828,554 | 79.7% | | |
| ID45 | 7 | N | | 5pM | 35 | 3,976,253 | 3,202,218 | 80.5% | | |
| ID45 | 8 | N | | 5pM | 35 | 3,858,756 | 3,108,958 | 80.6% | | |
| ID64 | 8 | Y | 339.052529 +/- 43.744176 | 6pM | 36 | 11,039,680 | 10,371,172 | 93.9% | 9,606,994 | 92.6% |
| ID66 | 3 | Y | 338.813991 +/- 43.742254 | 6pM | 36 | 12,317,750 | 11,577,286 | 94.0% | 10,789,809 | 93.2% |
| ID69 | 1 | Y | 338.685350 +/- 43.706918 | 6pM | 36 | 12,727,044 | 11,983,134 | 94.2% | 11,164,556 | 93.2% |
| ID69 | 2 | Y | 338.568591 +/- 44.134278 | 6pM | 36 | 11,085,642 | 10,421,860 | 94.0% | 9,697,772 | 93.1% |
| ID69 | 3 | Y | 338.675867 +/- 43.822840 | 6pM | 36 | 9,794,370 | 9,208,542 | 94.0% | 8,539,768 | 92.7% |
| ID69 | 5 | Y | 338.877557 +/- 43.778193 | 6pM | 36 | 13,776,136 | 12,969,038 | 94.1% | 12,094,156 | 93.3% |
| ID79 | 2 | Y | 339.291135 +/- 42.760396 | 6pM | 36 | 13,849,592 | 13,013,004 | 94.0% | 12,074,329 | 92.8% |
| ID79 | 3 | Y | 339.348849 +/- 42.578734 | 6pM | 36 | 14,059,650 | 13,126,898 | 93.4% | 12,115,612 | 92.3% |

The non-paired reads were sequenced on the Illumina Genome Analyzer I and the paired end reads were sequenced on the Genome Analyzer 2 platforms. The total coverage of the *H. arabidopsidis* Emoy2 v8.3 assembly is 46.0x nucleotide coverage through 10.5x non-paired read coverage and 35.5x nucleotide coverage through paired end reads. For the paired ends Y = yes and N = no.

## Appendix table 3.3: List of genes used to evaluate assemblies

```
gi|224993523|gb|ACN76441.1| heat shock transcription factor [Hyaloperonospora parasitica]
gi|222144621|gb|ACM46122.1| MAP kinase [Hyaloperonospora parasitica]
gi|209573498|gb|ACI62835.1| CFZ1-like protein [Hyaloperonospora parasitica]
gi|171879818|gb|ACB55623.1| avirulence protein [Hyaloperonospora parasitica]
gi|167047082|gb|ABZ10809.1| RXL96 [Hyaloperonospora parasitica]
gi|152963459|gb|ABS50086.1| putative effector protein Avh341 [Hyaloperonospora parasitica]
gi|108885465|gb|ABG23238.1| unknown [Hyaloperonospora parasitica]
gi|93139267|gb|ABE99946.1| NADH dehydrogenase subunit 1 [Hyaloperonospora parasitica]
gi|93139088|gb|ABE99881.1| beta-tubulin [Hyaloperonospora parasitica]
gi|89146243|gb|ABD62107.1| cytochrome oxidase subunit II [Hyaloperonospora parasitica]
gi|108885471|gb|ABG23241.1| putative small cys-rich protein [Hyaloperonospora parasitica]
gi|108885467|gb|ABG23239.1| retrotransposon element [Hyaloperonospora parasitica]
gi|108885463|gb|ABG23237.1| unknown [Hyaloperonospora parasitica]
gi|108885461|gb|ABG23236.1| putative small cys-rich protein [Hyaloperonospora parasitica]
gi|108885459|gb|ABG23235.1| putative membrane protein [Hyaloperonospora parasitica]
gi|108885457|gb|ABG23234.1| putative membrane protein [Hyaloperonospora parasitica]
gi|108885455|gb|ABG23233.1| unknown [Hyaloperonospora parasitica]
gi|108885453|gb|ABG23232.1| putative N-acetyltransferase-like protein [Hyaloperonospora
parasitica]
gi|66934640|gb|AAY58912.1| putative 3-isopropylmalate dehydratase large subunit
[Hyaloperonospora parasitica]
gi|66934639|gb|AAY58911.1| putative F-actin capping protein [Hyaloperonospora parasitica]
gi|66934638|gb|AAY58910.1| ras-like protein [Hyaloperonospora parasitica]
gi|66934637|gb|AAY58909.1| putative RXLR protein 12I13.1 [Hyaloperonospora parasitica]
gi|66934636|gb|AAY58908.1| putative methylene tetrahydrofolate dehydrogenase [Hyaloperonospora
parasitica]
gi|66934635|gb|AAY58907.1| putative dimeric dihydrodiol dehydrogenase [Hyaloperonospora
parasitica]
gi|66934634|gb|AAY58906.1| avirulence protein-like protein [Hyaloperonospora parasitica]
gi|66934633|gb|AAY58905.1| putative Myb-like protein [Hyaloperonospora parasitica]
gi|66934632|gb|AAY58904.1| avirulence protein [Hyaloperonospora parasitica]
gi|66934628|gb|AAY58903.1| putative LON protease [Hyaloperonospora parasitica]
gi|66934627|gb|AAY58902.1| putative CDC48/ATPase [Hyaloperonospora parasitica]
gi|66934626|gb|AAY58901.1| putative BAX inhibitor [Hyaloperonospora parasitica]
gi|66934624|gb|AAY58900.1| avirulence protein [Hyaloperonospora parasitica]
gi|58042874|gb|AAW63774.1| PPAT5 [Hyaloperonospora parasitica]
gi|58042862|gb|AAW63768.1| avirulence protein ATR13 [Hyaloperonospora parasitica]
gi|34922241|gb|AAQ83522.1| cysteine rich [Hyaloperonospora parasitica]
gi|34922215|gb|AAQ83519.1| unknown [Hyaloperonospora parasitica]
gi|34922209|gb|AAQ83518.1| putative carboxylase [Hyaloperonospora parasitica]
gi|34922199|gb|AAQ83517.1| putative dehydrogenase [Hyaloperonospora parasitica]
gi|34922186|gb|AAQ83516.1| putative homogentisate 1,2-dioxygenase [Hyaloperonospora parasitica]
gi|34922176|gb|AAQ83515.1| unknown [Hyaloperonospora parasitica]
gi|34922153|gb|AAQ83514.1| putative ATPase [Hyaloperonospora parasitica]
gi|34922145|gb|AAQ83513.1| unknown [Hyaloperonospora parasitica]
gi|34922136|gb|AAQ83512.1| unknown [Hyaloperonospora parasitica]
gi|34922130|gb|AAQ83511.1| putative carboxyltransferase [Hyaloperonospora parasitica]
gi|34922121|gb|AAQ83510.1| putative beta-glucosidase [Hyaloperonospora parasitica]
gi|34922115|gb|AAQ83509.1| putative fatty acid synthase alpha subunit [Hyaloperonospora
parasitica]
gi|34922091|gb|AAQ83506.1| putative serine/threonine protein kinase [Hyaloperonospora
parasitica]
gi|34922077|gb|AAQ83505.1| unknown [Hyaloperonospora parasitica]
gi|34922067|gb|AAQ83504.1| putative dnaK-type molecular chaperone [Hyaloperonospora parasitica]
gi|34922055|gb|AAQ83503.1| putative H+ translocating inorganic pyrophosphatase [Hyaloperonospora
parasitica]
gi|34922040|gb|AAQ83501.1| unknown [Hyaloperonospora parasitica]
gi|34922033|gb|AAQ83500.1| putative exo-1,3-beta-glucanase [Hyaloperonospora parasitica]
gi|33350990|gb|AAP49016.1| cytochrome oxidase subunit II [Hyaloperonospora parasitica]
```

List of genes used to evaluate the quality of Velvet assemblies using a k-mer lengths of 21 and 23. The genes highlighted in red were found full length in the assembly using a k-mer length of 21 and partially in the assembly of k-mer length 23. There were no genes that were complete in the k-mer 23 assembly and partially assembled in the k-mer 21 assembly.

## Appendix table 3.4: Final v8.3 assembly scaffold (AGP)

| | |
|---|---|
| Contig0 | Contig0 1 .. 360550 + NODE_1260_length_184_cov_18.402174 1 .. 67 + Contig0 360618 .. 596263 + NODE_1359_length_453_cov_3.960265 168 .. 477 + Contig0 596572 .. 1494027 + NODE_3875_length_453_cov_4.039735 151 .. 477 + Contig0 1494353 .. 2274569 |
| Contig1 | Contig1 1 .. 81705 + rev NODE_5373_length_2156_cov_8.420222 1 .. 2029 + Contig1 83735 .. 846734 + NODE_4843_length_554_cov_2.187726 1 .. 578 + Contig1 856266 .. 1003983 + rev NODE_15761_length_1234_cov_7.220421 1 .. 1079 + Contig1 1005063 .. 1243756 |
| Contig10 | Contig10 1 ..466923 + CU681819 1 ..106602 + Contig10 613066 ..end |
| Contig100 | Contig100 1 .. 192300 + NODE_4687_length_424_cov_2.448113 167 .. 448 + Contig100 192581 .. 223720 + NODE_97_length_484_cov_8.770661 173 .. 508 + Contig100 224055 .. 249180 |
| Contig103 | Contig103 1 .. 161684 + rev NODE_2829_length_509_cov_3.504912 1 .. 356 + Contig103 162041 .. 162141 + NODE_16332_length_839_cov_5.582837 1 .. 683 + Contig103 162825 .. 209324 |
| Contig1030 | Contig1030 1 .. 2313 + NODE_4475_length_887_cov_3.993236 187 .. 911 + Contig1030 2313 .. 2313 |
| Contig105 | Contig105 1 .. 111126 + NODE_1726_length_926_cov_8.438445 185 .. 950 + Contig105 111891 .. 233890 + NODE_1321_length_473_cov_4.175476 165 .. 497 + Contig105 234222 .. 248440 |
| Contig106 | CU694540 1 ..94388 + Contig106 66699 ..239699 + rev NODE_14441_length_581_cov_4.843373 1 .. 454 + Contig106 239649 .. End |
| Contig107 | Contig107 1 .. 32775 + rev NODE_21751_length_732_cov_3.777322 1 .. 495 + Contig107 33191 .. 266073 + NODE_2155_length_522_cov_3.963602 232 .. 546 + Contig107 266387 .. 267334 |
| Contig108 | Contig108 1 .. 43451 + NODE_903_length_1344_cov_8.246280 161 .. 1368 + Contig108 44658 .. 271342 |
| Contig1086 | Contig1086 1 .. 1994 + rev NODE_8762_length_518_cov_4.820463 1 .. 390 + Contig1086 1994 .. 1994 |
| Contig11 | CU855856 1 ..89035 + Contig11 23947 .. 139076 + NODE_5432_length_459_cov_3.289760 157 .. 483 + Contig11 139402 .. 418829 + NODE_202_length_436_cov_3.683486 161 .. 460 + Contig11 484216 .. 486217 + NODE_16417_length_18822_cov_7.432313 159 .. 18846 + Contig11 504904 .. 728412 + rev NODE_12193_length_566_cov_3.275618 1 .. 384 + Contig11 728797 .. 810063 + NODE_10815_length_819_cov_5.318681 152 .. 843 + |
| Contig110 | Contig110 1 .. 124736 + NODE_4998_length_837_cov_6.221027 165 .. 861 + Contig110 125432 .. 226695 |
| Contig1101 | Contig1101 1 .. 2539 + rev NODE_2735_length_890_cov_8.229214 1 .. 700 + Contig1101 2539 .. 2539 |
| Contig114 | Contig114 1 ..63194 + CU694290 1 ..104779 + NODE_11682_length_642_cov_6.291277 267 .. 666 |
| Contig116 | Contig116 1 .. 165748 + NODE_185_length_358_cov_88.776535 312 .. 382 + Contig116 165818 .. 168596 |
| Contig1162 | Contig1162 1 .. 2254 + NODE_5508_length_453_cov_2.346578 156 .. 477 + Contig1162 2254 .. 2254 |
| Contig117 | Contig117 1 ..93883 + CU611059 1 ..117161 + Contig117 211053 ..211182 |
| Contig1171 | Contig1171 1 .. 1 + NODE_1644_length_913_cov_25.886089 1 .. 165 + Contig1171 24 .. 2110 |
| Contig1173 | Contig1173 1 .. 3338 + NODE_2914_length_570_cov_2.647368 260 .. 594 + Contig1173 3338 .. 3338 |
| Contig1175 | Contig1175 1 .. 2833 + rev NODE_15069_length_574_cov_3.900697 1 .. 439 + Contig1175 2833 .. 2833 |
| Contig118 | Contig118 1 .. 168962 + rev NODE_15181_length_824_cov_4.258495 1 .. 679 + Contig118 169642 .. 171330 |
| Contig12 | Contig12 1 .. 549888 + NODE_5701_length_950_cov_4.801053 174 .. 974 + Contig12 550688 .. 710758 |
| Contig120 | Contig120 1 .. 83491 + NODE_14842_length_9445_cov_10.180731 164 .. 9469 + Contig120 92796 .. 159163 + rev NODE_18724_length_4059_cov_7.507514 1 .. 3882 + Contig120 163046 .. 233050 |
| Contig1205 | Contig1205 1 .. 4896 + NODE_7763_length_140_cov_2.771429 156 .. 164 + Contig1205 4896 .. 4896 |
| Contig121 | Contig121 1 .. 78816 + NODE_111_length_241_cov_90.585060 1 .. 113 + Contig121 78930 .. 210927 |
| Contig123 | Contig123 1 .. 69998 + rev NODE_10824_length_2433_cov_7.838471 1 .. 2295 + Contig123 72044 .. 148978 |
| Contig124 | Contig124 1 .. 285585 + rev NODE_8740_length_1156_cov_3.957613 1 .. 1020 + Contig124 285585 .. 285585 |
| Contig125 | Contig125 1 .. 68642 + rev NODE_16547_length_586_cov_3.849829 1 .. 458 + Contig125 69101 .. 282133 |
| Contig126 | Contig126 1 .. 34807 + NODE_19981_length_13774_cov_9.950269 244 .. 13798 + Contig126 48361 .. 189320 |
| Contig127 | Contig127 2 .. 2 + rev NODE_10953_length_1721_cov_35.972691 223 .. 1745 + Contig127 2 .. 142654 |
| Contig1271 | Contig1271 1 .. 2625 + rev NODE_10560_length_2317_cov_7.077687 1 .. 1598 + Contig1271 2625 .. 2625 |
| Contig129 | Contig129 1 .. 6794 + NODE_3210_length_512_cov_3.671875 198 .. 536 + Contig129 7132 .. 90119 + CU611060 1 ..87205 + NODE_15149_length_970_cov_5.123711 156 .. 994 |
| Contig13 | Contig13 1 .. 1 + NODE_673_length_135_cov_70.481483 1 .. 28 + Contig13 1 .. 8840 + rev NODE_16754_length_5480_cov_8.888868 1 .. 5344 + Contig13 14185 .. 150510 + rev NODE_14331_length_1013_cov_6.566634 1 .. 880 + Contig13 150941 .. 173180 + rev NODE_16384_length_596_cov_4.365772 1 .. 465 + Contig13 173646 .. 375899 + NODE_15321_length_1199_cov_7.063386 1 .. 1223 + Contig13 380434 .. 382543 + NODE_6651_length_629_cov_3.855326 1 .. 653 + Contig13 383653 .. 495818 + NODE_3625_length_817_cov_2.410037 166 .. 841 + Contig13 496658 .. 567628 + rev NODE_14804_length_3122_cov_7.659193 1 .. 2939 + Contig13 570568 .. 677457 + NODE_3145_length_570_cov_4.177193 158 .. 594 + Contig13 677893 .. 735227 + rev NODE_1942_length_552_cov_4.353261 1 .. 415 + Contig13 735227 .. 735227 |
| Contig130 | Contig130 1 ..183574 + CU694979 1 ..71487 |
| Contig131 | Contig131 1 .. 177771 + NODE_15152_length_1186_cov_6.473019 1 .. 1210 + Contig131 182830 .. 230277 + rev NODE_17398_length_8173_cov_10.797381 1 .. 8040 + Contig131 230277 .. 230277 |
| Contig1320 | Contig1320 1 .. 4497 + NODE_5512_length_676_cov_8.042899 216 .. 700 + Contig1320 4497 .. 4497 |
| Contig134 | Contig134 1 .. 74068 + rev NODE_1914_length_476_cov_3.495798 1 .. 349 + Contig134 74143 .. 154543 |
| Contig136 | Contig136 1 .. 148251 + rev NODE_2519_length_639_cov_5.893584 1 .. 511 + Contig136 148251 .. 148251 |
| Contig137 | Contig137 1 .. 114733 + NODE_6069_length_441_cov_2.342404 151 .. 465 + Contig137 114733 .. 114733 |
| Contig1374 | Contig1374 1 .. 2516 + rev NODE_1939_length_483_cov_3.627329 1 .. 318 + Contig1374 2516 .. 2516 |
| Contig140 | CU694305 1 ..90238 30318 ..88118 + Contig140 53553 ..end |
| Contig1400 | Contig1400 1 .. 2882 + rev NODE_15924_length_1379_cov_5.123278 1 .. 1239 + Contig1400 2882 .. 2882 |
| Contig141 | Contig141 1 .. 96985 + rev NODE_14154_length_562_cov_3.096085 1 .. 426 + Contig141 97412 .. 169079 |
| Contig142 | Contig142 1 .. 1 + NODE_5_length_628_cov_25.616241 1 .. 39 + Contig142 1 .. 2346 + NODE_166_length_1878_cov_29.269968 392 .. 1902 + Contig142 3206 .. 126722 |

| | |
|---|---|
| Contig1429 | Contig1429 1 .. 4705 + rev NODE_14497_length_5534_cov_10.494579 1 .. 5402 + Contig1429 4705 .. 4705 |
| Contig144 | Contig144 1 .. 19295 + NODE_6851_length_256_cov_4.273438 148 .. 280 + Contig144 19427 .. 158184 + rev NODE_7470_length_3907_cov_9.310469 1 .. 3749 + Contig144 158184 .. 158184 |
| Contig145 | Contig145 1 .. 139773 + rev NODE_18070_length_2015_cov_9.157816 1 .. 1878 + Contig145 141652 .. 177339 |
| Contig1487 | Contig1487 1 .. 2252 + NODE_4170_length_966_cov_8.107660 169 .. 990 + Contig1487 2252 .. 2252 |
| Contig15 | Contig15 1 ..361830 + rev NODE_14482_length_6646_cov_10.322148 1 .. 6511 + Contig15 368342 ..480077 + CU694978 1 ..87635 + Contig15 564182 ..585209 + CU6722411 ..97518 |
| Contig153 | Contig153 1 .. 100681 + rev NODE_14557_length_1463_cov_6.264525 1 .. 1316 + Contig153 101998 .. 149020 + NODE_4460_length_947_cov_3.472017 155 .. 971 + |
| Contig1550 | Contig1550 1 .. 1 + NODE_1245_length_1908_cov_16.900944 1 .. 1690 + Contig1550 1 .. 2611 + NODE_1929_length_489_cov_5.288343 216 .. 513 + |
| Contig1556 | Contig1556 1 .. 2670 + rev NODE_22940_length_699_cov_3.616595 1 .. 529 + Contig1556 2670 .. 2670 |
| Contig1562 | Contig1562 1 .. 1 + NODE_6932_length_150_cov_14.326667 1 .. 16 + Contig1562 1 .. 5779 |
| Contig1568 | Contig1568 1 .. 4145 + rev NODE_16781_length_1425_cov_5.206316 1 .. 1293 + Contig1568 4145 .. 4145 |
| Contig157 | Contig157 1 .. 22456 + rev NODE_14473_length_967_cov_4.770424 1 .. 774 + Contig157 23231 .. 79499 + rev NODE_8623_length_1012_cov_6.389328 1 .. 878 + Contig157 80378 .. 199122 |
| Contig1588 | Contig1588 1 .. 2573 + NODE_14455_length_9009_cov_10.351315 163 .. 9033 + Contig1588 2573 .. 2573 |
| Contig16 | Contig16 1 .. 363036 + NODE_2589_length_449_cov_2.723831 167 .. 473 + Contig16 363342 .. 503995 + NODE_599_length_476_cov_3.207983 160 .. 500 + Contig16 504335 .. 635027 |
| Contig162 | Contig162 1 .. 74894 + NODE_921_length_1251_cov_5.250999 153 .. 1275 + Contig162 76016 .. 128175 |
| Contig163 | Contig163 1 .. 46260 + NODE_4512_length_1378_cov_7.321480 196 .. 1402 + Contig163 47466 .. 156115 + NODE_8672_length_475_cov_3.225263 201 .. 499 + Contig163 156413 .. 159641 |
| Contig165 | Contig165 1 .. 106154 + NODE_16105_length_2188_cov_9.011883 178 .. 2212 + Contig165 108188 .. 144739 + rev NODE_14398_length_1141_cov_8.914110 1 .. 1014 + Contig165 145654 .. 207316 |
| Contig167 | Contig167 1 .. 24803 + rev NODE_5751_length_448_cov_2.770089 1 .. 317 + Contig167 25121 .. 132510 + NODE_1596_length_455_cov_3.991209 189 .. 479 + |
| Contig168 | Contig168 1 .. 92631 + NODE_11270_length_754_cov_3.249337 187 .. 778 + Contig168 93222 .. 140950 |
| Contig169 | rev CU694534 1 ..112195 + Contig132 72011 -128769 +  rev CU611061 1 ..103983 + Contig169 1 ..25405 |
| Contig1697 | Contig1697 1 .. 5140 + rev NODE_2413_length_563_cov_3.753108 1 .. 425 + Contig1697 5140 .. 5140 |
| Contig17 | Contig17 1 ..94921 + CU694961 1 ..109339 + Contig17 230767 ..467896 + NODE_1776_length_463_cov_4.142549 151 .. 487 + Contig17 494739 .. End |
| Contig171 | rev Contig150 48264  .. end + CU672242 1 ..111115 + Contig171 46174  .. end |
| Contig172 | Contig172 1 .. 1 + NODE_14713_length_1597_cov_17.507828 1 .. 1427 + Contig172 1 .. 108195 |
| Contig177 | Contig177 1 .. 56325 + rev NODE_16963_length_9998_cov_10.462193 1 .. 9863 + Contig177 59889 .. 110834 + NODE_197_length_379_cov_34.398418 308 .. 403 + Contig177 110929 .. 114000 + rev NODE_173_length_972_cov_19.322016 1 .. 462 + Contig177 114000 .. 114000 |
| Contig178 | Contig178 1 – 86198 + rev CU694995 1 ..127294 |
| Contig18 | Contig18 1 .. 427586 + rev NODE_14363_length_927_cov_5.319310 1 .. 753 + Contig18 428340 .. 466965 + NODE_518_length_445_cov_3.489888 155 .. 469 + Contig18 467279 .. 619892 |
| Contig182 | Contig182 1 .. 98351 + rev NODE_18487_length_4403_cov_9.712242 1 .. 4224 + Contig182 98351 .. 98351 |
| Contig183 | Contig183 1 .. 66926 + rev NODE_14529_length_2216_cov_7.519855 1 .. 2087 + Contig183 68364 .. 121939 + rev NODE_14488_length_550_cov_3.300000 1 .. 416 + Contig183 121939 .. 121939 |
| Contig1833 | Contig1833 1 .. 3009 + rev NODE_14637_length_1316_cov_10.629939 1 .. 1165 + Contig1833 3009 .. 3009 |
| Contig185 | Contig185 1 .. 73148 + rev NODE_23407_length_559_cov_3.688730 1 .. 430 + Contig185 73579 .. 94000 |
| Contig1861 | Contig1861 1 .. 2061 + rev NODE_16436_length_516_cov_1.891473 1 .. 389 + Contig1861 2061 .. 2061 |
| Contig187 | Contig187 1 ..75567 + CU694661 1 ..81981 |
| Contig1875 | Contig1875 1 .. 2169 + rev NODE_16353_length_1679_cov_5.949375 1 .. 1493 + Contig1875 2169 .. 2169 |
| Contig19 | Contig19 1 .. 491528 + rev NODE_3462_length_651_cov_11.815668 1 .. 218 + Contig19 491747 .. 520455 + CU855827 1 ..77273 + Contig19 596202 ..end |
| Contig2 | Contig2 1 .. 270750 + NODE_7416_length_463_cov_2.267819 199 .. 487 + Contig2 271038 ..458054  + rev CU694980 1 ..125054 + CU633974 8627 ..80422 + Contig2 689796 ..698255 + rev NODE_14692_length_1397_cov_7.984252 1 .. 1270 + + Contig2 699526 .. 800808 + NODE_1518_length_494_cov_3.611336 175 .. 518 + Contig2 801151 .. 1013382 + rev CU469389 + Contig2 1144127 ..end |
| Contig20 | NODE_7577_length_1084_cov_5.682657 1 .. 918 + Contig20 1 ..447093 + CU694536 1 ..98191 |
| Contig2023 | Contig2023 1 .. 2681 + NODE_2793_length_873_cov_5.423826 193 .. 897 + Contig2023 2681 .. 2681 |
| Contig205 | Contig205 1 .. 81309 + NODE_113_length_197_cov_43.746193 151 .. 221 + Contig205 81309 .. 81309 |
| Contig206 | Contig206 1 .. 12190 + NODE_5994_length_176_cov_4.835227 1 .. 95 + Contig206 12286 .. 74709 |
| Contig2068 | Contig2068 1 .. 2651 + NODE_17117_length_3584_cov_7.075614 151 .. 3608 + Contig2068 2651 .. 2651 |
| Contig21 | Contig21 1 .. 362966 + rev NODE_821_length_481_cov_3.532224 1 .. 341 + Contig21 363308 .. 647099 |
| Contig212 | Contig212 1 .. 77616 + NODE_1_length_438_cov_2.413242 151 .. 462 + Contig212 77616 .. 77616 |
| Contig214 | Contig214 1 .. 71605 + NODE_2650_length_503_cov_4.980119 193 .. 527 + Contig214 71605 .. 71605 |
| Contig217 | Contig217 1 ..47233 + CU638820 1 ..100007 |
| Contig218 | Contig218 1 .. 47278 + rev NODE_19452_length_1930_cov_5.424352 1 .. 1758 + Contig218 47278 .. 47278 |
| Contig2191 | Contig2191 1 .. 2182 + rev NODE_17793_length_591_cov_4.228426 1 .. 432 + Contig2191 2182 .. 2182 |
| Contig22 | Contig22 1 .. 153674 + rev NODE_313_length_471_cov_5.766454 1 .. 344 + Contig22 154019 .. 584877 |
| Contig220 | Contig220 1 .. 37380 + NODE_6471_length_492_cov_4.264228 157 .. 516 + Contig220 37739 .. 39033 + NODE_14678_length_1753_cov_6.051911 238 .. 1777 + Contig220 40572 .. 72659 |
| Contig222 | Contig222 1 .. 40233 + rev NODE_7630_length_683_cov_5.256223 1 .. 532 + Contig222 40446 .. 55851 |
| Contig225 | Contig225 1 .. 52862 + NODE_450_length_560_cov_3.657143 253 .. 584 + Contig225 52862 .. 52862 |
| Contig23 | Contig23 1 .. 379862 + rev NODE_2366_length_2653_cov_7.867320 1 .. 2489 + Contig23 381952 .. 596662 |

| | |
|---|---|
| Contig2361 | Contig2361 1 .. 1 + rev NODE_907_length_1688_cov_7.479265 540 .. 1712 + Contig2361 1 .. 2384 + rev NODE_15679_length_1433_cov_3.371249 1 .. 1297 + Contig2361 2384 .. 2384 |
| Contig24 | Contig24 1 .. 35558 + rev NODE_18212_length_4374_cov_3.618656 1 .. 4241 + Contig24 39800 .. 454541 |
| Contig2454 | Contig2454 1 .. 4305 + NODE_510_length_463_cov_3.053996 173 .. 487 + Contig2454 4305 .. 4305 |
| Contig248 | Contig248 1 .. 40993 + NODE_2121_length_451_cov_3.170732 159 .. 475 + Contig248 40993 .. 40993 |
| Contig25 | Contig25 1 .. 205574 + NODE_4588_length_443_cov_2.090293 154 .. 467 + Contig25 205887 .. 565665 |
| Contig26 | Contig26 1 .. 390489 + rev NODE_14341_length_1960_cov_21.487246 172 .. 1984 + Contig26 392302 .. 492606 + rev NODE_404_length_529_cov_7.674858 318 .. 19815 + Contig266 1 .. 17428 |
| Contig27 | Contig27 1 .. 205677 + NODE_1954_length_558_cov_4.743728 215 .. 582 + Contig27 206044 .. 217495 + NODE_14848_length_5464_cov_10.145498 1 .. 5488 + Contig27 225654 .. 359562 + NODE_7075_length_560_cov_5.392857 161 .. 584 + Contig27 359985 .. 600462 |
| Contig273 | Contig273 1 .. 1 + NODE_194_length_273_cov_18.494505 1 .. 122 + Contig273 1 .. 30874 |
| Contig276 | Contig276 1 .. 20623 + NODE_8636_length_449_cov_2.665924 153 .. 473 + Contig276 20943 .. 34795 + rev NODE_15662_length_872_cov_6.877294 1 .. 691 + Contig276 34795 .. 34795 |
| Contig277 | Contig277 1 .. 19644 + NODE_6017_length_482_cov_2.201245 1 .. 505 + Contig277 38258 .. 50896 + rev NODE_6778_length_3639_cov_10.760098 1 .. 3511 + Contig277 50896 .. 50896 |
| Contig28 | Contig28 1 ..139059 + CU855855 1 ..90842 + Contig28 249950 ..end |
| Contig281 | Contig281 1 .. 1 + NODE_1025_length_176_cov_5.210227 1 .. 14 + Contig281 1 .. 14704 |
| Contig288 | Contig288 1 .. 23612 + rev NODE_18215_length_1053_cov_6.152896 1 .. 910 + Contig288 24523 .. 50717 |
| Contig289 | Contig289 1 .. 21179 + NODE_4738_length_724_cov_5.578729 168 .. 748 + Contig289 21759 .. 31437 |
| Contig29 | Contig29 1 ..53101 + CU469394 1 ..94703 + Contig29 119034 .. 127032 + NODE_8469_length_1395_cov_8.497491 184 .. 1419 + Contig29 128267 .. 195517 + NODE_1420_length_473_cov_4.350951 159 .. 497 + Contig29 195855 .. 553050 + rev NODE_3037_length_3430_cov_10.689795 1 .. 3259 + Contig29 553050 .. end |
| Contig2969 | Contig2969 1 .. 4128 + NODE_1607_length_1260_cov_17.555555 287 .. 1284 + Contig2969 4128 .. 4128 |
| Contig297 | Contig297 1 .. 1343 + NODE_183_length_635_cov_7.535433 398 .. 659 + Contig297 1604 .. 25642 + NODE_3869_length_820_cov_6.934146 157 .. 844 + |
| Contig298 | Contig298 1 .. 26872 + NODE_2209_length_398_cov_8.100503 161 .. 422 + Contig298 26872 .. 26872 |
| Contig3 | Contig3 1 .. 487012 + rev NODE_1492_length_508_cov_3.267717 1 .. 357 + Contig3 487370 .. 513388 + rev CU694971 1 .. 99582 + Contig3 584628 .. 1101443 + rev NODE_5744_length_665_cov_4.942857 1 .. 534 + Contig3 1101443 .. 1101443 |
| Contig30 | Contig30 1 .. 71785 + NODE_2953_length_438_cov_3.100457 155 .. 462 + Contig30 72092 .. 475048 |
| Contig300 | Contig300 1 .. 20996 + NODE_4553_length_380_cov_5.907895 154 .. 404 + Contig300 20996 .. 20996 |
| Contig302 | Contig302 1 .. 32456 + rev NODE_14806_length_1283_cov_5.076384 1 .. 196859 |
| Contig304 | Contig304 1 .. 28811 + rev NODE_13616_length_971_cov_7.786818 1 .. 839 + Contig304 28811 .. 28811 |
| Contig3093 | Contig3093 1 .. 4482 + rev NODE_8201_length_662_cov_3.252266 1 .. 510 + Contig3093 4482 .. 4482 |
| Contig31 | Contig247 1 ..22880 + rev CU694967 1 ..79556 + Contig31 50686 .. 395994 + NODE_29_length_410_cov_26.929268 231 .. 13874 + Contig31 409638 .. end |
| Contig311 | Contig311 1 .. 29808 + NODE_14409_length_2558_cov_7.649726 151 .. 2582 + Contig311 29808 .. 29808 |
| Contig314 | Contig314 1 .. 24874 + rev NODE_14809_length_1026_cov_3.161793 1 .. 885 + Contig314 24874 .. 24874 |
| Contig319 | Contig319 1 .. 1 + NODE_126_length_514_cov_14.714007 1 .. 339 + Contig319 1 .. 8040 |
| Contig320 | Contig320 1 .. 24609 + rev NODE_14250_length_12494_cov_10.207219 1 .. 12365 + Contig320 24609 .. 24609 |
| Contig325 | Contig325 1 .. 1 + rev NODE_15163_length_2859_cov_7.077300 184 .. 2883 + Contig325 1 .. 21795 |
| Contig33 | Contig33 1 ..16824 + CU694981 1 ..104149 + Contig33 120975 ..537736 |
| Contig331 | Contig331 1 .. 14497 + rev NODE_2540_length_797_cov_11.115433 1 .. 664 + Contig331 14497 .. 14497 |
| Contig34 | Contig34 1 .. 366792 + NODE_4365_length_465_cov_3.135484 188 .. 489 + Contig34 367093 .. 496322 |
| Contig344 | Contig344 1 .. 1 + NODE_4_length_609_cov_9.159278 1 .. 107 + Contig344 1 .. 25673 + NODE_17461_length_1074_cov_2.853817 167 .. 1098 + |
| Contig348 | Contig348 1 .. 18496 + rev NODE_40_length_273_cov_9.593407 1 .. 145 + Contig348 18496 .. 18496 |
| Contig35 | Contig35 1 ..4444 + rev CU856318 1 ..95046 + Contig35 122119 ..371820 + CU694969 1 ..99415 + Contig35 421351 ..end |
| Contig360 | Contig360 1 .. 20124 + rev NODE_95_length_305_cov_98.324593 1 .. 102 + Contig360 20227 .. 22956 |
| Contig3624 | Contig3624 1 .. 792 + NODE_6964_length_603_cov_4.396351 261 .. 627 + Contig3624 1158 .. 4220 |
| Contig368 | Contig368 1 .. 12713 + NODE_15945_length_860_cov_6.098837 175 .. 884 + Contig368 12713 .. 12713 |
| Contig37 | Contig37 1 ..163042 + CU469400 1 ..76478 + Contig37 214366 ..316261 + CU469399 1 ..74137 + Contig37 367109 ..end |
| Contig3713 | Contig3713 1 .. 2677 + rev NODE_15390_length_3855_cov_9.522438 1 .. 3719 + Contig3713 2677 .. 2677 |
| Contig375 | Contig375 1 .. 18900 + rev NODE_14327_length_7204_cov_10.817463 1 .. 12098 + |
| Contig376 | Contig376 1 .. 12234 + rev NODE_5996_length_1195_cov_7.857740 1 .. 1046 + Contig376 12234 .. 12234 |
| Contig38 | rev CU694291 1 ..98065 + Contig38 124120 ..end |
| Contig381 | Contig381 1 .. 7907 + NODE_231_length_2455_cov_18.198370 2307 .. 2479 + Contig381 7907 .. 7907 |
| Contig386 | Contig386 1 .. 18175 + rev NODE_19030_length_3462_cov_7.522531 1 .. 3236 + Contig386 18175 .. 18175 |
| Contig4 | Contig4 1 .. 227908 + rev NODE_6936_length_1336_cov_4.700599 1 .. 1158 + Contig4 229067 .. 698377 + rev NODE_7354_length_445_cov_2.777528 1 .. 298 + Contig4 698676 .. 888725 + rev NODE_15515_length_1254_cov_3.856459 1 .. 1127 + Contig4 889553 .. 1101618 |
| Contig40 | Contig40 1 .. 43280 + NODE_4016_length_484_cov_2.291322 190 .. 508 + Contig40 43598 .. 341184 |
| Contig400 | Contig400 1 .. 15663 + rev NODE_18803_length_1280_cov_5.189063 1 .. 1148 + Contig400 15663 .. 15663 |
| Contig403 | Contig403 1 .. 19815 + NODE_1752_length_338_cov_27.875740 307 .. 362 + Contig403 19815 .. 19815 |
| Contig405 | Contig405 1 .. 15792 + rev NODE_8707_length_495_cov_2.846465 1 .. 346 + Contig405 15792 .. 15792 |

| | |
|---|---|
| Contig408 | Contig408 1 .. 13313 + rev NODE_6170_length_509_cov_3.911591 1 .. 367 + Contig408 13313 .. 13313 |
| Contig41 | Contig41 1 .. 403002 + rev NODE_1401_length_947_cov_1.885956 1 .. 820 + Contig41 403763 .. 459702 |
| Contig4148 | Contig4148 1 .. 4150 + NODE_5270_length_1027_cov_64.281403 155 .. 1051 + Contig4148 4150 .. 4150 |
| Contig417 | Contig417 1 .. 21320 + rev NODE_10144_length_607_cov_4.341022 1 .. 474 + Contig417 21320 .. 21320 |
| Contig42 | Contig42 1 .. 264262 + NODE_17133_length_1436_cov_7.683147 209 .. 1460 + Contig42 265513 .. 421951 |
| Contig421 | Contig421 1 .. 17897 + rev NODE_6062_length_473_cov_2.756871 1 .. 334 + Contig421 18132 .. 20192 |
| Contig425 | Contig425 1 .. 18055 + rev NODE_22337_length_579_cov_3.170985 1 .. 422 + Contig425 18055 .. 18055 |
| Contig426 | Contig426 1 .. 14412 + rev NODE_14428_length_5722_cov_10.105732 1 .. 5578 + Contig426 14412 .. 14412 |
| Contig427 | Contig427 1 .. 12424 + rev NODE_159_length_1934_cov_74.370735 1 .. 1627 + Contig427 12424 .. 12424 |
| Contig428 | Contig428 1 .. 17826 + rev NODE_16882_length_2436_cov_6.986042 1 .. 2309 + Contig428 17826 .. 17826 |
| Contig43 | Contig43 1 .. 311438 + rev NODE_8971_length_735_cov_3.473469 1 .. 564 + Contig43 312003 .. 314510 + rev NODE_16026_length_1676_cov_4.460024 1 .. 1357 + Contig43 315638 .. 424029 |
| Contig433 | Contig433 1 .. 15193 + NODE_3929_length_433_cov_3.422633 152 .. 457 + Contig433 15193 .. 15193 |
| Contig44 | Contig44 1 .. 81289 + rev NODE_6450_length_565_cov_3.194690 1 .. 383 + Contig44 81673 .. 301395 + CU694966 1 ..89842 |
| Contig45 | Contig45 1 ..183185 + CU469398 1 ..102410 + Contig45 284211 ..end |
| Contig456 | Contig456 1 .. 470 + NODE_5201_length_595_cov_4.863865 198 .. 619 + Contig456 891 .. 14190 |
| Contig46 | Contig46 1 .. 223023 + rev NODE_15347_length_1758_cov_7.529579 1 .. 1608 + Contig46 224632 .. 430272 |
| Contig463 | Contig463 1 .. 14184 + rev NODE_4746_length_509_cov_3.341847 1 .. 369 + Contig463 14184 .. 14184 |
| Contig47 | Contig47 1 .. 161095 + NODE_7366_length_580_cov_4.091379 159 .. 604 + Contig47 161540 .. 291716 + NODE_5325_length_440_cov_2.700000 153 .. 464 + Contig47 292027 .. 421817 |
| Contig477 | Contig477 1 .. 13758 + NODE_1064_length_489_cov_3.433538 194 .. 513 + Contig477 13758 .. 13758 |
| Contig48 | Contig48 1 .. 191047 + rev NODE_16025_length_1373_cov_6.369265 1 .. 1197 + Contig48 192245 .. 278537 + CU469405 1 ..78006 + NODE_157_length_613_cov_94.970634 35 .. 6659 |
| Contig484 | Contig484 1 .. 1 + NODE_672_length_424_cov_19.028301 1 .. 292 + Contig484 1 .. 7388 |
| Contig49 | Contig49 1 .. 331478 + rev NODE_7376_length_758_cov_5.889182 1 .. 617 + Contig49 332096 .. 358873 |
| Contig491 | Contig491 1 .. 8758 + NODE_4774_length_545_cov_4.724771 164 .. 569 + Contig491 8758 .. 8758 |
| Contig493 | Contig493 1 .. 12429 + rev NODE_14610_length_4792_cov_10.597245 1 .. 4629 + Contig493 12429 .. 12429 |
| Contig5 | Contig5 1 .. 285699 + rev NODE_8829_length_706_cov_4.325779 1 .. 538 + Contig5 286238 .. 310010 + rev NODE_21192_length_2880_cov_8.858334 1 .. 2741 + Contig5 312752 .. 613894 + rev NODE_1626_length_1749_cov_4.488851 1 .. 1593 + Contig5 615488 .. 835074 + NODE_3345_length_431_cov_2.190255 166 .. 455 + Contig5 835363 .. 879806 |
| Contig50 | CU694287 1 ..41447 + Contig50 1 ..407448 |
| Contig503 | Contig503 1 .. 11355 + NODE_15842_length_1746_cov_6.022337 159 .. 1770 + Contig503 11355 .. 11355 |
| Contig51 | Contig51 1 .. 375428 + NODE_18596_length_840_cov_2.229762 171 .. 864 + Contig51 375428 .. 375428 |
| Contig516 | Contig516 1 .. 6496 + NODE_3006_length_475_cov_3.305263 179 .. 499 + Contig516 6496 .. 6496 |
| Contig517 | Contig517 1 .. 10704 + rev NODE_20130_length_1934_cov_5.732162 1 .. 1804 + Contig517 10704 .. 10704 |
| Contig52 | Contig52 1 .. 110128 + NODE_6946_length_534_cov_3.404494 260 .. 558 + Contig52 110426 .. 305086 + CU672240 90046 ..109627 |
| Contig53 | Contig53 1 .. 191711 + NODE_1587_length_450_cov_4.173333 151 .. 474 + Contig53 192034 .. 305567 + CU855859 1 ..96477 |
| Contig535 | Contig535 1 .. 6465 + NODE_16523_length_2819_cov_6.350833 151 .. 2843 + Contig535 6465 .. 6465 |
| Contig536 | Contig536 1 .. 1 + NODE_812_length_1713_cov_25.736135 1 .. 1188 + Contig536 1 .. 9922 + NODE_19258_length_5867_cov_7.049088 155 .. 5891 + |
| Contig54 | Contig54 1 .. 158997 + rev NODE_1232_length_451_cov_5.971175 1 .. 317 + Contig54 159315 .. 193638 + rev NODE_1425_length_600_cov_3.035000 1 .. 402 + Contig54 194041 .. 357926 |
| Contig546 | Contig546 1 – 3064 + rev CU694974 1 – 115766 |
| Contig55 | Contig55 1 .. 118040 + rev NODE_22541_length_535_cov_2.796262 1 .. 405 + Contig55 118386 .. 231324 + rev NODE_16001_length_3402_cov_9.982364 1 .. 3253 + Contig55 234578 .. 350972 |
| Contig555 | Contig555 1 .. 5813 + NODE_8781_length_1029_cov_4.029154 153 .. 1053 + Contig555 5813 .. 5813 |
| Contig56 | Contig56 1 .. 74134 + rev NODE_15378_length_1820_cov_5.259890 1 .. 1643 + Contig56 75778 .. 360670 + NODE_13662_length_849_cov_4.440518 427 .. 873 + |
| Contig563 | Contig563 1 .. 3042 + rev NODE_4221_length_488_cov_6.399590 1 .. 328 + Contig563 3371 .. 11335 + rev NODE_15583_length_1275_cov_6.377255 1 .. 1079 + Contig563 11335 .. 11335 |
| Contig577 | Contig577 1 .. 1 + rev NODE_8515_length_2031_cov_10.483013 529 .. 2055 + Contig577 1 .. 5444 |
| Contig58 | Contig58 1 .. 175610 + NODE_14760_length_1979_cov_8.186963 173 .. 2003 + Contig58 177440 .. 415316 |
| Contig583 | Contig583 1 .. 1049 + rev NODE_8938_length_469_cov_3.882729 1 .. 293 + Contig583 1343 .. 11663 |
| Contig59 | Contig59 1 ..57252 + rev CU633976 1 ..89791 + Contig59 162575 ..306602 + CU855832 1 ..62978 + Contig59 360959 ..end |
| Contig6 | Contig6 1 ..179815 + rev CU633880 1 ..84452 + Contig6 271630 ..436793 + CU694993 1 ..95704 + Contig6 542523 .. 626158 + rev NODE_6117_length_477_cov_2.559748 1 .. 335 + Contig6 626494 .. 644888 + rev NODE_8339_length_520_cov_2.911538 1 .. 366 + Contig6 645255 .. end |
| Contig60 | CU694962 1 ..57507 + Contig60 37145 ..end |
| Contig605 | Contig605 1 .. 5777 + NODE_3221_length_729_cov_12.895747 283 .. 753 + Contig605 5777 .. 5777 |
| Contig61 | Contig61 1 .. 83644 + NODE_14339_length_1425_cov_4.524210 216 .. 1449 + Contig61 84877 .. 170306 + rev NODE_6993_length_1627_cov_4.542717 1 .. 1481 + Contig61 171788 .. 414948 |
| Contig628 | Contig628 1 .. 7554 + NODE_14511_length_3119_cov_10.034947 183 .. 3143 + Contig628 7554 .. 7554 |
| Contig630 | Contig630 1 .. 6355 + rev NODE_21838_length_844_cov_7.751185 1 .. 687 + Contig630 6355 .. 6355 |
| Contig633 | CU694760 1 ..210875 + Contig633 1269 ..4357 |

| | |
|---|---|
| Contig634 | Contig634 1 .. 5870 + NODE_3390_length_877_cov_20.280502 209 .. 901 + Contig634 5870 .. 5870 |
| Contig64 | Contig64 1 ..93731 + rev CU694537 1 ..73946 + Contig64 201331 ..end |
| Contig651 | Contig651 1 .. 2257 + NODE_8425_length_668_cov_6.769461 156 .. 692 + Contig651 2257 .. 2257 |
| Contig652 | Contig652 1 .. 1 + NODE_328_length_149_cov_56.013424 1 .. 11 + Contig652 1 .. 5412 + NODE_558_length_776_cov_42.896908 220 .. 800 + |
| Contig66 | Contig66 1 .. 132303 + rev NODE_15593_length_1145_cov_5.543231 1 .. 1169 + Contig66 174027 .. 195876 + rev NODE_9587_length_1151_cov_3.461338 1 .. 985 + Contig66 196862 .. end |
| Contig67 | Contig67 1 .. 42087 + NODE_4033_length_1090_cov_13.867890 158 .. 1114 + Contig67 43043 .. 96779 + rev NODE_14351_length_4131_cov_10.676834 1 .. 3990 + Contig67 100470 .. 142561 + rev NODE_4542_length_471_cov_3.895966 1 .. 342 + Contig67 142904 .. end |
| Contig670 | Contig670 1 .. 4887 + rev NODE_592_length_462_cov_2.448052 1 .. 328 + Contig670 5216 .. 6466 |
| Contig68 | Contig68 1 .. 47902 + NODE_14222_length_905_cov_6.382320 151 .. 929 + Contig68 48680 .. 339720 |
| Contig682 | Contig682 1 .. 5750 + rev NODE_9776_length_654_cov_3.660550 1 .. 517 + Contig682 6268 .. 7841 |
| Contig69 | Contig69 1 .. 136245 + NODE_110_length_464_cov_3.030172 151 .. 488 + Contig69 136582 .. 349856 + rev NODE_6644_length_464_cov_3.553879 1 .. 331 + Contig69 349856 .. 349856 |
| Contig697 | Contig697 1 .. 2838 + NODE_20545_length_4837_cov_5.825511 233 .. 4861 + Contig697 2838 .. 2838 |
| Contig70 | Contig70 1 .. 28866 + NODE_9917_length_687_cov_4.344978 324 .. 711 + Contig70 29193 .. 182126 + NODE_669_length_442_cov_1.737557 157 .. 466 + Contig70 182435 .. 263256 |
| Contig704 | Contig704 1 .. 3304 + rev NODE_4846_length_516_cov_4.172481 1 .. 386 + Contig704 3304 .. 3304 |
| Contig712 | Contig712 1 .. 1 + rev NODE_250_length_199_cov_27.703518 187 .. 223 + Contig712 1 .. 4532 |
| Contig73 | CU694289 1 ..92391 + Contig73 100949 ..end |
| Contig734 | Contig734 1 .. 2235 + rev NODE_979_length_875_cov_15.531428 1 .. 582 + Contig734 2818 .. 5444 |
| Contig74 | Contig74 1 ..162837 + CU469396 1 ..109766 |
| Contig744 | Contig744 1 .. 2705 + rev NODE_4146_length_187_cov_5.620321 1 .. 101 + Contig744 2705 .. 2705 |
| Contig75 | Contig75 1 .. 53501 + rev NODE_10034_length_1076_cov_4.166357 1 .. 945 + Contig75 54447 .. 102861 + NODE_218_length_434_cov_2.481567 151 .. 458 + Contig75 103168 .. 121409 + CU694973 1 ..147660 + Contig75 211709 ..end |
| Contig76 | Contig76 1 .. 137076 + rev NODE_1209_length_469_cov_2.648188 1 .. 325 + Contig76 137402 .. 349624 |
| Contig763 | Contig763 1 .. 2450 + NODE_477_length_468_cov_4.472222 166 .. 492 + Contig763 2450 .. 2450 |
| Contig782 | Contig782 1 .. 1715 + NODE_263_length_455_cov_3.597802 151 .. 479 + Contig782 2043 .. 5357 |
| Contig790 | Contig790 1 .. 1 + rev NODE_15486_length_1328_cov_19.017319 248 .. 1352 + Contig790 1 .. 3804 |
| Contig8 | rev CU694972 1 ..87649 + Contig8 33698 ..596124 + CU694663 1 ..12551 + rev CU855831 1 ..77771 + CU694976 46770 ..112370 + Contig8 789381 ..end |
| Contig805 | Contig805 1 .. 2094 + NODE_3_length_442_cov_4.106335 167 .. 466 + Contig805 2094 .. 2094 |
| Contig815 | Contig815 1 .. 495 + NODE_14462_length_2430_cov_7.737037 273 .. 2454 + Contig815 2676 .. 3678 |
| Contig823 | Contig823 1 .. 5520 + rev NODE_16145_length_4748_cov_7.206824 1 .. 4592 + Contig823 5520 .. 5520 |
| Contig824 | Contig824 1 .. 2447 + NODE_7518_length_474_cov_4.388186 156 .. 498 + Contig824 2447 .. 2447 |
| Contig829 | Contig829 1 .. 1 + NODE_951_length_1075_cov_94.086510 1 .. 800 + Contig829 1 .. 5492 |
| Contig84 | Contig84 1 .. 20234 + rev NODE_14460_length_807_cov_5.354399 1 .. 676 + Contig84 20911 .. 237125 |
| Contig849 | Contig849 1 .. 1 + NODE_935_length_748_cov_8.784760 1 .. 272 + Contig849 1 .. 3186 + rev NODE_16443_length_736_cov_5.957880 1 .. 453 + Contig849 3186 .. 3186 |
| Contig85 | Contig85 1 ..106054 + CU469395 1 .. 99525 + rev NODE_4246_length_599_cov_4.176961 1 .. 471 + CU469395 99697 .. 99915 + Contig85 204267 .. 234167 + rev CU469391 1 ..101110 + Contig85 332408 ..366117 + NODE_3332_length_461_cov_3.455531 160 .. 485 + Contig85 366442 .. end |
| Contig86 | Contig86 1 .. 147102 + rev NODE_14504_length_2196_cov_7.049181 1 .. 2066 + Contig86 149169 .. 283757 |
| Contig867 | Contig867 1 .. 1240 + NODE_195_length_531_cov_78.436913 444 .. 555 + Contig867 1351 .. 4591 |
| Contig87 | Contig87 1 .. 95692 + NODE_5703_length_632_cov_3.813291 306 .. 656 + Contig87 96042 .. 189393 |
| Contig88 | Contig88 1 ..115787 + CU855849 1 ..101004 + NODE_212_length_463_cov_3.153348 180 .. 487 |
| Contig898 | Contig898 1 .. 4661 + NODE_4070_length_472_cov_2.563559 156 .. 496 + Contig898 4661 .. 4661 |
| Contig9 | Contig9 1 .. 421407 + NODE_19464_length_662_cov_2.889728 155 .. 686 + Contig9 421938 .. 730399 + rev NODE_10648_length_578_cov_4.174740 1 .. 444 + Contig9 730844 .. 788031 |
| Contig90 | rev CU694960 1 ..84167 + Contig90 52642 ..end |
| Contig908 | Contig908 1 .. 2153 + NODE_1684_length_690_cov_5.473913 158 .. 714 + Contig908 2153 .. 2153 |
| Contig918 | Contig918 1 .. 2949 + rev NODE_6769_length_569_cov_2.644991 1 .. 422 + Contig918 2949 .. 2949 |
| Contig92 | Contig92 1 .. 125077 + rev NODE_2948_length_456_cov_2.539474 1 .. 296 + Contig92 125374 .. 227751 |
| Contig93 | Contig93 1 .. 90769 + NODE_6980_length_1044_cov_4.462644 200 .. 1068 + Contig93 91637 .. 286667 + NODE_15694_length_3257_cov_10.260670 184 .. 3281 + |
| Contig938 | Contig938 1 .. 1 + NODE_882_length_396_cov_8.035354 1 .. 235 + Contig938 1 .. 2231 |
| Contig94 | Contig94 1 .. 267283 + NODE_4344_length_781_cov_4.618438 203 .. 805 + Contig94 267885 .. 281273 |
| Contig945 | Contig945 1 .. 1260 + NODE_5483_length_478_cov_2.523013 160 .. 502 + Contig945 1602 .. 5202 |
| Contig948 | Contig948 1 .. 2041 + NODE_5136_length_472_cov_2.546610 158 .. 496 + Contig948 2041 .. 2041 |
| Contig949 | Contig949 1 .. 8025 + rev NODE_18477_length_1807_cov_6.795241 1 .. 1679 + Contig949 8025 .. 8025 |
| Contig95 | Contig95 1 .. 19929 + NODE_16536_length_2120_cov_7.550000 177 .. 2144 + Contig95 21896 .. 42883 + rev NODE_19117_length_6155_cov_7.971080 1 .. 6021 + Contig95 45305 .. 288207 + NODE_1691_length_452_cov_3.475664 155 .. 476 + |
| Contig960 | Contig960 1 .. 1301 + rev NODE_12728_length_562_cov_3.049822 1 .. 374 + Contig960 1676 .. 4775 |
| Contig97 | Contig97 1 .. 112333 + NODE_17945_length_8351_cov_10.547599 166 .. 8375 + Contig97 120542 .. 140751 + rev NODE_14657_length_6702_cov_10.688451 1 .. 6547 + Contig97 147299 .. 329294 |

| | |
|---|---|
| Contig977 | Contig977 1 .. 2670 + rev NODE_3598_length_992_cov_4.484879 1 .. 854 + Contig977 2670 .. 2670 |
| Contig983 | Contig983 1 .. 5312 + NODE_1192_length_10230_cov_10.357380 163 .. 10254 + Contig983 5312 .. 5312 |
| Contig990 | Contig990 1 .. 3597 + rev NODE_14389_length_650_cov_16.389231 1 .. 516 + Contig990 3597 .. 3597 |
| Contig991 | Contig991 1 .. 3138 + NODE_21273_length_2069_cov_5.269212 162 .. 2093 + Contig991 3138 .. 3138 |
| Contig992 | Contig992 1 .. 1 + NODE_17086_length_1456_cov_5.812500 1 .. 954 + Contig992 1 .. 2993 |
| Contig995 | Contig995 1 .. 2053 + NODE_3157_length_584_cov_6.989726 153 .. 608 + Contig995 2053 .. 20532 |
| SuperCU469403 | rev CU469403 2001 ..98732 + CU469388 1 ..117284 + CU469392 900 ..114210 + CU694662 13653 ..106040 |
| SuperCU694965 | CU694965 1 ..132537 + rev CU469401 12588 ..103744 |
| SuperCU694970 | CU694970 1 ..75255 + rev CU856298 66845 ..77235 |

## Appendix table 3.5: Bacterial contamination on contigs (top 50 contigs)

| Contig | Count of Hit | Sum of Length |
|---|---|---|
| Cu694975 | 321 | 66,621 |
| Cu694660 | 138 | 20,528 |
| SuperContig28 | 1 | 3791 |
| SuperContig3669 | 4 | 2028 |
| SuperContig3704 | 1 | 2014 |
| SuperContig2186 | 2 | 1552 |
| SuperContig1743 | 3 | 1543 |
| SuperContig2681 | 2 | 1170 |
| SuperContig4760 | 1 | 1160 |
| SuperContig2675 | 2 | 1128 |
| SuperContig3569 | 1 | 1128 |
| SuperContig2234 | 2 | 1109 |
| SuperContig4133 | 3 | 1104 |
| SuperContig779 | 3 | 1103 |
| SuperContig3136 | 2 | 1072 |
| SuperContig3910 | 2 | 1064 |
| SuperContig4139 | 2 | 1053 |
| SuperContig4648 | 2 | 1009 |
| SuperContig3695 | 2 | 982 |
| SuperContig2810 | 2 | 943 |
| SuperContig3812 | 2 | 908 |
| SuperContig3932 | 2 | 906 |
| SuperContig3644 | 2 | 895 |
| SuperContig4575 | 2 | 892 |
| SuperContig1461 | 4 | 864 |
| SuperContig2778 | 6 | 854 |
| SuperContig3010 | 2 | 835 |
| SuperContig4463 | 2 | 834 |
| SuperContig1805 | 1 | 831 |
| SuperContig2642 | 2 | 829 |
| SuperContig4419 | 2 | 823 |
| SuperContig4169 | 2 | 819 |
| SuperContig1305 | 2 | 809 |
| SuperContig3068 | 1 | 793 |
| SuperContig4602 | 1 | 782 |
| SuperContig1291 | 5 | 777 |
| SuperContig1423 | 4 | 766 |
| SuperContig3103 | 2 | 766 |
| SuperContig3595 | 2 | 759 |
| SuperContig2267 | 2 | 756 |
| SuperContig2346 | 3 | 743 |
| SuperContig4199 | 2 | 743 |
| SuperContig219 | 1 | 736 |
| SuperContig2046 | 1 | 718 |
| SuperContig3694 | 2 | 717 |
| SuperContig2134 | 3 | 706 |
| SuperContig4281 | 1 | 680 |
| SuperContig3165 | 2 | 670 |
| SuperContig2593 | 3 | 657 |
| SuperContig3524 | 1 | 655 |
| … | … | … |
| **Grand Total** | **859** | **185,062** |

## Appendix table 3.6: Bacterial contaminants (top 50 contigs)

| Best Hit | Total |
|---|---|
| **Xanthomonas campestris pv. campestris str. B100, complete genome** | **120** |
| **Xanthomonas axonopodis pv. citri str. 306, complete genome** | **108** |
| **Xanthomonas campestris pv. vesicatoria str. 85-10, complete genome** | **100** |
| **Xanthomonas oryzae pv. oryzae PXO99A, complete genome** | **92** |
| **Methylobacillus flagellatus KT, complete genome** | **58** |
| **Xanthomonas campestris pv. campestris str. ATCC 33913, complete genome** | **41** |
| **Flavobacterium johnsoniae UW101, complete genome** | **41** |
| Xanthomonas oryzae pv. oryzae MAFF 311018, complete genome | 11 |
| Acidovorax avenae subsp. citrulli AAC00-1, complete genome | 9 |
| Janthinobacterium sp. Marseille, complete genome | 9 |
| Pseudomonas fluorescens Pf0-1, complete genome | 7 |
| Variovorax paradoxus S110 chromosome 1, complete genome | 7 |
| Polaromonas naphthalenivorans CJ2, complete genome | 7 |
| Sphingomonas wittichii RW1, complete genome | 6 |
| Methylibium petroleiphilum PM1, complete genome | 6 |
| Leptothrix cholodnii SP-6, complete genome | 6 |
| Thiobacillus denitrificans ATCC 25259, complete genome | 6 |
| Pseudomonas fluorescens Pf-5, complete genome | 5 |
| Agrobacterium tumefaciens str. C58 chromosome linear, complete sequence | 5 |
| Azotobacter vinelandii DJ, complete genome | 5 |
| Pseudomonas entomophila L48, complete genome | 5 |
| Novosphingobium aromaticivorans DSM 12444, complete genome | 5 |
| Sphingopyxis alaskensis RB2256, complete genome | 4 |
| Xanthomonas campestris pv. campestris str. 8004, complete genome | 4 |
| Bordetella pertussis Tohama I, complete genome | 4 |
| Delftia acidovorans SPH-1, complete genome | 4 |
| Ralstonia pickettii 12J chromosome 1, complete sequence | 4 |
| Laribacter hongkongensis HLHK9, complete genome | 4 |
| Thauera sp. MZ1T, complete genome | 4 |
| Pseudomonas fluorescens SBW25, complete genome | 4 |
| Verminephrobacter eiseniae EF01-2, complete genome | 4 |
| Xanthomonas oryzae pv. oryzae KACC10331, complete genome | 4 |
| Pseudomonas putida GB-1, complete genome | 4 |
| Pseudomonas putida F1, complete genome | 4 |
| Burkholderia phytofirmans PsJN chromosome 1, complete genome | 3 |
| Erythrobacter litoralis HTCC2594, complete genome | 3 |
| Rhodoferax ferrireducens T118, complete genome | 3 |
| Nitrosomonas eutropha C91, complete genome | 3 |
| Chromobacterium violaceum ATCC 12472, complete genome | 3 |
| Nitrosospira multiformis ATCC 25196 chromosome 1, complete sequence | 3 |
| Rhizobium sp. NGR234, complete genome | 3 |
| Phenylobacterium zucineum HLK1, complete genome | 3 |
| Diaphorobacter sp. TPSY, complete genome | 3 |
| Burkholderia glumae BGR1 chromosome 1, complete genome | 3 |
| Cellvibrio japonicus Ueda107, complete genome | 3 |
| Pseudomonas putida KT2440, complete genome | 3 |
| Azoarcus sp. BH72, complete genome | 2 |
| Aromatoleum aromaticum EbN1, complete genome | 2 |
| Rhodopseudomonas palustris HaA2, complete genome | 2 |
| Herminiimonas arsenicoxydans, complete genome | 2 |
| … | … |
| Grand Total | 859 |

**Appendix table 3.7: *Arabidopsis thaliana* contamination**

| Contig | Count of Hit | Sum of Length |
|---|---|---|
| SuperContig2363 | 1 | 97 |
| SuperContig4001 | 2 | 1392 |
| SuperContig4406 | 2 | 339 |
| SuperContig591 | 2 | 128 |
| SuperContig657 | 1 | 94 |
| SuperContig84 | 2 | 572 |
| SuperCu856152 | 3 | 476 |
| Grand Total | 13 | 3098 |

| Hit | Total |
|---|---|
| 3 | 5 |
| 4 | 1 |
| 5 | 1 |
| chloroplast | 2 |
| mitochondria | 4 |
| | |
| Grand Total | 13 |

# Appendices for Chapter 4

## Appendix table 4.1: Go Terms

| GO Class ID | Definitions | Counts | Fractions |
|---|---|---|---|
| GO:0003674 | molecular_function | 9837 | 17.9% |
| GO:0008150 | biological_process | 6576 | 12.0% |
| GO:0005488 | binding | 5176 | 9.4% |
| GO:0005575 | cellular_component | 4243 | 7.7% |
| GO:0003824 | catalytic activity | 4030 | 7.3% |
| GO:0008152 | metabolism | 3972 | 7.2% |
| GO:0005623 | cell | 2100 | 3.8% |
| GO:0005622 | intracellular | 1460 | 2.7% |
| GO:0016787 | hydrolase activity | 1421 | 2.6% |
| GO:0005515 | protein binding | 1321 | 2.4% |
| GO:0000166 | nucleotide binding | 1206 | 2.2% |
| GO:0019538 | protein metabolism | 1153 | 2.1% |
| GO:0016740 | transferase activity | 1117 | 2.0% |
| GO:0009058 | biosynthesis | 1055 | 1.9% |
| GO:0003676 | nucleic acid binding | 959 | 1.8% |
| GO:0006139 | nucleobase, nucleoside, nucleotide and nucleic acid metabolism | 904 | 1.7% |
| GO:0006810 | transport | 809 | 1.5% |
| GO:0016301 | kinase activity | 573 | 1.0% |
| GO:0005737 | cytoplasm | 551 | 1.0% |
| GO:0004672 | protein kinase activity | 485 | 0.9% |
| GO:0006464 | protein modification | 407 | 0.7% |
| GO:0003677 | DNA binding | 369 | 0.7% |
| GO:0005215 | transporter activity | 362 | 0.7% |
| GO:0006412 | protein biosynthesis | 312 | 0.6% |
| GO:0005634 | nucleus | 307 | 0.6% |
| GO:0009056 | catabolism | 261 | 0.5% |
| GO:0008233 | peptidase activity | 249 | 0.5% |
| GO:0006259 | DNA metabolism | 245 | 0.5% |
| GO:0007154 | cell communication | 239 | 0.4% |
| GO:0005975 | carbohydrate metabolism | 234 | 0.4% |
| GO:0003723 | RNA binding | 217 | 0.4% |
| GO:0016043 | cell organization and biogenesis | 185 | 0.3% |
| GO:0015031 | protein transport | 165 | 0.3% |
| GO:0007165 | signal transduction | 161 | 0.3% |
| GO:0006629 | lipid metabolism | 149 | 0.3% |
| GO:0005198 | structural molecule activity | 144 | 0.3% |
| GO:0005840 | ribosome | 143 | 0.3% |
| GO:0006996 | organelle organization and biogenesis | 139 | 0.3% |
| GO:0006950 | response to stress | 130 | 0.2% |
| GO:0006811 | ion transport | 126 | 0.2% |

## Appendix table 4.1: Go Terms

| GO Class ID | Definitions | Counts | Fractions |
|---|---|---|---|
| **GO:0008289** | lipid binding | 93 | 0.2% |
| **GO:0004518** | nuclease activity | 88 | 0.2% |
| **GO:0005694** | chromosome | 72 | 0.1% |
| **GO:0005856** | cytoskeleton | 68 | 0.1% |
| **GO:0030234** | enzyme regulator activity | 67 | 0.1% |
| **GO:0005509** | calcium ion binding | 64 | 0.1% |
| **GO:0004721** | phosphoprotein phosphatase activity | 62 | 0.1% |
| **GO:0004871** | signal transducer activity | 56 | 0.1% |
| **GO:0003700** | transcription factor activity | 56 | 0.1% |
| **GO:0003774** | motor activity | 55 | 0.1% |
| **GO:0005739** | mitochondrion | 51 | 0.1% |
| **GO:0019725** | cell homeostasis | 50 | 0.1% |
| **GO:0005783** | endoplasmic reticulum | 43 | 0.1% |
| **GO:0006091** | generation of precursor metabolites and energy | 42 | 0.1% |
| **GO:0005794** | Golgi apparatus | 41 | 0.1% |
| **GO:0003682** | chromatin binding | 35 | 0.1% |
| **GO:0008135** | translation factor activity, nucleic acid binding | 34 | 0.1% |
| **GO:0005576** | extracellular region | 32 | 0.1% |
| **GO:0005216** | ion channel activity | 29 | 0.1% |
| **GO:0004872** | receptor activity | 28 | 0.1% |
| **GO:0016209** | antioxidant activity | 26 | 0.1% |
| **GO:0030246** | carbohydrate binding | 26 | 0.1% |
| **GO:0030528** | transcription regulator activity | 25 | 0.1% |
| **GO:0007049** | cell cycle | 25 | 0.1% |
| **GO:0005654** | nucleoplasm | 23 | 0.0% |
| **GO:0007010** | cytoskeleton organization and biogenesis | 22 | 0.0% |
| **GO:0008092** | cytoskeletal protein binding | 21 | 0.0% |
| **GO:0003779** | actin binding | 17 | 0.0% |
| **GO:0016023** | cytoplasmic membrane-bound vesicle | 16 | 0.0% |
| **GO:0005829** | cytosol | 15 | 0.0% |
| **GO:0005886** | plasma membrane | 14 | 0.0% |
| **GO:0005777** | peroxisome | 14 | 0.0% |
| **GO:0005635** | nuclear membrane | 11 | 0.0% |
| **GO:0040029** | regulation of gene expression, epigenetic | 11 | 0.0% |
| **GO:0030312** | external encapsulating structure | 9 | 0.0% |
| **GO:0007005** | mitochondrion organization and biogenesis | 9 | 0.0% |
| **GO:0005773** | vacuole | 9 | 0.0% |
| **GO:0005618** | cell wall | 7 | 0.0% |
| **GO:0005764** | lysosome | 6 | 0.0% |
| **GO:0030313** | cell envelope | 5 | 0.0% |
| **GO:0009628** | response to abiotic stimulus | 5 | 0.0% |

## Appendix table 4.1: Go Terms

| GO Class ID | Definitions | Counts | Fractions |
|---|---|---|---|
| **GO:0005815** | microtubule organizing center | 5 | 0.0% |
| **GO:0005730** | nucleolus | 4 | 0.0% |
| **GO:0000228** | nuclear chromosome | 4 | 0.0% |
| **GO:0016265** | death | 3 | 0.0% |
| **GO:0008219** | cell death | 3 | 0.0% |
| **GO:0009607** | response to biotic stimulus | 2 | 0.0% |
| **GO:0009653** | morphogenesis | 2 | 0.0% |
| **GO:0008283** | cell proliferation | 2 | 0.0% |
| **GO:0009605** | response to external stimulus | 1 | 0.0% |
| **GO:0019748** | secondary metabolism | 1 | 0.0% |
| **GO:0005615** | extracellular space | 1 | 0.0% |
| **GO:0030154** | cell differentiation | 1 | 0.0% |
| **Total** | | **54,903** | **100.0%** |

# Appendices for Chapter 5

## Appendix table 5.1: Distribution of percentage coverage (Perc cov) of genes

| Perc cov | Cala2 | Emco5 | Emoy2 | Emwa1 | Hind2 | Maks9 | Noco2 | Waco9 |
|---|---|---|---|---|---|---|---|---|
| **100** | 13,883 | 14,078 | 14,199 | 14,120 | 13,853 | 13,993 | 14,176 | 13,939 |
| **99** | 148 | 124 | 72 | 107 | 200 | 177 | 81 | 179 |
| **98** | 95 | 38 | 35 | 45 | 68 | 58 | 37 | 57 |
| **97** | 48 | 38 | 22 | 24 | 46 | 34 | 22 | 37 |
| **96** | 31 | 19 | 20 | 28 | 43 | 27 | 19 | 24 |
| **95** | 35 | 15 | 17 | 15 | 20 | 29 | 16 | 22 |
| **94** | 27 | 17 | 5 | 15 | 18 | 11 | 12 | 22 |
| **93** | 18 | 8 | 9 | 13 | 15 | 9 | 8 | 12 |
| **92** | 14 | 9 | 7 | 7 | 12 | 10 | 5 | 12 |
| **91** | 10 | 8 | 3 | 6 | 7 | 12 | 3 | 6 |
| **90** | 10 | 8 | 7 | 8 | 9 | 10 | 8 | 7 |
| **89** | 7 | 7 | 4 | 5 | 6 | 7 | 5 | 5 |
| **88** | 8 | 4 | 6 | 6 | 7 | 6 | 9 | 6 |
| **87** | 10 | 8 | 4 | 7 | 9 | 7 | 5 | 3 |
| **86** | 8 | 6 | 4 | 6 | 7 | 2 | 3 | 2 |
| **85** | 5 | 5 | 2 | 2 | 4 | 3 | 5 | 6 |
| **84** | 7 | 4 | 4 | 9 | 5 | 4 | 3 | 5 |
| **83** | 12 | 7 | 6 | 5 | 3 | 6 | 7 | 5 |
| **82** | 5 | 3 | 4 | 5 | 3 | 4 | 4 | 2 |
| **81** | 6 | 5 | 3 | 4 | 3 | 5 | 5 | 8 |
| **80** | 5 | 4 | 4 | 3 | 7 | 2 | 4 | 6 |
| **79** | 7 | 3 | 2 | 3 | 4 | 3 | 2 | 4 |
| **78** | 7 | 4 | 3 | 3 | 1 | 2 | 4 | 10 |
| **77** | 4 | 7 | 2 | 3 | 7 | 1 | 1 | 3 |
| **76** | 2 | 4 | 2 | 4 | 4 | 4 | 2 | 0 |
| **75** | 5 | 3 | 2 | 3 | 4 | 6 | 4 | 6 |
| **74** | 3 | 1 | 1 | 3 | 1 | 1 | 4 | 1 |
| **73** | 6 | 0 | 3 | 3 | 5 | 4 | 5 | 6 |
| **72** | 6 | 6 | 1 | 2 | 3 | 0 | 1 | 5 |
| **71** | 4 | 0 | 2 | 2 | 1 | 2 | 2 | 2 |
| **70** | 3 | 3 | 2 | 1 | 3 | 2 | 2 | 2 |
| **69** | 4 | 6 | 1 | 1 | 2 | 5 | 0 | 6 |
| **68** | 4 | 1 | 1 | 1 | 3 | 3 | 7 | 1 |
| **67** | 5 | 4 | 3 | 6 | 2 | 4 | 1 | 5 |
| **66** | 0 | 3 | 2 | 6 | 2 | 2 | 2 | 6 |
| **65** | 2 | 1 | 0 | 2 | 1 | 3 | 3 | 3 |
| **64** | 4 | 4 | 2 | 5 | 2 | 2 | 1 | 2 |
| **63** | 5 | 3 | 2 | 3 | 3 | 7 | 1 | 2 |
| **62** | 7 | 2 | 1 | 2 | 1 | 1 | 0 | 2 |
| **61** | 3 | 0 | 2 | 3 | 0 | 2 | 2 | 3 |
| **60** | 1 | 1 | 1 | 2 | 4 | 2 | 1 | 2 |
| **59** | 1 | 4 | 3 | 2 | 5 | 2 | 1 | 2 |
| **58** | 3 | 2 | 2 | 2 | 3 | 2 | 2 | 1 |
| **57** | 2 | 4 | 4 | 4 | 1 | 3 | 1 | 2 |
| **56** | 0 | 1 | 5 | 2 | 3 | 2 | 2 | 3 |
| **55** | 5 | 1 | 3 | 1 | 1 | 4 | 2 | 1 |
| **54** | 2 | 1 | 2 | 3 | 1 | 4 | 0 | 3 |
| **53** | 2 | 1 | 2 | 2 | 1 | 4 | 0 | 4 |
| **52** | 3 | 2 | 2 | 2 | 2 | 0 | 1 | 2 |
| **51** | 3 | 0 | 1 | 1 | 2 | 4 | 0 | 5 |

| Perc cov | Cala2 | Emco5 | Emoy2 | Emwa1 | Hind2 | Maks9 | Noco2 | Waco9 |
|---|---|---|---|---|---|---|---|---|
| **50** | 2 | 2 | 2 | 1 | 1 | 2 | 1 | 2 |
| **49** | 0 | 0 | 0 | 2 | 2 | 1 | 2 | 1 |
| **48** | 5 | 2 | 1 | 0 | 3 | 2 | 0 | 1 |
| **47** | 1 | 4 | 1 | 1 | 2 | 3 | 2 | 1 |
| **46** | 2 | 3 | 0 | 2 | 1 | 6 | 1 | 3 |
| **45** | 3 | 3 | 0 | 0 | 3 | 1 | 1 | 5 |
| **44** | 0 | 2 | 1 | 1 | 4 | 1 | 2 | 6 |
| **43** | 1 | 2 | 0 | 1 | 2 | 2 | 2 | 2 |
| **42** | 0 | 2 | 4 | 1 | 3 | 2 | 2 | 1 |
| **41** | 2 | 2 | 3 | 0 | 4 | 2 | 2 | 1 |
| **40** | 1 | 3 | 1 | 0 | 3 | 0 | 0 | 1 |
| **39** | 3 | 5 | 4 | 2 | 3 | 3 | 3 | 0 |
| **38** | 3 | 0 | 5 | 3 | 3 | 1 | 2 | 0 |
| **37** | 6 | 3 | 1 | 1 | 3 | 1 | 0 | 1 |
| **36** | 0 | 3 | 2 | 0 | 0 | 3 | 2 | 1 |
| **35** | 0 | 2 | 1 | 1 | 5 | 2 | 0 | 5 |
| **34** | 2 | 1 | 1 | 0 | 2 | 2 | 1 | 2 |
| **33** | 1 | 2 | 5 | 2 | 2 | 2 | 4 | 4 |
| **32** | 2 | 0 | 1 | 2 | 7 | 1 | 1 | 0 |
| **31** | 2 | 1 | 3 | 2 | 6 | 2 | 2 | 6 |
| **30** | 2 | 1 | 2 | 1 | 2 | 1 | 0 | 5 |
| **29** | 5 | 0 | 1 | 0 | 3 | 0 | 2 | 3 |
| **28** | 1 | 2 | 1 | 1 | 2 | 3 | 2 | 4 |
| **27** | 3 | 2 | 2 | 3 | 2 | 1 | 3 | 0 |
| **26** | 2 | 3 | 2 | 0 | 5 | 1 | 1 | 4 |
| **25** | 0 | 2 | 4 | 3 | 4 | 1 | 2 | 2 |
| **24** | 2 | 1 | 4 | 1 | 4 | 5 | 2 | 0 |
| **23** | 4 | 3 | 0 | 2 | 4 | 2 | 2 | 0 |
| **22** | 0 | 6 | 1 | 2 | 3 | 2 | 1 | 1 |
| **21** | 2 | 3 | 2 | 0 | 2 | 2 | 1 | 2 |
| **20** | 5 | 0 | 2 | 0 | 3 | 7 | 3 | 2 |
| **19** | 1 | 1 | 1 | 1 | 2 | 0 | 3 | 2 |
| **18** | 2 | 2 | 1 | 1 | 4 | 0 | 5 | 3 |
| **17** | 0 | 1 | 0 | 0 | 3 | 0 | 0 | 2 |
| **16** | 5 | 2 | 0 | 2 | 7 | 3 | 3 | 5 |
| **15** | 1 | 3 | 2 | 0 | 2 | 0 | 0 | 1 |
| **14** | 0 | 3 | 2 | 0 | 2 | 1 | 0 | 4 |
| **13** | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 5 |
| **12** | 2 | 1 | 2 | 1 | 1 | 2 | 3 | 3 |
| **11** | 1 | 1 | 1 | 0 | 3 | 3 | 3 | 3 |
| **10** | 2 | 1 | 2 | 1 | 5 | 0 | 2 | 4 |
| **9** | 0 | 1 | 0 | 0 | 2 | 3 | 1 | 0 |
| **8** | 2 | 2 | 3 | 3 | 4 | 4 | 3 | 4 |
| **7** | 0 | 1 | 1 | 4 | 2 | 1 | 1 | 3 |
| **6** | 0 | 2 | 2 | 2 | 2 | 2 | 0 | 2 |
| **5** | 1 | 1 | 2 | 4 | 5 | 0 | 5 | 2 |
| **4** | 0 | 0 | 0 | 0 | 5 | 0 | 1 | 2 |
| **3** | 0 | 0 | 0 | 0 | 1 | 1 | 3 | 0 |
| **2** | 2 | 1 | 1 | 0 | 2 | 0 | 0 | 1 |
| **1** | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| **0** | 6 | 6 | 8 | 11 | 18 | 0 | 9 | 11 |
| **Total** | **14,582** | **14,582** | **14,582** | **14,582** | **14,582** | **14,582** | **14,582** | **14,582** |

## Appendix table 5.2: Multi copy genes

The full of percentage coverage, mean coverage and Poisson CDF coverage can be found in attached file: “Appendix table 5.2: Multi copy genes.xlsx”

| Gene | Cala2 | Emco5 | Emoy2 | Emwa1 | Hind2 | Maks9 | Noco2 | Waco9 | Std Dev |
|---|---|---|---|---|---|---|---|---|---|
| *pasa_g28450* | 97.08 | 0 | 99.99 | 100 | 0 | 0 | 99.99 | 0 | 53.06794 |
| *HaRxLL107* | 0.89 | 0 | 99.99 | 92.32 | 99.02 | 0.01 | 99.87 | 0.77 | 52.10976 |
| *805568* | 99.93 | 99.62 | 2.14 | 7.14 | 4.7 | 94.03 | 0.12 | 96.41 | 50.30372 |
| *803972* | 0.48 | 0.01 | 99.95 | 99.99 | 0.48 | 17.07 | 99.97 | 0 | 50.19579 |
| *814281* | 99.87 | 99.99 | 25.14 | 3.19 | 99.97 | 0.61 | 0 | 99.99 | 50.18138 |
| *810011* | 4.95 | 99.99 | 91.21 | 99.98 | 6.41 | 0.81 | 98.07 | 2.9 | 50.10177 |
| *806880* | 4.17 | 0 | 97.58 | 93.05 | 0 | 0 | 99.99 | 0 | 49.76072 |
| *808490* | 9.62 | 0.82 | 99.83 | 99.15 | 2.96 | 6.84 | 99.84 | 1.51 | 49.38215 |
| *806813* | 0.05 | 99.05 | 99.57 | 99.3 | 0.01 | 36.84 | 97.16 | 0 | 49.36307 |
| *800333* | 0.5 | 0 | 99.34 | 92.58 | 1.83 | 0.22 | 82.72 | 95.45 | 49.33684 |

**Appendix table 5.2.11: Table of the genes displaying the most variance in expected Poisson CDF for coverage. Std Dev = standard deviation.**

| Gene | Cala2 | Emco5 | Emoy2 | Emwa1 | Hind2 | Maks9 | Noco2 | Waco9 | Std Dev |
|---|---|---|---|---|---|---|---|---|---|
| *814281* | 99.87 | 99.99 | 25.14 | 3.19 | 99.97 | 0.61 | 0 | 99.99 | 50.18137617 |
| *810011* | 4.95 | 99.99 | 91.21 | 99.98 | 6.41 | 0.81 | 98.07 | 2.9 | 50.10177271 |
| *808693* | 0 | 99.79 | 71.35 | 1.45 | 0 | 99.96 | 88.88 | 0.08 | 48.70677056 |
| *806770* | 99.81 | 99.99 | 5.7 | 0.01 | 0 | 100 | 52.45 | 100 | 48.5910171 |
| *805266* | 9.62 | 2.41 | 93.26 | 99.99 | 1.96 | 1.31 | 93.71 | 0.59 | 47.98286121 |
| *807909* | 0 | 97.69 | 93.41 | 64.27 | 0 | 99.99 | 95.62 | 2.51 | 47.56984968 |
| *801959* | 97.48 | 99.99 | 10.74 | 56.46 | 99.98 | 0 | 99.99 | 0 | 47.52614168 |
| *807217* | 1.84 | 0.61 | 15.15 | 0.08 | 43.11 | 99.58 | 99.59 | 99.99 | 47.40193334 |
| *HaRxL133* | 91.99 | 96.99 | 11.96 | 0.76 | 0.01 | 100 | 28.37 | 99.92 | 47.34252513 |
| *814216* | 83.57 | 99.99 | 75.02 | 0 | 0 | 0 | 99.6 | 88.02 | 46.89017686 |

**Appendix table 5.2.2: Table of the genes displaying the most variance in expected Poisson CDF for coverage, where the gene is expected to be single copy in the reference strain, Emoy2. Std Dev = standard deviation.**

| Gene | Cala2 | Emco5 | Emoy2 | Emwa1 | Hind2 | Maks9 | Noco2 | Waco9 | Std Dev |
|---|---|---|---|---|---|---|---|---|---|
| *809919* | 4.65 | 90.43 | 0.06 | 6.04 | 99.3 | 99.9 | 0.49 | 0.38 | 48.89147134 |
| *812547* | 99.97 | 0 | 0.83 | 0 | 83.67 | 99.7 | 15.33 | 14.11 | 46.40758203 |
| *803007* | 98.49 | 0 | 0.01 | 0.07 | 0 | 0 | 0 | 99.99 | 45.93477674 |
| *804317* | 84.82 | 81.17 | 0.54 | 0.02 | 99.86 | 99.99 | 65.73 | 99.99 | 42.5571619 |
| *804786* | 89.31 | 99.04 | 0.66 | 0 | 80.41 | 44.61 | 0 | 68.95 | 42.52956274 |
| *RXLR87* | 6.64 | 99.99 | 0.74 | 49.24 | 99.99 | 20.18 | 0 | 11.59 | 42.44680433 |
| *814239* | 96.69 | 1.42 | 0.95 | 69.16 | 0.99 | 19.83 | 2.02 | 1.09 | 37.64354486 |
| *RXLR35* | 67.29 | 0 | 0.55 | 0 | 94.38 | 36.77 | 0 | 49.79 | 36.89845516 |
| *809198* | 76.83 | 53.18 | 0.4 | 25 | 21.09 | 98.46 | 2.22 | 10.36 | 36.39590352 |
| *812075* | 74.93 | 83.58 | 0 | 0.02 | 1.78 | 1.76 | 0 | 2.6 | 36.2995899 |

**Appendix table 5.2.3: Table of the genes displaying the most variance in expected Poisson CDF for coverage, where the gene is expected to be hemizygous in the reference strain, Emoy2. Std Dev = standard deviation**

## Appendix table 5.3: SNP gene tables (top 20)

| | Secreted? | Transmemebrane helices? | Effector? | Ave Exon | Std dev Exon | Ave Exon:Intron | Std dev Exon:Intron | Ave Exon:Intron,Up,Down | Std dev Exon:Intron,Up,Down |
|---|---|---|---|---|---|---|---|---|---|
| 808490 | Y | - | - | 33.4 | 14.3 | 33.4 | 14.3 | 22.3 | 13.1 |
| 801090 | - | - | - | 24.4 | 13.4 | 24.4 | 13.4 | 4.5 | 2.9 |
| 808092 | - | - | - | 21.3 | 4.9 | 21.3 | 4.9 | 17.8 | 9.1 |
| 814172 | Y | - | - | 21.0 | 13.7 | 21.0 | 13.7 | 15.7 | 10.9 |
| 802192 | Y | - | - | 20.9 | 15.0 | 20.9 | 15.0 | 7.7 | 6.6 |
| eff_15410_g | Y | 1 | - | 20.9 | 14.4 | 20.9 | 14.4 | 7.4 | 5.8 |
| ceg_12014_g | Y | 3 | - | 20.3 | 16.6 | 5.8 | 4.2 | 2.7 | 2.1 |
| 814613 | - | - | - | 19.5 | 9.6 | 6.0 | 5.4 | 3.3 | 5.2 |
| 814861 | - | - | - | 19.5 | 9.4 | 19.5 | 9.4 | 3.9 | 2.3 |
| 813648 | - | - | - | 18.3 | 19.1 | 18.3 | 19.1 | 13.6 | 17.2 |
| 813379 | - | - | - | 18.0 | 6.9 | 18.0 | 6.9 | 15.4 | 9.3 |
| 803035 | Y | - | - | 17.9 | 15.1 | 17.9 | 15.1 | 12.4 | 12.3 |
| 807859 | - | - | - | 17.1 | 11.4 | 11.6 | 9.7 | 4.9 | 5.1 |
| 808876 | - | - | - | 16.9 | 4.1 | 16.9 | 4.1 | 13.7 | 7.0 |
| ATR1_Emoy2 | Y | - | ATR1 | 16.8 | 21.5 | 16.8 | 21.5 | 6.1 | 6.5 |
| 803927 | - | - | - | 16.4 | 5.4 | 8.5 | 2.2 | 4.5 | 3.4 |
| 800520 | - | - | - | 16.3 | 6.3 | 16.3 | 6.3 | 15.5 | 7.3 |
| 807858 | - | - | - | 16.3 | 10.4 | 5.2 | 4.0 | 3.0 | 2.2 |
| 808811 | Y | - | - | 16.3 | 12.9 | 16.3 | 12.9 | 4.0 | 4.4 |
| 804775 | - | - | - | 16.1 | 6.0 | 16.1 | 6.0 | 6.0 | 4.5 |

Full list of genes can be found in appendix file: “Appendix table 5.3, 5.4, 5.5 - SNPs.xlsx” under SNP tab.

## Appendix table 5.4: Heterozygous SNP gene table (top 20)

| | Secreted? | Transmemebrane helices? | Effector? | Ave Exon | Std dev Exon | Ave Exon:Intron | Std dev Exon:Intron | Ave Exon:Intron,Up,Down | Std dev Exon:Intron,Up,Down |
|---|---|---|---|---|---|---|---|---|---|
| 801090 | - | - | - | 23.1 | 14.1 | 23.1 | 14.1 | 4.8 | 4.5 |
| 802192 | Y | - | - | 20.0 | 16.0 | 20.0 | 16.0 | 7.0 | 6.9 |
| 814172 | Y | - | - | 17.5 | 11.1 | 17.5 | 11.1 | 15.0 | 11.8 |
| eff_15410_g | Y | 1 | - | 16.4 | 12.5 | 16.4 | 12.5 | 6.3 | 6.1 |
| 814613 | - | - | - | 16.4 | 12.0 | 4.0 | 4.8 | 1.1 | 0.5 |
| 814861 | - | - | - | 14.8 | 8.0 | 14.8 | 8.0 | 3.0 | 2.0 |
| 806792 | - | - | - | 14.5 | 10.4 | 14.5 | 10.4 | 9.4 | 8.8 |
| 814802 | - | - | - | 14.1 | 14.5 | 14.1 | 14.5 | 9.4 | 10.5 |
| 813648 | - | - | - | 13.8 | 11.9 | 13.8 | 11.9 | 9.4 | 8.5 |
| 808811 | Y | - | - | 12.6 | 9.8 | 12.6 | 9.8 | 4.8 | 7.1 |
| 808716 | - | - | - | 12.0 | 10.6 | 1.0 | 0.7 | 0.5 | 0.5 |
| 811880 | Y | - | - | 10.5 | 6.9 | 10.5 | 6.9 | 7.0 | 6.1 |
| 813542 | - | - | - | 10.3 | 7.8 | 10.3 | 7.8 | 5.1 | 4.6 |
| 810634 | Y | - | - | 9.8 | 7.6 | 9.8 | 7.6 | 3.5 | 3.7 |
| PHYT9337.2 | - | - | - | 9.6 | 5.0 | 6.5 | 4.9 | 5.0 | 4.7 |
| 807678 | - | - | - | 9.1 | 6.8 | 9.1 | 6.8 | 7.5 | 6.8 |
| 812044 | Y | - | - | 8.9 | 15.2 | 8.9 | 15.2 | 8.5 | 14.4 |
| PHYT4874.8 | - | - | - | 8.9 | 9.6 | 3.0 | 2.9 | 2.5 | 2.6 |
| 808490 | Y | - | - | 8.8 | 16.0 | 8.8 | 16.0 | 8.6 | 15.0 |
| 811478 | Y | - | HaRxLL36 | 8.8 | 5.9 | 8.8 | 5.9 | 4.8 | 4.9 |

Full list of genes can be found in appendix file: “Appendix table 5.3, 5.4, 5.5 - SNPs.xlsx” under Hets tab.

## Appendix table 5.5: Protein coding effect of SNPs (top 20)

| | Secreted? | Transmemebrane helices? | Effector? | Average Syn SNPs | SD | Average het non-syn SNP | SD | Average non-syn SNPs | SD | Count mutated start | Count early stop |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 808490 | Y | - | - | 0.4 | 1.1 | 8.5 | 15.5 | 33.0 | 15.2 | 0.0 | 0.0 |
| 802192 | Y | - | - | 0.1 | 0.4 | 9.0 | 7.5 | 20.8 | 15.1 | 0.0 | 0.0 |
| 808092 | - | - | - | 0.9 | 1.0 | 1.5 | 2.8 | 20.4 | 5.1 | 0.0 | 0.0 |
| 814172 | Y | - | - | 1.6 | 1.5 | 16.3 | 10.2 | 19.4 | 13.4 | 0.0 | 0.0 |
| 814861 | - | - | - | 0.6 | 0.5 | 12.9 | 6.9 | 18.9 | 9.2 | 0.0 | 0.0 |
| 803035 | Y | - | - | 0.1 | 0.4 | 0.6 | 1.4 | 17.8 | 15.1 | 0.0 | 0.0 |
| eff_15410_g | Y | 1 | - | 3.3 | 2.0 | 12.9 | 10.1 | 17.6 | 13.3 | 0.0 | 0.0 |
| 813379 | - | - | - | 0.5 | 0.5 | 3.1 | 4.9 | 17.5 | 6.5 | 0.0 | 0.0 |
| ceg_12014_g | Y | 3 | - | 2.9 | 2.0 | 0.1 | 0.4 | 17.4 | 15.6 | 0.0 | 4.0 |
| 807859 | - | - | - | 0.4 | 0.7 | 0.0 | 0.0 | 16.8 | 11.5 | 0.0 | 0.0 |
| ATR1_Emoy2 | Y | - | ATR1 | 0.0 | 0.0 | 0.0 | 0.0 | 16.8 | 21.5 | 0.0 | 0.0 |
| 803927 | - | - | - | 0.4 | 0.5 | 1.8 | 4.9 | 16.0 | 5.6 | 0.0 | 0.0 |
| 801090 | - | - | - | 8.8 | 3.5 | 15.0 | 9.8 | 15.6 | 10.1 | 0.0 | 0.0 |
| 800520 | - | - | - | 0.6 | 0.5 | 0.3 | 0.5 | 15.6 | 5.8 | 0.0 | 0.0 |
| 814613 | - | - | - | 4.1 | 1.7 | 13.0 | 9.2 | 15.4 | 8.7 | 2.0 | 0.0 |
| 814292 | - | - | - | 0.3 | 0.5 | 1.3 | 2.3 | 15.0 | 6.6 | 0.0 | 0.0 |
| 804775 | - | - | - | 1.3 | 0.7 | 1.0 | 2.1 | 14.9 | 5.7 | 0.0 | 0.0 |
| 811796 | - | - | - | 0.0 | 0.0 | 0.0 | 0.0 | 14.9 | 10.7 | 0.0 | 0.0 |
| 809040 | - | - | - | 0.4 | 1.1 | 4.8 | 5.6 | 14.6 | 6.7 | 0.0 | 0.0 |
| 807858 | - | - | - | 1.6 | 1.2 | 0.1 | 0.4 | 14.6 | 9.4 | 0.0 | 0.0 |

Full list of genes can be found in appendix file: “Appendix table 5.3, 5.4, 5.5 - SNPs.xlsx” under Syn tab.

## Appendix table 5.6: INDEL gene tables (top 20)

| | SP | TM | EF | Ave Exon | Std dev Exon | Ave Exon:Intron | Std dev Exon:Intron | Ave Exon:Intron,Up,Down | Std dev Exon:Intron,Up,Down |
|---|---|---|---|---|---|---|---|---|---|
| **801090** | - | - | - | 4.9 | 1.6 | 4.9 | 1.6 | 3.5 | 1.6 |
| **803674** | - | - | - | 4.6 | 3.2 | 0.7 | 0.3 | 0.7 | 0.3 |
| **814172** | Y | - | - | 4.4 | 0.9 | 4.4 | 0.9 | 4.1 | 1.2 |
| **811880** | Y | - | - | 3.9 | 1.2 | 3.9 | 1.2 | 3.9 | 1.2 |
| **807247** | - | - | - | 3.3 | 1.7 | 3.3 | 1.7 | 3.3 | 1.6 |
| **804903** | Y | - | - | 3.0 | 1.8 | 3.0 | 1.8 | 3.0 | 1.7 |
| **812377** | - | - | - | 3.0 | 1.3 | 2.2 | 1.3 | 2.2 | 1.2 |
| **PHYT2811.3** | Y | - | - | 2.9 | 2.2 | 1.4 | 0.7 | 1.4 | 0.7 |
| **PHYT9337.2** | - | - | - | 2.9 | 1.5 | 1.6 | 1.0 | 0.9 | 0.4 |
| **ceg_3124_g** | Y | - | - | 2.8 | 3.5 | 2.8 | 3.5 | 2.8 | 3.3 |
| **813447** | Y | 1 | - | 2.6 | 2.0 | 2.6 | 2.0 | 2.4 | 1.8 |
| **809897** | - | - | - | 2.5 | 5.2 | 2.5 | 5.2 | 2.5 | 4.8 |
| **806792** | - | - | - | 2.5 | 1.6 | 2.5 | 1.6 | 2.4 | 1.6 |
| **808490** | Y | - | - | 2.5 | 1.5 | 2.5 | 1.5 | 2.0 | 1.1 |
| **808716** | - | - | - | 2.5 | 1.2 | 2.2 | 0.7 | 2.0 | 0.7 |
| **PHYT2459.8** | - | - | - | 2.5 | 0.8 | 2.5 | 0.8 | 2.5 | 0.7 |
| **801705** | - | - | - | 2.5 | 0.5 | 2.5 | 0.5 | 2.3 | 0.6 |
| **809705** | - | - | - | 2.5 | 0.9 | 2.5 | 0.9 | 0.9 | 0.5 |
| **802220** | - | - | - | 2.5 | 0.9 | 2.5 | 0.9 | 1.0 | 0.4 |
| **PHYT4874.8** | - | - | - | 2.4 | 1.8 | 2.4 | 1.8 | 2.0 | 1.1 |

A full list of genes can be found in attached file: “Appendix table 5.6, 57 - INDELs.xlsx” under INDELs tab.

## Appendix table 5.7: Heterozygous INDEL gene tables (top 20)

| | SP | TM | EF | Ave Exon | Std dev Exon | Ave Exon:Intron | Std dev Exon:Intron | Ave Exon:Intron,Up,Down | Std dev Exon:Intron,Up,Down |
|---|---|---|---|---|---|---|---|---|---|
| **801090** | - | - | - | 4.9 | 1.6 | 4.9 | 1.6 | 2.8 | 1.6 |
| **803674** | - | - | - | 4.6 | 3.2 | 0.7 | 0.3 | 0.6 | 0.3 |
| **814172** | Y | - | - | 4.3 | 1.2 | 4.3 | 1.2 | 3.4 | 1.8 |
| **811880** | Y | - | - | 3.5 | 1.4 | 3.5 | 1.4 | 3.1 | 1.8 |
| **807247** | - | - | - | 3.1 | 1.4 | 3.1 | 1.4 | 2.9 | 1.6 |
| **804903** | Y | - | - | 3.0 | 1.8 | 3.0 | 1.8 | 2.6 | 1.9 |
| **PHYT9337.2** | - | - | - | 2.9 | 1.5 | 1.9 | 0.9 | 1.2 | 0.9 |
| **813447** | Y | 1 | - | 2.6 | 2.0 | 2.6 | 2.0 | 1.9 | 1.8 |
| **809897** | - | - | - | 2.5 | 5.2 | 2.5 | 5.2 | 2.5 | 4.8 |
| **806792** | - | - | - | 2.5 | 1.6 | 2.5 | 1.6 | 2.1 | 1.8 |
| **809705** | - | - | - | 2.5 | 0.9 | 2.5 | 0.9 | 0.7 | 0.5 |
| **802220** | - | - | - | 2.5 | 0.9 | 2.5 | 0.9 | 0.9 | 0.4 |
| **pasa_gi_SuperContig10_291** | - | - | - | 2.4 | 0.7 | 2.4 | 0.7 | 2.1 | 1.0 |
| **HaRxLCRN4** | Y | - | - | 2.3 | 1.8 | 2.3 | 1.8 | 2.1 | 1.8 |
| **805490** | - | - | - | 2.3 | 1.4 | 2.1 | 1.4 | 1.8 | 1.5 |
| **808811** | Y | - | - | 2.3 | 1.6 | 2.3 | 1.6 | 1.7 | 1.2 |
| **812153** | - | - | - | 2.3 | 1.4 | 2.3 | 1.4 | 1.9 | 1.1 |
| **807750** | - | - | - | 2.1 | 4.5 | 2.1 | 4.5 | 2.2 | 4.2 |
| **810634** | Y | - | - | 2.1 | 1.4 | 2.1 | 1.4 | 1.8 | 1.4 |
| **808716** | - | - | - | 2.1 | 1.4 | 1.8 | 0.8 | 1.5 | 0.8 |

A full list of genes can be found in attached file: “Appendix table 5.6, 57 - INDELs.xlsx” under Hets tab.

## Appendix table 5.8: DnaSP tables (top and bottom 20 Fu's Fs)

| Gene | Secreted? | Transmembrane? | Effector? | n | S | Eta | Hap | Hap / N | TajimaD | FuLiD* | FuLiF* | FuFs |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ceg_g20448 | - | - | - | 16 | 33 | 41 | 4 | 0.3 | 1.0 | 0.5 | 0.8 | 12.0 |
| eff_g3498 | Y | - | HaRxL21 | 16 | 21 | 21 | 3 | 0.2 | 2.6 | 1.6 | 2.1 | 11.8 |
| 807780 | - | - | - | 16 | 12 | 12 | 2 | 0.1 | 2.5 | 1.5 | 2.0 | 10.9 |
| 808682 | - | - | - | 16 | 18 | 18 | 3 | 0.2 | 2.4 | 1.5 | 2.1 | 10.4 |
| 803642 | - | - | - | 16 | 19 | 19 | 3 | 0.2 | 2.0 | 1.6 | 1.9 | 10.3 |
| 808594 | Y | - | HaRxLL38 | 14 | 36 | 36 | 4 | 0.3 | 1.2 | 1.6 | 1.7 | 10.3 |
| 807859 | - | - | - | 16 | 43 | 43 | 5 | 0.3 | 1.3 | 1.7 | 1.8 | 10.2 |
| 814245 | - | - | - | 16 | 13 | 13 | 2 | 0.1 | 1.3 | 1.5 | 1.6 | 9.9 |
| 801132 | Y | - | - | 16 | 56 | 56 | 6 | 0.4 | 1.2 | 1.7 | 1.8 | 9.6 |
| eff_g7948 | Y | - | - | 16 | 56 | 56 | 6 | 0.4 | 1.2 | 1.7 | 1.8 | 9.6 |
| 800248 | - | 1 | - | 16 | 24 | 24 | 4 | 0.3 | 2.2 | 1.6 | 2.0 | 9.5 |
| 801226 | - | - | - | 16 | 9 | 9 | 2 | 0.1 | 2.8 | 1.4 | 2.0 | 9.4 |
| 800673 | Y | - | - | 16 | 35 | 39 | 4 | 0.3 | -0.3 | 1.7 | 1.3 | 9.4 |
| 809692 | - | - | - | 16 | 32 | 32 | 4 | 0.3 | 0.5 | 1.6 | 1.5 | 9.3 |
| 811507 | - | - | - | 16 | 15 | 15 | 3 | 0.2 | 2.5 | 1.5 | 2.1 | 9.2 |
| 807947 | - | - | - | 16 | 16 | 16 | 3 | 0.2 | 2.1 | 1.5 | 1.9 | 9.2 |
| 807060 | - | - | - | 16 | 14 | 14 | 3 | 0.2 | 2.7 | 1.5 | 2.1 | 9.0 |
| 807061 | - | - | - | 16 | 14 | 14 | 3 | 0.2 | 2.7 | 1.5 | 2.1 | 9.0 |
| ATR1_Emoy2 | Y | - | ATR1 | 12 | 62 | 64 | 5 | 0.4 | 0.2 | 1.7 | 1.5 | 8.8 |
| 805211 | - | - | - | 16 | 22 | 26 | 4 | 0.3 | 1.1 | 0.6 | 0.9 | 8.8 |
| … | … | … | … | … | … | … | …. | … | … | … | … | … |
| 804837 | - | - | - | 16 | 4 | 5 | 8 | 0.5 | -0.2 | 0.5 | 0.3 | -4.2 |
| 804910 | - | - | - | 16 | 10 | 10 | 10 | 0.6 | -0.4 | 0.5 | 0.3 | -4.3 |
| ceg_9280_g | - | - | - | 16 | 4 | 6 | 9 | 0.6 | 0.2 | 0.0 | 0.0 | -4.4 |
| PHYT5312.14 | - | - | - | 16 | 16 | 16 | 12 | 0.8 | -0.1 | 0.0 | 0.0 | -4.4 |
| ceg_14750_g | Y | 13 | Emoy2_HpRXLR104 | 16 | 15 | 15 | 12 | 0.8 | 0.0 | -0.1 | 0.0 | -4.7 |
| 804903 | Y | - | - | 16 | 15 | 17 | 11 | 0.7 | -1.4 | -0.4 | -0.8 | -4.8 |
| eff_g19502 | Y | - | - | 16 | 15 | 17 | 11 | 0.7 | -1.4 | -0.4 | -0.8 | -4.8 |
| eff_g7027 | Y | - | - | 16 | 15 | 17 | 11 | 0.7 | -1.4 | -0.4 | -0.8 | -4.8 |
| 808661 | - | - | - | 14 | 3 | 5 | 8 | 0.6 | -0.6 | -0.2 | -0.4 | -5.2 |
| 801127 | - | - | - | 16 | 3 | 5 | 8 | 0.5 | -0.8 | 0.5 | 0.1 | -5.2 |
| pasa_gi_SuperContig2_159 | - | - | - | 16 | 9 | 9 | 11 | 0.7 | 0.3 | 0.9 | 0.9 | -5.3 |
| 806421 | - | - | - | 16 | 24 | 24 | 14 | 0.9 | 0.1 | 1.0 | 0.8 | -5.5 |
| pasa_883_g | - | - | - | 16 | 15 | 17 | 13 | 0.8 | 0.0 | 0.4 | 0.4 | -5.6 |
| 802816 | - | 8 | - | 16 | 7 | 8 | 11 | 0.7 | 0.5 | 0.8 | 0.8 | -5.7 |
| HaRxLL15 | - | - | - | 16 | 13 | 14 | 12 | 0.8 | -0.7 | -0.2 | -0.4 | -6.2 |
| 809011 | - | - | - | 16 | 7 | 8 | 11 | 0.7 | 0.1 | 0.8 | 0.7 | -6.3 |
| HaRxLL133 | - | - | - | 16 | 15 | 16 | 13 | 0.8 | -0.4 | 0.0 | -0.1 | -6.6 |
| 811403 | - | - | - | 16 | 17 | 20 | 13 | 0.8 | -1.4 | -1.6 | -1.8 | -7.2 |
| 812038 | - | - | HaRxLL163 | 16 | 7 | 9 | 12 | 0.8 | -0.1 | 0.4 | 0.3 | -7.8 |
| 808717 | - | - | - | 16 | 7 | 9 | 13 | 0.8 | 0.4 | 0.9 | 0.9 | -9.1 |

A full list of genes can be found in attached file: "Appendix table 5.8 - DnaSP tables.xlsx" under the All tab.

## Appendix table 5.9: PAML tables (top 20 codeml dN/dS)

| | Secreted? | Transmembrane? | Effector? | **yn00 Dn/Ds** | **codeml Dn/Ds** | **M3-M0** | **M2a-M1a** | **M7-M8** | **M8-M8a** |
|---|---|---|---|---|---|---|---|---|---|
| ceg_13464_g | - | - | - | 0.00 | 97.01 | 0.00 | 0.75 | 0.75 | 0.75 |
| 810037 | - | - | - | 0.00 | 50.50 | 0.00 | 0.00 | 0.00 | 0.00 |
| 810267 | - | - | - | 0.00 | 48.42 | 0.00 | 0.00 | 0.00 | 0.00 |
| 811885 | - | - | - | 0.00 | 46.40 | 0.00 | 0.00 | 0.00 | 0.00 |
| 801186 | - | - | - | 0.00 | 36.19 | 0.00 | 1.27 | 1.27 | 1.27 |
| 807911 | Y | - | HaRxLL108 | 0.00 | 34.95 | 0.00 | 1.23 | 1.23 | 1.23 |
| 811640 | - | - | - | 0.00 | 34.33 | 13.62 | 18.41 | 18.41 | 18.41 |
| ceg_15480_g | - | - | - | 0.00 | 32.26 | 0.00 | 0.00 | 0.00 | 0.00 |
| 802815 | - | - | - | 0.00 | 31.95 | 0.00 | 0.00 | 0.00 | 0.00 |
| 808320 | - | 1 | - | 0.00 | 28.05 | 0.00 | 0.00 | 0.00 | 0.00 |
| 800545 | - | - | - | 0.00 | 24.99 | 0.00 | 0.00 | 0.00 | 0.00 |
| pasa_gi_SuperContig157_18 | - | - | - | 0.00 | 24.78 | 0.00 | 1.33 | 1.33 | 1.33 |
| 811033 | - | - | - | 0.00 | 24.25 | 0.00 | 4.47 | 4.47 | 4.47 |
| 812953 | - | - | - | 0.00 | 23.81 | 0.00 | 1.82 | 1.82 | 1.83 |
| 805928 | - | - | - | 0.00 | 23.57 | 0.00 | 0.00 | 0.00 | 0.00 |
| 807000 | - | - | - | 0.00 | 23.37 | 0.00 | 0.00 | 0.00 | 0.00 |
| HaRxLL447 | Y | - | HaRxLL447 | 0.00 | 23.34 | 0.00 | 0.00 | 0.00 | 0.00 |
| 806586 | - | 3 | - | 0.00 | 23.08 | 0.00 | 0.00 | 0.00 | 0.00 |
| 805043 | - | - | - | 0.00 | 22.61 | 0.00 | 0.00 | 0.00 | 0.00 |
| 804639 | - | - | - | 0.00 | 22.48 | 0.00 | 0.76 | 0.76 | 0.76 |

A full list of genes can be found in attached file: “Appendix table 5.9 - PAML tables.xlsx” under the All tab.

# Abbreviations

*A. laibachii - Albugo laibachii*
*A. thaliana - Arabidopsis thaliana*
ADP - adenosine diphosphate
ARF - ADP Ribosylation Factors
BAC - bacterial artificial chromosome
bp - base pair
*C. elegans - Caenorhabditis elegans*
CDF - cumulative distribution function
cDNA - complementary deoxyribonucleic acid
CEG - core eukaryotic genes present at single or low copy numbers
CEGMA - Core Eukaryotic Genes Mapping Approach
CHXC - CHXC motif
CNV - copy number variation
D. melanogaster - Drosophila melanogaster
d.p.i - days post inoculation
DEPC - diethylpyrocarbonate
DNA - deoxyribonucleic acid
EST - expressed sequence tags
ETI - effector-triggered immunity.
ETS - effector-triggered susceptibility
flg22 - 22-amino acid sequence of the conserved N-terminal part of flagellin is known to activate plant defence mechanisms
FLS2 - FLAGELLIN SENSITIVE 2
GA - Genome Analyser
GFF3 - general feature format 3
GO - Gene Ontology
*H. sapiens - Homo sapiens*
HMM - hidden Markov model
*Hpa - Hyaloperonospora arabidopsidis*
HRI - Horticulture Research International
INDEL - insertion and/or deletion
IPTG - isopropyl β-D-1-thiogalactopyranoside
IUPAC - International Union of Pure and Applied Chemistry
KOG - core eukaryotic protein
MCMC - Monte Carlo Markov chain
MRCA - most recent common ancestor
NADH - nicotinamide adenine dinucleotide
NCBI - National Centre for Biotechnology Information
NR - non redundant
nt - nucleotide
*P. infestans - Phytophthora infestans*

*P. ramorum - Phytophthora ramorum*
*P. sojae - Phytophthora sojae*
PAMP - pathogen associated molecular patterns
PCEG - conserved single copy Phytophthora genes
PCR - polymerase chain reaction
PTI - PAMP triggered immunity
QQ-plots - quantile-quantile-plots
RNA - ribonucleic acid
RXLR - RXLR motif
*S. cerevisiae - Saccharomyces cerevisiae*
*S. pombe - Schizosaccharomyces pombe*
SNP - single nucleotide polymorphism
TAIR - The Arabidopsis Information Resource
TSL - The Sainsbury Laboratory
UTR - untranslated region
VBI - Virginia Bioinformatics Institute
VCF - variant call format

# References


Albers, C.A., Lunter, G., MacArthur, D.G., McVean, G., Ouwehand, W.H., and Durbin, R. (2011). Dindel: accurate indel calls from short-read data. Genome Research *21*, 961-973.

Allen, R.L., Bittner-Eddy, P.D., Grenvitte-Briggs, L.J., Meitz, J.C., Rehmany, A.P., Rose, L.E., and Beynon, J.L. (2004). Host-parasite coevolutionary conflict between Arabidopsis and downy mildew. Science *306*, 1957-1960.

Allen, R.L., Meitz, J.C., Baumber, R.E., Hall, S.A., Lee, S.C., Rose, L.E., and Beynon, J.L. (2008). Natural variation reveals key amino acids in a downy mildew effector that alters recognition specificity by an Arabidopsis resistance gene. Molecular Plant Pathology *9*, 511-523.

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

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research *25*, 3389-3402.

Animal and Plant Health Inspection Service (2002). Agricultural Bioterrorism Protection Act, D.o. Agriculture, ed., pp. 52383-52389.

Anisimova, M., Bielawski, J.P., and Yang, Z. (2001). Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. Molecular Biology and Evolution *18*, 1585-1592.

Anisimova, M., Bielawski, J.P., and Yang, Z. (2002). Accuracy and power of bayes prediction of amino acid sites under positive selection. Molecular Biology and Evolution *19*, 950-958.

Anisimova, M., Nielsen, R., and Yang, Z. (2003). Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. Genetics *164*, 1229-1236.

Armstrong, M.R., Whisson, S.C., Pritchard, L., Bos, J.I.B., Venter, E., Avrova, A.O., Rehmany, A.P., Böhme, U., Brooks, K., Cherevach, I.*, et al.* (2005). An ancestral oomycete locus contains late blight avirulence gene Avr3a, encoding a protein that is recognized in the host cytoplasm. Proceedings of the National Academy of Sciences of the United States of America *102*, 7766-7771.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T.*, et al.* (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature Genetics *25*, 25-29.

Attwood, T.K., Bradley, P., Flower, D.R., Gaulton, A., Maudling, N., Mitchell, A.L., Moulton, G., Nordle, A., Paine, K., Taylor, P.*, et al.* (2003). PRINTS and its automatic supplement, prePRINTS. Nucleic Acids Research *31*, 400-402.

Bailey, K., Cevik, V., Holton, N., Byrne-Richardson, J., Sohn, K.H., Coates, M., Woods-Toer, A., Aksoy, H.M., Hughes, L., Baxter, L.*, et al.* (2011). Molecular Cloning of ATR5(Emoy2) from Hyaloperonospora arabidopsidis, an Avirulence Determinant That Triggers RPP5-Mediated Defense in Arabidopsis. Molecular Plant-Microbe Interactions *24*, 827-838.

Bary, H.A.d. (1879). Die Erscheinung der Symbiose. Vortrag gehalten auf der Versammlung Deutscher Naturforscher und Aerzte zu Cassel (K. J. Trübner).

Bashiardes, S., Veile, R., Helms, C., Mardis, E.R., Bowcock, A.M., and Lovett, M. (2005). Direct genomic selection. Nature Methods *2*, 63-69.

Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E.L.*, et al.* (2004). The Pfam protein families database. Nucleic Acids Research *32*, D138-141.

Baxter, L., Tripathy, S., Ishaque, N., Boot, N., Cabral, A., Kemen, E., Thines, M., Ah-Fong, A., Anderson, R., Badejoko, W.*, et al.* (2010). Signatures of adaptation to obligate biotrophy in the Hyaloperonospora arabidopsidis genome. Science *330*, 1549-1551.

Bendtsen, J.D., Nielsen, H., von Heijne, G., and Brunak, S. (2004). Improved Prediction of Signal Peptides: SignalP 3.0. Journal of Molecular Biology *340*, 783-795.

Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, K.P., Evers, D.J., Barnes, C.L., Bignell, H.R.*, et al.* (2008). Accurate whole human genome sequencing using reversible terminator chemistry. Nature *456*, 53-59.

Birch, P.R., Rehmany, A.P., Pritchard, L., Kamoun, S., and Beynon, J.L. (2006). Trafficking arms: oomycete effectors enter host plant cells. Trends in Microbiology *14*, 8-11.

Birney, E., Clamp, M., and Durbin, R. (2004). GeneWise and Genomewise. Genome Research *14*, 988-995.

Bru, C., Courcelle, E., Carrere, S., Beausse, Y., Dalmar, S., and Kahn, D. (2005). The ProDom database of protein domain families: more emphasis on 3D. Nucleic Acids Research *33*, D212-215.

Boutemy, L. S., S. R. King, et al. (2011). "Structures of Phytophthora RXLR effector proteins: a conserved but adaptable fold underpins functional diversity." J Biol Chem 286(41): 35834-35842.

Burrows, M., and Wheeler, D.J. (1994). A Block-sorting Lossless Data Compression Algorithm. In Technical Report 124, D.E. Corporation, ed.

Butler, J., MacCallum, I., Kleber, M., Shlyakhter, I.A., Belmonte, M.K., Lander, E.S., Nusbaum, C., and Jaffe, D.B. (2008). ALLPATHS: de novo assembly of whole-genome shotgun microreads. Genome Research *18*, 810-820.

Cao, J., Schneeberger, K., Ossowski, S., Gunther, T., Bender, S., Fitz, J., Koenig, D., Lanz, C., Stegle, O., Lippert, C.*, et al.* (2011). Whole-genome sequencing of multiple Arabidopsis thaliana populations. Nature Genetics *43*, 956-963.

Chaisson, M.J., and Pevzner, P.A. (2008). Short read fragment assembly of bacterial genomes. Genome Research *18*, 324-330.

Chen, F., Mackey, A.J., Stoeckert, C.J., and Roos, D.S. (2005). OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. Nucleic Acids Research *34*, D363-D368.

Chen, K., Wallis, J.W., McLellan, M.D., Larson, D.E., Kalicki, J.M., Pohl, C.S., McGrath, S.D., Wendl, M.C., Zhang, Q., Locke, D.P.*, et al.* (2009). BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nature Methods *6*, 677-681.

Chevreux, B., Pfisterer, T., Drescher, B., Driesel, A.J., Muller, W.E., Wetter, T., and Suhai, S. (2004). Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. Genome Research *14*, 1147-1159.

Chinchilla, D., Bauer, Z., Regenass, M., Boller, T., and Felix, G. (2006). The Arabidopsis receptor kinase FLS2 binds flg22 and determines the specificity of flagellin perception. Plant Cell *18*, 465-476.

Cingolani, P. (2011). snpEff: Variant effect prediction http://snpeff.sourceforge.net.

Coghlan, A., Fiedler, T.J., McKay, S.J., Flicek, P., Harris, T.W., Blasiar, D., and Stein, L.D. (2008). nGASP--the nematode genome annotation assessment project. BMC Bioinformatics *9*, 549.

Cox, M.P., Peterson, D.A., and Biggs, P.J. (2010). SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. BMC Bioinformatics *11*, 485.

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

Darling, A.E., Mau, B., and Perna, N.T. (2010). progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS One *5*, e11147.

Darwin, C. (1859). On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life, 1st edn (London, John Murray).

Dodds, P.N., and Rathjen, J.P. (2010). Plant immunity: towards an integrated view of plant-pathogen interactions. Nature Reviews Genetics *11*, 539-548.

Douglas, A. (2010). The Symbiotic Habit (Princeton University Press).

Duplessis, S., Cuomo, C.A., Lin, Y.C., Aerts, A., Tisserant, E., Veneault-Fourrey, C., Joly, D.L., Hacquard, S., Amselem, J., Cantarel, B.L.*, et al.* (2011). Obligate biotrophy features unraveled by the genomic analysis of rust fungi. Proceedings of the National Academy of Sciences *108*, 9166-9171.

Epp, E., Vanier, G., Harcus, D., Lee, A.Y., Jansen, G., Hallett, M., Sheppard, D.C., Thomas, D.Y., Munro, C.A., Mullick, A.*, et al.* (2010). Reverse Genetics in Candida albicans Predicts ARF Cycling Is Essential for Drug Resistance and Virulence. PLoS Pathology *6*, e1000753.

Farrer, R.A., Kemen, E., Jones, J.D., and Studholme, D.J. (2009). De novo assembly of the Pseudomonas syringae pv. syringae B728a genome using Illumina/Solexa short sequence reads. FEMS Microbiology Letters *291*, 103-111.

Fay, J.C. (2011). Weighing the evidence for adaptation at the molecular level Trends in Genetics

Felsenstein, J. (1989). PHYLIP—phylogeny inference package. Cladistics *5*, 164-166.

Finn, R.D., Clements, J., and Eddy, S.R. (2011). HMMER web server: interactive sequence similarity searching. Nucleic Acids Research *39*, W29-37.

Freese, E. (1959). The Difference between Spontaneous and Base-Analogue Induced Mutations of Phage T4. Proceedings of the National Academy of Sciences *45*, 622-633.

Fu, Y.X. (1997). Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. Genetics *147*, 915-925.

Fu, Y.X., and Li, W.H. (1993). Statistical tests of neutrality of mutations. Genetics *133*, 693-709.

Garbeva, P., Silby, M.W., Raaijmakers, J.M., Levy, S.B., and Boer, W. (2011). Transcriptional and antagonistic responses of Pseudomonas fluorescens Pf0-1 to phylogenetically different bacterial competitors. International Society for Microbial Ecology Journal *5*, 973-985.

Goker, M., Riethmuller, A., Voglmayr, H., Weiss, M., and Oberwinkler, F. (2004). Phylogeny of Hyaloperonospora based on nuclear ribosomal internal transcribed spacer sequences. Mycological Progress *3*, 83-94.

Goldman, N., and Yang, Z. (1994). A codon-based model of nucleotide substitution for protein-coding DNA sequences. Molecular Biology and Evolution *11*, 725-736.

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

Guigo, R. (1998). Assembling genes from predicted exons in linear time with dynamic programming. Journal of Computational Biology *5*, 681-702.

Haas, B.J. (2010). TransposonPSI: An Application of PSI-Blast to Mine (Retro-)Transposon ORF Homologies http://transposonpsi.sourceforge.net/, B. Institute, ed.

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

Haas, B.J., Kamoun, S., Zody, M.C., Jiang, R.H., Handsaker, R.E., Cano, L.M., Grabherr, M., Kodira, C.D., Raffaele, S., Torto-Alalibo, T*., et al.* (2009). Genome sequence and analysis of the Irish potato famine pathogen Phytophthora infestans. Nature *461*, 393-398.

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

Hardy, G.H. (1908). Mendelian Proportions in a Mixed Population. Science *28*, 49-50.

Horton, H.R., Moran, L.A., Scrimgeour, K.G., Perry, M.D., and Rawn, J.D. (2006). Principles of Biochemistry 4th edition edn (Pearson Prentice Hall).

Horton, P., Park, K.-J., Obayashi, T., Fujita, N., Harada, H., Adams-Collier, C.J., and Nakai, K. (2007). WoLF PSORT: Protein Localization Predictor. Nucleic Acids Research.

Huang, X., Wang, J., Aluru, S., Yang, S.P., and Hillier, L. (2003). PCAP: a whole-genome assembly program. Genome Research *13*, 2164-2170.

Hudson, R.R. (1983). Properties of a neutral allele model with intragenic recombination. Theoretical Population Biology *23*, 183-201.

Hulo, N., Sigrist, C.J., Le Saux, V., Langendijk-Genevaux, P.S., Bordoli, L., Gattiker, A., De Castro, E., Bucher, P., and Bairoch, A. (2004). Recent improvements to the PROSITE database. Nucleic Acids Research *32*, D134-137.

Ilie, L., Fazayeli, F., and Ilie, S. (2011). HiTEC: accurate error correction in high-throughput sequencing data. Bioinformatics *27*, 295-302.

Ina, Y. (1995). New methods for estimating the numbers of synonymous and nonsynonymous substitutions. Journal of Molecular Evolution *40*, 190-226.

Jack, D.L., Paulsen, I.T., and Saier, M.H. (2000). The amino acid/polyamine/organocation (APC) superfamily of transporters specific for amino acids, polyamines and organocations. Microbiology *146 ( Pt 8)*, 1797-1814.

Jeck, W.R., Reinhardt, J.A., Baltrus, D.A., Hickenbotham, M.T., Magrini, V., Mardis, E.R., Dangl, J.L., and Jones, C.D. (2007). Extending assembly of short DNA sequences to handle error. Bioinformatics *23*, 2942-2944.

Jiang, R.H.Y., Tripathy, S., Govers, F., and Tyler, B.M. (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*, 4874-4879.

Joanna, M. (2011). Genetic drift. Current Biology *21*, R837-R838.

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

Kamoun, S. (2006). A catalogue of the effector secretome of plant pathogenic oomycetes. In Annual Review of Phytopathology, pp. 41-60.

Kamper, J., Kahmann, R., Bolker, M., Ma, L.J., Brefort, T., Saville, B.J., Banuett, F., Kronstad, J.W., Gold, S.E., Muller, O.*, et al.* (2006). Insights from the genome of the biotrophic fungal plant pathogen Ustilago maydis. Nature *444*, 97-101.

Kemen, E., Gardiner, A., Schultz-Larsen, T., Kemen, A.C., Balmuth, A.L., Robert-Seilaniantz, A., Bailey, K., Holub, E., Studholme, D.J., MacLean, D.*, et al.* (2011). Gene Gain and Loss during Evolution of Obligate Parasitism in the White Rust Pathogen of Arabidopsis thaliana. PLoS Biology *9*.

Kent, W.J. (2002). BLAT--the BLAST-like alignment tool. Genome Research *12*, 656-664.

Kimura, M. (1983). The Neutral Theory of Molecular Evolution (Cambridge, Cambridge University Press).

Kingman, J.F.C. (1982). The coalescent. Stochastic Processes and their Applications *13*, 235-248.

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

Kurtz, S., Phillippy, A., Delcher, A.L., Smoot, M., Shumway, M., Antonescu, C., and Salzberg, S.L. (2004). Versatile and open software for comparing large genomes. Genome Biology *5*, R12.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology *10*, R25.

Lazarevic, V., Whiteson, K., Huse, S., Hernandez, D., Farinelli, L., Osteras, M., Schrenzel, J., and Francois, P. (2009). Metagenomic study of the oral microbiota by Illumina high-throughput sequencing. Journal of Microbiological Methods *79*, 266-271.

Letunic, I., Goodstadt, L., Dickens, N.J., Doerks, T., Schultz, J., Mott, R., Ciccarelli, F., Copley, R.R., Ponting, C.P., and Bork, P. (2002). Recent improvements to the SMART domain-based sequence annotation resource. Nucleic Acids Research *30*, 242-244.

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

Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics *26*, 589-595.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009a). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078-2079.

Li, H., Ruan, J., and Durbin, R. (2008a). Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Research *18*, 1851-1858.

Li, R., Fan, W., Tian, G., Zhu, H., He, L., Cai, J., Huang, Q., Cai, Q., Li, B., Bai, Y.*, et al.* (2010). The sequence and de novo assembly of the giant panda genome. Nature *463*, 311-317.

Li, R., Li, Y., Kristiansen, K., and Wang, J. (2008b). SOAP: short oligonucleotide alignment program. Bioinformatics *24*, 713-714.

Li, R., Yu, C., Li, Y., Lam, T.W., Yiu, S.M., Kristiansen, K., and Wang, J. (2009b). SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics *25*, 1966-1967.

Librado, P., and Rozas, J. (2009). DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics *25*, 1451-1452.

Lomsadze, A., Ter-Hovhannisyan, V., Chernoff, Y.O., and Borodovsky, M. (2005). Gene identification in novel eukaryotic genomes by self-training algorithm. Nucleic Acids Res *33*, 6494-6506.

Lunter, G., and Goodson, M. (2011). Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. Genome Research *21*, 936-939.

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

MacLean, D., Jones, J.D.G., and Studholme, D.J. (2009). Application of 'next-generation' sequencing technologies to microbial genetics. Nature Reviews Microbiology *7*, 287-296.

Majoros, W.H., Pertea, M., Delcher, A.L., and Salzberg, S.L. (2005). Efficient decoding algorithms for generalized hidden Markov model gene finders. BMC Bioinformatics *6*, 16.

Martin, F., Aerts, A., Ahren, D., Brun, A., Danchin, E.G., Duchaussoy, F., Gibon, J., Kohler, A., Lindquist, E., Pereda, V.*, et al.* (2008). The genome of Laccaria bicolor provides insights into mycorrhizal symbiosis. Nature *452*, 88-92.

Maxam, A.M., and Gilbert, W. (1977). New method for sequencing DNA. Proceedings of the National Academy of Sciences *74*, 560-564.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M.*, et al.* (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research *20*, 1297-1303.

Mendel, G.J. (1865). Versuche über Pflanzenhybriden Paper presented at: Verhandlungen des naturforschenden Vereines (Brünn).

Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A.C., and Kanehisa, M. (2007). KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Research *35*, W182-W185.

Mouyna, I., Fontaine, T., Vai, M., Monod, M., Fonzi, W.A., Diaquin, M., Popolo, L., Hartland, R.P., and Latge, J.P. (2000). Glycosylphosphatidylinositol-anchored glucanosyltransferases play an active role in the biosynthesis of the fungal cell wall. Journal of Biological Chemistry *275*, 14882-14889.

Mukhtar, M.S., Carvunis, A.R., Dreze, M., Epple, P., Steinbrenner, J., Moore, J., Tasan, M., Galli, M., Hao, T., Nishimura, M.T.*, et al.* (2011). Independently evolved virulence effectors converge onto hubs in a plant immune system network. Science *333*, 596-601.

Nei, M., and Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Molecular Biology and Evolution *3*, 418-426.

Nielsen, R., and Yang, Z. (1998). Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics *148*, 929-936.

Ning, Z., Cox, A.J., and Mullikin, J.C. (2001). SSAHA: a fast search method for large DNA databases. Genome Research *11*, 1725-1729.

Nordborg, M. (2007). Coalescent Theory. In Handbook of Statistical Genetics (Wiley).

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., and Kanehisa, M. (1999). KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Research *27*, 29-34.

Orbach, M.J., Farrall, L., Sweigard, J.A., Chumley, F.G., and Valent, B. (2000). A telomeric avirulence gene determines efficacy for the rice blast resistance gene Pi-ta. Plant Cell *12*, 2019-2032.

Ossowski, S., Schneeberger, K., Clark, R.M., Lanz, C., Warthmann, N., and Weigel, D. (2008). Sequencing of natural strains of Arabidopsis thaliana with short reads. Genome Research *18*, 2024-2033.

Otto, T.D., Sanders, M., Berriman, M., and Newbold, C. (2010). Iterative Correction of Reference Nucleotides (iCORN) using second generation sequencing technology. Bioinformatics *26*, 1704-1707.

Parkinson, J., Anthony, A., Wasmuth, J., Schmid, R., Hedley, A., and Blaxter, M. (2004). PartiGene—constructing partial genomes. Bioinformatics *20*, 1398-1404.

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

Parra, G., Bradnam, K., Ning, Z., Keane, T., and Korf, I. (2009). Assessing the gene space in draft genomes. Nucleic Acids Research *37*, 289-297.

Pearl, F.M., Lee, D., Bray, J.E., Sillitoe, I., Todd, A.E., Harrison, A.P., Thornton, J.M., and Orengo, C.A. (2000). Assigning genomic sequences to CATH. Nucleic Acids Research *28*, 277-282.

Pelloux, J., Rustérucci, C., and Mellerowicz, E.J. (2007). New insights into pectin methylesterase structure and function. Trends in Plant Science *12*, 267-277.

Pertea, G., Huang, X., Liang, F., Antonescu, V., Sultana, R., Karamycheva, S., Lee, Y., White, J., Cheung, F., Parvizi, B.*, et al.* (2003). TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. Bioinformatics *19*, 651-652.

Pozzoli, U., Menozzi, G., Fumagalli, M., Cereda, M., Comi, G.P., Cagliani, R., Bresolin, N., and Sironi, M. (2008). Both selective and neutral processes drive GC content evolution in the human genome. BMC Evolutionary Biology *8*, 99.

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K.S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., *et al.* (2010). A human gut microbial gene catalogue established by metagenomic sequencing. Nature *464*, 59-U70.

Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R., and Lopez, R. (2005). InterProScan: protein domains identifier. Nucleic Acids Research *33*, W116-W120.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841-842.

Rehmany, A.P., Gordon, A., Rose, L.E., Allen, R.L., Armstrong, M.R., Whisson, S.C., Kamoun, S., Tyler, B.M., Birch, P.R., and Beynon, J.L. (2005). Differential recognition of highly divergent downy mildew avirulence gene alleles by RPP1 resistance genes from two Arabidopsis lines. Plant Cell *17*, 1839-1850.

Rehmany, A.P., Grenville, L.J., Gunn, N.D., Allen, R.L., Paniwnyk, Z., Byrne, J., Whisson, S.C., Birch, P.R., and Beynon, J.L. (2003). A genetic interval and physical contig spanning the Peronospora parasitica (At) avirulence gene locus ATR1Nd. Fungal Genetics and Biology *38*, 33-42.

Rehmany, A.P., Lynn, J.R., Tor, M., Holub, E.B., and Beynon, J.L. (2000). A comparison of Peronospora parasitica (Downy mildew) isolates from Arabidopsis thaliana and Brassica oleracea using amplified fragment length polymorphism and internal transcribed spacer 1 sequence analyses. Fungal Genetics and Biology *30*, 95-103.

Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: The European Molecular Biology Open Software Suite. Trends in Genetics *16*, 276-277.

Ronaghi, M., Nygren, M., Lundeberg, J., and Nyren, P. (1999). Analyses of secondary structures in DNA by pyrosequencing. Analytical Biochemistry *267*, 65-71.

Ronquist, F., and Huelsenbeck, J.P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics *19*, 1572-1574.

Rooney, H.C., Van't Klooster, J.W., van der Hoorn, R.A., Joosten, M.H., Jones, J.D., and de Wit, P.J. (2005). Cladosporium Avr2 inhibits tomato Rcr3 protease required for Cf-2-dependent disease resistance. Science *308*, 1783-1786.

Rougon-Cardaso, D.A. (2007). Identification and Characterization of Hyaloperonospora parasitica Genes Expressed during Infection of Arabidopsis thaliana (Norwich, University of East Anglia).

Rozowsky, J., Euskirchen, G., Auerbach, R.K., Zhang, Z.D., Gibson, T., Bjornson, R., Carriero, N., Snyder, M., and Gerstein, M.B. (2009). PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. Nature Biotechnology *27*, 66-75.

Sanger, F., Nicklen, S., and Coulson, A.R. (1977). DNA sequencing with chain-terminating inhibitors Proceedings of the National Academy of Sciences of the United States of America *74*, 5463-5467.

Schatz, M.C., Trapnell, C., Delcher, A.L., and Varshney, A. (2007). High-throughput sequence alignment using Graphics Processing Units. BMC Bioinformatics *8*, 474.

Schneeberger, K., Ossowski, S., Lanz, C., Juul, T., Petersen, A.H., Nielsen, K.L., Jorgensen, J.E., Weigel, D., and Andersen, S.U. (2009). SHOREmap: simultaneous mapping and mutation identification by deep sequencing. Nature Methods *6*, 550-551.

Scofield, S.R., Tobias, C.M., Rathjen, J.P., Chang, J.H., Lavelle, D.T., Michelmore, R.W., and Staskawicz, B.J. (1996). Molecular Basis of Gene-for-Gene Specificity in Bacterial Speck Disease of Tomato. Science *274*, 2063-2065.

Scordis, P., Flower, D.R., and Attwood, T.K. (1999). FingerPRINTScan: intelligent searching of the PRINTS motif database. Bioinformatics *15*, 799-806.

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

Smit, A.F.A., Hubley, R., and Green, P. (1996-2010). RepeatMasker Open-3.0 http://www.repeatmasker.org.

Soderlund, C., Humphray, S., Dunham, A., and French, L. (2000). Contigs Built with Fingerprints, Markers, and FPC V4.7. Genome Research *10*, 1772-1787.

Sohn, K.H., Lei, R., Nemri, A., and Jones, J.D. (2007). The downy mildew effector proteins ATR1 and ATR13 promote disease susceptibility in Arabidopsis thaliana. Plant Cell *19*, 4077-4090.

Sonnhammer, E.L., von Heijne, G., and Krogh, A. (1998). A hidden Markov model for predicting transmembrane helices in protein sequences. Proceedings of the International Conference on Intelligent Systems for Molecular Biology *6*, 175-182.

Spanu, P.D., Abbott, J.C., Amselem, J., Burgis, T.A., Soanes, D.M., Stuber, K., Ver Loren van Themaat, E., Brown, J.K., Butcher, S.A., Gurr, S.J.*, et al.* (2010). Genome expansion and gene loss in powdery mildew fungi reveal tradeoffs in extreme parasitism. Science *330*, 1543-1546.

Stanke, M., Diekhans, M., Baertsch, R., and Haussler, D. (2008). Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics *24*, 637-644.

Stein, L.D., Bao, Z., Blasiar, D., Blumenthal, T., Brent, M.R., Chen, N., Chinwalla, A., Clarke, L., Clee, C., Coghlan, A.*, et al.* (2003). The Genome Sequence of Caenorhabditis briggsae: A Platform for Comparative Genomics. PLoS Biology *1*, e45.

Stein, L.D., Mungall, C., Shu, S., Caudy, M., Mangone, M., Day, A., Nickerson, E., Stajich, J.E., Harris, T.W., Arva, A.*, et al.* (2002). The generic genome browser: a building block for a model organism system database. Genome Research *12*, 1599-1610.

Stenson, P., Mort, M., Ball, E., Howells, K., Phillips, A., Thomas, N., and Cooper, D. (2009). The Human Gene Mutation Database: 2008 update. Genome Medicine *1*, 13.

Stephens, M., and Scheet, P. (2005). Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. American Journal of Human Genetics *76*, 449-462.

Tajima, F. (1983). Evolutionary relationship of DNA sequences in finite populations. Genetics *105*, 437-460.

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics *123*, 585-595.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., and Kumar, S. (2011). MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Molecular Biology and Evolution *28*, 2731-2739.

Tang, X., Frederick, R.D., Zhou, J., Halterman, D.A., Jia, Y., and Martin, G.B. (1996). Initiation of Plant Disease Resistance by Physical Interaction of AvrPto and Pto Kinase. Science *274*, 2060-2063.

The Genome Institute (2011). http://genome.wustl.edu/ (St. Louis, Washington University).

Thomas, P.D., Campbell, M.J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., Diemer, K., Muruganujan, A., and Narechania, A. (2003). PANTHER: a library of protein families and subfamilies indexed by function. Genome Research *13*, 2129-2141.

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

Thompson, J.D., Higgins, D.G., and Gibson, T.J. (1994b). Improved sensitivity of profile searches through the use of sequence weights and gap excision. Computer Applications in the Biosciences : CABIOS *10*, 19-29.

Tian, M., Win, J., Song, J., van der Hoorn, R., van der Knaap, E., and Kamoun, S. (2007). A Phytophthora infestans cystatin-like protein targets a novel tomato papain-like apoplastic protease. Plant Physiology *143*, 364-377.

Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics *25*, 1105-1111.

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

Tsiamis, G., Mansfield, J.W., Hockenhull, R., Jackson, R.W., Sesma, A., Athanassopoulos, E., Bennett, M.A., Stevens, C., Vivian, A., Taylor J.D. and Murillo, J. (2000). Cultivar specific avirulence and virulence functions asigned to avrPphf in Pseudomonas syringae pv. phaseolicola, the cause of bean halo-blight disease. The EMBO Journal *19*, 3204-3214

Tyler, B.M., Tripathy, S., Zhang, X., Dehal, P., Jiang, R.H., Aerts, A., Arredondo, F.D., Baxter, L., Bensasson, D., Beynon, J.L.*, et al.* (2006). Phytophthora genome sequences uncover evolutionary origins and mechanisms of pathogenesis. Science *313*, 1261-1266.

Van der Biezen, E.A., and Jones, J.D. (1998). Plant disease-resistance proteins and the gene-for-gene concept. Trends in Biochemical Sciences *23*, 454-456.

Varshney, R.K., Nayak, S.N., May, G.D., and Jackson, S.A. (2009). Next-generation sequencing technologies and their implications for crop genetics and breeding. Trends in Biotechnology *27*, 522-530.

Voelkerding, K.V., Dames, S.A., and Durtschi, J.D. (2009). Next-generation sequencing: from basic research to diagnostics. Clinical Chemistry *55*, 641-658.

Wacey, D., Kilburn, M.R., Saunders, M., Cliff, J., and Brasier. M.D. (2011). Microfossils of sulphur-metabolizing cells in 3.4-billion-year-old rocks of Western Australia. Nature Geoscience *4*, 698-702.

Wan, J., Zhang, X.-C., and Stacey, G. (2008). Chitin signaling and plant disease resistance. Plant Signaling and Behavior *3*, 831-833.

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

Watson, J.D., and Crick, F.H. (1953a). Genetical implications of the structure of deoxyribonucleic acid. Nature *171*, 964-967.

Watson, J.D., and Crick, F.H. (1953b). Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature *171*, 737-738.

Weinberg, W. (1908). Über den Nachweis der Vererbung beim Menschen. Jahreshefte des Vereins für vaterländische Naturkunde in Württemberg *64*, 368-382.

White, J.L., and Kaper, J.M. (1989). A simple method for detection of viral satellite RNAs in small plant tissue samples. Journal of Virological Methods *23*, 83-93.

Win, J., Kanneganti, T.D., Torto-Alalibo, T., and Kamoun, S. (2006). Computational and comparative analyses of 150 full-length cDNA sequences from the oomycete plant pathogen Phytophthora infestans. Fungal Genetics and Biology *43*, 20-33.

Wong, W.S.W., Yang, Z., Goldman, N., and Nielsen, R. (2004). Accuracy and Power of Statistical Methods for Detecting Adaptive Evolution in Protein Coding Sequences and for Identifying Positively Selected Sites. Genetics *168*, 1041-1051.

Wootton, J.C., and Federhen, S. (1993). Statistics of Local Complexity in Amino Acid Sequences and Sequence Databases. Computers and Chemistry *17*, 149-163.

Wu, C.H., Nikolskaya, A., Huang, H., Yeh, L.S., Natale, D.A., Vinayaka, C.R., Hu, Z.Z., Mazumder, R., Kumar, S., Kourtesis, P.*, et al.* (2004). PIRSF: family classification system at the Protein Information Resource. Nucleic Acids Research *32*, D112-114.

Wu, R., and Kaiser, A.D. (1968). Structure and base sequence in the cohesive ends of bacteriophage lambda DNA. Journal of Molecular Biology *35*, 523-537.

Wu, R., and Taylor, E. (1971). Nucleotide sequence analysis of DNA. II. Complete nucleotide sequence of the cohesive ends of bacteriophage lambda DNA. Journal of Molecular Biology *57*, 491-511.

Xie, C., and Tammi, M.T. (2009). CNV-seq, a new method to detect copy number variation using high-throughput sequencing. BMC Bioinformatics *10*, 80.

Yaeno, T., H. Li, et al. (2011). "Phosphatidylinositol monophosphate-binding interface in the oomycete RXLR effector AVR3a is required for its stability in host cells to modulate plant immunity." Proc Natl Acad Sci U S A 108(35): 14682-14687.

Yang, Z. (2007). PAML 4: Phylogenetic Analysis by Maximum Likelihood. Molecular Biology and Evolution *24*, 1586-1591.

Yang, Z., and Bielawski, J.P. (2000). Statistical methods for detecting molecular adaptation. Trends in Ecology and Evolution *15*, 496-503.

Yang, Z., and Nielsen, R. (2000). Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Molecular Biology and Evolution *17*, 32-43.

Yucel, I., Slaymaker, D., Boyd, C., Murillo, J., Buzzell, R.I., and Keen, N.T. (1994). Avirules gene avrPphC from Pseudomonas syringae pv. paseolicola 3121: a plasmid-borne homologue or avrC closely linked to an avrD allele. Molecular Plant Microbe Interactions 7, 677-679

Zdobnov, E.M., and Apweiler, R. (2001). InterProScan--an integration platform for the signature-recognition methods in InterPro. Bioinformatics *17*, 847-848.

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

Zimmer, C. (2002). Evolution: The Thriumph of an Idea (New York, NY: Perennial).

Zipfel, C., Kunze, G., Chinchilla, D., Caniard, A., Jones, J.D., Boller, T., and Felix, G. (2006). Perception of the bacterial PAMP EF-Tu by the receptor EFR restricts Agrobacterium-mediated transformation. Cell *125*, 749-760.