SATURATION SEQUENCING, CHARACTERISATION AND MAPPING OF THE NBS-LRR RESISTANCE GENE FAMILY IN APPLE, *Malus* x *domestica* (Borkh.)

Joseph Mafofo

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor Philosophiae in the Faculty of Science, University of the Western Cape

Supervisor: Prof DJG Rees

August 2008

ABSTRACT

Saturation sequencing, characterisation and mapping of the NBS-LRR resistance gene family in apple, *Malus x domestica* (Borkh.)

J. Mafofo

PhD thesis, Department of Biotechnology, Faculty of Science, University of the Western Cape

To date five classes of resistance proteins have been identified in plants and these include the intracellular protein kinases, receptor-like protein kinases with extracellular leucine-rich repeat (LRR) domain, LRR proteins that encode membrane bound extracellular proteins, toxin reductase and intracellular LRR proteins with a nucleotide-binding site (NBS). These proteins recognise "invading pathogen" and in turn trigger defence response systems that act to protect plants from invading pathogens. The NBS-LRR genes which constitutes the major class encode a family of resistance proteins that are made up of a centrally located nucleotide binding site domain and a C-terminal leucine rich repeat receptor. This class of genes constitute the largest family of resistance genes identified in plants to date. They make up the majority of proteins involved in the plant basal and inducible defence systems against pathogen infection. The LRR domain has been classified as a receptor that recognises specific pathogen avirulence factors. This process in turn triggers the NBS domain to hydrolyse the α-phosphate bond of a bound

GTP molecule. The energy thus generated is used to initiate a cascade of signal transduction reactions that activates the hypersensitive response (HR). This HR resistance response involves the killing of all infected cells and thus localising the invading pathogen to the site of the entry thereby preventing the spread of the infection to new cells. Death of infected cells is manifested on leaf surfaces as necrotic lesions.

Genes in the NBS-LRR family are classified into TIR and non-TIR subfamilies based on the presence or absence of an N-terminal domain with functional homology to the *Drosophila* Toll/ mammalian interleukin 1 receptor-like domain respectively. Members of these subfamilies occur as clusters on the genome with genes responding to different pathogen infections. The question of how many genes make up the NBS-LRR family has not been answered in apples although estimates have been made in the *Arabidopsis* model plant using sequence data from the *Arabidopsis* Genome Project. The main aim of this project was to investigate the NBS-LRR gene copy number, identify markers linked to them and show tentative proof that their transcription levels are upregulated following infection of the plant by pathogens.

PCR amplification of RGAs from *Malus x domestica* (Borkh.) Anna and Golden Delicious genomic DNA were performed using a set of degenerate oligonucleotides complementary to the highly conserved P-loop and GLPL motifs flanking the NBS domain. A total of 661 RGAs were sequenced using cloning and the Sanger sequencing system and an additional 142 000 were generated using the 454 sequencing system. The 454 dataset was generated from sixteen samples of the *Malus x domestica* (Borkh.) cv.

Anna x Golden Delicious bin mapping population including the two parentals. The Sanger dataset contained sequences with an average length of ±520 nucleotides and was used for phylogenetic analyses of the NBS-LRR gene family. Inference of phylogeny using this dataset and control NBS-LRR genes from *Arabidopsis*, flax and rice were broadly distributed into TIR and non-TIR subfamilies with a total of 14 clusters. *In silico* translation of this data revealed that this gene family is made up of a large number of pseudogenes (~ 45%) that can be distinguished from functional genes by the presence of either premature termination codons or other frame-shift mutations within their open reading frames. Analysis of coding to silent mutation rates per cluster showed a set of clusters undergoing diversifying, neutral and purifying selection. Two of the three clusters undergoing purifying selection were detected as members of the non-TIR subfamily. Evidence of gene conversion was detected in eight RGA clusters, three of which belong to the non-TIR subfamily. About 43% of the gene conversion events detected were between functional genes, 30% between pseudogenes and 27% were between a pseudogene and a functional gene.

Sequence assemblies performed using both Sanger and 454 datasets gave a total of 278 contigs, 54 of which contained at least one sequence from the Sanger dataset and the rest were from the 454 dataset. Total contigs per sample for the 14 *Malus x domestica* (Borkh.) cvs. Anna x Golden Delicious bin mapping progeny ranged from 220 to ±400 and for the two sets of analyses up to a thousand sequences remained unassembled. These figures suggests that there are at least ±400 RGAs in the apple genome, understanding the exact number of unique singletons which would either show low copy RGAs or those

with a low PCR amplification efficiency is somehow complicated owing to possible artefacts from both amplification and sequencing processes. Such finer analyses would require mining the completed apple genome sequence.

The NBS-LRR transcriptome analysis experiments were performed using seedlings from the Carmine x Simpson and Lady Williams x Prima crosses. Carmine and Prima cultivars used as parents in these crosses were shown to possess disease resistance properties to powdery mildew and apple scab respectively. The design of these crosses therefore gives two populations in which the resistance traits are segregating and thus allows comparative analysis of the susceptible against the resistant transcriptome. Seedlings from the Lady Williams x Prima cross were inoculated with a virulent field isolate of *Venturia inaequalis* and those from the Carmine x Simpson cross were inoculated by exposure to airborne *Podosphaera leucotricha* spores. PCR amplification of candidate RGAs from RNA isolated before and two weeks following exposure to infection was thus performed to analyse the NBS-LRR transcriptome. DNA fragments generated from these PCR amplification experiments were purified and sequenced using the GS20 system and the data was used through assemblies to estimate genes transcribed in the presence of pathogen infection. This process showed sets of genes that are constitutively transcribed, repressed and relatively fewer genes that were induced by inoculation with the pathogen. A selection of these contigs were further analysed using quantitative real-time PCR and results showed some genes that appear to be linked to disease resistance in both scab and powdery mildew experiments. GD36, which showed marked up-regulation in scab-infected samples was located in RGA cluster 14 and subsequently mapped to LG1 in the

Anna x Golden Delicious genetic linkage map. AnRGA346 in RGA cluster 13 was shown to be up-regulated in both scab-infected and powdery mildew-infected samples and GD120 from cluster 13 showed marked up-regulated transcription in powdery mildew infected samples. These results indicate that the transcription of NBS-LRR genes does respond to pathogen infection and more work will be needed to investigate the precise function of these up-regulated genes.

Identification of SNPs was performed using two strategies. The first strategy used re-sequencing of selected RGA clusters from the *Malus x domestica* (Borkh.) cv Golden Delicious phylogenetic tree. In this experiment, oligonucleotides were designed to exclude the degenerate priming sites and also targeted on unique regions of the sequences in the selected clusters. These were used to selectively amplify genes from the selected clusters in *Malus x domestica* (Borkh.) cvs Anna and Golden Delicious. Sequence of assemblies of data from these oligonucleotides were used to identify a set of candidate SNPs which were then genotyped in the fourteen seedlings of the *Malus x domestica* (Borkh.) cvs. Anna x Golden Delicious bin mapping population using the SNaPshot™ method. In the second strategy, all sixteen samples of the bin mapping population were sequenced using the GS FLX (454 Life Sciences) system and candidate SNPs were detected by sequence assemblies.

Three SNPs segregating from the Anna cultivar were identified from one cluster using the re-sequencing strategy. These were mapped to Anna linkage group 5 (LG 5) on the existing *Malus x domestica* (Borkh.) Anna x Golden Delicious genetic linkage map in the

bin defined by the SSR markers CH02b12a and CH03a04. The GS FLX sequencing and sequence assemblies strategy facilitated for the identification of an additional eight mappable SNPs and these were positioned on linkage groups 1, 6, 12, 16 and 17 with variable degrees of precision. Of the nine mapped SNPs, four were co-localised with published NBS markers on loci containing powdery mildew and scab resistance genes. The GS FLX sequencing and assembly strategy identified a large number of candidate SNPs that were heterozygous in both parents and thus could not be mapped using the bin mapping approach.

Data from the GS FLX sequencing showed evidence of distorted sequence distribution in the assembled contigs. A subset of contigs contained significantly below average to no sequence representation from other progeny and this reduced the number of candidate SNPs that could be genotyped. Of the candidate SNPs that were successfully genotyped, the majority showed complete heterozygosity in all bin mapping progeny, which is characteristic of gene conversion and paralogous sequence variants. Although this experiment did not manage to map all identified RGAs, it provided a valuable collection of candidate SNPs in the parental datasets that can be genotyped using either SNaPshot™ or SNPLex™ methods.

KEYWORDS

Nucleotide binding site – leucine rich repeat (NBS-LRR), Resistance gene analogs (RGA), Hypersensitive response (HR), Sequencing, Phylogenetic, Cluster, Single

nucleotide polymorphism (SNP), Paralogous sequence variation (PSV), Transcriptome, Sequence assembly

DECLARATION

"I hereby declare that **"Saturation sequencing, characterisation and mapping of the NBS-LRR resistance gene family in apple, *Malus x domestica* (Borkh.)"** is my own work and that, to the best of my knowledge it contains no material previously published or written by another person nor material which to a substantial extent has been accepted for the award of any other degree or diploma in any other university or institute of higher learning. All the sources used and cited have been duly indicated in text and acknowledged by complete references thereafter."

Joseph Mafofo March 2008

Signed:……

ACKNOWLEDGEMENTS

My most heartfelt and sincere gratitude to Lynn (my wife) and our kids Isheanesu and Ruvarashe, thank you for everything - I love you guys. I would like to thank my supervisor, Prof. DJG Rees for both financial and academic support; I really appreciate all your efforts in getting me through my studies. I also wish to thank Dr Iwan Labuschagne (ARC-Nievroobij) for his valuable expertise in apple breeding and pathology, staff and students of the Biotechnology Department and last but certainly not least all the members of 'Jasper's lab'. This work was funded by DFPT, DTI-THRIP, and the European Union.

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

ABC	ATP binding cassette
ABI	Applied Biosystems
AFLP	Amplified fragment length polymorphism
AnRGA	Anna resistance gene analog
ATP	Adenosine triphosphate
Avr/ Avr	Avirulent protein/ avirulent gene
BAC	Bacteria artificial chromosome
BC-KA	Bonferroni –corrected Karlin and Altschul
BLAST	Basic local alignment search tool
bp	Base pair
CAPS	Cleaved - amplified polymorphic sequence
CDD	Conserved Domain Database
cDNA	complementary DNA
Cf	Cladosporium falvum resistance gene
cM	centimorgan
CNL	Coiled Coil –Nucleotide Binding Site – Leucine Rich Repeat
CTAB	N-cetyl-N,N,N-trimethyl-amino bromide
ctg	contig
cv(s)	cultivar(s)
dbEST	expressed sequence tag database
ddNTP	dideoxynucleotide triphosphate
DNA	deoxyribonucleic acid

dNTP	deoxynucleotide triphosphate
EDS1	enhanced disease susceptibility signalling protein 1
EDTA	ethylene diamine tetraacetic acid
EGF	epidermal growth factor
EST	expressed sequence tag
Exo I	Exonuclease I
FA	formaldehyde denaturing agarose
GD	Golden Delicious
gDNA	genomic deoxyribonucleic acid
GDR	Genome Database for Rosaceae
GS20	Genome Sequencer 20TM sequencing system
GSS	genome survey sequence
GTP	guanosine triphosphate
GTR	general time reversible
HKY	Hasegawa, Kishino and Yano model of nucleotide substitution
HR	hypersensitive response
HSAP	high scoring aligned pair
HyPhy	Hypothesis testing using Phylogenies
JTT	Jones, Taylor and Thornton model of amino acid substitution
Ka/Ks	non-synonymous to synonymous ratio
kb	kilobase
LB	Luria-Bertani
LG	linkage group

LRR	leucine rich repeat
MAPK	mitogen activated protein kinase
MAS	marker assisted selection
Mb	megabase
MCMC	Markov Chain Monte Carlo
MEGA	Molecular Evolutionary Genetics Analysis software
Myr	million years
NBS	nucleotide binding site domain
NBS-LRR	nucleotide binding site - leucine rich repeat
NCBI	National Centre for Biotechnology Information
nc-RNA	non-coding RNA
NDR1	non-race specific disease resistance protein 1
ORF	open reading frame
PAMP	pathogen associated molecular pattern
PAUP	Phylogenetic Analysis Using Parsimony
PCD	programmed cell death
PCR	polymerase chain reaction
PDB	protein data bank
PR	pathogenesis related
PSV	paralogous sequence variation
QTL	quantitative trait locus
R gene	Resistance gene
RAPD	random amplified polymorphic DNA

REL	random effects likelihood
RFLP	restriction fragment length polymorphis
RGA	resistance gene analog
RLK	receptor like kinase
RNA	ribonucleic acid
ROS	reactive oxygen species
RT	reverse transcriptase
RT-PCR	reverse transcriptase – polymerase chain reaction
SA	salicyclic acid
SAG	salicyclic acid ß-glucosidase
SAP	shrimp alkaline phosphatase
SAR	systemic acquired resistance
SLAC	single likelihood ancestor counting
SNP	single nucleotide polymorphism
SPR	subtree pruning and regrafting
SSAP	Single-sequence amplification polymorphism
SSCP	Single-stranded conformation polymorphism
SSR	simple sequence repeat
STAND	signal transduction ATPases with numerous domains
STY	serine/threonine/tyrosine
TBE	tris borate EDTA
TE	Tris-EDTA
TIR	Toll/Interleukin receptor

TM	transmembrane
TNL	Toll/Inteleukin -1 Receptor Nucleotide Binding Site – Leucine Rich Repeat
TTSS	type three secretion system
UV	ultra violet
V/cm	volts per centimetre
WAK	wall associated kinase
WRKY	transcription factors defined by the presence of one or more conserved motifs with the signature sequence WRKYGQK

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

1.1 Economics of Apple production

Apples constitute a very important fruit crop that is rated fourth in the world after all citrus, grapes and bananas both in terms of return on investment and food value. Figures quoted in the Deciduous Fruit Producer's Trust Annual Report of 2006 indicate that apple production takes up approximately 28% of the total hectarage under deciduous fruits in South Africa (Louw, 2006). This is second only to grapes with an estimated 30% contribution to the approximately 74 000 hectares under deciduous fruit production. South Africa produced a total of 3.1 million metric tonnes of all deciduous fruits in 2006 of which an estimated 780 000 were apples. The total world apple production was estimated at 62 million tonnes and South Africa was rated 15th. The graph in Figure 1.1(a) below shows world apple production figures of 2006 (Deciduous Fruit Production Trust Annual Report, 2006).

The gross value of South Africa's agricultural production was estimated at R70 million in 2005 and the Horticultural industry contributed approximately 7% to this figure. According to Statistics South Africa reports of 2007, deciduous fruit exports contributed an estimated 16% to the total agricultural earnings in 2005 and apples in particular were rated as number three at 35% followed by grapes and pears at 33 and 24% respectively. As a whole the agricultural sector represents about 8% of the country's total earnings through exports. Agriculture contributes approximately 3 and 7% to the country's gross domestic product and formal employment respectively.

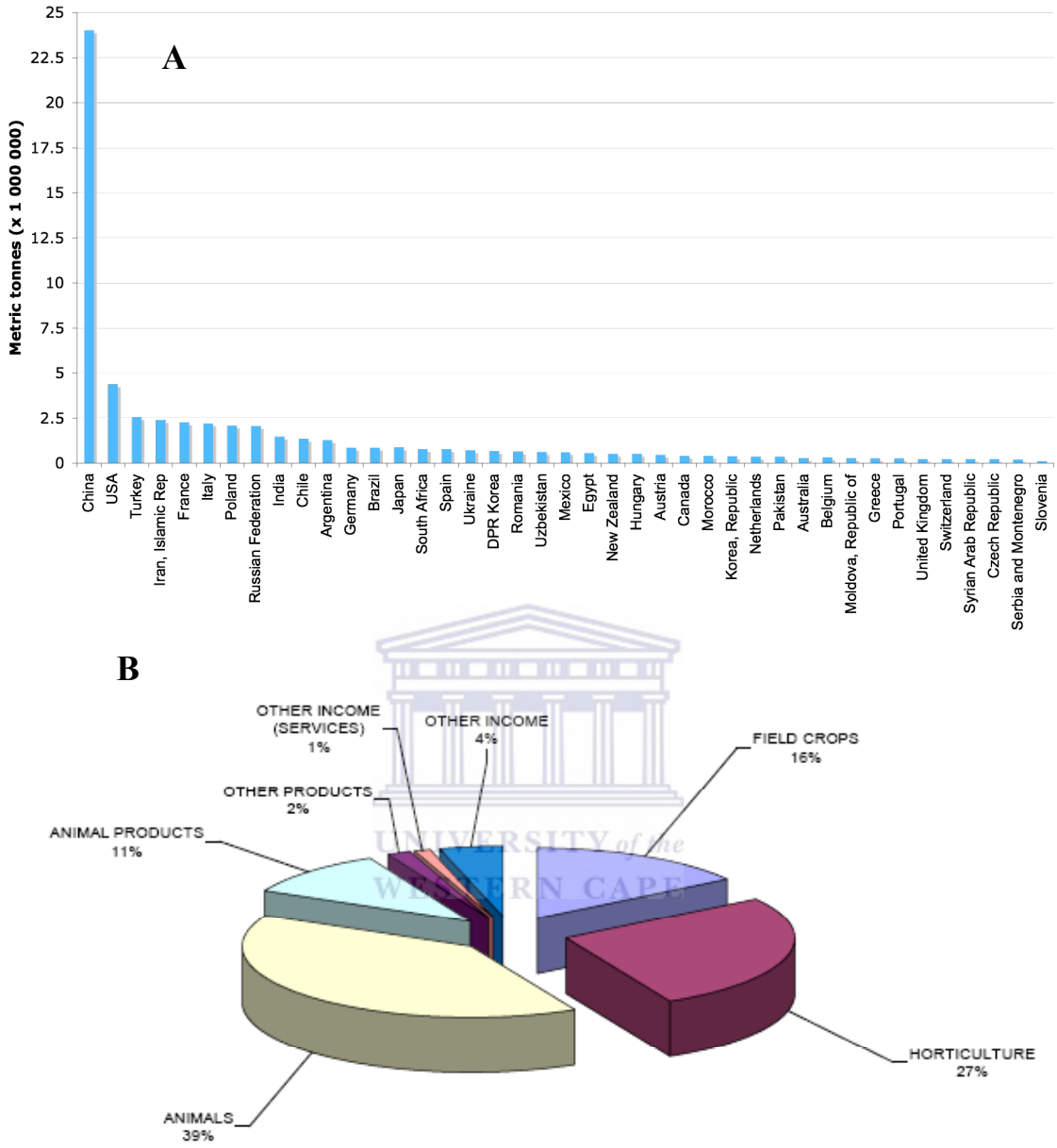


Figure 1.1. The importance and contribution of South Africa's horticultural industry. Figure (A) shows 42 top world producers of apples and (B) shows the percentage distribution of gross farming income by major division within agriculture (Statistics South Africa, 2007).

Although the agriculture sector contributes an estimated 7% as formal employment this figure increases to approximately 30% including seasonal casual labour and in a country that has an estimated 40% unemployment rate (unofficial estimates) this makes it a formidable force in job creation. According to production figures cited in the discussion above, this sector provides an important source of foreign currency earnings for the country.

Apples are not only important as a source of foreign currency; they also represent a very important and healthy food source. They have no fat and cholesterol and have very insignificant levels of sodium although they do contain a small amount of potassium. The nutritional composition of apples makes them a very healthy foodstuff and evidence to this effect has been provided over the years through laboratory-based experiments. There is evidence that apples could protect against formation of mammary tumours (Liu *et al.*, 2005) and neurodegenerative diseases such as Alzheimer's and Parkinson's (Rogers *et al.*, 2004; Tchantchou *et al.*, 2004; Chan *et al.*, 2006). Apples constitute a cheap source of antioxidants such as flavanoids that are well known for their ability to counteract the effect of free radicals. As a result apples are now being investigated in the laboratories for their potential use as dietary supplements against cancer, heart and other chronic diseases (Lewis *et al.*, 2004).

Historically apples have always been regarded as valuable components of a healthy diet. In commercial agriculture they are also valuable both on the domestic and in foreign markets. Apple orchards as well as related industry continue to provide employment both

to the low and high-income groups. Intervention from biotechnology especially focussed at improving yields and sustainability of production becomes thus a very useful exercise. This work is intended to make contributions to this industry through studying and understanding plant pathogen interaction. It is among a number of research efforts whose major goals are to understand the genetics of disease resistance and thus provide a useful knowledge base that could be useful in the generation of disease resistant apple varieties.

1.2 Current Trends in Apple Plant Biotechnology

The most devastating diseases of commercial importance against apples are caused by fungal (e.g. scab and powdery mildew) and bacterial (e.g. fireblight) pathogens (Gessler and Patocchi, 2007). Insects such as woolly apple aphids also pose a threat to apple orchards though they may not be as devastating and as widespread as fungal and bacterial diseases. To date two technologies, genetic engineering (generation of transgenic plants) and marker assisted breeding hold the key to the production of apple plants with durable broad-spectrum resistance. A number of trial transgenic plants have been produced using a variety of enzymes mostly targeted at chitin, a cell wall component found in fungi, bacteria and insects. Higher plants naturally produce very small quantities of chitinases and thus attempts to increase the level of expression gave anything from no effect to significant increases in fungal resistance though in some cases subsequent reduction in plant vigour was noted (Bolar *et al.*, 2001; Brunner *et al.*, 2003)

Experiments based on the over-expression of endogenous genes have been performed e.g. polyphenol oxidase in transgenic tomato which consequently produced a significant

reduction in disease severity (Li and Steffens, 2002). There are trials that used a range of enzymes with antimicrobial properties such as proteanase inhibitors (Maheswaran *et al.*, 2007), others managed to induce death of moth larvae on transgenic apple leaves expressing high levels of avidin or streptavidin (Markwick *et al.*, 2003). Although these experiments carry a lot of promise, the generation of genetically modified apples is a technology that the world market is not yet ready to embrace. This is mainly due to the uncertainty surrounding the question of a practical and acceptable minimal trial period needed before both certification by FAO and acceptance by the ordinary consumer.

The parallel focus in the last few years to date has been the identification of endogenous resistance genes and pyramiding these in commercial cultivars using marker-assisted breeding. In the early 1990s single dominant genes such as *Vf* and others resistant to apple scab were introduced into apple plants either through transgenic of classical breeding approaches (Gygax *et al*., 2004). The discovery of *Venturia inaequalis* races 6 and 7 that could overcome all known *Vf* genes transformed the general focus towards the multiple gene strategies that are capable of offering a broad spectrum resistance against multiple races of apple scab. V. *inaequalis* race 6 was found to be avirulent against *M. floribunda* 821 (*Vfh* gene) and race 7 against Golden Delicious and Prima carrying the *Vg* gene (Parisi *et al.*, 1993; Roberts and Crute, 1994).

The current focus thus is aimed at the identification of more resistance genes either against races 6, 7 and 8 or that possess multiple resistance to all races of *V. inaequalis*. Discovery of such non-host resistance genes not only against scab but for all apple

diseases will allow the generation of plants that give broad spectrum resistance to all pathovars of a given pathogen species endemic to a given geographical region. This has given rise to the need for in-depth knowledge of the chromosomal distribution and functions of resistance genes naturally occurring in the apple genome.

Traditional methods used for the selection of disease resistance traits require that the candidate plants be reasonably mature. Apples have a long juvenile phase and thus the process becomes expensive as disease susceptible plants are also maintained for years before they can be identified and discarded. PCR-based marker assisted selection (MAS) makes use of markers that are linked to traits of interest and allows a significant reduction in the waiting period from planting to the identification of seedlings carrying the requisite traits. Using PCR-based MAS techniques, seedlings are tested as soon as they produce leaf material following breeding for resistance. This technique allows the farmer or breeder to save both time and money.

Coupling MAS and functional genomics can allow bulk breeding for resistance facilitated by the ease associated with assessing thousands of progeny in a short time. The use of functional genomics in identifying candidate resistance genes/ resistance gene analogs (RGAs) provides a useful strategy for pyramiding functional genes. This approach provides another possible alternative to creating apple varieties that possess broad-spectrum disease resistance.

The *Arabidopsis* genome initiative has shown among other things that *Arabidopsis thaliana* has approximately 200 RGAs (Fluhr, 2001; McHale *et al.*, 2006). Targeted gene sequencing on *Malus* RGAs shows that sequence data generated in the last few years to date is far from being exhaustive. Sequencing these genes, identifying those specifying resistance to a variety of pathogens and markers tightly linked to them constitutes a very valuable strategy that can lead to the development of apple varieties with minimal demand for fungicides and other sprays.

1.3 Plants and the disease challenge

Plants and pathogens are involved in relationship characterised by complicated and dynamic phenomena in which either system is continually adapting their strategies for species survival (Richter and Ronald, 2000). An example of these phenomena is the feeding of aphids. Under normal circumstances a puncture on the phloem induces sieve-plate occlusion but aphids over time have managed to develop a two-saliva system that inhibits deposition of forisomes and thereby stops the subsequent plugging of sieve element puncture (Will *et al.*, 2007). There are a number of such intricate mechanisms both for and against pathogen infection.

The plant defence system has become multilayered starting from a generalised first line that is chiefly made up of physical and chemical barriers of constitutively expressed waxes, cell wall components, antimicrobial peptides, proteins and other non-proteinaceous secondary metabolites that deter mechanical invasion. Following mechanical injury plants initiate a series of defensive systems among which is the release

of volatile emissions that attract natural enemies of the aggressor, ethylene and jasmonic acid recruit systemin-related systems and all these are aimed at detracting the invader (Schmelz *et al.*, 2007). These primary barrier systems not only deter microbes, some are even toxic to multicellular invaders. The first line of defence is not always effective against microbial pathogens thus secondary and quaternary levels of defence made up of inter- and intra-cellular gene-encoded resistance mechanisms take over.

According to Hrazdina (1997), an analysis done on apple leaves, stems and roots showed major phenolic compounds such as phloridzin and phloretin that accumulate in millimolar quantities following infection with scab. Other major metabolites that follow this trend are p-coumaric and benzoic acid. The same work also showed an accumulation of biphenyl or dibenzofuran compounds in cell suspension cultures of *Malus floribunda* 821. These compounds however, are produced both in susceptible and resistant cultivars (Hrazdina, 1997) showing that they represent the first line of non-specific defence systems to fungal infections.

Despite the presence of these constitutively expressed defence systems in apples, a number of diseases still manage to cause serious problems for the farming community. The major diseases of economic importance are discussed in section 1.3.1.

1.3.1 Apple Diseases

Despite the complexity displayed at the primary defence level, the primary (constitutively expressed) defence system is still easily overcome by a number of pathogens that

subsequently cause serious diseases. In apples, most of these diseases destroy the aesthetics of the fruits and thus affect market appeal; others even manage to devastate orchards by destroying the fruit tree itself. The most commercially important diseases of apples are summarised in Table 1; the current status of two of these will also be discussed in detail.

Table 1: The most commercially important diseases of apples. Diseases are arranged in this table depending on the level of importance in commercial production.

Disease	Disease characteristics
Scab	Brown necrotic lesions on the surface of the leaves as well as the fruit.
Fireblight	Sudden wilting followed by shrivelling and blackening of leaves, shoots and the developing fruit
Powdery Mildew	Grey powdery depositions on the leaves and young shoots and fruit.
Woolly Apple Aphid	White spongy areas on the bark, which has the appearance of fungal infections. There is extensive galling on the root system.
Aphids	Distorted young shoots and leaves
Codling Moth	Maggots that grow in the fruit
Canker	Sunken discoloured blotches on the tree bark
Rust	Small yellow to orange spots on the upper side of leaves that turn black after a short while; affected areas appear blistered
Brown rot	Browning of fruit especially under storage

1.3.1.1 Apple Scab

Scab is a serious disease of commercial importance that affects orchards in all countries where apples are grown. It is caused by *Venturia inaequalis,* a haploid filamentous ascomycete whose life cycle enables it to adapt both to the host and the environment yearly (MacHardy *et al*., 2001; Boudichevskaia *et al*., 2006). This allows the fungus the potential to generate new pathotypes with a high frequency, work done to date has revealed three new races; 6, 7 and 8 that still present a serious challenge to farmers (Boudichevskaia *et al.*, 2006). Most commercial cultivars are susceptible to this disease although some show resistance to races 1 up to 5. The status of the research on scab resistance genes was discussed in section 1.2 above.

Mapping experiments have located the *Vf* gene to linkage group 1 (LG1) and a number of other scab resistance genes to a region on linkage group 2 (LG2); *Vh*2 and *Vr* using markers CH02b10, OPL19, OPZ13 (Gygax *et al*., 2004), *Vr*2 was located on LG2 about 43 cM from *Vh*2 using the marker CH02c02a (Patocchi *et al.*, 2004), *Vh*4, *Vh*8 and *Vbj* were also located within LG2 (Bus *et al.*, 2005). There are other genes of importance that have been located on other linkage groups such as *Vb* and *Vg* (LG12), *Vd* (LG10). However, LG2 has appeared to be very interesting to date given that Baldi *et al*. (2004) located 2 RGAs within 3.5 cM of *Vr*2 and Calenge (2004) located 6 NBS containing markers within 5 cM around *Vr*2 and 7 within 3 cM of *Vh*4. Based on these results it would appear that RGAs could facilitate the description of new genes of importance that haven't been identified to date.

1.3.1.2 Fireblight

Fireblight is among the most devastating diseases that attack plants in warm and humid climatic conditions. It is caused by *Erwinia amylovora* a member of the family *Enterobacteriaceae*, and is not restricted to apples only but also attacks many other plants (Halbwirth *et al*., 2003). Stages from fruit set to young fruits are also affected hence governments make it mandatory that all affected plants be destroyed to avoid an uncontrolled spread of the disease.

Work has been done in developing a set of molecular markers linked to fireblight resistance genes. Calenge *et al*. (2005) classified fireblight resistance as polygenic and identified a number of QTLs mapping on LG3, 7, 12, 13 and LG5 (Peil *et al*., 2006). A fire blight resistance QTL (FBF7) was identified on linkage group 7 of apple cultivar Fiesta and was shown to improve resistance in susceptible cultivars (Khan *et al.*, 2006; Khan *et al.*, 2007). LG12 also contains the scab resistance genes Vb and Vg and a number of NBS markers have been placed on this linkage group (Calenge *et al.*, 2005). The exact mechanism of resistance seems very complex at present and judging by the general absence of literature on the subject more work is required if fireblight research is to match the current status in other diseases of economic importance.

Several other genes that confer resistance to diseases mentioned above have also been mapped onto apple linkage groups. These include powdery mildew resistance genes *Pl-w* (LG8), *Pl-d* and *Pl-1* (LG12) and *Pl-2* (LG11) using SSR, RAPD, AFLP and isoenzymes in the case of *Pl-w* (James *et al.*, 2004; Gardiner *et al*., 2007). Woolly apple aphid genes,

Er-1 and *Er-3* were both located on LG8 although positions for *Er-m* and *Er-4* have not been determined yet. The determination of loci where these genes are located makes it possible for breeders to pyramid genes of economic importance using backcrossing. Subsequently, PCR-based diagnostic tools such as SSRs can then be used to verify the presence such genes. This technology is set to revolutionise apple breeding by providing the platform for the pyramiding of resistance genes and production of cultivars with multiple resistance specificities.

1.4 Plant Resistance Genes

Although plants can make use of preformed barriers (e.g.) toxic saponins, that target membrane sterols in some pathogens, other plant pathogens have little or none of these targets e.g. *Phytophthora* (Baldauf *et al.*, 2000). Plant genomes encode a set of complex networks of inducible and constitutively expressed genes. Those that are constitutively expressed provide the basal defense system that ensures an ever-present defense arsenal in each cell. The non-host complex of defense genes ensures a broad-spectrum resistance targeted at the microbe-derived structural components referred to as pathogen-associated molecular patterns (PAMPs). These could be lipopolysaccharides, flagellin, elongation factor Tu or DNA methylation patterns characteristic of the invading pathogen (Ham *et al.*, 2007; Krzymowska *et al.*, 2007). These are always structural components of microbes with a stable evolutionary status and thus are reliable perception targets. The genome encoded non-host defense systems lead through a complicated network of signal transduction pathways to an inducible defense including the HR system that protects the plant from infection by all isolates of a given microbe (Nurnberger and Lipka, 2005). The

primary role of the HR system is to stop the spread of the infection through killing of infected cells thereby containing the invading pathogen to the site of entry.

Plant pathogens have co-evolved with a core group of hosts and as such their ability to counter plant defences entails compatibility and the opposite of this results in a non-host relationship (Ham *et al.*, 2007). About five classes of R genes have been identified in plants and among the core of their function is the ability to scan and recognise microbial type III effectors. These R gene classes include the NBS-LRR protein kinases, extracytoplasmic-LRR transmembrane proteins (eLRR-TM), LRR, and toxin reductase (Lee, Seo, Rodriguez-Lanetty and Lee, 2003). The first three classes in this list are discussed in sections 1.4.1 through 1.4.3 of this review.

1.4.1 NBS-LRR resistance genes

The NBS-LRR genes constitute the largest and most diverse family of resistance genes in plants (Wroblewski *et al.*, 2007). These encode for a family of proteins with a centrally located nucleotide-binding site (NBS) and a C-terminal leucine-rich repeat (LRR) region. This family of plant resistance genes are hypothesised to function in a classical gene-for-gene interaction in which pathogen elicitors are recognised by the C-terminal LRR receptor region and a hypersensitivity response is activated. This model of interaction is also characteristic of a number of pathogen-host relationships and was first described by Flor (1971). NBS-LRR proteins can be subdivided into two major subfamilies based on the presence of either the amino-terminal Toll/interleukin-1 receptor (TIR) or the coiled-coil (CC) regions (McHale *et al.*, 2006). The classification into TIR-NBS-LRR (TNL) and CC-NBS-LRR (CNL) protein subfamilies is shown in Figure 1.2 below.

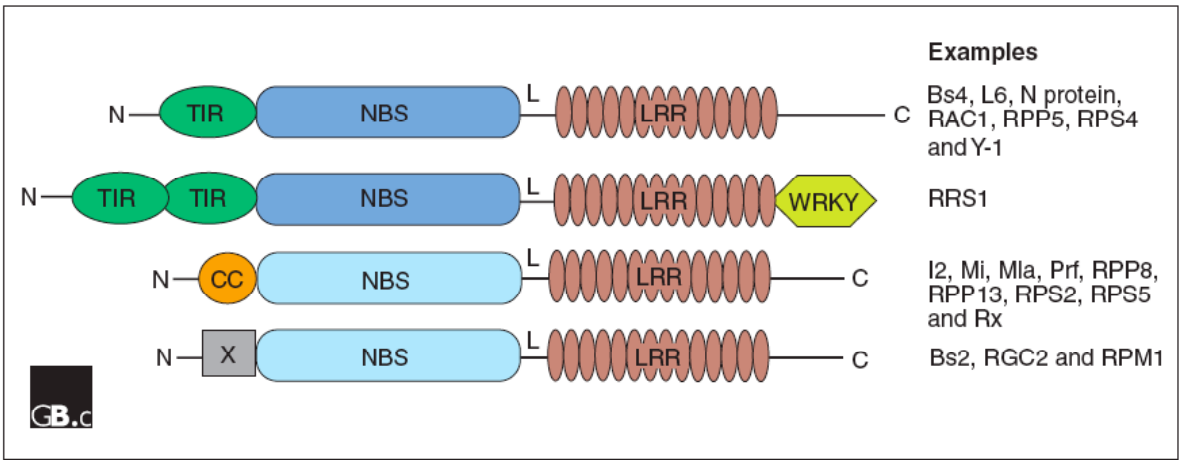


Figure 1.2. Graphical representation of the major domains of the NBS-LRR protein. The proteins included in this diagram are RAC1, RPP5, RPS4, RRS1, RPP8, RPP13, RPS2, RPS5 and RPM1 from *Arabidopsis*; Y-1 and Rx (potato); Mla (barley); RGC2 (lettuce) and Bs2 (pepper). X represents the unknown N terminal additions in some of the cloned NBS – type proteins, L, linker motif (McHale *et al.*, 2006).

Some authors talk of a leucine zipper (LZ) molecule as the amino-terminal region found in the non-TIR subfamily (Dinesh-Kumar *et al.*, 2000; Lee *et al.*, 2003). These two N-terminal regions participate in pathogen recognition and also determine which signal transduction pathway is used in the activation of the defence arsenal. *Drosophila* and mammals share a class of transmembrane receptors that function both in innate and adaptive immunity (Zhou *et al.*, 2007). The Toll-like class of receptors are pattern-recognition receptors (PRR) that are often found on immune cells e.g. B and dendritic cells, macrophages etc and are used to identify pathogen associated molecular patterns, PAMPs (Agrawal and Kandimalla, 2004). Table 2 below gives an example of R genes that have been identified and classified to date.

Table 2. NBS-LRR resistance gene classes, sources and the respective pathogens they are directed against.

Class	R proteins	Pathosystem	Source	Reference
TIR	N	TMV	Tobacco	(Whitham *et al.*, 1994)
	L	Rust	Flax	(Lawrence *et al.*, 1995)
	M	Rust	Flax	(Anderson *et al.*, 1997)
	RPP5	Bacterial	*Arabidopsis*	(Parker *et al*., 1997a)
Non-TIR	RPS2	*P. syringae*	*Arabidopsis*	(Mindrinos *et al.*, 1994)
	RPS5	*P. syringae*	*Arabidopsis*	(Parker *et al.*, 1997b)
	Prf	*P. syringae* spp	Tomato	(Salmeron *et al.*, 1996)
	RPP8	Downy mildew	*Arabidopsis*	(McDowell *et al*., 1998)
	Mi	Root knot nematode	Tomato	(Vos *et al.*, 1998)
	Rx1	PVX	Potato	(Bendahmane *et al.*, 1999)

In addition to the N-terminal substitutions, the NBS-LRR class contains a highly conserved hydrophobic GLPL(AL) motif between the NBS and LRR domains. Although the function of this motif has not been elucidated to date, its amino acid composition makes it ideal for anchoring the R-protein to the cell membrane. Mutations targeted at this motif compromise the resistance of the R gene, the G→E point mutation mentioned in Dodds *et al*. (2001) is the only example to date and it was shown to completely abolish P2 resistance in flax. Figure 1.3 below gives a graphical representation of the putative structure of RPS4, which is from the CNL subfamily of *Arabidopsis*.

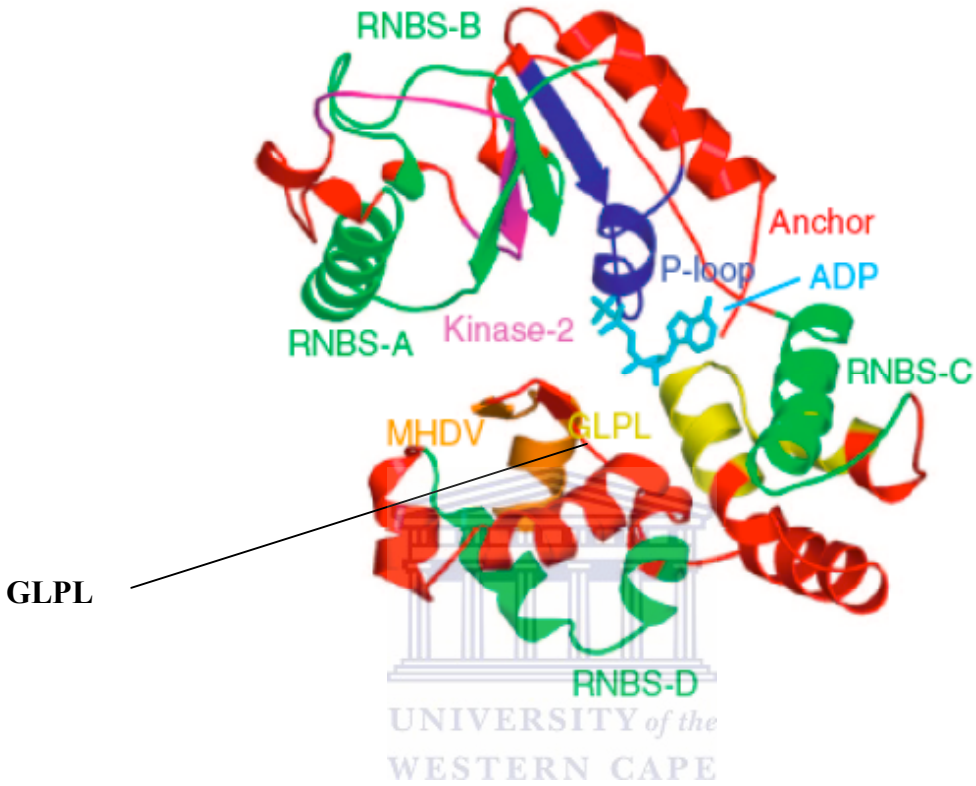


Figure 1.3. The predicted structure of *Arabidopsis* RPS4 NBS domain. The GLPL(AL) hydrophobic domain is located next to the LRR (not included in this figure) domain possibly for anchoring the protein to the cell membrane (McHale *et al.*, 2006).

The NBS domain contains three highly conserved structural motifs that include the P-loop, kinase-2 and kinase-3a (Bozkurt *et al.*, 2007). The P-loop motif [GWGGGGG**K**/T/S] is believed to interact with Mg^{2+} ions; the kinase-2 motif [LIVLDD] is characterised by four consecutive hydrophobic amino acids followed by aspartate residues the first of which is believed to interact with the third phosphate of

ATP and thereby act in a phosphotransfer reaction; the kinase-3a motif [DWFGxGSRIITTTRI] contains a tyrosine or arginine residue (in most cases) and is involved in binding ribose or purine sugar molecules of the nucleotide triphosphates (Dinesh-Kumar, 2000; and references therein). The 'x' in the motifs primary structure represents positions that are not highly conserved and can be substituted for by any amino acid without compromising the function of the R protein.

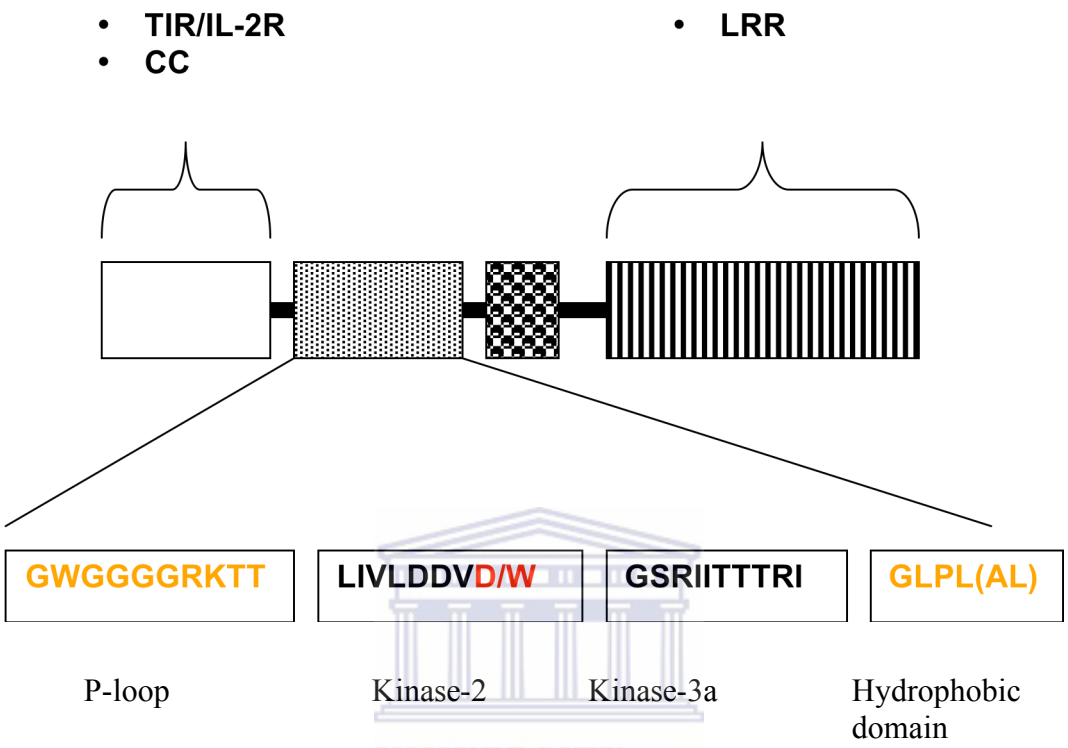


Figure 1.4. A simplified diagram of the motif organisation in the NBS domain. There are three highly conserved motifs P-loop, kinase-2 and kinase-3a, the hydrophobic domain separates the NBS domain from the leucine-rich repeats (LRR). D/W represents the last amino acid residue of the kinase-2 which is either D or W for TIR or non-TIR respectively.

The NBS domain also contains such conserved motifs as the NKxD, DxxG and (C/S)Ax that are predicted to participate in the binding of either ATP or GTP. The NBS domain of the mammalian Apaf-1 and its nematode functional homolog, CED-4, bears a striking resemblance to the plant R proteins (Van der Biezen, 1998; Dinesh-Kumar, 2000;

Belkhadir, Subramaniam and Dangl, 2004). These two proteins bind ATP (or dATP) in the presence of cytochrome c (cyt c) resulting in the induction of apoptosis (Jiang and Wang, 2000). The NBS domain is well conserved in functional homologs such as Apaf-1, R proteins and CED-4 hence is commonly referred to as the NB-ARC (**N**ucleotide **B**inding adaptor that is shared by NOD-LRR proteins, **A**paf-1, **R** proteins and **C**ED-4) domain (McHale *et al.*, 2006). Molecular modelling of the ARC domain of plant NBS-LRR proteins based on the crystal structure of Apaf-1 suggests that there are two sub-domains, the N-terminal helical ARC1 and an ARC2 that is a C-terminal winged helix sub-domain (Albrecht and Takken, 2006; McHale *et al*., 2006).

It is believed that due to this structural homology, the RGA NBS domain in plants might act as an intra-molecular signal transducer or molecular switch. It has several conserved motifs that are characteristic of the 'signal transduction ATPases with numerous domains', STAND family of ATPases (McHale *et al.*, 2006). The STAND family of proteins has been known to have members that function as molecular adaptors in the disease signalling pathways. The LRR region recognises pathogen-derived ligands and thereafter the NB-ARC generates energy for a conformational change that triggers a cascade of downstream pathways to activate the defence responses (DeYoung and Innes, 2006). One example of the NBS-LRR protein oligomerization mechanism was observed in the tobacco N protein in response to pathogen elicitors (Mestre, 2006).

The exact progression of events is not known and other models with supporting explanations are available as will be shown later in this review. The model that proposes

structural modification of the NBS domain following sensing of virulence factors by the LRR is supported in part by the function of Prf in the Prf/Pto signalling pathway in tomato resistance to the bacterial speck disease (Fredrick, 1998). Pto is a serine-threonine kinase that requires Prf, an NBS-LRR type protein for its function. Recent evidence however, refutes this model and rather reverses the sequence of events as regards the function of bound nucleotides in pathogen recognition. Work done by Dodds (2006) with the flax L6 protein and the yeast two-hybrid system suggests that the presence of a bound nucleotide is required before the L6 protein can adopt a recognition competent conformation. Whether this recent piece of evidence strengthens or weakens the proposed function of plant R proteins as molecular switches still needs to be closely investigated as more advanced protein studies become possible.

1.4.1.2 Functions of the NBS-LRR proteins

NBS-LRR proteins with a Toll/IL-1R-like N-terminal domain generally use a downstream PAD4 and EDS1 (enhanced disease susceptibility) signalling protein for signal transduction whereas the class with a coiled-coil N-terminal domain employs the NDR1 (non-race-specific disease resistance) proteins (Aarts *et al*., 1998; (Hu *et al.*, 2005). Aarts *et al*. (1998) conducted mutation experiments on the EDS1 protein and were able to show diminished activity conferred by RPP proteins and not RPM1. Although other genes from the RPP locus are affected by mutations in the NDR1 protein indicating some degree of cross reactivity, overall the differential requirements for signalling molecules is thought to suggest the existence of two quasi-independent pathways specifying resistance to clusters of pathogens (Aarts *et al.*, 1998).

The C-terminal region of the LRR domain has been implicated in pathogen *Avr* recognition function although to date no direct physical protein-protein interactions have been proved satisfactorily to support this model. In NBS-LRR proteins that possess an extra-cytoplasmic LRR, they have been predicted to possess a β-sheet structure with hyper-variable regions that are subject to diversifying selection pressures (Halterman and Wise, 2004; Van der Biezen, 1998). This supports their function in specific pathogen recognition and binding of the potentially highly variable *Avr* gene products of the pathogen.

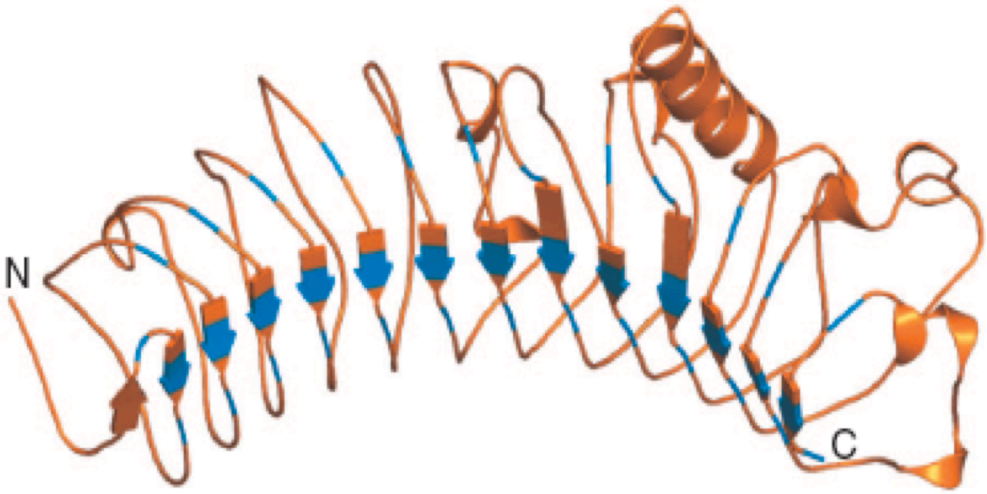


Figure 1.5. Predicted structure of the LRR domain of NBS-LRR proteins. This structure was generated by threading the RPS5 LRR onto bovine Decorin (PDB code lxku). ß-sheets are represented as arrows; conserved aliphatic residues are shown in blue (McHale *et al*., 2006).

In its native structure, the conserved hydrophobic leucine residues of the LRR domain project into the hydrophobic core of the folded protein, whereas the hydrophilic amino acid residues in the spaces between the conserved leucines form a solvent exposed surface that is believed to function in ligand recognition and binding (Mondragon - Palamino *et al*., 2002; Van der Biezen, 1998).

Studies carried out on the *Solanum tuberosum* and *Capricum annuum* proteins Rx and Bs2 respectively proved that expression of protein fragments of either CC-NB-ARC plus LRR or CC plus NB-ARC-LRR emulate the function of the full - length molecule CC-NB-ARC-LRR (Moffett *et al*., 2002; Leister *et al*., 2005). These fragments undergo physical intra-molecular interactions between ARC and LRR domains, which in the case of the Rx protein is disrupted in the presence of the PVX coat protein (CP). Rairdan and Moffett (2006) suggest that this disruption plays a role in the function of R proteins. Results from these experiments also showed that this disruption coincides with signal initiation and might be required for multiple rounds of recognition leading to signal amplification.

1.4.2 Receptor-like kinases (RLKs)

Searches on TAIR (The *Arabidopsis* Information Resource) reveals 1748 loci and 2442 distinct gene models for kinases and only 8 loci and 20 distinct gene models for receptor-like kinases (RLKs). Where a genetic locus here is a positional unit on the chromosome that segregates as a single or quantitative trait and a gene model is defined as any gene product from computational prediction, mRNA sequencing or genetic characterization (Rhee *et al.*, 2003). These represent a super-family with various members involved in diverse physiological functions. In animals, RLKs function predominantly as growth factor receptors that regulate developmental processes in homeostasis (Becraft, 2002). Most of the known animal receptor protein kinases amplified from genomic DNA using a set of degenerate primers targeted at the kinase active site have been shown to be predominantly tyrosine kinases. Currently in plants no classical tyrosine kinases have

been cloned to date. Genome-wide analysis of sequence data in *Arabidopsis* revealed that all serine/threonine kinases in plants probably belong to a dual specificity STY kinase family (Rudrabhatla *et al.*, 2006). Phosphorylation at the tyrosine residues seems to be carried out by STY (serine/threonine/tyrosine) class of kinases exemplified by the AtLecRK2 (salt-inducible ethylene receptor kinase) that was shown to autophosphorylate on the serine/threonine and tyrosine residues (He *et al.*, 2004) and the SERK (somatic embryogenesis receptor-like kinase) that also phosphorylates serine, threonine and tyrosine residues (Mu *et al.*, 1994; Rudrabhatla *et al.*, 2006).

The known functions of RLKs in plants include phytohormone responses, reproduction (self-incompatibility), developmental regulation and plant defence systems among others. There is a possibility that this range of functions will grow as more of these genes keep being discovered. The RLK family has members with variable extracellular domains (ectodomains), such as transmembrane domains and cytoplasmic protein kinase catalytic domains among others (Becraft, 2002).

The *Arabidopsis* model has shown different RLK families and other plant genes and others are still being discovered. Functional homologues of animal kinases have also been shown thereby suggesting a possible common ancestry. There are however, a few *Arabidopsis* RLK proteins that carry ectodomains that are unfamiliar to other eukaryotes such as thaumatin or light repressible RLKs LRPK (Champion *et al.*, 2004). A few examples are discussed briefly in the following text (Torii *et al.*, 1996; Clark *et al.*, 1997; Li and Chory, 1997; Jinn *et al.*, 2000).

1.4.2 LRR-RLK

This subgroup represents one of the most diverse members of the RLK class with an extracellular leucine rich repeat (LRR) domain. This domain has been implicated in a number of protein-protein interactions although in most cases this has not been proven. Proteins with the LRR domain have been implicated in a number of developmental processes. These include ERECTA for the regulation of organ shape, CLAVATA1 controls differentiation of cells at the shoot meristem, HAESA for the floral abscission processes and BRI1 that is involved in the brassinosteroid perception (Torii *et al.,* 1996; Clark *et al.,* 1997; Li and Chory 1997; Jinn *et al.*, 2000). The rice Xa21 resistance gene that confers resistance to *Xanthomonas oryzae* pv. *oryzae* also belongs to this group of RLKs. Other sources place the tomato *Cf* gene in this subfamily although it lacks the cytoplasmic protein kinase catalytic domain.

1.4.2 (b) CR4 class

This group is made up TNFR (tumor necrosis factor receptor)-like repeats linked to amino acids that display a loose similarity to the RCC GTPase group. This group contains the maize CRINLY4 (CR4) gene product that plays a role in the normal cell differentiation of the epidermis (Becraft, 1996).

1.4.2 (c) PR5

Arabidopsis PR5K is the most commonly known member of the group of PR5 RLKs. It is made up of a protein with sequence similarity to the pathogenesis response protein 5

(PR5). PR5 is induced under pathogen attack and as such this group of RLKs is thought to belong to the PR group of proteins.

1.4.2 (d) WAKs (EGF-)

The wall-associated kinases (WAKs) represent a transmembrane class of RLKs that is made up of an N-terminal EGF (epidermal growth factor)-like structure, Ca^{2+} - binding EGF repeat (EGF-Ca^{2+}) and EGF2-like domains (He *et al.*, 1999; Verica and He, 2002; (Decreux and Messiaen, 2005). The EGF-like repeats are most likely involved in protein-protein interactions due to the presence of the extracytoplasmic cysteines, which probably form conserved disulphide bonds. This conformation has been found in a variety of animal extracytoplasmic receptor domains. The *Arabidopsis* WAK1-5 have been shown to have these extracytoplasmic EGF repeats and as such have been implicated in disease resistance (Decreux and Messiaen, 2005).

In *Arabidopsis,* most of the WAKs are serine /threonine kinases (STKs) (Shiu and Bleecker, 2001) and are found covalently linked to cell wall pectin hence the close association with the cell wall. This association requires enzymatic digestion, boiling or boiling detergent and reductants to dissociate. Whether this linkage is a physical stabilisation association or that it may play a role in their functionality remains unknown. However, WAKs provide a direct physical connection between the cell and the extracellular matrix (ECM) forming a critical link that allows the cell to have a clear perception of its environment and thus enable it to adapt to its immediate surroundings.

WAKs are associated with a number of different functions in both animals and plants; in the latter they are involved in development, hormone perception, cell expansion, sporophytic self-incompatibility and disease resistance. Antisense suppression of *Arabidopsis* WAK2 and 4 (Wagner and Kohorn, 2001; Lally *et al.*, 2001) was shown to halt cell expansion and result in the production of dwarf plants (Kohorn *et al.*, 2006). However, whether or not the antisense constructs under the dexamethasone promoter only managed to suppress these 2 WAKs is debatable. Kohorn *et al*. (2006) produced a mutated form of WAK2 (wak2-1) and used it to trace events in cell expansion; the mutant had the effect of lowering vacuolar invertase expression. This confirms unlike the first experiment that WAK2 in some way regulates vacuolar invertase expression and is linked to levels of solutes in the vacuole which affect the maintenance of cell turgor pressure hence cell expansion (Koch, 2004; Kohorn *et al.*, 2006).

The WAK1 transcript was found to be inducible by methyl jasmonate, ethylene (Schenk *et al.*, 2000) and salicylic acid (He *et al.*, 1998) and further experimentation showed that it is essential for the survival of plants in a medium containing salicylic acid. Whether or not there is a link between WAK1 and the TIR-NBS-LRR signal transduction pathways leading to systemic acquired resistance still needs to be investigated. However, WAK1 seems to play the role of a caretaker protein that ensures cell survival after release of the secondary level defence signal molecules. WAK1 was also found to be essential for the survival of plants following *Pseudomonas syringae* (Zhang *et al.*, 2005) and *Alternaria brassicicola* (Schenk *et al.*, 2000) infection.

WAKs occur in multiple gene clusters in the plant genome (Verica and He, 2002) and are believed to be the result of gene duplications (Zhang, 2005). Broadly the two gene systems occur in clusters with members that can be grouped according to nucleotide sequence similarities. Some of the genes such as WAK1 – 5 were found to occur in a tandem array (He *et al.*, 2002).

1.4.3 eLRR-TM

Extracytoplasmic-LRR-transmembrane, eLRR-TM is a class of resistance proteins that is made up of an extracellular leucine rich repeat and a putative transmembrane domain. Members of this class are broadly classified under receptor-like proteins (RLPs) and include Vf and the tomato Cf proteins that confer resistance to apple *Venturia inaequalis* and tomato *Cladosporium falvum* respectively (Belfanti *et al.*, 2004). The structure is made up of a cysteine-rich region separated from the transmembrane domain by a stretch of approximately 27 leucine repeats (Kruijt *et al.*, 2005). Members of this class of R genes detect race specific elicitors secreted by the pathogen and activate an HR system that stops the spread of infection. These genes occur in clusters of tandem repeated homologs.

1.5 Resistance Mechanisms

1.5.1 Pathogen Recognition

The plant – pathogen interactions as shown earlier are mediated by specific mechanisms that result in either infection and subsequent disease progression or suppression of the disease by restricting the invading pathogen to the site of entry. To ensure broad-spectrum pathogen recognition the plant genome encodes a large repertoire of R-genes that are either induced or constitutively expressed. The former set of genes should be capable of recognising a wide range of phytopathogens in order to afford a plant the necessary protection. The presence of pattern recognition receptors, TIR or CC whose functional homologs in animal systems allow recognition of pathogen associated molecular patterns (PAMPs) have been shown to recognise lipopolysaccharides (LPS) in dicot plants (Desaki *et al.*, 2006). The Toll-like receptors also serve as non-self recognition receptors and are widespread in animal innate immunity systems, needless to say the patterns they recognise should not share conserved motifs with compounds endogenous to the defending system (Brunner *et al.*, 2002).

How do the NBS-LRR type proteins recognise the presence of invading pathogens? Recent work done in *Nicotiana benthamiana* using *Arabidopsis* RPS5 (a CNL type R protein) showed among other things that the LRR region inhibits constitutive signalling of programmed cell death (Ade *et al.*, 2007). According to Ade *et al*. (2007), a phosphorylated form of PBS1, which is the primary target of *Pseudomonas syringae* AvrPphB proteolytic activity, is always bound to the CC-NBS complex together with a

molecule of ADP. *P. syringae* secretes AvrPphB during infection that targets and cleaves PBS1. Cleavage of PBS1 results in a conformational change in the complex and cleavage products possibly bind the LRR domain. This in turn removes the LRR inhibition of the NBS domain thereby allowing it to exchange ADP for ATP resulting in activation of RPS5. A similar model was worked out for TIR-NBS-LRR proteins, making this the current working model in plant inducible defence systems mediated by NBS-LRR proteins.

Whether this model eliminates the alleged role of LRR 'receptors' as primary pathogen recognition systems is still a matter that requires more experimentation. Nevertheless, R-genes with a multiplicity of recognition specificities for defined *Avr* genes are being isolated in various plants (Odjakova, 2001). Mutation studies, especially gain of function mutations have provided evidence for the requirement of specific amino acid residues in the LRR domains to achieve resistance. In tomato *Mi*, an intracellular CNL class R-gene, mutation studies of amino acids 984 – 986 showed that they are part of a conserved functional motif that has a central role in nematode recognition (Hwang and Williamson, 2003).

Magnaporte grisea Avr-Pita, a zinc metalloprotease was shown to interact directly with the intact LRR domain encoded by the rice Pi-ta R protein (Fields and Song, 1989). A mutated Pi-ta protein which does not confer resistance did not interact with this elicitor thus giving the only evidence to date that direct binding to the LRR is essential for biological function (Bryan *et al.*, 2000; Belkhadir *et al.*, 2004).

It has been shown for Apaf-1 in animals system that binding of the WD-40 repeat (C-terminal repeat) domain to the NBS-ARC domain prevents spontaneous self-association thus negatively regulating the function of Apaf-1 (Hu *et al.*, 1998; Hu *et al.*, 1999). Binding to ATP/ dATP in the presence of cytochrome c induces a conformational change that promotes homooligomerization and subsequently the formation of an apoptosome. If this model is used as an analogy to the R protein system, then binding of the LRR domain to the NBS-ARC maintains the complex in an inactive form. Thus an association between an Avr protein and the LRR domain distorts its conformation, which in turn allows the NBS domain to bind ATP triggering self-association to form a signalosome similar to the apoptosome in the case of Apaf-1 (Rairdan and Moffett, 2006). This allows the NBS-LRR proteins to detect pathogen invasion and trigger the HR response.

These two models as exemplified here by the RPS5 and Pi-ta systems might explain the roles played by the N-terminal domain (TIR/CC) and the LRR domain in pathogen recognition and subsequently the activation of defence systems. The three examples used above describe mechanisms observed in *P. syringae* (a gram negative bacteria), *Magnaporte grisea* (a fungus in the Ascomycota phyla) and nematode interactions (tomato Mi proteins), AvrPphB, Avr-Pita and the *Mi* R gene systems respectively. The involvement of the CC and TIR systems were proven for RPS5 and RPP1A for bacterial and viral defence systems respectively (Ade *et al.*, 2007). The "Guard Hypothesis" model supports the pre-association of an endogenous molecule and R genes and explains how

targeting this particular molecule by pathogen *Avr* factors results in activation of the defence system (McDowell, 2004; McHale *et al.*, 2006).

1.5.1.1 Guard Hypothesis

This hypothesis proposes the existence of an endogenous group of substances called 'guardees'; these are basically virulence factors that would act as primary targets of pathogen-derived proteins (Tornero *et al*., 2002; McDowell and Woffenden, 2003; Wang, 2005). Upon infection, the pathogen targets this group of substances, guardees, which they would modify to create an enabling environment in order to establish the infection state. R proteins are thought to guard these proteins and any modification thereof would thus trigger defence responses.

This model was used to explain the RPM1 mediated resistance in which it detects phosphorylation of the RPM1-interacting Protein 4 (RIN4) by the elicitors AvrB and AvrRpm1 from *Pseudomonas syringae* pv. *glycinea* and *maculicola,* respectively (McHale *et al*., 2006). According to literature reviewed in McHale *et al*. (2006) AvrRpt2, a protease that constitutes a second elicitor from *P. syringae* pv. *tomato* cleaves RIN4 thereby arresting the effects of RPM1. However, RPS2, which is a second CNL R protein detects the disappearance of RIN4 and elicits a defence response. There is a growing body of evidence that supports the 'guard hypothesis' as a mechanism used in the interaction of R proteins with their corresponding elicitors (McHale *et al*., 2006 and references therein). The RPS5-AvrPphB interaction reviewed in section 1.5.1 above is another example of this model.

Most of the interactions in which a 'guardee' has been identified show a confirmed involvement of the N-terminal domain, either the TIR or the CC domain. The LRR interactions that follow this model have not been worked out conclusively. The lack of detectable protein-protein interactions involving the LRR domain is currently being viewed as evidence for the requirement of multiple interdependent associations leading to a successful activation of HR in plants (Bogdanove, 2002).

1.5.1.2 Non-host recognition

Non-host resistance defines an interaction in which all varieties of a plant species are resistant to all strains of a particular pathovar through the use of a universal resistance mechanism. This differs from the intraspecies specific variability observed in R-gene-mediated resistance. Non-host resistance makes use of the recognition of specific exogenous elicitors or plant cell wall derived (endogenous) elicitors (Odjakova, 2001) such as chito-oligomers from chitin that is ubiquitous in fungi and thus is thought to induce this class of resistance and possibly basal defences. Other elicitors of this nature could be glycoproteins or other substances derived from enzymatic activities on preformed precursors such as hydrolysis of polymeric substances (Nurnberger *et al.*, 2004; Nurnberger and Lipka, 2005).

1.6 Signal Perception

Most pathogen infections in plant either via stomatal openings or otherwise proceed to establish an infection state in the apoplastic tissues. In gram-negative bacteria infections e.g. *Pseudomonas, Erwinia, Xanthomonas and Ralstonia* an infection proceeds through the bacterial type III secretion system (TTSS). These TTS systems enable pathogens to inject products of the *hrp* (HR and pathogenicity) genes into the cytosol through the plant cell wall, which in turn promotes host plant susceptibility to the given infection (Collmer *et al.*, 2000; Staskawicz *et al.*, 2001; Hauck *et al.*, 2003). The TTS systems of DC3000 (*P. syringae* pv. tomato strain) for example is said to secrete more than 30 effector proteins in the host cell and their cumulative effect is intended to alter cellular processes thus allowing the pathogen to multiply and accumulate in the apoplastic regions (Hauck, 2003; Abramovitch, 2003). However, though the primary function of these effector proteins is to promote host susceptibility, they may also trigger defence responses when detected by the plant surveillance system in surface exposed and cytosolic receptors.

Studies carried so far have shown that the downstream signal transduction pathways are highly branched, partially redundant, and partly contain feedback loops and homeostatic control of defence related proteins (Iunes, 1998; Peart, 2002). Dissecting and analysing the functions of each of the motifs contained in the NBS domain and both functional and structural homology to the human Apaf-1 and nematode CED-4 systems have provided insight into some of these mechanisms.

1.7 The Hypersensitivity Response

The inducible plant disease resistance is mediated by a system of R proteins each of which has receptors that either match corresponding pathogen Avr proteins or detects pathogen associated molecular patterns (Odjakova and Hadjiivanova, 2001). The resultant recognition takes the form of gene-for-gene interactions whose sole purpose is to trigger a series of defence mechanisms.

Whether or not the interaction is direct or indirect, a rapid programmed death of the infected cells is activated to contain invading pathogens to the site of infection (Kosak and Jones, 1996; McDowell, 2003). The series of events culminating in programmed cell death constitutes the hypersensitivity response, which is manifested as necrotic lesions on the surfaces of leaves, stems or fruits (Krzymowska *et al.*, 2007). HR can also give rise to 'local acquired resistance' (LAR) also characterised by local tissue reinforcement and production of antimicrobial compounds among other local physiological alterations that render neighbouring cells refractory to further pathogen invasion (McDowell, 2003; Krzymowska *et al*., 2007).

The initial reactions following *R-Avr* formation comprises the changes in plasma membrane permeability leading to Ca^{2+} and H_3O^+ influx and K^+ and Cl^- efflux (McDowell, 2000). This pattern of ion fluxes induces the production of reactive oxygen intermediates, ROIs, ($\bullet O_2^-$ H_2O_2 and $\bullet OH$) catalysed by membrane-linked NADPH oxidase and/or apoplastic-localised peroxidases (Zimmermann *et al.*, 1997; Somssich and Hahlbrock, 1998; Delledonne *et al.*, 2001; Mittler *et al.*, 2004). ROS have been shown to

possess direct antimicrobial activity and also to function in cell wall reinforcing processes in an HR response. Furthermore, they are believed to activate defence gene expression and induce the long lasting systemic acquired response, SAR, (Hoeberichts, 2003). This dual role of ROIs is dose dependent with high doses triggering HR-related programmed cell death (PCD) whereas low doses induce antioxidant enzymes and block cell-cycle progression (Hoeberichts, 2003).

All these early reactions trigger a systemic acquired response, SAR, which is long lasting and acts to prime the whole plant for a more generalised response against a broad spectrum of pathogens (Dong, 2001). This delayed defence system is characterised by elevated levels of endogenous salicylic acid (SA), further production of ROIs and other unstable radicals, there is also an increased PR-1 and PR-2 gene expression (Cutt and Klessig, 1992; Malamy and Klessig, 1992; Dorey *et al*., 1997; McDowell and Woffenden, 2003).

1.8 PR gene expression

In addition to R genes and other genes encoding signal transduction proteins, plants also have downstream defence genes encoding among other things pathogenesis-related, PR proteins (Odjakova, 2001). More than eleven PR gene families have been characterised in different plants with various anti-pathogenic activities. These PR gene systems generate a range of secondary metabolites and other substances essential in plant defences. Some of the PR proteins are enzymes involved in the generation of phytoalexins, tissue repair, lignification, and oxidative stress protection.

Different pathogenesis-related gene families function in various biological activities though generally they all secrete their effector proteins into either the intercellular space (acidic proteins) or the vacuole (basic proteins). The activities of PR-1 remain elusive although it has been found to target the pathogen's plasma membrane. PR-2 has 1,3-ß-glucanase activity and thus targets cell wall glucan and chitin, which occur in cell walls of most higher fungi (Odjakova, 2001). Activation of the PR system and the subsequent accumulation of pathogenesis-related proteins represent the single major quantitative change in protein composition in non-inoculated plant parts that upon challenge exhibits acquired resistance (Odjakova, 2001).

1.8.1 Caspase-like PCD

Programmed cell death in plants serves a number of physiological functions from selective removal of reproductive structures to destruction of infected cells. This phenomenon is present in animals though research for parallel systems in plants is now focused on deciphering the differences and similarities between these pathways. In animals there is generation of ROS and other compounds and upregulation of the ubiquitin-proteosome pathway leading to cytoplasm condensation and regression, chromosomal degradation and finally cell death. Studies in tobacco leaves have also detected ROS production, cyt c release and caspase-3-like protease activation in the early phases of heat shock induced PCD (Vacca *et al.*, 2007).

According to Vacca *et al*. (2007) the increase in ROS production is a consequence of the impaired oxidative metabolism in the mitochondria and other cellular components. The antioxidant enzymes and proteosome are then evoked so as to maintain ROS. There is associated release of cytochrome c for the activation of a putative caspase-like cascade.

1.8.2 MAPK and defence signalling

In animal and yeast cells, mitogen-activated protein kinase (MAPK) cascades form a crucial link that transduces stimuli from extracellular receptors (Zhang, 2001). Up to 20 MAPKs have been observed in *Arabidopsis*. The characteristic Thr-Glu-Tyr (TQY) activation motif has been identified in most MAPK families except members of the subfamily V (Zhang, 2001). These molecules have also been identified in plants including the tobacco salicylic acid-induced protein kinase (SIPK) using the yeast two-hybrid system (Liu, 2000). There are also others such as the NtMEK2 (tobacco MAPKK) upstream of SIPK, wounding induced protein kinase (WIPK) and the alfalfa salt stress-induced mitogen activated protein kinase (SIMKK).

In the systems studied thus far, MAPKKs are activated by phosphorylation of serine/ threonine residues of the S/TxxxS/T conserved motif by MAPKKKs. This means that several unrelated MAPKKKs could activate the MAPK cascade; this scheme explains how signals perceived by a plethora of divergent MAPKKKs converge to a MAPK cascade (Zhang, 2001). Components of the plant MAPK cascade have been defined through homology to animal and yeast systems. In general these pathways in plants have

been described for such processes as cytokinesis (Bögre *et al.*, 1999) and phytohormone signalling (Zhang, 2001).

In animal and yeast systems, activation of the MAPK cascade leads to the phosphorylation of transcription factors. Although no plant based substrates of MAPK have been defined to date, translocation of MAPK into the nucleus of *Petroselinum crispum* cells after treatment with Pep25 elicitor (Ligterink, 1997) give support to its role as predicted in comparison with animal and yeast systems.

The MAPK cascade although not defined conclusively in plants, seem to be the signal transduction pathway in the plant defence signalling. The MAPK cascade and related pathways are represented in the diagram in Figure 1.6.

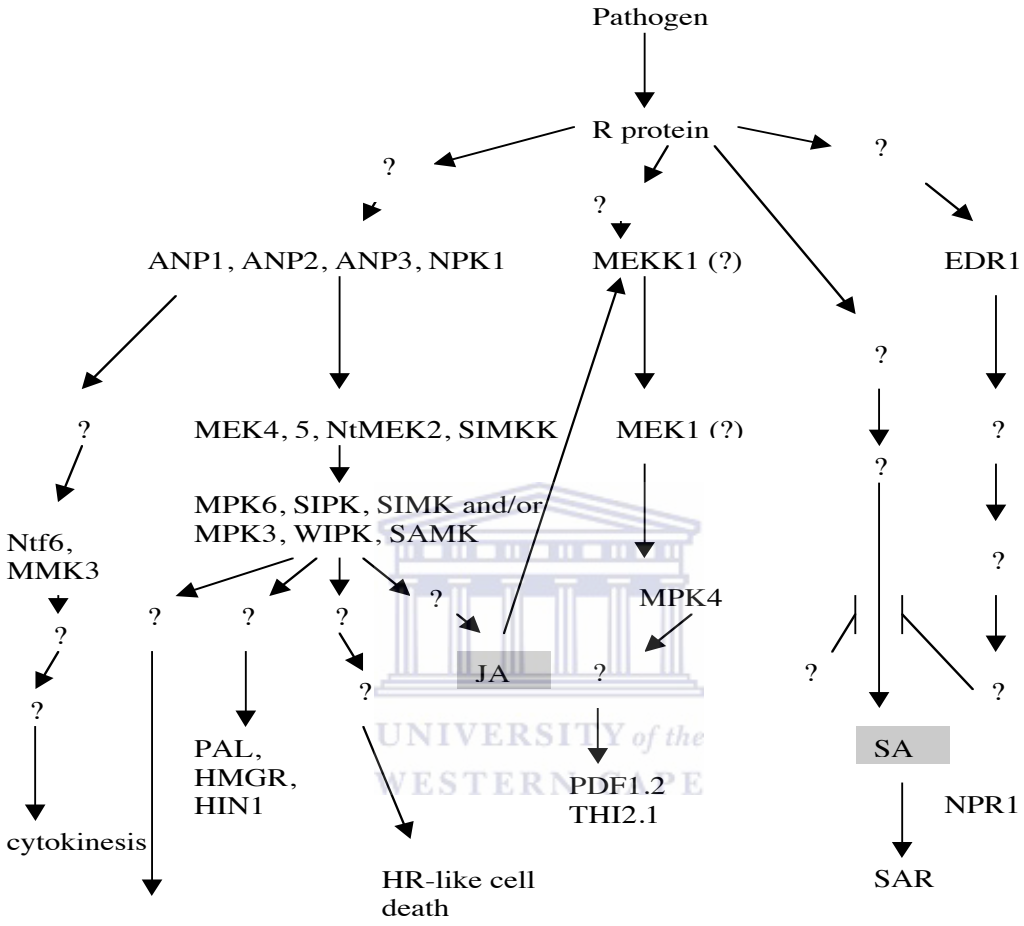


Figure 1.6. Signal transduction cascade involving various MAPKs in plants. Details used to construct this diagram were obtained from reviews and publications (Zhang and Klessig, 2001; Daxberger *et al.*, 2007; Takahashi *et al.*, 2007).

1.9 Systemic Acquired Resistance (SAR)

Systemic acquired resistance (SAR) in plants is a broad-spectrum pathogen defence mechanism that primes the whole plant against subsequent infection. It is characterised by an elevated expression of pathogenesis-related proteins. There is also accumulation of a number of compounds such as ß-1-3-glucanases and chitinases (Plymale *et al.*, 2007) that form a broad host defence system against phytopathogens. Long distance signal transmission is required to protect the whole plant is provided by salicylic acid. However, activation of the SAR has been shown to inhibit wounding responses meaning plants expressing this form of resistance become vulnerable to herbivory (Durrant and Dong, 2004; Plymale *et al.*, 2007). Section 1.9.1 below explains in detail the role of SA in the induction of SAR.

1.9.1 Salicylic acid and SAR induction

Salicylic acid is the key defence response component in *Arabidopsis* (McDowell and Dangl, 2000). Transgenic plants containing the *Nah*G gene (a gene that codes for the salicylate hydroxylase enzyme that hydrolyses SA to catechol) were compromised in their ability to activate systemic acquired resistance, SAR, (Ryals, 1996). SA was shown to stimulate the translocation of NPR1 to the nucleus where it is assumed to act as a transcriptional co-factor for the expression of PR-1 (Ham *et al.*, 2007). Induction of the PR gene systems, notably PR-1 and PR-5 has also been shown to occur as a result of the accumulation of SA (Krzymowska *et al.*, 2007).

It is clear that SA plays a significant role in inducing downstream expression of pathogenesis related genes. SA was detected in the phloem and studies using Carbon-14 Labeled benzoic acid showed proved not only the translocation but also the subsequent induction of PR genes in other parts of the plant other than the one infected (Molders *et al.*, 1996). It appears that SA plays a significant role both in the activation of local leaf necrosis hence localisation of the infection to the infected part and establishing acquired resistance in other parts of the plant. Other experiments also provide evidence of salicylic acid transportation through the translocation of ^{18}O-labelled SA in TMV infected tobacco plants (Shulaev *et al.*, 1995; Rocher *et al.*, 2006).

SA accumulated following infection also converted to salicylic acid β glucosidase, SAG, its conjugate form. Although not proven to date, SAG might act as a stored chemical messenger present in primed plants. Subsequent attacks could cause disruption of cell walls thereby releasing β glucosidase that cleaves SAG freeing SA, which allows a faster and more potent defence response.

1.10 WRKY superfamily of plant transcription factors

The WRKY group of proteins were first defined in sweet potato, wild oat, parsley and *Arabidopsis*. In all the cases they were seen to bind specifically to the DNA motif characterised by the signature sequence [(T)(T)TGAC(C/T)] known as the W box (Eulgem *et al.*, 1999; Eulgem *et al.*, 2000). Controlled infection or treatment of plants with SA induced a rapid WRKY gene expression in a variety of plants (Eulgem *et al.*, 2000). This shows that WRKY proteins are an integral part of the plant defence system.

These genes provide transcriptional regulation of genes involved in defence related pathways. Evidence that implicates this group of transcription factors as inducers defence-related gene expression has been steadily accumulating in literature (Eulgem, 2006; Li *et al.*, 2006; Ulker *et al.*, 2007).

PR gene systems and the regulatory NPR1 genes were shown to possess typical W-box elements in their promoter regions (Zheng *et al.*, 2007). Although this could be viewed as circumstantial evidence, more direct evidence of compromised disease resistance following suppression of WRKY gene expression has also been provided in recent years (AbuQamar *et al.,* 2006). Virus-induced silencing of WRKY proteins in tobacco was shown to reduce the resistance function of the tobacco N gene against TMV (Liu *et al.*, 2004).

In parsley WRKY1, WRKY2 and WRKY3 were shown to specifically bind W boxes designated W1, W2 and W3 contained in the promoter regions of PR genes encoding the PR-10 class of proteins. Qualitative analysis showed that these factors are transiently expressed following fungal infection and that their production is rapid and localised at the point of pathogen entry (Eulgem, 1999).

1.11. Generation of Specificity in the NBS-LRR type R proteins

1.11.1. RGA Evolutionary Dynamics

Plants are exposed to a myriad of pathogens and given their lack of a rapid circulatory system; it means each cell has to be equipped with a set of defence-related genes. For survival of the species, plants have to encode a pro-active system of defence that scouts for and detect pathogen-derived elicitors. This requires a robust and dynamic system that is capable of directed evolutionary regulation of its receptors. It is becoming clear that resistance genes occur in clusters on the genome and that paralogs in each cluster evolve by cross-over events, recombination and gene conversion (Richter and Ronald, 2000).

An analysis of these clusters using the ratio of coding to silent mutations (Ka/Ks ratio) show a high level of positive selection for expressed functional R-genes (Botella *et al.*, 1998; Sun *et al.*, 2006). Other methods of analysis targeting recombination events and gene conversion have also supported the existence of constant evolutionary drive in functional clusters. This among other things suggests that the current set of plant defence genes discovered thus far, notably the NBS-LRR and RLKs are constantly being refined and that the classical Darwinian 'survival of the fittest' phenomenon defines the dynamics observed in these gene families.

1.11.2.1 Adaptive divergence

Analysis of NBS-LRR paralogs in gene clusters show evidence of intergenic exchange, as one of the mechanisms driving gene evolution. However, intergenic exchange cannot be the only mechanism to explain the generation of allelic variations (Bergelson, 2001; (Michelmore and Meyers, 1998). Nevertheless, NBS-LRR clusters have been shown to harbour a large number of pseudogenes as a reservoir of useful sequence variation. These genes do not code for functional genes and have been shown to mutate at a rate higher than normal (Michelmore, 1998), whether these are driven by challenges present in the environment or the observed mutations are just random is not clear at the moment. What is clear, however, is that a large repertoire of paralogs in one given gene cluster has the potential to increase the likelihood of generating novel sequence patterns through exchange thus exploiting clustering as a means to achieving important allelic variations.

Observations made by Bergelson *et al.* (2001) in rates of adaptive evolution shows that genetic exchange contributed to the generation of new adaptive alleles. This observation also justifies the school of thought that suggests that 'selection' has been the driving force behind the generation of allelic polymorphisms. This assumption presupposes the continual challenge of genomes with ancestral pathogens (carrying ancestral *Av*r-genes) in a cyclic fashion, if functional/ specific alleles are to remain effective. Selection brings about the spread of new and novel resistance alleles in hosts thereby reducing the frequency of ancestral R-alleles (frequency-dependent selection) and in the absence of a continual challenge, isolated infections might result in serious outbreaks.

These evolutionary dynamics are observed in the LRR domain polymorphisms. However, to better understand them, more genetic data is needed together with short-term disease dynamics (Bergelson *et al.*, 2001; Moore and Purugganan, 2005). However, results from mutagenesis or transgenic complementation analyses show that in most cases resistance is a function of a single dominant gene in a cluster, the 'dominant functional gene' theory (Michelmore and Meyers, 1998).

If generation of allelic polymorphisms were ascribed to intergenic exchange as a means to keep up with pathogen evolution then high levels of polymorphisms would be expected between orthologs. This is based on the assumption that R-genes occur as haplotypes (Michelmore and Meyers, 1998). Comparisons between paralogs and orthologs in the NBS-LRR gene family have shown that orthologs are more homologous compared to paralogs, which is the reverse of the expectations under the above premise.

Michelmore and Meyers (1998) propose an initial involvement of duplication and intergenic exchange following meiotic recombination events, the adaptive nature of the NBS-LRR genes then occurs through substitutions and deletions; unequal crossing-over occurs at a low rate which might not be enough to homogenise sequences. A new model that involves the interplay of illegitimate recombination (IR) has been suggested to explain recombination in the highly divergent LRR domain (Wicker *et al.*, 2007). In this model the presence of short homologous fragments (or motifs) that flank a highly divergent region is required to produce either random duplication or deletion events.

1.11.2.2 Neutral theory of Molecular Evolution

Under the neutral theory, polymorphism is regarded as a transient phase of molecular evolution (Hudson *et al.*, 1987). The evolutionary history of defence systems is shaped by the necessity to either recognise all forms of *Avr* specificities or evade recognition and escape infection. However, since the later case is highly unlikely due to the presence of non-host defence systems, theories that address survival and maintenance of species have to address non-synonymous substitutions that bring about new receptor binding specificities. The theory of neutral evolution is in contrast to natural selection in shaping the course of adaptive evolution as put forward by Darwin, in this theory random genetic drift is regarded as a result of the cumulative effects of neutral or near neutral mutations under continued input of new mutations (Kimura, 1991).

Kimura's theory envisions three types of mutations, neutral in which the amino acid sequence is unaltered and thus function remains unaffected, deleterious mutations that are subsequently eliminated by selection and beneficial mutations, which are very rare enough to be neglected. Effectively this means that only neutral evolution is observed.

It has been shown that resistance systems rely on the advantage of rare allele forms and that this prevents the loss of alleles by genetic drift thus leading to the maintenance of sometimes very old alleles (De Mita *et al.*, 2006). In a departure from neutral evolution on a population level, directional selection that entails a rapid accumulation of non-synonymous substitutions leading eventually to selected synonymous mutations may produce a selective sweep that induces hitchhiking effects around the locus under

selection (Barton, 2000). This is often observed in defence systems when plants are challenged by pathogens. On a species level these changes remain neutral given that they are episodic and are restricted to limited sites on the protein sequence.

1.11.3. Birth-and-Death Model of RGA evolution

As has been indicated above, Michelmore and Meyer (1998) first proposed the 'birth and death' model as an attempt to explain the generation of R-gene specificities. Features of this model include the alteration of specificities due to inter-allelic recombination and gene conversion in the LRR region. However, these two processes acting on paralogs are rare thus further changes are attributed to mutations acting on solvent exposed surfaces or inter-allelic unequal crossing-over or gene conversion due to mispairing.

Recent studies using the NBS domain sequences of six species within the family *Solanaceae* have provide experimental evidence of the birth-and-death model in action (Couch *et al.*, 2006). The continued random mutation and inter-allelic recombination could facilitate selection of variants with improved specificities. For this phenomenon to be successful, there has to be a repertoire of possible specificities any one of which could be passed onto purifying selection mechanisms to generate possible functional receptors. This increases the likelihood of sequence similarity, which in turn facilitates unequal crossing-over and either duplications or deletions, which have been shown to affect single or blocks of genes (Kuang *et al.*, 2004). The whole process obviously results in high levels of sequence similarity, which in turn increases instability hence subsequent rounds of duplication and deletion processes.

However, as the intergenic regions diverge the frequency of unequal crossing-over is reduced and the generated variants stabilize as haplotypes. The duplicated genes also diverge as random mutations come into play and thus Avr ligand binding specificities become altered. This in turn results in some R-genes losing their ligand binding capabilities and further developing into pseudogenes, a state where mutation rates are allowed to be high.

According to this model, genetic exchange between paralogs as a model cannot justify the frequency of generation of specificities. An example here is the flax L locus, which only has a single gene but 13 allelic specificities, which would seem to suggest that evolution of resistance occurs primarily at the single gene level (Ellis *et al.*, 1997).

1.11.4. Role of pseudogenes

Pseudogenes are present in gene families and most notably among resistance genes. Possibly they are maintained as potential reservoirs of genetic variation useful in rapid generation of specificity through genetic exchange between paralogs (Michelmore and Meyers, 1998). Functional genes could be lost or become fixed as pseudogenes in populations that are not continually exposed to a particular pathogen (Rose *et al.*, 2005; Lin *et al.*, 2007). These genes are said to evolve at a faster rate than their functional counterparts therefore could be indicators of rapid evolution of a gene family. Expression of these genes has also been detected in other organisms where their primary role is to stabilise the expression of genes encoding functional proteins (Yano *et al.*, 2004).

1.12 Differential expression of plant R genes

1.12.1 Microarrays

The combination of genomics and gene expression studies holds the key to elucidating the complex networks of gene systems at play in a variety of physiological processes including but not limited to the ones discussed in this chapter thus far. Microarrays have emerged as the gene expression analysis tool of choice for the simultaneous detection of expression profiles either for all genes in a given organism or a subset thereof. This technology, as is the case with all other powerful tools of its nature, is subject to the quality of the experimental design. Currently microarray chips can contain 385 000 cDNA probes that are 60 nucleotides long on a 17.4 x 13 mm glass slide thus covering a considerable percentage of an organism's inducible genes (NimbleGen Systems Inc.). This layout can allow comparative analysis of expression systems under two different treatment regimens at a time through the use of two fluorescent dyes; imaging of the colours on each spot thus can reveal the up-regulated or down-regulated genes per treatment. If these two results are rationalised, the technique has the power to unlock and elucidate complex gene networks underlying given physiological processes.

The disadvantage here is that errors or oversights in experimental design could have the potential to produce large numbers of genes in which some of the genes identified might be artefacts that have no direct link to the physiological process under review; or the key genes identified might be unknown. In this case the limitation becomes designing statistical analyses necessary to extract useful information from microarray data.

However, advantages to this technique far outweigh the disadvantages and as is the case with any upcoming technology, experimental designs need to be properly streamlined so as to extract the maximum possible information from its application. To date microarray studies have uncovered useful novel genes and possible alterations in pathways and mechanisms. Zimmerli *et al.* (2004) was able to show that photosynthesis and carbon metabolism-related transcripts were repressed in infected plants, and that this might actually imply a re-allocation of resources to defence responses. Rammonell *et al* (2002) used microarrays to show the involvement of non-host responses following infection of plants by chitin containing pathogens. Through controlled inoculation of two sets of plants with chito-oligomers and SA, and monitoring temporal changes in gene expression, it was observed that inoculation with the former induced signal transduction pathways that are different from those used in the gene-for-gene defence response.

cDNA and oligonucleotide probes are designed to the unique regions of a transcript and this eliminates the possibility of cross-hybridisation between closely related genes thereby allowing analysis of independent expression patterns. The fact that these are synthetic reagents means also that they can be made in bulk and so ensure reproducible results assuming good quality controls (Wullschleger and Difazio, 2003). A number of other expression patterns have been elucidated using this technique and some of them useful in the study of plant defence systems will be considered in section 1.12.2.

1.12.2 Applications of microarrays to defence systems in plants

Maleck *et al.* (2000) studied expression profiles of defence related genes under 14 different conditions relating to SAR (e.g. SA treatment, bacterial inoculations and various mutants such as altered NPR1 genes (npr1)) on *Arabidopsis* EST microarray chips. Each chip covered 10 000 ESTs representing about 25 – 30% of all *Arabidopsis* genes. Differential expression profiles where demonstrated in 413 ESTs (~300 genes) on the assayed conditions. This experiment allowed the first clustering of PR gene systems based on their regulation patterns. The genes with at least ~ 4 WRKY binding sites per promoter region were classified under the PR-1 regulon and the non-PR-1 regulated genes with promoter regions having less than two WRKY factors.

These analyses showed also for the first time that responses characteristic of compatible interactions overlap with SAR-associated gene expression (Maleck *et al*., 2000). This would mean subsequent responses within the first few hours of the initial infection are aided by a broad spectrum and rapid response. The role of different WRKY factors in co-regulation of PR-1 regulon genes was also confirmed (Rowland and Jones, 2001). There are other groundbreaking discoveries in the area of plant defence systems through the microarray technology, however these were not done in apples and as such will not be reviewed in detail here.

1.12.3 Microarrays and transcriptome analysis

Plants activate a number of genes following infection, for example a total of 149 R genes are potentially expressed in the *Arabidopsis thaliana* genome following invasion by a pathogens (Belkhadir *et al.*, 2004). These activation events are characterised by a series of transcription re-programming leading to either programmed cell death or establishment of a generalised broad-spectrum resistance mechanism and longer lasting priming of the whole plant. Monitoring temporal and spatial events in these systems with the use of proper probing techniques allowed the discovery of novel pathways in a variety of physiological challenges.

The use of expression profiling systems e.g. microarrays and high throughput cDNA sequencing have the ability to investigate thousands of gene systems simultaneously. This has the potential to give intricate detail on the defence transcriptome and allied regulatory networks (Eulgem, 2005). The interconnectivity between basal defences in R-gene systems and SAR has been established through the involvement of common signal molecules (Glazebrook, 2001; Dong, *et al*., 2001). However, though the three defence systems overlap at some point, R-gene systems and SAR tend to be more robust with strong and rapid responses compared to basal defences activated by the same signal transducers. Table 3 summarises the achievements of microarray-mediated transcriptome dissection mostly in *Arabidopsis*. In addition to assigning roles to genes the table, they were very instrumental in understanding redox systems in the HR response especially the generation of reactive oxygen species and related gene pathways (Sagi and Fluhr, 2006).

Table 3. Defence related transcription factors discovered using microarrays. This table was taken from Eulgem (2005)

Transcription factor type	Size of Arabidopsis family	Key features	Consensus core motif of binding sites	Comments
ERF	56	One ERF-DNA binding domain	GCCGCC (GCC box)	Subfamily of AP2 transcription factors; activators and repressors
R2R3 Myb	125	Two repeats of Myb domain (R2 and R3)	Type I: (T/C)AAC(T/G)G Type II: G(G/T)T(A/T)G(G/T)T	Predominating subfamily of Myb factors in plants
TGA bZIP	10	One basic DNA binding domain; leucine zipper protein dimerization motif	TGACGTGA (TGA box), this motif usually occurs as direct repeats	Subfamily of bZIPs; activators and repressors; form homo- or hetero-dimmers
NPR1	6	Ankyrin repeat domain; BTB POZ domain	No DNA binding sites	Interacts with TGA-bZIPs
Whirly	3	Whirly domain	GTCAAAAA/T	Forms homo-tetramers; Binds to single-stranded DNA
WRKY	74	One or two WRKY binding domains	(T)GACC/T (W box)	Activators and repressors

1.13 Mapping of resistance genes

There are a number of techniques that have been adopted for studying the way genes or indeed alleles are arranged on the genome. Among these is the use of molecular markers to 'tag' genes or clusters of genes encoding a given trait. The localisation of 'single' or 'quantitative' trait loci relies a great deal on the establishment of a robust linkage map that not only covers the whole genome but also offers a high resolution around loci of interest both for the purposes of positional cloning or marker assisted breeding.

To date microsatellite markers have emerged as the tool of choice in this regard. These are codominant, have conserved flanking regions, are PCR-based and offer superior portability between cultivars or indeed between species as shown in synteny mapping (Decroocq *et al.*, 2003). Studies have managed to establish the transferability of microsatellite markers between apples and pears (synteny mapping), the advantage thereof above AFLPs, RFLPs and RAPDs is the possibility of studying corresponding resistance gene loci in different species. A study carried out with *Malus* and *Prunus* NBS-LRR genes identified synteny for the genes linked to the powdery mildew resistance loci (Xu *et al.*, 2007).

Single nucleotide polymorphic sites (SNPs) can now be used to locate gene positions using the high-resolution microsatellite linkage maps. This has been facilitated by the recent development of bin mapping technologies that make use of a defined set of mapping populations. These provide resources necessary to either saturate existing linkage maps or position genes in a framework map; the critical aspect becomes the use of a defined set of seedlings in a mapping population.

1.14 Status of the *Malus* genomics

Despite all the progress made in *Malus* research, sequencing of the apple genome has not been completed to date. Current information seems to suggest that there are two groups sequencing the apple genome. These include the Department of Natural Resources and Environmental Sciences (University if Illinois at Urbana-Champaign) currently in possession of a BAC library that is already fingerprinted and ordered into contigs. The second group in the Genetics and Molecular Biology Department at the Istituto Agrario San Michele all'Adige in Italy is currently doing a 4X coverage using Sanger sequencing and 10X with the 454 Life Sciences (www.bioinfo.wsu.edu/gdr/). Release of the draft sequence from the Italian group is expected towards end of 2008 or early 2009. However, individual and isolated sequencing initiatives have managed to produce and deposit 262108 nucleotide sequences in databases including GeneBank. A general nucleotide search in GeneBank gives a total of 256280 sequences made up of 1343 core nucleotides, 254902 ESTs and 35 GSS. Currently there are 14626 gene-orientated total cluster sets of transcript sequences and genome database for Rosaceae (GDR) shows a total of 250907 ESTs after filtering and these have an average length of about 583 nucleotides.

Determination of the type and frequency of simple sequence repeats using the CUGIssr.pl program shows 46663 genes that contain one or more microsatellites with the total number of candidate SSRs found in this exercise currently at 58319 from 657 motifs. For this exercise SSRs are defined as dinucleotides repeated at least 5 times, trinucleotides repeated at least 4 times, tetranucleotides repeated at least 3 times, or pentanucleotides repeated at least 3 times.

The AutoSNP software package (Savage *et al*., 2005) was used to determine the type and frequency of single nucleotide polymorphisms (SNPs) in the GDR Unigene Assembly (v3) and 14298 SNPs were found in the 23868 contigs analysed. The frequency was calculated at 0.07/100 base pairs and the constitution currently has 7060, 3836 and 3402 total transitions, transversions and indels respectively (www.rosaceae.org). Table 4 summarises the current status of the Entrez records as accessed in March 2007.

Table 4. The current state of apple genomics. Figures in this table were quoted from GDR (www.rosaceae.org), these figures were quoted in December, 2007.

Database Name	Data available
Nucleotide	256,280
Protein	1,135
Genome Projects	1 (private initiative, Italy)
UniGene	14,626
UniSTS	28
HomoloGene	3,424

1.15 AIMS AND OBJECTIVES OF THE PROJECT

Plant resistance gene analogs (RGAs) represent a very important route to the discovery and characterisation of novel resistance genes in apples. A number of studies have confirmed that the NBS-encoding region of RGAs contains valuable sequence data that allows construction of informative phylogenetic analyses (Calenge *et al*., 2005). Not much has been done on the comparative analysis of expression profiles for cloned RGAs although phylogenetic analyses to determine the clustering pattern have been done though with smaller datasets (Lee *et al*., 2003; Baldi *et al*., 2004). Research has confirmed that RGAs constitute the best markers for resistance genes. To date they have been used as markers to identify loci containing genes of resistance to scab, Powdery Mildew (*Pl2* and Pl_{MIS}) and Woolly Apple Aphid (*Er*3) among others (Dunemann *et al*., 1996; Bus *et al.*, 2002).

The main aim of this project was to investigate the gene copy number and identify markers that are linked to NBS-LRR genes that show some expression under controlled infection of the plant. This requires that transcriptome analysis and mapping of these genes be performed though as proof of concept for future research that could lead to the discovery of novel apple resistance genes.

Specific objectives under this aim

1. To amplify candidate RGAs from Anna and Golden Delicious apple cultivars using a set of degenerate primers flanking the NBS domain of the NBS-LRR genes, clone and sequence using the Sanger sequencing system;
2. To determine the RGA clustering pattern using a range of phylogenetic tools and analyse evolutionary pressures acting on individual clusters;
3. To identify 3 clusters with the most genes, perform multiple alignments for each and design specific primers, which are then used to amplify RGAs from Anna and Golden Delicious genomic DNA. The PCR products are cloned and sequenced using the Sanger sequencing system and the data is used to identify SNPs for mapping these three clusters in the Anna x Golden Delicious Bin mapping population;
4. To amplify RGAs from Anna and Golden Delicious apple cultivars using the framework set in objective 1 and perform direct sequencing of PCR amplicons using the pyrosequencing system (454 Life Sciences)
5. To perform sequence assembly of the datasets from objectives 1 and 4 and use the results to estimate the RGA copy number in the apple NBS-LRR resistance gene family;
6. To perform two glasshouse infection trials for apple scab and powdery mildew using the Lady Williams x Prima and Carmine x Simpson seedlings respectively and collect samples before and after the inoculations;

7. To isolate RNA from samples collected in objective 6, convert it to cDNA and amplify candidate RGAs for direct PCR amplicons-sequencing using the GS20 Sequencing System (454 Life Sciences).
8. To perform sequence assemblies per infection trial using data generated in objective 1 and 7. This process gives an indication of the apple NBS-LRR transcriptome under scab and powdery mildew infections.
9. To select representative contigs containing genes constitutively present before and after pathogen inoculation and those present only after the inoculation process. Specific primers to be designed from these contigs and used to assess transcription in the same set of samples using the quantitative real-time PCR technique. Each seedling assessed to use as control material sampled from that same plant before the inoculation process.

Answers for these specific objectives are arranged in chapters and the order thereof doesn't follow the order set for the objectives. Chapters are set in a way that optimises the full utilisation of results obtained at each stage of the experimental design.

CHAPTER 2: MATERIALS AND METHODS

2.1 LIST OF MATERIALS AND SUPPLIERS

Materials	Suppliers
ABI PRISM SNaPshot™ Multiplex Kit	Applied Biosystems
40% Acrylamide-Bis (37.5:1) ready-to-use solution	Sigma
Agarose D1LE	Whitehead Scientific
Ampicillin	Roche
5-bromo-4-chloro-3-indolyl-ß-D-galactoside (X-gal)	Roche
Boric acid (Orthoboric acid)	Merck
BSA (Bovine serum albumin)	Sigma
ß-mercaptoethanol	Sigma
Chloroform	BDH - Merck
CTAB (N-cetyl-N,N,N-trimethyl-amino bromide)	Saarchem
DMSO (Dimethyl sulphoxide)	Roche Diagnostics
DEPC (Diethyl pyrocarbonate)	Sigma
dNTPs	ABgene
EDTA (ethylene diamine tetraacetic acid)	Merck
Ethidium Bromide	Sigma
Exonuclease I (Exo I)	Fermentas
Ficoll 400	Sigma

Formaldehyde	Merck
Glycerol	Saarchem
GeneScan-120 LIZ size standard	Applied Biosystems
Herring sperm DNA	Roche
Hi-Di formamide	Applied Biosystems
Hydrochloric acid	Saarchem
Isoamyl alcohol	Merck
Isopropanol (propan-2-ol)	BDH - Merck
IPTG	Fermentas
L – lysine monohydrochloride	Aldrich
LabelStar Array Kit	QIAGEN
LightCycler Fast Start DNA Master^{PLUS} SYBR Green I	Roche Applied Science
Magnesium chloride hexahydrate	Riedel-de Haën
MOPS	Sigma
Oligonucleotides	Inqaba Biotechnology
pGEM-T Easy Vector	Promega
PIPES	Sigma
PVP-40 and PVP-360 (polyvinyl pyrollidone)	Sigma
Potassium chloride	Saarchem
Proteinase K	Roche Diagnostics
QIAquick® Gel Extraction kit	QIAGEN
QIAEX II Gel Extraction kit	QIAGEN
QIAshredder™ columns	QIAGEN

QuantiTect® Reverse Transcription Kit	QIAGEN
RNase A	Roche
RNase AWAY®	Molecular BioProducts
RNeasy Plant Mini Kit	QIAGEN
Shrimp Alkaline Phosphatase (SAP)	Fermentas
Sodium acetate	Saarchem
Sodium chloride	Saarchem
Sodium diethyldithiocarbamate	Sigma
Sodium dodecyl sulphate	Sigma
Sodium metabisulfite	Sigma
Sodium pyrophosphate	Sigma
Sodium sulphite	NT Laboratory Supplies
T4 DNA Ligase	Fermentas, Promega
TEMED	Sigma
Tris(hydroxymethyl aminomethane)	Merck
Triton X-100	BDH – Merck
Trizol Reagent	Invitrogen
Tryptone	Biolab - Merck
Tween 20	Merck
Yeast Extract	Biolab – Merck

2.2 LIST OF SOLUTIONS

RLT, RPE and RW1 are commercial reagents supplied as components of the RNeasy Plant Mini Kit. Buffers RLT and RPE are prepared by adding 143 mM ß-mercaptoethanol and 4 volumes of 96% ethanol respectively.

QG and PE are commercial buffers supplied as components of the QIAquick® Gel Extraction Kit. Buffer PE is prepared by adding 4 volumes of 96% ethanol to 1 volume of reagent PE concentrate (v/v).

QX1 buffer and QIAEX II silica-gel particles are commercial reagents supplied as components of the QIAEX II Gel Extraction Kit.

RNase-free water was prepared by adding 0.1% diethyl pyrocarbonate in deionised water, incubate at 37ºC overnight and autoclave at 121ºC for 20 minutes.

gDNA Wipeout buffer is a commercial product supplied as a component of the QuantiTect® Reverse Transcription kit.

2X CTAB	2% (w/v) CTAB, 100 mM Tris-Cl (pH 8.0), 20 mM EDTA (pH 8.0), 1.4 M NaCl, 2% PVP-40, 0.06% Na_2SO_3
75% Ethanol	75% ethanol, 25% H_2O (v/v)
10% APS	10% (w/v) ammonium persulfate in H_2O
CIA	Chloroform: isoamyl alcohol (24:1 v/v)
5X RNA loading buffer	0.16% (v/v) saturated aqueous bromophenol blue, 4 mM EDTA, pH 8.0, 0.89 M Formaldehyde (37%), 20% Glycerol, 4X FA gel buffer, 7.71 M Formamide
6X DNA loading buffer	50% (v/v) glycerol, 0.25% (w/v) Bromophenol blue, 0.25% (w/v) Xylene Cyanol FF, 5mM EDTA
Inoue Transformation Buffer	55 mM $MnCl_2$•$4H_2O$, 15 mM $CaCl_2$•$2H_2O$, 250 mM KCl, 10 mM PIPES (pH 6.7)
IPTG (100 mM)	1.2 g in 50 ml filter deionised water
LB	1% Tryptone, 0.5% Yeast extract, 0.17 M NaCl
10X PCR buffer	500 mM KCl, 100 mM Tris-Cl (pH 9.0), 1% Triton X-100, 25 mM $MgCl_2$
SOB	2% (w/v) Tryptone, 0.5% (w/v) yeast extract, 8.55 mM NaCl
10X TBE	0.89 M Tris(hydroxymethyl) aminomethane, 0.89 M Boric acid, 0.04 M EDTA pH8.3
TE	10 mM Tris-Cl, 1 mM EDTA pH 8.0

Ligase buffer	30 mM Tris-Cl (pH 8.0), 10 mM $MgCl_2$, 10 mM DTT, 1 mM ATP, 5% polyethylene glycol
Diffusion buffer	0.5 M ammonium acetate; 10 mM magnesium acetate; 1 mM EDTA, pH 8.0; 0.1% SDS.
X-gal	100 mg 5-bromo-4-chloro-3-indolyl-ß-D-galactoside, 2 ml N,N´-dimethyl-formamide

***In all cases solutions were made using deionised water unless otherwise indicated.

2.3 Bacterial cultures

Bacterial Strains

Table 5. The bacterial strain used

Strain	Genotype
E. coli XL1-Blue	RecA1, end A1, gyrA96, thi-1, hsdR17, supE44, relA1, lac[F'proAB, lacI^{q}ΔM15, Tn10(tet^{r})]

2.3.2 Storage of bacterial strains

A single colony picked from a streaked plate was grown overnight in fresh LB broth to near saturation. The cultures were then mixed with 20% (v/v) sterile glycerol before being stored at -80ºC in 1 ml aliquots.

2.4 Cloning vectors

2.4.1 pGEM-T Easy Vector System

The pGEM-T® Easy Vector is prepared by cutting Promega's pGEM-T® Easy Vectors with *Eco*R V and adding 3' terminal thymidine to both ends of the linearised vector. It is these 3'-T overhangs that not only prevent re-circularisation of the vector, but also allow high efficiency ligation to PCR products. This principle works well in conjunction with certain *Taq* polymerases that add a template independent adenine to the 3'-terminal of a PCR product.

This is a high copy number vector containing T7 and SP6 RNA Polymerase promoters flanking the multiple cloning site within the α-peptide coding region of the gene encoding the β-galactosidase enzyme. Insertional inactivation of this gene allows recombinant clones to be screened on indicator plates through blue/white selection. The vector also has pUC/M13 oligonucleotide binding sites flanking the multiple cloning region, this allows for screening for recombinants using M13 forward and reverse oligonucleotides. Sequencing of the insert fragment is made easy through the use of T7, SP6 or M13 oligonucleotides.

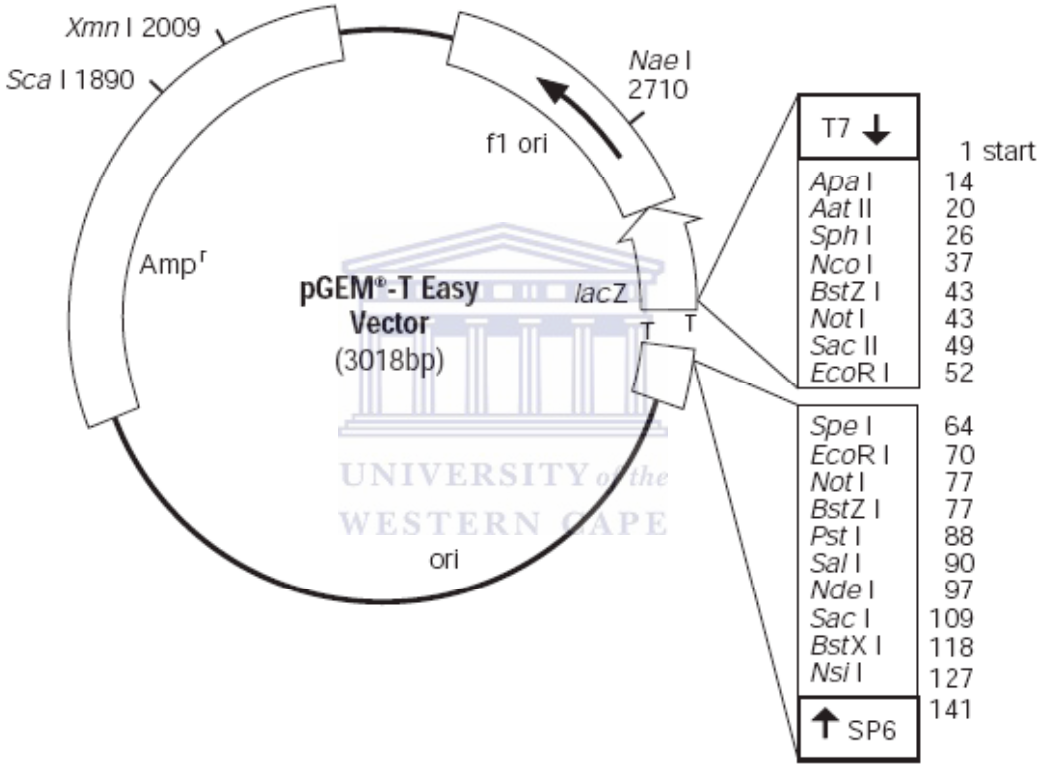


Figure 2.1. The circular map of pGEM-T® Easy Vector showing the structure of the multiple cloning site (mcs) (Promega Technical Manual No. 042).

2.5 Sampling of Plant Material

Apple leaf material was obtained from Bienne Donne Experimental farm (ARC Infruitec-Nietvoorbij) and other experimental farms within the Western Cape region. *Malus x domestica* leaf bulks made up of tender young leaves were obtained from both juvenile and mature plants that had not been used in experimental or controlled infection trials. Leaves were rinsed in distilled water before collection. All labelled ziplock bags containing the samples were submerged in ice water before being transported to the laboratory. Samples were stored at -20ºC.

2.5.1 Anna and Golden Delicious leaf bulks

Leaves were cleaned under normal room temperature tape water to remove dust and other possible contaminants. Midribs were removed and the leaf lamina was crushed in liquid nitrogen using a pestle and mortar. The powder was pooled into five 15 ml tubes, sealed with parafilm and then stored at -20°C until needed. This was only done for the bulked samples to get a homogenous mixture of the different leaves.

2.5.2 Samples for expression analysis

Seedlings of the two crosses, Simpson x Carmine and Lady Williams x Prima were monitored closely for a period of two months at Bienne Donne Experimental farm (ARC Infruitec-Nietvoorbij). Lady Williams x Prima seedlings were maintained under sterile greenhouse conditions before and after inoculation with *Venturia inaequalis* spores. Sack cloths soaked in a cocktail of fungicides were hung around the seedlings. Carmine x

Simpson seedlings were maintained under sterile glasshouse conditions before exposure to *Podosphaera leucotricha* spores. After sampling uninfected leaves, seedlings were transferred out of the glasshouse for natural powdery mildew infection. Seedlings were rinsed with sterile sprinkler water and leaves were collected in labeled zip-lock plastic bags. These were immediately snap frozen in liquid nitrogen and stored at -80°C. Seedlings were photographed before and after infection.

2.6 Nucleic acid isolation

2.6.1 DNA extraction

2.6.1.1 Genomic DNA extraction

Leaves with their midribs removed were crushed in liquid nitrogen using a pestle and mortar for samples other than the bulks. DNA was extracted using the CTAB method (Jobes *et al.*, 1995) with a few alterations to maximise on both quantity and quality. Roughly 0.2 g of ground leaf powder was transferred to a 2 ml tube and mixed with 1 ml pre-warmed 2X CTAB supplemented with 0.06% Na_2SO_3. The suspension was incubated in a 65°C waterbath for 1 hour. Recombinant Proteinase K was added to the homogenate to a final concentration of 0.1mg/ml and incubated for 30 minutes at 37°C. Nucleic acids were separated from organic debri through solvent – solvent partitioning in an equal volume of chloroform : isoamyl alcohol (24:1 v/v). An equal volume of ice-cold isopropanol was added and mixed by inversion 5 times then tubes were incubated at -20ºC for 20 minutes. The precipitated nucleic acids were recovered by centrifugation at 16000 x g for 10 minutes. The nucleic acid pellet was washed twice in 500 µl of 75%

ethanol and the nucleic acid pellet recovered each time by centrifugation at 10000 x g for 2 minutes, and the ethanol discarded after each washing step. The nucleic acid pellet was air dried for 15 minutes and resuspended in 1X TE buffer. The co-precipitated RNA was removed by a 30-minute incubation with RNase A (Roche). The RNase A enzyme and degraded RNA were removed by precipitation with 0.56 volumes of ice-cold isopropanol for 2 minutes and centrifugation at 10000 x g for 5 minutes. Purified DNA was resuspended in 1X TE buffer and stored at +4ºC.

2.6.1.2 DNA extraction from agarose gels

DNA was resolved on an agarose gel (section 2.12) and DNA fragments were visualised on a UV transilluminator at 312 nm wavelength. Desired fragments were excised from the agarose gel and placed in pre-weighed 1.5 ml tubes. The amount of agarose per tube was estimated as the difference of masses before and after addition of the agarose gel slice. Three volumes of buffer QG were added per one volume of agarose gel (v/w) then incubated at 50ºC for 10 minutes (or until the agarose gel had dissolved completely). After complete dissolution of the agarose gel, one volume of isopropanol (v/v) was added and mixed by inverting the tube. The solution was added into a MiniElute column (Qiagen) placed in a 2 ml collection tube and centrifuged at 10 000 x g for a minute. The flow through was discarded and a fresh 500 µl of buffer QG was added into the MiniElute column and centrifuged at 10 000 x g for a minute. After discarding the flow through from the collection tube, 750 µl of buffer PE was added to the MiniElute column and centrifuged at 10 000 x g for a minute. The flow through was discarded and the empty column was centrifuged at 14 000 x g for a minute to remove residual buffer PE.

The column was placed in a 1.5 ml tube and 15 µl of pre-warmed 1X TE buffer was added to the column and incubated at 25ºC for a minute then centrifuged at 10 000 x g for a minute. The supernatant containing purified DNA was stored at -20ºC.

2.6.1.3 DNA extraction from polyacrylamide gels

DNA was resolved on a 6% polyacrylamide gel and DNA fragments were visualized on a UV transilluminator (412 nm wavelength). Desired fragments were excised from the gel placed into 1.5 ml pre-weighed tubes. The amount of gel per tube was estimated as the difference of masses before and after addition of the gel slice. Two volumes of diffusion buffer were added per one volume of gel (v/w) before incubation for 30 minutes at 50ºC. The mixture was transferred to a QIAshredder™ column (Qiagen) placed in a 2 ml tube and centrifuged at 14 000 x g for a minute. The recovered filtrate was then mixed with 3 volumes of buffer QX1 and 10 µl of QIAEX II silica-gel particles before incubating the reaction for 10 minutes at 25ºC with constant mixing. The suspension was then centrifuged for 0.5 minutes at 14 000 x g to pellet the QIAEX II particles bound to the DNA. The pellet was washed twice with 0.5 ml of buffer PE before air-drying for 20 minutes. The bound DNA was then eluted by adding 20 µl of 1X TE, incubating at 25ºC for 10 minutes then centrifuging 14 000 x g for 0.5 minutes. The supernatant containing the purified DNA was transferred into a sterilized 1.5 ml tube and stored at -20ºC.

2.6.2 Total RNA extraction

2.6.2.1 Pre-treatment of glass and plastic-ware

All glass and plastic-ware to be used in the extraction and handling of RNA were soaked overnight in 0.1% DEPC (diethyl pyrocarbamate). These were sterilized by autoclaving at 121°C for 20 minutes before baking to dry at 100°C dry heat. The treatment included pestle and mortars, micropipette tips and Eppendorf tubes. Micropippettes, gel electrophoresis apparatus and other such material were wiped with RNase AWAY® (Molecular BioProducts).

2.6.2.2 (a) Column based extraction of Total RNA

Leaf material was crushed under liquid nitrogen, 0.1 grams of the powder were transferred to RNase-free 1.5-ml tubes and extracted using the RNeasy Plant Mini Kit (QIAGEN) according to the manufacturer's instructions. The crushed leaf material was mixed with 450 µl of buffer (RLT) and vortexed thoroughly to homogenize the crushed material. The suspension was applied to a QIAshredder spin column placed in a 2-ml collection tube and centrifuged for 2 minutes at 16000 x g. The supernatant was transferred to an RNase free 1.5-ml tube and mixed with 225 µl of absolute ethanol, mixed thoroughly by pipetting before being loaded into an RNeasy mini column (placed in a 2-ml collection tube). RNA was bound to the column by centrifugation at 10 000 rpm for 15 seconds, subsequent washes once in 500 µl buffer (RW1) and twice in 700 and 500 µl buffer (RPE) respectively were carried out at these centrifugation conditions. Total

RNA was eluted in 40 µl RNase-free water (RNeasy Mini Kit, QIAGEN) and stored at -20°C.

2.6.2.2 (b) Trizol® Reagent extraction of total RNA

Leaves were ground in liquid nitrogen and the powder was transferred to a 1.5 ml tube on ice. Leaf powder in 1.5 ml tubes was removed from ice and immediately mixed with 1 ml of Trizol® reagent. The mixture was homogenized for 15 seconds before centrifugation at 12 000 x g for 10 minutes to remove debri. The supernatant was transferred into a clean 1.5 ml tube and 200 µl of chloroform was added. The solution was mixed thoroughly and incubated at room temperature for 3 minutes. Tubes were centrifuged at 12 000 x g for 15 minutes to separate the organic and aqueous phases. The top aqueous phase was transferred to a clean 1.5 ml tube, mixed with 500 µl of isopropanol and incubated at room temperature for 10 minutes. The precipitation reaction was centrifuged at 12 000 x g for 10 minutes and then the supernatant was discarded. The pellet was washed in 75% ethanol and centrifuged for 5 minutes at 7 500 x g. A partial drying step was included for 10 minutes following which the pellet was mixed with 20 µl of RNase-free water, incubated at 60ºC for 10 minutes and mixed to resuspend the RNA pellet.

2.7 cDNA synthesis

2.7.1 Elimination of Genomic DNA

First strand cDNA synthesis was performed using the QuantiTect® Reverse Transcription kit (Qiagen) according to the manufacturer's instructions. Template RNA and gDNA Wipeout Buffer were thawed on ice, mixed gently then centrifuged at 10000 x g for 3 seconds to collect at the bottom. The genomic DNA elimination reaction was set up on ice by mixing 1X gDNA Wipeout Buffer, 1 µg template RNA and the reaction volume was made up to 14 µl using RNase-free water. The reaction was incubated at 42ºC for 2 minutes then placed on ice.

2.7.2 cDNA synthesis reaction

The reaction was performed on ice according to the manufacturer's instructions. A 1X final concentration of Quantiscript RT Buffer, 1 µl Quantiscript Reverse Transcriptase and 1 µl RT Primer Mix were added to the genomic DNA elimination reaction in a total volume to 20 µl. The reaction was incubated at 42ºC for 20 minutes then the Quantiscript Reverse Transcriptase enzyme was inactivated by incubation at 95ºC for 5 minutes. The cDNA was then stored at -20ºC.

2.8 Polymerase chain reaction

2.8 (a) Optimisation of PCR annealing temperature

A master mix containing all the PCR amplification reagents including target DNA was prepared on ice for twelve reactions (section 2.8.1). These were aliquoted into 0.2 ml thin-walled tubes and amplified through 35 cycles of 95°C for 30 seconds, annealing temperature for 10 seconds and 72°C for 30 seconds. Each of the twelve amplification reactions was subjected to a unique annealing temperature in the range 55 to 65°C in a gradient thermal cycler. Products of the amplification reaction were resolved through agarose gel electrophoresis (section 2.12.1). The optimal annealing temperature was identified by the position of a reaction on the thermal cycler block that produces a single DNA fragment of the expected size.

2.8 (b) Optimisation of oligonucleotide concentration

Individual PCR master mixes were prepared for 12 reactions with all reagents as described in section 2.8.1 with the exception of oligonucleotides. Twelve serial dilutions of oligonucleotides were performed for individual reactions from a 10 µM stock. Thermal amplification was carried out using optimal annealing temperatures (section 2.8. (a)) as described in section 2.8.1. PCR amplification products were resolved by agarose gel electrophoresis (section 2.12.1) and the optimal oligonucleotide concentrations were identified as the dilutions producing a single DNA fragment of the expected size.

2.8.1 Standard PCR

PCR reactions were setup in a total volume of 25 µl with a final concentration of 10 µM dNTP, 0.8 µM oligonucleotide, ~50 nanograms DNA, 1X PCR buffer and the total volume was made up to 25 µl with autoclaved deionised water. The thermal cycling was run at 94°C for 4 minutes for the initial denaturation step followed by 35 cycles of 94°C for 30 seconds, annealing temperature set at 53°C for 30 seconds and the oligonucleotide extension at 72°C for 60 seconds followed by a final extension at 72°C for 10 minutes.

2.8.2 Quantitative real-time PCR

A PCR master mix was prepared containing 2µl of LightCycler FastStart DNA Master^{PLUS} SYBR Green I reaction mix (Roche Applied Science), 0.25 µM of each oligonucleotides, ~30 ng of cDNA and autoclaved deionised water was added to a final volume of 20 µl. The negative control contained all components of the master mix and the DNA was substituted by autoclaved deionised water. The PCR mixes were transferred to glass capillaries that had been pre-cooled in centrifuge adaptors and centrifuged at 700 x g for 5 seconds. Capillaries were placed in the LightCycler carousel and then loaded into the LightCycler machine. The reaction was performed using the LightCycler Software version 3 package and Table 6 shows the PCR cycles and parameters used.

LightCycler software package

The LightCycler version 3 software package was used to perform quantitative real-time PCR analyses. This package is made up of LightCycler Software version 3.5 (Roche

GmbH), with data analysis version 3.5.28, graph works version 10.0.7 (Idaho Technology Inc., 1998 and 1999 respectively) and LightCycler3 Run version 5.32 (Idaho Technology Inc., 1998 and Roche GmbH, 1999).

Table 6. Thermal cycling conditions for quantitative real-time PCR. The asterisk '*' used in the table indicates PCR cycles at which the apparatus was set to acquisition mode.

Step	Temperature (°C)	Time (Seconds)	Cycles	Acquisition mode
Denaturation	95	240	1	None
PCR amplification	94	15	45	*Single per cycle
	57	10		
	72*	15		
Melting curve	95	0	1	
	65	15		
	95*	0		Continuous
Cooling	40	30	1	

2.8.3 Colony PCR

White colonies were picked using sterile yellow tips and arrayed on a 96-well plate in 20 µl of autoclaved deionised water. From this cell suspension, 2 µl were used as a template in a reaction containing 50 µM dNTP mix, 0.32 µM M13 forward and reverse oligonucleotides, 0.04 µg/µl bovine serum albumin, 1X PCR buffer and autoclaved deionised water was added to make up the reaction volume to 25 µl. Thermal cycling was performed as follows; initially denaturation was set at 94ºC for 4 minutes followed by 35 cycles of 94ºC for 45 seconds, annealing at 63ºC for 30 seconds and extension at 72ºC for 45 seconds followed by a final 72ºC extension step for 10 minutes.

2.9 Ligation

Ligation reactions were performed using 8.3 ng of the PCR product (0.05 pmol) and 50 ng of pGEM-T Easy Vector (0.05 pmol), 1 U of T4 DNA ligase, T4 ligase buffer and autoclaved deionised water added to a total reaction volume of 10 µl. The ligation reaction was initially incubated at 25°C for half 1 hour then for 16 hours at 4°C. Alternatively ligation reactions were also incubated at 25°C for 4 hours.

2.10 Transformation

2.10.1 Preparation of E. coli XL1-Blue competent cells for chemical transformation

A single colony from a streaked plate was inoculated into 25 ml of SOB broth and allowed to grow at 37°C for 6 hours in a shaking incubator at 150 rpm. This was used as a starter culture 4 ml of which was used to inoculate a 250 ml volume of SOB in a 2-litre flask. This was grown overnight at 20°C and 100 rpm to an OD of 0.5 at A_{600}. The cells were cooled to 4°C on ice and were then centrifuged at 3000 x g for 10 minutes at 4°C. The bacterial pellet was gently resuspended into 80 ml of Inoue transformation buffer at 4°C. Bacteria cells were recovered by centrifugation at 3000 x g for 10 minutes at 4°C. The supernatant was discarded and the cells were resuspended in 20 ml Inoue transformation buffer supplemented with 1.5 ml DMSO. The resuspended bacteria were incubated on ice for 10 minutes before being aliquoted in 200 µl volumes into pre-frozen 1.5 ml tubes that were snap frozen in liquid nitrogen and stored at -80°C.

2.10.2 Heat shock transformation

A total of 3 µl of the ligation reaction was transferred to a 1.5 ml tube and pre-incubated on ice for 3 minutes. A fresh tube of frozen chemically competent XL1 - Blue cells was placed on ice until just thawed then 100 µl was immediately transferred to the pre-chilled ligation reaction. The transformation reaction was mixed thoroughly though gently and incubated on ice for 20 minutes. The transformation reaction was heat shocked a 42ºC

for 70 seconds, incubated on ice for 2 minutes then equilibrated to room temperature for a further 2 minutes. A total of 500 µl LB was added and incubated at 37ºC for 45 minutes. The transformation reaction was pelleted by centrifugation at 10000 x g for 1 minute, resuspended in 100 µl LB and plated on indicator plates containing 0.1 mg/ml ampicillin, 80 µg/ml X-gal and 0.5 mM IPTG. Plates were then incubated at 37°C for 16 hours.

2.11 Preparation of glycerol stocks of recombinants

Bacteria colonies picked into wells of a 96-well plate were mixed with 150 µl of LB. Plates were incubated at 37°C on a shaker for 6 hours before addition of 15% (v/v) sterile autoclaved glycerol. These glycerol stocks were then stored at -80°C.

2.12. Gel electrophoresis

2.12.1 Standard agarose gel electrophoresis

PCR fragments ranging from 0.5 to 0.8 kb were resolved on 1% agarose gels. Agarose gels were prepared by melting the appropriate amount of agarose in 0.5X TBE (half strength), allowing it to cool to 50°C before adding ethidium bromide to 0.5 ng/ml. Sample DNA to be resolved was mixed with 0.25 volumes of 6X loading dye then loaded into wells of the gel. DNA size standards were also loaded for size estimation. Electrophoresis was performed in 0.5X TBE buffer at 6 V/cm for genomic DNA and 10 V/cm for smaller fragments and DNA was visualised on a UV transilluminator.

2.12.2 Denaturing agarose gel electrophoresis

A 1.2% denaturing agarose gel for a 10 x 10 x 0.7 cm caster was prepared by melting 0.6 g of agarose in 50 ml of 1X FA gel buffer (section 2.2). The molten agarose was equilibrated to 65ºC in a waterbath. Formaldehyde (37%) and ethidium bromide were added to a final concentration of 0.44 M and 0.2 mg/ml respectively and thoroughly mixed by swirling. The molten gel was poured into a caster and allowed to set at room temperature. The gel was pre - equilibrate in 1X FA gel buffer for 30 minutes prior to running the electrophoresis.

2.12.3 Polyacrylamide gel electrophoresis

A 6% polyacrylamide gel was prepared by mixing 2.25 ml of 40% acrylamide-bis (37.5:1) ready-to-use solution, 0.6X TBE, 0.06% APS and 105 mM urea. The ingredients were mixed thoroughly before the solution was made up to a 15 ml final volume using distilled water. The solution was mixed with 0.132 mM TEMED then poured into the gel casting apparatus to polymerise. Each well of the gel was rinsed by squirting a jet of 1X TBE using a hypodermic syringe and a 24-gauge hypodermic needle to remove excess urea before loading the samples. Electrophoresis was then performed at 10 volts/cm for 2 hours after which post-staining was performed for 15 minutes in 0.5 mg/ml ethidium bromide in 1X TBE buffer.

2.12.3 DNA size standards

Most of the agarose gels were run to resolve nucleic acid fragments under 1.0 Kb in length and for such purposes the pTZ/*Hinf* I marker was used. The profile for the pTZ/*Hinf* I marker resolved at 10 V/cm for 1 hour on a 0.5X TBE buffered 1.0% agarose gel shows 6 bands at 1.2, 0.5, 0.39, 0.36, 0.2 and 0.07 kb and thus was suitable for the purposes of this work. A lambda/*Hind* III marker resolved at 6 V/cm for 2 hours on a 20 cm long 0.8% agarose gel shows 7 bands with the profile 23, 9, 6, 4, 2.3, 2 and 0.65 kb and thus was used to estimate sizes of isolated genomic DNA.

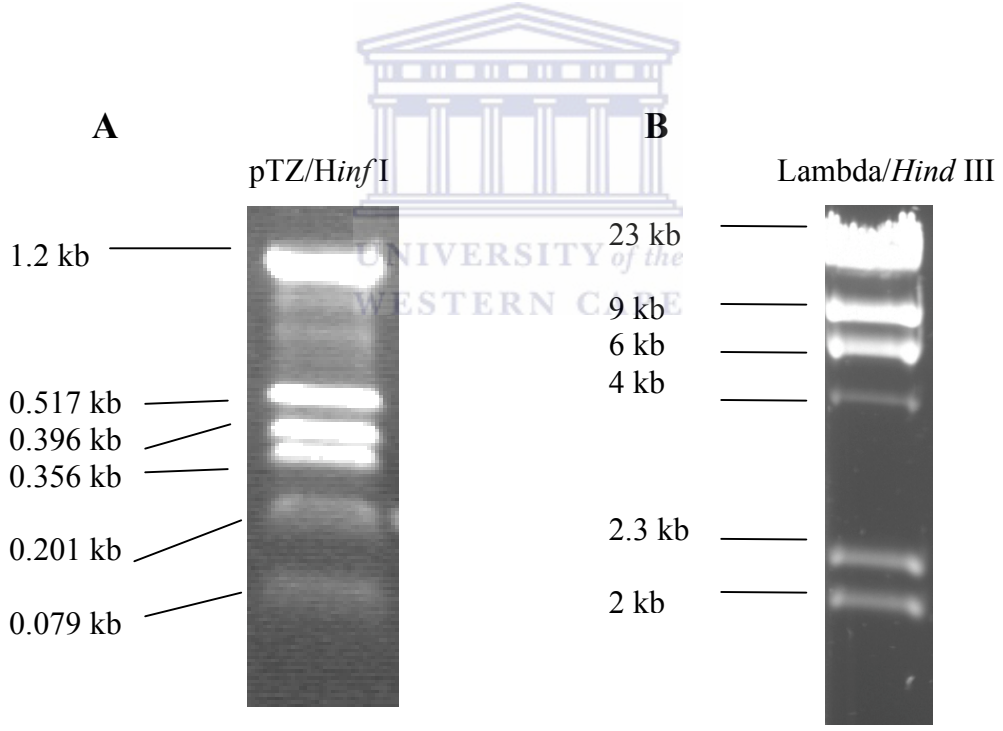


Figure 2.2. pTZ/*Hinf* I and lambda/*Hind* III DNA size standards. A is the pTZ plasmid digested with *Hinf* I restriction endonuclease; B is lambda DNA digested using *Hind* III restriction endonuclease.

2.13 Sequencing and sequence analysis

2.13.1 Sequencing

Glycerol stocks of colonies with desired inserts were streaked on ampicillin plates and incubated at 37ºC for 16 hours. Plates were sealed on the edges using parafilm and sent to Inqaba Biotechnical Industries (Pty) Ltd, South Africa. Plasmid preps were performed using the Mini Prep Kit (Fermentas) and single direction thermal amplification was done using the BigDye® Terminator version 3.1 Cycle sequencing kit (Applied Biosystems). Sequence by electrophoresis was performed on the Spectrumedix Genetic Analysis System (USA).

2.13.2 Sequence analysis

Sequence chromatograms were viewed using the 4Peaks program (Griekspoor and Groothuis, 2005). Sequence data was stored both as text files in fasta format and as sequence chromatogram files. In the text format the flanking vector sequences were deleted manually and each sequence was then displayed in the proper orientation based on results from BLAST homology searches (Altschul *et al.*, 1997) and ORF predictions in Sequence Analysis software version 1.6.0 (Gilbert, updated 2006).

2.13.2.1 Computer Programs and tools used in data analysis

Sequence homology searches were performed using the *tblastx* algorithm of the BLAST program (Altschul *et al.*, 1997). Open reading frames were predicted using a combination

of Sequence Analysis Software version 1.6.0 and the Pasteur Institute's GETORF program (http://bioweb.pasteur.fr/seqanal/interfaces/getorf.html).

Multiple alignments were performed using Clustalx version 1.83 (Thompson *et al*., 1997) and MEGA version 3.1 (Kumar *et al.*, 2004). Manual editing of multiple sequence alignments were performed using JalView version 2.1.1 (Clamp *et al.*, 2004).

Phylogenetic analyses were performed using MrBayes version 3.12 (Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003), PAUP*4 (Swofford, 2003). For these analyses multiple sequence alignment output files were opened in TextEdit and manually converted to nexus format. The GTR + I + gamma substitution model of DNA evolution was used in MrBayes analyses. Phylogenetic trees were manipulated and displayed in Mesquite software version 2.01 (Madison and Madison, 2007).

Estimation of gene conversion and Ka/Ks values were performed using Geneconv version 1.81 (Sawyer, 1999) and the HyPhy molecular evolution analysis platform (Pond *et al.*, 2005) respectively. Site-to-site inference of amino acid conservation was performed using McRate version 1.0.1 (Mayrose *et al*., 2004).

2.13.2.2 Sequence assembly and SNP discovery

Sequence assemblies were performed using CodonCode Aligner version 2.0.1 (CodonCode Corporation, Dedham, MA). In all cases an initial assembly was performed using default settings to assess the quality of the assembly and identify parameters that

required optimization. Trimming of vector sequence where necessary was performed using the pGEM-5Zf(+)[X65308.2:2982-3000-49] option in the Univec library of the software. The assembly process and comparison of contigs were performed using the built-in assembly algorithm and ClustalW options respectively.

2.14 High throughput SNP genotyping using SNaPshot™

2.14.1 Design of the SNP oligonucleotides

Two sets of oligonucleotides were designed, the first set with a pair of standard oligonucleotides flanking the nucleotide region containing the candidate SNP. The second set with a single extension oligonucleotide nested within the nucleotide fragment flanked by the first set. The second oligonucleotide was designed to terminate one nucleotide upstream of the candidate SNP. In cases where was more than one candidate SNP in the defined nucleotide fragment, extension oligonucleotides were designed to vary by four nucleotides in length.

2.14.2 Preparation of PCR template

A standard PCR amplification was performed on genomic DNA to generate nucleotide fragments flanking the candidate SNPs using the first set of oligonucleotides described in section 2.14.1. PCR products were purified from residual primer dimers and dNTPs in a reaction containing 2 and 5 Weiss Units of Exo I and SAP respectively, 1X Exo I and SAP buffers, 15 µl of PCR product and autoclaved deionised water to a final reaction volume of 22 µl. The reaction was incubated at 37°C for 1 hour and the enzymes were

deactivated by incubation at 80°C for 15 minutes. The purified nucleotide fragments were stored at -20°C.

2.14.3 Preparation of the SNaPshot™ reactions and thermal cycling

Reactions for the positive and negative controls were prepared using the control template and autoclaved deionised water respectively. For these reactions and the nucleotide fragments purified as described in section 2.14.2 were prepared as outlined in Table 7.

Table 7. Preparation of reaction mixes for the SNaPshot extension PCR amplification. All reagent volumes used are in µl.

Reagent	Positive	Negative	Sample
SNaPshot Multiplex Ready reaction mix	2	2	2
PCR product	-	-	3
SNaPshot Multiplex control template	2	-	-
SNaPshot Multiplex control primer mix	1	1	-
SNaPshot oligonucleotides	-	-	1
Sterile de-ionized water	5	7	4

The Multiplex control primer mix contains 20A, 28G/A, 36G, 44T, 52C/T and 60C oligonucleotides, where the number denotes the size of the primer and the letter is the candidate SNP. All the oligonucleotides for the samples were standardized to 0.02 pmol

final concentration. The SNaPshot Multiplex Ready reaction mix contains AmpliTaq® DNA polymerase, fluorescently labeled ddNTPs and PCR reaction buffer.

All reaction mixes were set up on ice. The reactions were run through a standard thermal cycling program with 40 cycles of 96°C for 10 seconds, 50°C for five seconds and 60°C for 30 seconds. Post extension products were purified by adding one Weiss unit of SAP, 1X SAP buffer and incubated at 37°C for one hour. The enzyme was deactivated by incubation at 80°C for 15 minutes.

2.14.4 Electrophoresis of the ddNTP termination reaction

SNaPshot extension products were electrophoresed using the ABI Prism 3130xl Genetic Analyzer (Applied Biosystems). Reactions for electrophoresis were prepared by mixing 1.0 µl of purified SNaPshot extension products (as described in section 2.14.3), 9 µl of Hi-Di formamide and 0.4 µl GeneScan-120 LIZ size standard. The reactions were vortexed, centrifuged at 10000 x g for four seconds then loaded into wells of a 96-well PCR plate and denatured at 95ºC for five minutes. Reactions were incubated on ice for two minutes then centrifuged at 1700 x g for 4 seconds to collect at the bottom and remove bubbles.

Fragment sizes were analysed on the ABI Prism 3130xl Genetic Analyzer using Any5Dye as Dye Set and FragmentAnalysis36_POP7 modules. Fragment sizes were analysed using GeneMapper software version 4.0 (Applied Biosystems).

CHAPTER 3

CLONING AND SEQUENCING OF RESISTANCE GENE ANALOGS

CONTENTS

CHAPTER 3: CLONING AND SEQUENCING OF RESISTANCE GENE ANALOGS

3.1 INTRODUCTION

The largest family of disease resistance (R) genes isolated in plants to date encodes proteins made up of a centrally located nucleotide binding site (NBS) domain and a C-terminal leucine rich repeat (LRR) receptor (Wroblewski *et al.*, 2007). The NBS domain of these R proteins may act as molecular switches that initiates a signal transduction pathway following a successful identification of pathogen elicitors by the receptor (Michelmore and Meyers, 1998). This family of R proteins is broadly divided into two subfamilies, the TIR and non-TIR due to the presence or absence of an N-terminal domain with structural similarity to the Toll/IL-1R proteins that recognises pathogen associated molecular patterns in *Drosophila* and mammalian systems (H. Zhou *et al.*, 2007). The NBS-LRR family represents the most diverse family of R proteins with members that play an important role in viral, fungal, bacterial, nematode and oomycete disease resistance. Examples of these proteins have been described in *Arabidopsis* (Reuber and Ausubel, 1996), tomato (Hennin *et al.*, 2001) and flax (Anderson *et al.*, 1997).

In order to isolate the NBS-LRR resistance gene analogs (RGA), a PCR-based strategy using a set of degenerate primers targeted at the P-loop and GLPL motifs has been used successfully. This approach has enabled the isolation of many RGAs from a number of plants including those in Malus genus. In 2003 43 RGAs were isolated from *Malus*

prunifolia, M. buccata, *Malus x domestica* (Borkh.) cvs Hong-ok and Fuji (Lee *et al.*, 2003), 30 from *Malus x domestica* (Borkh.) cv. Florina (Baldi *et al.*, 2004) and a direct submission of 351 isolated from *M. floribunda* cv. 821, *Malus x domestica* (Borkh.) cvs. Pinkie and A172-2 (Rikkerink *et al*., 2006). Genome analyses in *Arabidopsis* gave a total of 150 RGAs from a genome that is 125 Mb and has five chromosomes with an estimated 25 000 genes (TAIR; (Henry *et al.*, 2006; Rampitsch and Srinivasan, 2006). Estimates in apple to date shows that the genome is 750 Mb and has 17 chromosomes (Liebhard *et al.*, 2003) and thus might have more RGAs compared to *Arabidopsis*.

The level of accuracy in sequence analysis by nature increases with an increase in the volume of data analysed. This chapter describes the isolation of RGAs from *Malus x domestica* (Borkh.) cultivar Golden Delicious using the PCR-based approach. Sequence analyses to estimate the NBS-LRR family characteristics are then performed using a range of Bioinformatic tools.

3.2 Sequencing of RGAs

3.2.1 Design of the oligonucleotides

Arabidopsis and flax R proteins together with translations of putative *Malus x domestica* (Borkh.) resistance gene nucleotide sequences were accessed from GeneBank. Multiple sequence alignments were performed using ClustalX and manually edited as necessary using JalView and results are shown in Figure 3.1. A set of degenerate oligonucleotides was designed on the consensus sequence of the P-loop and GLPL motifs highlighted in orange colour in Figure 3.1. The forward oligonucleotide NBSF was designed on the P-

loop motif and the reverse NBSR1 from the GLPLAL motif. A third oligonucleotide NBSR2 was a modification of NBSR1 in which all the degeneracy codes W, H, N, Y, V and D except R were replaced by inosine residues to simplify the synthesis. The oligonucleotide sequences are shown in Table 8.

Table 8. **RGA forward and reverse oligonucleotide sequences.** Inosine residues are indicated in red font in NBSR2 showing positions that were altered to reduce the complexity of NBSR1

PRIMER NAME	PRIMER SEQUENCE
NBSF	GGWATGGGWGGWRTHGGWAARACHAC
NBSR1	ARNWYYTTVARDGCVARWGGVARWCC
NBSR2	ARIIIITTIARIGCIARIGGIARICC

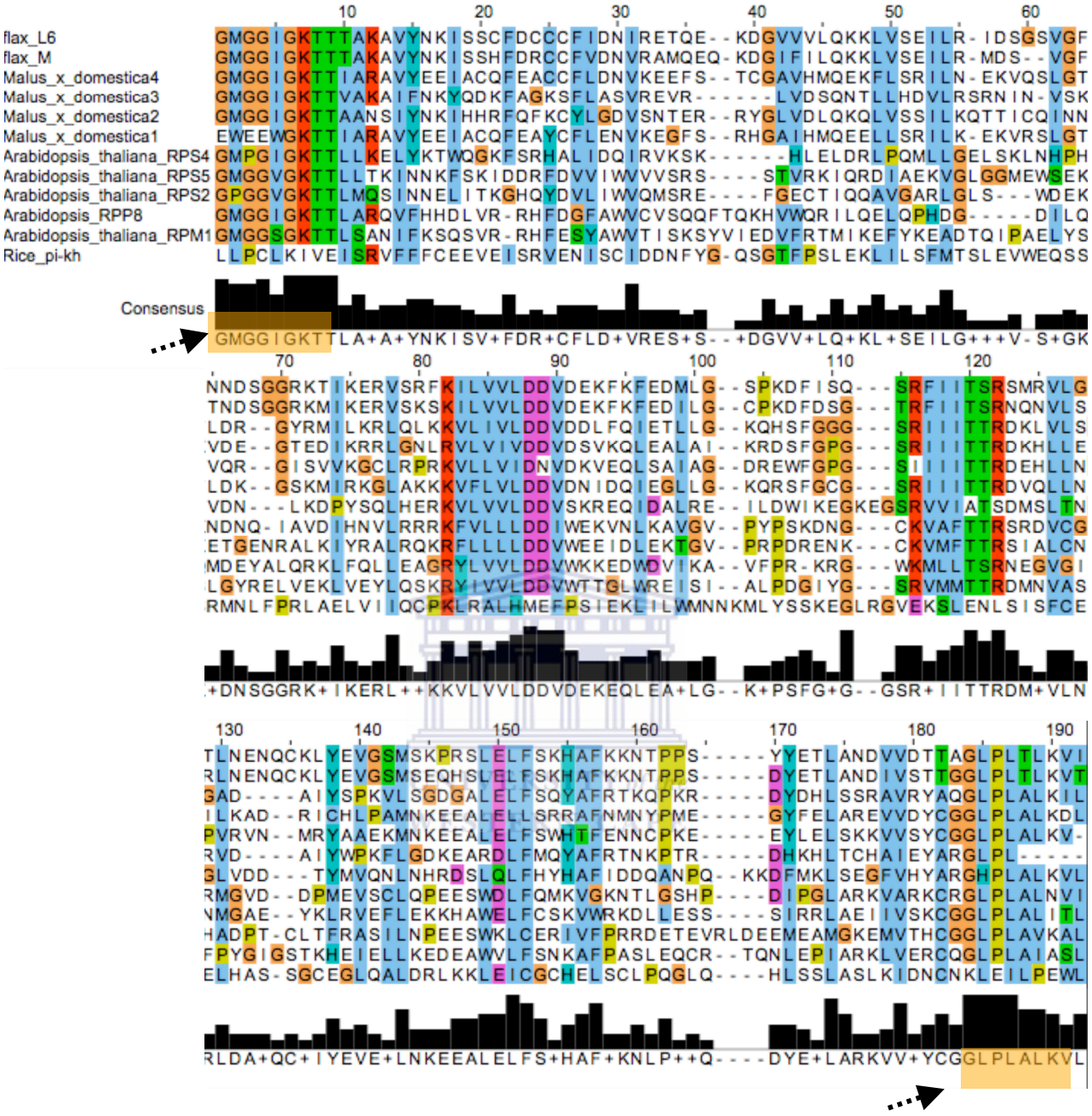


Figure 3.1. Multiple sequence alignment of R protein sequences showing positions of the forward and reverse oligonucleotides. GMGGGIGKTT and GLPLALKV amino acid sequences are highlighted here in orange. The bar graphs at the bottom of the alignment show the level of conservation of the particular amino acid residue.

3.2.2 Sample collection and DNA extraction

Leaf samples were collected from Golden Delicious trees as described in section 2.5 and genomic DNA was isolated from these leaves using the CTAB method as described in section 2.6.1. Figure 3.2 shows an agarose gel electrophoresis image of the isolated genomic DNA as clean high molecular weight fragments of nucleic acid co-migrating with the 23 kb fragment of the lambda/*Hind* III size standard.

3.2.3 PCR amplification of the NBS domains of NBS-LRR R genes

PCR amplifications were performed using conditions described in section 2.8.1. The NBSF and NBSR2 oligonucleotides gave amplification products that co-migrated with the 0.517 kb fragment of the DNA size standard. The expected fragment size that encompasses the whole NBS domain of the NBS-LRR genes ranges from 510 to 520 bases as evidenced by the indicated positions of the oligonucleotides in Figure 3.1. Figure 3.3 shows the successful amplification of the NBS domain of the NBS-LRR family of resistance genes using oligonucleotides NBSF and NBSR2.

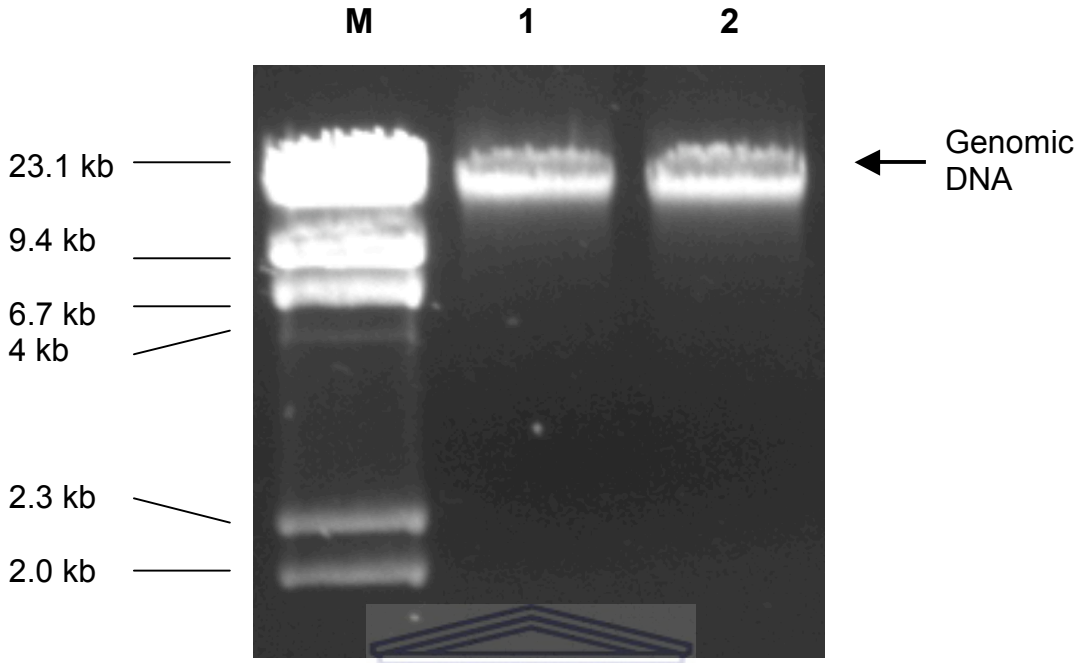


Figure 3.2. Isolation of genomic DNA from leaves of *Malus x domestica* (Borkh.). A 1% agarose gel electrophoresis image showing genomic DNA isolated from cultivars Anna and Golden Delicious in lanes 1 and 2 respectively and lane M shows the profile of lambda/*Hind* III DNA size standard.

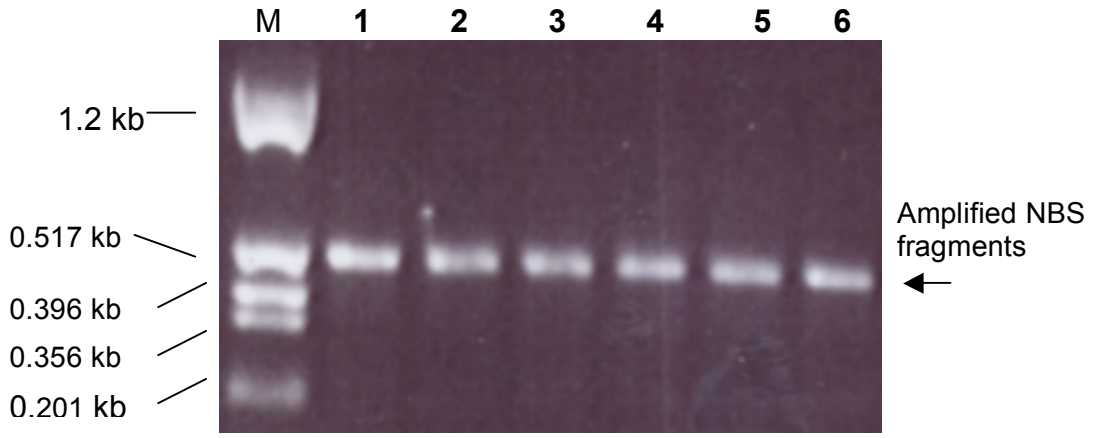


Figure 3.3. PCR amplification of the NBS domain fragments of NBS-LRR genes. A 1% agarose gel electrophoresis image showing results of the PCR amplification of cultivars Anna and Golden Delicious using the NBSF and NBSR2 oligonucleotides. Lanes 1, 2 and 3 have Golden Delicious and lanes 4, 5 and 6 have Anna products; the profile of pTZ/*Hinf* I DNA size standard is in lane M.

3.2.4 Colony PCR screening and selection of clones for sequencing

PCR amplification products of cultivars Anna and Golden Delicious described in section 3.2.3 and shown in Figure 3.3 showed single fragments per lane and no detectable primer dimers on the 1% agarose gel electrophoresis. These fragments were ligated into pGEM-T® Easy Vector (section 2.9) and transformed into competent XL1-Blue cells (section 2.10.2). These cells have the lacI^{q}ΔM15 gene that α-complements with the α-peptide coding region of β-galactosidase in pGEM-T® Easy Vector thus enabling colour screening of recombinant colonies. Recombinants were picked and arrayed into wells of

96-well microtiter plates before screening for colonies carrying the NBS DNA fragment inserts using colony PCR (section 2.8.3). Figure 3.4 shows an agarose gel electrophoresis image of colony PCR results. All colonies carrying inserts with sizes that were ± 700 bp were selected for sequencing. Colony PCR of an empty vector using M13 universal oligonucleotides gives a DNA fragment that is ± 200 bp due to amplification of DNA regions flanking the cloning site (Figure 2.1). Lane 5 in Figure 3.4 shows a DNA fragment that is ± 517 bp since it co-migrated at approximately the same level with the 517 bp fragment of the pTZ/*Hinf* I DNA size standard. This clone was carrying an insert of ± 317 bp after subtracting ± 200 bp of the DNA flanking that insert. Colonies carrying inserts of this size were not considered for sequencing.

3.2.5 DNA Sequencing

Colonies carrying inserts of the correct sizes were streaked on ampicillin plates and cultured for 16 hours at 37°C. The plates were then transported to Inqaba Biotechnical Industries (Pty) Ltd for sequencing (section 2.13.1). A total of 136 sequences were generated for a bulked sample, 261 and 265 for Anna and Golden Delicious (section 2.5.1) respectively.

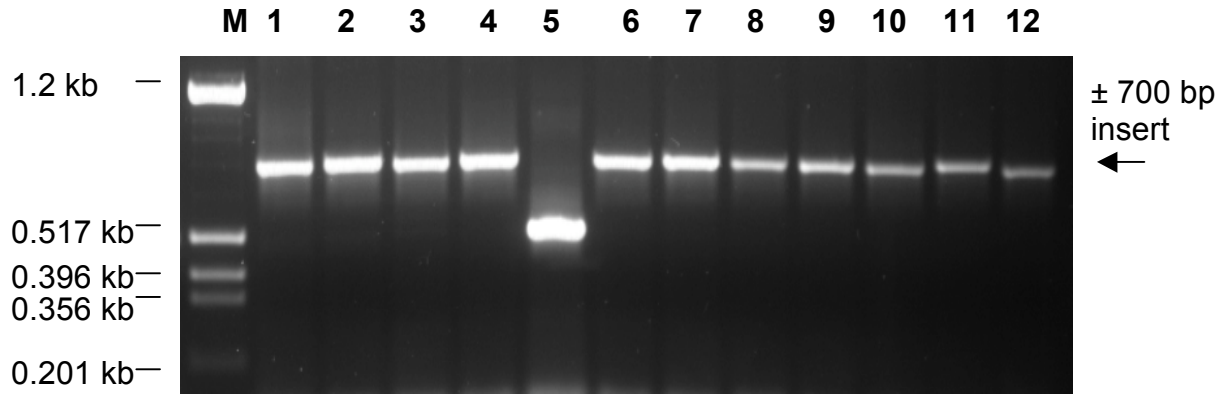


Figure 3.4. Colony PCR screening for XL1-Blue recombinant colonies carrying the NBS fragments. A 1% agarose gel electrophoresis image showing results of colony PCR screening of 12 of the recombinants. Lane M shows the profile of the pTZ/*Hinf* I DNA size standard and lanes 1 to 12 with the exception of lane 5 represent recombinants with insert sizes estimated at ± 700 bp. Lane 5 has a product estimated at about 517 bp.

3.3 Bioinformatic analysis of the sequences

3.3.1 Preliminary analysis of the sequence data

Candidate RGA sequences were screened for errors by visual inspection and those with a large ratio of signal to background noise were sent for re-sequencing. Sequences were manually edited to remove vector sequence. Homology searches of the cleaned sequences were performed using BLAST against the NCBI non-redundant database and for this step

the tblastx algorithm (Altschul *et al*., 1997) was used. All searches were carried out using default settings of the critical parameters such as the expect values (E-values). Homology searches were performed to confirm both sequence identity and orientation. For sequences in reverse orientation, Sequence Analysis software (section 2.13.2.1) was used to generate the reverse complements.

The sole selection criterion here was taken to be local alignments extending over the whole sequence in the best alignment of BLAST's tblastx search results and not just the priming regions. Sequences satisfying this criterion were then annotated as RGAs and thus were taken for further analysis.

3.3.2 Alignment-based editing of trace files

Using the procedure outlined above 682 sequences made up of 136, 261 and 285 from a bulked mix of *Malus x domestica* Borkh cvs. Anna and Golden Delicious, Anna alone and Golden Delicious alone respectively were confirmed as NBS-LRR RGAs through Blast homology searches. These sequences were aligned using the ClustalW option of MEGA3.1 (Kumar *et al.*, 2004) and the multiple sequence alignment was assessed manually. Trace files of all the sequences causing gaps in the alignment were analysed again to correct duplications and/ or insertions, the ambiguity character 'N' was used to replace a nucleotide base to correct miscalling of nucleotide by the sequencing software. Several alignments and trace file editing sessions were performed until the quality of the remaining sequences in the alignment was satisfactory.

3.3.4 Multiple sequence alignment and phylogenetic analysis

3.3.4.1 Multiple alignment and test of substitution pattern homogeneity

Cycles of multiple sequence alignments of the 285 Golden Delicious sequences were performed using ClustalX and each result file imported into MEGA3.1 for manual editing to remove sequences distorting the alignment. Edited alignment files were re-imported into ClustalX for re-alignment with the default 'delay divergent sequences' option at 30%. These rounds of multiple alignments and manual editing were done up to and until the final ClustalX multiple sequence alignment result was sufficiently acceptable with all the conservation indicators in appropriate columns based on expected primary structure of the NBS domain.

Phylogenetic analyses assume a homogeneous pattern of nucleotide substitution and any violation to this basic assumption introduces serious biases that upset the related inferences. The 'analysis of substitution homogeneity' method in MEGA3.1 was used to assess the aligned sequences.

Results from this analysis were used in confirming the exclusion of very divergent sequences from the final alignment file. This process confirmed the exclusion of 20 sequences from the final Golden Delicious multiple sequence alignment. Consequently chromatograms of these 20 sequences also showed a high signal to noise ratio.

A final multiple sequence alignment was performed for the Golden Delicious RGAs including NBS domain sequences from 6 previously characterised resistance genes, 5 *Arabidopsis* R genes (*RPP4, RPM1, RPS2, RPS5* and *RAC1*) and the flax *L6* gene as

controls for the downstream phylogenetic analyses. The P-loop and GLPL motifs were deleted from the final alignment since they constituted the priming site and could potentially introduce bias into the analyses. A total of 271 of sequences were used and the pairwise and multiple alignment parameters for gap opening and gap extension penalties were set to 25, 7.0 (pairwise) and 10, 10 (multiple) respectively to make the alignment more stringent since an extensive pre-cleanup had been done. The IUB DNA scoring matrix and a transition weight of 0.5 were left as default settings.

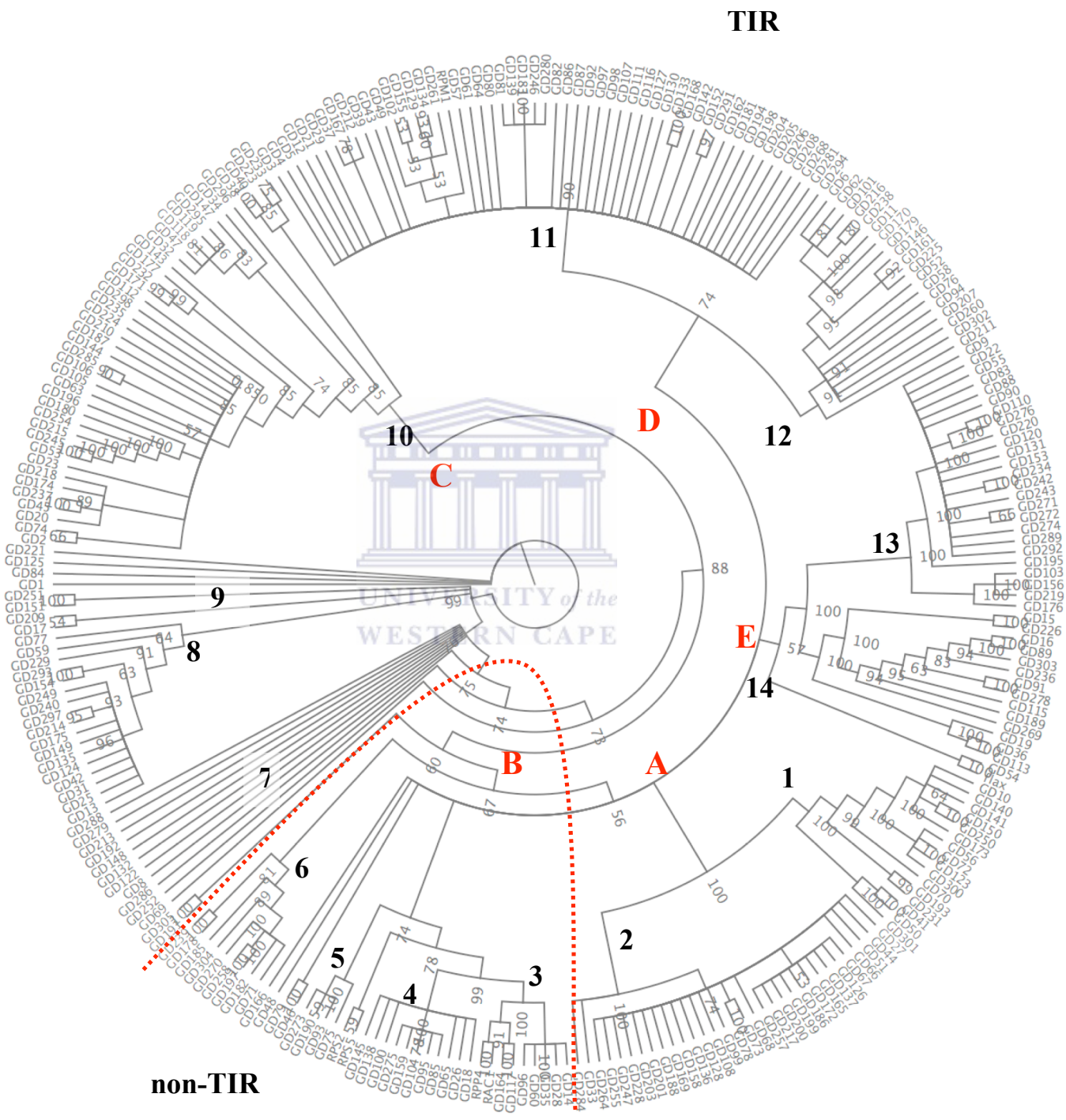
3.3.4.2 Phylogenetic tree construction for *cv*. Golden Delicious RGAs

The multiple sequence alignment from MEGA3.1 was used as the input file for inference of phylogeny using MrBayes version 3.1.2 (section 2.13.2.1). MrBayes is a program for the Bayesian inference of phylogeny using the Markov chain Monte-Carlo algorithm, MCMC (Ronquist *et al.*, 2005). The substitution model was set to the GTR + I + gamma model, the general time reversible model with rate variation set to invgamma (employs both gamma distributed rate variation and a proportion of invariable sites (Rogers, 2001; Mateiu and Rannala, 2006). The program uses Metropolis-coupled MCMC to approximate posterior probabilities of phylogenetic trees and the posterior probability density of the model parameters through several simulation cycles (Ronquist and Huelsenbeck, 2003). Two simultaneous and completely independent analyses that start from different random trees are run. The program compares sampled trees from the two parallel runs then summarises them to one tree at the end of the analysis. In this analysis the estimation was stopped at the 96% confidence interval (at this point the average standard deviation of split frequencies had reached 0.037657).

The extensive level of simulations allowed in this process eliminated the need to test for the reliability of branching through bootstrapping. To achieve a standard deviation of 0.037657 required 3 x 10^6 cycles with sampling done every 100^{th} cycle meaning a total of 60 000 trees were sampled. This process is more rigorous compared to bootstrapping with 1000 replicates in a standard bootstrapping process.

The consensus tree generated through the 50% majority-rule was manually edited to convert final posterior probabilities to percentage support. The use of percentage support as a measure of the reliability of tree branching was assumed to simplify interpretation of the result. The phylogenetic tree was displayed as a circular tree using the Mesquite software version 1.12 (section 2.13.2.1). The inferred phylogenetic tree is displayed in Figure 3.5. The tree shows 14 clusters although 2 (7 and 9) represent groups of highly divergent sequences. The whole dataset is broadly divided into 2 subfamilies, TIR and non-TIR based on controls from *Arabidopsis* and flax characterised genes (section 3.3.4.1).

Figure 3.5. The phylogenetic tree of Golden Delicious RGAs inferred using MrBayes. There are 5 major branches labelled A through E, which are further subdivided into 14 clusters, clusters 7 and 9 represent unresolved highly divergent sequences. The thick dotted red line divides TIR and non-TIR groups based on control sequences used in the alignment.



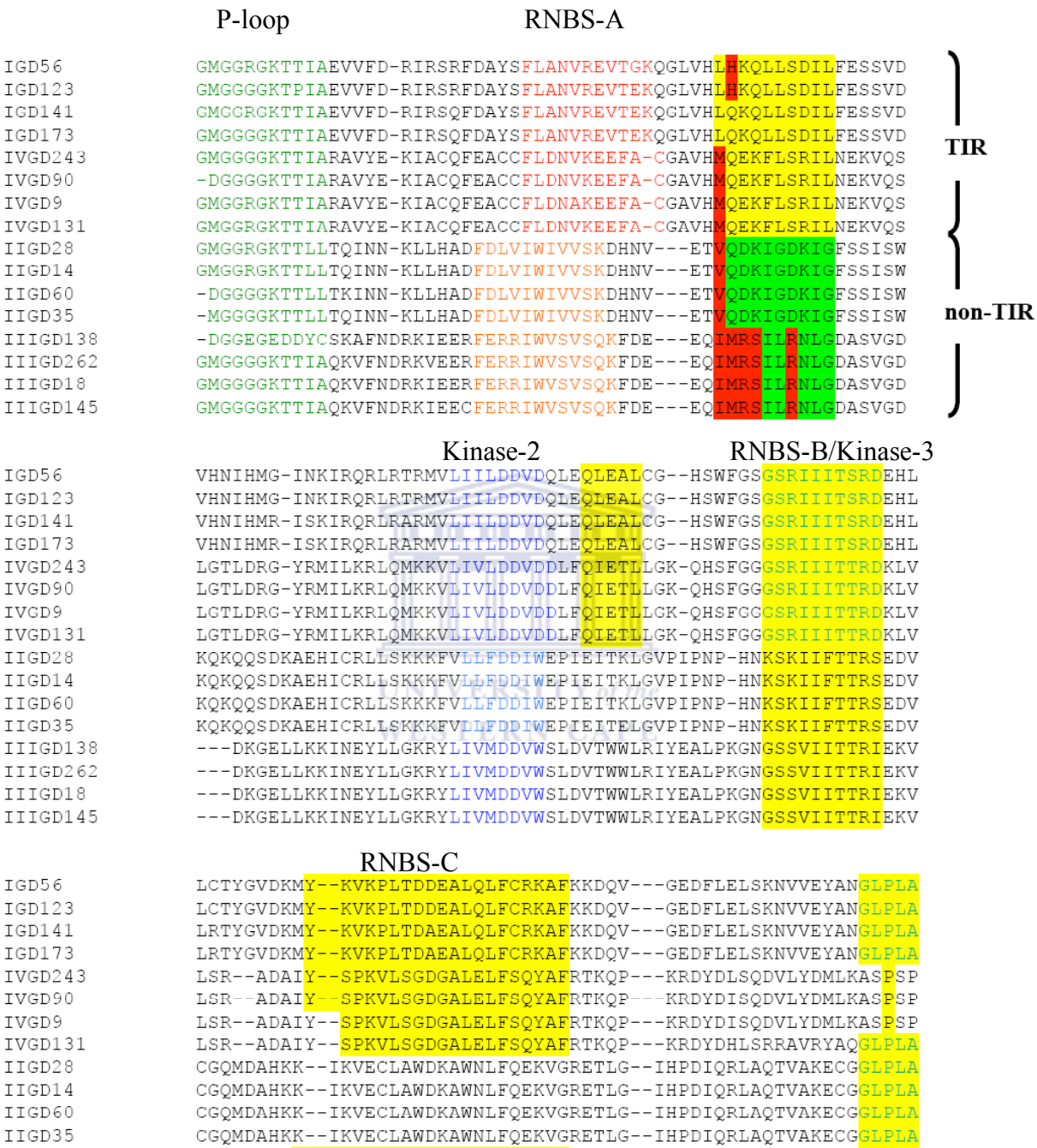
The phylogenetic analysis split sequences into TIR and non-TIR subfamilies as shown in Figure 3.5, the basis for this subdivision is elaborated in Table 9 and Figure 3.6. A total of 16 representative sequences from each of the two subfamilies were selected and translated into their respective amino acid sequences. Homology searches were performed against the PROSITE (Bairoch, 1991, 1992, 1993) and PRINTS (Attwood and Beck, 1994; Attwood, 2002) databases to determine the conserved motifs. Results showed the disease resistance signature motifs, P-loop (GMGGKTT), GLPL and kinase-2 (LI(V/I)LDDV(W/D)Q) motifs (Lee *et al.*, 2003; Xu *et al.*, 2005, 2007). The kinase-2 motif showed a diagnostic characteristic in which the presence of aspartic acid (D) and tryptophan (W) residues occurred exclusively in the TIR and non-TIR subfamilies respectively. The RNBS-B/Kinase-3 motif was also shown to vary significantly between the two subfamilies. The conservation of the GLPL motif did not show any variation between TIR and non-TIR subfamilies, however four members of the TIR subfamily had ASPS in place of GLPL in sequences analysed in Figure 3.6.

The alteration in the GLPL motif to ASPS in four of the RGAs analysed in Figure 3.6 effectively changes the overall hydrophobicity of the motif. Based on the side chain chemistry of the amino acids and the hydrophobicity scales (Wolfenden *et al.*, 1981; Kyte and Doolittle, 1982), L is strongly hydrophobic whereas S is uncharged but hydrophilic. The mutation from GLPL to ELPL (aliphatic to acidic amino acid residue) was shown to abolish P2 resistance in flax (Dodds *et al.*, 2001) showing that this motif is essential in determining recognition specificity. However, it is not yet known how it affects functionality between TIR and non-TIR subfamilies. However, using Bioinformatics these differences provide a useful tool for confirming the distinction between TIR and non-TIR subfamilies in a set of sequences.

Table 9. Conserved amino acid motif patterns in TIR and non-TIR subfamilies. The yellow colour highlights polymorphic sites.

Motif	TIR	Non-TIR
P-Loop motif/Kinase-1	GMGGRGKTTIA	M/DGGGGKTTLL
Kinase-2	LIILDDVD	LLFDDIWS
	LIVLDDVD	LIVMDDVWS
Kinase-3/RNBS-B	GSRIIITSRD	GSSVIITTRI
	GSRIIITTRD	KSKIIFTTRS
RNBS-A	FLANVREVTE/GK	FDLVIWIVVSK
	FLDNVKEEFA-C	FERRIWVSVSQK
GLPLA	GLPLA	GLPLA
	ASPSP	

Figure 3.6. Multiple sequence alignment of TIR and non-TIR sequences showing differences in motif structure. The yellow colour is used to highlight conserved motifs that show some differences between TIR and non-TIR subfamilies. A combination of yellow and green denotes motifs where no significant similarities are present and the red colour shows similarities between some and not all sequences (the later is used here in a chosen region to highlight more differences between the TIR and non-TIR subfamilies).


P-loop
RNBS-A
IGD56 GMGGRGKTTIAEVVFD-RIRSRFDAYSFLANVREVTGKQGLVHLHKQLLSDILFESSVD
IGD123 GMGGGGKTPIAEVVFD-RIRSRFDAYSFLANVREVTEKQGLVHLHKQLLSDILFESSVD
IGD141 GMGGRGKTTIAEVVFD-RIRSQFDAYSFLANVREVTEKQGLVHLQKQLLSDILFESSVD
IGD173 GMGGGGKTTIAEVVFD-RIRSQFDAYSFLANVREVTEKQGLVHLQKQLLSDILFESSVD
IVGD243 GMGGGGKTTIARAVYE-KIACQFEACCFLDNVKEEFA-CGAVHMQEKFLSRILNEKVQS
IVGD90 -DGGGGKTTIARAVYE-KIACQFEACCFLDNVKEEFA-CGAVHMQEKFLSRILNEKVQS
IVGD9 GMGGRGKTTIARAVYE-KIACQFEACCFLDNAKEEFA-CGAVHMQEKFLSRILNEKVQS
IVGD131 GMGGRGKTTIARAVYE-KIACQFEACCFLDNVKEEFA-CGAVHMQEKFLSRILNEKVQS
IIGD28 GMGGRGKTTLLTQINN-KLLHADFDLVIWIVVSKDHNV---ETVQDKIGDKIGFSSISW
IIGD14 GMGGRGKTTLLTQINN-KLLHADFDLVIWIVVSKDHNV---ETVQDKIGDKIGFSSISW
IIGD60 -DGGGGKTTLLTKINN-KLLHADFDLVIWIVVSKDHNV---ETVQDKIGDKIGFSSISW
IIGD35 -MGGGGKTTLLTQINN-KLLHADFDLVIWIVVSKDHNV---ETVQDKIGDKIGFSSISW
IIIGD138 -DGGEGEDDYCSKAFNDRKIEERFERRIWVSVSQKFDE---EQIMRSILRNLGDASVGD
IIIGD262 GMGGGGKTTIAQKVFNDRKVEERFERRIWVSVSQKFDE---EQIMRSILRNLGDASVGD
IIIGD18 GMGGGGKTTIAQKVFNDRKIEERFERRIWVSVSQKFDE---EQIMRSILRNLGDASVGD
IIIGD145 GMGGGGKTTIAQKVFNDRKIEECFERRIWVSVSQKFDE---EQIMRSILRNLGDASVGD
TIR
non-TIR
Kinase-2
RNBS-B/Kinase-3
IGD56 VHNIHMG-INKIRQRLRTRMVLIILDDVDQLEQLEALCG--HSWFGSGSRIIITSRDEHL
IGD123 VHNIHMG-INKIRQRLRTRMVLIILDDVDQLEQLEALCG--HSWFGSGSRIIITSRDEHL
IGD141 VHNIHMR-ISKIRQRLRARMVLIILDDVDQLEQLEALCG--HSWFGSGSRIIITSRDEHL
IGD173 VHNIHMR-ISKIRQRLRARMVLIILDDVDQLEQLEALCG--HSWFGSGSRIIITSRDEHL
IVGD243 LGTLDRG-YRMILKRLQMKKVLIVLDDVDDLFQIETLLGK-QHSFGGGSRIIITTRDKLV
IVGD90 LGTLDRG-YRMILKRLQMKKVLIVLDDVDDLFQIETLLGK-QHSFGGGSRIIITTRDKLV
IVGD9 LGTLDRG-YRMILKRLQMKKVLIVLDDVDDLFQIETLLGK-QHSFGGGSRIIITTRDKLV
IVGD131 LGTLDRG-YRMILKRLQMKKVLIVLDDVDDLFQIETLLGK-QHSFGGGSRIIITTRDKLV
IIGD28 KQKQQSDKAEHICRLLSKKKFVLLFDDIWEPIEITKLGVPIPNP-HNKSKIIFTTRSEDV
IIGD14 KQKQQSDKAEHICRLLSKKKFVLLFDDIWEPIEITKLGVPIPNP-HNKSKIIFTTRSEDV
IIGD60 KQKQQSDKAEHICRLLSKKKFVLLFDDIWEPIEITKLGVPIPNP-HNKSKIIFTTRSEDV
IIGD35 KQKQQSDKAEHICRLLSKKKFVLLFDDIWEPIEITELGVPIPNP-HNKSKIIFTTRSEDV
IIIGD138 ---DKGELLKKINEYLLGKRYLIVMDDVWSLDVTWWLRIYEALPKGNGSSVIITTRIEKV
IIIGD262 ---DKGELLKKINEYLLGKRYLIVMDDVWSLDVTWWLRIYEALPKGNGSSVIITTRIEKV
IIIGD18 ---DKGELLKKINEYLLGKRYLIVMDDVWSLDVTWWLRIYEALPKGNGSSVIITTRIEKV
IIIGD145 ---DKGELLKKINEYLLGKRYLIVMDDVWSLDVTWWLRIYEALPKGNGSSVIITTRIEKV
RNBS-C
IGD56 LCTYGVDKMY--KVKPLTDDEALQLFCRKAFKKDQV---GEDFLELSKNVVEYANGLPLA
IGD123 LCTYGVDKMY--KVKPLTDDEALQLFCRKAFKKDQV---GEDFLELSKNVVEYANGLPLA
IGD141 LRTYGVDKMY--KVKPLTDAEALQLFCRKAFKKDQV---GEDFLELSKNVVEYANGLPLA
IGD173 LRTYGVDKMY--KVKPLTDAEALQLFCRKAFKKDQV---GEDFLELSKNVVEYANGLPLA
IVGD243 LSR--ADAIY--SPKVLSGDGALELFSQYAFRTKQP---KRDYDLSQDVLYDMLKASPSP
IVGD90 LSR--ADAIY--SPKVLSGDGALELFSQYAFRTKQP---KRDYDISQDVLYDMLKASPSP
IVGD9 LSR--ADAIY--SPKVLSGDGALELFSQYAFRTKQP---KRDYDISQDVLYDMLKASPSP
IVGD131 LSR--ADAIY--SPKVLSGDGALELFSQYAFRTKQP---KRDYDHLSRRAVRYAQGLPLA
IIGD28 CGQMDAHKK--IKVECLAWDKAWNLFQEKVGRETLG--IHPDIQRLAQTVAKECGGLPLA
IIGD14 CGQMDAHKK--IKVECLAWDKAWNLFQEKVGRETLG--IHPDIQRLAQTVAKECGGLPLA
IIGD60 CGQMDAHKK--IKVECLAWDKAWNLFQEKVGRETLG--IHPDIQRLAQTVAKECGGLPLA
IIGD35 CGQMDAHKK--IKVECLAWDKAWNLFQEKVGRETLG--IHPDIQRLAQTVAKECGGLPLA
IIIGD138 AQKMGVKEARSHWPKCLSKDDSWLLFRKIAFAENGGECMYPELENVGKEIVEKCKGLPFA
IIIGD262 AQKMGVKEARSHWPKCLSKDDSWLLFRKIAFAENGGECMYPELENVGKEIVEKCKGLPLA
IIIGD18 AQKMGVKEARSHWPKCLSKDDSWLLFRKIAFAENGGECMYPELENVGKEIVEKCKGLPLA
IIIGD145 AQKMGVKEARSHWPKCLSKDDSWLPFRKIAFAENGGECMYPELENVGKEIVEKCKGLPLA

3.3.4.3 Classification of genes into clusters

Clustering of R-genes is a common phenomenon that results from tandem duplications of gene paralogs (Meyers *et al.*, 2005). The continual challenge by pathogens plays a significant role in facilitating natural selection, the pressure thus generated contributes to homogenisation of the cluster and as a result sequences from the same cluster group together under phylogenetic reconstruction. These clusters probably represent a set of genes that are located on the same locus on a chromosome and can be any number of genes above two (Mondragon-Palomino and Gaut, 2005). As shown earlier in Figure 3.5, the classification into respective clusters was achieved through the Bayesian inference of phylogeny.

The percentage reliability of tree branching for all 12 resolved clusters were in the range 90 – 100%. As mentioned in section 3.3.4.2 the tree was generated using a rigorous phylogenetic inference method and thus clusters were assumed represent the clustering found on the apple genome. The model used in generating the phylogenetic tree in Figure 3.5 allowed both nucleotide substitution rate variation and a proportion of invariable sites (section 3.3.4.2). This allowed the analysis to be sensitive to conserved motifs, which subsequently become identified as invariable sites. The percentage support for the 5 major branches ranges from 75 – 89% (Figure 3.5).

Two 'clusters' indicated as 7 and 9 in Figure 3.5 represent unresolved internal nodes with long branches and both are positioned at the 2 extremes of the phylogenetic tree. Homology searches were performed again for sequences in these nodes against the NCBI

non-redundant database and all of them matched NBS-LRR resistance genes confirming that they are indeed RGAs that are either single genes occupying isolated loci or could be paralogs that arose from one ancestor through a single gene duplication event. Table 10 shows the number of clusters generated from section 3.3.4.2 and the number of RGAs for each of the clusters using the order adopted in Figure 3.5.

Table 10. Classification of RGAs into clusters and the number of functional to pseudogenes per cluster. Identification of functional to pseudogenes is explained in section 3.3.4.4.

Cluster	Number of sequences	Functional genes	Pseudo- genes
1	16	11	5
2	32	24	8
3	7	4	3
4	11	11	0
5	4	0	4
6	9	6	3
8	17	14	3
10	39	27	12
11	53	2	51
12	19	5	14
13	20	7	13
14	9	7	2

3.3.4.4 Prediction of open reading frames

According to Michelmore and Meyers (1998), pseudogenes make up a potential reservoir of useful elicitor recognition specificity variations. These genes are said to arise by unequal crossover, recombination between paralogs, other spurious mutations and gene conversions in regions other than those determining specificity. Research on VH pseudogenes in chicken showed that chicken has one functional gene coding for the heavy chain variable region of immunoglobulins, V_H locus and 80 pseudo- genes located at the 5' side of V_H1 and that diversity is generated by gene conversion between this gene and the pseudogenes (Ota, 1995). It was also shown in mice that a pseudo- gene, *Makorin1-p1* (AB219438) consisting of four stop codons within the best coding region (forming two ORFs 126 and 804 base pairs) had a stabilizing effect on the expressed transcripts of the corresponding functional gene (Hirotsune *et al.*, 2003). Based on these apparent roles of pseudogenes in other organisms it was equally crucial that they be carefully analysed to eliminate mis-characterisation due to errors in sequence chromatograms (traces).

In this analysis, genes were preliminarily characterised into 'functional' and 'pseudo' genes based on the occurrence of at least one termination codon in the best ORF. Sequence Analysis software was used as the starting point in estimating the coding potential of the RGA sequences. The most important assumption in this analysis was that the PCR amplified NBS domains of candidate R genes are contained within a coding sequence. This means the presence of termination codons within this domain results in truncated protein fragments and as such the gene becomes a pseudo-gene. RGAs giving

complete or uninterrupted ORFs were re-analysed using the GETORF program (section 2.13.2.1). This program computes ORFs in all the six frames and outputs the best translation in a frame that matches the set parameters including the expected minimum size of the translated amino acid sequence. Comparison of Sequence Analysis and GETORF results allowed a confirmation of the sequences classified as either functional or pseudo- genes. Table 10 in section 3.3.4.3 shows the number of functional against pseudogenes per cluster.

According to results in Table 10, the ratio of functional to pseudo-genes varies per cluster. The percentage of pseudogenes ranges from zero for cluster 4 to 100% for cluster 5. This percentage representation of pseudogenes per cluster is irrespective of the total number of RGAs in the clusters. Cluster 11 has 53 RGAs with only 2 as functional and 51 (96%) as pseudogenes, whereas cluster 5 has 4 RGAs all of which are pseudogenes. One example of an RGA from the set of functional genes was used in the web-based NCBI ORF Finder program to assess the primary structure of the sequences against the database and the result is shown in Figure 3.7 shows the result of the ORF search and the subsequent translation into an amino acid sequence.

One common observation in most of the RGA amino acid sequences is the random occurrence of methionine residues within the NBS domain. In *Arabidopsis* they form part of the highly conserved MHDV motif whose function has not been deciphered to date (Chini and Loake, 2005).

Figure 3.7. NCBI ORF Finder-generated domain sequence of an uninterrupted RGA ORF. RGA sequence GD203 from cluster 2 was used here.

The sequence of the NBS domain given in Figure 3.7 was subsequently searched against the NCBI CDD pfam conserved domain database and Figure 3.8 shows its alignment to the best nine matching proteins. The hits include *Arabidopsis* RPS2 (gi30173240) protein specifying resistance to *Pseudomonas syringae* 2, gi46395938 and gi46395604, which are annotated as probable disease resistance proteins. The I2C-2 protein located on

the wilt disease resistance locus (gi75318196), gi75102252 which is the L6tr R protein for flax rust resistance related to the *Arabidopsis* bacterial resistance gene RPS2 and the tobacco viral resistance gene N (Lawrence *et al*., 1995), giIZ6T_A chain A of the apoptotic protease-activating factor 1 that is predicted to have a DEATH (pfam00531) domain-like fold (Riedl *et al*., 2005), gi29839510 (RPM1) *Pseudomonas syringe* protein 3, gi29839509 (RPP13-like) and RPP8-like protein 1 (gi29839442) (Marchler-Bauer *et al*., 2005). This was further confirmation that the genes sequenced in section 3.2 were indeed NBS type resistance gene analogs.

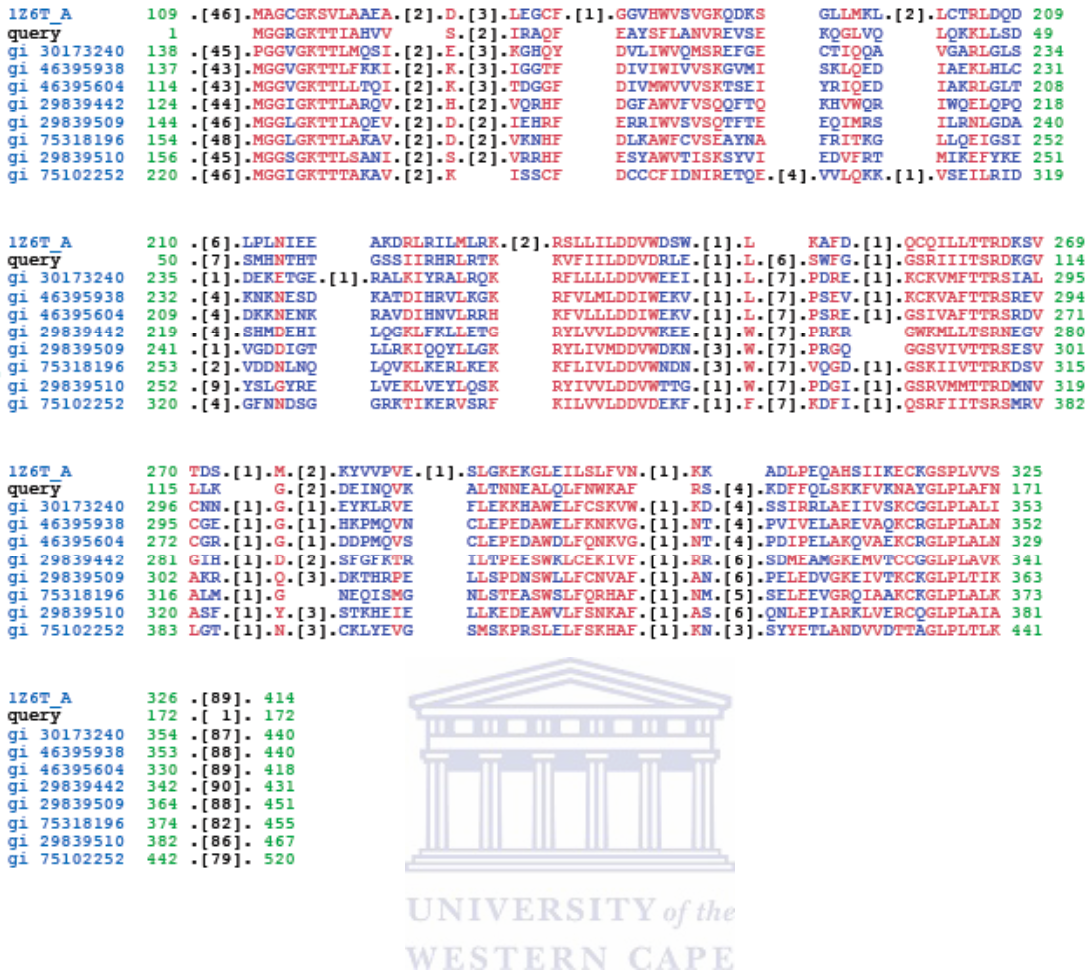


Figure 3.8. The NCBI CDD pfam alignment of RGA GD203 with 9 closely related genes containing the NBS-ARC domain (Marchler-Bauer *et al.*, 2007). The red colour scheme indicates amino acid residues that are highly conserved at those sites and blue represents those that are not.

3.3.5 Assessment of the non-synonymous /synonymous values per cluster

The ratio of non-synonymous (Ka) to synonymous (Ks) rates of substitution gives an estimate of the selective pressure exerted on a cluster of genes. The ratio of Ka/Ks > 1

represents evidence of positive selection in the history of the cluster, Ka/Ks < 1 is purifying selection and Ka/Ks = 1 shows neutral selection. Random coding substitutions alter the primary structure of proteins and thus have a high chance of introducing deleterious mutations that could affect an organism's fitness. However, they could also be beneficial in generating new specificities for protein – protein interactions. Genes in clusters with a history of purifying selection most likely encode functional R proteins and the occurrence of positive or coding selection might lead to deleterious mutations that could abolish protein function. This means clusters with evidence of purifying selection should be identified as candidates for analysis for potential R genes of interest. Neutral selection assumes evolutional stability in which the ratio of coding to non-coding mutation is balanced. For these sequence data the assessments of the Ka/Ks ratios were calculated per cluster using methods provided by Datamonkey (section 2.13.2.1).

The single likelihood ancestor counting (SLAC) and random effects likelihood (REL) methods were used. The SLAC method is a modified derivative of the Suzuki-Gojobori counting approach. It uses tree topologies and branch lengths along with a codon-based substitution model to reconstruct the ancestral sequences and infer the number of changes that have taken place along the phylogeny. The REL method, which is an improvement of the Nielsen-Yang methodology, allows the Ka/Ks rates to vary across sites independently thereby achieving a reduction in inference of false positives. All the methods used the HKY85 model (Hasegawa *et al.*, 1985) of codon substitution bias. Ka/Ks values are presented in Table 11, selection per site and the associated P-values are provided.

Table 11. **The non-synonymous to synonymous rates of substitution for RGAs per cluster**. Clusters are based on the order set up in Figure 3.5 with cluster 10 split to 10a (subset with partially resolved branches) and 10b, based on the topology of the branching structure. Positive, Purifying and Neutral indicates the selection pressure detected in respective clusters.

Cluster	Global Ka/Ks	Comment
1	1.034	Neutral
2	2.382	Positive
3	2.086	Positive
4	2.499	Positive
6	0.817	Purifying
8	0.739	Purifying
10a	0.651	Purifying
10b	1.065	Neutral
12	1.920	Positive
13	0.805	Purifying
14	1.323	Positive

Results in Table 11 show that clusters 2, 3, 4, 12 and 14 are undergoing positive selection. The extent of this selection pressure increases from cluster 3 to 2, 4 and 12 with 14 being the lowest in that series. Clusters 6, 8, 10a and 13 show evidence of purifying selection in which 10a is more pronounced followed by 8 and 6 in that order. Clusters 1 and 10b have ratios of the order Ka/Ks ≈ 1, this could mean near-neutrality and that there are sites undergoing positive selection though overall these balance out with sites under synonymous substitutions (Chamary *et al.*, 2006). Clusters 1 and 10b have ratios of 1.034 and 1.065 respectively.

Clusters 7 and 9 were not analysed here since they are made up of RGAs that do not form strict clusters but rather groups of unrelated genes. Clusters 5 and 11 with zero and two functional genes could not be analysed due to limitations in the analysis program. Only clusters with 5 or more functional genes could be used. Unbiased calculation of nonsynonymous to synonymous mutations uses sequences that are not interrupted by termination codons within the ORF. Secondly the reliability of the results obtained increases with the increase in number of sequences per analysis (Pond *et al.*, 2005).

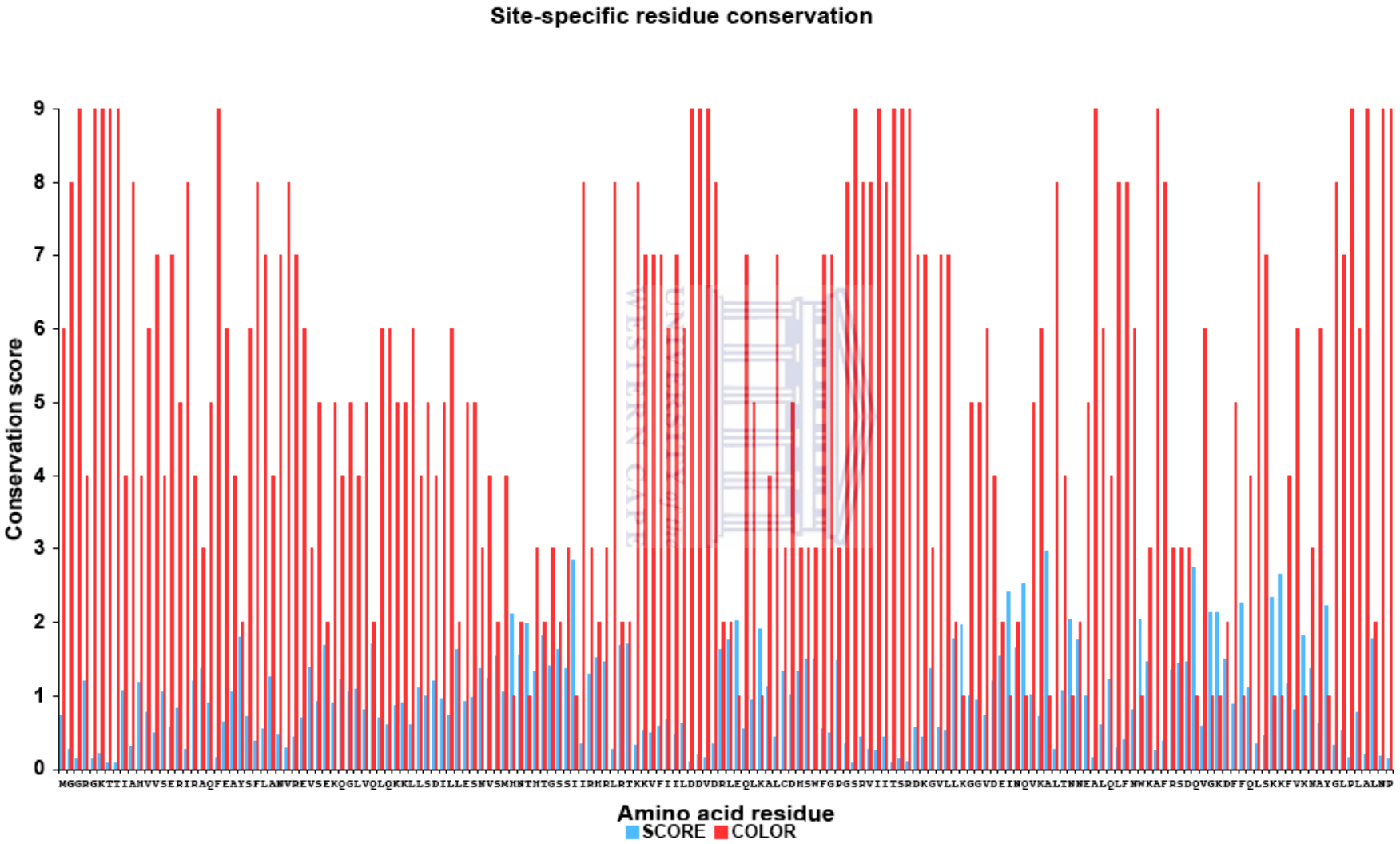
3.3.6 Site – specific inference of amino acid conservation in the NBS domain

The NBS domain is approximately 520 bases long (about 174 amino acid residues) and is made up of a series of well-characterised motifs (Pal *et al.*, 2007). Based on results shown in Table 11, out of the 11 clusters tested, 5 show evidence of positive selection. As discussed in section 3.3.5, mutations that affect the amino acid primary structure challenge the fitness of an organism. To assess codons that have the highest probability of being mutated, a site-to-site inference of conservation was performed using McRate software (section 2.13.2.1). For this analysis, a total of 20 representative amino acid sequences (section 3.3.4.4) selected from the clusters in Figure 3.5 were used.

McRate is a Bayesian method that uses Markov Chain Monte Carlo methodology to assess all possible phylogenetic trees and model parameters. The program calculates the evolutionary rate of each site and normalizes it using the average rates across all sites (Mayrose *et al.*, 2005). A total of 12800 inference cycles were performed using the JTT model of amino acid replacements (Jones *et al.*, 1992; Mayrose *et al.*, 2005) with among-site rate variation set to gamma. Rates were inferred every 10 cycles to give a total of 1280 inferences. Results were plotted into a graph and are shown in Figure 3.9.

Using results of Figure 3.9, the $G_8L_7P_9L_6A_9L_2N_9P_9$ is invariable on amino acid residues GxPxAxxNP where x denotes positions that are variable. The P-loop motif, $M_6G_8G_9R_4G_9K_9T_9T_9I_4A_8$ is invariable at xGGxGKTTxA. According to Dodds *et al.* (2001), a mutation in the GLPL motif (G to E) was enough to abolish rust resistance in flax.

Figure 3.9. A plot of the site-specific amino acid evolutionary rates in the NBS domain. The amino acid residues are plotted in a sequence on the x-axis, the red bars represent estimates of relative conservation and blue bars represent site-specific rate variation scores. The two quantities are inversely proportional to each other at conserved and variable sites.


Site-specific residue conservation
Conservation score
0
1
2
3
4
5
6
7
8
9
MGGRGKTTIAMVVSERIRAQFEAYSFLAMVREVSEKQGLVQLQKKLLSDILLESMVSMMNTMTGSSIIRMRLRTKKVFIILDDVDRLEQLKALCDMSWFGPGSRVIITSRDKGVLLKGGVDEIMQVKALIMNEALQLFMWKAFRSDQVGKDFFQLSKKFVKMAYGLPLALNP
Amino acid residue
SCORE COLOR

3.3.7 Detection of gene conversion

The analysis of gene conversion was performed using the methods implemented in GENECONV version 1.81 (Sawyer, 1999). Gene conversion is defined as any process that causes a segment of DNA to be copied onto another segment of DNA (Sawyer, 1989). The program scans for gene conversion by finding identical fragments between pairs of sequences in an alignment. Pairwise and global P values are calculated to assess the statistical significance of the observed gene conversion event. Evidence for gene conversion is taken from the global P values which are from comparisons between each high-scoring aligned pair (HSAP) with all other possible HSAP fragments for the entire alignment (Sawyer, 1999). In this method pairwise P values represent the proportion of permuted alignments for which the maximum HSAP score for a given pair is greater than or equal to the original fragment score. Figure 3.10 shows an example from the gene conversion event detected in cluster 1 between GD70 and 140.

Table 12 shows GENECONV results obtained in the cluster specific analysis of gene conversion. The table gives global P values for the significant inner fragments; inner fragments give evidence of a possible gene conversion event between ancestors of two sequences in the alignment.

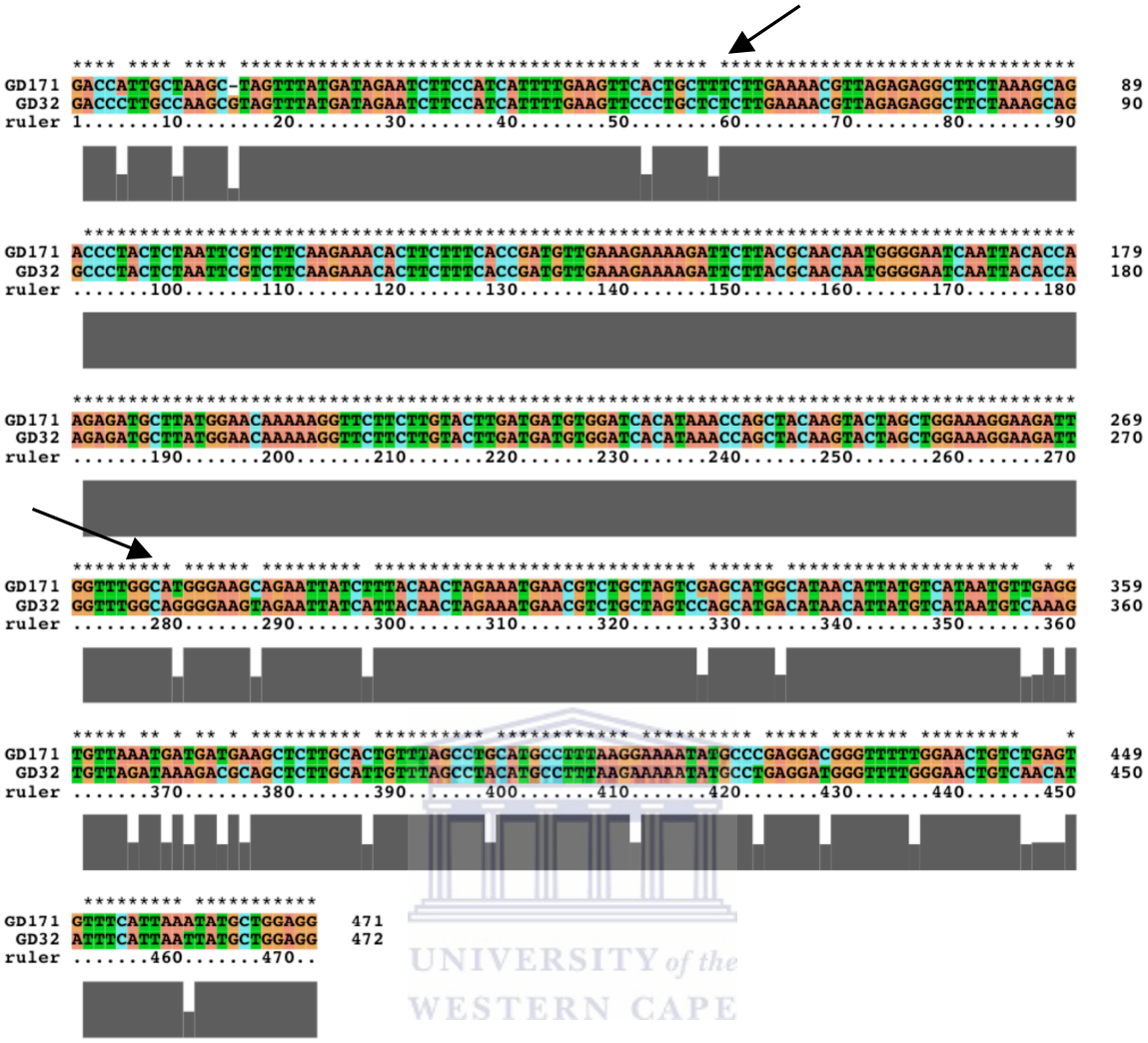


Figure 3.10. Evidence of gene conversion between GD32 and GD171 of cluster 10b. The HSAP for these two sequences is flanked by arrow 1 and 2; the quality bar graph indicates a perfect match in this region.

The two RGA sequences shown in Figure 3.10, GD32 and GD171 share a region of perfect alignment from position 59 to 281 although this is 114 to 310 with priming regions. According to results in Table 12, there is significant evidence of gene conversion (99.9% confidence interval or p-value = 0.001) in a region of 197 bases between GD32

and GD171. GENECONV considers all events with multiple-comparison corrected P-values of 0.05 or smaller meaning from 95% confidence interval and above as confirmation of gene conversion. All the clusters in Table 12 show gene conversion although the most significant evidence is in cluster 10b (the highly resolved branch of cluster 10 as shown in Figure 3.5) and between GD182:GD239 in cluster 6, GD20:GD105 in cluster 10a, GD6:GD179 in cluster 12 and GD110:GD220 in cluster 13.

Gene conversion events in clusters 5, 8 and 13 represent exchanges between RGAs whose ORFs are interrupted by termination codons. Events in clusters 1 and 10b on the other hand are entirely between RGAs whose ORFs are not interrupted by termination codons. The rest of the clusters in Table 12 have evidence of gene conversion events between RGAs interrupted and not interrupted by termination codons. However, two out of the most significant nine cases are between RGAs with and without termination codon interruptions. In all clusters where gene conversion was detected it appears to be a random process between any two RGAs regardless of whether they are functional or pseudo-genes. Clusters 2, 3, 11 and 14 had no significant events above the 90% confidence interval. Table 12 shows a tabulated form of the data discussed here.

Table 12. Evidence of gene conversion between genes in the same cluster. The table shows identities of sequences in significant pairwise comparisons; Simulated and BC KA P-values. 'Begin' and 'End' denote the gene conversion breakpoints and 'Length' of the fragments are also provided. F and P represent functional and pseudo-gene respectively.

Clusters	Sequence Names		Sim Pvalue	BC KA Pvalue	Aligned Offsets		
					Begin	End	Length
1	GD140;GD70	F;F	0.09	0.90145	1	282	282
4	GD65;GD138	F;F	0.044	> 1.0	42	519	478
5	GD75;GD273	P;P	0.054	0.15182	240	436	197
6	GD182;GD239	F;F	0.021	0.18477	1	53	53
	GD270;GD178	P;F	0.052	0.23828	1	53	53
8	GD13;GD59	P;P	0.024	0.41668	45	303	259
	GD149;GD59	P;P	0.024	0.41668	45	303	259
	GD214;GD59	P;P	0.024	0.41668	45	303	259
10a	GD20;GD105	P;F	0.004	0.22204	114	403	290
	GD20;GD237	P;P	0.035	0.89712	265	419	155
	GD20;GD187	P;F	0.039	> 1.0	114	423	310
10b	GD171;GD32	F;F	0.001	0.0278	114	310	197
	GD38;GD223	F;F	0.001	0.03109	147	327	181
	GD147;GD223	F;F	0.005	0.07115	312	403	92
	GD143;GD34	F;F	0.005	0.07549	14	242	229
	GD34;GD233	F;F	0.006	0.09045	449	497	49
	GD38;GD223	F;F	0.027	0.20643	28	145	118
	GD143;GD47	F;F	0.032	0.24074	14	273	260
	GD47;GD233	F;F	0.034	0.2594	312	403	92
	GD32;GD223	F;F	0.041	0.33086	312	403	92
	GD232;GD223	F;F	0.048	0.34643	449	479	31
12	GD6;GD179	F;P	0.006	0.0772	259	367	109
	GD170;GD58	P;F	0.012	0.11863	110	186	77
	GD170;GD302	P;F	0.012	0.11863	110	186	77
	GD170;GD94	P;P	0.014	0.14966	110	186	77
13	GD110;GD220	P;P	0.005	0.05783	82	144	63

3.3.8 Analysis of gene duplication in the NBS-LRR R genes

Using results of Figure 3.5 all branches displaying polytomy (hard or soft) at the terminal nodes and composed of al least 4 RGAs were pruned to reduce the number of sequences used in this analysis. The justification here is that sequences showing a very high level of percentage identity do not add value to the reconstruction of the phylogenetic path. Therefore removing them from the analysis does not adversely affect the tree topology in the final result.

This process reduced the number of sequences from 265 to 174 that were subsequently aligned in ClustalX. Modeltest version 3.8 (Posada and Crandall, 1998; Posada, 2006) was used to select the model that best fits the RGA sequence data. Out of the 56 models tested using PAUP*4 and evaluated on the Modeltest server, HKY85+G was selected using hierarchical likelihood ratio tests (hLRTs) with the best negative $\log_e$ likelihood score of 19975.9004.

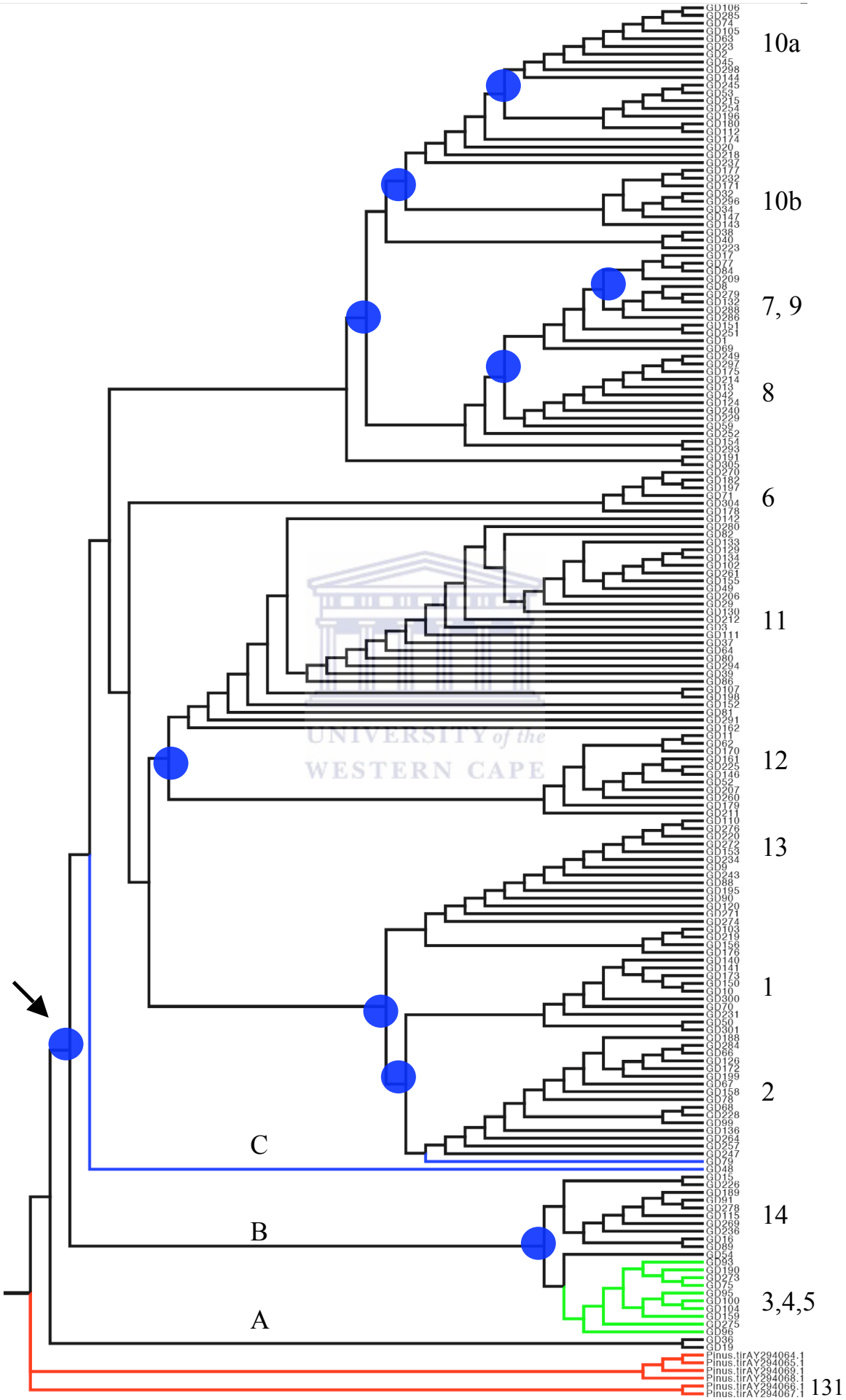
To reconstruct the phylogenetic history of RGAs using the maximum likelihood algorithm a starting tree was constructed using parsimony with topology optimisation done through sub-tree pruning and re-grafting (SPR). The HKY85+G model of substitution was then fitted to the data to estimate base frequencies, the transition/ transversion (Ti/tv) ratio and the gamma distribution shape parameter. Distance models optimise branch lengths and thus improve inference of genetic distances in molecular evolutionary studies. Thus estimates obtained here were incorporated in inferring the phylogenetic tree under the maximum likelihood (ML) criterion. The tree was rooted

using 6 *Pinus monticolor* RGAs as the outgroup. The phylogenetic tree was displayed and manipulated using Mesquite (section 2.13.2.1) and is shown in Figure 3.11.

This analysis shows a set of putative gene duplication points (indicated by blue circles in the tree) that could have resulted in increasing the observed gene copy number in the NBS-LRR family. Branch lengths are optimised through the use of the HKY85 + G distance matrix. The 'cluster' indicated using the black perpendicular line in Figure 3.11 is made up of sequences that do not show up as a cluster in the unrooted MrBayes tree (Figure 3.5). In Figure 3.11 this group has one of the shortest branch lengths, which could mean that they are a more recent product of gene duplication. It is important to note here that generation of clusters is a process that takes millions of years and not in the life cycle of the apple trees.

The most accurate inference of gene duplication points in the generation of both the copy number and clusters in apple would require a correlation of the gene and species tree. Such a high level analysis is not within the scope of this work, however, Figure 3.11 does help in identifying clusters that are most likely pro-orthologues. Branches A (GD36 and GD19) and B (clusters 14 and 3/4/5) in Figure 3.11 appear to be predate all the other clusters in the family. Branch B composed of TIR and non-TIR (the later indicated in green) appear to belong to the earliest genes that gave rise to the NBS-LRR family based on results of Figure 3.11. However, such an assumption would require the use of full length NBS-LRR and rigorous analyses methods to confirm.

Figure 3.11. Estimating molecular evolution and gene duplication in the NBS-LRR gene family. The red branches represent the *Pinus monticolor* RGA outgroup, green represents the non-TIR subfamily and blue shows RGAs that are intermediate between TIR and non-TIR based on the kinase 2 motif (section 3.3.4.2). The numbers represent clusters based on results of Figure 3.5. The black arrow indicates the putative earliest gene duplication event that gave rise to TIR and non-TIR subfamilies.



According to this tree, the non-TIR subfamily belongs to the earliest NBS-LRR genes that most probably gave rise to the defined clusters of R genes through tandem duplications. Some of the RGAs in the non-TIR subfamily differ in the characteristic RNBS-A, kinase-2 and kinase-3 motifs. GD93, 273, 190 and 75 have the RNBS-A, kinase-2 and kinase-3 as FNCCAWITVSK, SKRLVVVLDDVWD and GSRIILTTRNED respectively. The sequence valine-tryptophan-aspartate in the kinase-2 motif differs from the highly conserved valine-tryptophan-serine of the other non-TIR genes as shown in Table 9. The C-terminal amino acid residue of the kinase-2 motif is said to distinguish with 95% certainty whether a gene belongs to the non-TIR (tryptophan) or TIR (aspartate) subfamily (Baldi *et al*., 2004). Consequently these non-TIR genes form a cluster that is isolated from the other non-TIR genes with the kinase-2 motif as NKKVLLVLDDVWS.

The majority of the TIR genes after branch B in Figure 3.11 have the sequences KKVFIILDDVD/ KKVLRLDDVD for the kinase-2 motif with the exception of GD48, which has the sequence KKALLSMVI. These changes in amino acid primary structure not only confirm divergence of genes through coding mutations but also suggest possible alterations in function. Cluster 14 and the non-TIR subfamily show up early in the tree as progeny from the putative gene duplication indicated by the black arrow in Figure 3.11, consequently the *Arabidopsis* control genes RPM1 (cluster 14), RPS2 and RPS5 (cluster 3,4,5) are found in these clusters (Figure 3.5).

3.4 DISCUSSION

Cloning and sequencing of RGAs was based on a set of degenerate oligonucleotides targeted at the Walker-A/ P-loop, GGVGKTT, and the hydrophobic GLPL motifs (Lee *et al.*, 2003; Baldi *et al.*, 2004; Xu *et al.*, 2005; Bozkurt *et al.*, 2007). The P-loop motif is highly conserved in all nucleotide-binding proteins that bind GTP/ ATP for hydrolysis of the γ-phosphoryl group (Pfeifer *et al.*, 2001; Pal *et al.*, 2007). GLPL/GxP is a hydrophobic motif located on the C-terminal end of the NBS domain of NBS-LRR proteins, ATPase and ABC transporters (D. Peter Tieleman, 2001; Takken *et al.*, 2006). In ATPases and ABC transporters however, it is present as GxP with high conservation focus on glycine and proline in the transmembrane helix (D. Peter Tieleman, 2001). This motif however, has been shown in literature as highly conserved in all NBS-LRR proteins (Tao *et al*., 2000) hence its use as the priming region for PCR amplification. There are other members of the NBS-containing proteins that have a highly conserved GLPL motif in non-plant organisms such as CARD12 (Ipaf1/Clan), NOD1, NOD2 and Apaf-1 that play important roles in programmed cell death, cytokine processing and activation of NF-κB (Lu *et al*., 2005). The level of structural similarity with these proteins is shown in Figure 3.12. Based on these factors it was reasonable to expect selective amplification of R-gene NBS domains in plants with no co-amplification of ATPases/ GTPases or other proteins that possess an ABC domain.

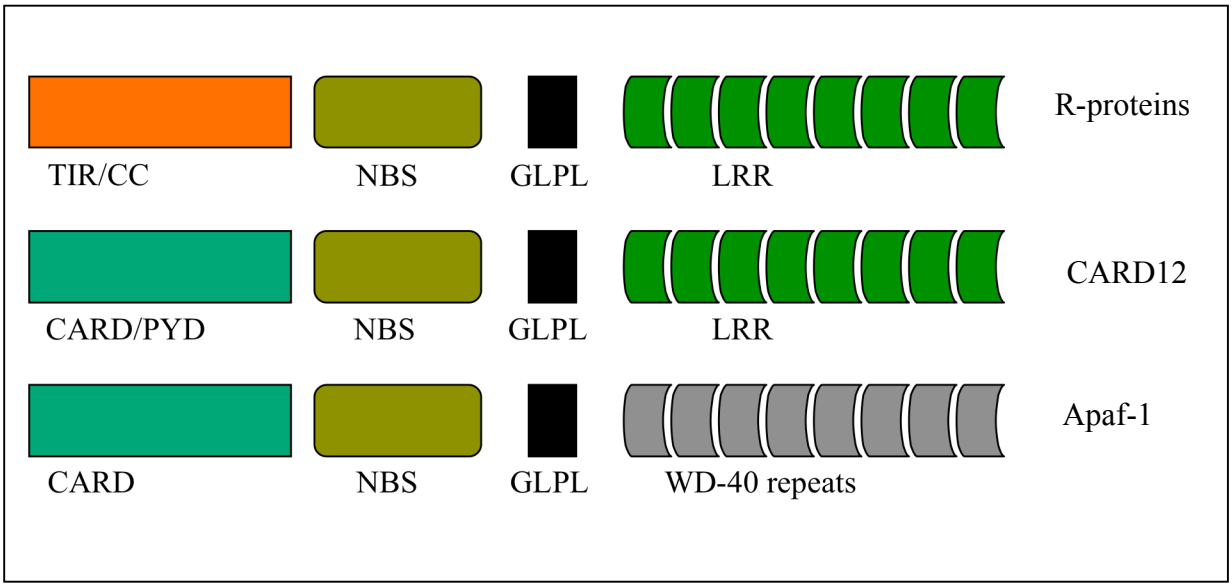


Figure 3.12. **Graphical comparison of domain structures in members of the NBS-LRR protein family.** Apaf-1 has WD-40 repeats where CARD12 and R-proteins have an LRR domain although. This figure was constructed using evidence from literature (Y. Hu *et al.*, 1998; Lee *et al.*, 2003; Lu *et al.*, 2005).

Inserts with sizes corresponding to the PCR amplified RGAs were identified by colony PCR screening (section 2.8.3) and these were then subsequently sequenced. Sequences were characterised initially through Blast homology searches against the NCBI non-redundant database (Altschul *et al.*, 1997). A series of multiple sequence alignments were performed using ClustalX followed by manual editing (section 2.13.2.1), each round of alignment and editing enabling parallel editing of artifacts in sequence chromatograms. Priming sites were removed to eliminate bias in multiple alignments that were subsequently used for phylogenetic analyses, determination of selection pressure and

gene conversion events. Given that degenerate primers were used for the amplification, it is reasonable to assume that in some cases primer hybridization to target regions were not fully complimentary and as such a certain level of sequence inaccuracies might have been incurred in the PCR amplification process.

The inference of phylogeny performed using MrBayes (section 3.3.4.2) resulted in the division of sequences into TIR and non-TIR subfamilies (Figure 3.5). The number of TIR sequences was almost double that of the non-TIR members (clusters 3, 4 and 5 in Figure 3.5). This broad division is based on a number of striking differences in the conserved motifs, the most diagnostic of which is the kinase-2 motif. The kinase-2 motif of the TIR subfamily has the signature sequence LI(I/V)LDDVD that is replaced by L(I/L)(V/F)(-/M)DD(V/I)W in the non-TIR subfamily (Table 9). The presence of either an aspartate/ tryptophan residue (D/W) in the kinase-2 motif constitutes one of the distinguishing criteria between the two subfamilies. There are also relatively conserved differences in the kinase-3 domain with GSRIIIT(T/S)RD exclusively conserved in members of the TIR subfamily whereas non-TIR members were divided between GSSVIITTRI and KSKIITTRS. Results obtained from the PROSITE and PRINTS motif scanning methods (section 3.3.4.2) show that the important sequence in the kinase-3 motif is the TTR (PROSITE accession number PDOC00005), which is defined as a protein kinase C phosphorylation site (Kishimoto *et al.*, 1985; Woodgett *et al.*, 1986). There are other significant differences in the primary structures of the conserved motifs between TIR and non-TIR as shown in section 3.3.4.2.

The TIR and non-TIR subfamilies were further subdivided into clusters based on sequence homology (Figure 3.5 and 3.11). According to Mondragon-Palamino and Gaut (2005), a cluster is made up of sequences of genes that are probably physically located at the same locus on a chromosome. This is an attractive model that allows for cluster specific tandem gene duplication events to increase numbers of possible pathogen recognition specificities. Under this model it would be reasonable to assume that the largest clusters are those that respond to endemic pathogen species. Selection of mutation-induced sequence variability could lead to generation of new R gene specificities. Exposure of these clusters to repeated rounds of equal and unequal crossover and recombination events could then cause a wave of homogenization through concerted evolution (Leister, 2004). A number of pseudogenes were also identified in clusters generated through the introduction of premature termination codons. According to Michelmore and Meyers (1998) the presence of these genes increases a cluster's level of robustness in the generation of pathogen recognition specificities through a significantly high rate of coding mutations.

A number of authors describe pseudogenes as genes whose ORFs are interrupted by either termination codons or frame shifts, and that this set of genes provide valuable reservoirs of genetic variability in a gene family (Michelmore and Meyers, 1998; Bergelson *et al.*, 2001; Balakirev and Ayala, 2003; Liu and Ekramoddoullah, 2003). According to Michelmore and Meyers (1998) this phenomenon is well elaborated in the chicken immunoglobulin V_H genes that are made up of one functional gene and 80 pseudogenes, the fast and vast potential of genetic mutations in this case allows for

generation of a number of specificities that can be kept silent up to and until the need arises.

Clusters in this dataset ranged from 0:11 (pseudogenes : uninterrupted ORFs) in cluster 4 to 51:2 in cluster 11 (Table 10). Other clusters have ratios in between these values as shown in Table 10. In the whole Golden Delicious dataset there are 126 pseudogenes to 139 uninterrupted ORFs (47.5% pseudogenes). If this was to be ascribed to PCR and sequencing errors then it would mean 1 bad sequence reaction in every 2.105 performed. The error rate for PCR using *Taq* DNA Polymerase without 3'-5' exonuclease activity is 1.5×10^{-5} for every nucleotide polymerized and 0.7×10^{-5} for substitution error and frame shift (Kwiatowski *et al.*, 1991). Thus taking potential PCR error rate into consideration, the NBS-LRR family of genes still possesses a high ratio of pseudo- to functional genes.

It is estimated that between 2 - 3% and 0.5 – 1% of the human and mouse genomes respectively are made up of pseudogenes (Yano *et al.*, 2004) and literature confirms a functional role for pseudogenes in a range of gene families including plant R genes. Liu and Ekramoddoullah (2003) found that almost half of the RGAs sequenced in western white pine (*Pinus monticola*) were made up of pseudogenes. Analysis of cDNA sequences obtained following infection confirmed transcription of a large number of genes with premature termination codons. There is no evidence of translation of pseudogenes into protein to date, however, Yano *et al*. (2004) show that pseudogenes constitute a posttranscriptional mRNA regulation system. In these studies, silencing of

the non-coding RNAs (nc-RNAs from pseudogenes) resulted in destabilization of the transcription of the human *Makorin1* gene.

3.4.1 Selection and gene conversion events in R genes

Sequences were edited using the Sequence Analysis software (section 2.13.2.1) to determine the correct ORFs. *In silico* translation of the un-truncated nucleotide sequences was performed given although degenerate oligonucleotides might bias nucleotide sequences they still conserve the amino acid residues at priming sites. Only sequences that gave complete ORFs that were not interrupted by termination codons were considered for calculation of Ka/Ks values. This is because the NBS domain is a small portion in the middle of a protein coding sequence.

The calculation was performed using the HyPhy program (section 3.3.5) and values of 2.09 and 2.50 were obtained for clusters 3 and 4 respectively while cluster 5 with only 1 uninterrupted ORF could not be included in this analysis. Calculation of Ka/Ks values infers the selection history along an evolutionary path and thus uses a phylogenetic tree, which requires a minimum of 4 sequences per calculation. These 3 clusters constitute the non-TIR subfamily. According to results of this analysis, the non-TIR subfamily of *Malus x domestica* is under positive or diversifying selection (Noel *et al.*, 1999; Liu and Ekramoddoullah, 2003).

The TIR sub-family is subdivided into clusters 1, 2, 6, 8, 10, 10b, 12, 13 and 14 (Figure 3.5). Clusters 2, 12 and 14 had Ka/Ks values of the order 2.382, 1.920 and 1.323

respectively meaning they are under diversifying selection. The value of 1.323 for cluster 14 would suggest near-neutral selection (Ka/Ks = 1.0), however, positive selection means an excess of coding over silent mutations. This result suggests that coding mutations for cluster 14 are 30% higher than silent mutations, although the level of positive selection here would be far lower than in the other two clusters it is still significantly higher than clusters 1 and 10b with values of 1.034 and 1.065 respectively which are undergoing neutral selection. Clusters 6, 8, 10a and 13 with values of 0.817, 0.739, 0.651 and 0.805 respectively are classic examples of clusters under purifying selection.

Clusters 7 and 9 have members that are very divergent and do not form a properly defined cluster (Figure 3.5). Further homology searches using BLAST against the non-redundant database confirmed that these were indeed RGAs that were too divergent to fit in the other clusters. This phenomenon was also observed in other RGA phylogenetic studies (Mondragon-Palamino *et al*., 2002). However, these still possess a proportion of functional to pseudogenes and whether they represent single genes scattered along random positions on the genome or form part of a highly divergent tandem array remains unclear at this point. Cluster 11 with 53 RGAs has only 2 that are uninterrupted ORFs, determination of selection pressure also could not be performed in this case.

RGAs in the non-TIR subfamily are all under positive selection, whether this suggests a common locus still needs to be investigated. The TIR subfamily constitutes the larger of the two subfamilies with 11 clusters with individual clusters under a unique different

selection pressure. However, proper confirmation of genome distribution would need to be tested by mapping and linkage analysis.

In the evolution of gene families, short segment gene conversion events form a very important force and these events are said to be more frequent than point mutations (S. Sawyer, 1989; S. A. Sawyer, 1999, 2000). GENECONV (section 2.13.2.1) was used to detect evidence of past gene conversion events in cluster based alignments. Gene conversion was detected in all the clusters except 2, 3, 11 and 14. Calculations were not done for clusters 7 and 9, which do not form conventional clusters (Figure 3.5). In clusters 1, 4 and 10b gene conversion events were only detected between functional genes. Clusters 5, 8 and 13 on the other hand show evidence of gene conversion involving only pseudogenes. Gene conversion in clusters 6, 10a and 12 were detected between functional and pseudogenes. The ratio of functional to pseudogenes was not shown to have any significant influence on the prevalence of fragment swapping between two genes. Figure 3.10 demonstrates the significance of the gene conversion between RGAs GD32 and GD171. In this case the putative shared fragment forms a perfect alignment between the two genes although flanking regions show significant variation between the participating pair.

It can be assumed therefore based on results of Table 12 and Figure 3.10 that the occurrence of gene duplication is a random process that is not based on the ratio between functional to pseudogenes. However, GENECONV cannot distinguish between unequal

crossover and gene conversion and as such these values may not be regarded as absolute proof of gene conversion alone acting on the clusters analysed.

3.4.2 Investigating gene duplication in apple RGAs

According to plant evolutionary history (Chaw *et al.*, 2004) gymnosperms exemplified by Pinus diverged earlier from the angiosperms before the monocot – dicot split. However, the presence of TIR and non-TIR subfamilies in monocots (Budak *et al.*, 2006) and in *Pinus* (Liu and Ekramoddoullah, 2003) shows the co-existence of both TIR and non-TIR as pro-orthologs in the plant genomes. However, the relative representation in dicots such as *Malus* shows almost double the number of TIR compared to non-TIR, which is the reverse of what is observed in monocots such as the *Agrostis* and *Oryza* species.

Ka/Ks values for TIR and non-TIR subfamilies show that both are experiencing diversifying though more pronounced in non-TIR. Results of Figure 3.11 show co-existence of TIR and non-TIR in the earlier branches of a tree rooted using *Pinus monticolor* RGAs. The purifying selection in some clusters in the TIR subfamily could be as a consequence of the continual challenge exerted by endemic pathogens. The relative representation could be related to the type of pathogens endemic to the monocots and dicots. More work still needs to be done to address the issue of relative representation of TIR and non-TIR between the monocots and dicots, although such work is out of the scope of this work.

CHAPTER 4

SATURATION SEQUENCING AND COMPARATIVE ANALYSIS OF RGA ORTHOLOGS

CONTENTS

CHAPTER 4: SATURATION SEQUENCING AND COMPARATIVE ANALYSIS OF RGA ORTHOLOGS

4.1 INTRODUCTION

The *Arabidopsis* genome has been shown to encode approximately 200 defence related genes and of these 149 belong to the NBS-LRR family (Meyers *et al.*, 2003; Trognitz and Trognitz, 2005). Extrapolation from such studies have led to the assumption that plants with large genomes are likely to have a copy number that ranges from 400 - 1000. About 535 *Malus* NBS-LRR genes have been deposited in GeneBank and related studies have shown that the copy number per cluster varies even among varieties of the same species (Leister *et al.*, 1998; Trognitz and Trognitz, 2005).

Comparative analysis of RGAs has been performed in *Malus* and the results have shown among other things that although these genes show a high degree of sequence homology among different species, gene copy numbers per cluster differs even among varieties of the same species (Leister *et al.*, 1998). The level of sequence homology among orthologous genes was shown to be higher than that observed in paralogs, which means phylogenetic analyses using sequences from different species still manages to recover the respective clustering order per species in a comparative study.

PCR amplification of candidate RGAs in different plants using degenerate oligonucleotides targeted at the P-loop and GLPL motifs, cloning and sequencing the

inserts has become the conventional strategy for targeted sequencing of NBS domains of the NBS-LRR gene family (Lee *et al.*, 2003; Zhou *et al.*, 2004; Xu *et al.*, 2005; Palomino *et al.*, 2006; Bozkurt *et al.*, 2007). As described in Lee *et al.* (2003), the use of degenerate oligonucleotides as described above constitutes a non-exhaustive approach to sequencing RGAs. Bias in PCR amplification against genes that either are not easy to amplify or are present in low copy numbers distorts the percentage representation of the sequences obtained.

The aim of this chapter is to analyse sequence representation under two different sequencing technologies. The underlying assumption is that the priming regions although highly conserved in the NBS-LRR gene family still show sequence variations (Figure 3.9) that might limit hybridisation efficiency with the degenerate oligonucleotides. However, the resulting level of bias is expected to suppress relative copies of affected genes and not totally excluding them from the amplification product. If this is the case then high throughput direct sequencing of PCR amplification products should recover sequences of all genes flanked by the P-loop and GLPL motifs used in designing the degenerate oligonucleotides.

4.2 The RGA database

A comprehensive database was constructed using RGAs sequenced in this project and others accessed from GenBank. This database also contained sequences from *Malus floribunda, Malus prunifolia, Malus buccata* and *Pinus monticolor*; the *Pinus* sequences

were used for the purposes of rooting the species analysis tree. This allowed for a broader analysis of the NBS domain of R genes from apple.

4.2.1 Sequencing of RGAs

The 661 sequences generated as described in section 3.2 and constituted as follows; Golden Delicious (265), Anna (260) and a bulked sample of Golden Delicious and Anna mixed leaf samples (136) were used here. The sequencing procedure was described in chapter 2 and 3. Analysis of the signal to noise ratio in sequence chromatograms and pairwise comparisons were done to eliminate sequences that were either too identical or were very homologous to sequences generated earlier. Using this procedure the sequence dataset was reduced from the original 661 to 515 unique sequences for subsequent analyses.

4.2.2 Incorporated RGA sequences from the public database

A total of 443 sequences were accessed from GenBank for *Malus floribunda* (141), *Malus baccata* (5), *Malus prunifolia* (17) and *Malus x domestica* (280). Literature on these sequences suggested that they were produced using degenerate primers targeted also on the P-loop and the GLPL motifs (Lee *et al*., 2003; Baldi *et al*., 2004; Rikkerink *et al*., 2006 (direct submission to GenBank)). These were screened initially using length after removing priming sites as the critical parameter; only sequences at least 490 nucleotides were used. The second screening parameter was percentage divergence that was determined using the Clustax version 1.83 delay divergence setting.

4.3 Phylogenetic analysis of the RGA orthologs

4.3.1 Sequence assembly and removal of duplicates

The complete data set from Anna and Golden Delicious, including those from the bulked sample sequencing were assembled using the Staden Package version 1.6.0 (Bonfield and Staden, 1996; Staden *et al.*, 2000). This assembly was performed using sequence data in fasta format and as such pre-processing in Pregap4 was only set up to include general configuration, initialise experiments and setup Gap4 shotgun assembly modules. Maximum pads and percentage mismatch were set at 5 and 2.0 respectively, this allowed for a stringent control of sequences permitted into each contig. However, parameters were relaxed enough to only group duplicates in a contig and thus achieve a significant reduction of repetition in subsequent analyses.

In the Gap4 assembly, variations between constituent readings in a contig were highlighted using background colour. This process facilitated for a clearer distinction between sequences incorrectly assumed to be duplicates of each other. Contigs with too many random mismatches were split manually with the number of splits dependent on the most likely sequence groupings within the contig being analysed. This process generated 205 contigs; consensus sequences from these were used in downstream analyses. Most important to note here is that the assembly process was not intended to produce contigs that represent 'genes' where constituent sequences become alleles of that gene. The reason for this exercise was to reduce the number of sequences and thereby allow phylogenetic trees from this data to be informative without losing relative resolution.

4.3.2 Generation and refinement of multiple sequence alignments

Two separate sequence alignments were performed using ClustalX. The first alignment was made up of data from *Malus x domestica* (GenBank accessed data and consensus sequences from contigs assembled in section 4.3.1) and the second alignment had representative sequences from *Malus x domestica* (consensus sequences from assembled contigs), *Malus buccata*, *Malus prunifolia, Malus floribunda* and *Pinus monticolor*.

The ClustalX alignments were manually edited in JalView (section 2.13.2.1). This process included the removal of terminal gaps, all gap columns, priming sites and incorrectly aligned sites. Priming sites were removed from all sequences both from the Anna/ Golden Delicious and the GenBank datasets. This was performed to eliminate the possibility of biased phylogenetic analyses. According to site-specific conservation scores determined in section 4.5 the P-loop and GLPL motifs are highly conserved thus their exclusion from the analyses does not adversely affect inference of phylogeny.

4.3.3 Inference of phylogeny for the comprehensive RGA dataset

Two phylogenetic trees were constructed from this data set, the first set was made up entirely of sequences from *Malus x domestica* and the second tree had representative sequences from *Malus x domestica*, *M. buccata*, *M. prunifolia, M. floribunda* and *Pinus monticolor* making up the outgroup. Both trees were inferred in PAUP* version 4.0 with the reliability of the branching order confirmed through the sub-tree pruning and regrafting, SPR, algorithm.

In the *Malus x domestica* phylogeny reconstruction, a parsimony tree was used to estimate model parameters given the data. The transition/ transversion rate ratio, base frequencies and among-site rate heterogeneity were estimated using the HKY85 model. These parameters were then used to construct a tree under the maximum likelihood criterion. The series of steps in this protocol were formulated using the suggestions in the PAUP version 3.1 manual (section 2.13.2.1). A hundred trees were generated and from this a final consensus tree was obtained using the 50% majority rule. The graphical manipulation was performed using the Mesquite software (section 2.13.2.1). This tree is shown in Figure 4.1. Clusters in this phylogenetic tree show a high degree of sequence homology not only among genes from different cultivars within *Malus x domestica* (Borkh.) but also among different species of the *Malus* genus. This confirms sequence conservation of in NBS-LRR genes after divergence.

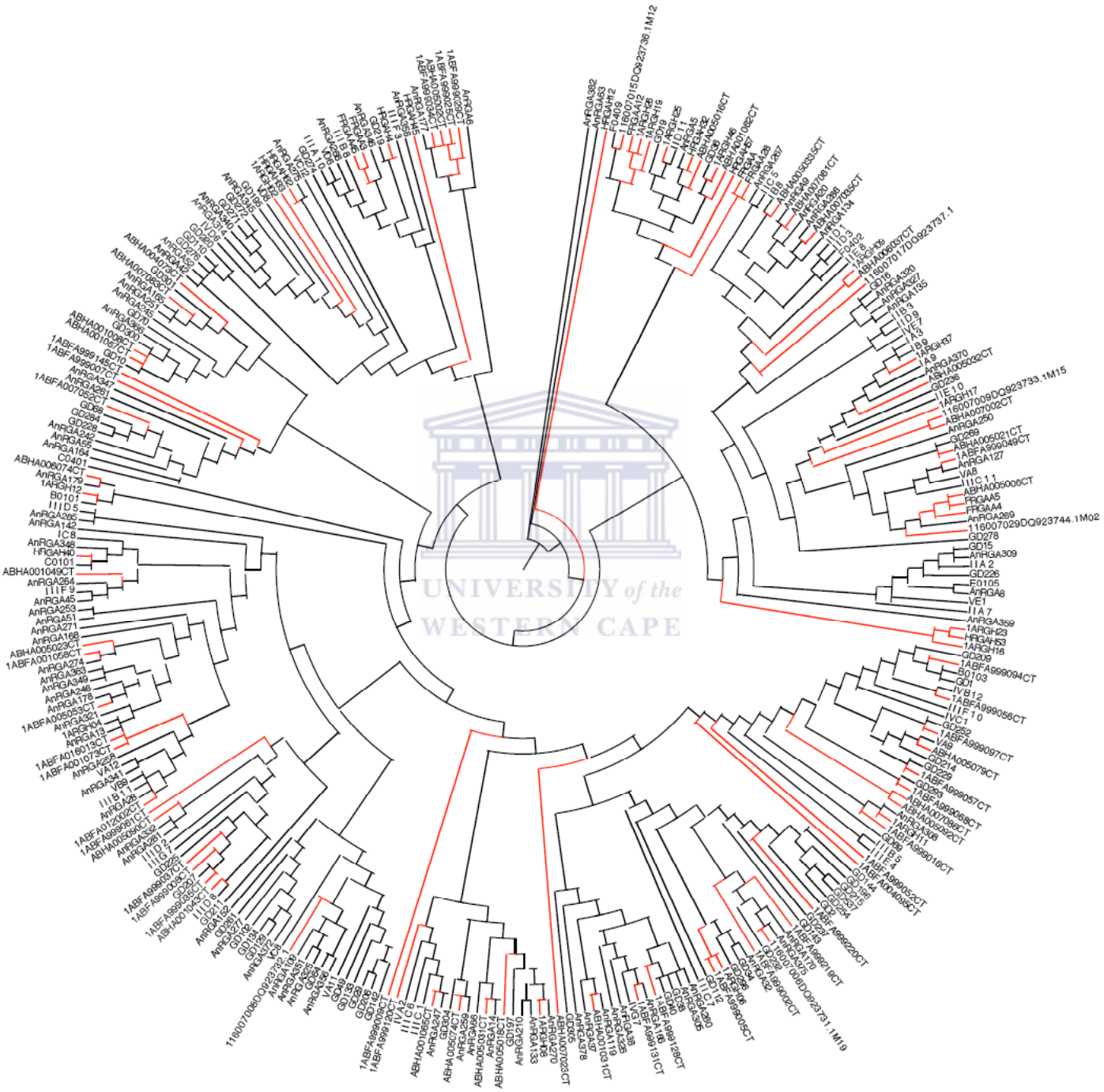
The comparative analysis of orthologs was also performed in PAUP* version 4, though with a slightly different protocol. In this analysis, an initial sample of 1000 trees was estimated using parsimony with the number of reps set to 100 and all parsimony uninformative sites excluded. In the second step trees were randomised before being added at random to search for the optimum tree. This process generated a total of 8929 trees that were then were filtered to obtain the best trees while permanently deleting those that were not successful in the filtering process.

This reduced the number from 8929 to 403 trees that were then rooted using the *Pinus monticolor* outgroup. Finally a consensus tree was produced using the 50% majority rule.

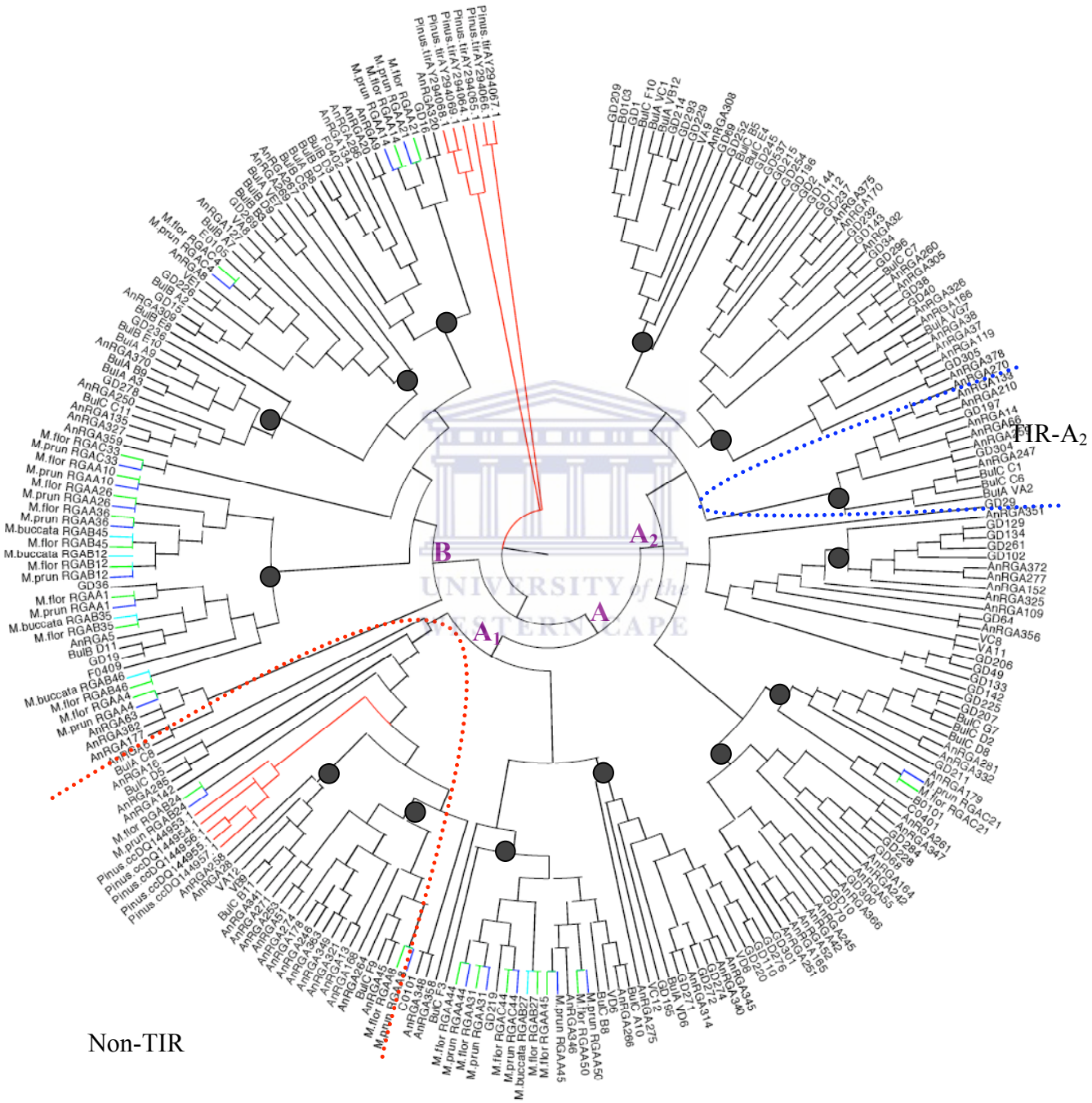
The consensus tree was manipulated and displayed using Mesquite (section 2.13.2.1). All other species besides *Malus x domestica* were colour coded to allow for an easy analysis of the inferred relationships. This tree is shown below in Figure 4.2. This phylogenetic tree shows a total of 14 clusters, 9 of which have representative genes from other *Malus* species. The non-TIR genes of *Pinus monticolor* were grouped together in one cluster with cultivars Anna and Golden Delicious of *Malus x domestica* (Borkh.). Some of the clusters were composed entirely of *Malus x domestica* (Borkh.) cultivars Anna and Golden Delicious. There are also clusters composed entirely of *M. prunifolia* and *M. floribunda* and a few that are biased towards either Golden Delicious or Anna genes. Whether this represents functional genes that are active in the particular plant or that the sample size was not large enough to be representative cannot be deduced from this result.

Overall, the phylogenetic tree in Figure 4.2 shows that gene constitution of NBS-LRR per cluster varies depending on the species being analysed. There are however, clusters that share homologous sequences through conservation of orthologs although the average number per cluster varies. Most of the clusters in both figures 4.1 and 4.2 contain sub-clusters that are specific to either different cultivars or species. Analysis of genes sequenced from the same species (Figure 4.1) show that sequence homology per cluster is maintained though the average numbers vary from cultivar to cultivar. The reasons for variations in sequence number per cluster among cultivars of the same species could be simplified by analysing gene function following pathogen infections. This could reveal the clusters containing genes that are active in those cultivars and hence the need for expanding gene copy numbers.

Figure 4.1. Comparative analysis of RGAs within *Malus x domestica* (Borkh.) species. Samples with the prefix ABFA and ABHA are from cultivars A172-2 and Pinkie respectively (Rikkerink *et al*., 2006 – direct submission); ARGH samples are from Florina (Baldi *et al*., 2004); FRGA and HRGA are from Fuji and Hong-ok respectively (Lee *et al*., 2003) and samples with the GI numbers starting with 1160070 are from Florina (Sordo and Baldi, 2006 – direct submission). The red coloured branches represent all sequences accessed from GenBank.


AnRGA382
AnRGA453
1RGAH2
1160070015DQ923736.1M12
FRGAA12
1ARGH16
GD19
1ARGH25
IID.1
IVC3
1ARGH45
ABHA005016CT
GD35
1ARGH46
ABHA001062CT
FRGAA7
FRGAA4
FRGAA28
AnRGA267
IID.5
IIB.6
ABHA005035CT
AnRGA49
ABHA007061CT
AnRGA20
AnRGA266
ABHA007065CT
AnRGA134
IID.3
IID.2
IIE5H05
ABHA006937CT
1160070117DQ923737.1
GD16
AnRGA390
AnRGA327
AnRGA135
IIB.9
IVC7
IA9
1ARGH37
IA9
AnRGA370
ABHA005032CT
GD236
IIE10
1ARGH17
1160070090DQ923733.1M15
ABHA007002CT
AnRGA250
GD269
ABHA005021CT
1ABFA999049CT
AnRGA127
VA8
IIIC11
ABHA005006CT
FRGAA5
FRGAA4
AnRGA269
1160070290DQ923744.1M02
GD278
GD15
AnRGA309
IIA2
GD226
E0105
AnRGA8
VE1
IIA7
AnRGA359
1ARGH23
HRGAH53
1ARGH16
GD209
1ABFA999094CT
B0103
GD1
IVB12
1ABFA999058CT
IIIF10
IVC7
GD252
1ABFA999097CT
VA9
ABHA005079CT
GD214
GD229
GD223
1ABFA999057CT
ABHA007068CT
ABHA005032CT
AnRGA308
1ABFA999016CT
IIIB5
GD254
1ABFA999052CT
A00409SCT
GD229
GD234
1ABFA999220CT
1ABFA999218CT
AnRGA375
GD229
1160070060DQ923731.1M19
GD229
AnRGA432
1ABFA999092CT
GD12
1ABFA999005CT
GD206
AnRGA280
AnRGA305
GD12
1ABFA999129CT
GD206
GD208
1ABFA999191CT
1ABFA999166
VB7
AnRGA38
AnRGA325
ABHA007031CT
AnRGA119
AnRGA37
ABHA007028CT
AnRGA378
GD205
ABHA007022CT
AnRGA270
1ARGH08
AnRGA133
AnRGA210
GD197
ABHA005018CT
AnRGA14
AnRGA259
ABHA007047CT
GD304
GD207
1ABFA999120CT
1ABFA999005CT
IIIC6.1
IVA2
ABHA001065CT
AnRGA343
GD204
ABHA005074CT
AnRGA466
AnRGA259
ABHA007047CT
AnRGA14
ABHA005018CT
GD197
AnRGA210
1160070060DQ923732.1
AnRGA100
AnRGA351
AnRGA255
GD206
VA17
GD139
GD208
GD202
AnRGA378
GD139
GD191
GD271
AnRGA152
GD261
GD211
IIID.8
1ABFA999095CT
1ABFA999065CT
GD221
ABHA001043CT
GD205
GD206
1ABFA999005CT
IIIG.7
GD205
1ABFA999095CT
IIID.2
AnRGA261
ABHA005090CT
1ABFA012004CT
AnRGA262
IIIB.11
VB9
AnRGA241
VA12
1ABFA001073CT
1ABFA016013CT
AnRGA259
1ARGH13
AnRGA321
1ABFA005053CT
AnRGA178
AnRGA246
AnRGA349
AnRGA363
AnRGA274
1ABFA001058CT
ABHA005023CT
AnRGA168
AnRGA271
AnRGA51
AnRGA253
AnRGA45
IIIF9
AnRGA264
C0101
ABHA001049CT
HRGAH40
IC8
AnRGA348
AnRGA142
AnRGA265
B0101
1ARGH12
1ARGA179
ABHA006074CT
C0401
AnRGA164
AnRGA255
AnRGA242
GD228
GD284
GD89
1ABFA007052CT
AnRGA281
AnRGA347
1ABFA999007CT
1ABFA999145CT
ABHA001067CT
GD306
GD70
AnRGA165
AnRGA251
AnRGA358
GD307
ABHA007036CT
ABHA001098CT
AnRGA369
GD308
GD202
GD206
GD208
GD192
GD202
GD206
GD208
GD192
GD202
GD206
GD208
GD192

Figures 4.2. Comparative analysis of orthologs in the *Malus* genus. Leaves and branches in this phylogenetic tree are colour-coded with members of each species having the same colour. *Malus x domestica* cultivars are not colour-coded, M. prun (*M. prunifolia*) in blue, M. flor (*M. floribunda*) in green and *M. buccata* in sky blue. The branches in red are for *Pinus monticolor* and were used here to root the tree. The clusters demarcated by the two dotted red boundaries constitute the non-TIR subfamily. Black dots at internal nodes are used here to indicate gene clusters. TIR-A_2 represents a set of NBS-LRR genes that could not be classified either as TIR or non-TIR.


TIR-A2
Non-TIR
A2
B
A
A1

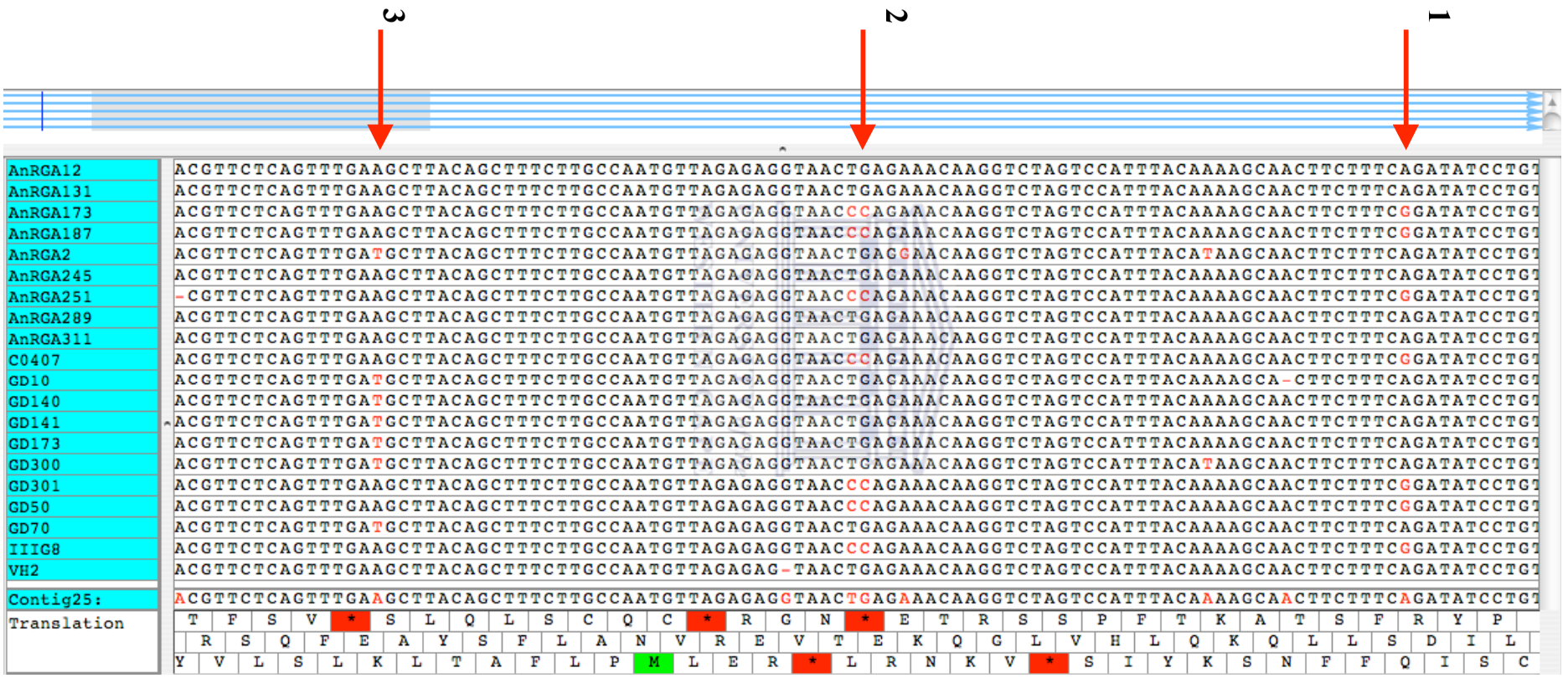
4.4 SNP discovery and re-sequencing of sequence clusters

Chromatograms of all sequences from Anna and Golden Delicious were assembled using CodonCode Aligner (section 2.13.2.1). Assembly parameters were set up to group sequences most likely to occur as alleles of the same gene. Given the errors associated with DNA fragment amplification using *Taq* polymerase, random mismatches were expected in assembled contigs. This problem was taken into account and assembly conditions were relaxed to accommodate an average of one error per 100 bases while simultaneously recovering genes as contigs.

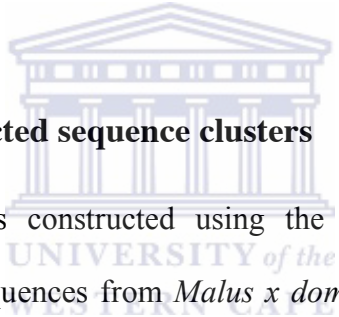
Prior to removing priming sites, the minimum percentage identity and maximum unaligned end gap settings were set to 90 and 70% respectively to accommodate sequencing artefacts and PCR related random mismatches. Sequences were then unassembled and re-assembled again with minimum percentage identity and maximum unaligned end gap settings adjusted to 95 and 75% respectively. This process gave a set of contigs that contained sequence reads most likely to be multiple copies of the same gene.

No reference sequences were identified as consensus chromatograms for this analysis given that the NBS-LRR gene family is made up of a large number of sequences that group into different clusters (section 3.3.4.3 and Figure 3.5). A *de novo* assembly was performed in which all sequences were compared against all and candidate SNP positions were highlighted by mutation detection options in the software. Examples of candidate SNP positions detected are shown in Figure 4.3.

Figure 4.3. **Sequence assembly of *Malus x domestica* (Borkh.) cultivars Anna and Golden Delicious RGAs showing candidate SNPs**. Thick red arrows in the figure indicate candidate SNPs. Sequence identities are abbreviations of the source and number of the clone; where AnRGA and GD denote Anna and Golden Delicious respectively and roman numerals are used for sequences from the bulked dataset.


3
2
1
AnRGA12
AnRGA131
AnRGA173
AnRGA187
AnRGA2
AnRGA245
AnRGA251
AnRGA289
AnRGA311
C0407
GD10
GD140
GD141
GD173
GD300
GD301
GD50
GD70
IIIG8
VH2
Contig25:
Translation

Candidate SNPs 1 and 2 (G/A and T/C) indicate that there is one gene with two alleles in this assembly. These are polymorphic in cultivar Anna with an approximately 30% distribution. Position 3 seems to complement 1 and 2. Alternatively the occurrence of other random mismatches (bases in red in Figure 4.3) could suggest that these are merely a mixture of two homologous genes with a possible two alleles each. To confirm whether RGAs do indeed contain SNPs and not that contigs such as Figure 4.3 are a mixture of highly homologous genes assembled together due to suboptimal assembly parameters, re-sequencing of genomic DNA was performed.



4.4.1 Re-sequencing of selected sequence clusters

A neighbour-joining tree was constructed using the HKY85 model of nucleotide substitution for all the 661 sequences from *Malus x domestica* (Borkh.) cultivars Anna and Golden Delicious (tree not shown). Sequences from large clusters made up exclusively of RGAs from either cultivar Anna or Golden Delicious were selected and aligned separately using ClustalX. Three sets of unique primers were designed from each of these alignments using regions of high conservation just downstream to the initial degenerate priming sites. The ‘majority rule’ policy was applied at alignment columns with more than one possible nucleotide. Table 13 shows sequences of the oligonucleotides.

Table 13. Oligonucleotides designed to re-sequence *Malus x domestica* cv. Anna and Golden Delicious clusters for SNP analysis.

Oligonucleotide ID	Oligonucleotide Sequence	Product size (Bases)
GD1F	GACGACCATTGCTAAGCTA	462
GD1R	TGGAATCGTCCGAGTGTT	
GD2F	CTGAAAGGATACGTGCTCA	412
GD2R	TAGGAGTGACCAGGTTGG	
Anna3F	GAAGACGACACTTGCTCA	361
Anna3R	TGCAAAGATTATGGGAGCC	

4.4.2 PCR amplification and cloning of amplicons

Genomic DNA was extracted from cultivars Anna and Golden Delicious as described in section 2.6.1. The integrity of the extracted DNA and its relative concentration was assessed by agarose gel electrophoresis (section 2.12). PCR amplifications were performed using oligonucleotides GD1, GD2 and Anna3. The standard PCR protocol in section 2.8.1 was altered with the reduction of dNTP final concentration to 25 µM and an increase in the elongation time to 2 minutes at 72°C. Using these conditions 35 amplification cycles were performed with an annealing temperature of 56°C (determined empirically). Results of the PCR amplification are shown in Figure 4.4. The

oligonucleotides were designed to flank DNA fragments of the order 462, 412 and 361 base pairs for GD1, GD2 and Anna3 respectively. Figure 4.4 shows PCR amplification products with the expected sizes as confirmed by the DNA size standard in lane M.

The PCR products were ligated into pGEM-T Easy Vector and transformed into XL1-Blue cells (sections 2.9 and 2.10 respectively). The transformation reaction was plated on Ampicillin/ X-gal indicator plates and incubated at 37°C for 16 hours. Recombinants were screened using blue/white selection. The positive white colonies were screened again using colony PCR (section 2.8.3). The results of colony PCR amplification are shown in Figure 4.5, colonies carrying the desired inserts have the sizes ± 660, ± 610 and ± 561 for oligonucleotides GD1, GD2 and Anna3 respectively (reasons for increased sizes are explained in section 3.2.4). These results show that the fragments containing the candidate SNPs (section 4.4) were successfully amplified and cloned into XL1-Blue cells (section 2.10).

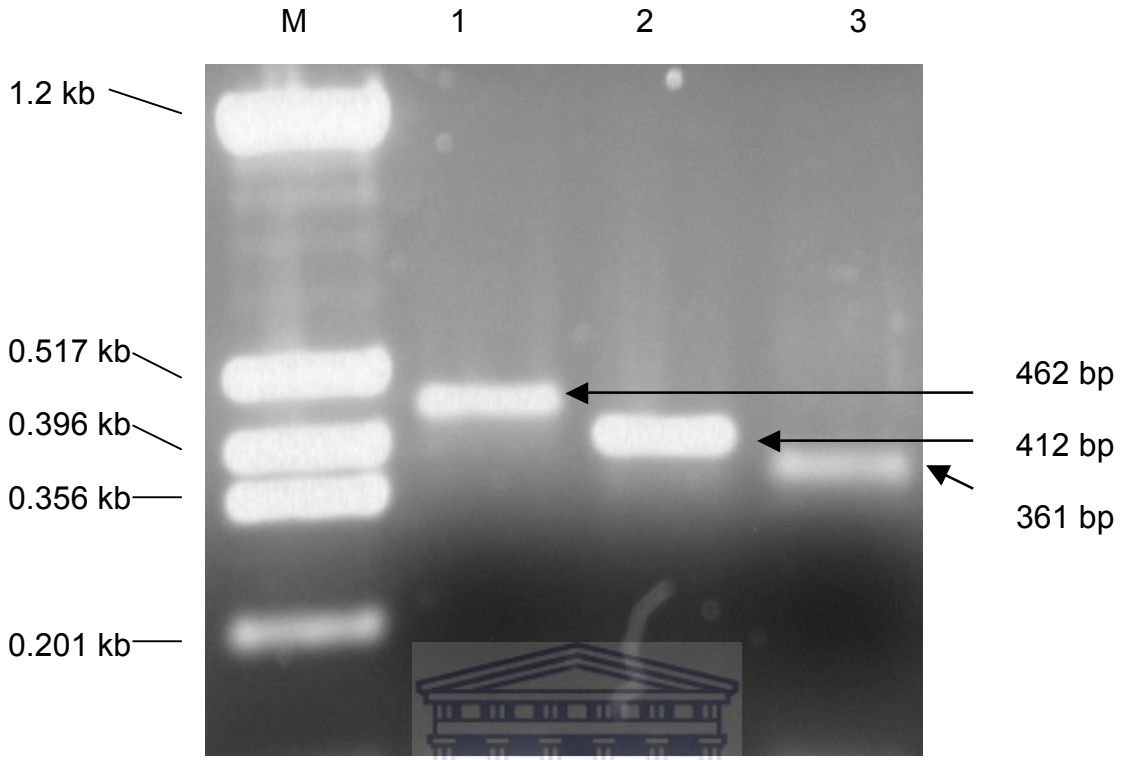
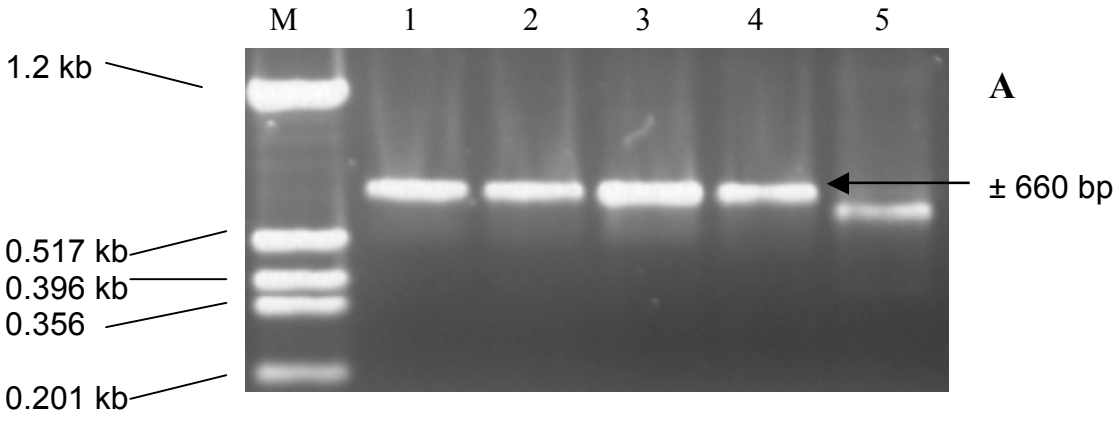
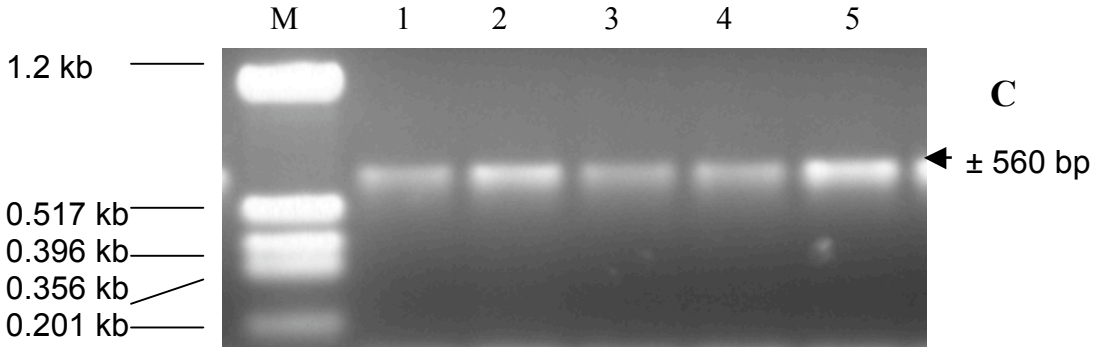


Figure 4.4. PCR amplification of DNA fragments from *Malus x domestica* cv. Anna genomic DNA using oligonucleotides GD1, GD2 and Anna3. Lane M shows the profile of pTZ/*Hinf* I DNA size standard, lane 1, 2 and 3 shows amplification products generated using oligonucleotides GD1, GD2 and Anna3 respectively.

Figure 4.5. Colony PCR screening of recombinants for inserts with DNA fragments generated from oligonucleotides GD1, GD2 and Anna3. Agarose gel images A, B and C show results for oligonucleotides GD1, GD2 and Anna3 respectively on *Malus x domestica* (Borkh.) cv. Anna genomic DNA. Lane M in each of the 3 gels represents the pTZ/*Hinf* I DNA size standard profile.





4.4.3 Sequencing of the recombinant colonies

Oligonucleotides GD1, GD2 and Anna3 were used to amplify target RGAs in both *Malus x domestica* (Borkh.) cvs. Anna and Golden Delicious genomic DNA. Recombinant clones carrying the desired inserts as confirmed with colony PCR were sequenced at Inqaba Biotechnical Industries (Pty) Ltd using M13 universal oligonucleotides (section 2.13). A total of 61 sequences were generated for the three oligonucleotide pairs in both cultivars. The sequences were inspected for confirmed to be DNA fragments of candidate RGAs through homology searches in GenBank using Blast (section 2.13.2). Inspection of the sequence chromatograms showed a high signal to noise ratio as shown in Figure 4.6. These were taken as good quality sequences and thus were used for SNP detection in section 4.5.1. The number of sequences generated per cultivar per oligonucleotide pair is shown in Table 14.

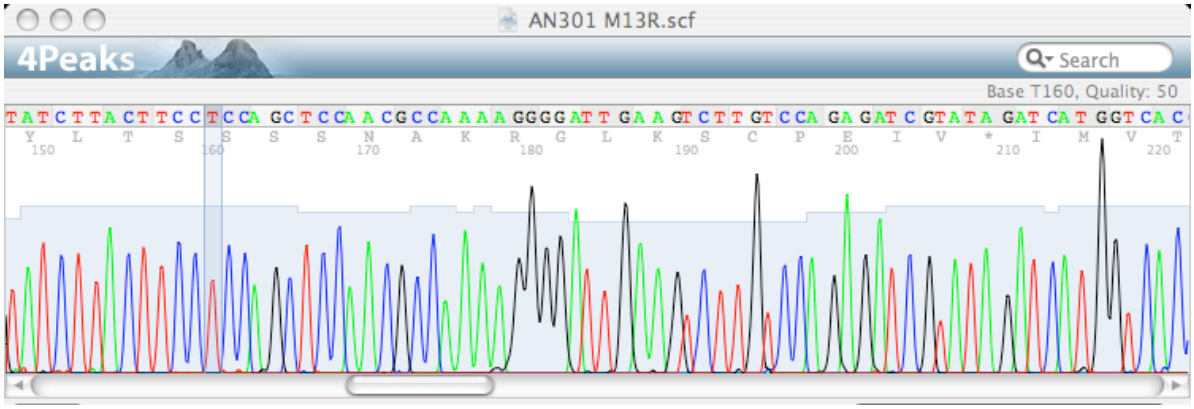


Figure 4.6. An example chromatogram of the sequences used for SNP detection.

Table 14. The number of sequences generated per cultivar for each oligonucleotide pair used in the re-sequencing of RGAs.

Cultivar	Primer IDs	Number of sequences
Anna	GD1	11
	GD2	11
	Anna3	15
Golden Delicious	GD1	23
	GD2	23
	Anna3	18

4.4.4 Sequencing of PCR amplicons using the GS20 Technology

4.4.4.1 Design of fusion oligonucleotides with the GS20 A and B tags

Degenerate oligonucleotides NBSF and NBSR-2 were re-designed to incorporate GS20 sequencing tags A and B on the 5' end of each oligonucleoide. The GS20 sequencing tags are 19 nucleotides in length and both terminate in a 'TCAG' key that is detected by the sequencing software and serves to initiate the sequencing reaction. These oligonucleotides were renamed by adding the identity of the respective tag at the end of the original name and are shown in Table 15.

Table 15. Sequences of the re-designed degenerate oligonucleotides containing 5' GS20 sequencing tags. The nucleotides fragment in red font constitutes the tag in each case and the underlined bases are the sequencing key.

Primer ID	Primer Sequence
NBSFA	GCCTCCCTCGCGCCATCAGGGWATGGGWGGWRTHGGWAARACHAC
NBSR-2B	GCCTTGCCAGCCCGCTCAGARIIIITTVARIGCIARIGGIARICC

4.4.4.2 Tail-PCR amplification of RGAs

Fusion degenerate oligonucleotides NBSFA and NBSR-1B were used for PCR amplification of candidate RGAs from *Malus x domestica* (Borkh.) cvs. Anna and Golden Delicious genomic DNA (section 2.12.1). The annealing temperature for the NBSF and R-1 oligonucleotide pair was used here to amplify the same target region and introduce the GS20 sequencing tags, with A on the 5' end and B on the 3' end of each PCR amplification product. These amplification products were resolved using agarose gel electrophoresis (section 2.12) and purified from the gel (section 2.6.1.2). The purified products are shown in Figure 4.7. The size of the PCR amplification product was

expected to be ± 550 bases due to the addition of 19 nucleotides on either end and the expected DNA fragments are indicated in Figure 4.7. Extraction of the fragments from agarose gels purifies them from contaminating template genomic DNA, residual oligonucleotides and dNTPs and gives a relatively pure product as shown in Figure 4.7. Recovery from agarose gel purification varied, 5 µl of the purified DNA was loaded per well.

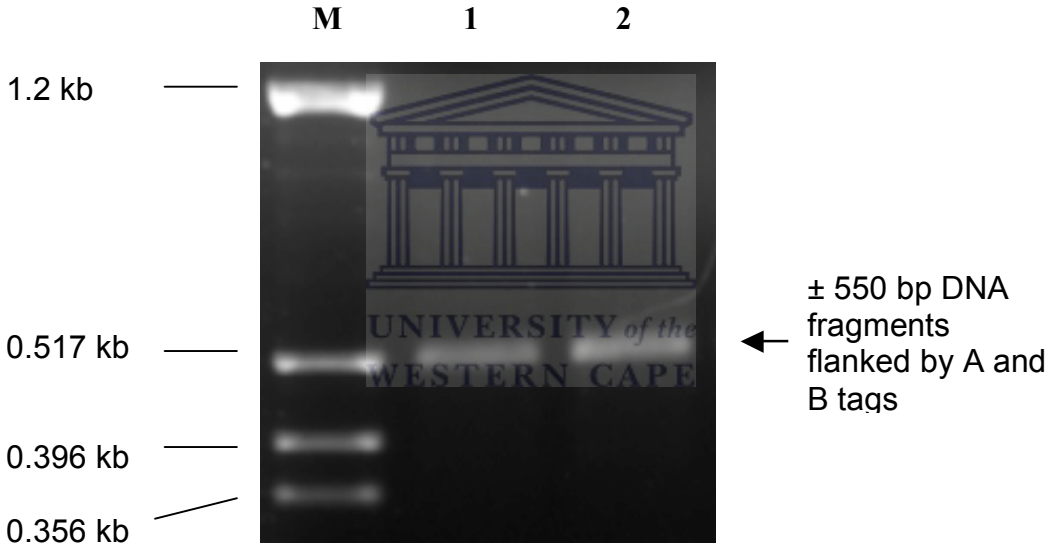


Figure 4.7. Agarose gel - purified products of the tail-PCR amplification using fusion oligonucleotides with 5' GS20 A and B sequencing tags. Lanes 1 and 2 contain purified DNA fragments from *Malus x domestica* (Borkh.) cvs. Anna and Golden Delicious respectively. Lanes M shows the pTZ/*Hinf* I DNA size standard profile.

4.4.4.4 Sequencing of fragments of RGAs using the GS20 System

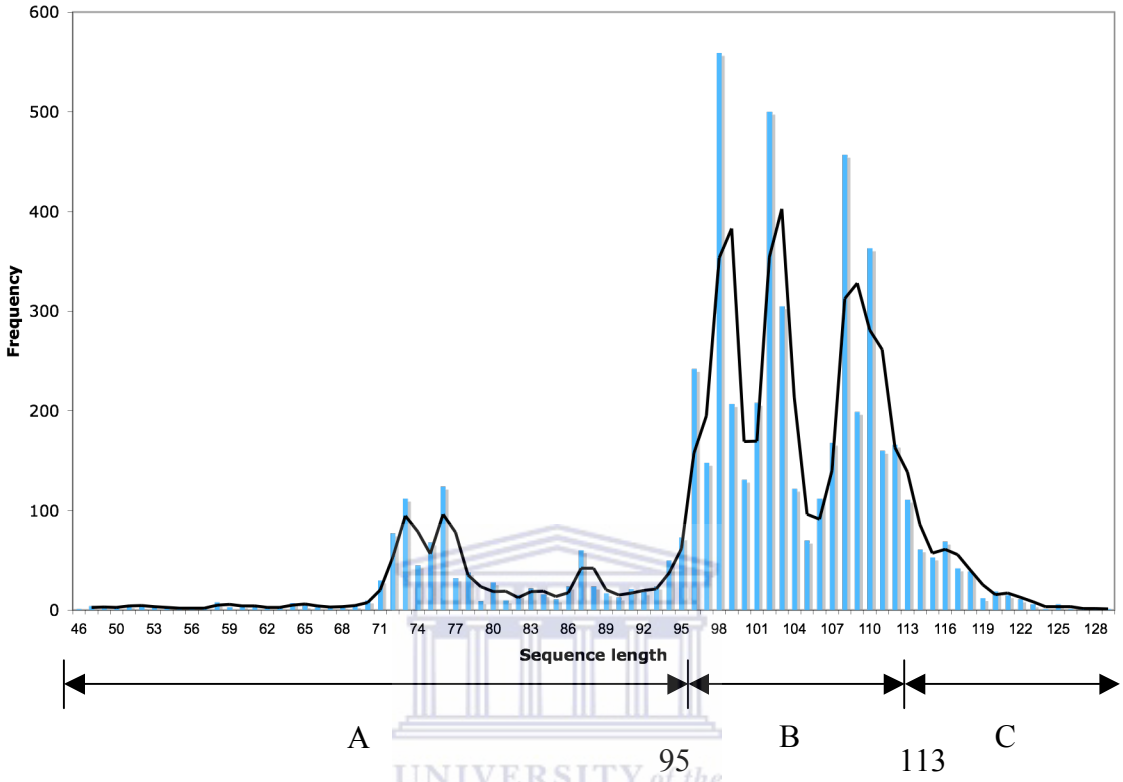
The GS20 Sequencing System (454 Life Sciences) allows high throughput direct sequencing of PCR amplicons and generates short fragments with an average length of between 115 to 130 bases. The work in section 3.3.6 was used to determine the region of the NBS domain that gives the most phylogenetically informative sequence data. Results of Figure 3.9 show that the region between amino acid 133 and 174 (corresponding to the position between nucleotide 399 and 522) is sparsely populated with amino acid conservation scores of between 8 and 9 and contain a higher proportion of scores between 3 and 5 (section 3.3.6). The higher proportion of very low amino acid conservation scores shows hyper-variability while the presence of very highly conserved sites maintains the conserved role of motifs in that region. This region was selected for sequencing short fragments of RGAs that distinguishes between different NBS-LRR genes and allows estimation of copy number in the gene family.

PCR amplification products (Figure 4.7) were sequenced from the B tag (3' - end of the NBS domain) at Inqaba Biotechnical Industries (Pty) Ltd. *Malus x domestica* (Borkh.) cvs. Anna and Golden Delicious were sequenced and a total of 5620 sequences were obtained. The sequence data was analysed to remove all reads containing ambiguous base calls (Ns) then sequence lengths were plotted on a graph against frequency of occurrence to produce the result displayed in Figure 4.8(I). According to Figure 4.8(I), the sequence lengths can be used to classify data into three broad categories in that can be used to estimate sequence quality. These are shown in the graph as region A comprising of reads that have up to 77 bases, B with 96 to 112 bases and C with 113 bases and above.

Sequences in region A were shown to contain a high percentage of double-primed reads whereas those in C contained sequencing artefacts predominantly characterised by errors in sequencing through homopolymers. Region B contained the correct sequence reads as confirmed by BLAST homology searches. Results of the sequence filtering exercise are displayed in Table 16.

The distribution by length of sequences in the three categories was of the order 19, 73 and 8% for the regions (Figure 4.8(I)) A, B and C respectively. Individual sequences were classified as members of the apple NBS-LRR gene family through sequence assemblies (section 4.5.2). Collapsing the homopolymers in all sequence reads to analyse the rate of premature termination of pyrosequencing produced a graph that approximates to a normal distribution of sequencing cycles (Figure 4.8(II). This analysis shows that homopolymer-related sequencing artefacts produced by pyrosequencing could give rise to longer than average sequence reads per average number of cycles. However, data with such sequencing artefacts can be eliminated through classification of sequence data by length then characterising a sample from each size category using BLAST homology searches.

Figure 4.8. Sequence length distribution for the 454 data and analysis of frequency related sequencing accuracy. (I) Regions A (up to 95 bases), B (96 to 112 bases) and C (113 and above) show the main size related quality classifications of the sequence data. (II) Analysis of sequencing progress per fragment length in the absence of homopolymers.



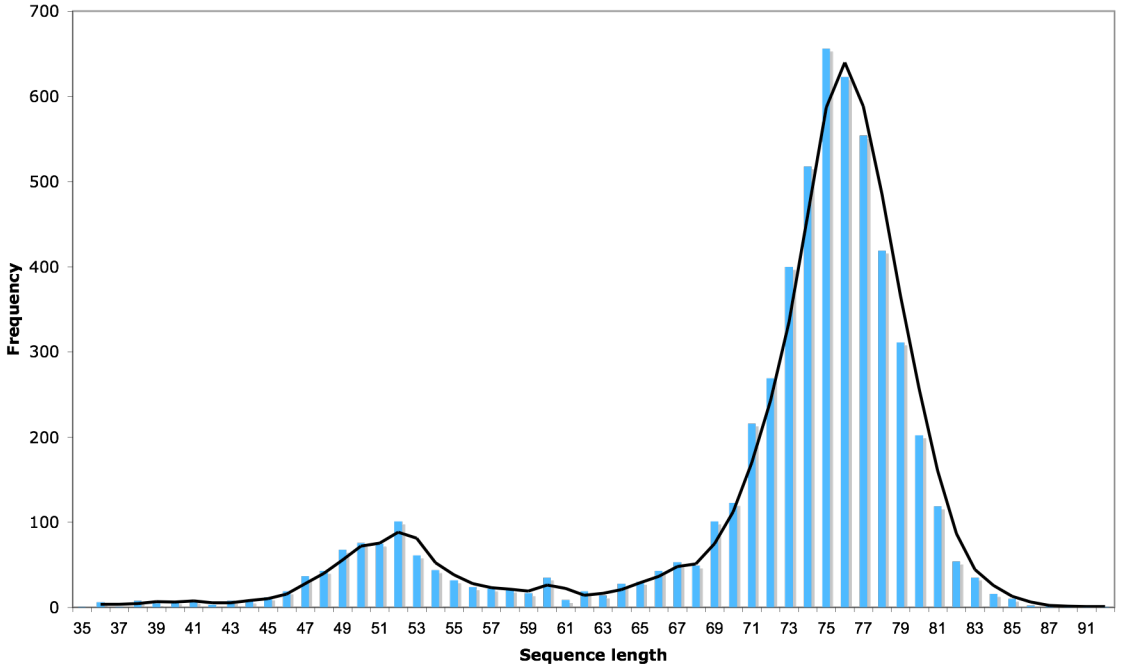


Table 16. Classification of GS20 sequenced data based on DNA fragment size and amount of ambiguous base calls. Only those sequences classified in the 'average of 110 bases' column were considered for sequence assembly.

Cultivar	Total	Data with sequencing artefacts	RGAs	Ambiguous base calls
Anna	3326	1333	1993	31
Golden Delicious	2294	864	1430	18
Total	5620	2197	3423	49

4.5. Sequence assembly and mutation detection

4.5.1 SNP detection in the re-sequenced dataset

The 61 sequences obtained for oligonucleotides GD1, GD2 and Anna3 (section 4.4.3) were trimmed of vector sequences and analysed using Blast homology searches (section 2.13.2). These were then assembled using CodonCode Aligner (section 2.13.2.1). The percentage identity for both the alignment and assembly options was set at 95% for the pre-assembly. A total of 9 contigs were obtained, 3 for each oligonucleotide pair and 4 sequences were unassembled. Sequences that remained unassembled were not regarded as singletons due to relatively bad sequence runs compared to the other 57 distributed into 9 contigs.

This result shows that the oligonucleotides although designed using sequences from clusters showing a high degree of sequence homology, they were not unique to single genes hence the distribution into different contigs. Two of the oligonucleotides, GD1 and GD2 were designed from clusters 11 and 2 (Figure 3.5). However, it was assumed that each of the contigs recovered per oligonucleotide represents a single gene and thus sequences within them constituted possible alleles of that gene. This assumption justifies the mixed gene alternative explanation for multiple SNP-like mismatches identified in section 4.4 (Figure 4.3). The initial predictions of candidate SNPs per contig were performed using CodonCode Aligner software. The predicted candidates were then manually analysed and a decision was made to either accept or reject them.

Contig 1 (GD1) with 26 sequences contained 7 predicted candidate SNPs and only one was polymorphic in both *Malus x domestica* (Borkh.) cvs. Anna and Golden Delicious though it had a low frequency in both cultivars. The other two were only polymorphic in cultivar Anna and not Golden Delicious and thus cannot be used for linkage mapping. Two other contigs from GD1, contig 4 and 5 had 3 and 6 sequences respectively and candidate SNP analysis on them was regarded as uninformative. However, random differences that are cultivar-specific were noted, giving the assumption that oligonucleotide GD1 amplifies a possible three genes from cluster 11 (Figure 3.5).

Contig 7 (Anna3) with 10 sequences showed 8 candidate SNPs that were polymorphic in both cultivars, Anna and Golden Delicious. The other two contigs showed random

mutations distributed throughout the entire contig. Although candidate SNPs were detected in these, they were classified as polymorphic in one not both cultivars.

Contig 2 (GD2) had 21 sequences that were highly homologous sequences with a total of 11 random mutations in the whole contig. The second contig from GD2 had 9 sequences (4 from Golden Delicious and 5 from Anna). Only cultivar specific mutations were detected at two positions and these were not accepted as candidate SNPs. The last contig for GD2 had only two sequences and thus could not be analysed. According to these results it seems GD2 oligonucleotides amplify DNA fragments from a possible 3 genes in cluster 2 (Figure 3.5). Figures 4.9 and 4.10 show graphical representations of contig 1 and 7 for GD1 and Anna3 respectively.

Figure 4.9. Candidate SNPs identified in contig 1 with sequences from the GD1 oligonucleotide pair. The figure shows 2 of the 3 candidate SNPs and these are highlighted in red font and indicated by red arrows 1 and 2. The amino acid translation in the third ORF was used to show that these are RGA sequences in their positive orientation.

2
1

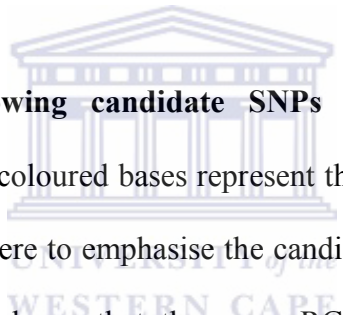
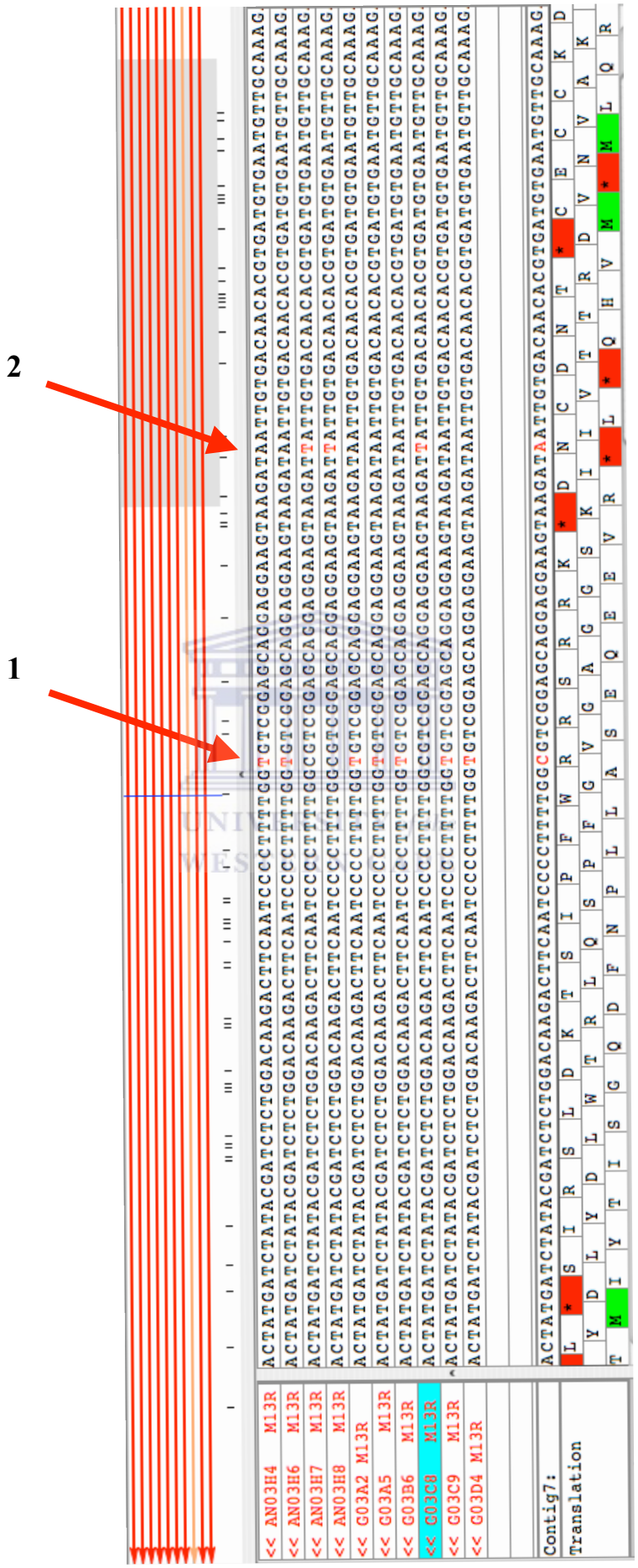


Figure 4.10. Contig 7 showing candidate SNPs identified from the Anna3 oligonucleotide pair. The red coloured bases represent the polymorphic nucleotides and the thick red arrows are used here to emphasise the candidate SNPs. The correct ORF in the second translation frame shows that these are RGA sequences in their positive orientation.


2
1
<< AN03H4 M13R
<< AN03H6 M13R
<< AN03H7 M13R
<< AN03H8 M13R
<< G03A2 M13R
<< G03A5 M13R
<< G03B6 M13R
<< C03C8 M13R
<< G03C9 M13R
<< G03D4 M13R
Contig7:
Translation

4.5.2 Analysis of sequencing coverage: full-length NBS domain sequences

The initial dataset with 661 sequences (section 3.2.5) was assembled again using CodonCode Aligner (section 2.13.2.2). Phred20 quality scores were used for screening trace files and minimum percent identity, minimum overlap length and word length were set at 95, 50 and 25 respectively. The assembly process and comparison of contigs were performed using the built-in assembly algorithm and ClustalW options respectively.

Contigs were manually edited and orientated in the proper 5' to 3' direction using the consensus translation in all three forward frames. The assembly was passed through several rounds of manual editing, un-assembly and re-assembly respectively to incorporate unassembled trace files. This process was terminated three cycles after the point where the number of unassembled trace files had stopped decreasing. A comparison of individual contigs against each other was performed to confirm the reliability of sequence distribution in the assembly. At this point sequences that were not incorporated into contigs were taken to be singletons within the sequenced dataset. This assembly was made up of a total of 50 contigs and 179 singletons from the cloned and Sanger sequenced dataset, Table 17 shows details of this assembly project.

Table 17. Sequence assembly of the Sanger dataset.

Cultivar	Total representation	Contigs with 2 sequences	Singletons
Anna	17	10	88
Golden Delicious	16	7	51
Bulked sample	2	2	30
Anna + Golden Delicious	14	1	

4.5.3 Comparative assembly of the GS20 and Sanger sequenced datasets

The 661 full-length NBS domain sequences (section 3.2.5) were then added to sequences generated using the GS20 sequencing system. These were assembled using CodonCode Aligner (section 2.13.2.2). Assembly parameters included minimum percentage identity, minimum overlap length, word length, minimum score and maximum unaligned gap and these were set to 90%, 25, 15, 10 and 75% respectively. The average aligned region was about 100 bases, the reverse priming site in the GS20 sequencing dataset was not removed prior to the assembly meaning that an average of about 350 bases remain unaligned ends.

This assembly generated a total of 266 contigs, translation of the consensus sequences in each contig was performed in three forward frames and the contig orientation was set

based on the GLPL motif in the reverse oligonucleotide of the GS20 sequence data. Since priming sites were trimmed from the full-length NBS domain dataset prior to the assembly, occurrence of the P-loop motif in the consensus translation was regarded as indication of sequencing errors in the GS20 sequence data. The presence of such artefacts resulted in either individual sequences or the whole contig being deleted depending on how the errors were distributed in the contig. All sequences introducing too many gaps in a contig were either deleted or the contig was split to remove the individual sequence. After removing sources of error from individual contigs, the whole sequence assembly was unassembled. Subsequent data training assemblies were performed using the approach described here before the final sequence assembly.

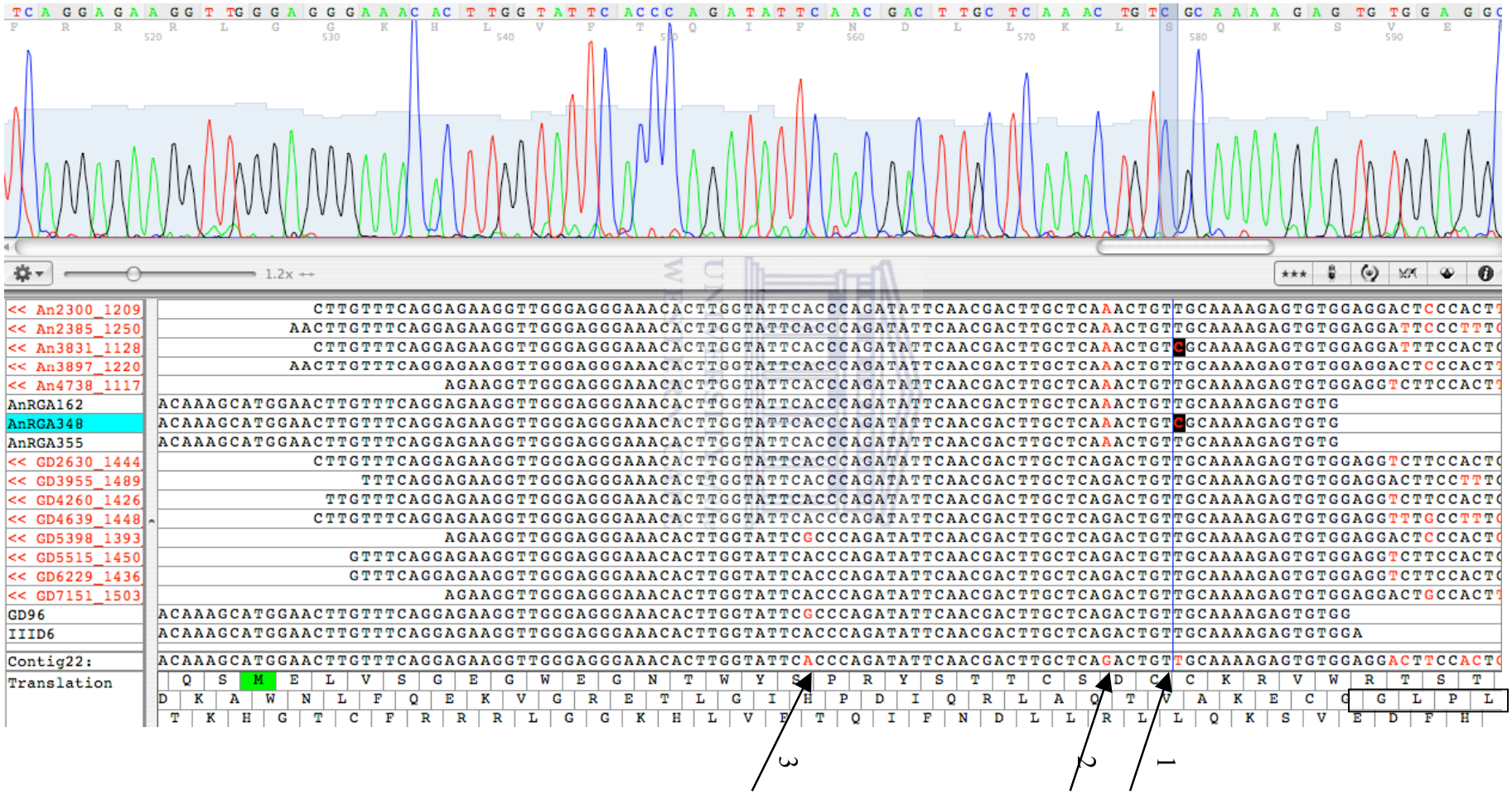
The final assembly was performed using stringent parameters, the minimum percentage identity, bandwidth (maximum gap size), maximum unaligned end gap and word length parameters at 95%, 25, 98 and 15 respectively. A comparative analysis of all contigs was performed (section 2.13.2.2) to confirm the reliability of individual contigs. A total of 278 contigs were obtained and these comprised 54 that had at least one full-length NBS domain sequence, 224 made up entirely of GS20 sequence data and 841 singletons. The quality of contigs made up of both GS20 and full-length NBS domain data is shown in Figure 4.11. The overall distribution of sequences in contigs is presented in Table 18. A total of 2859 sequence reads were incorporated into 278 contigs leaving 841 singletons. A total of 34 contigs contained both GS20 and Sanger data whilst only 20 contigs were constituted entirely of Sanger data. This shows that Sanger and GS20 datasets can be

assembled together though the challenge becomes that of making a clear distinction between mutations and sequencing errors.

Table 18. Sequence distribution in the 278 contigs of GS20 and Sanger datasets

Number of sequence reads	Contigs/Singletons
1	841
2	82
3	43
4 – 6	56
7 – 9	28
10 – 19	41
20 – 30	11
31 – 60	16
61 – 100	6
> 100	2

Figure 4.11. A contig containing both GS20 (454 Life Sciences) and Sanger sequence datasets. The second translation frame shows sequence orientation and position of the reverse oligonucleotide highlighted in the box. The chromatogram is from the Sanger sequenced AnRGA348 read showing a candidate SNP (nucleotides highlighted in black in the assembly image and blue in the sequence chromatogram). Sequence identities in red and black font are from the GS20 and Sanger datasets respectively. The sequence chromatogram does not match the scale of the assembly image, it is added here to show prove the reliability of data used in this analysis. Arrows labelled 1, 2 and 3 are candidate SNPs.


1.2x
<< An2300_1209
<< An2385_1250
<< An3831_1128
<< An3897_1220
<< An4738_1117
AnRGA162
AnRGA348
AnRGA355
<< GD2630_1444
<< GD3955_1489
<< GD4260_1426
<< GD4639_1448
<< GD5398_1393
<< GD5515_1450
<< GD6229_1436
<< GD7151_1503
GD96
IIID6
Contig22:
Translation
3
2
1

The 95% minimum sequence identity used in the assembly takes into account homopolymer-related errors associated with GS20 sequencing and also problems associated with PCR amplification of the DNA fragments. At higher minimum percentage identities and/ longer 'word lengths' the number of contigs recovered was significantly reduced and also more singletons were observed. The effect of mismatches that were characteristic of the reverse priming site (3' end of the assembly shown in Figure 4.11) were also taken into account.

Assembled contigs showed a number of candidate SNP-like mutations as indicated by arrows 1, 2 and 3 in Figure 4.11. Arrows 1 and 3 show two candidate SNP-like mutations that are heterozygous in cultivar Anna and Golden Delicious respectively and not in both. The third candidate SNP-like mutation (arrow 2) is homozygous in both cultivars. However, these and other SNP-like mutations could not be verified in GS20 data due to the lack of sequence chromatograms that could be used to distinguish between mutations and sequencing artefacts.

4.7. DISCUSSION

Targeted sequencing of gene families using PCR products although widely used as a technique of choice generates data that in some cases is subject to amplification and sequencing artefacts. PCR amplification biases the percentage representation of members of a gene family. The presence of secondary structures or variations in nucleotide conservation at a given site (Figure 3.9) results in differences in the priming efficiency. Subsequently differences in priming efficiency in turn affect PCR amplification efficiency. The result becomes a PCR product with a percentage representation that is biased against genes that are altered in the priming sites. Cloning amplification products relies on the concentration equilibrium between copy numbers of each amplified fragment and the available amount of cloning vector. The consequence of which is that DNA fragments that are under-represented in the PCR product will also form fewer successful ligation reactions. Transformation into cells and the numbers of transformants that survive to form bacterial colonies either reduces the possibility of some fragments being sequenced or selects against some. The GS20 sequencing technology eliminates the cloning step and allows direct high throughput sequencing of PCR amplicons where all possible amplified gene copies are sequenced regardless of percentage representation in the PCR product if sequence depth is deep enough.

In order to estimate the possible size of the NBS-LRR gene family in *Malus x domestica*, the GS20 system was used to sequence the 3' ends of the PCR amplicons and generate reads with an average length of 100 nucleotides (section 4.4.4.4). Analysis of site-specific evolutionary rates of the NBS domain revealed that this part of the domain has a fair

representation of highly conserved and hypervariable regions. High throughput sequencing applied here should give an indication of the approximate size of the NBS-LRR family since absolute sequence identity in the hypervariable region should expected in alleles of a gene and not possible between two genes. The 5 620 sequences from the GS20 system combined with 661 from cloning of PCR products gave a total of 6 280. The advantage with this dataset was that it provided optimal length of the cloning and sequencing system and the numbers from the high throughput GS20 sequencing system. Assembling these two datasets in a single CodonCode Aligner assembly allowed a better resolution between multiple copies of the same gene and mixed genes. This process generated 278 contigs, 54 of which had at least one copy from the 'full-length' dataset and 224 from the GS20 sequence dataset. This result shows the extent to which sequencing of cloned PCR products as a strategy could be limiting the number of genes being recovered from the NBS-LRR family. However, the assembly also gave 841 singletons, 35 from the 'full-length' dataset and 806 from GS20 sequencing. One of the problems in estimating saturation becomes the question of how to treat the singletons. These could be artefacts from the sequencing process or genuine single genes isolated on the genome or both. This question will only be resolved by sequencing and related assembly of the apple genome.

A comparative analysis of *Malus* RGA sequences available in GenBank and from the sequencing done in this project shows that genes in the NBS-LRR family occur as clusters in the phylogenetic tree (Figure 3.5). This has also been confirmed by various analyses performed on this family to date (Michelmore and Meyers, 1998; Pan *et al.*,

2000a; Baldi *et al.*, 2004; Xu *et al.*, 2007). These clusters represent genes that have a common recent origin and that multiplied through tandem gene duplication and thus could share common loci on the linkage maps (Pan *et al.*, 2000b). Results of comparative analyses performed on clusters between cultivars in the same species and between species have shown that gene copy number per cluster varies (McHale *et al.*, 2006). Possibly this level of variability reflects cluster activity which could be linked to the range of pathogens a particular species has been exposed to.

Comparative analysis within the *Malus x domestica* (Borkh.) species (Figure 4.1) shows a significant conservation of clusters among cultivars. Sequences accessed from GeneBank for 3 different cultivars were all incorporated into already existing clusters from cultivars Anna and Golden Delicious. The number of clusters was shown to increase as previously un-clustered sequences become incorporated into new clusters with addition of more sequences from other cultivars. This suggests a high degree of homology among orthologs from different cultivars, which is expected given that these cultivars are selections from a common parent and that the timescale of genetic evolution takes millions of years to accumulate to levels that can be detected by phylogenetic analysis.

Given the level of sequence homology even among different species of the same genus, a comparative analysis of *Malus* orthologs performed using *Malus x domestica*, *M. prunifolia*, *M. buccata* and *M. floribunda* had to be rooted using a very distant species. *Pinus* (a member of the gymnosperms) diverged from Angiosperms about 137 – 152 Myr (million years) before the proposed monocot – dicot split (Chaw *et al.*, 2004). Because of

this distance in time and evolutionary history between the two species *Pinus monticolor* RGAs were used as an outgroup in the comparative analysis of *Malus* species RGAs (Figure 4.2). The phylogenetic tree in Figure 4.2 inferred with molecular clock enforced suggests co-existence of TIR and non-TIR sub-families as branches A and B. Branch B is exclusively composed of TIR genes that also duplicate to produce 4 TIR clusters indicated by black dots. Branch A however duplicates to give A_1 and A_2, where the former duplicates to give four clusters of TIR and non-TIR (two of each). The split in that branch is indicated by the red dotted line in Figure 4.2. The sub-branch A_2 however, gave rise to 6 clusters 5 of which are made up of TIR genes and the sixth cluster (indicated by a blue dotted line) composed of genes that cannot be classified as TIR or non-TIR. These genes have been labelled as TIR-A_2 in Figure 4.2. Comparative analysis of conserved motifs in genes under branch A_1 shows that whereas the majority of TIR genes encode a signature LI(I/V)LDDVDQ kinase-2 motif (Table 9 and Figure 3.6), the TIR- A_2 genes encode a signature LLVLDDVDD. In the kinase-3 motif they encode a signature GSRIVIxTRN as opposed to GSRIIITTRD found in the TIR genes. There are other subtle differences between TIR and TIR- A_2 genes that are not mentioned here. It appears from results of Figure 4.2 and section 3.3.5 that the TIR sub-family has clusters that are undergoing diversifying selection and creating new subfamilies of the NBS-LRR genes.

However, the *Pinus monticolor* non-TIR genes are located within branch A_1 together with those from *Malus*. This either implies homoplastic evolutionary paths or that division into TIR and non-TIR predates the conifer – Angiosperm split. However, this statement cannot be confirmed in this work. It is certain though that gene duplications did

occur later after speciation to increase the size of the sub-families in *Malus*.

SNP discovery using this data set seems informative after re-sequencing using cluster specific oligonucleotides. All the three oligonucleotides designed from multiple alignments of all sequences from selected clusters (clusters 2 and 11 of Figure 3.5) amplified mixed genes. Results of sequence assemblies of the re-sequenced RGAs gave 3 contigs per cluster (section 4.5.1). Identification of candidate SNPs was performed per individual cluster and a total of 4 were identified from contigs 1 and 7 from GD1 and Anna3 oligonucleotide pairs respectively (Figures 4.9 and 4.10). These candidate SNPs are going to be tested in chapter 6.

CHAPTER 5

ANALYSIS OF EXPRESSION PROFILES OF THE RGA SEQUENCE DATABASE

CONTENTS

CHAPTER 5: ANALYSIS OF EXPRESSION PROFILES OF THE RGA SEQUENCE DATABASE

5.1 INTRODUCTION

The NBS-LRR resistance gene family is made up of clusters of genes that are responsible for a variety of defence related functions. To date a number of studies have been conducted on this family through sequencing of ESTs, quantitative/ semi-quantitative RT-PCR technologies and microarrays (Maleck *et al.*, 2000; Budak *et al.*, 2006; Cheung *et al.*, 2006) with various results being obtained. Budak *et al.* (2006) investigated the constitutively expressed *Agrostis* species NBS-LRR genes and reported the first monocot non-TIR subfamily of genes using ESTs. Cheung *et al.* (2006) working on *Medicago truncatula* used a combination of GS20 sequencing and Sanger sequencing technologies to uncover a novel set of ESTs that had not been identified before with the conventional cDNA clone library and Sanger sequencing approach. What has become apparent is that the high throughput sequencing capacity of the GS20 system gives a significant advantage in gene discovery and annotation.

It has already been shown in chapters 3 and 4 that RGAs are distributed into clusters and that these clusters in turn represent genes that could be located in physical proximity on the genome. The conservation of cluster identity among cultivars of the same species has been demonstrated in chapter 4, meaning all cultivars of *Malus x domestica* (Borkh.) probably share the same clusters though activities of individual genes reflect the range of pathogens the plant has been exposed to.

It is also clear from earlier work that assemblies of targeted high throughput sequencing of gene families have the potential to group genes into contigs. Using this information this chapter combines high throughput targeted sequencing of RGAs and the sensitivity of quantitative real-time PCR to investigate the transcriptome of selected members of the NBS-LRR genes following experimental inoculation of apple seedlings with fungal pathogens. This work aims to identify genes that encode functional resistance proteins that can be bred into commercial apple varieties through breeding programs.

5.2 Sample collection

Uninfected apple seedlings grown under greenhouse conditions were used in this experiment. These were progeny from the two crosses Carmine x Simpson and Lady Williams x Prima segregating for powdery mildew and scab trials respectively. In these crosses Carmine is known to be resistant to powdery mildew and Prima is resistant to scab. Seedlings budded on rootstocks were allowed to grow for three months under sterile greenhouse conditions with filtered air, sterile water, fungicide sprays and restricted access from the outside. Uninfected leaf samples were collected from these seedlings before they were either control-infected with spores of *Venturia inaequalis* (apple scab) or allowed to stand in the open for infection with spores of *Podosphaera leucotricha* (powdery mildew). The *Venturia inaequalis* strain used for the scab infections was a field isolate cultured and maintained at Bienne Donne Experimental farm (section 2.5). The powdery mildew fungus (*Podosphaera leucotricha*) is an obligate parasite and cannot be isolated and maintained on nutrient agar.

Seedlings infected with apple scab were grouped into 5 classes using the Chavalier scale (Chevalier *et al.*, 1991). Details of how the classification was carried out are presented in Figure 5.1. Classes 1 and 2 classified as showing a hypersensitive reaction and class 4 which shows symptoms of susceptibility following scab infection trials were sampled for analysis; class 3 shows a characteristic chlorotic reaction that is not compatible with the gene-for-gene disease resistance model and thus was not considered in this analysis. Class 1 shows typical signs of a hypersensitive response characterised by the formation of necrotic lesions. The response reaction in class 2 is characterised more by chlorotic lesions although they do not develop to be as severe as those shown in classes 3a and 3b (Figure 5.1). Class 4 seedlings show no evidence of a resistance reaction and the leaves are characterised by clear sporulating lesions (Figure 5.1).

Seedlings infected with powdery mildew were also selected based on the extent of fungal growth and severity of powdery blotches on the leaf surfaces. Most of the seedlings displayed an infection phenotype characteristic of systemic acquired resistance. Some seedlings showed symptoms of severe infection at the bottom, which grew progressively milder up the plant, in other seedlings the top leaves were uninfected. Samples were collected from five seedlings and details of the symptoms are given in Table 19 and Figure 5.2.

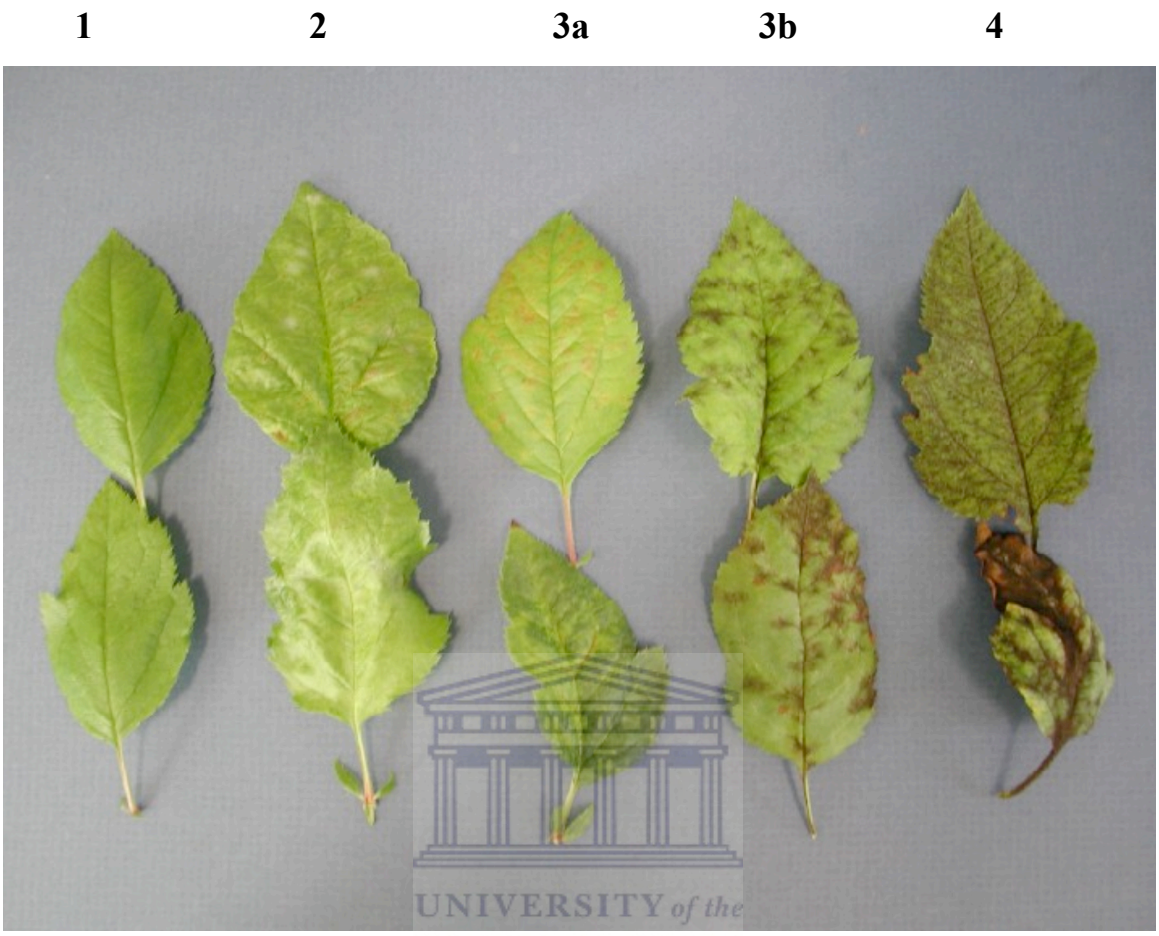


Figure 5.1. Scab infected leaves from seedlings classified into 5 categories using the Chavalier scale. Class 1 and 4 seedlings show a characteristic hypersensitive response characterised by necrotic lesions where class 1 is resistant and 4 is susceptible; class 2 seedlings are resistant and thus show mild disease symptoms; Class 3 is split into 3a and 3b characterised by chlorotic lesions with symptoms more severe in 3b.

Table 19. Seedlings of the Carmine x Simpson cross sampled for transcriptome analysis. Leaves were sampled at two positions showing differential resistance phenotypes on the same plant in all seedlings. These were representative of the range of symptoms displayed in this cross.

Seedling	Description of the seedling phenotype after infection
19-G12	The bottom had a weak fungal invasion although the leaves had a yellowish colour. The top leaves had pronounced necrotic lesions. Generally the whole plant had a mild infection as evidenced by fewer whitish fungal spores.
10-C2	The bottom part of the seedling was extensively infected with whitish fungal spores covering about 85% of the infected leaves. The top leaves were clean with no sign of infection.
19-B14	The seedling had stunted growth and the general infection levels were generally uniform from top to bottom.
19-G8	A small fraction of the total leaf count was uniformly infected from top to bottom with the infected leaves showing extensive fungal invasion. The plant had a lateral shoot with all the leaves showing total susceptibility to the infection.

Figure 5.2. Powdery mildew infected seedlings representative of the range of symptoms in the Carmine x Simpson cross. 'A' has infected leaves at the bottom and uninfected leaves at the top; 'B' shows a mild infection at the bottom, uninfected leaves at the top and two shoots that are susceptible; 'C' has generalised distribution of the same symptoms on the whole plant with mild fungal colonisation; 'D' has stunted growth with bottom leaves showing chlorotic lesions.

A

B

C

D

Leaf samples were collected from all seedlings before and after infection and selection of the candidates with representative symptoms was performed after comparative analysis of the observed phenotypic responses. Leaves were collected into re-sealable plastic bags then snap frozen in liquid nitrogen. These samples were then stored at -80ºC.

5.3 Sequencing of cDNA PCR amplicons

5.3.1 RNA extraction

Isolation of total RNA from uninfected and infected leaves (section 5.2) was performed using the RNeasy Plant Mini Kit (Qiagen) as described in section 2.6.22(a). Isolations for scab and powdery mildew samples were performed separately. This was done to avoid contaminating the uninfected samples. Pipettes were treated with RNaseAway® in between the two batches to reduce the likelihood of cross contamination. Isolated total RNA was resolved using denaturing agarose gel electrophoresis (section 2.12.2). The agarose gel is shown in Figure 5.3, showing intact 25S and 18S ribosomal RNA fragments. This shows that the isolated RNA was not degraded and could be used for PCR amplification.

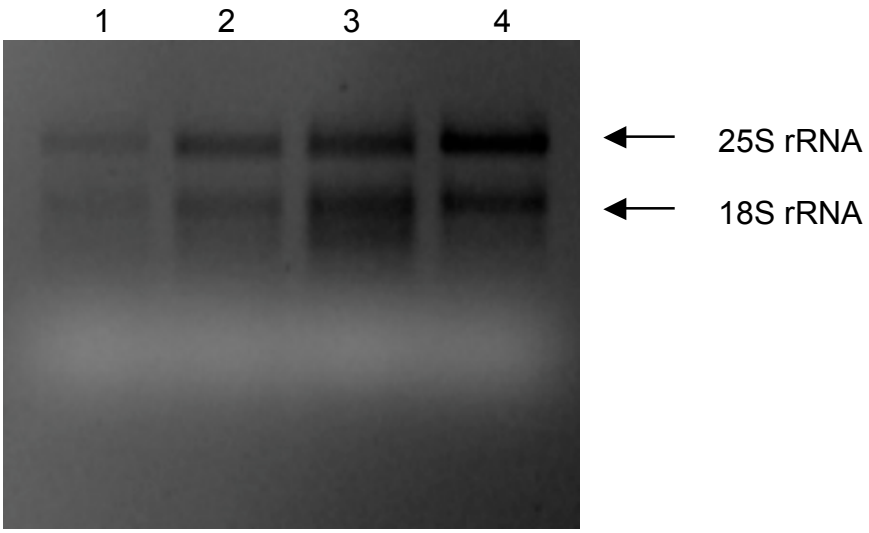


Figure 5.3. Total RNA isolation from powdery mildew infected seedlings. Lanes 1 to 4 consist of Mildew uninfected, Mildew infected, Scab uninfected and Scab infected respectively. The 28S and 18S RNA bands are evident in the agarose gel images.

5.3.2 PCR amplification of RNA extracts

First strand synthesis of cDNA was performed as described in section 2.7. The cDNA was used as a template for the PCR amplification of the NBS domain of candidate NBS-LRR gene transcripts using fusion oligonucleotides NBSFA and NBSR-2B (section 4.4.4.1). In this case PCR amplification was targeted at the NBS domain of mRNA transcribed from functional NBS-LRR genes either in the presence or absence of pathogen infection. The amplification products are shown in Figure 5.4 as DNA fragments that are ± 550 nucleotides in length and shown here co-migrating with the

0.517 kb fragment of the pTZ/*Hinf* I DNA size standard. The DNA fragments were visualised with minimal exposure to long wavelength UV light and excised from the agarose gel. The amplified nucleotide fragments were then recovered from gel plugs through gel extraction (section 2.6.1.2).

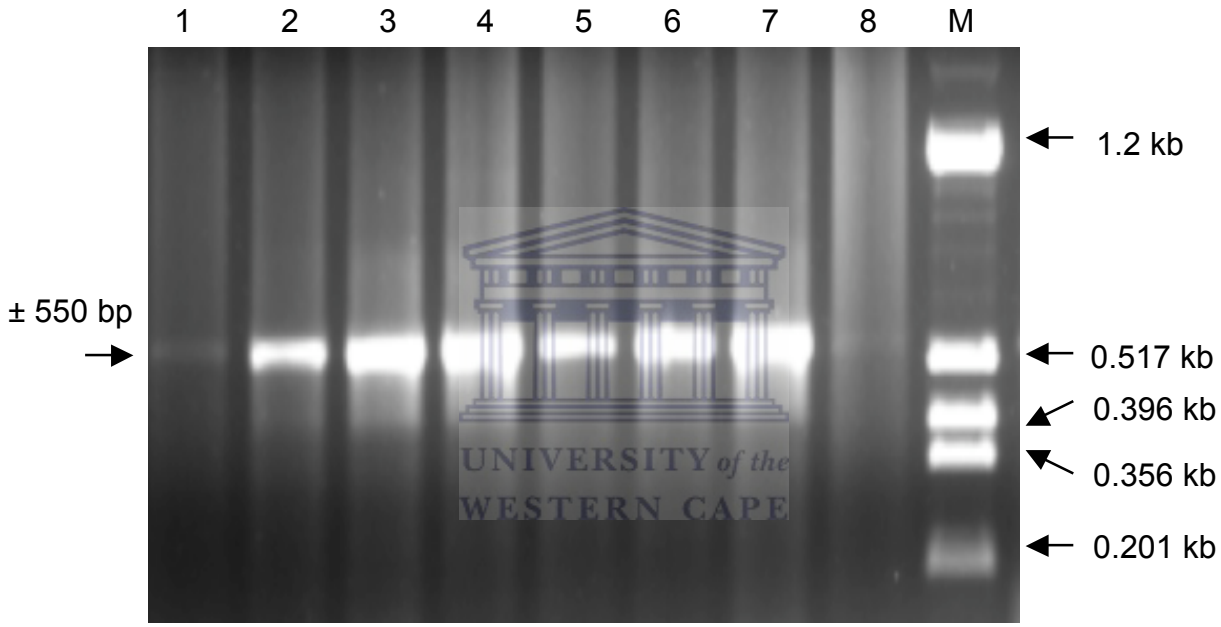


Figure 5.4. PCR amplicons from cDNA of infected and uninfected apple leaves. Lane M shows the profile of the pTZ/*Hinf* I DNA size standard. The expected DNA fragments in lanes 1 through 8 are shown co-migrating with the 0.517 kb fragment of the DNA size standard. Lanes 1 – 4 contains samples from Powdery Mildew samples 19-B14 and 10-C2 before and after infection (Table 19); lanes 5 through 8 class 1 and 4 samples before and after infection with scab.

5.3.3 High throughput sequencing of PCR amplicons

The purified PCR products from cDNA flanked by the GS20 sequencing tags (A on the 5' end and B on the 3' end) were sent for sequencing at Inqaba Biotechnical Industries (Pty) Ltd using the GS20 sequencing system. Emulsion PCR (emPCR) was performed using the GS emPCR Kit III (Roche) containing SPRI beads capable of specific binding to the A tag. Sequencing was then performed from the B tag, which was selected based on the level of sequence variability shown in the site-specific amino acid conservation plot (Figure 3.9). A total of 8811 sequence reads were generated, 4422 and 4389 sequence reads for powdery mildew and apple scab experimental infections respectively. The sequences had an average length of 110 nucleotides that encompasses the GLPL and parts of the RNBS-C motif (Figure 3.9). The GLPL and RNBS-C motifs play a role in binding ATP and the amino acid conservation in the RNBS-C motif is highly variable between TIR and non-TIR genes of the same plant species (Meyers *et al.*, 2003; Takken *et al.*, 2006). The sequence between these two motifs is hypervariable as shown in Figure 3.9, meaning this length of sequence could be used to differentiate between genes in the NBS-LRR family and subsequently to profile the transcriptome.

5.3.4 Comparative analysis of genomic and cDNA sequence datasets

The GS20 sequence data was labelled to indicate the respective treatment applied to the sample. Sequence data for both powdery mildew and scab experiments were pooled to give two datasets containing infected and uninfected data per infection trial. Each set was then analysed together with the 661 full-length NBS domain sequences from chapter 3 and 4. Table 20 shows the format used in generating the two datasets.

Table 20. The format used in generating apple scab and powdery mildew transcriptome sequence assembly datasets.

Infection trial	Sequence information	Number of sequences
Mildew	Infected	1424
	Uninfected	2998
	Full-length NBS domain (ABI sequenced)	661
	Total	5083
Scab	Infected	1995
	Uninfected	2394
	Full-length NBS domain (ABI sequenced)	661
	Total	5050

Sequences were assembled with CodonCode Aligner using parameters defined earlier (section 4.5.3). The average aligned regions were about 100 bases in length and the reverse priming site in the GS20 sequence dataset was not removed prior to the assembly to allow for the orientation of contigs in the 'sense' direction.

Translation of the consensus sequence in three open reading frames was also used in this assembly to allow for an easier inspection of individual contigs (section 4.5.3). Several consecutive data training assemblies were performed with each assembly facilitating for the removal of data with sequencing artefacts. Contigs with an average length of 70 nucleotides or less were discarded (section 4.4.4.4). Each contig in the data training assemblies was inspected individually before being passed on to subsequent assemblies. This process eliminated data with ambiguous nucleotide calls and sequences introducing single gaps in the alignments. The multiple rounds of assembly used here allowed for the incorporation of data that had not been assembled in preceding cycles possibly as a result of errors in respective contigs. The process was terminated at a point where no further reduction in the number of singletons could be achieved and contigs appeared optimal for transcriptome analysis.

The final assembly was accepted as representative of the possible transcriptome profile of the NBS-LRR genes in cultivars Anna and Golden Delicious. This transcriptome analysis was limited to NBS-LRR genes whose NBS domain is flanked by the P-loop and GLPL motifs complementary to the oligonucleotides used to generate these sequences. Given that the same set of oligonucleotides was used to generate both the transcriptome analysis and the full-length NBS domain datasets, this analysis was also intended to probe the former to identify sequences that are most likely genes coding for resistance against powdery mildew and apple scab in apple trees. Results of this analysis in Table 21 show a total of 101 contigs in the apple scab dataset, 15 of which are made up of sequences of mRNA present exclusively following inoculation of the plant with the *Venturia*

inaequalis field isolate (section 5.2). There are 31 contigs comprised of sequences from mRNA present in the plant before the inoculation and were possibly suppressed as they do not show after the inoculation. There are also 25 contigs made up of sequences that appear in the plant both before and after infection, which a possible indication of genes that are constitutively transcribed in the plant.

In the powdery mildew dataset, a total of 142 contigs were obtained. There are 27 contigs comprised of genes occurring in the uninfected plant, 8 detectable following exposure to airborne powdery mildew spores (section 5.2) and 38 that appear to be constitutively transcribed. In both apple scab and powdery mildew cases a few genes were induced by pathogen infection compared to those that were constitutively transcribed. However, the analysis was only applied to contigs containing 4 or more sequences, those with 3 or less were not regarded as enough evidence of gene transcription. Details of this analysis are presented in Table 21.

The cDNA sequences identical to any of the reads in the full length NBS domain dataset were referred to as Hits. The number of Hits on any full-length NBS domain sequence(s) was taken as indication of transcription with more Hits showing the relative level of response to the particular treatment. The number of Hits on the full-length NBS sequences varied from as few as one cDNA sequence to more than 100 per contig. A total of 8 for apple scab and 5 for powdery mildew were obtained following infection and these were equated to genes whose transcription was induced by the respective pathogen. Selected results of the assembly were further analysed using quantitative real time PCR

to evaluate whether or not high throughput GS20 sequencing of PCR amplicons constitutes a reliable analysis of the NBS-LRR transcriptome following pathogen infection.

Table 21. Results of the NBS-LRR transcriptome analysis performed using high throughput PCR amplicon sequencing. Total contigs, non-TIR and Hits on Sanger sequences in this table refers to the total number of contigs obtained in the assembly in both infected and uninfected cases, contigs with non-TIR sequences and specific matches to the full length NBS sequences respectively.

Pathogen	Total contigs	Treatment	Number of contigs	Hits on Sanger sequences	GS20 only contigs	Contigs with 2/3 sequences	Non-TIR
Scab	101	Un-infected	31	11	20	30	2
		Infected	15	8	7		1
		Both	25	11	14		
Mildew	142	Un-infected	27	6	21	69	1
		Infected	8	5	3		
		Both	38	12	26		1

5.4 Quantitative real-time PCR

5.4.1 Oligonucleotide design

5.4.1.1 Contig specific oligonucleotides

The datasets from apple scab and powdery mildew experiments including the full-length NBS domain sequences were assembled together using the parameters described in section 4.5.3. Four contigs containing at least one full-length NBS domain sequence and an additional three made up entirely of GS20 sequence data were selected from this assembly. These included AnRGA346, GD36, GD120, GD211, contig 114, 122 and 142, where the first four contigs were renamed using the identity of one of its full-length NBS domain sequence. These contigs were selected as follows, two contained genes transcribed in the absence of infection and suppressed following the infection, four were made up of genes induced by the infection (two induced by apple scab and two by powdery mildew infections), the last contig from this selection was made up of genes induced by both apple scab in Lady Williams x Prima seedlings and by powdery mildew in Carmine x Simpson seedlings. Oligonucleotides specific to each of these contigs were designed in order to avoid cross hybridisations between them. These oligonucleotides and their sequences are shown in Table 22. Melting temperatures for the oligonucleotides were between 58 – 62ºC to ensure that comparative analyses could be performed simultaneously.

5.4.1.2 Cluster specific oligonucleotides

A set of 9 clusters with a percentage sequence homology above 85 was selected from the phylogenetic tree in chapter 3 (Figure 3.5). Sequences from these clusters were re-aligned using ClustalX and a pair of oligonucleotides was designed on highly conserved regions within the clusters, no degeneracy was allowed in the terminal 3 bases at the 3' end thus making the oligonucleotides specific only to genes from the respective clusters. This was done to eliminate the possibility of cross hybridisations between clusters and thus ensure a relatively high resolution between clusters. These oligonucleotides are labelled with a 'Clst' prefix in Table 22. Clusters 7 and 9 are not represented here given that they do not form real clusters (section 3.3.4.3). Clusters with sequences represented in the chosen contigs are indicated as such in Table 22.

Table 22. Oligonucleotides used in quantitative real-time PCR to evaluate the NBS-LRR transcriptome. The first seven oligonucleotide pairs were designed from contigs that were identified using GS20 sequencing and the last nine designed from clusters in Figure 3.5. Details of the oligonucleotides designed from the GS20 sequencing experiment are given in the table. F and R represent forward and reverse oligonucleotides respectively.

Oligonucleotide ID	Details	Sequence	
GD120_2r_c157	Uninfected	F	GGAAGAGTTCACTTGTGGTG
		R	TTTTGGGTTGCTTTGTTCTAAAGG
Contig142_sc	Scab	F	CTTTAGCCAGTATGCCTTTAGAAC
		R	TTTTTAAGTGCGAGTGGGAGAC
AnRGA346_sc.m	Scab, Mildew	F	CTAGTCGCTTCCTTGAAAATATCAG
		R	CCGCCTAAAGAGTTCCAGAG
GD211_2r_un_c130	Uninfected	F	TGATAGCATTTCTTTTCATTTCGAAG
		R	AGGCTCTCAAGTTAAAAAGTTTAAGAG
GD36_inf04_c39	Scab	F	AAATGTGAGGGAAAGGAAATTGG
		R	GGCATGCCAACTAAGTAGCTG
Contig114	Mildew	F	CTCTTTAATTGGAGAGCCTTTAGGAG
		R	TTTTGAGTGCCAGAGGGAG
Contig122_rinf	Mildew	F	AGATCGAAGTCGAACTTGAATTCATC
		R	AGGCCAGTGGGAAACC
Clst1		F	GCATTGTGTGACCATAGTTGG
		R	GAGCTGAAGAGCTTCATTGTTAG
Clst3		F	TGCAGGCTATTGAGCAAGAA
		R	ATGAGCATCCATCTGACCAC
Clst4		F	GGAGCTTGGATGTTACTTGG
		R	AGCTCTGGATACATACACTCACC
Clst5		F	TTGTGTTGGATGATGTGTGG
		R	CAATGACTCAAGCTGTGATGG
Clst10		F	CACCAAGAGATGCTTATGGAAC
		R	AATGTTATGCCATGCTCGAC
Clst11		F	TCTGGCTAAATAGCTCAAGAGC
		R	TTGGTTTGGTAAGGGGAGTAGA
Clst12	(GD211)	F	GGTGTACAGAGGAAGCTTTCC
		R	TAGCAAGCGTTCATCTCTAGTC
Clst13	(GD120, AnRGA346)	F	TGTGGTGCCGTACATATGC
		R	GCTCCATCACCACTTAGAACC
Clst14	(GD36)	F	CAGCTGCCAAAGCCATTTAT
		R	GAAACCGATACCTTCATCAACA

5.4.2 Oligonucleotide optimisation

PCR amplification conditions for all the 16 oligonucleotides (Table 22) were optimised using genomic DNA from cultivars Anna and Golden Delicious as template. Optimisation of annealing temperatures (section 2.8(a)) was performed using pre-selected temperature settings at 56.0, 56.7, 57.4, 58.2, 59.9 and 60.7ºC based on the melting temperatures for the oligonucleotides (section 5.4.1.1). An optimal annealing temperature of 58ºC was identified for all reactions as indicated in lane 4 of Figure 5.5. Optimal oligonucleotide concentrations for all reactions were identified as 0.5 µM (section 2.8(b)).

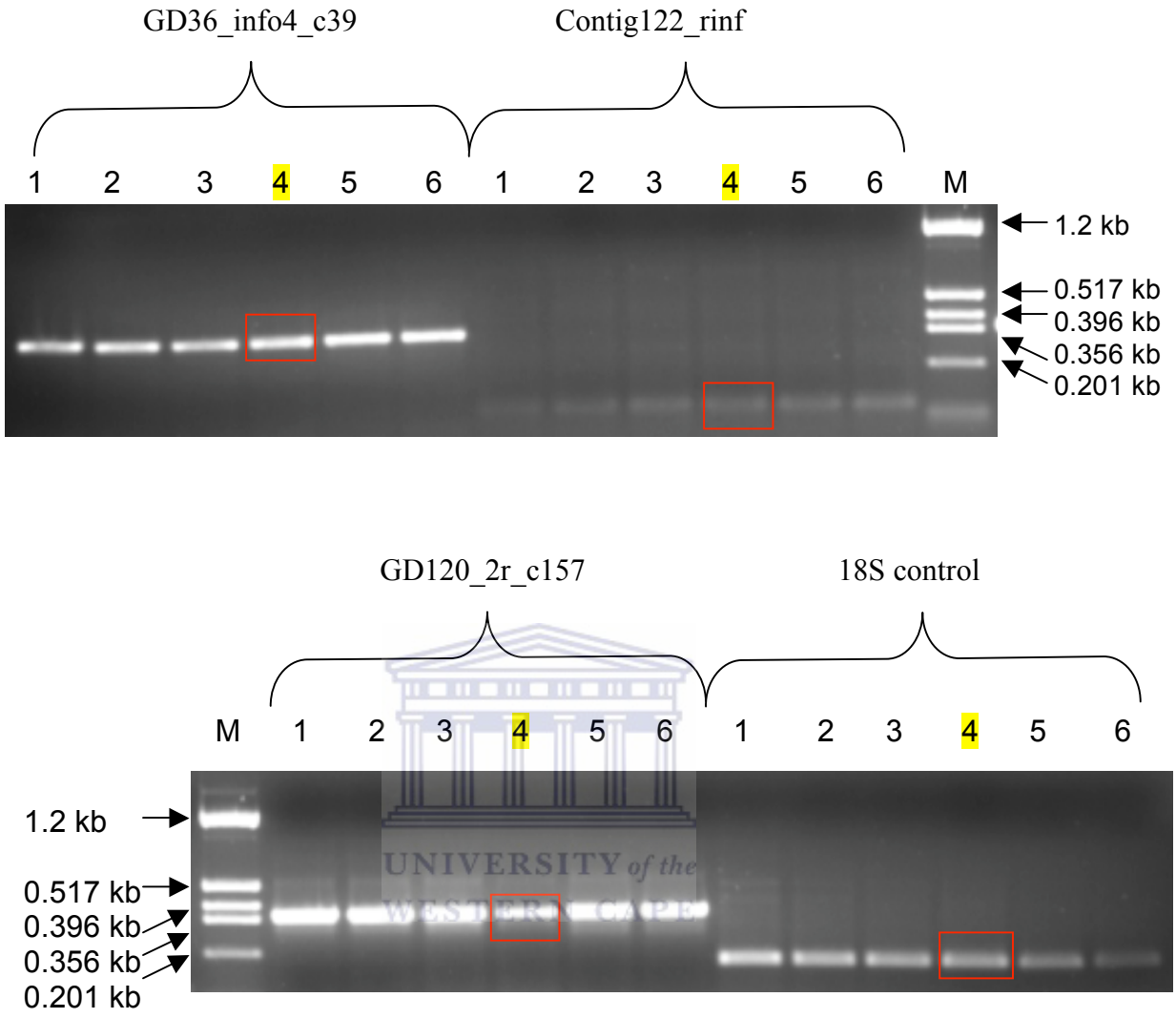


Figure 5.5. Optimisation of the annealing temperatures for 4 of the 16 oligonucleotide pairs. Lanes 1 through 6 represent PCR products that were amplified at 56.0, 56.7, 57.4, 58.2, 59.9 and 60.7ºC respectively. Lane M contains the profile of the pTZ/*Hinf* I DNA size standard.

5.4.3 Quantitative real time RT-PCR

Quantitative real-time PCR reactions were performed using all the 16 oligonucleotide pairs on cDNA produced from seedlings inoculated with apple scab (*Venturia inaequalis* field isolate) and those inoculated with airborne powdery mildew (*Podosphaera leucotricha*) spores including those that were not infected. Amplification reactions were performed in duplicates for individual samples. Four-sample dilution points 1, 1:10, 1:100 and 1:1000 were used to plot the standard curve using 18S ribosomal RNA oligonucleotides as the endogenous reference gene. Master mixes and cycle conditions were set up as described in section 2.8.2. Table 23 shows samples used for this analysis including the treatment and class of each sample where applicable.

Table 23. Samples used for quantitative real-time PCR amplification.

Sample ID	Cross	Disease
19-B14 19-E12 19-G12 10-C2 10-C8	Carmine x Simpson	Powdery mildew
A1-53 (Class 1) C10-52 (Class 2) F10-46 (Class 4)	Lady Williams x Prima	Apple scab

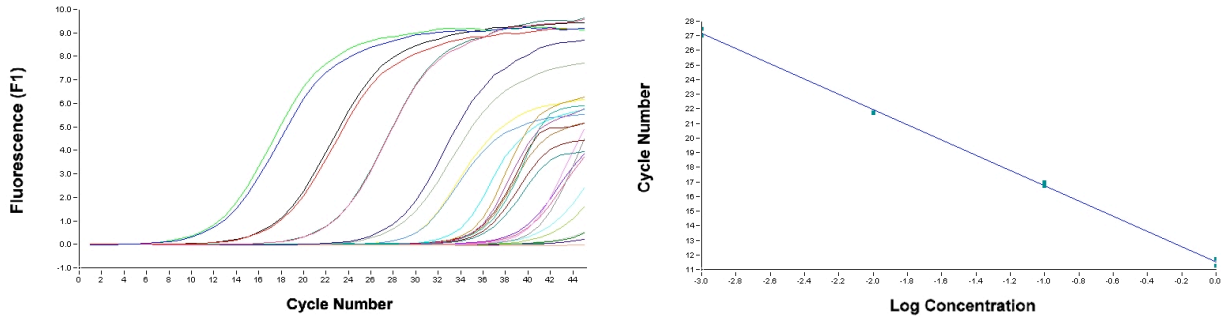
The reactions were performed using LightCycler software version 3 (section 2.8.2) that generates a standard curve for the endogenous reference gene by plotting dilution points on a two-dimensional plane as the logarithm of concentration against cycle number. The final curve is a plot of the input dilution points against observed crossing point above background. This allows determination of sample concentrations by direct extrapolation from the standard curve, where the unknowns are expressed as a ratio of the endogenous reference gene.

Using this software, four points were fitted using the fit points analysis method and arithmetic baseline adjustment options were used. The fit points method converts a sample's exponential curve to a straight line by plotting the logarithm of fluorescence against the cycle number (log-line). The arithmetic baseline adjustment method is the default option for SYBR Green I experiments, it calculates the mean of the lowest five measured points per sample and then subtract it from all measured points of that sample. The arithmetic baseline adjustment option is recommended for analyses in which there are sample – to - sample background variations. A 10-fold dilution of the sample cDNA was used and the analyses were performed in duplicate to assess the reproducibility of the crossing points (Ct values) (Jain *et al.*, 2006). Figure 5.6 shows the amplification and standard curves for some of the samples. The plot of the logarithm of cDNA concentrations against cycle number gave a straight line through all four fitted points and thus indicating a good level of precision in the dilution points.

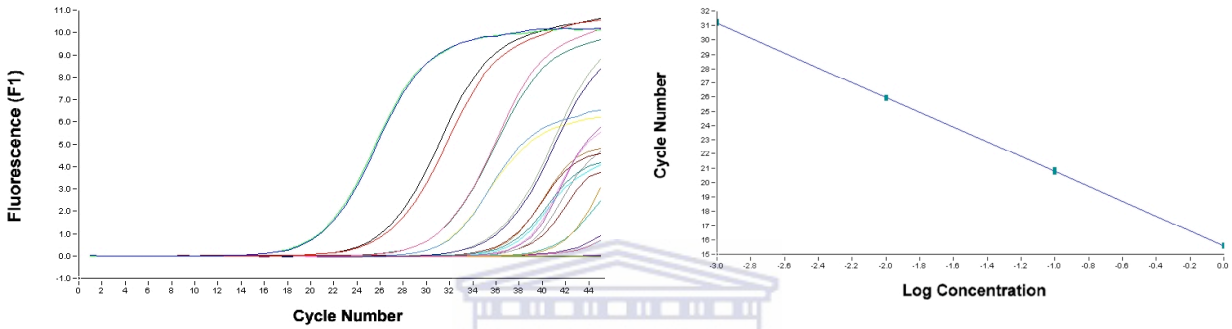
Figure 5.6. Amplification and standard curves for the quantitative real-time PCR analyses. Images A through D show the quantitative real-time PCR amplification and standard curves for the genes monitored in this experiment. Figure A is seedling 19-E12 before infection with powdery mildew; B and C are for seedling A1-53 before and after inoculation with the *Venturia inaequalis* field isolate respectively; and D is seedling C-2T after infection with the airborne powdery mildew fungal spores. The slope, linear coefficient (r) and PCR efficiency (E) values for each of the standard curves are shown below.

A: Slope = -5.208	Y intercept = 11.52	r = -1.0	E = 1.56
B: Slope = -5.197	Y intercept = 15.59	r = -1.0	E = 1.56
C: Slope = -5.32	Y intercept = 23.54	r = -1.0	E = 1.54
D: Slope = -5.199	Y intercept = 20.63	r = -1.0	E = 1.56

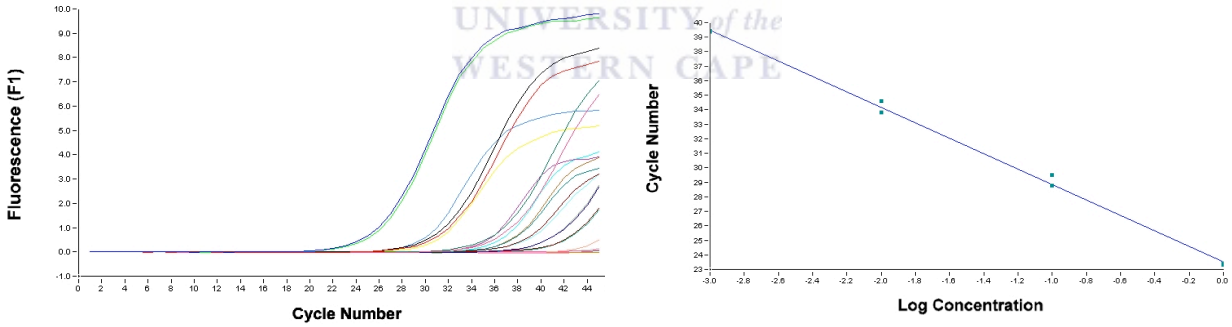
A



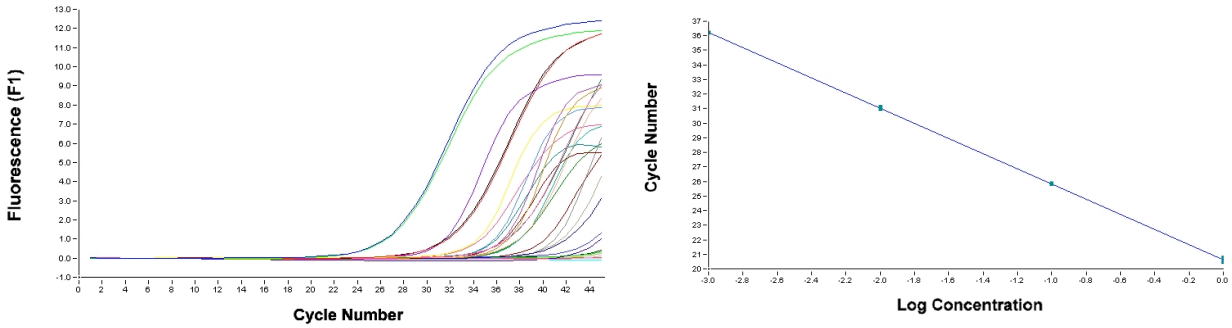
B



C



D



The accuracy of the relative transcriptome analysis based on quantitative real-time RT-PCR relies on the normalisation of raw CP values through the use of endogenous reference genes. A number of studies have been done to assess the stability of expression of a number of potential reference genes both in plants and animal systems. The work on rice endogenous reference genes showed that 18S and 25S rRNA are indeed ideal in plant systems because they show stable expression patterns under environmental stress and hormone treatment conditions (Jain *et al.*, 2006). Their use however is subject to two conditions 1) first strand cDNA synthesis should be done using random hexamers and 2) in assaying for low expressed transcripts, a 1:10 dilution should be used to contain their expression within the dynamic range of real-time PCR. Both these conditions were met in this analysis.

There are other endogenous reference genes that can also be used to normalise expression profiles of inducible gene systems in quantitative real-time PCR experiments. These include GAPDH, ß-tubulin, ß-actin and ubiquitin etc. Work done in sugarcane showed that expression levels of ß-actin and ß-tubulin were tissue specific and variations in expression level were detected between different tissues (Iskandar *et al.*, 2004). The expression levels of GAPDH were found to be stable in both sugarcane and rice (Iskandar *et al.*, 2004; Jain *et al.*, 2006), however, comparative analysis in different tissues showed that rRNA appears more stable compared to GAPDH in different tissues and at different developmental stages (Zhang and Hu, 2007).

Having established a stable endogenous reference gene as a control the inter-assay correlation of results can be assessed using PCR efficiencies obtained in respective assays. According to (Wilkening and Bader, 2004), the maximum possible PCR efficiency is 2, which represents replication of every PCR product in every cycle, and the minimum is 1 where no amplification occurs. PCR efficiencies are affected by factors such as sample purity, concentration and oligonucleotide dimers, similar experiments have reported efficiencies in the range 1.4 to 1.9 (Caldana *et al.*, 2007). The inter-assay correlation of PCR efficiencies should thus reflect the degree to which assays are comparative. Calculations of these values were performed here using Rasmussen's (2001) equation as shown below.

$$E = 10^{[-1/\text{-slope}]}$$

Where E is the PCR efficiency and 'slope' is the gradient of the standard curve and using this equation, the observed average efficiency for the assays is 1.56.

Calculation of relative transcription levels was performed using the MS Excel implementation of the relative expression software tool - multiple condition solver version 2 (REST – MCS© version 2) by Pfaffl *et al*. (2005). This software uses the equation;

$$R = \frac{(E_{target})^{-\text{CPtarget (Mean control – Mean sample)}}}{(E_{ref})^{-\text{CPref (Mean control – Mean sample)}}}$$

where R is the relative expression ratio, E_{target} and E_{ref} are the PCR efficiencies of the target and reference genes respectively; CP_{target} and CP_{ref} are the crossing points for the target and reference genes respectively (Pfaffl *et al.*, 2005).

Transcription ratios for genes assayed in seedlings of the two crosses were calculated using the uninfected state of the respective seedling as the reference. Normalisation of raw CP values was performed using the 1:10 dilution of the 18S rRNA endogenous reference gene. Results are given in Figure 5.7 for transcription profiles in Lady Williams x Prima seedlings and 5.8 for seedlings of the Carmine x Simpson cross. Figure 5.7 shows transcription ratios for classes 1, 2 and 4 that are described as hypersensitive, resistant and susceptible respectively in the Chevalier scale (section 5.2). The transcription levels of gene/oligonucleotide pair Ctg142 shows differential transcriptional levels for classes 1 (seedling A1-53) and class 4 (F10-46) in which the former is twofold higher. Class 2 (C10-52) however, shows suppression of Ctg142.

There is induction of GD36 in class 1 and consequently the same gene/oligonucleotide pair is suppressed in class 4 seedlings. Class 2 seedlings do not show either induction or suppression of GD36 following inoculation with the *V. inaequalis* field isolate. There is no correlation between transcription levels of Clst14 and GD36 although according to results of Figure 3.5 the latter is a member of cluster 14. This suggests that clusters are mixtures of genes with different functions and that oligonucleotides that hybridize to the whole cluster provide answers that are not useful with quantitative real-time PCR.

In Figure 5.8, seedlings from the selected set showed a range of responses with the set of oligonucleotides used. Seedling 19-G12 selected due to the presence of necrotic lesions on some of the leaves showed induction of gene/oligonucleotide pair Ctg142, which was more pronounced in the top leaves. Seedling 19-B14 showed stunted growth and a generalized distribution of whitish lesions on curled leaves. This seedling shows about a sevenfold reduction in the transcription levels of AnRGA346, Ctg114 and GD120 in the top leaves, although transcription levels in the bottom leaves are higher than all the other seedlings from the Carmine x Simpson cross. Seedling 10-C2 with extensive whitish lesions on the bottom leaves and uninfected/resistant top leaves also showed high transcription levels of AnRGA346, Ctg114 and GD120 although the difference in levels between the top and bottom levels were about 20% fold. The difference in transcription levels between seedlings 19-B14 and 10-C2 could reflect the strength of induction of the systemic acquired resistance in top leaves. Very low transcription levels of AnRGA346, Ctg114 and GD120 in the top compared to the bottom leaves following infection with powdery mildew could also be due to suppression of these genes hence the severity of infection in seedling 19-B14. However, more genes detected by the GS20 sequencing system and related assemblies (section 5.3.4 and Table 21) need to be analyzed using quantitative real time PCR to show how they change between uninfected and infected plants.

Figure 5.7. Analysis of transcription of selected NBS-LRR genes in seedlings of the Lady Williams x Prima cross following infection with *V. inaequalis*. The relative transcription ratios above uninfected states (y=0) are shown with grey bars representing class 1, orange for class 2 and blue for class 4. The important genes to note in this graph are AnRGA346, Ctg142, GD36 and GD120 identified using GS20 sequencing as indicated in Table 21;

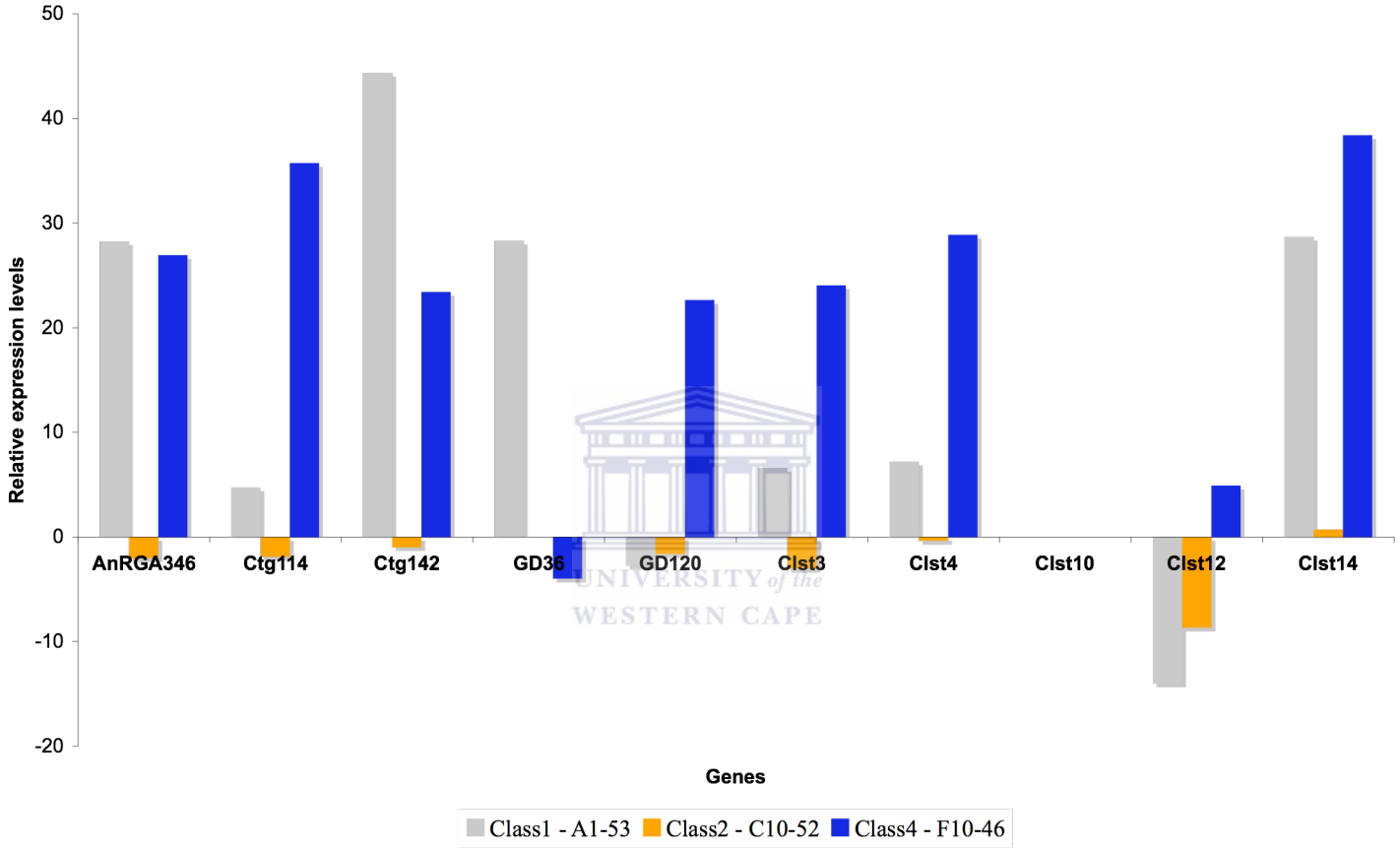
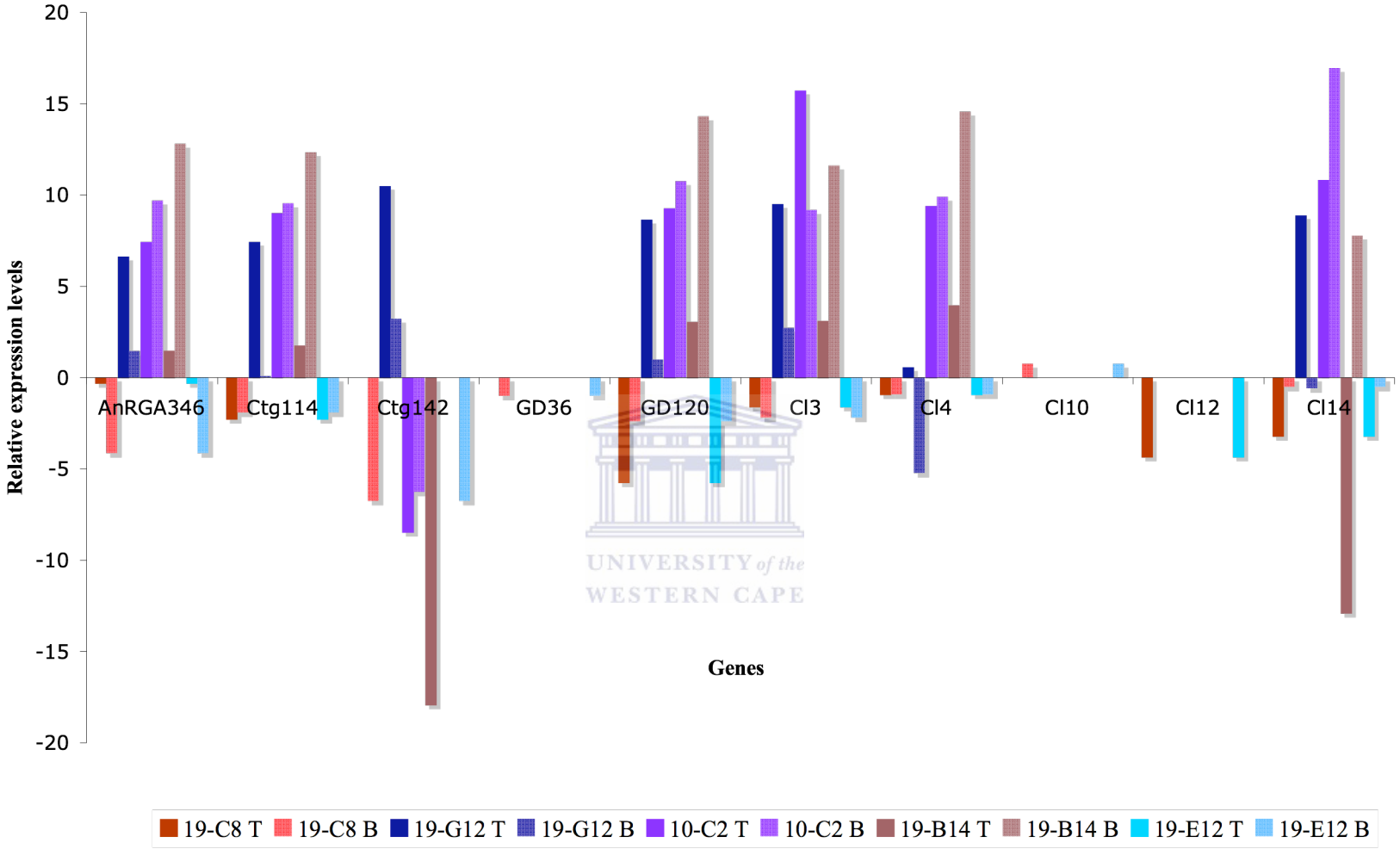


Figure 5.8. Analysis of transcription of selected NBS-LRR genes following powdery mildew infection in seedlings of the Carmine x Simpson cross. The relative expression ratios above uninfected states (y=0) are shown for the top (solid colour) and bottom (same colour with specks of white) leaves per seedling. Important genes to note in this graph are AnRGA346, Ctg114 and GD120 identified using GS20 sequencing.



5.5 DISCUSSION

The high throughput sequencing capacity of the GS20 system (454 Life Sciences) as used in section 5.3.4 enables high coverage of genes flanked by a P-loop and GLPL motifs. Its application in transcriptome analysis is based on the ability to clone PCR amplicons using emulsion PCR and thus sequence all available amplification products. This system has so far been applied in *Medicago truncatula* where thousands of novel transcripts were identified (Cheung *et al*., 2006). This technique is used here to investigate the NBS-LRR family transcriptome following infection of apple plants with *Podosphaera leucotricha* and *Venturia inaequalis*. The use of oligonucleotides flanking the NBS domain to amplify candidate NBS-LRR genes in cDNA generated large volumes of sequences representing genes activated in such conditions. Needless to say, the quality and reliability of results obtained in this analysis depend to a large extent on the quality of RNA isolated and subsequently the reliability of the reverse transcription system (Wilkening and Bader, 2004).

The RNA used was analysed for integrity and degradation using denaturing agarose gel electrophoresis and the quality and results are shown in Figure 5.3. The RNA concentrations were standardised to ensure that the starting templates were in a comparable range. Reverse transcription was performed using a mixture of oligo(dT) and random hexamers and all first strand synthesis reaction incubation times at 42ºC were all performed for two minutes to ensure that biases in the final cDNA copy number were kept minimal.

For GS20 sequencing, class 1 seedlings from Lady Williams x Prima and bottom leaves of Carmine x Simpson seedlings showing a typical hypersensitive reaction following infection were used for scab and powdery mildew trials respectively. A total of 8811 sequence reads were generated for this whole experiment. CodonCode sequence assemblies were performed against the 661 full-length NBS domain dataset. The assembly of scab infected and uninfected candidate RGAs gave a total of 101 contigs 25 of which were present before and after infection and thus were regarded as constitutively transcribed genes. Contigs with sequences of genes that were upregulated only after infection were assumed to be genes actively involved in the plant inducible defence system were estimated at 15 while 31 showed evidence of genes that were suppressed by the infection.

The same analysis performed for powdery mildew cDNA sequences gave a total of 142 contigs of which 38 were constitutively transcribed, 8 were induced by the infection process and 27 showed evidence of suppression. Genes that were assumed to be under suppression were those detected only in uninfected seedlings. These could be suppressed by pathogen-encoded mechanisms or alternatively they could be responsible for the negative regulation of inducible genes in the absence of pathogen infection.

Sequence homology between GS20 and full-length NBS domain data was used to filter the latter dataset for possible functional genes. However, a few contigs containing termination codons in the optimal open reading frame were identified in the assembly. Pseudogenes in the NBS-LRR family are generated mostly either through frame-shifts or

insertion of premature termination codons within the open reading frame (Meyers *et al.*, 2003); others could also arise from deletions and other forms of mutation. A small number of contigs with sequences that show evidence of constitutive transcription and those that appear to be suppressed following infection had pseudogenes. Work done in *Drosophila Ste* gene regulation reveals that pseudogenes mapping to the *Suppressor of Stellate* locus [*Su(Ste)*] regulate expression of Ste genes through repression of transcription and altering the splicing of *Ste* gene primary transcripts (Balakirev and Ayala, 2003). Thus, transcription of *Su(Ste)* pseudogenes provides competitive interaction with *Ste* for the positively acting transcription factors. This model could apply to plant defence systems as well. WRKY transcription factors have been implicated in regulating transcription of defence genes through the interplay of NPR1 and it has been shown that some of these transcription factors are constitutively present in the cytosol (Eulgem, 2006). It could be possible that transcription of pseudogenes in the absence of infection does provide a means for negative regulation of functional disease resistance genes in the absence of infection; this however, requires further experimentation.

The changes in expression levels observed in the GS20 analysis were verified with quantitative real-time PCR. For scab infection trials, results of quantitative real-time PCR correlate with classes of resistance (Figure 5.1) in classes 1 and 4 for AnRGA346, GD36 and Ctg142. GD120 and AnRGA346 were designed against contigs that show evidence of transcription following either scab or mildew infection, however, GD120 appears to be repressed in class 1. Class 2 seedlings have a completely different transcription profile with all the genes tested here. According to the Chevalier scale (Chevalier *et al.*, 1991)

class 2 is characterised by chlorotic lesions with no sporulation (Figure 5.1) and this class is closer to class 3 characterised by chlorotic lesions with pronounced sporulation in subclass 3b. The cause of differences in foliar disease or response symptoms and subsequently the source of resistance in these classes will also need to be investigated.

In powdery mildew trials (Figure 5.8), seedlings showed a variety of responses although results for AnRGA346, Ctg114 and GD120 were as expected from results of GS20 sequencing. The AnRGA346 and GD120 responses were similar though transcription levels were relatively higher in the later. Seedling 19-G12 was the only mildew-infected sample to show significant transcription levels for Ctg142. Consequently this seedling belonged to a group that showed weak fungal growth on the bottom leaves and a generally mild level of infection (section 5.2). The top leaves on these seedlings showed necrotic lesions in their response to infection. In scab-infected seedlings, class 1 had almost double the transcription level of Ctg142 compared to class 4. Clearly transcription levels of Ctg142 correlate with the Chevalier scale for Classes 1 and 4 in scab infected seedlings and its presence in 19-G12 could explain why these seedlings display a somewhat superior level of resistance to infection compared to the rest of the powdery mildew infected set. However, more genes still need to be tested from those observed in the GS20 analysis before gene functions could be inferred from this analysis.

Seedling 10-C2 from the class of seedlings characterised by extensive white fungal sporulation on the bottom leaves while those at the top are uninfected showed significant levels of AnRGA346, GD120 and Ctg114 after infection. The expression of Ctg142 was

repressed in these seedlings. Seedlings 19-B14 and 19-E12 belonged to a class that showed a generalised infection over the whole plant although a few bottom leaves had more whitish lesions on the back of the leaves (section 5.2). This class of seedlings also showed stunted and discoloured lateral or axial dormant shoots characterised by extensive sporulation. All the genes tested in this analysis showed evidence of total repression for seedling 19-E12 although 19-B14 had high levels of AnRGA346, Ctg114 and GD120 although Ctg142 was repressed. According to this result, seedling 19-E12 and 19-B14 do not belong in the same class although they present with comparable symptoms.

The differences in levels of transcription for top and bottom leaves also varied from one seedling to the next. Infection levels reflect the rate of progression of fungal invasion and progression of disease over the plant; this provides a level of internal control of expression analysis. Seedlings 10-C2 and 19-B14 show relatively low transcription levels for the top leaves compared to those at the bottom. On the contrary seedling 19-G12 had high levels of transcription at the top and phenotypic data shows hypersensitive responses on the top leaves. Minor differences in transcription levels between top and bottom leaves were noted for sample 10-C2 classified under seedlings showing extensive infection at the bottom and uninfected at the top (section 5.2). This could mean that the levels of transcription of some of these genes possibly following SAR induction determines the severity of fungal invasion for the top leaves

More quantitative real-time PCR analyses need to be done to investigate the source of resistance in seedlings classified under class 2 by the Chevalier scale. These seedlings showed total suppression of all genes tested in this analysis although they show a level of disease resistance superior to class 4. Possibly there might be two dominant defence mechanisms in which presence determines resistance and absence or presence in a mutated form shows total susceptibility. If this assumption is true then classes 1 and 4 have a different mode of defence to classes 2 and 3.

The response to powdery mildew infection appears to be encoded by a complex of genes that offer intermediate levels of defence. According to results obtained here, level of transcription, the number of genes transcribed and the respective difference in levels between the primary area of entry and other parts of the plant seem to affect the observed class of resistance. The difference in transcription levels of defence genes between the top and bottom suggests differences in the efficiency of long distance signal transduction systems.

The oligonucleotides designed from multiple sequence alignments of sequences in clusters give the overall function of a cluster. The problem here becomes cases in which there are genes that are both induced and suppressed in the same cluster. Results from such cases show an overall reduction in activity for the whole cluster. Transcription of AnRGA346 was detected in both powdery mildew and scab infected seedlings, although GD120 was more in powdery mildew. These two genes are members of cluster 13 although the Clst13 oligonucleotides showed an insignificant transcription level (not

shown in this chapter). Possibly there are other genes in this cluster whose transcription is suppressed thereby showing an overall cluster activity that is the average of induced and suppressed transcription levels. GD36 from cluster 14 is transcribed exclusively in class 1 infected seedlings though the Clst14 primers show a higher transcription level in class 4s. According to these results, oligonucleotides designed from multiple sequence alignments of sequences in a cluster show transcription levels of a family of genes. Other strategies that target individual genes using data from more comprehensive transcriptome sequencing projects could be employed to analyse absolute gene expression profiles. This will allow a greater understanding of gene/ gene cluster dynamics in disease resistance.

CHAPTER 6

ANALYSIS AND MAPPING OF CANDIDATE SNPS FROM THE RGA DATA

CONTENTS

CHAPTER 6: ANALYSIS AND MAPPING OF CANDIDATE SNPS FROM THE RGA DATA

6.1 INTRODUCTION

Molecular markers have contributed a lot to the efficiency of breeding for disease resistance, fruit quality and other traits of agronomic importance as well as in positional cloning of genes of interest. In breeding for disease resistance and fruit quality, markers developed from known functional genes or ESTs have proved valuable in the candidate gene mapping approach. Several attempts have been made to develop molecular markers for the genetic linkage mapping of R genes in plants with various levels of success. The NBS-profiling technique in which genomic DNA is digested using a restriction enzyme and degenerate NBS-specific together with adapter-specific oligonucleotides are used to amplify RGAs linked to known restriction endonuclease digestion sites has been developed (van der Linden *et al.*, 2004). This technique was used to generate a set of RGA specific markers and has thus far has been applied to apple, potato and lettuce (Calenge *et al.*, 2005; van der Linden *et al.*, 2004).

Other techniques such as sequence-specific amplification polymorphism (SSAP) were used to map resistance gene loci in barley, tomato and soybean (Hayes and Saghai Maroof, 2000; Waugh *et al.*, 1997). BAC physical maps were used in *Prunus* in which specific gene fragments were hybridized to arrayed peach BAC physical libraries and 42 R-gene locations were mapped with success (Lalli *et al.*, 2005). Baldi *et al.* (2004)

developed a set of SSCP and CAPS markers in apple using NBS-specific oligonucleotides designed on unique sites per cluster. These were located on 11 out of 17 linkage groups using the Fiesta x Discovery mapping population.

To date it has become clear that the use of techniques specifically targeting the NBS-LRR gene family have a higher chance of identifying R genes that can be used in breeding for resistance. It has thus far been established (chapter 3) that NBS-LRR genes occur as clusters of genes related by descent and that these clusters can be reconstructed using phylogenetic tools. A total of 661 RGAs have been cloned and sequenced (chapter 3 and 4) and were shown to distribute into 14 clusters. Each cluster is made up of a set of genes and their alleles (chapter 4) and the level of sequence homology between sequences in each cluster has been shown to be very high. Given the large number of genes in the NBS-LRR family and the apparent low level of sequence homology between clusters, high resolution linkage mapping of members of this family require a high throughput strategy. Methods that have been described thus far requires the design of a set of oligonucleotides for each gene within individual clusters, performing PCR amplifications for each one and revealing polymorphisms by sequencing or restriction endonuclease digestion followed by resolving DNA fragments on agarose gel electrophoresis.

This chapter describes the development of a candidate SNP-based technique that identifies and maps a set of markers in a Bin mapping (Howad *et al.*, 2005) population made up of sixteen progeny. It has been shown here (chapter 4) and in *Lolium perenne* (Xing *et al*., 2007), that the NBS domain of NBS-LRR genes contains a high density of

candidate SNPs (an average of one in every 33 bases) in some clusters though very low in others. SNP identification work presented in this chapter is performed in the *Malus x domestica* (Borkh.) Anna x Golden Delicious Bin mapping population that has been developed for linkage mapping using SSRs.

6.2 Sample collection from Bin mapping population

A set of 87 F_1 plants in the *Malus x domestica* (Borkh.) Anna x Golden Delicious mapping population was used to construct a genetic linkage map with 205 microsatellite markers. All the plants with at least 20% missing data were then removed from the initial population of 87 individuals. A set of fourteen plants 11, 12, 16, 19, 51, 52, 130, 276, 320, 327, 330, 335, 344 and 353 showing the most informative recombination breakpoints per linkage group were then selected to make up the Bin mapping population (van Dyk PhD thesis, UWC, 2008). Leaf samples were collected from these plants and also from *Malus x domestica* (Borkh.).

6.3 SNP identification by sequencing

6.3.1 DNA Extraction

Genomic DNA was isolated from sixteen samples from the Anna x Golden Delicious bin mapping population including the two parents using the CTAB method (section 2.6.1). The quality of the DNA was assessed through agarose gel electrophoresis (section 2.12.1) and results of the first eight extractions are shown in Figure 6.1. Figure 6.1 shows high

molecular weight genomic DNA co-migrating with the 23 kb fragment of the lambda/*Hind* III size standard. The DNA isolates obtained were thus accepted as sufficiently good quality for PCR amplification and were then stored at +4ºC.

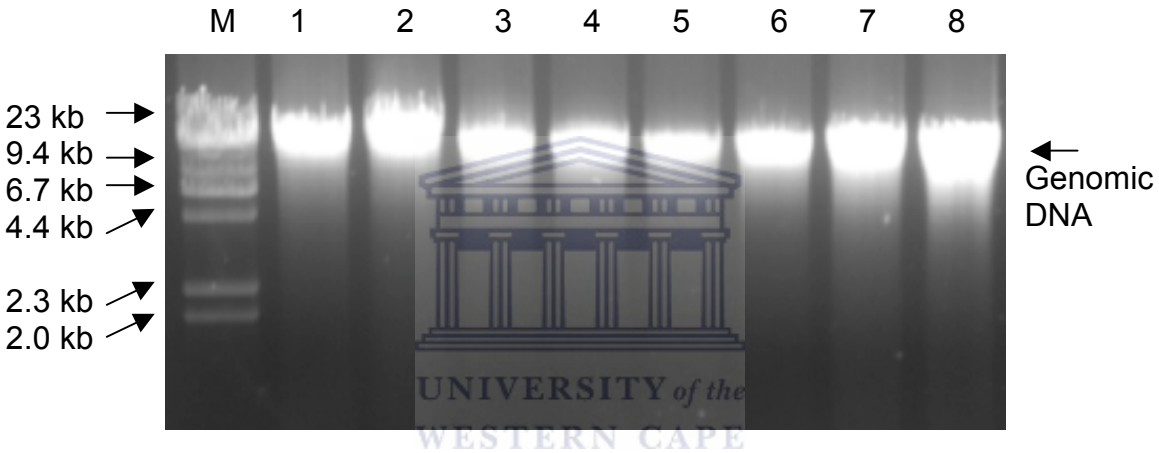


Figure 6.1. Some of the genomic DNA extracted from the bin mapping plants, Lanes 1 through 8 show genomic DNA isolations for samples Anna, Golden Delicious, 11, 12, 16, 19, 51 and 52. Lane M shows the profile of Lambda/*Hind* III DNA size standard.

6.3.2 PCR amplification and DNA fragment purification

PCR amplification of DNA fragments containing candidate SNPs were performed using oligonucleotides GD1 and Anna3 (section 4.4.1). Reactions were set up as described in section 4.4.2 with each 25 µl reaction containing ~50 ng of template DNA (Figure 6.1). Oligonucleotides GD1 and Anna3 flank DNA fragments that contain candidate SNPs identified in contigs 1 and 7 (section 4.5.1) respectively. PCR amplification products were resolved using agarose gel electrophoresis (section 2.12.1) and the results are shown in Figure 6.2. PCR amplification of genomic DNA using the Anna3 and GD1 oligonucleotide pairs produces fragments that are ± 362 and ± 462 nucleotides in length respectively (Figure 4.4). These DNA fragments are indicated by the arrows in Figure 6.2 A and B and thus confirm that the amplification process produced the required DNA fragments from each of the fourteen bin mapping seedlings. The PCR amplification products were then purified from the agarose gel (section 2.6.1.2) and stored at - 20ºC.

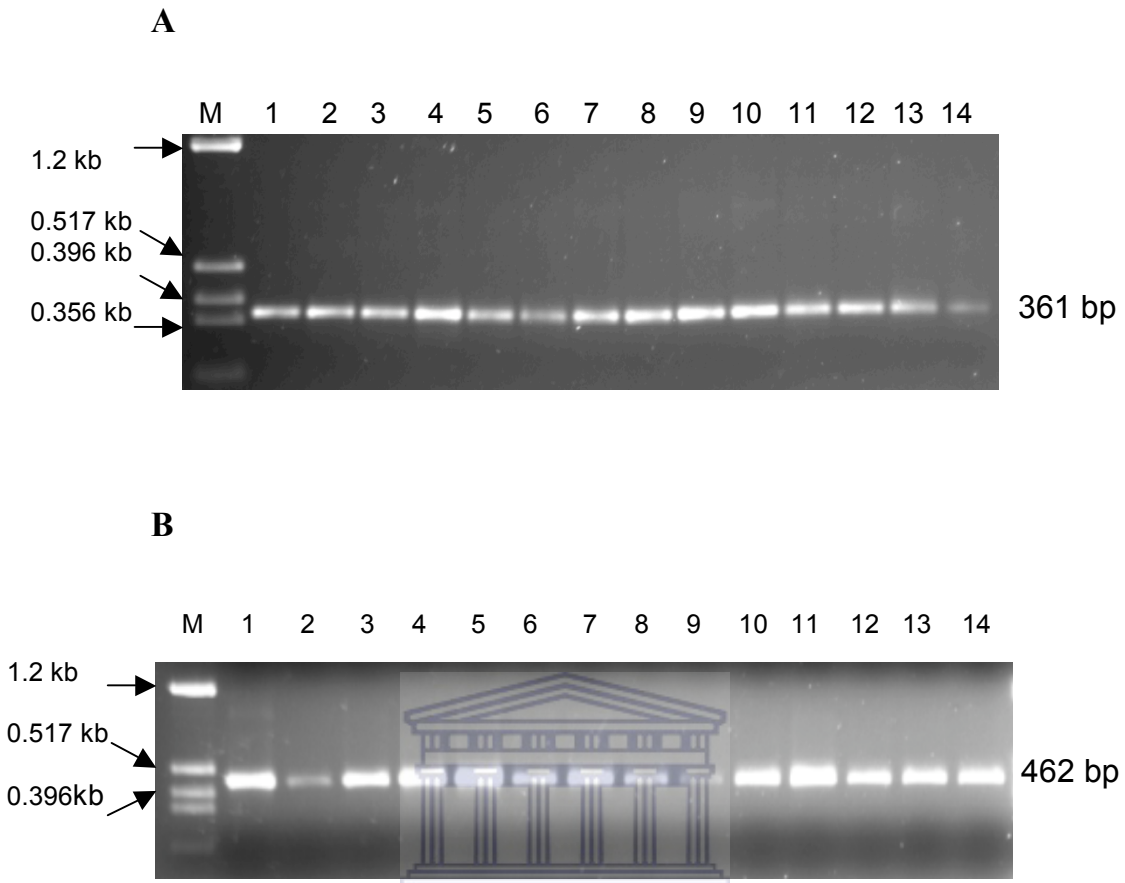


Figure 6.2. PCR amplification of DNA fragments containing candidate SNPs from the six bin mapping seedlings. A shows DNA fragments produced using the Anna3 oligonucleotide pair and B from the GD1 pair. Lanes 1 through 14 in both agarose gel images contain samples 11, 12, 16, 19, 51, 52, 130, 276, 320, 327, 330, 335, 344 and 353; lane M shows the partial profile of the pTZ/*Hinf* I DNA size standard.

6.3.3 Sequencing of PCR amplicons

Aliquots of the gel purified DNA fragments from section 6.3.2 were sequenced (section 2.13.1) using both the forward and reverse oligonucleotides used in section 6.3.2 for the

PCR amplification of these DNA fragments. A total of twenty-eight sequences were obtained, fourteen from each of the oligonucleotides Anna3 and GD1.

6.3.4 Identification and mapping of candidate SNPs

Sequences from the six bin mapping seedlings (section 6.3.3) and from cultivars Anna and Golden Delicious Anna3 and GD1 sets (section 4.4.3) were assembled together using CodonCode Aligner (section 2.13.2.1). Sequences from the bin mapping seedlings were incorporated into assemblies corresponding to contigs 1 and 7 identified in section 4.5.1.

6.3.4.1 The GD1 oligonucleotide set

Sequences from all 14 seedlings were assembled together with those from contig 1 (section 4.4.3) as shown in Figure 6.3. A total of 7 candidate SNPs were observed in the 462 nucleotide long contig thus giving an average of one SNP per 66 nucleotides. All the observed SNPs were homozygous in Golden Delicious; three were heterozygous in Anna and four were homozygous in both Anna and Golden Delicious as shown in Table 24. Results shown in Table 24 also show that all candidate SNPs that were heterozygous in Anna were also heterozygous in the bin mapping samples.

Figure 6.3. The assembly of sequences generated using the GD1 oligonucleotides showing the position of SNPs 5 and 6. AN01 and G01 prefixes are used here for data generated from sequencing DNA fragments amplified from Anna and Golden Delicious respectively using the GD1 oligonucleotide pair; the bin mapping samples include samples 11, 12, 16, 19, 51, 52, 130, 276, 320, 327, 330, 335, 344 and 353 in that order. The insert chromatograms are included to demonstrate the distinction between 'R' (sample 51) and G/G (sample 12) alleles for SNP 5.

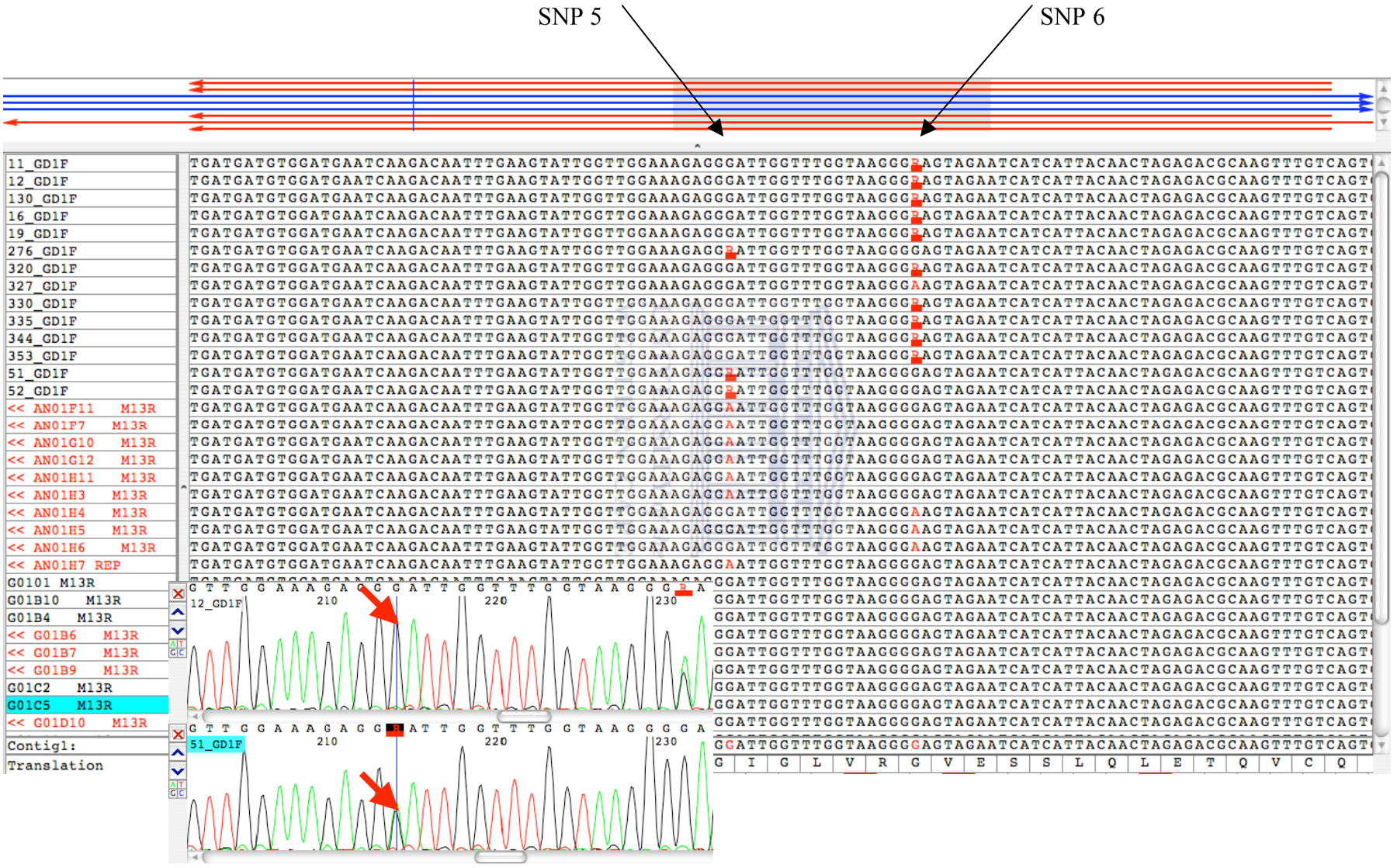

SNP 5
SNP 6
11_GD1F
12_GD1F
130_GD1F
16_GD1F
19_GD1F
276_GD1F
320_GD1F
327_GD1F
330_GD1F
335_GD1F
344_GD1F
353_GD1F
51_GD1F
52_GD1F
<< AN01F11 M13R
<< AN01F7 M13R
<< AN01G10 M13R
<< AN01G12 M13R
<< AN01H11 M13R
<< AN01H3 M13R
<< AN01H4 M13R
<< AN01H5 M13R
<< AN01H6 M13R
<< AN01H7 REP
G0101 M13R
G01B10 M13R
G01B4 M13R
<< G01B6 M13R
<< G01B7 M13R
<< G01B9 M13R
G01C2 M13R
G01C5 M13R
<< G01D10 M13R
Contig1:
Translation
12_GD1F
51_GD1F
210
220
230
G I G L V R G V E S S L Q L E T Q V C Q

Table 24. The distribution of SNPs observed in the GD1 series of Anna, Golden Delicious and 14 bin mapping samples. The number column represents the total number of observed SNPs; G. Delicious and Anna represent *Malus x domestica* (Borkh.) cultivars Golden Delicious and Anna respectively and Bin mapping samples includes samples 11, 12, 16, 19, 51, 52, 130, 276, 320, 327, 330, 335, 344 and 353. The candidate SNPs indicated with (*) show evidence of segregation in the bin mapping seedlings.

SNPs	Nucleotides	G. Delicious	Anna	Bin mapping samples
1	**A/A**	**A/A**	**A/A**	**All A/A**
2	**G/T**	**T/T**	**G/G**	**All G/T**
3*	**C/A**	**A/A**	**C/A**	**C/A**
4	**C/A**	**A/A**	**C/C**	**All C/A**
5*	**G/A**	**G/G**	**G/A**	**G/A**
6*	**G/A**	**G/G**	**G/A**	**G/A**
7	**G/A**	**A/A**	**G/G**	**All G/A**

Candidate SNPs 3, 5 and 6 marked by (*) in Table 24 were selected for mapping using the fourteen seedling bin mapping population to calculate possible locations on the Anna x Golden Delicious genetic linkage map. The SNP genotypes for individuals in this mapping population are displayed in Table 25.

Table 25. SNP genotypes identified for individual samples in the Anna x Golden Delicious bin mapping population. The 'Bins' columns show the overall genetic inheritance of the heterozygous SNPs in the bin mapping population (section 6.4.5).

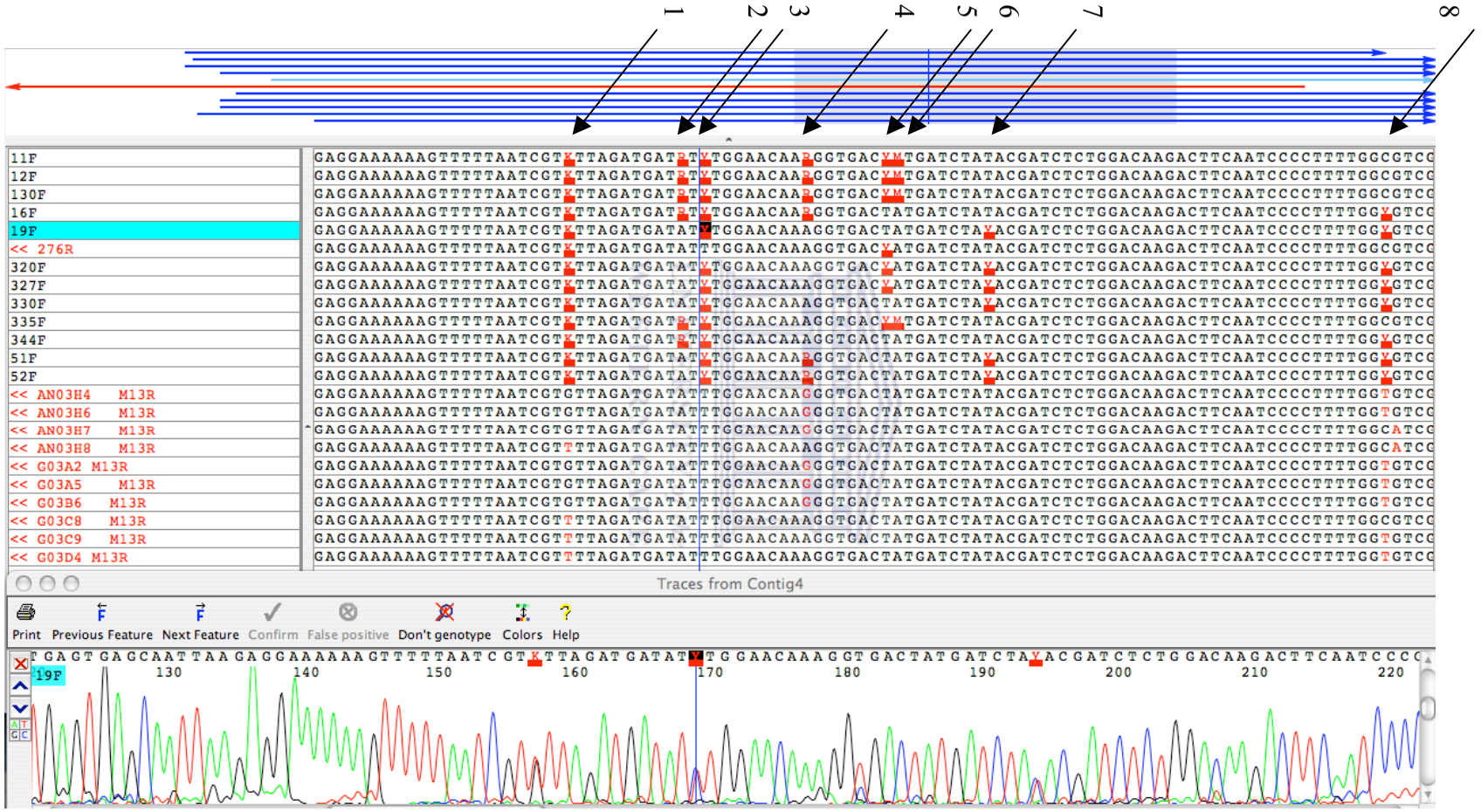
Sample	SNP3	Bins	SNP5	Bins	SNP6	Bins
Anna	A/C	1	A/G	1	A/G	1
GD	A/A	0	G/G	0	G/G	0
11	A/C	1	G/G	0	A/G	1
12	A/C	1	G/G	0	A/G	1
16	A/C	1	G/G	0	A/G	1
19	A/C	1	G/G	0	A/G	1
51	A/A	0	A/G	1	G/G	0
52	A/A	0	A/G	1	G/G	0
130	A/C	1	G/G	0	A/G	1
276	A/A	0	A/G	1	G/G	0
320	A/C	1	G/G	0	A/G	1
327	A/C	1	G/G	0	A/G	1
330	A/C	1	G/G	0	A/G	1
335	A/C	1	G/G	0	A/G	1
344	A/C	1	G/G	0	A/G	1
353	A/C	1	G/G	0	A/G	1

6.3.4.2 The Anna3 oligonucleotide set

Sequence assemblies were performed using the forward and reverse sequences of the fourteen - seedling bin mapping population including the parental data in contig 7 (section 4.5.1). Although the data from the fourteen bin mapping seedlings were incorporated into contig 7 (section 4.5.1), the SNP distribution pattern shows the presence of mixed genes in the Anna3 PCR amplification products (Figure 6.4). The SNPs in this figure do not show a clear pattern of segregation and an attempt to genotype these markers would lead to unreliable results.

Figure 6.4 also shows a sequence chromatogram of seedling 19 indicating the quality of SNPs 1, 3 and 7. This sequence constitutes one example of the PCR amplicon sequencing confirming the quality of the data used in the sequence assembly. Candidate SNPs 2, 5 and 6 do not appear to segregate from the parental data present in the contig. Although candidate SNPs 1, 4 and 7 show evidence of segregating from the parental data, their inheritance pattern however, makes it impossible to genotype in the bin mapping population. Each of the fourteen seedlings was sequenced using both the forward and reverse oligonucleotides so as to validate sequence quality and as indicated in section 6.3.2 DNA fragments were resolved on an agarose gel before being purified. This suggests co-amplification of mixed genes that could not be resolved by agarose gel electrophoresis. Because of this reason the Anna3 sequencing results of the bin mapping seedlings could not be used as markers in linkage mapping.

Figure 6.4. Assembly of data generated using PCR amplicon sequencing of seedlings of the bin mapping population. Arrows 1 through 8 indicates positions of the candidate SNPs considered in this analysis. The sequence chromatogram shows the quality of the PCR amplicon sequencing data.


1
2
3
4
5
6
7
8
11F
12F
130F
16F
19F
<< 276R
320F
327F
330F
335F
344F
51F
52F
<< AN03H4 M13R
<< AN03H6 M13R
<< AN03H7 M13R
<< AN03H8 M13R
<< G03A2 M13R
<< G03A5 M13R
<< G03B6 M13R
<< G03C8 M13R
<< G03C9 M13R
<< G03D4 M13R
Traces from Contig4
Print
Previous Feature
Next Feature
Confirm
False positive
Don't genotype
Colors
Help
19F
130
140
150
160
170
180
190
200
210
220

6.4 Genotyping SNPs 3, 5 and 6 using the SNaPshot™ method

6.4.1 SNaPshot™ extension oligonucleotides

SNaPshot™ extension oligonucleotides were designed on contig 1 (Figure 4.9) for candidate SNPs 3, 5 and 6 (Tables 24 and 25). The extension oligonucleotides were designed to contain 18 to 25 nucleotides complimentary to and terminating at one base upstream of the candidate SNP base. Addition of non-specific 'ACTG' bases at the 5′ end of the complimentary stretch of nucleotides was used to make up the required lengths of the oligonucleotides. The extension oligonucleotides and their sequences are shown in Table 26.

Table 26. Design of the SNaPshot™ extension oligonucleotides

SNP ID	Oligonucleotide sequence
SNP3M_GD1_41	$(ACTG)_4$TCGAAGTTTAGCAGTTTTCTTGCTA
SNP5R_GD1_48	$(ACTG)_5$AC)AATTTGAAGTATTGGTTGGAAAGAGG
SNP6R_GD1_32	$(ACTG)_4$ATTGGTTTGGTAAGGG

6.4.2 Preparation of SNaPshot™ extension products

The GD1 oligonucleotide set was used to amplify DNA fragments containing SNPs 3, 5 and 6 from the genomic DNA of fourteen bin mapping seedlings (section 6.3.2). These included samples 11, 12, 16, 19, 51, 52, 130, 276, 320, 327, 330, 335, 344 and 353. The

amplification products were resolved using agarose gel electrophoresis and DNA fragments migrating at ~462 bp as compared to the DNA size standard were excised out of the agarose gel. These DNA fragments were purified (section 2.6.1.2) and analysed by agarose gel electrophoresis, results are shown in Figure 6.5. Each of the fourteen lanes in Figure 6.5 contain a single un-degraded DNA fragment band that is purified from both oligonucleotide dimers and smears and could be used successfully as templates for SNaPshot™ assays.

Figure 6.5. Agarose gel purified DNA fragments produced by PCR amplification of genomic DNA using the GD1 oligonucleotide set. Lane M shows the profile of the pTZ/*Hinf* I DNA size standard and lanes 1 through 14 show the purified DNA fragments from seedlings 11, 12, 16, 19, 51, 52, 130, 276, 320, 327, 330, 335, 344 and 353.

6.4.3 SNaPshot™ and SNP-specific controls

SNaPshot™ Multiplex control template and oligonucleotides were used to set up the positive control for fragment size and colour schemes. Clones with inserts whose DNA sequences contained candidate SNPs 3, 5 and 6 from the re-sequenced dataset (chapter 4) were used as genotype controls. These represented templates with known SNPs and were

used here to assess the sensitivity and accuracy of the SNaPshot™ assays. Figure 6.6 shows the multiple sequence alignment containing the SNP positions under analysis, Table 27 and Figure 6.6 show the SNP genotypes for the control clones as identified through sequencing. Figure 6.6 gives a clear indication of SNPs at the positions indicated by arrows labelled SNP3, SNP5 and SNP6 and as such were accepted as satisfactory controls for the SNaPshot™ genotyping assays. The three SNPs indicate the presence of two haplotypes and thus the genotyping experiments are also internally controlled.

Table 27. Genotypes of the clones used as positive controls as detected by multiple sequence alignment

Clone	SNP3	SNP5	SNP6
AnRGA148	CC	GG	AA
AnRGA278	AA	AA	GG
AnRGA319	AA	AA	GG
AnRGA324	AA	AA	GG
AnRGA325	CC	GG	AA

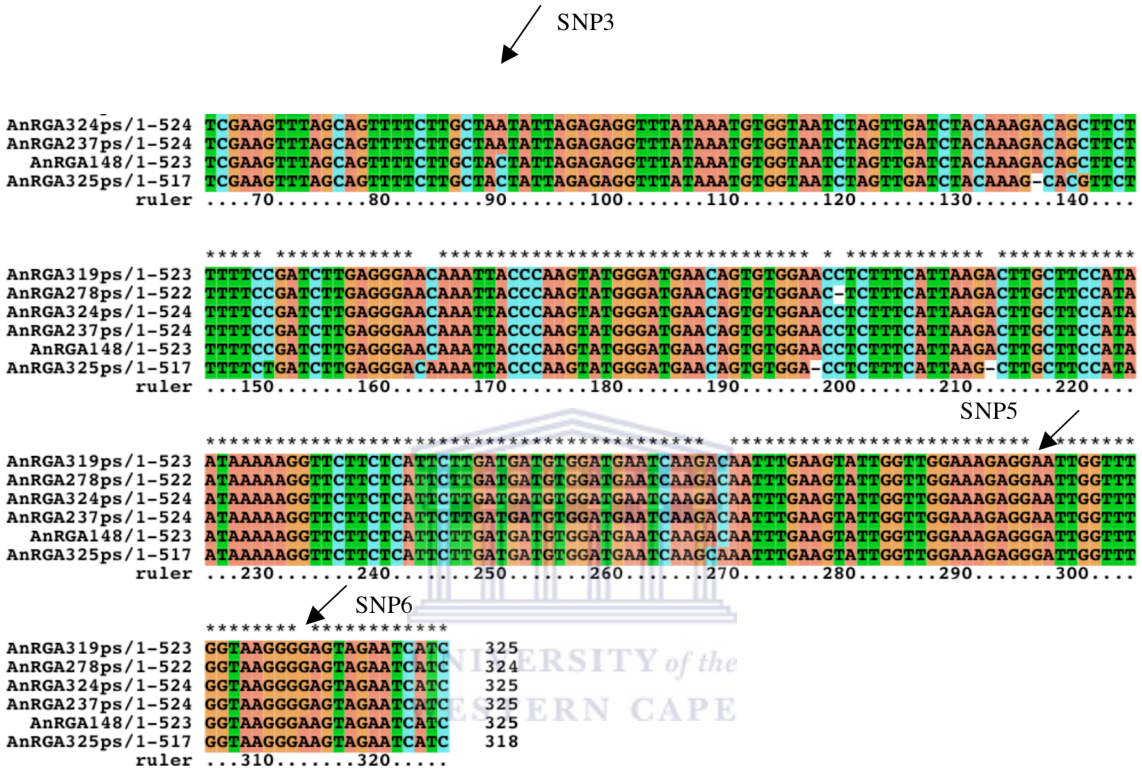


Figure 6.6. Multiple alignment of sequences from clones used as positive controls in evaluating the accuracy of the SNaPshot assays. The positions of the SNPs are indicated in the figure as SNP3 (A/C), 5 (A/G) and 6 (G/A) respectively.

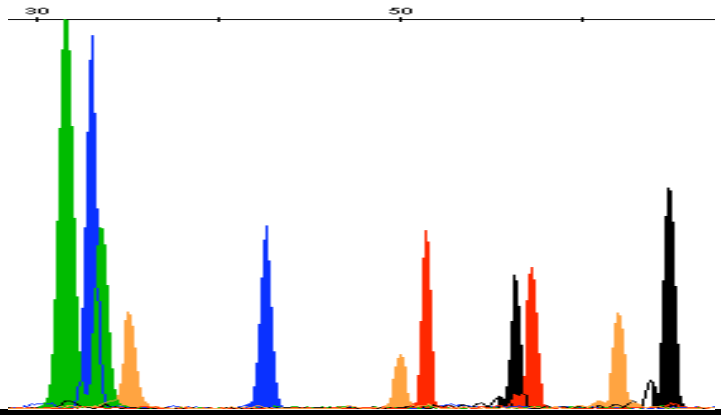
6.4.4 SNaPshot™ extension PCR and product purification

SNaPshot™ reaction mixes, oligonucleotide extension and post-extension purification of products were performed using the protocols described in section 2.14.3. Purified template DNA (Figure 6.5) and 0.2 µM extension oligonucleotides (Table 26) were used to set up the SNaPshot™ assays.

6.4.5 SNaPshot™ extension product fragment analysis

A total of 2.0 µl of purified SNaPshot™ extension products were mixed with 9 µl of Hi-Di formamide (Applied Biosystems) and 0.4 µl of GeneScan LIZ120 size standard (Applied Biosystems). These were denatured at 95ºC for 2 minutes and analysed using the ABI 3130xl Genetic Analyzer (Applied Biosystems) as described in section 2.14.4. Tables 28 (control clones), 29, 30 and 31 show results for SNaPshot™ extension oligonucleotides SNP3M_GD1_41, SNP5R_GD1_48 and SNP6R_GD1_32 on GD1 amplified DNA fragments respectively. The expected fragment sizes were offset from the manufacturer's observed sizes (Figure 6.7). The manufacturer's sizes were detected using POP-4 polymer and results shown here were detected by the ABI 3130xl Genetic Analyzer/ POP-7 polymer system. However, these fragment sizes were reproducible and as such SNP genotypes could be scored in single-oligonucleotide assay reactions.

The 'A' genotypes were detected as green (dR6G) peaks, 'C' as black (dTAMRA™) and 'G' as blue (dR110) peaks. These colour schemes for the associated SNP genotypes were confirmed by the sequences of the control clones (Figure 6.6).



Observed using POP-7™	32	33.15C 33.78A	42	51	56G 57T	64
Primer	20A	28G/A	36G	44T	52C/T	60C
Relative product size (POP-4™)	24	29	39	46	53	62
Size difference	8	4	3	5	4	2

Figure 6.7. Fragment analysis results of the SNaPshot™ assay positive control oligonucleotides analyzed using GeneScan LIZ120 size standard. The influence of the dye on the mobility shift of the DNA fragments altered the observed from the manufacturer's expected relative product sizes.

The observed and expected sizes (Figure 6.7) differed for all the eight mixed fragments and all fluorescent dyes in the manufacturer's positive control. The expected fragment sizes were obtained using POP-4™ (Applied Biosystems) and these were altered by analysis using POP-7™. However, these sizes were consistent in subsequent analyses and could be used as reliable fragment size controls for subsequent SNaPshot™ assays.

Having established the consistency of the fragment sizing ranges using POP-7™, sequenced control clones (Figure 6.6) were used as genotype controls (Table 27) to guide the genotyping assays. Figure 6.8 shows results from two of these clones representing the two heterozygous alleles detected in Golden Delicious cluster 11 (Figure 3.5). These results defined the dTAMRA™ background detected in most SNP3 peaks and thus provided the basis for accepting results of the bin mapping seedling assays.

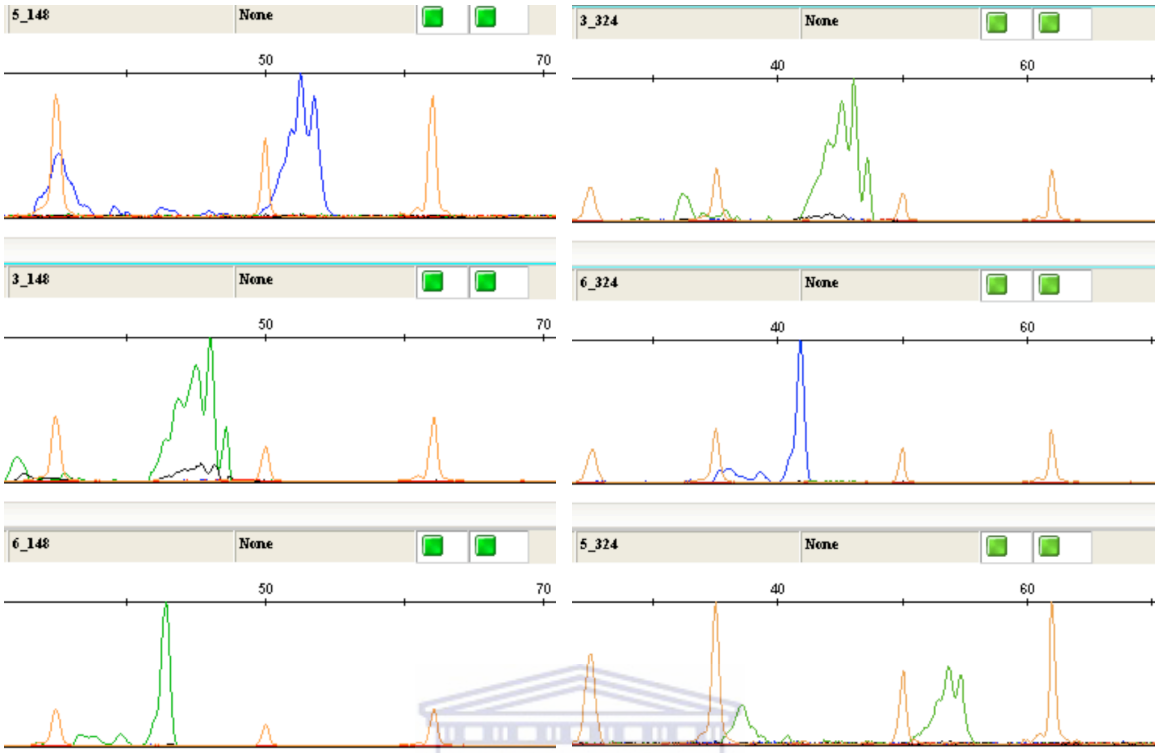


Figure 6.8. SNaPshot™ assay results for the positive control clones representing the two alleles identified in Golden Delicious cluster 11 of Figure 3.5. Each pane shows fragment peaks for SNPs 3, 5 and 6 for controls AnRGA148 and AnRGA324.

Genotypes for all the five control clones detected by the SNaPshot™ assays are presented in Table 28. These results confirm the specificity of these assays as shown by the strong correlation with amplicon sequencing results. Subsequently this process was used as a basis to validate genotypes of the fourteen bin mapping seedlings. In addition to the validation using results of the control clones, genotypes for the bin mapping seedlings were also internally controlled as shown by results presented in Tables 29, 30 and 31 that show SNP3, 5 and 6 genotypes for the fourteen bin mapping seedlings.

Table 28. SNaPshot™ results for the positive control clones for SNPs 3, 5 and 6.

SNP	Dye	Sample	Size	SNP genotype
SNP3	dTAMRA™	AnRGA148	46.36	**C**
	dR6G	AnRGA278	46.01	**A**
	dR6G	AnRGA319	46.02	**A**
	dR6G	AnRGA324	46.13	**A**
	dTAMRA™	AnRGA325	46.51	**C**
SNP5	dR6G	AnRGA148	42.81	**A**
	dR110	AnRGA278	41.83	**G**
	dR110	AnRGA319	41.87	**G**
	dR110	AnRGA324	41.73	**G**
	dR6G	AnRGA325	42.86	**A**
SNP6	dR110	AnRGA148	53.45	**G**
	dR6G	AnRGA278	53.57	**A**
	dR6G	AnRGA319	53.69	**A**
	dR6G	AnRGA324	53.66	**A**
	dR110	AnRGA325	53.54	**G**

Multiplexing the three extension oligonucleotides gave non-specific results in which known genotypes of the positive control clones could not be reproduced. The A genotypes gave very strong peaks of the dR6G fluorescent dye at all the assayed injection volumes, while other genotypes were suppressed. This problem could not be corrected since ddATP concentrations in the commercial SNaPshot™ multiplex ready mix reagent are not possible to alter. The problem of altered fragment sizes and the A genotype peak structure also limits the usefulness of this method for multiplexing and although this could be corrected through the use of POP-4™ polymer, such an exercise however, was not useful given that alternating between POP-7™ and POP-4™ polymer on the ABI 3130xl Genetic Analyser becomes an expensive way of operating the equipment.

Table 29. SNaPshot™ results for SNP3 on the bin mapping population and the positive control clones. Genotype assignments are supported by results of the positive control clones as described for Table 28.

Dye	Sample	Size	Genotype
dR6G	Anna	46.46	**A**
dTAMRA™	Anna	46.84	**C**
dR6G	Golden Delicious	46.57	**A**
dR6G	11	46.71	**A**
dTAMRA™		46.12	**C**
dR6G	12	46.55	**A**
dTAMRA™		46.65	**C**
dR6G	16	46.68	**A**
dTAMRA™		47.04	**C**
dR6G	19	46.56	**A**
dTAMRA™		46.92	**C**
dR6G	51	46.62	**A**
dR6G	52	46.56	**A**
dR6G	130	46.68	**A**
dTAMRA™		46.92	**C**
dR6G	276	46.66	**A**
dR6G	320	46.49	**A**
dTAMRA™		45.66	**C**
dR6G	327	46.59	**A**
dTAMRA™		45.88	**C**
dR6G	330	46.46	**A**
dTAMRA™		46.94	**C**
dR6G	335	46.59	**A**
dTAMRA™		46.95	**C**
dR6G	344	46.36	**A**
dTAMRA™		46.59	**C**
dR6G	353	46.56	**A**
dTAMRA™		46.92	**C**

Table 30. SNaPshot™ results for SNP5 on the fourteen - seedling bin mapping population. Genotype assignments are supported by results of the positive control clones (Figure 6.6)

Dye	Sample	Size	Genotype
dR110	Anna	53.74	G
dR6G	Anna	54.38	A
dR110	Golden Delicious	53.88	G
dR110	11	53.04	G
dR110	12	53.13	G
dR110	16	53.25	G
dR110	19	53.19	G
dR110	51	53.16	G
dR6G		54.14	A
dR110	52	53.87	G
dR6G		53.98	A
dR110	130	53.84	G
dR110	276	53.08	G
dR6G		54.07	A
dR110	320	53.14	G
dR110	327	53.11	G
dR110	330	53.08	G
dR110	335	53.07	G
dR110	344	53.44	G
dR110	353	53.33	G

Table 31. SNaPshot™ results for SNP6 on the fourteen - seedling bin mapping population. Genotype assignments are supported by results of the positive control clones (Figure 6.6)

Dye	Sample	Size	Genotype
dR110	Anna	41.36	**G**
dR6G	Anna	42.4	**A**
dR110	Golden Delicious	41.94	**G**
dR110	11	42.24	**G**
dR6G		43.24	**A**
dR110	12	42.19	**G**
dR6G		43.3	**A**
dR110	16	42.18	**G**
dR6G		43.19	**A**
dR110	19	41.93	**G**
dR6G		43.3	**A**
dR110	51	42.11	**G**
dR110	52	42.19	**G**
dR110	130	42.3	**G**
dR6G		43.4	**A**
dR110	276	42.38	**G**
dR110	320	42.23	**G**
dR6G		43.23	**A**
dR110	327	42.24	**G**
dR6G		43.24	**A**
dR110	330	42.18	**G**
dR6G		43.29	**A**
dR110	335	42.26	**G**
dR6G		43.26	**A**
dR110	344	41.74	**G**
dR6G		42.75	**A**
dR110	353	41.84	**G**
dR6G		42.86	**A**

SNaPshot™ assay results for SNPs 3 and 6 in Tables 29 and 31 respectively show an identical segregation pattern in which polymorphisms in SNP3 also correlated with SNP6. SNP5 genotypes (Table 30) are in reverse phase of SNPs 3 and 6, which is expected given that the alleles represent haplotypes identified from one gene and thus specify the same location on the genetic linkage map. This is expected given that all three SNPs were identified in a 462 bp DNA sequence from genes occurring on the same cluster (Figure 3.5). These results also confirm the specificity and sensitivity of the SNaPshot™ assays. The data were then used to position RGA cluster 11 (Figure 3.5) on the Anna x Golden Delicious genetic linkage map.

6.4.6 Linkage analysis and genetic mapping

The alleles of the three SNPs for each of the fourteen seedlings were re-coded to either '+' representing heterozygotes (AC or AG) that showed clear inheritance of SNPs from both parent and '-' representing homozygotes (AA or GG) in which the inheritance of either SNP could not be traced to an individual parent. The joint genotypes were obtained as '----++-+------', '++++--+-++++++' and '----++-+------' for SNPs 3, 5 and 6 respectively where 3 and 6 are identical and 5 is a repulsion of the first two. The joint genotypes for each SNP were then compared to the genotype of the bin mapping set present in the Anna x Golden Delicious framework map (van Dyk PhD thesis, UWC, 2008). The genotype codes for SNPs 3, 5 and 6 are identical to those for microsatellite marker CH03a04 (83.9 cM) on linkage group 5 (LG5) of the Anna x Golden Delicious genetic linkage map (in the Anna parental map) as illustrated in Table 32 and Figure 6.9.

The bin around CH03a04 (83.9 cM) spans a 4.8 cM region to CH02b12 (79.1 cM) and it is estimated that these SNPs map within that region.

Table 32. Microsatellite markers mapped on *Malus x domestica* (Borkh.) cv Anna linkage group 5 showing the position of the RGAcl1 marker. Missing genotypes are represented by '0s' in the table and cM represents the genetic positions in centiMorgans for the microsatellite markers in the linkage group.

Marker	cM	11	12	16	19	51	52	130	276	320	327	330	335	344	353
Hi22f12	0.0	+	-	+	-	-	-	-	+	-	+	-	-	-	-
SA-A340	11.8	+	0	+	0	-	0	0	0	-	-	-	-	0	-
CH03a09	20.9	+	-	+	-	-	+	+	-	-	-	-	-	-	-
CH05e06	26.3	+	+	+	**0**	-	+	+	0	0	+	-	-	-	0
SA-A401	41.4	-	+	+	-	-	+	+	+	-	-	-	-	-	-
CH03e03	50.4	-	+	+	-	-	+	+	+	-	-	-	-	-	-
Hi04d02	52.3	-	+	+	-	-	+	+	-	-	-	-	-	-	-
CH05f06	63.2	-	+	0	-	-	+	+	-	-	-	-	-	-	-
CH05d04	67.9	-	+	-	-	+	+	0	-	-	-	-	-	-	0
CH02b12	79.1	-	-	0	-	+	+	-	-	-	-	-	-	-	-
RGAcl11		-	-	-	-	+	+	-	+	-	-	-	-	-	-
CH03a04	83.9	-	-	-	-	+	+	-	+	-	-	-	-	-	-

The exact position of this marker cannot be ascertained although the joint genotypes for SNPs 3, 5 and 6 (Table 32) suggest that RGAcl11 has binary codes identical to microsatellite marker CH03a04 (83.9 cM). Marker RGAcl11 lies within the 4.8 cM bin defined by the recombination breakpoints at microsatellite markers CH02b12 (79.1) and CH03a04 (83.9).

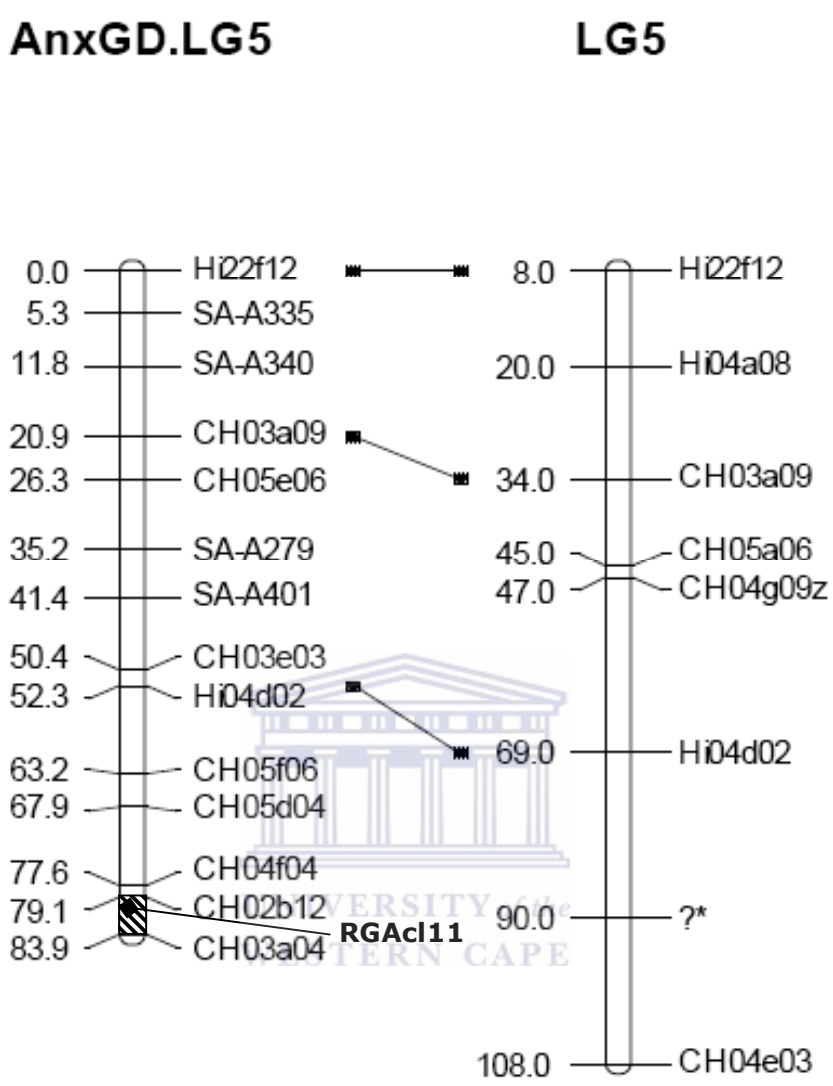


Figure 6.9. Genetic position of RGAcl11 on the Anna x Golden Delicious linkage map. RGAcl11 represents **RGA cl**uster **11** (Figure 3.5) and indicates the location specified by the joint genotypes of SNPs 3, 5 and 6.

6.5 DISCUSSION

As described in section 6.1, a number of methods have been used to convert RGAs into molecular markers. However, these methods either target one RGA at a time through the use of either specific or degenerate oligonucleotides for targeted gene amplification which could be coupled to restriction endonucleases to produce either CAPS or SSCPs. Given the gene copy number in the NBS-LRR family, these methods are limited in their capacity for producing reliable SNP data useful in genetic mapping of RGAs. In this chapter, two methods, SNP detection by sequencing and SNaPshot™ multiplex ready mix reagent kit (Applied Biosystems) assays were analysed as potential high throughput mapping techniques.

A set of oligonucleotides designed from multiple sequence alignments of genes from selected clusters (Figure 3.5) were used to amplify DNA fragments of the NBS domain from the Anna x Golden Delicious bin mapping population. SNP identification by sequencing (section 4.5.1) was used to genotype the Anna x Golden Delicious bin mapping population to confirm their segregation and to position them on an existing genetic linkage map.

This technique was useful in identifying candidate SNPs in the bin mapping seedlings and the two methods produced identical results for the GD1 oligonucleotide set. Sequences generated from the Anna3 oligonucleotide set showed a characteristic mixed gene contig in which no clear segregation pattern could be deciphered. This is despite the fact that the quality of chromatograms from these sequences was satisfactory with high

signal to noise ratios (Figure 6.4). The cluster targeted in this analysis appears to be made up of more than one gene each of which has its own alleles. This result complicates SNP identification and mapping through PCR amplicon sequencing and it also confirms that RGA clusters are made up of genes physical proximity to each other and sharing the same locus. This demonstrates the need for high throughput PCR amplicon sequencing such a those offered by 454 Life Science technologies to resolve these clusters and analyse SNPs in individual genes making up such clusters.

Genotyping experiments using the SNaPshot™ method as a way of confirming segregation of SNPs identified through sequencing/ sequence assemblies seems more robust as compared to PCR amplicon sequencing using Sanger sequencing. The sequencing process suppresses heterozygote base calls depending on the threshold of peak heights set in the base-caller for the alternative nucleotide per site. Lowering this threshold to allow detection of low frequency SNPs might lead to base - calling of background noise. This limits the capacity of Sanger sequencing to detect all SNPs in any given gene. High throughput amplicon sequencing, however can be used as an alternative given its ability to target individual molecules in a PCR amplification product as templates for clonal amplification. Of the two platforms tested in this chapter, the SNaPshot™ method appears to be more sensitive. The use of known RGA clones with known SNP genotypes as positive controls proved very valuable in determining sample genotypes from SNaPshot™ assays.

The average SNP frequency was estimated at one in every ± 66 nucleotides and the three SNPs identified from the GD1 oligonucleotide set define two haplotypes of a gene in which SNP5 is a repulsion of 3 and 6. These were positioned on linkage group 5 (LG5) in a 4.8 cM bin between microsatellite markers CH02b12 and CH03a04 on the *Malus x domestica* (Borkh.) Anna x Golden Delicious genetic linkage map. LG5 was shown to contain the microsatellite marker CH04e03 on the Discovery x TN10-8 genetic linkage map and this marker has been linked to partial resistance to apple scab (Calenge *et al.*, 2004). This linkage group also contains a microsatellite marker, CH03e03, which is regarded as a diagnostic marker for resistance to fireblight and an RGA marker NBS2R11 (Calenge *et al.*, 2005). Results of Figure 6.9 (marker RGAcl11) proves that sequencing and sequence assemblies as a process can be used to identify SNPs that can be converted into markers and can be mapped on existing genetic linkage maps. The combination of SNP identification by sequencing and genotyping using SNaPshot™ assays constitutes a useful strategy that can be used to map all RGA clusters identified in chapter 3 (Figure 3.5) that do contain SNPs. However, problems associated with SNP identification in clusters with mixed genes (Anna3 oligonucleotide set) strengthen the need for high throughput PCR amplicon sequencing (454 Life Sciences) as a strategy to resolve SNPs in such clusters.

CHAPTER 7

HIGH THROUGHPUT SNP IDENTIFICATION USING PCR AMPLICON SEQUENCING

CONTENTS

CHAPTER 7: HIGH THROUGHPUT SNP IDENTIFICATION USING PCR AMPLICON SEQUENCING

7.1 INTRODUCTION

It has been established in published literature and partly in the preceding chapters that RGAs form a superfamily of genes with members that are physically distributed as clusters in the plant genome. Individual clusters are made up of either one gene with multiple alleles or a mosaic of genes that specify resistance to various pathogens (Michelmore and Meyers, 1998). PCR amplification products generated from cluster-based oligonucleotides are made up of a complex of genes and Sanger sequencing of these products produces a consensus sequence of all the genes in the amplified cluster. This gives a sequence read with numerous ambiguous base calls and super-imposed bases that could be mistaken for SNPs. As shown for Anna3 - primed amplification products (chapter 6), sequence assemblies of data from such clusters shows candidate SNPs that cannot be genotyped. This demonstrates the limits of SNP identification through sequencing of PCR amplification products using cluster-specific oligonucleotides.

This chapter describes high throughput sequencing of PCR amplified RGAs using degenerate oligonucleotides described in chapter 3. This data is used to identify SNPs through sequence assemblies that incorporate sequence data from plants used in the *Malus x domestica* (Borkh.) Anna x Golden Delicious bin mapping population. All

identified segregating SNPs will then be mapped using the *Malus x domestica* (Borkh.) Anna x Golden Delicious genetic linkage map (chapter 6).

7.2 Production of RGAs flanked by A and B sequencing tags

Fusion degenerate oligonucleotides NBSFA and NBSR-2B (section 4.4.4.2) were used to amplify RGAs from *Malus x domestica* (Borkh.) cvs. Anna, Golden Delicious and fourteen of the seedlings from the bin mapping population (11, 12, 16, 19, 51, 52, 130, 276, 320, 327, 330, 335, 344 and 353). PCR amplification was performed as described in section 4.4.4.2. The amplification products were resolved using agarose gel electrophoresis (section 2.12.13) and then purified from the gel (section 2.6.1.2). A 3 µl aliquot from some of the purified amplification products was then analysed through agarose gel electrophoresis (Figure 7.1).

The expected product size from PCR amplification of RGAs using the NBSFA and NBSR-2B oligonucleotide set is ±550 (section 4.4.4.2). The purification process effectively removed residual oligonucleotides and thus ensuring the recovery of amplification products with minimal contamination from small DNA fragments. The arrowhead in Figure 7.1 indicates the expected product sizes as compared to the pTZ/*Hinf* I DNA size standard.

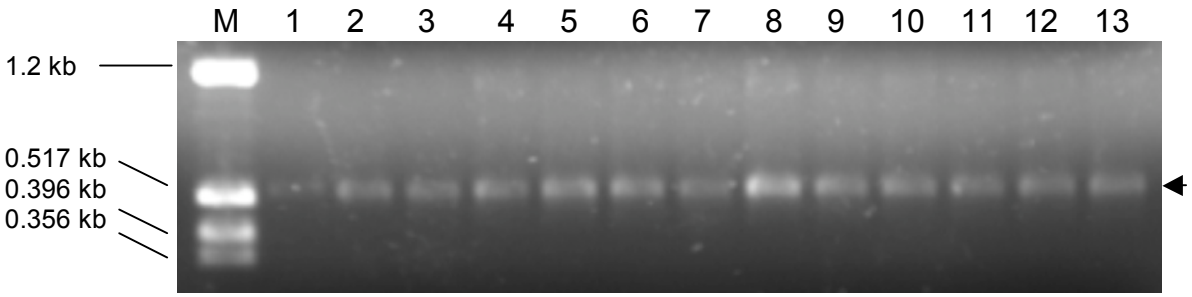


Figure 7.1. PCR amplification products of the bin mapping seedlings amplified using fusion oligonucleotides NBSFA and NBSR-1B. Lanes 1 through 13 show the PCR amplification products from the bin mapping set and lane M shows the profile of the pTZ/ *Hinf* I DNA size standard.

7.3 Sequencing of candidate RGAs using the GS FLX System

A total of 136370 sequence reads were generated with the statistical mode for the read length estimated in the range 252 to 261 nucleotides. The read length distribution was in the range 40 to 285 nucleotides although the data tended to form a peak in the ranges 200 to 276 nucleotides. Sequence length distribution was plotted on a 2D graph (x-axis) against frequency on (y-axis) and two such plots are shown in Figure 7.2. These plots represent length distributions in the two parental samples Anna and Golden Delicious Table 32.

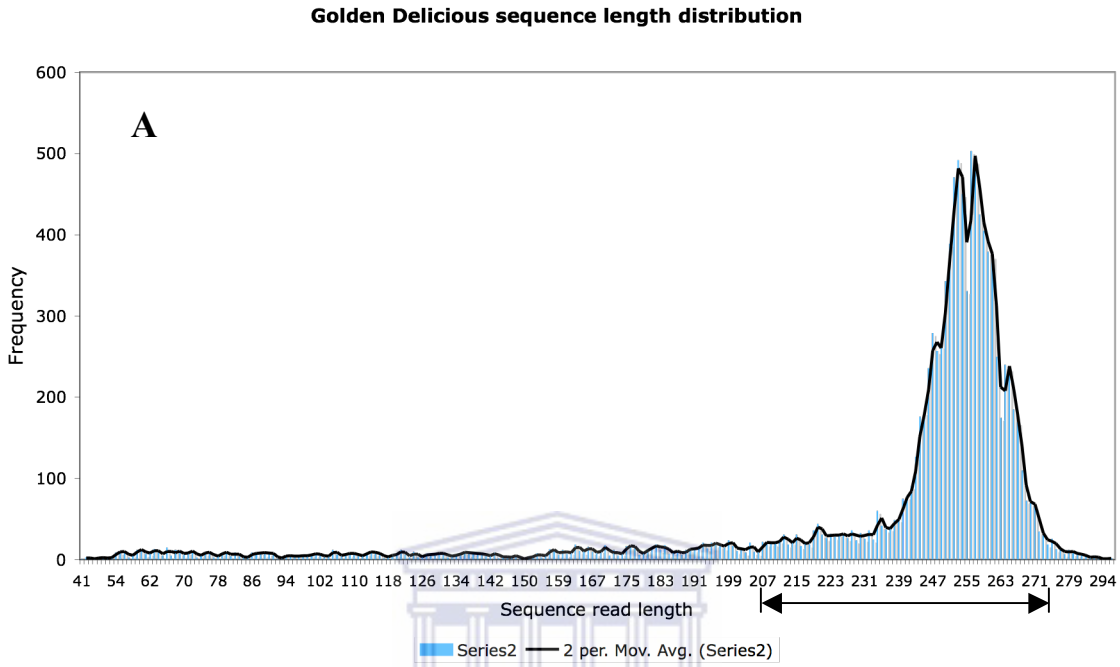
Table 32. Data generated from the 454 sequencing experiment. Samples in the table are arranged in descending order of number of sequences obtained per sample; the percentages of sequence reads that are 100 nucleotides or shorter are provided in brackets.

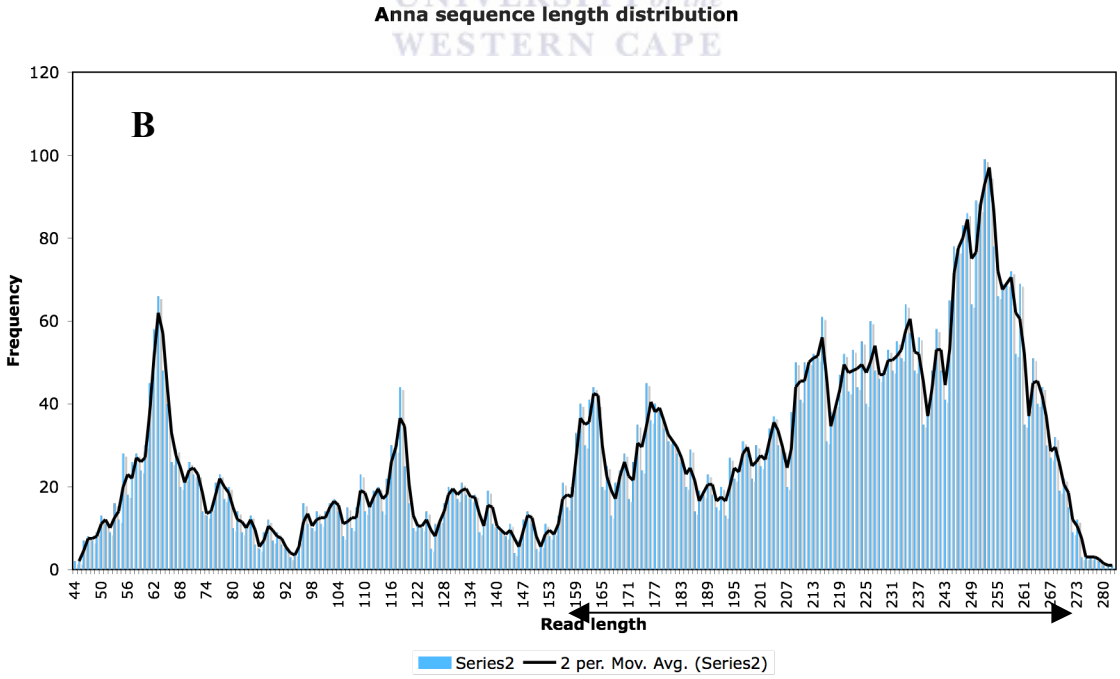
Samples	Total sequences	≤ 100 bases
327	4782	717 (15.0%)
11	4975	571 (12.0%)
12	4989	587 (12.0%)
330	5273	713 (14.0%)
335	5949	756 (13.0%)
Anna	6617	991 (15.0%)
130	7646	663 (9.00%)
320	7797	838 (11.0%)
344	8689	927 (11.0%)
52	9777	667 (7.00%)
276	10311	540 (5.00%)
51	10387	709 (7.00%)
Golden Delicious	10918	351 (3.0%)
19	11465	428 (4.00%)
353	11772	647 (6.00%)
16	15023	643 (4.00%)

Sequences were prepared for analyses by clipping off the priming region (about 26 nucleotides) this reduced the non-specific bias ascribable to the degenerate oligonucleotides used to amplify these genes. Sequence reads that were 74 nucleotides or shorter after clipping off the priming region were then selected and analysed using stringent BLAST homology searches with the 'Expect threshold' parameter set to 1.0 and 'word size' reduced to 16 from the default 28. Significant matches to the R-gene NBS domain were observed down to 35 nucleotides (or 61 before end-clipping). However, sequence reads that are 100 nucleotides and shorter also contained artefacts such as sequencing adaptors and or forward oligonucleotides. Sequencing was designed to generate an average of 250 nucleotides from the reverse oligonucleotide and the 454 sequencing software is designed to filter off the adaptor sequence.

BLAST homology searches for sequence reads 101 nucleotides and longer revealed that this dataset was composed exclusively of the 3' fragments of the NBS domain of RGAs. Based on these preliminary analyses, the sequence dataset was accepted as 3' fragments of apple RGAs and thus used for downstream analyses. Three of the samples in this 16 sample sequencing experiment gave data with individual numbers not exceeding 5000 reads and as shown in Table 32 they are samples 11, 12 and 327. However, analysis of the length distribution demonstrated that all data conformed to the same typical format in which lengths span the range ~40 to ~285 nucleotides with a peak in the region 240 to 270 nucleotides.

Figure 7.2. Frequency distribution plots for (A) Golden Delicious and (B) Anna samples showing sequence read length plotted against number of times it appears in the dataset. The x-axis represents the read length ranges and the y-axis represents the frequency of occurrence in the datasets. The trendline is a moving average that considers two points at a time.





The area demarcated by the double arrowhead in Figures 7.2 A and B makes up about 87% and 72% of the data in those two datasets respectively. These sequences showed significant homology to the 3' end of the NBS domain of NBS-LRR type R genes. In the Golden Delicious sequence data, sequence reads with lengths in the range 230 – to – 270 nucleotides constituted about 80% of the total dataset compared to the same range in Anna which represents about 35% of sequence reads in that dataset.

The 454 FLX sequencing system uses a hundred cycles of each nucleotide, meaning a hypothetical sequence with no base repeats gives a length of 100 bases per full run. The difference in sequence lengths above the hypothetical 100 nucleotides becomes a function of the percentage of the individual sequence reads represented by homopolymeric regions. Thus sequence reads that are shorter than 100 nucleotides or significantly longer than the statistical mode per dataset represent potential sequencing artefacts. These two ranges were assessed using BLAST homology searches against the NCBI non-redundant database to determine the average percentage error rate. Only sequences shorter than 45 nucleotides did not produce Hits to NBS-LRR resistance genes and as such these were excluded from the analysis. Very few reads contained the forward oligonucleotide and or the 454 sequencing tag although sequencing was performed from the 3' end/ reverse oligonucleotide, these were also disregarded from the analysis. All assessed reads above the statistical mode produced significant hits to the NBS-LRR resistance genes. Based on these assessments, the remaining dataset was thus accepted as ideal for SNP analysis and subsequent mapping onto the existing Anna x Golden Delicious genetic linkage map.

7.4 Sequence assembly to detect SNPs in the RGA data

7.4.1 Anna and Golden Delicious sequence assemblies

Anna and Golden Delicious datasets were initially assembled separately in CodonCode aligner using default parameters. The resultant contigs with aligned sequences were used to clip off priming regions and to discard all reads containing the forward oligonucleotide an/ 5' sequencing tag. The cleaned datasets were then assembled with only the minimum percentage identity adjusted up to 95%. All contigs from this second assembly were analysed individually to eliminate gaps, individual reads showing exaggerated random mutations and whole contigs either showing the forward oligonucleotide/ sequencing tag and those shorter than average and not producing significant homology to NBS-LRR genes. Significant homology to NBS-LRR genes was assessed through BLAST searches against the NCBI non-redundant database.

The processed assemblies were then reassembled with minimum percentage identity and bandwidth (maximum gap size allowed in alignment) adjusted to 96% and 15 respectively. The resultant assemblies were then re-assembled separately at a minimum percentage identity of 97% to assess for any possible re-arrangements. The results of these two parallel sequence assemblies are shown in Table 33.

Table 33. Results of the Anna and Golden Delicious individual and combined sequence assemblies. Table 33A reports statistics on the two individual assemblies and 33B reports statistics for the combined assembly.

A

Statistics for the sequence assemblies for individual samples

Sample	Contigs			Unassembled
	Total	2 reads	3 reads	
Anna	211	36	28	976
GD	234	42	26	

B

Distribution of the 329 contigs produced in the combined sequence assembly

Sample	Contigs			Unassembled
	Total	2 reads	3 reads	
Anna only	77	26	17	790
GD only	97	33	20	
Combined	146	5	4	

Subsets of both Anna and Golden Delicious individual contigs merged in the combined sequence assembly to produce 149 contigs composed of sequence reads from both samples. A total of 186 sequences previously unassembled in the individual sequence assemblies were incorporated into contigs with either two or more reads thus reducing the unassembled set 790. Further assessment of the unassembled set showed that although 572 of the unassembled were shorter than 100 nucleotides, they were significantly homologous to NBS-LRR genes as determined by BLAST searches against the NCBI non-redundant database.

7.4.2 SNP identification and segregation analysis

A total of 210 polymorphic SNPs were identified in 89 of the 146 contigs produced from the Anna and Golden Delicious sequence assembly. The distribution of SNPs per contig in this assembly ranged from as low as one to as high as five in an average length of 220 nucleotides. The majority of these SNPs however, were heterozygous in both Anna and Golden Delicious and thus were disregarded from this analysis since they could not be mapped using the bin mapping approach. Only one SNP per contig was selected for segregation analysis and genetic linkage mapping. This reduced the number of SNPs that were analysed to a total of 24.

Contigs in the combined sequence assembly (section 7.4.1) were renamed using the 'AnGD' prefix and the contig number was left unaltered. Contigs that did not contain SNPs were eliminated from the downstream analyses of segregation. Each of the 14 bin mapping progeny were then processed individually to remove priming regions (section

7.4.1) before being assembled one sample at a time on top of the Anna and Golden Delicious SNP-containing dataset. Figure 7.3 A and B shows contig views for one of the candidate SNPs identified through using this approach.

Figure 7.3A is a highly trimmed version of the parent contig containing four reads per seedling for the purposes of displaying the image on an A4 paper. Figure 7.3B is a portion of the untrimmed contig showing data from only three of the fourteen seedlings. Figure 7.3B is inserted here to show the distribution of the alleles as a fraction of the total assembled data per seedling (the parent contig before trimming contains a total of 549). Although the assembly is characterised by sparsely populated random mutations, the consistent single base mutation indicated by the red arrow in Figure 7.3A was accepted as evidence of a polymorphic SNP and thus treated as such in subsequent analyses. These were thus scored as SNPs in all contigs where they occurred provided the same phenomenon was detected in the Anna and Golden Delicious sequence reads. The presence of both mutating nucleotides (e.g. G and T in Figure 7.3) in a seedling was accepted as evidence of heterozygosity for that SNP and the occurrence of either of the two forms alone (e.g. either G or T alone in Figure 7.3) was evidence of homozygosity. This scheme was used to assess and score the pattern of segregation for all candidate SNPs identified in the Anna and Golden Delicious assembly.

Figure 7.3. Sequence assembly contig view showing the segregation pattern for SNP-017. This figure shows the genotypes of 7 out of the 14 bin mapping progeny, the sequence representation has been reduced to four for each of the displayed progeny. The 'red arrow' indicates the position and genotype of SNP_017. The sequence identities are labelled with the prefix BS (bin mapping seedling) followed by the seedling ID (e.g. BS_327_ for bin mapping seedling 327), the numbers immediately following the seedling IDs were generated from by the sequencing software and are not significant in this analysis.

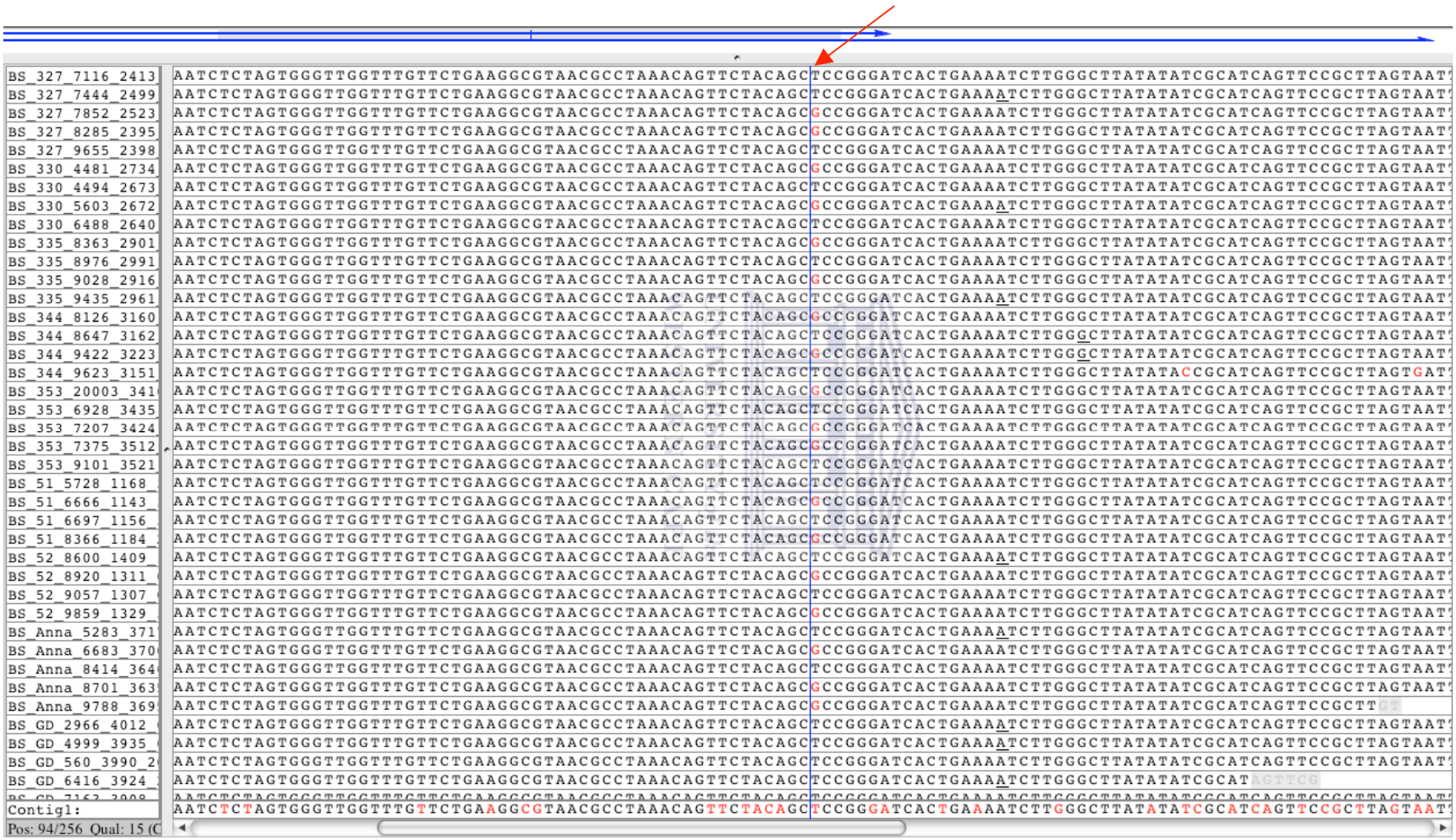

BS_327_7116_2413
BS_327_7444_2499
BS_327_7852_2523
BS_327_8285_2395
BS_327_9655_2398
BS_330_4481_2734
BS_330_4494_2673
BS_330_5603_2672
BS_330_6488_2640
BS_335_8363_2901
BS_335_8976_2991
BS_335_9028_2916
BS_335_9435_2961
BS_344_8126_3160
BS_344_8647_3162
BS_344_9422_3223
BS_344_9623_3151
BS_353_20003_341
BS_353_6928_3435
BS_353_7207_3424
BS_353_7375_3512
BS_353_9101_3521
BS_51_5728_1168_
BS_51_6666_1143_
BS_51_6697_1156_
BS_51_8366_1184_
BS_52_8600_1409_
BS_52_8920_1311_
BS_52_9057_1307_
BS_52_9859_1329_
BS_Anna_5283_371
BS_Anna_6683_370
BS_Anna_8414_364
BS_Anna_8701_363
BS_Anna_9788_369
BS_GD_2966_4012_
BS_GD_4999_3935_
BS_GD_560_3990_2
BS_GD_6416_3924_
Contig1:
Pos: 94/256 Qual: 15 (C

Figure 7.3B. A portion of the untrimmed contig view showing sequence reads from three of the fourteen seedlings in the contig. The underlined nucleotides were corrected to match the consensus (labelled as contig 1), this indicates 454 sequencing artefacts associated with homopolymer repeats.

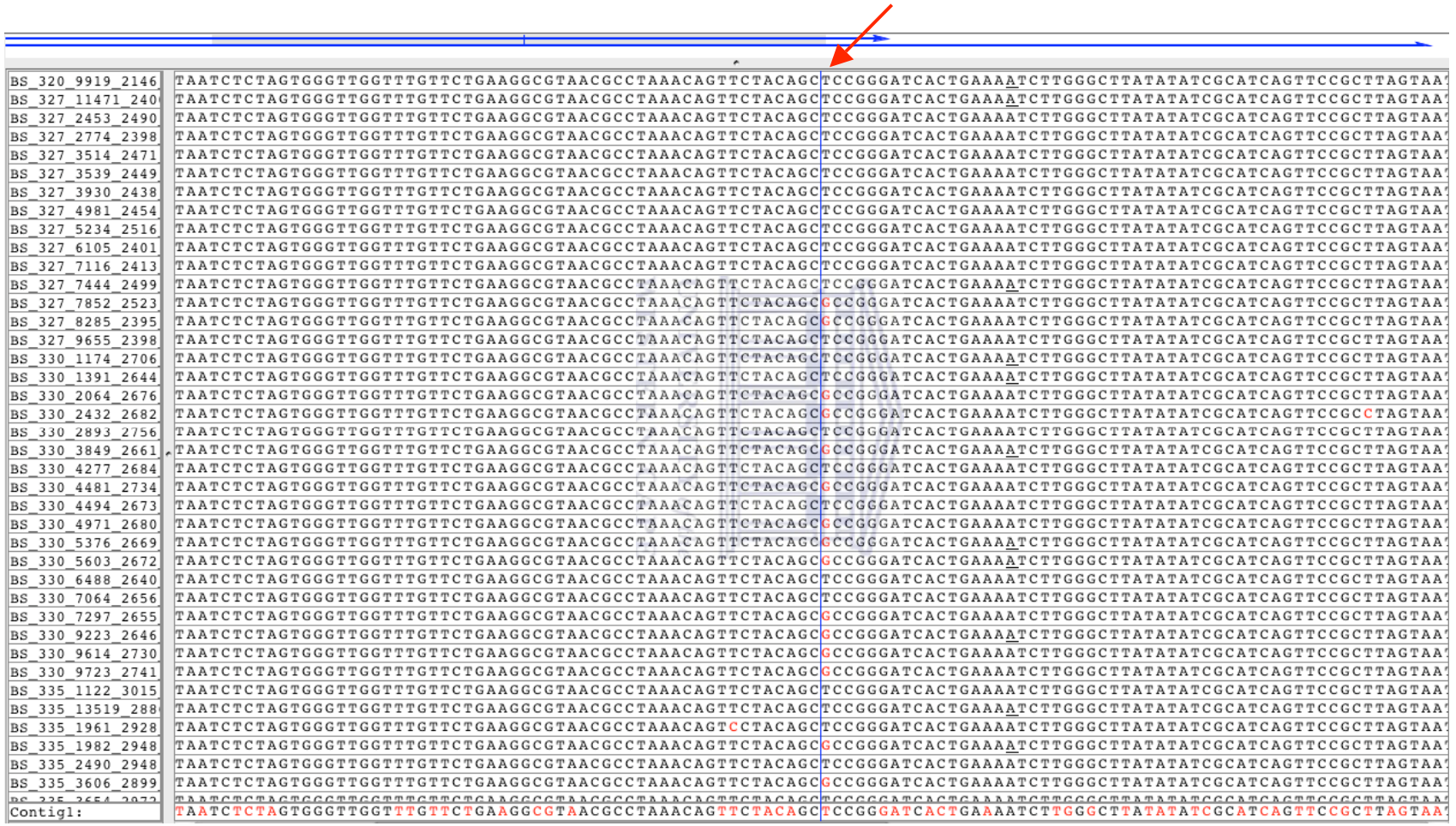

BS_320_9919_2146
BS_327_11471_240
BS_327_2453_2490
BS_327_2774_2398
BS_327_3514_2471
BS_327_3539_2449
BS_327_3930_2438
BS_327_4981_2454
BS_327_5234_2516
BS_327_6105_2401
BS_327_7116_2413
BS_327_7444_2499
BS_327_7852_2523
BS_327_8285_2395
BS_327_9655_2398
BS_330_1174_2706
BS_330_1391_2644
BS_330_2064_2676
BS_330_2432_2682
BS_330_2893_2756
BS_330_3849_2661
BS_330_4277_2684
BS_330_4481_2734
BS_330_4494_2673
BS_330_4971_2680
BS_330_5376_2669
BS_330_5603_2672
BS_330_6488_2640
BS_330_7064_2656
BS_330_7297_2655
BS_330_9223_2646
BS_330_9614_2730
BS_330_9723_2741
BS_335_1122_3015
BS_335_13519_288
BS_335_1961_2928
BS_335_1982_2948
BS_335_2490_2948
BS_335_3606_2899
Contig1:
TAATCTCTAGTGGGTTGGTTTGTTCTGAAGGCGTAACGCCTAAACAGTTCTACAGCTCCGGGATCACTGAAAATCTTGGGCTTATATATCGCATCAGTTCCGCTTAGTAA

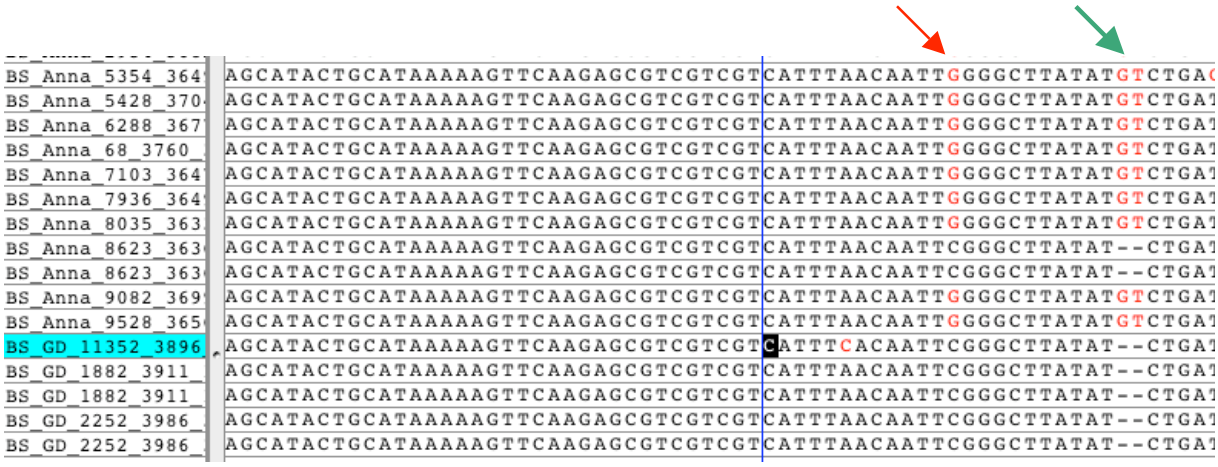
Table 34. Analysis of segregation for nine candidate SNPs with the potential of mapping to the existing Anna/Golden Delicious genetic linkage map. The 1/0 binary code was used to indicate hetero- /homo- zygous SNP genotype inheritance respectively, '?' shows insufficient data, '-' denotes missing data and the parental genotypes are written in the order Anna/Golden Delicious with first six being polymorphic in Anna and the last three polymorphic in Golden Delicious.

SNP	Parental genotype	11	12	16	19	51	52	130	276	320	327	330	335	344	353
017	TG/TT	0	-	1	0	1	1	0	1	0	1	1	1	1	1
018	AG/AA	0	-	0	1	0	1	1	1	0	1	1	1	1	1
035	TA/TT	0	0	1	0	1	1	0	1	-	0	1	1	1	1
105	GA/GG	0	1	1	?	1	1	0	1	0	0	1	0	1	0
125	CA/CC	0	-	1	0	1	1	1	1	1	1	0	1	1	1
198	TC/TT	1	1	-	1	0	1	1	0	1	1	1	0	1	1
119	TT/TC	0	1	-	0	1	1	0	1	1	0	1	1	1	1
210	GG/GT	1	1	1	1	1	1	1	0	1	0	1	1	0	1
276	CC/CT	1	0	1	1	0	1	0	0	0	0	0	0	1	1

Table 35. Candidate SNPs that do not match the currently existing Anna/Golden Delicious genetic linkage map. These SNPs could not be placed on the existing genetic linkage map; the hyphen '-' is used here to denote missing data.

SNP	Parental genotype	11	12	16	19	51	52	130	276	320	327	330	335	344	353
Polymorphic in Anna															
040	CG/CC	-	1	1	1	1	1	1	1	1	1	1	1	1	1
112	TC/TT	0	1	0	1	1	1	1	1	-	0	-	-	1	1
146	CA/CC	0	0	1	1	0	1	0	-	1	1	1	0	1	1
151	CG/CC	1	0	0	1	1	1	1	1	-	1	0	1	1	1
242	AG/AA	1	1	1	1	1	0	0	0	0	1	1	0	1	1
336	CA/CC	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Polymorphic in Golden Delicious															
003	TT/TC	1	1	1	1	0	1	1	1	1	1	1	1	1	1
006	GG/GA	1	0	0	0	1	1	0	1	0	1	1	0	1	1
045	AA/AC	1	0	0	0	0	1	0	-	0	1	1	0	1	1
062	CC/GC	1	1	1	1	1	1	-	1	1	1	1	1	1	1
081	CC/CT	1	0	1	0	0	0	0	0	1	0	0	0	0	0
116	CC/CA	1	0	1	1	0	0	1	0	1	1	0	0	1	1
182	TT/TG	1	1	1	1	0	1	1	1	0	1	0	0	1	1
288	AA/AG	1	1	1	1	1	1	1	1	1	1	1	1	1	1
322	GG/GA	0	0	0	0	0	1	0	1	0	0	0	0	0	0

The hyphen '-' in both tables highlights missing data meaning sequence reads in those contigs were either relatively too few and homozygous as compared to numerical representations from other progeny or alternatively they were not represented at all. In cases where the genotypes were heterozygous, the number of reads from each progeny was not a significant factor. However, the number of inherited polymorphic nucleotides expressed as a percentage of the total reads for progeny with the largest number of sequences per SNP was considered as guidelines for all homozygous genotypes. Cases where a decision could not be made with reasonable statistical significance were indicated as missing data. Candidate SNPs 003, 040, 062, 288 and 336 in Table 35 showing 100% heterozygosity in all the fourteen seedling can either be showing a case of gene duplication rather than allelic segregation or insufficient data from the homozygous parent. Candidate SNPs 040 and 062 were paired to indels as shown in Figure 7.4 and as such could not be dismissed as artefacts.



```
BS_Anna_5354_364  AGCATACTGCATAAAAAGTTCAAGAGCGTCGTCGTCATTTAACAATTGGGGCTTATATGTCTGA
BS_Anna_5428_370  AGCATACTGCATAAAAAGTTCAAGAGCGTCGTCGTCATTTAACAATTGGGGCTTATATGTCTGA
BS_Anna_6288_367  AGCATACTGCATAAAAAGTTCAAGAGCGTCGTCGTCATTTAACAATTGGGGCTTATATGTCTGA
BS_Anna_68_3760_  AGCATACTGCATAAAAAGTTCAAGAGCGTCGTCGTCATTTAACAATTGGGGCTTATATGTCTGA
BS_Anna_7103_364  AGCATACTGCATAAAAAGTTCAAGAGCGTCGTCGTCATTTAACAATTGGGGCTTATATGTCTGA
BS_Anna_7936_364  AGCATACTGCATAAAAAGTTCAAGAGCGTCGTCGTCATTTAACAATTGGGGCTTATATGTCTGA
BS_Anna_8035_363  AGCATACTGCATAAAAAGTTCAAGAGCGTCGTCGTCATTTAACAATTGGGGCTTATATGTCTGA
BS_Anna_8623_363  AGCATACTGCATAAAAAGTTCAAGAGCGTCGTCGTCATTTAACAATTCGGGCTTATAT--CTGA
BS_Anna_8623_363  AGCATACTGCATAAAAAGTTCAAGAGCGTCGTCGTCATTTAACAATTCGGGCTTATAT--CTGA
BS_Anna_9082_369  AGCATACTGCATAAAAAGTTCAAGAGCGTCGTCGTCATTTAACAATTGGGGCTTATATGTCTGA
BS_Anna_9528_365  AGCATACTGCATAAAAAGTTCAAGAGCGTCGTCGTCATTTAACAATTGGGGCTTATATGTCTGA
BS_GD_11352_3896  AGCATACTGCATAAAAAGTTCAAGAGCGTCGTCGTCATTTCACAATTCGGGCTTATAT--CTGA
BS_GD_1882_3911_  AGCATACTGCATAAAAAGTTCAAGAGCGTCGTCGTCATTTAACAATTCGGGCTTATAT--CTGA
BS_GD_1882_3911_  AGCATACTGCATAAAAAGTTCAAGAGCGTCGTCGTCATTTAACAATTCGGGCTTATAT--CTGA
BS_GD_2252_3986_  AGCATACTGCATAAAAAGTTCAAGAGCGTCGTCGTCATTTAACAATTCGGGCTTATAT--CTGA
BS_GD_2252_3986_  AGCATACTGCATAAAAAGTTCAAGAGCGTCGTCGTCATTTAACAATTCGGGCTTATAT--CTGA
```

Figure 7.4. Candidate SNP paired to a two-base indel. The red and the green arrows are inserted here to indicate the candidate SNP and indel respectively.

7.4.3 Bin mapping of segregating SNPs

Potential locations for SNPs 017, 018, 035, 105, 125, 198, 210 and 276 were determined by matching the segregation pattern to those in the existing Anna/Golden Delicious genetic linkage map (Table 36), though with varying degrees of precision. The genotypes for sample 12 were largely disregarded due to apparent distortions in the distribution of sequence reads. The worst match was between SNP-125 and the SSR marker CH05g03 in LG 17 where three points of disagreements in the observed pattern of segregation. The match between SNP-276 and SSR marker CH01d09 in LG 12 showed two disagreements and two points of missing data in the later. SNPs 017 and 035 both had two possible locations in LG 1 (Hi21g05) and LG 4 (CH02c02b) as their closest matches. SNP-018 was shown to be the reverse phase of SSR marker CH03c01 in LG 6. This phenomenon is made possible by the uncertainty in tracing the parentage of the monomorphic nucleotide (A) in the genotype given that the parental genotypes are GA/AA. Complete genetic linkage map tables are attached as an appendix.

Table 36. Determination of the genetic linkage positions of SNPs identified through sequence assemblies by association mapping. The yellow highlighting indicates positions with disagreements and '-' indicate missing data. LG refers to the linkage group to which the SNP below it was positioned

Locus		11	12	16	19	51	52	130	276	320	327	330	335	344	353
LG1															
Hi21g05	(ll,lm)	0	1	1	0	1	1	0	1	0	1	1	1	1	1
SNP_017		**0**		**1**	**0**	**1**	**1**	**0**	**1**	**0**	**1**	**1**	**1**	**1**	**1**
SNP_035		**0**		**1**	**0**	**1**	**1**	**0**	**1**	**-**	**0**	**1**	**1**	**1**	**1**
SNP_125		**0**		**1**	**0**	**1**	**1**	**1**	**1**	**1**	**1**	**0**	**1**	**1**	**1**
CH03g12x	(ll,lm)	0	1	1	0	1	1	0	0	0	1	1	1	0	1
Hi02b10x	(ll,lm)	0	1	1	-	1	1	0	0	0	1	1	1	0	1
Hi02c07	(ll,lm)	0	1	1	0	1	1	0	-	0	1	1	1	0	0
SA-A369	(ll,lm)	0	1	1	-	1	1	0	0	0	1	1	1	0	0
AG11	(ll,lm)	0	1	0	0	1	1	0	0	0	1	1	1	1	0
SA-A490		-	1	-	1	1	1	1	1	-	1	1	0	1	1
NZmsCN879773	(ll,lm)	0	1	1	1	1	1	1	1	0	1	1	0	1	0
Hi07d08	(ll,lm)	0	1	1	1	1	1	1	0	0	1	1	0	1	0
CH05g08	(ll,lm)	0	1	1	1	1	1	1	0	0	1	1	0	1	0
LG4															
05g8	(ll,lm)	1	1	-	0	1	1	1	1	0	1	1	0	1	0
CH04e02	(ll,lm)	1	1	1	0	1	-	1	1	0	1	1	0	0	0
Hi01e10	(ll,lm)	1	1	1	0	1	1	1	1	0	1	1	0	0	0
SA-A687	(ll,lm)	1	1	-	-	1	1	1	1	0	1	1	1	0	0
CH01b09b	(ll,lm)	0	1	1	0	1	1	1	1	0	1	1	1	0	0
Hi23g08	(ll,lm)	0	1	1	0	1	1	1	1	0	1	1	1	0	0
GD162 / A8		0	1	1	0	1	1	1	1	0	1	1	1	0	0
CH02h11a	(ll,lm)	0	1	1	0	1	1	1	1	0	1	1	1	-	-
CH01d03	(ll,lm)	0	1	1	0	1	1	1	1	0	1	1	1	-	0
SA-A615		1	1	1	0	1	1	1	1	-	1	1	1	0	0
Hi23g02	(ll,lm)	1	1	1	0	1	1	1	1	0	-	1	1	0	0
CH05d02	(ll,lm)	0	1	1	1	1	1	1	1	1	1	1	1	1	0
CH02c02b	(ll,lm)	0	1	1	0	1	1	0	1	0	1	1	1	1	1
SNP_017		**0**	**-**	**1**	**0**	**1**	**1**	**0**	**1**	**0**	**1**	**1**	**1**	**1**	**1**
SNP_035		**0**	**-**	**1**	**0**	**1**	**1**	**0**	**1**	**-**	**0**	**1**	**1**	**1**	**1**
SNP_125		**0**	**-**	**1**	**0**	**1**	**1**	**1**	**1**	**1**	**1**	**0**	**1**	**1**	**1**

LG2															
SNP_105		**0**	**-**	**1**	**-**	**1**	**1**	**0**	**1**	**0**	**0**	**1**	**0**	**1**	**0**
Hi22d06	(ll,lm)	0	1	1	1	1	0	0	1	0	0	1	0	1	0
CH02f061	(ll,lm)	0	1	1	1	1	0	0	1	0	0	1	0	1	0
AJ251116-SSR	(ll,lm)	0	1	1	1	1	1	0	1	0	0	0	0	1	0
SA-A389	(ll,lm)	-	1	1	1	1	1	0	-	0	0	0	0	1	0
SA-A513		0	1	-	1	1	1	0	1	0	0	0	0	1	0
CH03d10	(ll,lm)	0	1	-	1	0	1	0	1	0	0	0	1	1	0
CH05e03	(ll,lm)	0	1	0	1	0	1	0	0	0	1	0	1	1	0
CH02a04y	(ll,lm)	0	1	0	1	0	1	0	0	0	0	0	1	0	0
CH02c061	(ll,lm)	0	1	0	1	0	1	0	0	0	0	0	0	0	0

LG6															
Hi05d10	(ll,lm)	1	1	1	0	1	0	1	1	0	-	0	0	1	1
CH03d12	(ll,lm)	1	1	1	0	1	0	0	1	0	0	0	0	0	1
Hi04d10	(ll,lm)	1	1	1	0	1	-	0	1	0	0	0	0	0	1
SA-A484		1	0	-	0	1	0	0	1	-	0	0	0	0	1
Hi03a03x	(ll,lm)	1	0	1	0	1	-	0	0	-	0	0	0	0	1
CH03c01	(ll,lm)	1	0	1	0	1	0	0	0	1	0	0	0	0	-
SNP_018		**0**		**0**	**1**	**0**	**1**	**1**	**1**	**0**	**1**	**1**	**1**	**1**	**1**
Hi07b06	(ll,lm)	1	0	1	0	1	-	0	0	1	0	0	0	0	1

LG12															
SA-A219	(nn,np)	1	1	0	1	0	1	0	1	0	1	1	0	-	0
28f4	(nn,np)	1	1	1	1	0	1	0	1	1	1	1	0	-	0
CH01g121	(nn,np)	1	0	1	1	0	1	0	1	1	0	1	0	0	0
CH04d02	(nn,np)	1	0	1	1	0	1	0	1	1	0	1	0	-	0
SA-A656		0	0	1	1	0	1	0	1	1	0	1	0	0	0
CH05d11	(nn,np)	1	0	1	1	0	1	0	1	1	0	1	0	0	0
CH01d09	(nn,np)	1	0	1	1	0	1	0	-	-	0	1	0	1	0
SNP_276		**1**	**0**	**1**	**1**	**0**	**1**	**0**	**0**	**0**	**0**	**0**	**0**	**1**	**1**
CH01f021	(nn,np)	1	0	1	1	0	1	-	1	1	0	1	0	1	0
CH02h11b	(nn,np)	1	0	1	1	0	1	0	1	1	0	1	0	1	0

LG16															
CH02d10a	(ll,lm)	1	1	-	-	0	1	1	1	1	1	1	0	1	1
SA-A343	(ll,lm)	1	1	0	1	0	1	1	1	1	1	1	0	1	1
CH05b06x	(ll,lm)	1	1	0	1	0	1	1	1	1	1	1	0	1	1
Hi12a02	(ll,lm)	1	1	0	1	0	1	1	0	1	1	1	0	1	1
SNP_198		**1**	**-**	**-**	**1**	**0**	**1**	**1**	**0**	**1**	**1**	**0**	**1**	**1**	**1**
Hi02h08	(ll,lm)	1	1	0	1	0	-	1	0	1	1	0	0	1	1
CH05e04	(ll,lm)	1	1	0	1	-	1	1	-	1	1	0	0	1	1
SA-A601	(ll,lm)	1	1	0	0	-	1	1	1	1	1	0	0	1	1
SA-A728	(ll,lm)	1	0	-	-	0	1	0	1	1	1	0	0	1	1
SA-A494	(ll,lm)	1	0	0	0	0	1	0	1	1	1	0	0	1	1
CH05a04	(ll,lm)	1	0	0	0	0	0	0	0	1	1	0	0	1	1
Hi02b10y	(ll,lm)	1	0	0	-	-	0	0	-	1	1	0	0	1	1
SA-A680		1	0	0	0	1	0	0	1	1	0	1	0	1	1
CH04f10	(ll,lm)	1	0	0	-	-	-	0	-	-	0	1	0	-	-
SA-A267	(ll,lm)	1	0	0	0	1	0	0	0	1	0	1	0	1	1
SA-A430	(ll,lm)	0	0	1	1	0	0	-	-	-	1	-	-	-	-

LG17															
CH05g03	(nn,np)	0	0	-	1	0	0	0	1	1	0	1	0	1	0
CH02g04	(nn,np)	0	0	1	1	0	0	0	1	1	0	1	1	1	1
CH04c06	(nn,np)	0	0	1	1	0	0	0	0	1	0	1	1	0	-
CH01h011	(nn,np)	0	0	1	1	0	0	0	0	1	0	1	1	0	1
SA-A736	(nn,np)	0	1	1	1	1	0	0	0	1	0	1	1	0	1
SA-A413		0	1	-	1	1	0	0	0	-	0	1	1	0	1
SA-A236	(nn,np)	0	1	1	0	1	0	0	1	1	0	1	1	0	1
Hi07h02	(nn,np)	0	1	1	0	1	-	0	-	1	0	1	1	0	1
SA-A234	(nn,np)	0	1	1	1	1	1	0	0	1	0	1	1	0	-
SNP_210		**1**	**1**	**1**	**1**	**1**	**1**	**1**	**0**	**1**	**0**	**1**	**1**	**0**	**1**

7.5 DISCUSSION

High throughput sequencing of RGA PCR amplicons as a strategy for the identification of SNPs in the 3'-end region of the NBS domain was tested in this chapter. The deep sequencing capacity of the 454 system theoretically aims at capturing individual DNA fragments in a PCR amplification product, immobilise these beads and achieve clonal propagation of each molecule individually in a virtual micro-reactor. Work done in this chapter shows that the deep sequencing capacity of the 454 system relies to a large extent on the molar concentrations of individual molecules in a given reaction. This has the effect of magnifying PCR amplification distortions thereby giving sequence reads whose numerical representation per run is distorted largely in favour of genes with either a high copy number or with high PCR efficiencies.

A dataset containing 136 370 sequence reads was generated from sixteen samples made up of two parents and fourteen progeny of the Anna x Golden Delicious bin mapping population. Seven of the samples contributed reads in the range ~10000 to 15000 and the rest were in the range ~5000 to 8700. Read lengths ranged between ~40 to ~300 nucleotides with the distribution varying depending on total size of the dataset (Table 32) although the statistical mode was approximately 255 ± 3 nucleotides for all samples. There was a gradual reduction in sequence reads that are 100 nucleotides and shorter as the size of the dataset increased.

The priming regions were deleted from all the sequences to eliminate non-specific bias in the sequence assemblies. Sequencing artefacts such as reads containing the entire 45

nucleotides of the forward fusion oligonucleotide (the 23mer degenerate oligonucleotide linked to the 19mer sequencing tag), reads with above average random nucleotide substitutions per contig and entire contigs that did not produce a significant hit to NBS-LRR genes in the NCBI non-redundant database were all eliminated from the assemblies. Comparative analysis of sequence assemblies from sample – to - sample revealed that as the size of the dataset increases the number of contigs increased steadily as the number of unassembled (singletons) decreased. The Anna x Golden Delicious combined assembly (Table 33) showed that even with large datasets there are subsets of RGAs that are unique to individual cultivars and do not share significant homology between cultivars.

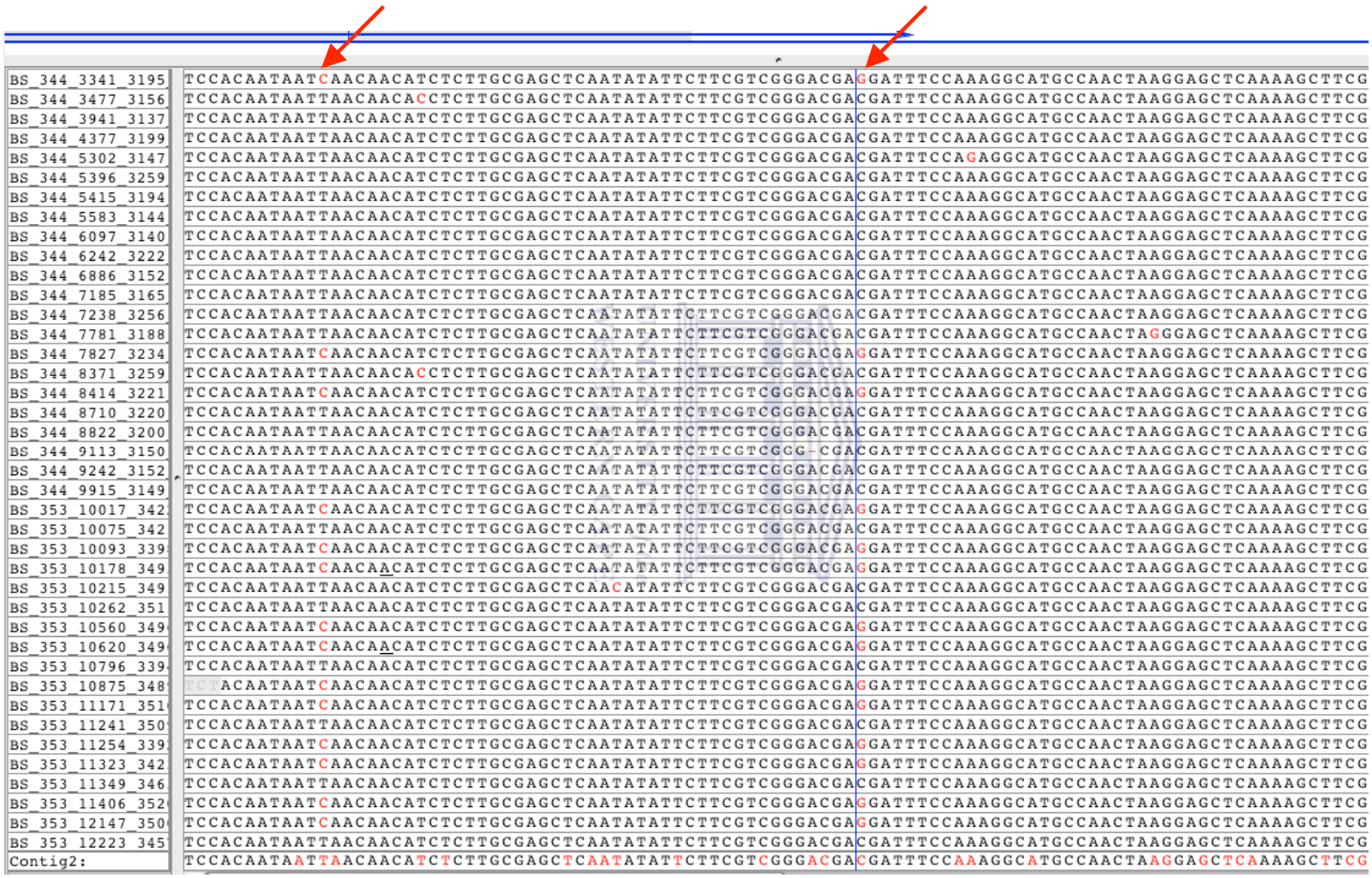
For the subsets of genes that do share homology between cultivars, a set of contigs was obtained that contained sequences from both Anna and Golden Delicious. These were used to identify polymorphic SNPs whose pattern of segregation could be used to map those genes using the Anna x Golden Delicious bin mapping population. A total of 210 candidate SNPs were identified in 89 contigs, however, the majority of these SNPs were heterogygous in both parents and as such could not be used in bin mapping. A set of 24 candidate SNPs that were homozygous in one parent and heterozygous in the other were identified (Tables 34 and 35). Only eight of these could be mapped to the existing Anna x Golden Delicious genetic linkage map though with variable degrees of precision (Table 36).

The drop in the number of mappable 'SNPs' from 24 to 8 in this analysis can be explained in part by the occurrence of gene conversion (section 3.3.7), unequal crossover

and tandem gene duplication (Michelmore and Meyers, 1998; Baumgarten *et al.*, 2003; Leister, 2004) in the NBS-LRR gene family. Gene duplication and gene conversion events increase the percentage sequence homology among duplicons to a level where it approaches allelic variation in gene families characterised by clusters of highly duplicated genes. This phenomenon creates paralogous sequence variants (PSVs), which are by nature mutations originating from multiple sites and as such will always give genotypes that are 'heterozygote-like' (Hurles *et al.*, 2002; Fredman *et al.*, 2004). Candidate SNPs 040, 336, 003, 062 and 288 (Table 35) represent such a phenomenon.

Candidate SNP_062 (Figure 7.5) shows the distribution of the G/C mutation (paired to C/T) between seedlings 344 and 353. In this figure there is an approximately 1:1 distribution between G and C (or C and T) for seedling 353 whereas in seedling 344 there are only 3 Gs out of 22 sequences. The presence of 'G' in sequences from seedling 344 results in a heterozygote genotype and such a phenomenon with variable allelic ratios is characteristic of the segregation pattern for SNP-062. The sequence alignment in Figure 7.5 was consistent at 98% minimum percent identity, meaning that such distortions cannot be resolved by increasing the stringency of the sequence assemblies. Disregarding the low frequency mutation could resolve the problem of overestimating heterozygosity, however, given the problem of amplification and sequencing distortions that are magnified in 454 sequencing platforms and the variable ratio of such mutations it remains impossible to significantly purify unique SNPs from PSVs using sequencing.

Figure 7.5. Segregation distortion in the contig containing SNP-062. The two red arrows indicate paired 'SNPs' with a distorted segregation pattern between seedlings 344 and 353.


BS_344_3341_3195 TCCACAATAATCAACAACATCTCTTGCGAGCTCAATATATTCTTCGTCGGGACGAGGATTTCCAAAGGCATGCCAACTAAGGAGCTCAAAAGCTTCG
BS_344_3477_3156 TCCACAATAATTAACAACACCTCTTGCGAGCTCAATATATTCTTCGTCGGGACGACGATTTCCAAAGGCATGCCAACTAAGGAGCTCAAAAGCTTCG
BS_344_3941_3137 TCCACAATAATTAACAACATCTCTTGCGAGCTCAATATATTCTTCGTCGGGACGACGATTTCCAAAGGCATGCCAACTAAGGAGCTCAAAAGCTTCG
BS_344_4377_3199 TCCACAATAATTAACAACATCTCTTGCGAGCTCAATATATTCTTCGTCGGGACGACGATTTCCAAAGGCATGCCAACTAAGGAGCTCAAAAGCTTCG
BS_344_5302_3147 TCCACAATAATTAACAACATCTCTTGCGAGCTCAATATATTCTTCGTCGGGACGACGATTTCCAGAGGCATGCCAACTAAGGAGCTCAAAAGCTTCG
BS_344_5396_3259 TCCACAATAATTAACAACATCTCTTGCGAGCTCAATATATTCTTCGTCGGGACGACGATTTCCAAAGGCATGCCAACTAAGGAGCTCAAAAGCTTCG
BS_344_5415_3194 TCCACAATAATTAACAACATCTCTTGCGAGCTCAATATATTCTTCGTCGGGACGACGATTTCCAAAGGCATGCCAACTAAGGAGCTCAAAAGCTTCG
BS_344_5583_3144 TCCACAATAATTAACAACATCTCTTGCGAGCTCAATATATTCTTCGTCGGGACGACGATTTCCAAAGGCATGCCAACTAAGGAGCTCAAAAGCTTCG
BS_344_6097_3140 TCCACAATAATTAACAACATCTCTTGCGAGCTCAATATATTCTTCGTCGGGACGACGATTTCCAAAGGCATGCCAACTAAGGAGCTCAAAAGCTTCG
BS_344_6242_3222 TCCACAATAATTAACAACATCTCTTGCGAGCTCAATATATTCTTCGTCGGGACGACGATTTCCAAAGGCATGCCAACTAAGGAGCTCAAAAGCTTCG
BS_344_6886_3152 TCCACAATAATTAACAACATCTCTTGCGAGCTCAATATATTCTTCGTCGGGACGACGATTTCCAAAGGCATGCCAACTAAGGAGCTCAAAAGCTTCG
BS_344_7185_3165 TCCACAATAATTAACAACATCTCTTGCGAGCTCAATATATTCTTCGTCGGGACGACGATTTCCAAAGGCATGCCAACTAAGGAGCTCAAAAGCTTCG
BS_344_7238_3256 TCCACAATAATTAACAACATCTCTTGCGAGCTCAATATATTCTTCGTCGGGACGACGATTTCCAAAGGCATGCCAACTAAGGAGCTCAAAAGCTTCG
BS_344_7781_3188 TCCACAATAATTAACAACATCTCTTGCGAGCTCAATATATTCTTCGTCGGGACGACGATTTCCAAAGGCATGCCAACTAGGGAGCTCAAAAGCTTCG
BS_344_7827_3234 TCCACAATAATCAACAACATCTCTTGCGAGCTCAATATATTCTTCGTCGGGACGAGGATTTCCAAAGGCATGCCAACTAAGGAGCTCAAAAGCTTCG
BS_344_8371_3259 TCCACAATAATTAACAACACCTCTTGCGAGCTCAATATATTCTTCGTCGGGACGACGATTTCCAAAGGCATGCCAACTAAGGAGCTCAAAAGCTTCG
BS_344_8414_3221 TCCACAATAATCAACAACATCTCTTGCGAGCTCAATATATTCTTCGTCGGGACGAGGATTTCCAAAGGCATGCCAACTAAGGAGCTCAAAAGCTTCG
BS_344_8710_3220 TCCACAATAATTAACAACATCTCTTGCGAGCTCAATATATTCTTCGTCGGGACGACGATTTCCAAAGGCATGCCAACTAAGGAGCTCAAAAGCTTCG
BS_344_8822_3200 TCCACAATAATTAACAACATCTCTTGCGAGCTCAATATATTCTTCGTCGGGACGACGATTTCCAAAGGCATGCCAACTAAGGAGCTCAAAAGCTTCG
BS_344_9113_3150 TCCACAATAATTAACAACATCTCTTGCGAGCTCAATATATTCTTCGTCGGG-CGACGATTTCCAAAGGCATGCCAACTAAGGAGCTCAAAAGCTTCG
BS_344_9242_3152 TCCACAATAATTAACAACATCTCTTGCGAGCTCAATATATTCTTCGTCGGGACGACGATTTCCAAAGGCATGCCAACTAAGGAGCTCAAAAGCTTCG
BS_344_9915_3149 TCCACAATAATTAACAACATCTCTTGCGAGCTCAATATATTCTTCGTCGGGACGACGATTTCCAAAGGCATGCCAACTAAGGAGCTCAAAAGCTTCG
BS_353_10017_342 TCCACAATAATCAACAACATCTCTTGCGAGCTCAATATATTCTTCGTCGGGACGAGGATTTCCAAAGGCATGCCAACTAAGGAGCTCAAAAGCTTCG
BS_353_10075_342 TCCACAATAATTAACAACATCTCTTGCGAGCTCAATATATTCTTCGTCGGGACGACGATTTCCAAAGGCATGCCAACTAAGGAGCTCAAAAGCTTCG
BS_353_10093_339 TCCACAATAATCAACAACATCTCTTGCGAGCTCAATATATTCTTCGTCGGGACGAGGATTTCCAAAGGCATGCCAACTAAGGAGCTCAAAAGCTTCG
BS_353_10178_349 TCCACAATAATCAACAACATCTCTTGCGAGCTCAATATATTCTTCGTCGGGACGAGGATTTCCAAAGGCATGCCAACTAAGGAGCTCAAAAGCTTCG
BS_353_10215_349 TCCACAATAATTAACAACATCTCTTGCGAGCTCAACATATTCTTCGTCGGGACGACGATTTCCAAAGGCATGCCAACTAAGGAGCTCAAAAGCTTCG
BS_353_10262_351 TCCACAATAATTAACAACATCTCTTGCGAGCTCAATATATTCTTCGTCGGGACGACGATTTCCAAAGGCATGCCAACTAAGGAGCTCAAAAGCTTCG
BS_353_10560_349 TCCACAATAATCAACAACATCTCTTGCGAGCTCAATATATTCTTCGTCGGGACGAGGATTTCCAAAGGCATGCCAACTAAGGAGCTCAAAAGCTTCG
BS_353_10620_349 TCCACAATAATCAACAACATCTCTTGCGAGCTCAATATATTCTTCGTCGGGACGAGGATTTCCAAAGGCATGCCAACTAAGGAGCTCAAAAGCTTCG
BS_353_10796_339 TCCACAATAATTAACAACATCTCTTGCGAGCTCAATATATTCTTCGTCGGGACGACGATTTCCAAAGGCATGCCAACTAAGGAGCTCAAAAGCTTCG
BS_353_10875_348 ACAATAATCAACAACATCTCTTGCGAGCTCAATATATTCTTCGTCGGGACGAGGATTTCCAAAGGCATGCCAACTAAGGAGCTCAAAAGCTTCG
BS_353_11171_351 TCCACAATAATCAACAACATCTCTTGCGAGCTCAATATATTCTTCGTCGGGACGAGGATTTCCAAAGGCATGCCAACTAAGGAGCTCAAAAGCTTCG
BS_353_11241_350 TCCACAATAATTAACAACATCTCTTGCGAGCTCAATATATTCTTCGTCGGGACGACGATTTCCAAAGGCATGCCAACTAAGGAGCTCAAAAGCTTCG
BS_353_11254_339 TCCACAATAATCAACAACATCTCTTGCGAGCTCAATATATTCTTCGTCGGGACGAGGATTTCCAAAGGCATGCCAACTAAGGAGCTCAAAAGCTTCG
BS_353_11323_342 TCCACAATAATCAACAACATCTCTTGCGAGCTCAATATATTCTTCGTCGGGACGAGGATTTCCAAAGGCATGCCAACTAAGGAGCTCAAAAGCTTCG
BS_353_11349_346 TCCACAATAATTAACAACATCTCTTGCGAGCTCAATATATTCTTCGTCGGGACGACGATTTCCAAAGGCATGCCAACTAAGGAGCTCAAAAGCTTCG
BS_353_11406_352 TCCACAATAATCAACAACATCTCTTGCGAGCTCAATATATTCTTCGTCGGGACGAGGATTTCCAAAGGCATGCCAACTAAGGAGCTCAAAAGCTTCG
BS_353_12147_350 TCCACAATAATCAACAACATCTCTTGCGAGCTCAATATATTCTTCGTCGGGACGAGGATTTCCAAAGGCATGCCAACTAAGGAGCTCAAAAGCTTCG
BS_353_12223_345 TCCACAATAATTAACAACATCTCTTGCGAGCTCAATATATTCTTCGTCGGGACGACGATTTCCAAAGGCATGCCAACTAAGGAGCTCAAAAGCTTCG
Contig2: TCCACAATAATTAACAACATCTCTTGCGAGCTCAATATATTCTTCGTCGGGACGACGATTTCCAAAGGCATGCCAACTAAGGAGCTCAAAAGCTTCG

As discussed above, high throughput sequencing using the 454 system randomly magnifies PCR amplification distortions. This was also observed in the sequence dataset generated for sample 12 in which despite having comparative total sequence numbers to sample 11 showed a serious case of sequence distribution distortions. This sample either had none or was poorly represented in some contigs that were used to assess the pattern of segregation in the bin mapping population. This generally was not a problem in cases where sample 12 was heterozygous, however, to ascertain true homozygosity requires a large enough dataset per contig. As a result of this distortion, homozygous genotypes of sample 12 were largely disregarded in determining bin-mapping positions for the SNPs.

To assess the extent of amplification distortions, all contigs containing either Anna or Golden Delicious only sequence reads in the Anna x Golden Delicious assembly (section 7.4.1 and Table 33 B) were assembled with the progeny datasets. In this assembly the segregation pattern of all contigs containing SNP-like mutations were scored and the results are presented in Table 37. In this table '1/0' shows heterozygous/ homozygous genotypes respectively and '-' denotes total absence of sequence reads from the respective progeny. In cases where homozygous genotypes were scored, the 'yellow highlight' shows instances where sequence read representation was lower than 10 per contig from the respective progeny. Results of this table show that these SNPs though segregating in other progeny confirm the amplification distortions characterising the whole dataset. It can be assumed from these results that these contigs represent genes that are either present in very low copy number or that they have a low PCR efficiency with the amplification conditions used in this analysis. The sequence numbers per contig are

not enough to make an assessment of whether any of these represent null alleles. Contig 061 and 130 that show TT genotypes (as opposed to TC) seems to suggest segregation from a TT/TC Golden Delicious genotype. Contigs 178 and 195 are present only in Golden Delicious and not in any of the other 15 datasets assessed here including Anna. Results displayed in Table 37 could be analysed further through either SNaPshot™ or SNPLex assays in bin mapping population to ascertain whether they are truly present in Golden Delicious only or that they are also present in the other fifteen genomes.

Table 37. Segregation of candidate SNPs identified in contigs with sequence reads from either Anna or Golden Delicious. The pattern of segregation for these candidate SNPs was assessed in all the 14 bin mapping seedlings. The hyphen '-' represents no missing data from the respective seedling, the yellow highlight shows homozygotes determined from 10 sequence reads or below.

Contig	SNP	11	12	16	19	51	52	130	276	320	327	330	335	344	353
Anna SNPs															
041	CA	-	-	-	-	-	0	-	1	-	-	-	-	-	-
056	TC	1	-	0	0	0	-	-	0	0	1	1	0	0	0
061	TC	1	0	0	0	0	0	0	0	TT	1	-	-	-	-
123	TC	1	0	0	0	0	0	0	0	0	1	0	0	1	0
130	TC	-	-	-	-	-	-	-	-	TT	1	-	TT	TT	-
156	TA	1	-	0	-	0	1	1	1	1	1	1	1	1	1
163	TC	-	0	1	1	-	1	-	-	-	-	-	-	-	-
160	GA	-	1	1	0	0	1	1	1	-	-	-	1	-	0
Golden Delicious SNPs															
019	CA	-	0	0	0	1	0	1	0	1	0	0	0	0	0
052	GA	-	-	0	-	-	-	-	-	-	-	-	-	-	-
128	GA	0	1	0	0	0	0	0	0	0	1	0	0	0	1
178	GA	-	-	-	-	-	-	-	-	-	-	-	-	-	-
195	GT	-	-	-	-	-	-	-	-	-	-	-	-	-	-
196	TC	-	-	-	-	-	-	-	-	-	1	-	-	-	-
228	GA	-	-	0	-	-	-	-	-	-	0	0	0	0	-
233	GA	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Assessment of cluster distribution through sequence assemblies of the full length RGAs against the 454 dataset did not produce useful results. Although significant homologies existed between the two datasets, most of the full-length RGAs were incorporated into contigs containing candidate SNPs that were polymorphic in both parents and thus could not be mapped using the bin mapping approach. Only four SNPs, SNP-006, 112, 125 and 210 (Table 35) were located to clusters 3, 13, 14 and 5 (Figure 3.5) respectively. However, these were not enough to ascertain whether genes occurring in the same phylogenetic cluster also share the same linkage group.

CHAPTER 8: DISCUSSION

8.1 Sequencing and sequence analysis

Cloning and sequencing of RGAs was based on a set of degenerate oligonucleotides targeted at the P-loop and GLPL motifs flanking the NBS domain of the NBS-LRR resistance genes. Although the NBS domain is common in most nucleotide-binding proteins that hydrolyse either GTP or ATP (Pfeifer *et al.*, 2001; Pal *et al.*, 2007), co-conservation with the GLPL in plants appears to be unique to genes in the immune response systems (Tao *et al.*, 2000). These oligonucleotides thus amplify genes belonging to the NBS-LRR resistance family in plants with no co-amplification of ATPases/ GTPases or other proteins that possess an ABC domain.

Sequences generated from amplification of genomic DNA using these oligonucleotides were confirmed to be NBS-LRR genes through homology searches with the tblastx algorithm of the NCBI Blast search tool (Altschul *et al.*, 1997). *In silico* translation of these sequences to their corresponding amino acid sequences revealed that the NBS-LRR gene family contains a large number of pseudogenes. These were identified primarily by the presence of premature termination codons within a sequence's best ORF.

Inference of phylogeny using sequences from *Malus x domestica* (Borkh.) cultivar Golden Delicious was performed and these divided genes into two major subfamilies. The larger subfamily with almost double the gene copy number belongs to genes with an N-terminal Toll/IL-1R (TIR) domain and the second subfamily with genes belonged to the class of resistance genes lacking the TIR domain. The TIR domain is a common

feature in animal immune cells where it occurs associated with the plasma membrane (on the surfaces of macrophages, B and dendritic cells) and also recognizes pathogen associated molecular patterns (PAMPs) (Nurnberger *et al.*, 2004). In these cells they play a significant role in innate immune responses where they recognize patterns as opposed to race-specific recognition as in plants. In plants these genes have been implicated in inducible defense systems, which might explain the large size of the family. This division into subfamilies was based on the presence of highly conserved motifs within genes of the two subfamilies. An analysis of these motifs was performed and the major differences were localized to the kinase-2, kinase-3 and the RNBS-C motifs although other variable regions do exist within the NBS domain.

The phylogenetic analyses further grouped the genes into 12 true clusters and two groups of unrelated genes (singletons) based on sequence homology. These clusters are made up of genes that are physically located on the same locus on a chromosome and the unrelated singletons are possibly genes scattered singly on the genome (Mondragon-Palomino and Gaut, 2005). In this analysis the two groups of unrelated genes appear to be cluster together possibly as a consequence of long-branch attraction in calculating the tree topology. Analysis of the rate of pseudo- to functional- genes per cluster showed that these numbers vary greatly from 0:11 in cluster 4 to 51:2 in cluster 11 (Figure 3.5) and a range of combinations in between.

Assessments of coding to silent mutation rates were performed to give an indication of the evolutionary selection pressure per cluster. Results revealed that clusters within the

non-TIR subfamily were undergoing diversifying selection whereas those in the TIR subfamily showed a mix of clusters undergoing diversifying, neutral and purifying selection pressures. Whether this means non-TIR genes are a small subfamily within the NBS-LRR family still evolving rapidly remains to be worked out.

Evidence of gene conversion was detected in 9 of the 12 true clusters shown in Figure 3.5, the remaining two groups of unrelated genes were not analysed. As shown in table 12 cluster 10 was divided into 10a and 10b where the former revealed three incidences of gene duplication involving one pseudogene (GD20) and three other genes (GD105, GD187 and GD237). In this case two of the three identified incidences involved two different functional genes and the third incidence was with another pseudogene. Sub-cluster 10b showed ten incidences of gene duplication in which all participants were functional genes. Cluster 8 had evidence of three gene duplication events involving four pseudogenes. In other clusters not discussed here where evidence of gene duplication was detected, at least one event involved the interplay of a pseudogene. This might provide evidence that pseudogenes do play a role as reservoirs of genetic variability and possibly facilitate rapid generation of novel resistance gene specificities (Michelmore and Meyers, 1998). It is important to note here that these events take several millions of years of concerted evolution and involve several rounds of gene duplication, selection pressure, genetic recombination and other evolutionary forces (Michelmore and Meyers, 1998; Bergelson *et al.*, 2001; Balakirev and Ayala, 2003; Liu and Ekramoddoullah, 2003).

A comparative analysis of *Malus* orthologs performed using *Malus x domestica* (Borkh.),

M. prunifolia, *M. floribunda* and *M. buccata* with *Pinus monticolor* as the outgroup showed a high level of conservation of clusters. The phylogenetic tree constructed from sequences of these species produced clusters with representative genes from all the species. However, some clusters had a better representation and a few only contained genes from *Malus x domestica* (Borkh.) cvs Anna and Golden Delicious and this was mainly because these two cultivars had on average more sequences in the analysis. Clusters in this analysis were defined as groups of sequences having a common recent origin and most probably produced through a single duplication event (Pan *et al.*, 2000a). In other researches these genes have been shown to cluster on the same loci in linkage maps (Michelmore and Meyers, 1998; Pan *et al.*, 2000b; Gowda *et al.*, 2002; Xu *et al.*, 2007) although other studies show that even very homologous genes may be located on different loci or even different chromosomes (Baldi *et al.*, 2004).

Although it is clear that there are a number of evolutionary forces acting on clusters it would appear that the effect of intraspecific - haplotype divergence produces higher gene variations as compared to interspecific divergence that acts on orthologs and as a result homology between related genes in different species is conserved (Sun *et al.*, 2006). Comparative analysis of clusters between cultivars of the same species shows that gene copy number per cluster varies as shown between *Malus x domestica* (Borkh.) cultivars Anna and Golden Delicious (Figure 4.1) where some clusters have on average more sequences than the other (McHale *et al.*, 2006). Other genes accessed from GeneBank (Figure 4.1) did not belong to definite clusters were also detected in data accessed from GeneBank (Figure 4.1) and thus appears to be an established phenomenon in the NBS-

LRR gene family. This could suggest that the NBS-LRR gene family contains single genes scattered on isolated loci and whereas other loci contain groups of genes in tandem arrays or clusters scattered across the genome (Leister, 2004).

8.2 Saturation sequencing of the NBS-LRR gene family

PCR amplification, cloning and sequencing as a strategy leads to underestimation of gene copy numbers. This is related mainly to the average representation of each gene in a PCR product, which in turn depends on the average efficiency of amplification of each gene. The rates of coding to silent mutations detected in table 11 (section 3.3.5) and the rate of site – to - site amino acid residue conservation shown in Figure 3.9 (section 3.3.6) prove that even highly conserved motifs have sites that variable. This means the hybridisation efficiencies of degenerate oligonucleotides targeted at conserved sites are not comparable among different genes. Because of these factors sequence representation in targeted gene sequencing is biased towards genes that have high PCR amplification efficiencies. This technique either under-represents or misses other genes completely. In chapter 3 use of the GS20 sequencing technology allowed for sequencing of all possible gene copies in a single PCR amplification product thereby increasing coverage of all genes flanked by the P-loop and GLPL motifs. This eliminates ligation into a cloning vector, transformation into cells and losses associated with those methods.

The complete dataset of 661 sequences generated through cloning and sequencing of PCR amplification products (section 3.2) assembled into a total of 54 contigs where a contig represents a single gene and sequences making up that contig were thus assumed

to be alleles or multiple copies of that gene. Using the GS20 sequencing technology, a total of 5620 sequences from the 3′ end of the NBS domain with an average length of 110 bases were produced (section 4.4.4.4). Sequence assemblies of the two datasets from sections 3.2 and 4.4.4.4 gave a total of 278 contigs, 54 with at least one gene from section 3.2 and an additional 224 from the GS20 dataset. Based on this analysis and results of chapter 7, it was estimated that the NBS-LRR gene copy number for genes with an NBS domain that is flanked by a P-loop motif on the 5′-end and a GLPL motif on the 3′-end in the family is at least ~400. This approximation is made taking into account the uncertainty around the problem of unassembled genes/singletons in the generated dataset. Whether these all represent real single genes/ genes with a low PCR efficiency in the apple genome or that a given percentage represents reads highly distorted with sequencing artefacts could not be ascertained. The GS FLX sequencing performed in chapter 7 presents a rigorous analysis of the NBS-LRR gene copy number in the family. Independent high throughput sequencing of sixteen samples of the *Malus x domestica* (Borkh.) Anna x Golden Delicious bin mapping population gave contigs numbers in the range 220 to ~400 and up to a thousand singletons. Assuming random gene amplification and sequencing distortions then this dataset should give a close estimate of the family size. However, giving an exact copy number for genes in this family would require shotgun sequencing and assembly of the apple genome.

8.3 NBS-LRR transcriptome analysis

The use of GS20 (454 Life Sciences) sequencing of cDNA from infected and uninfected apple seedlings proved to be a very useful strategy in identifying genes being transcribed

following infection. Using these cDNA samples as templates in PCR amplification of RGAs facilitated for targeted sequencing of NBS-LRR genes that are either constitutively transcribed or those induced following pathogen invasion. This method has so far enabled the identification of thousands of novel transcripts in *Medicago truncatula* (Cheung *et al.*, 2006). Here it enabled identification of NBS-LRR genes that are either constitutively transcribed or those induced by apple scab and powdery mildew infections.

Assembling these sequences against those sequenced in section 3.2 facilitated the identification of candidate genes that may play a role in apple scab and powdery mildew disease resistance. The induced set had fewer genes as compared to those that are constitutively transcribed and most probably responsible for triggering the defence response in the presence of Avr factors. There was also a set of genes that were only detected in uninfected samples and not present in samples assayed after infection. These were assumed to represent genes that are either repressed following the infection or constitute a form of negative regulation mechanism that stops transcription of functional genes in the absence of infection possibly through RNAi mechanisms. Figure A1 (Appendix) shows an example of a contig that contains genes that were repressed following infection and these contain stop codons in the open reading frame as shown by accurate translation of the GLPL motif in the GS20 dataset.

A selected set of genes (contigs) identified using GS20 sequencing and sequence assemblies were analysed further with quantitative real-time PCR. Transcription of Ctg142 (a contig with sequences that appeared to be induced by pathogen infection) was

shown to be associated with seedlings that produce necrotic lesions following infection. According to the Chevalier scale, necrotic lesions are indicative of a hypersensitive response and as a result genes from this contig could be interesting to study in detail for the identification of genes of resistance. Seedling 19-G12 from the *Malus x domestica* (Borkh.) Carmine x Simpson cross was the only powdery mildew infected seedling that showed transcription of Ctg142 and consequently shows a resistance phenotype characterised by a hypersensitive response. Transcription of GD36 was only detected in apple scab infected class 1 seedlings, which represents plants resistant to apple scab. Transcription of AnRGA346 was common between scab and mildew infected seedlings. Although no conclusive evidence can be reached as to whether or not these genes are solely or partially responsible for the observed response to infection, the work sets up a possible method for the identification of genes responsible for disease resistance following infection with these two pathogens. Seedlings classified under class 2 according to the guidelines set up in the Chevalier scale showed total suppression of genes tested here including Ctg142 and GD36 that were transcribed in class 1 and to a lesser extent in class 4 seedlings. According to results of chapter 5, class 2 seedlings either express a different set of genes to classes 1 and 4 or further experimentation with a larger sample size is needed in future to study the genes responsible for the observed disease resistance phenotype.

Mildew infected seedlings that showed a total suppression of all the genes assayed here such as seedling 19-E12 belonged to a class that showed a generalised infection over the

whole plant although a few bottom leaves had more whitish lesions on the back of the leaves (section 5.2) which can be assumed to be total susceptibility to the infection.

More work will be done to test all the genes identified through GS20 sequencing and related sequence assemblies. This analysis and discussion does not infer absolute roles for NBS-LRR genes sequenced here because such detail would require functional studies. This work proves that high throughput sequencing of PCR amplification products generated through RT-PCR of RNA isolated before and after inoculating plants with pathogens can be a useful method of identifying genes that play a role in disease resistance. Identification of these genes in a segregating population could be used a strategy to identify and select for seedlings that can be propagated and used in breeding for resistance.

8.4 SNP genotyping in the NBS-LRR gene family

The RGA sequence data was also used to identify and map candidate SNPs in the NBS-LRR gene family. Analysis of data generated though re-sequencing of selected RGA clusters (section 4.4) provided preliminary evidence of the presence of candidate SNPs. Two of the three clusters selected for this analysis showed polymorphic SNPs and oligonucleotides designed to selectively amplify genes from these clusters were also used to amplify DNA fragments from a set of six *Malus x domestica* (Borkh.) Anna x Golden Delicious bin mapping seedlings (van Dyk *et al.*, In press). These PCR amplification products were purified to eliminate potential PCR artefacts and thereby ensure generation of reliable sequence data that could be analysed using sequence assemblies.

Sequence assemblies for data generated in this process (amplified using Anna3 and GD1 oligonucleotides) identified three segregating SNPs in cluster 11 of Figure 3.5 (GD1 oligonucleotide set). In the GD1 sequence assembly, only five of the six sequences from the bin mapping samples assembled in the same contig with sequences previously assembled in contig 1 (Figure 4.9). This suggests that SNP identification by sequencing could require the use of high throughput sequencing of PCR amplicons thereby circumvent complications brought about by DNA fragments containing mixed genes that present problems with Sanger sequencing. Sequence assemblies performed in section 4.5.1 showed that oligonucleotides designed to selectively amplify genes from individual clusters also amplified mixed genes. Sequencing of PCR products containing mixed genes gives sequences with high levels of ambiguous base calls and such data cannot be incorporated in sequence assemblies as shown in this case by the exclusion of the sequence from seedling 102. This problem was high in products produced using the Anna3 oligonucleotide pair in which only one sequence (seedling 51) assembled together with sequences of contig 7 (section 4.5.1).

Analysis of segregation and positioning of SNPs on an existing *Malus x domestica* (Borkh.) Anna x Golden Delicious genetic linkage map was performed using a set of fourteen bin mapping seedlings. This experiment was performed using SNaPshot™ assays and controls obtained from clones generated in chapter 3 (Figure 6.5). These assays proved to be more sensitive in detecting SNP genotypes for all seedlings used in the experiment as compared to Sanger sequencing of PCR amplicons. The parental genotypes for *Malus x domestica* (Borkh.) cvs Anna and Golden Delicious showed

polymorphisms in *Malus x domestica* (Borkh.) cv Anna only for all the three SNPs and thus all segregation in all the progeny were from that one parent. These SNPs were mapped to LG5 in the vicinity of microsatellite marker CH03a04 on the *Malus x domestica* (Borkh.) Anna x Golden Delicious genetic linkage map (van Dyk, *PhD thesis*, 2007). LG5 to which the GD1 SNPs were assigned was previously shown to either contain molecular markers tightly linked to disease resistance traits such as partial resistance to scab (CH03a04) and work by Calenge *et al*. (2005) also mapped an RGA marker (NBS2R11) in the same region through NBS-profiling.

8.4.1 High throughput sequencing and SNP mapping

High throughput sequencing and identification of SNPs through sequence assemblies using the fourteen bin mapping progeny of the *Malus x domestica* (Borkh.) Anna x Golden Delicious cross was performed. As discussed in Chapter 7 this process generated a complex dataset in which PCR amplification distortions were magnified. Consequently some of the sample specific datasets failed to produce representative sequence reads that would allow analysis of segregation patterns for the bin mapping progeny. This problem was exaggerated in the dataset produced for sample 12 of the mapping population whose sequence numbers were either non-existent or very low in contigs that contained candidate SNPs targeted for linkage mapping.

Pooling sequence reads from independent sequencing experiments to increase the numbers was not considered as an alternative. This was mostly because independent sequencing experiments from the same sample or a different PCR amplification merely

duplicates amplification distortions and produces a dataset biased in favour of genes with either a high copy number of PCR efficiency. This was confirmed by incorporating 1607 sequence reads from an additional sequencing experiment with the result that contigs with large numbers of reads were amplified further and little to no additions were observed for contigs with smaller numbers.

Datasets generated in this experiment confirm that the NBS-LRR family has a gene copy number in the range 230 to 300. Sequence assemblies for *Malus x domestica* (Borkh.) cultivars Anna, Golden Delicious and sample 16 of the bin mapping progeny with datasets in the order 6617, 10918 and 15023 gave contigs numbers of the order 211, 234 and 285 respectively. However, an increase in the number of sequences for these three samples achieved a reduction in the number of singletons from 995 (*Malus x domestica* (Borkh.) cv Anna), 481 (*Malus x domestica* (Borkh.) cv Golden Delicious and 413 (sample 16). Combined assemblies between two datasets achieved a further reduction in the unassembled (singleton) dataset. This could mean that the remaining genes that are currently not represented in these datasets are either genes with very low copy numbers or low PCR amplification efficiency.

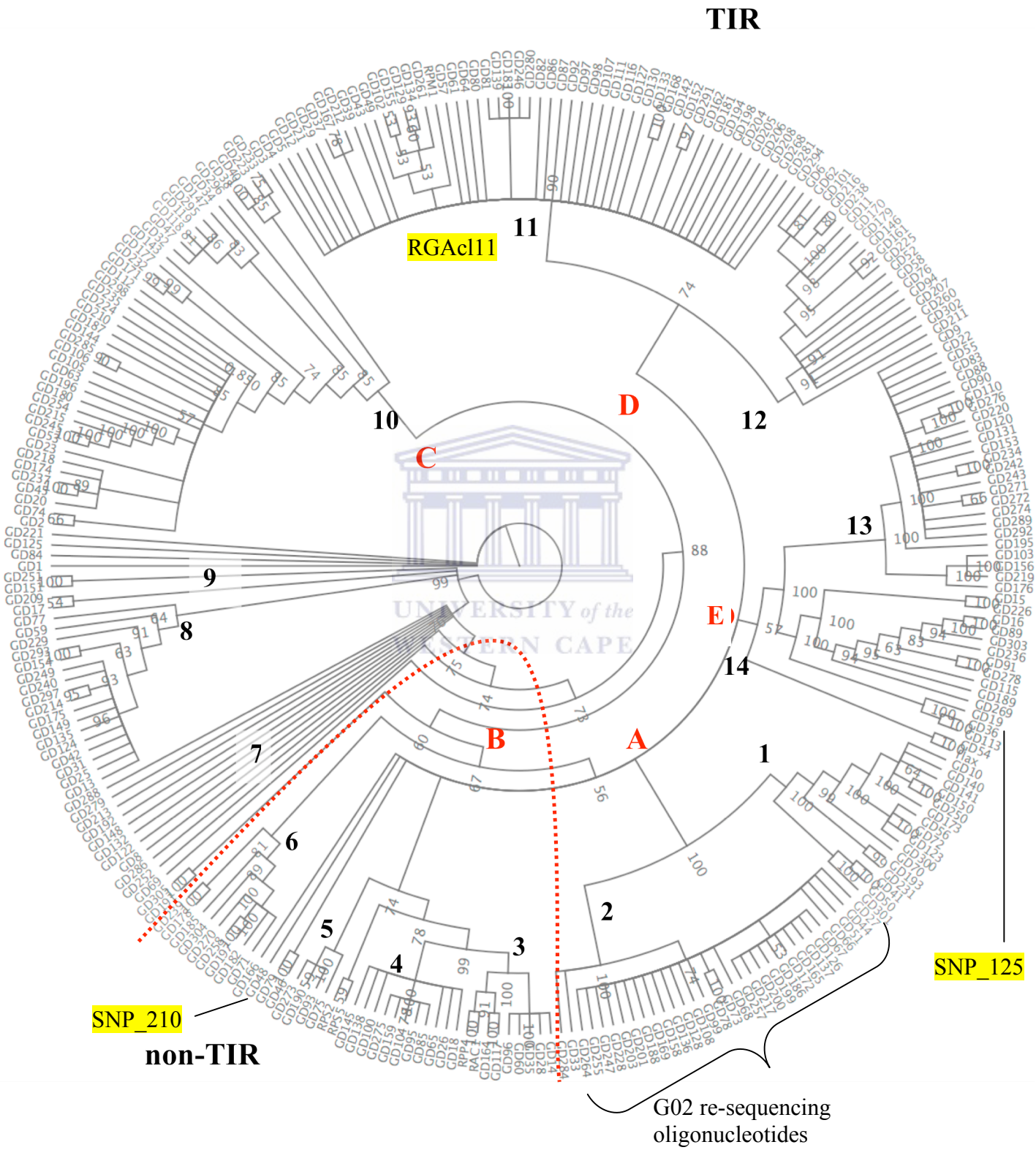
Although SNP mapping by sequencing generated a large number of candidate SNPs, mapping these with appreciable precision requires an alternative PCR-based method. The comparison between SNaPshot™ assays and high throughput sequencing shows that these two methods should be used together. High throughput sequencing could be used to identify candidate SNPs in the parental genotypes and assessment of the pattern of

segregation could then be performed using either SNaPshot™ assays or other PCR-based methods.

8.4.2 Comparative analysis of mapped RGAs

A total of nine SNPs have been mapped to eight linkage groups, LG 1, 2, 4, 5, 6, 12, 16 and 17. Three of the SNPs 017, 035 and 125 mapped to the same position on both LG1 and LG4 because the SSR markers CH02c02b (LG4) and Hi21g05 (LG1) share an identical segregation pattern using the 14 bin mapping seedlings used here. SNP_210 on LG17 is in reverse phase with alleles for seedlings 11 and 130 of SSR marker SA-A234 with the exception of missing data for seedling 353. Three of the nine SNPs have been placed on RGA phylogenetic tree (Figure 8.1) either due to sequence homology between members of the respective clusters and 454 sequence reads containing the SNPs or because the re-sequencing oligonucleotides were designed from alignments of produced using members of the respective clusters. More SNPs were identified from contigs that incorporated both 454 and Sanger data, however these either did not map to the available *Malus x domestica* (Borkh.) Anna x Golden Delicious genetic linkage map are not shown either due too much missing data (e.g. SNP_112 in cluster 13) or due to lack of matching SSR markers (e.g. SNP_006 in cluster 3). A comprehensive analysis of the SNP - cluster distribution is most probably going to be resolved by completion of the apple genome sequence.

Figure 8.1. Cluster distribution of the SNPs that have been mapped thus far. The SNPs highlighted in yellow are linked to the cluster from which they were produced. Cluster 2 was used to design the G02 oligonucleotide set used in Chapter 6.



A total of four SNPs were co-localized in bins to which other NBS markers have been mapped before (Figure 8.2). SNPs 017, 035 and 125 were localised to LG1 in the bin defined by SSR markers Hi21g05 (Silfverberg-Dilworth *et al.*, 2006) and CH03g12x (Liebhard *et al.*, 2003). No other NBS markers have been localised to this position, however LG1 does contain the NBS marker ARGH34 (Baldi *et al*., 2004) and an SSR marker CH-*Vf*1 that is tightly linked to the *Vf* gene (Vinatzer *et al.*, 2004).

SNP_105 was co-localised with sixteen NBS markers; ARGH37 and ARGH17 (Baldi *et al*., 2004) and NBS3M1-4, NBS2M9, NBS2M10, NBS2R9, NBS3M3, NBS2M4, NBS3M1b, NBS2M2, NBS2M3, NBS3M2, NBS2M20, NBS2M17, NBS2M7 and NBS3M19 (Calenge *et al.*, 2005) in the region where the *Vr*2 gene was mapped (Patocchi *et al.*, 2004) on LG2. LG2 also contains the SCAR marker *Vh*8 (Bus *et al.*, 2005) and *Vr (Patocchi et al., 2004)* and two more NBS markers ARGH46 (Baldi *et al*., 2004) and Pto-kin-x (Naik *et al.*, 2006) on the bottom half of this linkage group.

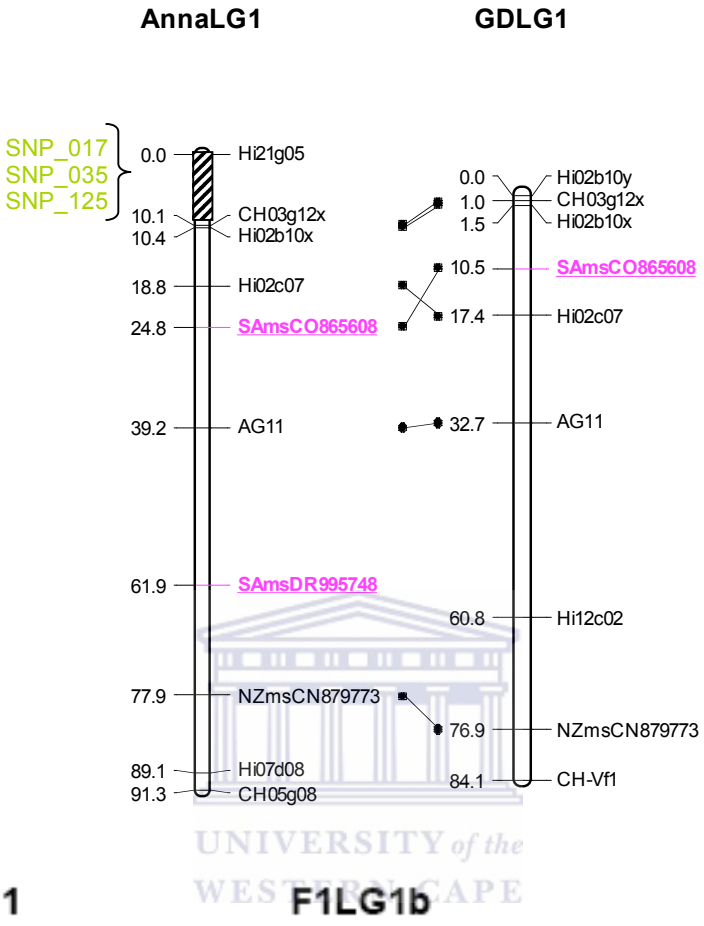
SNP RGAcl11 was co-localised with NBS2R11 (Calenge *et al*., 2005) and ARGH23b (Baldi *et al*., 2004) NBS markers on the bottom half of LG5. This part of LG5 contains a QTL linked to SSR marker CH04e03 that was shown to be associated with broad spectrum scab resistance (Calenge *et al.*, 2004). SNP_276 was localised on LG12 in the bin defined by SSR markers CH01d09 and CH01g12a (Liebhard *et al.*, 2003). This region contains both scab and mildew resistance genes *Vg* and *Pl-d* respectively located equidistant from SSR marker CH01g12 (James *et al.*, 2004; Erdin *et al.*, 2006). Two

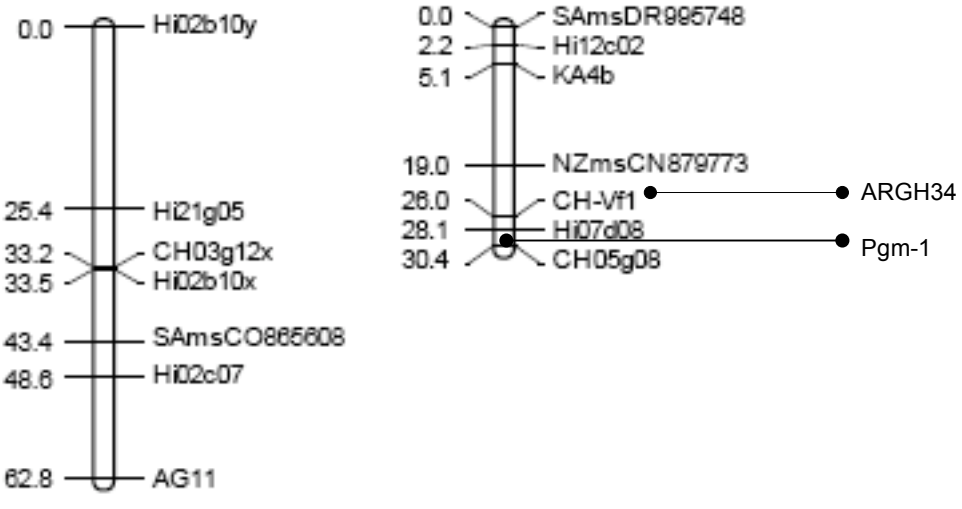
more NBS markers NLRR-INV and NLRR-INV-x (Naik *et al.*, 2006) were also mapped towards the bottom end of LG12.

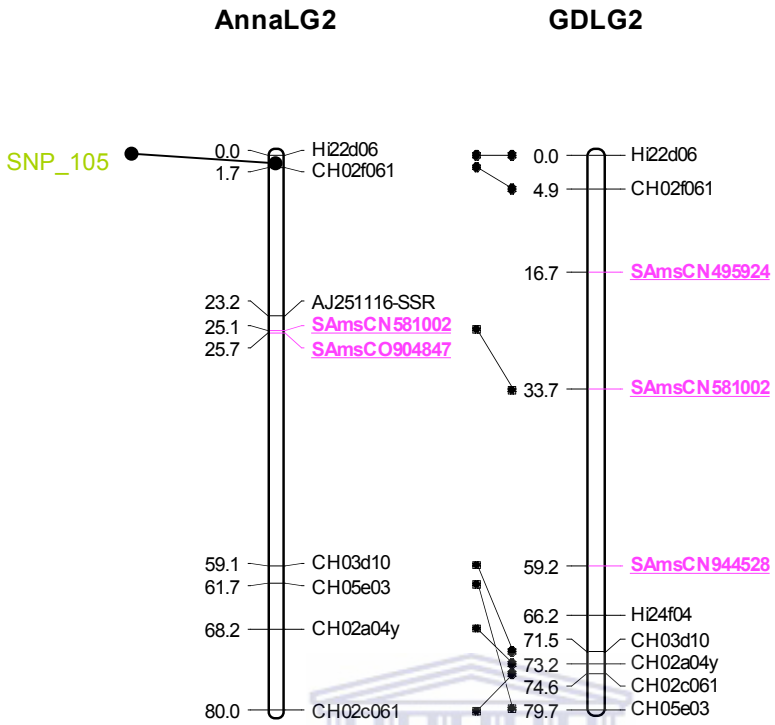
SNP_198 was co-localised to SSR marker Hi12a02 (Silfverberg-Dilworth *et al.*, 2006) on LG16. The NBS marker NBS3M17 (Calenge *et al*., 2005) was also mapped onto LG16 though on the bottom. SNP_210 was co-localised with SSR marker SA-A234 close to Hi07h02 (Silfverberg-Dilworth *et al.*, 2006). There are also four NBS markers ARGH30, ARGH02 (Baldi *et al*., 2004), NBS2M6 and NBSR16 mapped to the same region containing a QTL for powdery mildew resistance (Calenge *et al*., 2005).

The SNP mapping experiments were able to generate markers useful in identifying loci containing genes previously described in apple disease resistance. Three of these markers were also linked to RGA clusters in the phylogenetic tree (Figure 8.1). This provides a valuable resource in understanding the RGA cluster distribution with respect to NBS-LRR characterised genes. Such knowledge could be used to identify and characterise novel genes that offer new resistance specificities to apple pathogens.

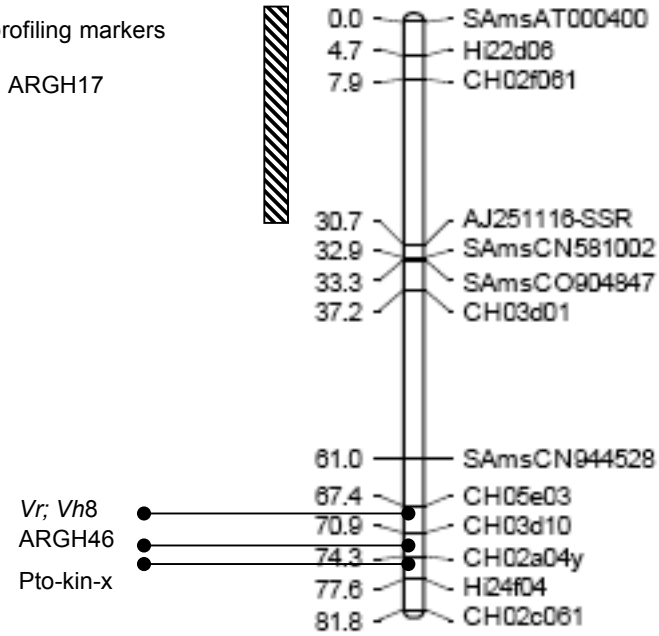
Figure 8.2. Review of *Malus x domestica* (Borkh.) Anna x Golden Delicious genetic linkage map against published markers. The SNPs from chapters 6 and 7 are shown in green on the AnnaLG and GDLG are parental linkage groups from *Malus x domestica* (Borkh.) cvs Anna and Golden Delicious respectively; a comparative analysis of the SNP positions to established markers are shown on the *Malus x domestica* (Borkh.) Anna x Golden Delicious integrated linkage groups 'F1LG'. The 'hatched bar' defines either the region or the bin to which markers have been located on the respective linkage groups. SNP_210 was placed on GDLG17 using the SSR marker SA-A234 mapped through the bin mapping approach. Integrated LGs 1 and 16 are broken due to lack of proper alignment and low marker resolution; only the portions with informative markers are shown here.

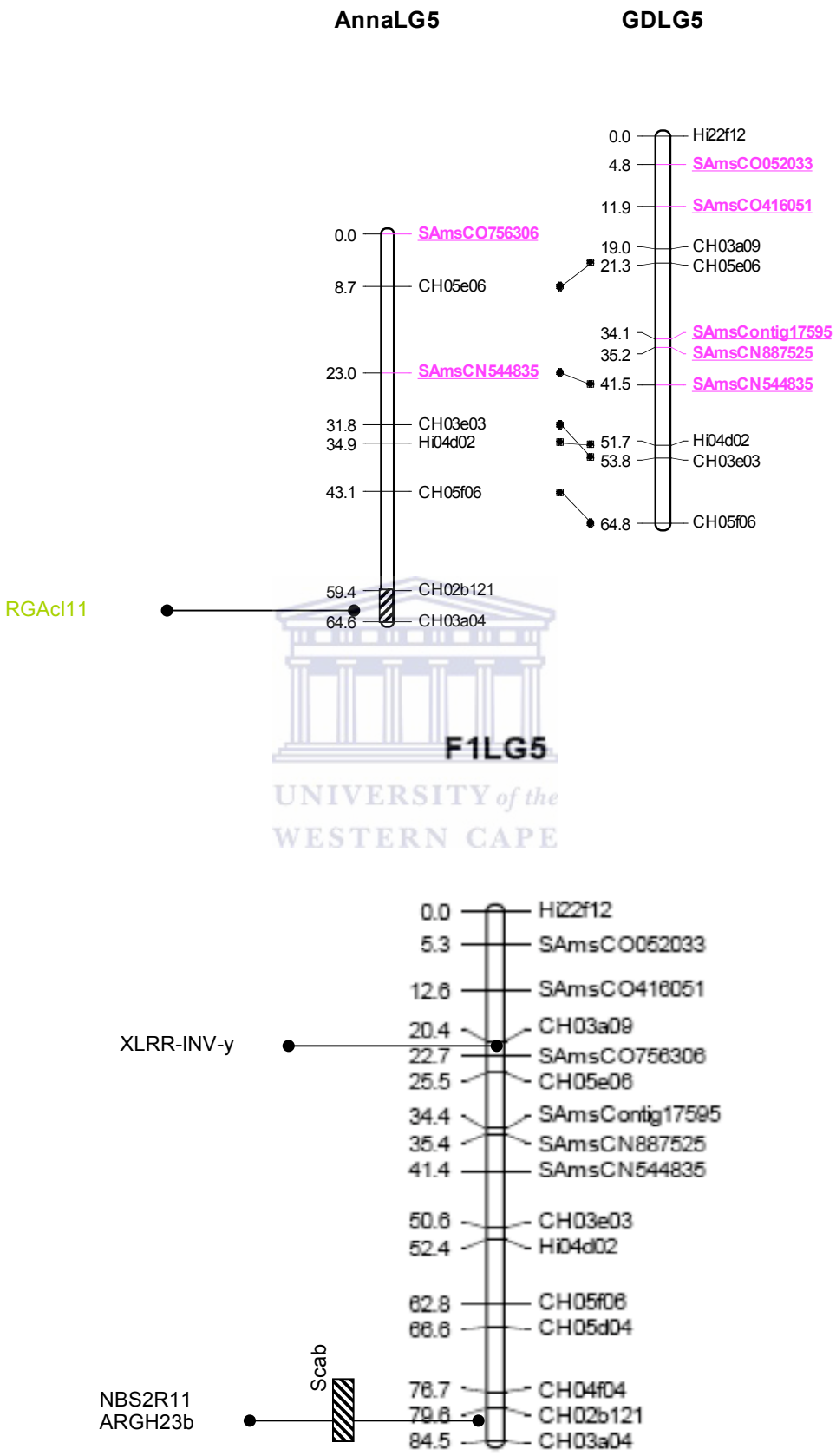


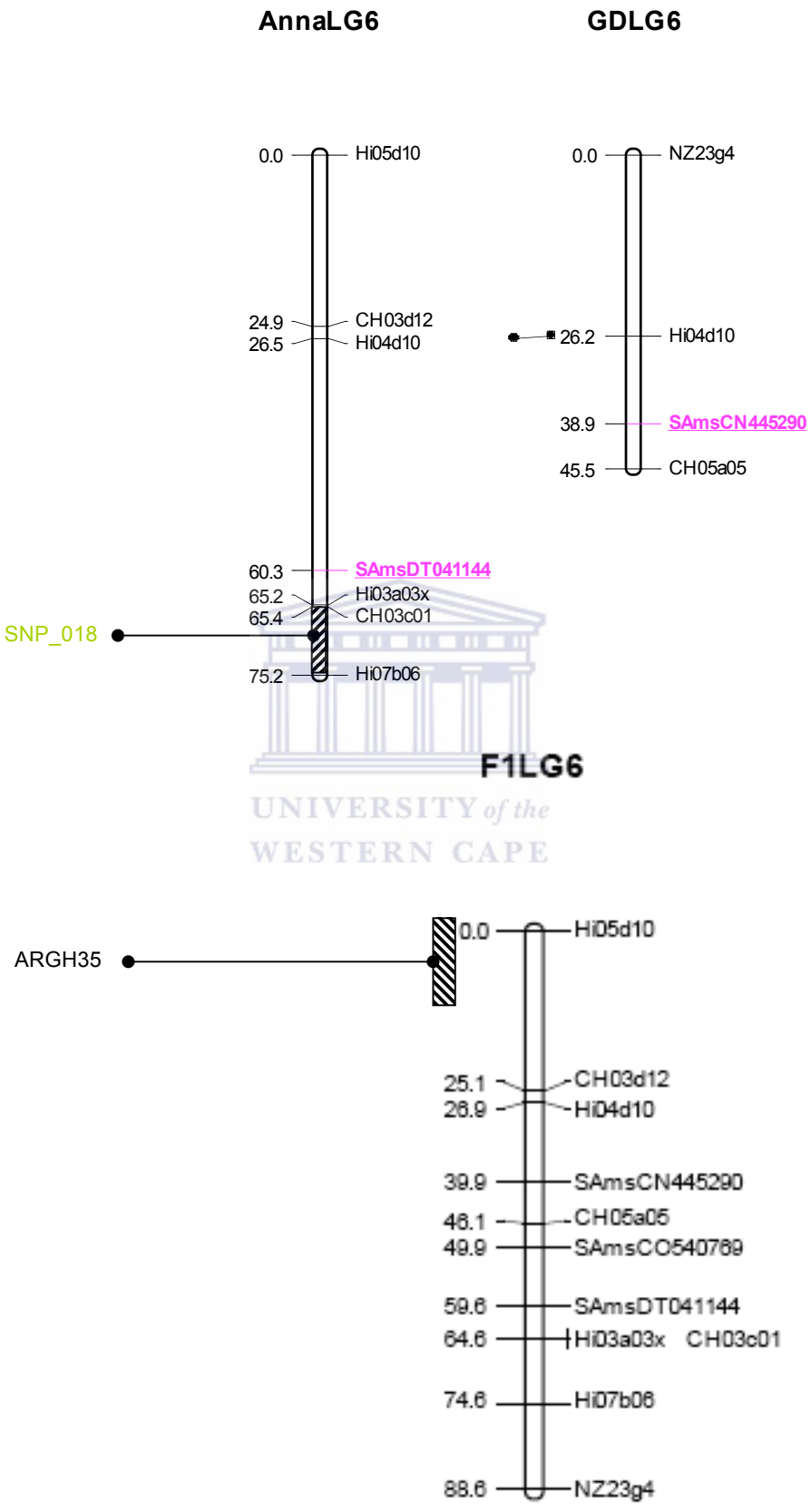


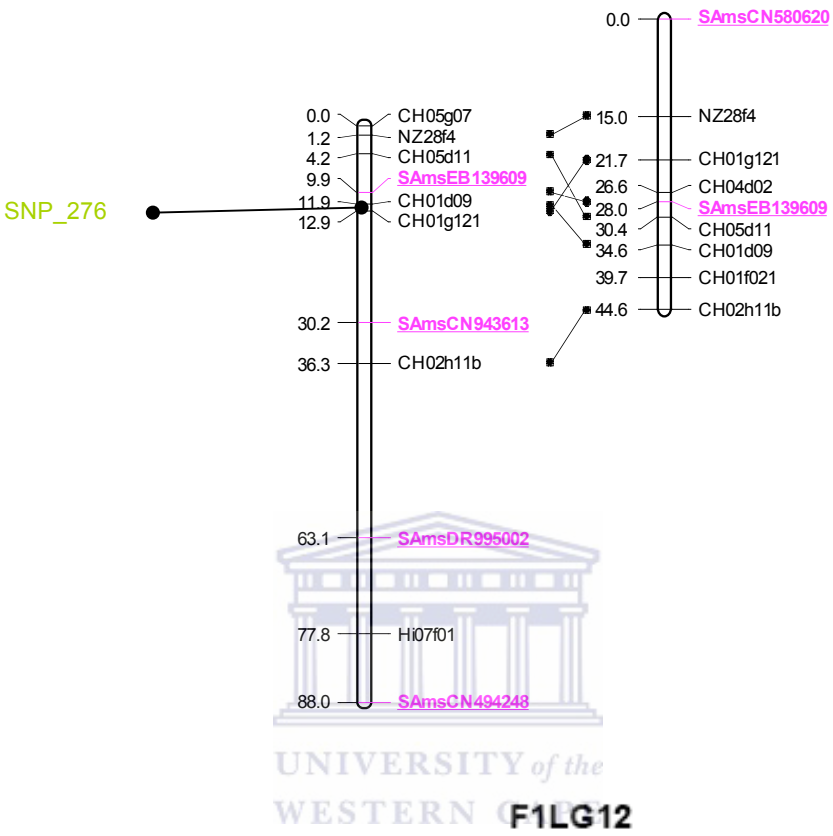


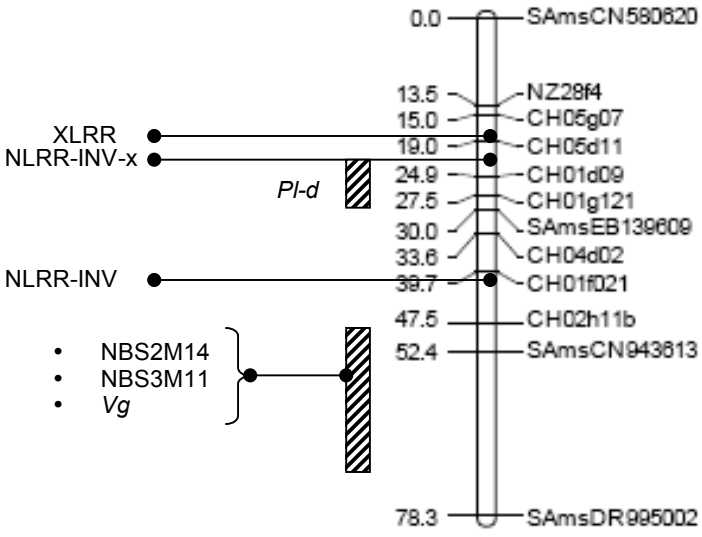
- 14 NBS profiling markers
- ARGH37; ARGH17
- *Vr*2

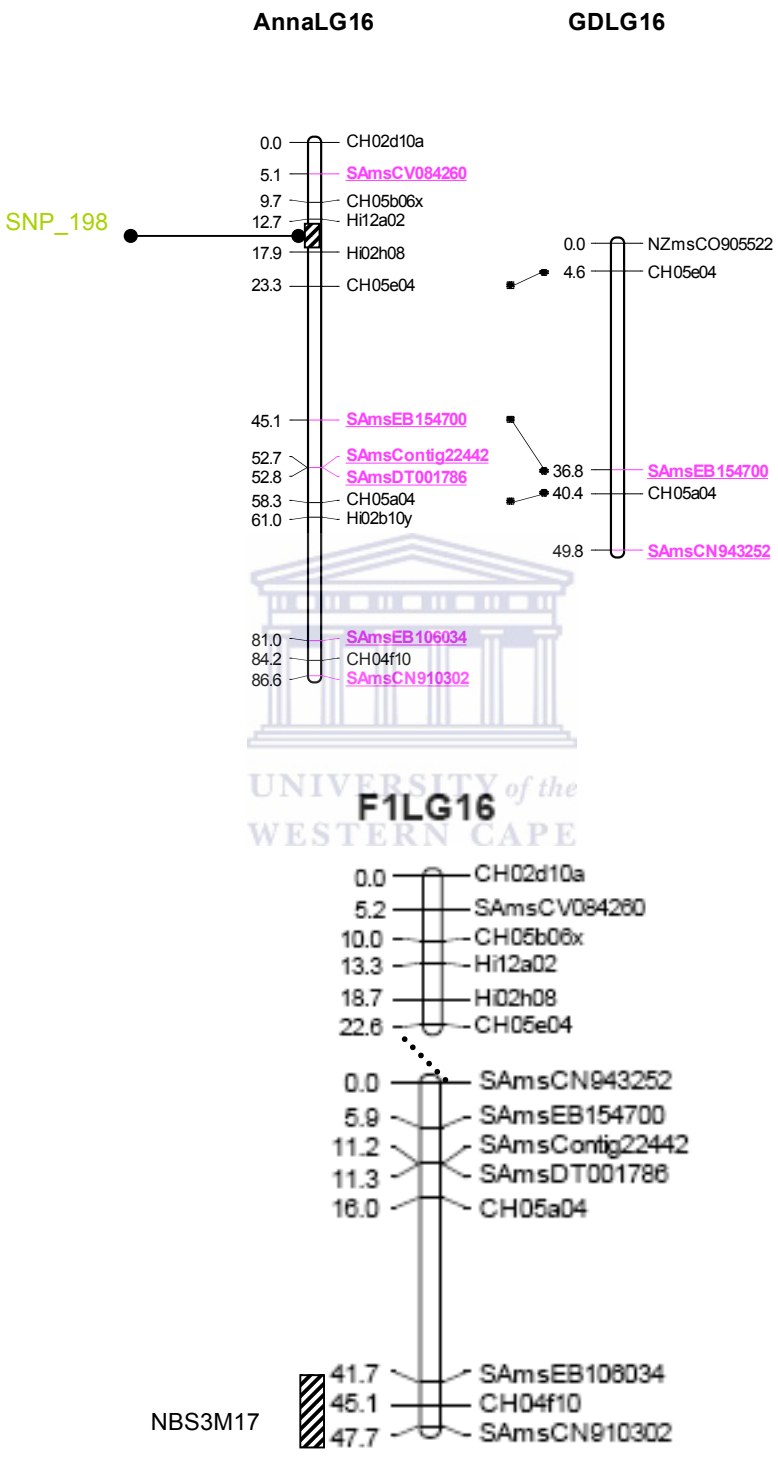


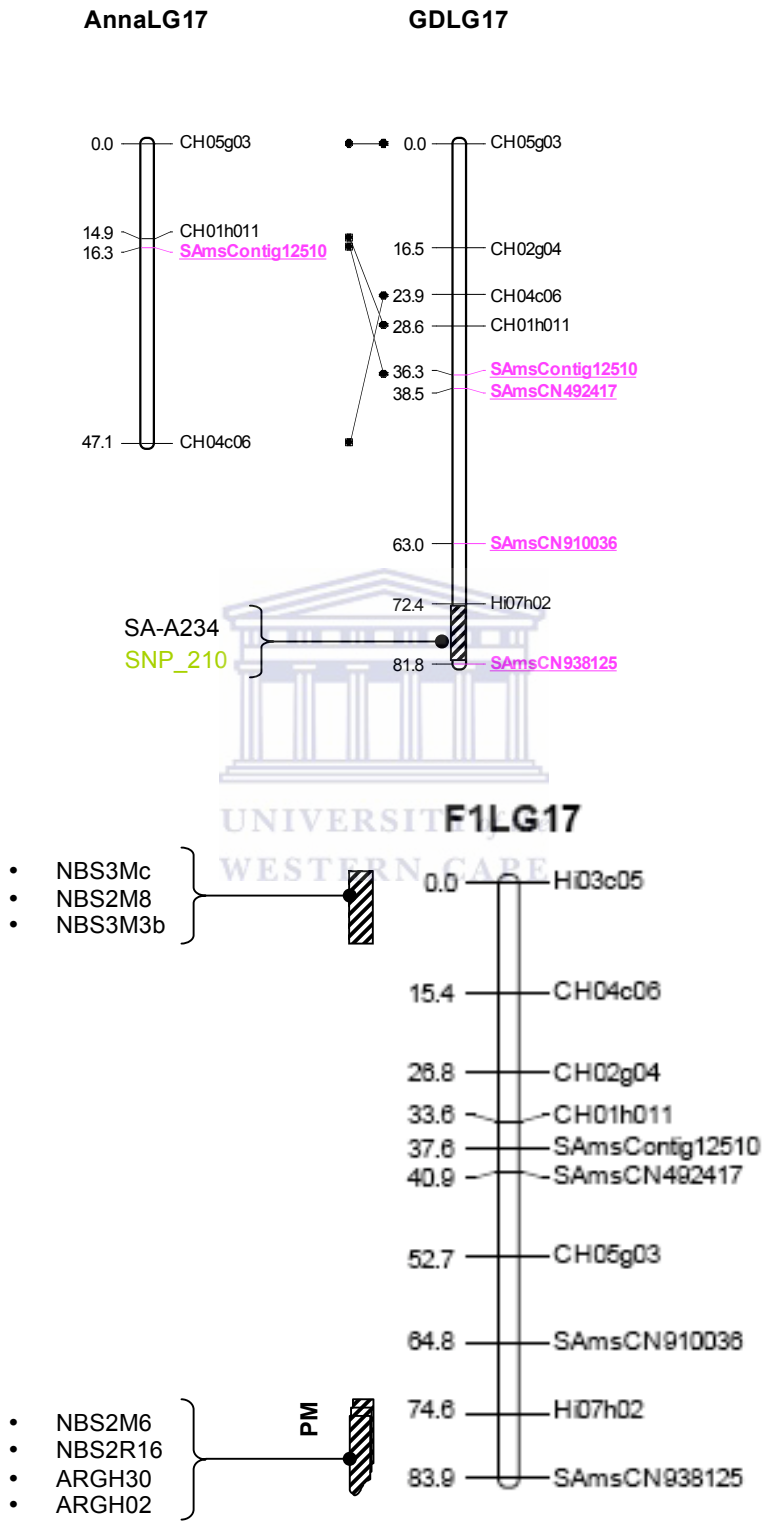





AnnaLG12
GDLG12
SNP_276
0.0 CH05g07
1.2 NZ28f4
4.2 CH05d11
9.9 SAmsEB139609
11.9 CH01d09
12.9 CH01g121
30.2 SAmsCN943613
36.3 CH02h11b
63.1 SAmsDR995002
77.8 Hi07f01
88.0 SAmsCN494248
0.0 SAmsCN580620
15.0 NZ28f4
21.7 CH01g121
26.6 CH04d02
28.0 SAmsEB139609
30.4 CH05d11
34.6 CH01d09
39.7 CH01f021
44.6 CH02h11b
F1LG12


XLRR
NLRR-INV-x
Pl-d
NLRR-INV
• NBS2M14
• NBS3M11
• Vg
0.0 SAmsCN580620
13.5 NZ28f4
15.0 CH05g07
19.0 CH05d11
24.9 CH01d09
27.5 CH01g121
30.0 SAmsEB139609
33.6 CH04d02
39.7 CH01f021
47.5 CH02h11b
52.4 SAmsCN943613
78.3 SAmsDR995002





CONCLUSION AND FUTURE RESEARCH

A number of techniques have been used in this work and valuable lessons have been learnt in the process. In designing future work it would be useful to follow up on the experiments carried out in this work and to make use of the valuable resources generated. Transcriptome analysis using high throughput sequencing of targeted NBS-LRR RNA transcripts following pathogen infections has generated a large sequence dataset with genes that could play a significant role in disease resistance. These would need to be analysed further using quantitative real-time PCR techniques on infected and uninfected samples. It would also be useful to use real-time PCR to understand the sequence of events from pathogen infection to generation of HR symptoms on the leaves. This might help to understand whether all functional NBS-LRR genes are R-genes or some might have a role in signal transduction or PR mechanisms. The presence of a large number of pseudogenes raises questions as to whether this is just a consequence of the evolution of the gene family or possibly transcription of such genes could also provide competitive gene regulation systems.

Some of the SNPs identified in chapter 7 could not be mapped using the bin mapping approach. These can be tested on a full mapping population together with SNP-like mutations identified in the Sanger sequenced dataset. This might resolve the problem of unmapped SNPs and possibly increase the number of clusters with mapped RGAs. Such a result is useful in describing the cluster distribution on the apple genome and also provides useful markers that can be used in the apple genome assembly.

REFERENCES

Aarts, N., Metz, M., Holub, E., Staskawicz, B. J., Daniels, M. J., and Parker, J. E. (1998). Different requirements for EDS1 and NDR1 by disease resistance genes define at least two R gene-mediated signaling pathways in *Arabidopsis*. *Proceedings of National Academy of Science U.S.A, 95*(17), 10306-10311.

AbuQamar, S., Chen, X., Dhawan, R., Bluhm, B., Salmeron, J., Lam, S., Dietrich, R. A., and Mengiste, T. (2006). Expression profiling and mutant analysis reveals complex regulatory networks involved in *Arabidopsis* response to *Botrytis* infection. *The Plant Journal, 48*(1).

Ade, J., DeYoung, B. J., Golstein, C., and Innes, R. W. (2007). Indirect activation of a plant nucleotide binding site-leucine-rich repeat protein by a bacterial protease. *Proceedings of National Academy of Science U.S.A, 104*(7), 2531-2536.

Agrawal, A. A. (2004). Plant defense and density dependence in the population growth of herbivores. *American Naturalist, 164*(1), 113-120.

Agrawal, S., and Kandimalla, E. R. (2004). Role of Toll-like receptors in antisense and siRNA [corrected]. *Nature Biotechnology, 22*(12), 1533-1537.

Agrawal, V., Zhang, C., Shapiro, A. D., and Dhurjati, P. S. (2004). A dynamic mathematical model to clarify signaling circuitry underlying programmed cell death control in *Arabidopsis* disease resistance. *Biotechnology Progress, 20*(2), 426-442.

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*(17), 3389-3402.

Anderson, P. A., Lawrence, G. J., Morrish, B. C., Ayliffe, M. A., Finnegan, E. J., and Ellis, J. G. (1997). Inactivation of the flax rust resistance gene M associated with loss of a repeated unit within the leucine-rich repeat coding region. *The Plant Cell, 9*(4), 641-651.

Attwood, T. K. (2002). The PRINTS database: a resource for identification of protein families. *Briefings in Bioinformatics, 3*(3), 252-263.

Attwood, T. K., and Beck, M. E. (1994). PRINTS--a protein motif fingerprint database. *Protein Engineering, 7*(7), 841-848.

Bae, H., Kim, M. S., Sicher, R. C., Bae, H. J., and Bailey, B. A. (2006). Necrosis- and ethylene-inducing peptide from *Fusarium oxysporum* induces a complex cascade of transcripts associated with signal transduction and cell death in *Arabidopsis*. *Plant Physiology, 141*(3), 1056-1067.

Bairoch, A. (1991). PROSITE: a dictionary of sites and patterns in proteins. *Nucleic Acids Research, 19 Suppl*, 2241-2245.

Bairoch, A. (1992). PROSITE: a dictionary of sites and patterns in proteins. *Nucleic Acids Research, 20 Suppl*, 2013-2018.

Bairoch, A. (1993). The PROSITE dictionary of sites and patterns in proteins, its current status. *Nucleic Acids Research, 21*(13), 3097-3103.

Balakirev, E. S., and Ayala, F. J. (2003). Pseudogenes: are they "junk" or functional DNA? *Annual Reviews of Genetics, 37*, 123-151.

Baldauf, S. L., Roger, A. J., Wenk-Siefert, I., and Doolittle, W. F. (2000). A kingdom-level phylogeny of eukaryotes based on combined protein data. *Science, 290*(5493), 972-977.

Baldi, P., Patocchi, A., Zini, E., Toller, C., Velasco, R., and Komjanc, M. (2004). Cloning and linkage mapping of resistance gene homologues in apple. *TAG Theoretical and Applied Genetics, 109*(1), 231-239.

Barton, N. H. (2000). Genetic hitchhiking. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences, 355*(1403), 1553-1562.

Baumgarten, A., Cannon, S., Spangler, R., and May, G. (2003). Genome-level evolution of resistance genes in *Arabidopsis thaliana*. *Genetics, 165*(1), 309-319.

Becraft, P. W. (2002). Receptor kinase signaling in plant development. *Annual Review of Cell and Developmental Biology, 18*(1), 163-192.

Belfanti, E., Silfverberg-Dilworth, E., Tartarini, S., Patocchi, A., Barbieri, M., Zhu, J., Vinatzer, B. A., Gianfranceschi, L., Gessler, C., and Sansavini, S. (2004). The HcrVf2 gene from a wild apple confers scab resistance to a transgenic cultivated variety. *Proceedings of the National Academy of Sciences, 101*(3), 886-890.

Belkhadir, Y., Subramaniam, R., and Dangl, J. L. (2004). Plant disease resistance protein signaling: NBS-LRR proteins and their partners. *Current Opinion in Plant Biology, 7*(4), 391-399.

Bendahmane, A., Kanyuka, K., and Baulcombe, D. C. (1999). The Rx gene from potato controls separate virus resistance and cell death responses. *The Plant Cell, 11*(5), 781-792.

Bergelson, J., Kreitman, M., Stahl, E. A., and Tian, D. (2001). Evolutionary dynamics of plant R-genes. *Science, 292*(5525), 2281-2285.

Bogdanove, A. J. (2002). Protein-protein interactions in pathogen recognition by plants. *Plant Molecular Biology, 50*(6), 981-989.

Bögre, L., Calderini, O., Binarova, P., Mattauch, M., Till, S., Kiegerl, S., Jonak, C., Pollaschek, C., Barker, P., and Huskisson, N. S. (1999). A MAP Kinase is activated late in plant mitosis and becomes localized to the plane of cell division. *The Plant Cell Online, 11*, 101-114.

Bolar, J. P., Norelli, J. L., Harman, G. E., Brown, S. K., and Aldwinckle, H. S. (2001). Synergistic activity of endochitinase and exochitinase from *Trichoderma atroviride* (*T. harzianum*) against the pathogenic fungus (*Venturia inaequalis*) in transgenic apple plants. *Transgenic Research, 10*(6), 533-543.

Bonfield, J. K., and Staden, R. (1996). Experiment files and their application during large-scale sequencing projects. *DNA Sequence, 6*(2), 109-117.

Botella, M. A., Parker, J. E., Frost, L. N., Bittner-Eddy, P. D., Beynon, J. L., Daniels, M. J., Holub, E. B., and Jones, J. D. (1998). Three genes of the *Arabidopsis* RPP1 complex resistance locus recognize distinct *Peronospora parasitica* avirulence determinants. *The Plant Cell, 10*(11), 1847-1860.

Boudichevskaia, A., Flachowsky, H., Peil, A., Fischer, C., and Dunemann, F. (2006). Development of a multiallelic SCAR marker for the scab resistance gene Vr1/Vh4/Vx from R12740-7A apple and its utility for molecular breeding. *Tree Genetics & Genomes, 2*(4), 186-195.

Bozkurt, O., Hakki, E. E., and Akkaya, M. S. (2007). Isolation and sequence analysis of wheat NBS-LRR type disease resistance gene analogs using degenerate PCR primers. *Biochemical Genetics, 45*(5-6), 469-486.

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

Brunner, K., Montero, M., Mach, R. L., Peterbauer, C. K., and Kubicek, C. P. (2003). Expression of the ech42 (endochitinase) gene of *Trichoderma atroviride* under carbon starvation is antagonized via a BrlA-like cis-acting element. *FEMS Microbiology Letters, 218*(2), 259-264.

Bryan, G. T., Wu, K. S., Farrall, L., Jia, Y., Hershey, H. P., McAdams, S. A., Faulk, K. N., Donaldson, G. K., Tarchini, R., and Valent, B. (2000). A single amino acid difference distinguishes resistant and susceptible alleles of the rice blast resistance gene Pi-ta. *The Plant Cell, 12*(11), 2033-2046.

Budak, H., Su, S., and Ergen, N. (2006). Revealing constitutively expressed resistance genes in *Agrostis* species using PCR-based motif-directed RNA fingerprinting. *Genetical Research, 88*(3), 165-175.

Bus, V., White, A., Gardiner, S., Weskett, R., Ranatunga, C., Samy, A., Cook, M., and Rikkerink, E. (2002). An update on apple scab resistance breeding in New Zealand. *Acta Horticulturae*, 43-48.

Bus, V. G., Laurens, F. N., van de Weg, W. E., Rusholme, R. L., Rikkerink, E. H., Gardiner, S. E., Bassett, H. C., Kodde, L. P., and Plummer, K. M. (2005). The Vh8 locus of a new gene-for-gene interaction between *Venturia inaequalis* and the wild apple *Malus sieversii* is closely linked to the Vh2 locus in *Malus pumila* R12740-7A. *The New Phytologist, 166*(3), 1035-1049.

Caldana, C., Scheible, W.-R., Mueller-Roeber, B., and Ruzicic, S. (2007). A quantitative RT-PCR platform for high-throughput expression profiling of 2500 rice transcription factors. *Plant Methods, 3*(1), 7.

Calenge, F., Faure, A., Goerre, M., Gebhardt, C., Van de Weg, W. E., Parisi, L., and Durel, C. E. (2004). Quantitative Trait Loci (QTL) analysis reveals both broad-spectrum and isolate-specific QTL for scab resistance in an apple progeny challenged with eight isolates of *Venturia inaequalis*. *Phytopathology, 94*(4), 370-379.

Calenge, F., Van der Linden, C. G., Van de Weg, E., Schouten, H. J., Van Arkel, G., Denance, C., and Durel, C. E. (2005). Resistance gene analogues identified through the NBS-profiling method map close to major genes and QTL for disease resistance in apple. *TAG Theoretical and Applied Genetics, 110*(4), 660-668.

Chamary, J. V., Parmley, J. L., and Hurst, L. D. (2006). Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nature Reviews: Genetics, 7*(2), 98-108.

Champion, A., Kreis, M., Mockaitis, K., Picaud, A., and Henry, Y. (2004). *Arabidopsis* kinome: after the casting. *Functional and Integrative Genomics, 4*(3), 163-187.

Chan, M. C., Sung, J. J., Lam, R. K., Chan, P. K., Lai, R. W., and Leung, W. K. (2006). Sapovirus detection by quantitative real-time RT-PCR in clinical stool specimens. *Journal of Virological Methods, 134*(1-2), 146-153.

Chaw, S. M., Chang, C. C., Chen, H. L., and Li, W. H. (2004). Dating the monocot-dicot divergence and the origin of core eudicots using whole chloroplast genomes. *Journal of Molecular Evolution, 58*(4), 424-441.

Cheung, F., Haas, B. J., Goldberg, S. M., May, G. D., Xiao, Y., and Town, C. D. (2006). Sequencing *Medicago truncatula* expressed sequenced tags using 454 Life Sciences technology. *BMC Genomics, 7*, 272.

Chevalier, M., Lespinasse, Y., and Renaudin, S. (1991). A microscopic study of the different classes of symptoms coded by the Vf gene in apple for resistance to scab (*Venturia inaequalis*). *Plant Pathology, 40*, 249-256.

Chini, A., and Loake, G. J. (2005). Motifs specific for the ADR1 NBS-LRR protein family in *Arabidopsis* are conserved among NBS-LRR sequences from both dicotyledonous and monocotyledonous plants. *Planta, 221*(4), 597-601.

Chini, A., and Loake, G. J. (2005). Motifs specific for the ADR1 NBS–LRR protein family in *Arabidopsis* are conserved among NBS–LRR sequences from both dicotyledonous and monocotyledonous plants. *Planta, 221*(4), 597-601.

Clamp, M., Cuff, J., Searle, S. M., and Barton, G. J. (2004). The Jalview Java alignment editor. *Bioinformatics, 20*(3), 426-427.

Collmer, A., Badel, J. L., Charkowski, A. O., Deng, W. L., Fouts, D. E., Ramos, A. R., Rehm, A. H., Anderson, D. M., Schneewind, O., and van Dijk, K. (2000). Colloquium Paper *Pseudomonas syringae* Hrp type III secretion system and effector proteins. *Proceedings of the National Academy of Science U.S.A, 97*(16), 8770-8777.

Couch, B. C., Spangler, R., Ramos, C., and May, G. (2006). Pervasive purifying selection characterizes the evolution of I2 homologs. *Molecular Plant-Microbe Interactions, 19*(3), 288-303.

Cutt, J. R., and Klessig, D. F. (1992). Pathogenesis-related proteins. *Genes Involved in Plant Defense. T. Boller and F. Meins, eds. Springer-Verlag, Vienna*, 209-243.

D. Peter Tieleman, I. H. S. M. R. U. M. S. P. S. (2001). Proline-induced hinges in transmembrane helices: Possible roles in ion channel gating. *Proteins: Structure, Function, and Genetics, 44*(2), 63-72.

Daxberger, A., Nemak, A., Mithofer, A., Fliegmann, J., Ligterink, W., Hirt, H., and Ebel, J. (2007). Activation of members of a MAPK module in beta-glucan elicitor-mediated non-host resistance of soybean. *Planta, 225*(6), 1559-1571.

De Mita, S., Santoni, S., Hochu, I., Ronfort, J., and Bataillon, T. (2006). Molecular evolution and positive selection of the symbiotic gene NORK in *Medicago truncatula*. *Journal of Molecular Evolution, 62*(2), 234-244.

Decreux, A., and Messiaen, J. (2005). Wall-associated kinase WAK1 interacts with cell wall pectins in a calcium-induced conformation. *Plant and Cell Physiology, 46*(2), 268-278.

Decroocq, V., Fave, M. G., Hagen, L., Bordenave, L., and Decroocq, S. (2003). Development and transferability of apricot and grape EST microsatellite markers across taxa. *TAG Theoretical and Applied Genetics, 106*(5), 912-922.

Delledonne, M., Zeier, J., Marocco, A., and Lamb, C. (2001). Signal interactions between nitric oxide and reactive oxygen intermediates in the plant hypersensitive disease resistance response. *Proceedings of the National Academy of Sciences*, 231178298.

Desaki, Y., Miya, A., Venkatesh, B., Tsuyumu, S., Yamane, H., Kaku, H., Minami, E., and Shibuya, N. (2006). Bacterial lipopolysaccharides induce defense responses associated with programmed cell death in rice cells. *Plant and Cell Physiology, 47*(11), 1530-1540.

DeYoung, B. J., and Innes, R. W. (2006). Plant NBS-LRR proteins in pathogen sensing and host defense. *Nature Immunology, 7*(12), 1243-1249.

Dinesh-Kumar, S. P., Tham, W. H., and Baker, B. J. (2000). Structure-function analysis of the tobacco mosaic virus resistance gene N. *Proceedings of the National Academy of Sciences, 97*(26), 14789.

Dodds, P. N., Lawrence, G. J., and Ellis, J. G. (2001). Six amino acid changes confined to the leucine-rich repeat beta-strand/beta-turn motif determine the difference between the P and P2 rust resistance specificities in flax. *The Plant Cell, 13*(1), 163-178.

Dong, X. (2001). Genetic dissection of systemic acquired resistance. *Current Opinion in Plant Biology, 4*(4), 309-314.

Dorey, S., Baillieul, F., Pierrel, M. A., Saindrenan, P., Fritig, B., and Kauffmann, S. (1997). Spatial and temporal induction of cell death, defense genes, and accumulation of salicylic acid in tobacco leaves reacting hypersensitively to a fungal glycoprotein elicitor. *Molecular Plant-Microbe Interactions, 10*, 646-655.

Dunemann, F., Bräcker, G., Markussen, T., and Roche, P. (1996). *Identification of molecular markers for the major mildew resistance gene Pl2 in apple.* Paper presented at the Eucarpia Symposium on Fruit Breeding and Genetics.

Durrant, W. E., and Dong, X. (2004). Systemic acquired resistance. *Annual Review of Phytopathology, 42*, 185-209.

Ellis, J., Lawrence, G., Ayliffe, M., Anderson, P., Collins, N., Finnegan, J., Frost, D., Luck, J., and Pryor, T. (1997). Advances in the molecular genetic analysis of the Flax-Flax rust interaction. *Annual Review of Phytopathology, 35*(1), 271-291.

Erdin, N., Tartarini, S., Broggini, G. A. L., Gennari, F., Sansavini, S., Gessler, C., and Patocchi, A. (2006). Mapping of the apple scab-resistance gene Vb. *Genome, 49*(10), 1238-1245.

Eulgem, T. (2005). Regulation of the *Arabidopsis* defense transcriptome. *Trends in Plant Science, 10*(2), 71-78.

Eulgem, T. (2006). Dissecting the WRKY web of plant defense regulators. *PLoS Pathogens, 2*(11), e126.

Eulgem, T., Rushton, P. J., Robatzek, S., and Somssich, I. E. (2000). The WRKY superfamily of plant transcription factors. *Trends in Plant Science, 5*(5), 199-206.

Eulgem, T., Rushton, P. J., Schmelzer, E., Hahlbrock, K., and Somssich, I. E. (1999). Early nuclear events in plant defence signalling: rapid gene activation by WRKY transcription factors. *The EMBO Journal, 18*, 4689-4699.

Evans, K., and James, C. (2003). Identification of SCAR markers linked to Pl-w mildew resistance in apple. *TAG Theoretical and Applied Genetics, 106*(7), 1178-1183.

Fields, S., and Song, O. (1989). A novel genetic system to detect protein-protein interactions. *Nature, 340*(6230), 245-246.

Flor, H. H. (1971). Current status of the gene-for-gene concept. *Annual Review of Phytopathology, 9*(1), 275-296.

Fluhr, R. (2001). Sentinels of disease. Plant resistance genes. *Plant Physiology, 127*(4), 1367-1374.

Fredman, D., White, S. J., Potter, S., Eichler, E. E., Den Dunnen, J. T., and Brookes, A. J. (2004). Complex SNP-related sequence variation in segmental genome duplications. *Nature Genetics, 36*, 861-866.

Gardiner, S. E., Bus, V. G. M., Rusholme, R. L., Chagne, D., and Rikkerink, E. H. A. (2007). Apple. In C. Kole (Ed.), *Genome Mapping and Molecular Breeding in Plants* (Vol. 4, pp. 1 - 61). Berlin Heidelberg: Springer-Verlag.

Gessler, C., and Patocchi, A. (2007). Recombinant DNA technology in apple. *Advances in Biochemical Engineering and Biotechnology, 107*, 113-132.

Glazebrook, J. (2001). Genes controlling expression of defense responses in *Arabidopsis*—2001 status. *Current Opinion in Plant Biology, 4*(4), 301-308.

Gowda, B., Miller, J., Rubin, S., Sharma, D., and Timko, M. (2002). Isolation, sequence analysis, and linkage mapping of resistance-gene analogs in cowpea (*Vigna unguiculata* L. Walp.). *Euphytica, 126*(3), 365-377.

Gygax, M., Gianfranceschi, L., Liebhard, R., Kellerhals, M., Gessler, C., and Patocchi, A. (2004). Molecular markers linked to the apple scab resistance gene Vbj derived from *Malus baccata* jackii. *TAG Theoretical and Applied Genetics, 109*(8), 1702-1709.

Halbwirth, H., Fischer, T. C., Roemmelt, S., Spinelli, F., Schlangen, K., Peterek, S., Sabatini, E., Messina, C., Speakman, J. B., and Andreotti, C. (2003). Induction of antimicrobial 3-deoxyflavonoids in pome fruit trees controls fire blight. *Zeitschrift fur Naturforschung. Section C, Biosciences, 58*(11/12), 765-770.

Halterman, D. A., and Wise, R. P. (2001). A single-amino acid substitution in the sixth leucine-rich repeat of barley MLA6 and MLA13 alleviates dependence on RAR1 for disease resistance signaling. *The Plant Journal, 38*(2).

Ham, J. H., Kim, M. G., Lee, S. Y., and Mackey, D. (2007). Layered basal defenses underlie non-host resistance of *Arabidopsis* to *Pseudomonas syringae* pv. *phaseolicola*. *The Plant Journal, 51*(4), 604-616.

Hammond-Kosack, K. E., and Jones, J. D. (1996). Resistance gene-dependent plant defense responses. *The Plant Cell, 8*(10), 1773-1791.

Hasegawa, M., Kishino, H., and Yano, T.-a. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution, 22*(2), 160-174.

Hauck, P., Thilmony, R., and He, S. Y. (2003). A *Pseudomonas syringae* type III effector suppresses cell wall-based extracellular defense in susceptible *Arabidopsis* plants. *Proceedings of the National Academy of Sciences U.S.A, 100*(14), 8577-8582.

Hayes, A. J., and Saghai Maroof, M. A. (2000). Targeted resistance gene mapping in soybean using modified AFLPs. *TAG Theoretical and Applied Genetics, 100*(8), 1279-1283.

He, X. J., Zhang, Z. G., Yan, D. Q., Zhang, J. S., and Chen, S. Y. (2004). A salt-responsive receptor-like kinase gene regulated by the ethylene signaling pathway encodes a plasma membrane serine/threonine kinase. *TAG Theoretical and Applied Genetics, 109*(2), 377-383.

He, Z. H., He, D., and Kohorn, B. D. (1998). Requirement for the induced expression of a cell wall associated receptor kinase for survival during the pathogen response. *The Plant Journal, 14*(1), 55-63.

Hennin, C., Hofte, M., and Diederichsen, E. (2001). Functional expression of Cf9 and Avr9 genes in *Brassica napus* induces enhanced resistance to *Leptosphaeria maculans*. *Molecular Plant-Microbe Interactions, 14*(9), 1075-1085.

Henry, Y., Bedhomme, M., and Blanc, G. (2006). History, protohistory and prehistory of the *Arabidopsis thaliana* chromosome complement. *Trends in Plant Science, 11*(6), 267-273.

Hirotsune, S., Yoshida, N., Chen, A., Garrett, L., Sugiyama, F., Takahashi, S., Yagami, K., Wynshaw-Boris, A., and Yoshiki, A. (2003). An expressed pseudogene regulates the messenger-RNA stability of its homologous coding gene. *Nature, 423*(6935), 91-96.

Hishida, T., Iwasaki, H., Yagi, T., and Shinagawa, H. (1999). Role of Walker motif A of RuvB protein in promoting branch migration of holliday junctions: Walker motif A mutations affect ATP binding, ATP hydrolyzing, and DNA binding activities of RuvB. *Journal of Biological Chemistry, 274*(36), 25335-25342.

Hoeberichts, F. A., and Woltering, E. J. (2003). Multiple mediators of plant programmed cell death: interplay of conserved cell death mechanisms and plant-specific regulators. *Bioessays, 25*(1), 47-57.

Howad, W., Yamamoto, T., Dirlewanger, E., Testolin, R., Cosson, P., Cipriani, G., Monforte, A. J., Georgi, L., Abbott, A. G., and Arus, P. (2005). Mapping with a few plants: Using selective mapping for microsatellite saturation of the *Prunus* reference map. *Genetics, 171*(3), 1305-1309.

Hu, G., deHart, A. K., Li, Y., Ustach, C., Handley, V., Navarre, R., Hwang, C. F., Aegerter, B. J., Williamson, V. M., and Baker, B. (2005). EDS1 in tomato is required for resistance mediated by TIR-class R genes and the receptor-like R gene Ve. *The Plant Journal, 42*(3), 376-391.

Hu, G., deHart, A. K. A., Li, Y., Ustach, C., Handley, V., Navarre, R., Hwang, C.-F., Aegerter, B. J., Williamson, V. M., and Baker, B. (2005). EDS1 in tomato is required for resistance mediated by TIR-class R genes and the receptor-like R gene Ve. *The Plant Journal, 42*(3).

Hu, Y., Benedict, M. A., Ding, L., and Nunez, G. (1999). Role of cytochrome c and dATP/ATP hydrolysis in Apaf-1-mediated caspase-9 activation and apoptosis. *The EMBO Journal, 18*(13), 3586-3595.

Hu, Y., Ding, L., Spencer, D. M., and Nunez, G. (1998). WD-40 repeat region regulates Apaf-1 self-association and procaspase-9 activation. *The Journal of Biological Chemistry, 273*(50), 33489-33494.

Hudson, R. R., Kreitman, M., and Aguade, M. (1987). A test of neutral molecular evolution based on nucleotide data. *Genetics, 116*(1), 153-159.

Huelsenbeck, J. P., and Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics, 17*(8), 754-755.

Hurles, M. (2002). Are 100, 000(quote) SNPs (quote) Useless? *Science, 298*(5598), 1509-1509.

Hwang, C. F., Bhakta, A. V., Truesdell, G. M., Pudlo, W. M., and Williamson, V. M. (2000). Evidence for a role of the N terminus and leucine-rich repeat region of the

Mi gene product in regulation of localized cell death. *The Plant Cell, 12*(8), 1319-1329.

Iskandar, H. M., Simpson, R. S., Casu, R. E., Bonnett, G. D., Maclean, D. J., and Manners, J. M. (2004). Comparison of reference genes for quantitative real-time polymerase chain reaction analysis of gene expression in sugarcane. *Plant Molecular Biology Reporter, 22*(4), 325-337.

Jain, M., Nijhawan, A., Tyagi, A. K., and Khurana, J. P. (2006). Validation of housekeeping genes as internal control for studying gene expression in rice by quantitative real-time PCR. *Biochemical and Biophysical Research Communications, 345*(2), 646-651.

James, C. M., Clarke, J. B., and Evans, K. M. (2004). Identification of molecular markers linked to the mildew resistance gene Pl-d in apple. *TAG Theoretical and Applied Genetics, 110*(1), 175-181.

Jensen, R. A. (2001). Orthologs and paralogs - we need to get it right. *Genome Biology, 2*(8), INTERACTIONS1002.

Jiang, X., and Wang, X. (2000). Cytochrome c promotes caspase-9 activation by inducing nucleotide binding to Apaf-1. *The Journal of Biological Chemistry, 275*(40), 31199-31203.

Jinn, T. L., Stone, J. M., and Walker, J. C. (2000). HAESA, an Arabidopsis leucine-rich repeat receptor kinase, controls floral organ abscission. *Genes and Development, 14*(1), 108-117.

Jobes, D. V., Hurley, D. L., and Thien, L. B. (1995). Plant DNA Isolation: A method to efficiently remove polyphenolics, polysaccharides, and RNA. *Taxon, 44*(3), 379-386.

Jones, D. T., Taylor, W. R., and Thornton, J. M. (1992). The rapid generation of mutation data matrices from protein sequences. *Computer Applications in the Biosciences, 8*(3), 275-282.

Jones, J. D. (2001). Putting knowledge of plant disease resistance genes to work. *Current Opinion in Plant Biology, 4*(4), 281-287.

Kamoun, S. (2001). Nonhost resistance to *Phytophthora*: novel prospects for a classical problem. *Current Opinion in Plant Biology, 4*(4), 295-300.

Khan, M., Duffy, B., Gessler, C., and Patocchi, A. (2006). QTL mapping of fire blight resistance in apple. *Molecular Breeding, 17*(4), 299-306.

Khan, M. A., Durel, C. E., Duffy, B., Drouet, D., Kellerhals, M., Gessler, C., and Patocchi, A. (2007). Development of molecular markers linked to the Fiesta linkage group 7 major QTL for fire blight resistance and their application for marker-assisted selection. *Genome, 50*(6), 568-577.

Kimura, M. (1991). The neutral theory of molecular evolution: a review of recent evidence. *Japanese Journal of Genetics, 66*(4), 367-386.

Kishimoto, A., Nishiyama, K., Nakanishi, H., Uratsuji, Y., Nomura, H., Takeyama, Y., and Nishizuka, Y. (1985). Studies on the phosphorylation of myelin basic protein by protein kinase C and adenosine 3':5'-monophosphate-dependent protein kinase. *The Journal of Biological Chemistry, 260*(23), 12492-12499.

Kobe, B., and Deisenhofer, J. (1995). A structural basis of the interactions between leucine-rich repeats and protein ligands. *Nature, 374*(6518), 183-186.

Koch, K. (2004). Sucrose metabolism: regulatory mechanisms and pivotal roles in sugar sensing and plant development. *Current Opinion in Plant Biology, 7*(3), 235-246.

Kohorn, B. D., Kobayashi, M., Johansen, S., Riese, J., Huang, L. F., Koch, K., Fu, S., Dotson, A., and Byers, N. (2006). An *Arabidopsis* cell wall-associated kinase required for invertase activity and cell growth. *The Plant Journal, 46*(2), 307-316.

Koonin, E. V. (2005). Orthologs, paralogs, and evolutionary genomics. *Annual Review of Genetics, 39*, 309-338.

Kruijt, M., De Kock, M. J. D., and De Wit, P. J. G. M. (2005). Receptor-like proteins involved in plant disease resistance. *Molecular Plant Pathology, 6*(1), 85-97.

Krzymowska, M., Konopka-Postupolska, D., Sobczak, M., Macioszek, V., Ellis, B. E., and Hennig, J. (2007). Infection of tobacco with different *Pseudomonas syringae* pathovars leads to distinct morphotypes of programmed cell death. *The Plant Journal, 50*(2), 253-264.

Kuang, H., Woo, S. S., Meyers, B. C., Nevo, E., and Michelmore, R. W. (2004). Multiple genetic processes result in heterogeneous rates of evolution within the major cluster disease resistance genes in lettuce. *The Plant Cell, 16*(11), 2870-2894.

Kumar, S., Tamura, K., and Nei, M. (2004). MEGA3: Integrated software for molecular evolutionary genetics analysis and sequence alignment. *Briefings in Bioinformatics, 5*(2), 150-163.

Kwiatowski, J., Skarecky, D., Hernandez, S., Pham, D., Quijas, F., and Ayala, F. J. (1991). High fidelity of the polymerase chain reaction. *Molecular biology and evolution, 8*(6), 884-887.

Kyte, J., and Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology, 157*(1), 105-132.

Lalli, D. A., Decroocq, V., Blenda, A. V., Schurdi-Levraud, V., Garay, L., Le Gall, O., Damsteegt, V., Reighard, G. L., and Abbott, A. G. (2005). Identification and mapping of resistance gene analogs (RGAs) in *Prunus*: a resistance map for *Prunus*. *TAG Theoretical and Applied Genetics, 111*(8), 1504-1513.

Lally, D., Ingmire, P., Tong, H. Y., and He, Z. H. (2001). Antisense expression of a cell wall-associated protein kinase, WAK4, inhibits cell elongation and alters morphology. *The Plant Cell, 13*(6), 1317-1332.

Lawrence, G. J., Finnegan, E. J., Ayliffe, M. A., and Ellis, J. G. (1995). The L6 gene for flax rust resistance is related to the *Arabidopsis* bacterial resistance gene RPS2 and the tobacco viral resistance gene N. *The Plant Cell, 7*(8), 1195-1206.

Lee, S. Y., Seo, J. S., Rodriguez-Lanetty, M., and Lee, D. H. (2003). Comparative analysis of superfamilies of NBS-encoding disease resistance gene analogs in cultivated and wild apple species. *Molecular Genetics and Genomics, 269*(1), 101-108.

Leister, D. (2004). Tandem and segmental gene duplication and recombination in the evolution of plant disease resistance genes. *Trends in Genetics, 20*(3), 116-122.

Leister, D., Kurth, J., Laurie, D. A., Yano, M., Sasaki, T., Devos, K., Graner, A., and Schulze-Lefert, P. (1998). Rapid reorganization of resistance gene homologues in cereal genomes. *Proceedings of the National Academy of Science U.S.A, 95*(1), 370-375.

Lewis, N., and Ruud, J. (2004). Apples in the American diet. *Nutrition and Clinical Care, 7*(2), 82-88.

Li, J., Brader, G., Kariola, T., and Palva, E. T. (2006). WRKY70 modulates the selection of signaling pathways in plant defense. *The Plant Journal, 46*(3), 477-491.

Li, J., and Chory, J. (1997). A putative leucine-rich repeat receptor kinase involved in brassinosteroid signal transduction. *Cell, 90*(5), 929-938.

Li, L., and Steffens, J. C. (2002). Overexpression of polyphenol oxidase in transgenic tomato plants results in enhanced bacterial disease resistance. *Planta, 215*(2), 239-247.

Liebhard, R., Koller, B., Gianfranceschi, L., and Gessler, C. (2003). Creating a saturated reference map for the apple (*Malus* × *domestica* Borkh.) genome. *TAG Theoretical and Applied Genetics, 106*(8), 1497-1508.

Ligterink, W., Kroj, T., Zur Nieden, U., Hirt, H., and Scheel, D. (1997). Receptor-mediated activation of a MAP kinase in pathogen defense of plants. *Stress activated MAPKs in plants, 277*, 2054-2057.

Lin, F., Chen, S., Que, Z., Wang, L., Liu, X., and Pan, Q. (2007). The blast resistance gene pi37 encodes a nucleotide binding site leucine-rich repeat protein and is a member of a resistance gene cluster on rice chromosome 1. *Genetics, 177*(3), 1871-1880.

Liu, J. J., and Ekramoddoullah, A. K. (2003). Isolation, genetic variation and expression of TIR-NBS-LRR resistance gene analogs from western white pine (*Pinus monticola* Dougl. ex. D. Don.). *Molecular Genetics and Genomics, 270*(5), 432-441.

Liu, R. H., Liu, J., and Chen, B. (2005). Apples prevent mammary tumors in rats. *The Journal of Agriculture and Food Chemistry, 53*(6), 2341-2343.

Liu, Y., Schiff, M., and Dinesh-Kumar, S. P. (2004). Involvement of MEK1 MAPKK, NTF6 MAPK, WRKY/MYB transcription factors, COI1 and CTR1 in N-mediated resistance to tobacco mosaic virus. *The Plant Journal, 38*(5), 800-809.

Liu, Y., Schiff, M., Marathe, R., and Dinesh-Kumar, S. P. (2002). Tobacco Rar1, EDS1 and NPR1/NIM1 like genes are required for N-mediated resistance to tobacco mosaic virus. *The Plant Journal, 30*(4), 415-429.

Liu, Y., Zhang, S., and Klessig, D. F. (1998). Molecular cloning and characterization of a tobacco MAP kinase kinase that interacts with SIPK. *Proceedings of the National Academy of Science U.S.A, 95*, 7433-7438.

Louw, D. (2006). Information and statistics. *Deciduous Fruit Producers' Trust Annual Report,* 33 -35.

Lu, C., Wang, A., Wang, L., Dorsch, M., Ocain, T. D., and Xu, Y. (2005). Nucleotide binding to CARD12 and its role in CARD12-mediated caspase-1 activation. *Biochemical and Biophysical Research Communications, 331*(4), 1114-1119.

Maheswaran, G., Pridmore, L., Franz, P., and Anderson, M. A. (2007). A proteinase inhibitor from *Nicotiana alata* inhibits the normal development of light-brown apple moth, *Epiphyas postvittana* in transgenic apple plants. *Plant Cell Reports, 26*(6), 773-782.

Malamy, J., and Klessig, D. F. (1992). Salicylic acid and plant disease resistance. *The Plant Journal, 2*(5), 643-654.

Maleck, K., Levine, A., Eulgem, T., Morgan, A., Schmid, J., Lawton, K. A., Dangl, J. L., and Dietrich, R. A. (2000). The transcriptome of *Arabidopsis thaliana* during systemic acquired resistance. *Nature Genetics, 26*(4), 403-410.

Marchler-Bauer, A., Anderson, J. B., Derbyshire, M. K., DeWeese-Scott, C., Gonzales, N. R., Gwadz, M., Hao, L., He, S., Hurwitz, D. I., and Jackson, J. D. (2007). CDD: a conserved domain database for interactive domain family analysis. *Nucleic Acids Research, 35*(Database issue), D237.

Markwick, N. P., Docherty, L. C., Phung, M. M., Lester, M. T., Murray, C., Yao, J. L., Mitra, D. S., Cohen, D., Beuning, L. L., Kutty-Amma, S., and Christeller, J. T. (2003). Transgenic tobacco and apple plants expressing biotin-binding proteins are resistant to two cosmopolitan insect pests, potato tuber moth and lightbrown apple moth, respectively. *Transgenic Research, 12*(6), 671-681.

Martinez, C., Baccou, J. C., Bresson, E., Baissac, Y., Daniel, J. F., Jalloul, A., Montillet, J. L., Geiger, J. P., Assigbetsé, K., and Nicole, M. (2000). Salicylic acid mediated by the oxidative burst is a key molecule in local and systemic responses of cotton challenged by an avirulent race of *Xanthomonas campestris* pv malvacearum. *Plant Physiology, 122*(3), 757.

Mateiu, L., and Rannala, B. (2006). Inferring complex DNA substitution processes on phylogenies using uniformization and data augmentation. *Systematic Biology, 55*(2), 259-269.

Mayrose, I., Mitchell, A., and Pupko, T. (2005). Site-specific evolutionary rate inference: taking phylogenetic uncertainty into account. *Journal of Molecular Evolution, 60*(3), 345-353.

McDowell, J. M. (2004). Convergent evolution of disease resistance genes. *Trends in Plant Science, 9*(7), 315-317.

McDowell, J. M., Dhandaydham, M., Long, T. A., Aarts, M. G., Goff, S., Holub, E. B., and Dangl, J. L. (1998). Intragenic recombination and diversifying selection contribute to the evolution of downy mildew resistance at the RPP8 locus of *Arabidopsis*. *The Plant Cell, 10*(11), 1861-1874.

McHale, L., Tan, X., Koehl, P., and Michelmore, R. W. (2006). Plant NBS-LRR proteins: adaptable guards. *Genome Biology, 7*(4), 212.

Mestre, P., and Baulcombe, D. C. (2006). Elicitor-mediated oligomerization of the tobacco N disease resistance protein W in box. *The Plant Cell, 18*(2), 491-501.

Meyers, B. C., Kaushik, S., and Nandety, R. S. (2005). Evolving disease resistance genes. *Current Opinion in Plant Biology, 8*(2), 129-134.

Meyers, B. C., Kozik, A., Griego, A., Kuang, H., and Michelmore, R. W. (2003). Genome-wide analysis of NBS-LRR-encoding genes in *Arabidopsis*. *The Plant Cell, 15*(4), 809-834.

Michelmore, R. W., and Meyers, B. C. (1998). Clusters of resistance genes in plants evolve by divergent selection and a birth-and-death process. *Genome Research, 8*(11), 1113-1130.

Mindrinos, M., Katagiri, F., Yu, G.-L., and Ausubel, F. M. (1994). The *A. thaliana* disease resistance gene RPS2 encodes a protein containing a nucleotide-binding site and leucine-rich repeats. *Cell, 78*(6), 1089-1099.

Mittler, R., Vanderauwera, S., Gollery, M., and Van Breusegem, F. (2004). Reactive oxygen gene network of plants. *Trends in Plant Science, 9*(10), 490-498.

Moffett, P., Farnham, G., Peart, J., and Baulcombe, D. C. (2002). Interaction between domains of a plant NBS-LRR protein in disease resistance-related cell death. *The EMBO Journal, 21*(17), 4511-4519.

Molders, W., Buchala, A., and Metraux, J. P. (1996). Transport of salicylic acid in tobacco necrosis virus-infected cucumber plants. *Plant Physiology, 112*(2), 787-792.

Mondragon-Palomino, M., and Gaut, B. S. (2005). Gene conversion and the evolution of three leucine-rich repeat gene families in *Arabidopsis thaliana*. *Molecular biology and evolution, 22*(12), 2444-2456.

Moore, R. C., and Purugganan, M. D. (2005). The evolutionary dynamics of plant duplicate genes. *Current Opinion in Plant Biology, 8*(2), 122-128.

Morris, E. R., and Walker, J. C. (2003). Receptor-like protein kinases: the keys to response. *Current Opinion in Plant Biology, 6*(4), 339-342.

Naik, S., Hampson, C., Gasic, K., Bakkeren, G., and Korban, S. S. (2006). Development and linkage mapping of E-STS and RGA markers for functional gene homologues in apple. *Genome, 49*(8), 959-968.

Noel, L., Moores, T. L., van Der Biezen, E. A., Parniske, M., Daniels, M. J., Parker, J. E., and Jones, J. D. (1999). Pronounced intraspecific haplotype divergence at the RPP5 complex disease resistance locus of *Arabidopsis*. *The Plant Cell, 11*(11), 2099-2112.

Nurnberger, T., Brunner, F., Kemmerling, B., and Piater, L. (2004). Innate immunity in plants and animals: striking similarities and obvious differences. *Immunological Reviews, 198*(1), 249-266.

Nurnberger, T., and Lipka, V. (2005). Non-host resistance in plants: new insights into an old phenomenon. *Molecular Plant Pathology, 6*(3), 335-345.

Odjakova, M., and Hadjiivanova, C. (2001). The complexity of pathogen defense in plants. *Bulgarian Journal of Plant Physiology, 27*, 101-109.

Ota, T., and Nei, M. (1995). Evolution of immunoglobulin VH pseudogenes in chickens. *Molecular biology and evolution, 12*(1), 94.

Pal, A., Chakrabarti, A., and Basak, J. (2007). New motifs within the NB-ARC domain of R proteins: Probable mechanisms of integration of geminiviral signatures within the host species of *Fabaceae* family and implications in conferring disease resistance. *Journal of Theoretical Biology, 246*(3), 564-573.

Palomino, C., Satovic, Z., Cubero, J. I., and Torres, A. M. (2006). Identification and characterization of NBS-LRR class resistance gene analogs in faba bean (*Vicia faba* L.) and chickpea (*Cicer arietinum* L.). *Genome, 49*(10), 1227-1237.

Pan, Q., Liu, Y. S., Budai-Hadrian, O., Sela, M., Carmel-Goren, L., Zamir, D., and Fluhr, R. (2000a). Comparative genetics of nucleotide binding site-leucine rich repeat resistance gene homologues in the genomes of two dicotyledons: Tomato and *Arabidopsis*. *Genetics, 155*(1), 309-322.

Pan, Q., Liu, Y. S., Budai-Hadrian, O., Sela, M., Carmel-Goren, L., Zamir, D., and Fluhr, R. (2000b). Comparative genetics of nucleotide binding site-leucine rich repeat resistance gene homologues in the genomes of two dicotyledons: Tomato and *Arabidopsis* (Version 1).

Parisi, L., Lespinasse, Y., Guillaumes, J., and KrÜGer, J. (1993). A new race of *Venturia inaequalis* virulent to apples with resistance due to the *Vf* gene. *Phytopathology, 83*(5), 533-537.

Parker, J. E., Coleman, M. J., Szabo, V., Frost, L. N., Schmidt, R., van der Biezen, E. A., Moores, T., Dean, C., Daniels, M. J., and Jones, J. D. (1997a). The *Arabidopsis*

downy mildew resistance gene RPP5 shares similarity to the toll and interleukin-1 receptors with N and L6. *The Plant Cell, 9*(6), 879-894.

Parker, J. E., Coleman, M. J., Szabo, V., Frost, L. N., Schmidt, R., van der Biezen, E. A., Moores, T., Dean, C., Daniels, M. J., and Jones, J. D. G. (1997b). The *Arabidopsis* downy mildew resistance gene RPP5 shares similarity to the Toll and Interleukin-1 receptors with N and L6. *The Plant Cell, 9*(6), 879-894.

Pastuglia, M., Roby, D., Dumas, C., and Cock, J. M. (1997). Rapid induction by wounding and bacterial infection of an S gene family receptor-like kinase gene in *Brassica oleracea*. *The Plant Cell, 9*(1), 49-60.

Patocchi, A., Bigler, B., Koller, B., Kellerhals, M., and Gessler, C. (2004). Vr2: a new apple scab resistance gene. *TAG Theoretical and Applied Genetics, 109*(5), 1087-1092.

Peart, J. R., Lu, R., Sadanandom, A., Malcuit, I., Moffett, P., Brice, D. C., Schauser, L., Jaggard, D. A. W., Xiao, S., and Coleman, M. J. (2002). Ubiquitin ligase-associated protein SGT1 is required for host and nonhost disease resistance in plants. *Proceedings of the National Academy of Sciences U.S.A, 99*(16), 10865.

Pfaffl, M. W., Horgan, G. W., Vainshtein, Y., and Avery, P. (2005). Relative quantification. *http://rest.gene-quantification.info/*.

Pfeifer, F., Zotzel, J., Kurenbach, B., Roder, R., and Zimmermann, P. (2001). A p-loop motif and two basic regions in the regulatory protein GvpD are important for the repression of gas vesicle formation in the archaeon *Haloferax mediterranei*. *Microbiology, 147*(1), 63-73.

Plymale, R. C., Felton, G. W., and Hoover, K. (2007). Induction of systemic acquired resistance in cotton foliage does not adversely affect the performance of an entomopathogen. *Journal of Chemical Ecology, 33*(8), 1570-1581.

Pond, S. L. K., Frost, S. D. W., and Muse, S. V. (2005). HyPhy: hypothesis testing using phylogenies. *Bioinformatics, 21*(5), 676-679.

Posada, D. (2006). ModelTest Server: a web-based tool for the statistical selection of models of nucleotide substitution online. *Nucleic Acids Research, 34*(Web Server issue), W700-703.

Posada, D., and Crandall, K. A. (1998). MODELTEST: testing the model of DNA substitution. *Bioinformatics, 14*(9), 817-818.

Rairdan, G. J., and Moffett, P. (2006). Distinct domains in the ARC region of the potato resistance protein Rx mediate LRR binding and inhibition of activation. *The Plant Cell, 18*(8), 2082-2093.

Ramonell, K. M., Zhang, B., Ewing, R. M., Chen, Y., Xu, D., Stacey, G., and Somerville, S. (2002). Microarray analysis of chitin elicitation in *Arabidopsis thaliana*. *Molecular Plant Pathology, 3*(5), 301-311.

Rampitsch, C., and Srinivasan, M. (2006). The application of proteomics to plant biology: a review. *Canadian Journal of Botany, 84*(6), 883-892.

References, S., Mu, J., Lee, H., and Kao, T. (1994). Characterization of a pollen-expressed receptor-like kinase gene of *Petunia inflata* and the activity of its encoded kinase. *The Plant Cell, 6*(5), 709-721.

Reuber, T. L., and Ausubel, F. M. (1996). Isolation of *Arabidopsis* genes that differentiate between resistance responses mediated by the RPS2 and RPM1 disease resistance genes. *The Plant Cell, 8*(2), 241-249.

Rhee, S. Y., Beavis, W., Berardini, T. Z., Chen, G., Dixon, D., Doyle, A., Garcia-Hernandez, M., Huala, E., Lander, G., Montoya, M., Miller, N., Mueller, L. A., Mundodi, S., Reiser, L., Tacklind, J., Weems, D. C., Wu, Y., Xu, I., Yoo, D., Yoon, J., and Zhang, P. (2003). The *Arabidopsis* Information Resource (TAIR): a model organism database providing a centralized, curated gateway to *Arabidopsis* biology, research materials and community. *Nucleic Acids Research, 31*(1), 224-228.

Richter, T. E., and Ronald, P. C. (2000). The evolution of disease resistance genes. *Plant Molecular Biology, 42*(1), 195-204.

Roberts, A. L., and Crute, I. R. (1994). Apple scab resistance from *Malus floribunda* 821 (Vf) is rendered ineffective by isolates of *Venturia inaequalis* from *Malus floribunda*. *Norwegian Journal of Agricultural Sciences, 17*, 403-406.

Rocher, F., Chollet, J.-F., Jousse, C., and Bonnemain, J.-L. (2006). Salicylic acid, an ambimobile molecule exhibiting a high ability to accumulate in the phloem. *Plant Physiology, 141*(4), 1684-1693.

Rogers, E. J., Milhalik, S., Ortiz, D., and Shea, T. B. (2004). Apple juice prevents oxidative stress and impaired cognitive performance caused by genetic and dietary deficiencies in mice. *Journal of Nutrition Health and Aging, 8*(2), 92-98.

Rogers, J. S. (2001). Maximum likelihood estimation of phylogenetic trees is consistent when substitution rates vary according to the invariable sites plus gamma distribution. *Systematic Biology, 50*(5), 713-722.

Rombel, I., Peters-Wendisch, P., Mesecar, A., Thorgeirsson, T., Shin, Y. K., and Kustu, S. (1999). MgATP binding and hydrolysis determinants of NtrC, a bacterial enhancer-binding protein. *Journal of Bacteriology, 181*(15), 4628.

Ronquist, F., and Huelsenbeck, J. P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics, 19*(12), 1572-1574.

Ronquist, F., Huelsenbeck, J. P., and van der Mark, P. (2005). MrBayes 3.1 Manual. *School of Computational Science, Florida State University*.

Rose, L. E., Langley, C. H., Bernal, A. J., and Michelmore, R. W. (2005). Natural variation in the Pto pathogen resistance gene within species of wild tomato (*Lycopersicon*). I. Functional analysis of Pto alleles. *Genetics, 171*(1), 345-357.

Rowland, O., and Jones, J. D. (2001). Unraveling regulatory networks in plant defense using microarrays. *Genome Biology, 2*(1), REVIEWS1001.

Rudrabhatla, P., Reddy, M. M., and Rajasekharan, R. (2006). Genome-wide analysis and experimentation of plant serine/ threonine/tyrosine-specific protein kinases. *Plant Molecular Biology, 60*(2), 293-319.

Rushton, P. J., Macdonald, H., Huttly, A. K., Lazarus, C. M., and Hooley, R. (1995). Members of a new family of DNA-binding proteins bind to a conserved cis-element in the promoters of α-Amy2 genes. *Plant Molecular Biology, 29*(4), 691-702.

Ryals, J. A., Neuenschwander, U. H., Willits, M. G., Molina, A., Steiner, H. Y., and Hunt, M. D. (1996). Systemic Acquired Resistance. *The Plant Cell, 8*(10), 1809.

Sagi, M., and Fluhr, R. (2006). Production of reactive oxygen species by plant NADPH oxidases. *Plant Physiology, 141*(2), 336.

Salmeron, J. M., Oldroyd, G. E., Rommens, C. M., Scofield, S. R., Kim, H. S., Lavelle, D. T., Dahlbeck, D., and Staskawicz, B. J. (1996). Tomato Prf is a member of the

leucine-rich repeat class of plant disease resistance genes and lies embedded within the Pto kinase gene cluster. *Cell, 86*(1), 123-133.

Saraste, M., Sibbald, P. R., and Wittinghofer, A. (1990). The P-loop--a common motif in ATP-and GTP-binding proteins. *Trends in Biochemical Sciences, 15*(11), 430-434.

Sawyer, S. (1989). Statistical tests for detecting gene conversion. *Molecular biology and evolution, 6*(5), 526-538.

Sawyer, S. A. (1999). GENECONV: a computer package for the statistical detection of gene conversion. *Distributed by the author, Department of Mathematics, Washington University in St. Louis, available at http://www. math. wustl. edu/sawyer*.

Sawyer, S. A. (2000). GENECONV: statistical tests for detecting gene conversion–Version 1.81. *Department of Mathematics, Washington University, St. Louis, Mo*.

Schenk, P. M., Kazan, K., Wilson, I., Anderson, J. P., Richmond, T., Somerville, S. C., and Manners, J. M. (2000). Coordinated plant defense responses in *Arabidopsis* revealed by microarray analysis. *Proceedidings of the National Academy of Science U.S.A, 97*(21), 11655-11660.

Schmelz, E. A., LeClere, S., Carroll, M. J., Alborn, H. T., and Teal, P. E. (2007). Cowpea chloroplastic ATP synthase is the source of multiple plant defense elicitors during insect herbivory. *Plant Physiology, 144*(2), 793-805.

Shen, Q. H., Zhou, F., Bieri, S., Haizel, T., Shirasu, K., and Schulze-Lefert, P. (2003). Recognition specificity and RAR1/SGT1 dependence in barley Mla disease resistance genes to the powdery mildew fungus. *The Plant Cell, 15*(3), 732-744.

Shiu, S. H., and Bleecker, A. B. (2001). Receptor-like kinases from *Arabidopsis* form a monophyletic gene family related to animal receptor kinases. *Proceedings of the National Academy of Sciences*, 181141598.

Shulaev, V., Leon, J., and Raskin, I. (1995). Is salicylic acid a translocated signal of systemic acquired resistance in tobacco? *The Plant Cell, 7*(10), 1691-1701.

Silfverberg-Dilworth, E., Matasci, C., Van de Weg, W., Van Kaauwen, M., Walser, M., Kodde, L., Soglio, V., Gianfranceschi, L., Durel, C., Costa, F., Yamamoto, T., Koller, B., Gessler, C., and Patocchi, A. (2006). Microsatellite markers spanning the apple (*Malus x domestica* Borkh.) genome. *Tree Genetics & Genomes, 2*(4), 202-224.

Somssich, I. E., and Hahlbrock, K. (1998). Pathogen defence in plants -- a paradigm of biological complexity. *Trends in Plant Science, 3*(3), 86-90.

Staden, R., Beal, K. F., and Bonfield, J. K. (2000). The Staden package, 1998. *Methods in Molecular Biology, 132*, 115-130.

Staskawicz, B. J., Mudgett, M. B., Dangl, J. L., and Galan, J. E. (2001). Common and contrasting themes of plant and animal diseases. *Science, 292*(5525), 2285-2289.

Sun, X., Cao, Y., and Wang, S. (2006). Point mutations with positive selection were a major force during the evolution of a receptor-kinase resistance gene family of rice. *Plant Physiology, 140*(3), 998-1008.

Takahashi, Y., Nasir, K. H., Ito, A., Kanzaki, H., Matsumura, H., Saitoh, H., Fujisawa, S., Kamoun, S., and Terauchi, R. (2007). A high-throughput screen of cell-death-inducing factors in *Nicotiana benthamiana* identifies a novel MAPKK that

mediates INF1-induced cell death signaling and non-host resistance to *Pseudomonas cichorii*. *The Plant Journal, 49*(6), 1030-1040.

Takken, F. L. W., Albrecht, M., and Tameling, W. I. L. (2006). Resistance proteins: molecular switches of plant defence. *Current Opinion in Plant Biology, 9*(4), 383-390.

Tameling, W. I., Elzinga, S. D., Darmin, P. S., Vossen, J. H., Takken, F. L. W., Haring, M. A., and Cornelissen, B. J. C. (2002). The tomato R gene products I-2 and Mi-1 are functional ATP binding proteins with ATPase activity. *The Plant Cell, 14*(11), 2929-2939.

Tao, Y., Yuan, F., Leister, R. T., Ausubel, F. M., and Katagiri, F. (2000). Mutational analysis of the *Arabidopsis* nucleotide binding site-leucine-rich repeat resistance gene *RPS2*. *The Plant Cell, 12*(12), 2541-2554.

Tchantchou, F., Graves, M., Ortiz, D., Rogers, E., and Shea, T. B. (2004). Dietary supplementation with 3-deaza adenosine, N-acetyl cysteine, and S-adenosyl methionine provide neuroprotection against multiple consequences of vitamin deficiency and oxidative challenge. *NeuroMolecular Medicine, 6*(2), 93-103.

Torii, K. U., Mitsukawa, N., Oosumi, T., Matsuura, Y., Yokoyama, R., Whittier, R. F., and Komeda, Y. (1996). The *Arabidopsis* ERECTA gene encodes a putative receptor protein kinase with extracellular leucine-rich repeats. *The Plant Cell, 8*(4), 735-746.

Traut, T. W. (1994). The functions and consensus motifs of nine types of peptide segments that form different types of nucleotide-binding sites. *FEBS Journal, 222*(1), 9-19.

Trognitz, F., and Trognitz, B. R. (2005). Survey of resistance gene analogs in *Solanum caripense*, a relative of potato and tomato, and update on R gene genealogy. *Molecular Genetics and Genomics, 274*(6), 595-605.

Ulker, B., Shahid Mukhtar, M., and Somssich, I. E. (2007). The WRKY70 transcription factor of *Arabidopsis* influences both the plant senescence and defense signaling pathways. *Planta, 226*(1), 125-137.

Vacca, R. A., Valenti, D., Bobba, A., de Pinto, M. C., Merafina, R. S., De Gara, L., Passarella, S., and Marra, E. (2007). Proteasome function is required for activation of programmed cell death in heat shocked tobacco Bright-Yellow 2 cells. *FEBS Letters, 581*(5), 917-922.

Van Der Biezen, E. A., and Jones, J. D. G. (1998). Plant disease-resistance proteins and the gene-for-gene concept. *Trends in Biochemical Sciences, 23*(12), 454-456.

van der Linden, C. G., Wouters, D. C., Mihalka, V., Kochieva, E. Z., Smulders, M. J., and Vosman, B. (2004). Efficient targeting of plant disease resistance loci using NBS profiling. *TAG Theoretical and Applied Genetics, 109*(2), 384-393.

van Dyk, M. M., Baison, J., Hove, P., Marondedze, C., Klein, A., Soeker, K., and Rees, D. J. G. (In press). Bin mapping of EST-SSRs and EST-SNPs in apple (*Malus x domestica* Borkh.). *Acta Horticulturae*.

van Dyk, M. M., Labuschagne, I. F., and Rees, D. J. G. (2007). Genetic linkage map construction and the identification of QTL's affecting time and number of initial vegetative budbreak in apple (*Malus x domestica* Borkh.). [Thesis]. *Thesis*.

Verica, J. A., and He, Z. H. (2002). The cell wall-associated kinase (WAK) and WAK-like kinase gene family. *Plant Physiology, 129*(2), 455-459.

Vinatzer, B. A., Patocchi, A., Gianfranceschi, L., Tartarini, S., Zhang, H. B., Gessler, C., and Sansavini, S. (2001). Apple contains receptor-like genes homologous to the *Cladosporium fulvum* resistance gene family of tomato with a cluster of genes cosegregating with Vf apple scab resistance. *Molecular Plant-Microbe Interactions, 14*(4), 508-515.

Vinatzer, B. A., Patocchi, A., Tartarini, S., Gianfranceschi, L., Sansavini, S., and Gessler, C. (2004). Isolation of two microsatellite markers from BAC clones of the Vf scab resistance region and molecular characterization of scab-resistant accessions in *Malus* germplasm*. *Plant Breeding, 123*(4).

Vos, P., Simons, G., Jesse, T., Wijbrandi, J., Heinen, L., Hogers, R., Frijters, A., Groenendijk, J., Diergaarde, P., Reijans, M., Fierens-Onstenk, J., Both, M. d., Peleman, J., Liharska, T., Hontelez, J., and ZabeauMarc. (1998). The tomato Mi-1 gene confers resistance to both root-knot nematodes and potato aphids. *Nature Biotechnology, 16*(13), 1365-1369.

Wagner, T. A., and Kohorn, B. D. (2001). Wall-associated kinases are expressed throughout plant development and are required for cell expansion. *The Plant Cell, 13*(2), 303-318.

Walker, J. E., Saraste, M., Runswick, M. J., and Gay, N. J. (1982). Distantly related sequences in the alpha-and beta-subunits of ATP synthase, myosin, kinases and other ATP-requiring enzymes and a common nucleotide binding fold. *The EMBO Journal, 1*(8), 945.

Ward, E. R., Uknes, S. J., Williams, S. C., Dincher, S. S., Wiederhold, D. L., Alexander, D. C., Ahl-Goy, P., Metraux, J. P., and Ryals, J. A. (1991). Coordinate gene activity in response to agents that induce systemic acquired resistance. *The Plant Cell, 3*(10), 1085.

Waugh, R., McLean, K., Flavell, A. J., Pearce, S. R., Kumar, A., Thomas, B. B., and Powell, W. (1997). Genetic distribution of Bare-1-like retrotransposable elements in the barley genome revealed by sequence-specific amplification polymorphisms (S-SAP). *Molecular and General Genetics, 253*(6), 687-694.

Whitham, S., Dinesh-Kumar, S. P., Choi, D., Hehl, R., Corr, C., and Baker, B. (1994). The product of the tobacco mosaic virus resistance gene N: Similarity to toll and the interleukin-1 receptor. *Cell, 78*(6), 1101-1115.

Wicker, T., Yahiaoui, N., and Keller, B. (2007). Illegitimate recombination is a major evolutionary mechanism for initiating size variation in plant resistance genes. *The Plant Journal, 51*(4), 631-641.

Wilkening, S., and Bader, A. (2004). Quantitative real-time polymerase chain reaction: Methodical analysis and mathematical model. *Journal of Biomolecular Technology, 15*(2), 107-111.

Will, T., Tjallingii, W. F., Thonnessen, A., and van Bel, A. J. (2007). Molecular sabotage of plant defense by aphid saliva. *Proceedings of the National Academy of Science U.S.A, 104*(25), 10536-10541.

Wolfenden, R., Andersson, L., Cullis, P. M., and Southgate, C. C. (1981). Affinities of amino acid side chains for solvent water. *Biochemistry, 20*(4), 849-855.

Woodgett, J. R., Gould, K. L., and Hunter, T. (1986). Substrate specificity of protein kinase C. Use of synthetic peptides corresponding to physiological sites as probes for substrate recognition requirements. *FEBS Journal, 161*(1), 177-184.

Wroblewski, T., Piskurewicz, U., Tomczak, A., Ochoa, O., and Michelmore, R. W. (2007). Silencing of the major family of NBS-LRR-encoding genes in lettuce results in the loss of multiple resistance specificities. *The Plant Journal*.

Wullschleger, S. D., and Difazio, S. P. (2003). Emerging use of gene expression microarrays in plant physiology. *Comparative and Functional Genomics, 4*(2), 216-224.

Xing, Y., Frei, U., Schejbel, B., Asp, T., and Lubberstedt, T. (2007). Nucleotide diversity and linkage disequilibrium in 11 expressed resistance candidate genes in *Lolium perenne*. *BMC Plant Biology, 7*, 43.

Xu, Q., Wen, X., and Deng, X. (2005). Isolation of TIR and non-TIR NBS--LRR resistance gene analogues and identification of molecular markers linked to a powdery mildew resistance locus in chestnut rose (*Rosa roxburghii* Tratt). *TAG Theoretical and Applied Genetics, 111*(5), 819-830.

Xu, Q., Wen, X., and Deng, X. (2007). Phylogenetic and evolutionary analysis of NBS-encoding genes in *Rosaceae* fruit crops. *Molecular Phylogenetics and Evolution, 44*(1), 315-324.

Yano, Y., Saito, R., Yoshida, N., Yoshiki, A., Wynshaw-Boris, A., Tomita, M., and Hirotsune, S. (2004). A new role for expressed pseudogenes as ncRNA: regulation of mRNA stability of its homologous coding gene. *Journal of Molecular Medicine, 82*(7), 414-422.

Zhang, S., and Klessig, D. F. (2001). MAPK cascades in plant defense signaling. *Trends in Plant Science, 6*(11), 520-527.

Zhang, Z., and Hu, J. (2007). Development and validation of endogenous reference genes for expression profiling of Medaka (*Oryzias latipes*) exposed to endocrine disrupting chemicals by quantitative real-time RT-PCR. *Toxicological Sciences, 95*(2), 356.

Zheng, Z., Mosher, S. L., Fan, B., Klessig, D. F., and Chen, Z. (2007). Functional analysis of *Arabidopsis* WRKY25 transcription factor in plant defense against *Pseudomonas syringae*. *BMC Plant Biology, 7*, 2.

Zhou, H., Gu, J., Lamont, S. J., and Gu, X. (2007). Evolutionary analysis for functional divergence of the Toll-like receptor gene family and altered functional constraints. *Journal of Molecular Evolution*.

Zhou, J., Loh, Y. T., Bressan, R. A., and Martin, G. B. (1995). The tomato gene Pti 1 encodes a serine/threonine kinase that is phosphorylated by Pto and is involved in the hypersensitive response. *Cell, 83*(6), 925-935.

Zhou, J., Tang, X., and Martin, G. B. (1997). The Pto kinase conferring resistance to tomato bacterial speck disease interacts with proteins that bind a cis-element of pathogenesis-related genes. *The EMBO Journal, 16*(11), 3207-3218.

Zhou, T., Wang, Y., Chen, J. Q., Araki, H., Jing, Z., Jiang, K., Shen, J., and Tian, D. (2004). Genome-wide identification of NBS genes in japonica rice reveals significant expansion of divergent non-TIR NBS-LRR genes. *Molecular Genetics and Genomics, 271*(4), 402-415.

Zhu, J., Dong, C. H., and Zhu, J. K. (2007). Interplay between cold-responsive gene regulation, metabolism and RNA processing during plant cold acclimation. *Current Opinion in Plant Biology, 10*(3), 290-295.

Zimmer, A., Lang, D., Richardt, S., Frank, W., Reski, R., and Rensing, S. A. (2007). Dating the early evolution of plants: detection and molecular clock analyses of orthologs. *Molecular Genetics and Genomics*.

Zimmerli, L., Stein, M., Lipka, V., Schulze-Lefert, P., and Somerville, S. (2004). Host and non-host pathogens elicit different jasmonate/ethylene responses in *Arabidopsis*. *The Plant Journal, 40*(5), 633-646.

Zimmermann, S., Nurnberger, T., Frachisse, J. M., Wirtz, W., Guern, J., Hedrich, R., and Scheel, D. (1997). Receptor-mediated activation of a plant Ca(2+)-permeable ion channel involved in pathogen defense. *Proceedings of the National Academy of Science U.S.A, 94*(6), 2751-2755.

Zuo, K. J., Qin, J., Zhao, J. Y., Ling, H., Zhang, L. D., Cao, Y. F., and Tang, K. X. (2007). Over-expression GbERF2 transcription factor in tobacco enhances brown spots disease resistance by activating expression of downstream genes. *Gene, 391*(1-2), 80-90.

APPENDIX

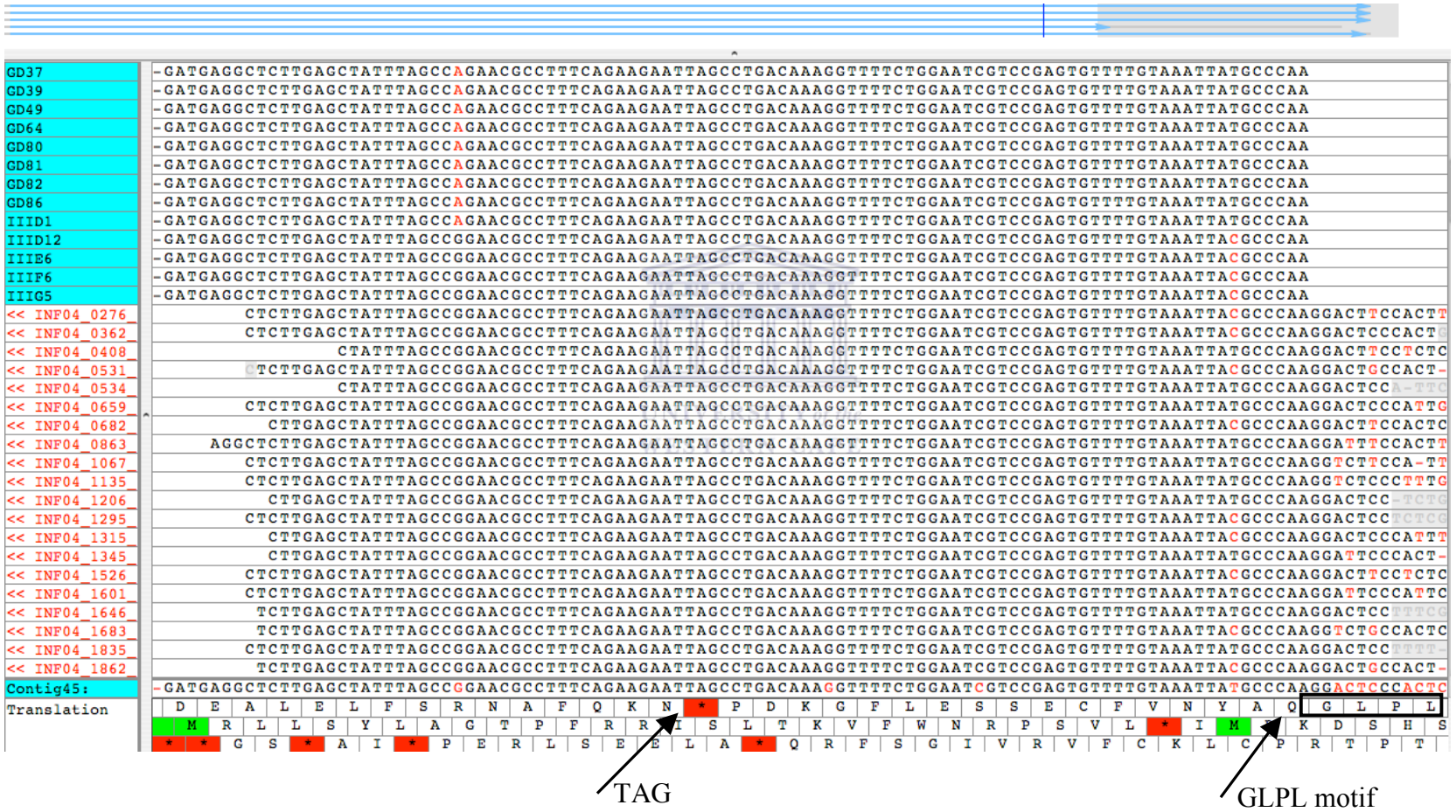


Figure A1. Sequence assembly of the transcriptome data showing evidence of a transcribed pseudogene.

Table A1. ***Malus x domestica*** **(Borh.) cv Anna bin mapping table**

LG	Locus	Classification	11	12	16	19	51	52	130	276	320	327	330	335	344	353
1	Hi21g05	(ab,cd)	0	1	1	0	1	1	0	1	0	1	1	1	1	1
1	CH03g12x	(ab,cd)	0	1	1	0	1	1	0	0	0	1	1	1	0	1
1	Hi02b10x	(ab,cd)	0	1	1	-	1	1	0	0	0	1	1	1	0	1
1	Hi02c07	(ab,cd)	0	1	1	0	1	1	0	-	0	1	1	1	0	0
1	SA-A369	(ab,cd)	0	1	1	-	1	1	0	0	0	1	1	1	0	0
1	AG11	(ab,cd)	0	1	0	0	1	1	0	0	0	1	1	1	1	0
1	SA-A490		-	1	-	1	1	1	1	1	-	1	1	0	1	1
1	NZmsCN879773	(ab,cd)	0	1	1	1	1	1	1	1	0	1	1	0	1	0
1	Hi07d08	(ab,cd)	0	1	1	1	1	1	1	0	0	1	1	0	1	0
1	CH05g08	(ab,cd)	0	1	1	1	1	1	1	0	0	1	1	0	1	0
2	Hi22d06	(ab,cd)	0	1	1	1	1	0	0	1	0	0	1	0	1	0
2	CH02f061	(ab,cd)	0	1	1	1	1	0	0	1	0	0	1	0	1	0
2	AJ251116-SSR	(ab,cd)	0	1	1	1	1	1	0	1	0	0	0	0	1	0
2	SA-A389	(ab,cd)	-	1	1	1	1	1	0	-	0	0	0	0	1	0
2	SA-A513		0	1	-	1	1	1	0	1	0	0	0	0	1	0
2	CH03d10	(ab,cd)	0	1	-	1	0	1	0	1	0	0	0	1	1	0
2	CH05e03	(ab,cd)	0	1	0	1	0	1	0	0	0	1	0	1	1	0
2	CH02a04y	(ab,cd)	0	1	0	1	0	1	0	0	0	0	0	1	0	0
2	CH02c061	(ab,cd)	0	1	0	1	0	1	0	0	0	0	0	0	0	0
3	CH03g07	(ab,cd)	1	1	1	1	0	1	0	-	1	-	0	1	0	1
3	Hi03d06	(ab,cd)	-	1	1	1	0	-	0	1	-	1	1	1	0	1
3	Hi04c10z	(ab,cd)	1	1	1	1	0	1	-	1	1	1	1	1	0	-
3	SA-A310	(ab,cd)	-	1	1	1	0	1	0	0	1	1	1	1	0	0
3	Hi04c10y	(ab,cd)	1	1	1	1	0	1	-	0	1	1	1	1	0	-
3	SA-A293		1	1	1	1	0	1	0	1	1	1	1	1	0	0
3	CH03g12y	(ab,cd)	1	1	1	1	0	1	0	1	1	1	1	1	0	0
4	05g8	(ab,cd)	1	1	-	0	1	1	1	1	0	1	1	0	1	0
4	CH04e02	(ab,cd)	1	1	1	0	1	-	1	1	0	1	1	0	0	0
4	Hi01e10	(ab,cd)	1	1	1	0	1	1	1	1	0	1	1	0	0	0
4	SA-A687	(ab,cd)	1	1	-	-	1	1	1	1	0	1	1	1	0	0
4	CH01b09b	(ab,cd)	0	1	1	0	1	1	1	1	0	1	1	1	0	0
4	Hi23g08	(ab,cd)	0	1	1	0	1	1	1	1	0	1	1	1	0	0
4	GD162 / A8		0	1	1	0	1	1	1	1	0	1	1	1	0	0
4	CH02h11a	(ab,cd)	0	1	1	0	1	1	1	1	0	1	1	1	-	-

LG	Locus	Classification	11	12	16	19	51	52	130	276	320	327	330	335	344	353
4	CH01d03	(ab,cd)	0	1	1	0	1	1	1	1	0	1	1	1	-	0
4	SA-A615		1	1	1	0	1	1	1	1	-	1	1	1	0	0
4	Hi23g02	(ab,cd)	1	1	1	0	1	1	1	1	0	-	1	1	0	0
4	CH05d02	(ab,cd)	0	1	1	1	1	1	1	1	1	1	1	1	1	0
4	CH02c02b	(ab,cd)	0	1	1	0	1	1	0	1	0	1	1	1	1	1
5	Hi22f12	(ab,cd)	1	0	1	0	0	0	0	1	0	1	0	0	0	0
5	SA-A340	(ab,cd)	1	-	1	-	0	-	-	-	0	0	0	0	-	0
5	CH03a09	(ab,cd)	1	0	1	0	0	1	1	0	0	0	0	0	0	0
5	SA-A282		0	1	1	0	0	1	1	1	0	0	0	0	0	0
5	CH05e06	(ab,cd)	1	1	1	-	0	1	1	-	-	1	0	0	0	-
5	SA-A401	(ab,cd)	0	1	1	0	0	1	1	1	0	0	0	0	0	0
5	CH03e03	(ab,cd)	0	1	1	0	0	1	1	1	0	0	0	0	0	0
5	Hi04d02	(ab,cd)	0	1	1	0	0	1	1	0	0	0	0	0	0	0
5	CH05f06	(ab,cd)	0	1	-	0	0	1	1	0	0	0	0	0	0	0
5	CH02b121	(ab,cd)	0	0	-	0	1	1	0	0	0	0	0	0	0	0
5	CH03a04	(ab,cd)	0	0	0	0	1	1	0	1	0	0	0	0	0	0
6	Hi05d10	(ab,cd)	1	1	1	0	1	0	1	1	0	-	0	0	1	1
6	CH03d12	(ab,cd)	1	1	1	0	1	0	0	1	0	0	0	0	0	1
6	Hi04d10	(ab,cd)	1	1	1	0	1	-	0	1	0	0	0	0	0	1
6	SA-A484		1	0	-	0	1	0	0	1	-	0	0	0	0	1
6	Hi03a03x	(ab,cd)	1	0	1	0	1	-	0	0	-	0	0	0	0	1
6	CH03c01	(ab,cd)	1	0	1	0	1	0	0	0	1	0	0	0	0	-
6	Hi07b06	(ab,cd)	1	0	1	0	1	-	0	0	1	0	0	0	0	1
7	CN444794-SSR	(ab,cd)	0	1	0	0	1	-	1	1	-	0	1	0	1	0
7	SA-A718	(ab,cd)	0	1	0	0	1	1	1	-	1	0	1	0	0	0
7	SA-A233	(ab,cd)	-	1	-	-	-	1	0	0	1	0	1	0	1	0
7	SA-A425		-	1	-	0	1	1	0	0	-	0	1	1	1	0
7	SA-A332	(ab,cd)	0	-	1	0	1	1	0	0	-	0	1	1	1	0
7	SA-A658		0	-	-	0	1	1	0	0	0	0	1	1	1	0
7	Hi05b09	(ab,cd)	0	0	1	0	1	1	0	0	0	0	1	1	1	0
8	Hi04b12	(ab,cd)	0	1	1	1	1	1	-	0	0	1	1	1	0	-
8	CH01c06	(ab,cd)	1	1	1	1	0	0	0	0	0	1	1	1	0	0
8	Hi20b03	(ab,cd)	1	1	0	1	0	0	1	1	0	1	1	1	-	0
8	CH05a02y	(ab,cd)	1	1	0	1	-	0	1	0	0	1	1	-	-	0
8	CH01e12	(ab,cd)	1	0	1	-	0	0	0	0	0	-	1	-	1	0
8	SA-A599	(ab,cd)	1	0	0	0	0	0	1	0	0	1	1	0	1	1

LG	Locus	Classification	11	12	16	19	51	52	130	276	320	327	330	335	344	353
9	GD142 / A6		0	1	1	0	1	0	0	0	0	1	0	1	0	1
9	Hi05e07	(ab,cd)	0	1	1	0	1	1	1	0	1	0	0	1	1	0
9	Hi01d01	(ab,cd)	-	1	1	1	1	-	1	0	-	1	1	-	1	0
9	Hi04a05	(ab,cd)	0	1	1	0	1	1	1	1	1	0	1	1	1	0
9	SA-A379	(ab,cd)	1	0	0	1	0	1	1	0	0	0	1	1	1	0
10	SA-A715	(ab,cd)	1	1	1	-	0	0	1	0	0	1	0	0	1	0
10	CH02b07	(ab,cd)	0	0	-	1	0	0	1	0	0	1	0	1	1	0
10	Hi02d04	(ab,cd)	0	1	1	1	0	0	1	0	0	1	0	1	1	0
10	SA-A659b		-	1	-	1	0	0	1	1	0	1	0	1	1	0
10	CH02a08	(ab,cd)	0	1	1	0	0	0	0	0	0	1	0	1	1	0
10	CH01e09b	(ab,cd)	0	1	1	0	0	0	0	0	0	1	0	1	1	0
10	CH02a10	(ab,cd)	0	1	1	0	0	0	0	0	0	1	0	1	1	0
10	CH01f12	(ab,cd)	0	1	1	0	0	0	0	0	0	1	0	1	1	0
10	CH02c11	(ab,cd)	0	1	1	0	0	0	0	0	0	1	0	1	-	0
10	SA-A326	(ab,cd)	0	0	0	-	0	-	1	-	-	0	0	1	-	1
10	SA-A502	(ab,cd)	0	0	0	1	0	1	1	-	0	0	1	1	1	0
10	SA-A492		0	0	0	1	0	1	1	-	-	0	1	1	1	0
10	SA-A247	(ab,cd)	-	-	-	-	-	-	-	-	-	0	-	1	-	-
10	SA-A462		-	0	0	1	0	1	1	-	0	0	0	1	1	1
10	MS06G03		-	0	1	1	0	1	1	0	0	0	0	1	0	1
11	CH04a12	(ab,cd)	-	-	0	1	0	-	0	1	0	0	1	1	0	0
11	CH04g07	(ab,cd)	1	1	0	1	1	0	0	1	0	0	0	1	0	1
11	SA-A485		1	1	-	1	1	0	0	1	-	0	0	1	0	1
11	SA-A756	(ab,cd)	1	1	1	1	1	0	1	-	0	-	1	1	-	1
12	CH05g07	(ab,cd)	1	0	0	0	0	0	1	1	0	0	0	0	1	1
12	28f4	(ab,cd)	1	0	0	0	0	0	1	1	0	0	0	0	-	1
12	CH05d11	(ab,cd)	1	0	0	0	0	0	1	1	0	0	0	0	1	1
12	SA-A656		1	1	0	0	0	0	1	0	0	0	0	0	1	1
12	CH01d09	(ab,cd)	0	1	0	0	0	0	1	-	-	0	0	0	1	1
12	CH01g121	(ab,cd)	0	1	0	0	0	0	1	1	0	0	0	0	1	1
12	SA-A245	(ab,cd)	0	0	0	0	0	0	1	0	0	0	1	0	1	1
12	CH02h11b	(ab,cd)	0	1	0	0	0	0	1	0	0	0	1	0	1	1
12	SA-A505		1	1	-	0	0	0	1	0	0	0	1	0	1	0
12	Hi07f01	(ab,cd)	0	-	0	0	0	-	-	1	-	-	-	-	-	-
12	SA-A202		0	0	0	0	0	0	0	0	0	0	0	0	1	1

LG	Locus	Classification	11	12	16	19	51	52	130	276	320	327	330	335	344	353
13	CH05h05	(ab,cd)	1	1	1	1	1	1	0	-	-	0	1	0	1	1
13	SA-A392	(ab,cd)	1	1	1	1	1	1	0	1	1	0	1	-	1	1
13	CH02g01	(ab,cd)	0	0	1	1	0	0	0	-	1	0	1	0	1	1
13	GD147 / A7		-	0	-	1	0	0	0	1	1	0	1	0	1	1
13	Hi04f09	(ab,cd)	1	0	1	1	0	0	-	1	1	0	1	0	1	-
13	NH009b	(ab,cd)	1	0	1	1	0	0	0	1	1	1	1	0	0	1
13	SA-A372	(ab,cd)	1	-	1	-	0	0	0	1	1	1	0	0	0	1
13	CH03h03	(ab,cd)	1	0	1	1	0	1	0	1	1	1	0	0	0	1
14	CH04f06	(ab,cd)	0	1	1	1	1	0	0	1	1	1	1	0	0	1
14	NZmsEB146613	(ab,cd)	0	0	1	0	0	1	1	0	0	1	0	1	0	1
14	Hi08c05	(ab,cd)	-	1	1	-	1	1	1	1	1	1	1	0	0	1
14	SA-A30	(ab,cd)	0	0	-	-	0	1	1	1	1	1	0	-	1	0
14	CH01g05	(ab,cd)	1	1	1	1	0	1	1	0	0	1	0	0	0	0
14	CH03g04	(ab,cd)	1	1	1	1	1	1	1	1	0	1	-	0	-	0
14	CH05g11	(ab,cd)	1	1	1	1	1	1	1	1	0	1	0	0	0	0
14	Hi02d11	(ab,cd)	1	1	-	-	1	1	1	1	0	1	0	0	0	0
14	SA-A397	(ab,cd)	0	1	0	0	0	1	1	1	0	1	0	0	0	0
14	SA-A523	(ab,cd)	1	1	1	1	1	1	1	1	0	1	0	0	0	0
14	SA-A222	(ab,cd)	1	1	1	1	1	1	1	0	-	1	0	-	0	0
14	SA-A324		1	1	-	0	1	1	1	1	1	1	1	0	0	0
14	SA-A652		1	1	-	0	1	1	1	1	-	1	1	0	0	0
14	Hi03a03y	(ab,cd)	1	1	0	0	1	-	1	0	-	1	1	0	0	0
15	SA-A238	(ab,cd)	0	0	0	-	0	1	1	1	-	0	-	1	-	-
15	SA-A182	(ab,cd)	-	0	-	-	-	1	1	-	-	1	1	1	-	1
15	02b1	(ab,cd)	-	0	0	1	-	1	1	-	1	1	1	1	0	-
15	SA-A516		-	0	0	1	0	1	0	0	1	1	1	1	0	0
15	CH03b10	(ab,cd)	1	1	0	0	1	0	1	1	0	1	0	0	1	0
15	SA-A244	(ab,cd)	1	1	1	0	-	0	1	0	0	-	1	0	1	0
15	SA-A478		1	1	1	0	1	0	1	0	0	1	1	0	1	0
15	SA-A535	(ab,cd)	1	1	1	0	1	0	1	-	0	1	1	0	1	0
15	SA-A354		-	0	-	1	1	0	1	1	1	1	1	0	1	1
15	CH02c09	(ab,cd)	-	0	1	1	0	0	1	1	1	-	1	0	0	1
16	CH02d10a	(ab,cd)	1	1	-	-	0	1	1	1	1	1	1	0	1	1
16	SA-A343	(ab,cd)	1	1	0	1	0	1	1	1	1	1	1	0	1	1
16	CH05b06x	(ab,cd)	1	1	0	1	0	1	1	1	1	1	1	0	1	1
16	Hi12a02	(ab,cd)	1	1	0	1	0	1	1	0	1	1	1	0	1	1
16	Hi02h08	(ab,cd)	1	1	0	1	0	-	1	0	1	1	0	0	1	1

LG	Locus	Classification	11	12	16	19	51	52	130	276	320	327	330	335	344	353
16	CH05e04	(ab,cd)	1	1	0	1	-	1	1	-	1	1	0	0	1	1
16	SA-A601	(ab,cd)	1	1	0	0	-	1	1	1	1	1	0	0	1	1
16	SA-A728	(ab,cd)	1	0	-	-	0	1	0	1	1	1	0	0	1	1
16	SA-A494	(ab,cd)	1	0	0	0	0	1	0	1	1	1	0	0	1	1
16	CH05a04	(ab,cd)	1	0	0	0	0	0	0	0	1	1	0	0	1	1
16	Hi02b10y	(ab,cd)	1	0	0	-	-	0	0	-	1	1	0	0	1	1
16	SA-A680		1	0	0	0	1	0	0	1	1	0	1	0	1	1
16	CH04f10	(ab,cd)	1	0	0	-	-	-	0	-	-	0	1	0	-	-
16	SA-A267	(ab,cd)	1	0	0	0	1	0	0	0	1	0	1	0	1	1
16	SA-A430	(ab,cd)	0	0	1	1	0	0	-	-	-	1	-	-	-	-
17	CH05g03	(ab,cd)	0	1	-	0	1	1	1	1	1	0	1	1	1	0
17	CH01h011	(ab,cd)	1	1	0	0	1	1	1	0	1	0	1	0	0	0
17	SA-A736	(ab,cd)	1	1	0	0	1	1	1	0	1	0	1	0	0	0
17	CH04c06	(ab,cd)	0	1	0	0	1	1	0	0	0	0	1	0	0	-

Table A2. ***Malus x domestica*** **(Borkh.) cv Golden Delicious bin mapping table**

LG	Locus	Classification	11	12	16	19	51	52	130	276	320	327	330	335	344	353
1	Hi02b10y	(ab,cd)	0	1	1	-	-	0	1	-	1	0	0	0	1	1
1	CH03g12x	(ab,cd)	0	1	1	0	0	0	1	1	1	0	0	0	1	1
1	Hi02b10x	(ab,cd)	0	1	1	-	0	0	1	1	1	0	0	0	1	1
1	SA-A369	(ab,cd)	0	1	1	-	0	0	1	1	1	0	0	1	1	1
1	Hi02c07	(ab,cd)	0	1	1	0	0	0	1	-	1	0	0	1	1	1
1	AG11	(ab,cd)	0	1	1	0	1	0	1	1	0	0	0	1	1	0
1	Hi12c02	(ab,cd)	0	0	1	0	1	0	1	0	0	0	0	1	1	0
1	NZmsCN879773	(ab,cd)	0	0	1	0	1	0	0	0	0	0	0	1	1	0
1	CH-Vf1	(ab,cd)	0	0	1	0	1	0	0	0	0	0	0	1	1	0
2	Hi22d06	(ab,cd)	1	1	1	1	1	0	1	1	1	1	1	1	0	0
2	CH02f061	(ab,cd)	0	1	1	1	1	0	1	1	1	1	1	1	0	0
2	SA-A314	(ab,cd)	0	1	1	1	0	0	0	0	-	1	0	1	0	1
2	SA-A389	(ab,cd)	-	1	0	1	0	1	1	-	1	1	1	0	1	0
	SA512		0	1	-	0	1	1	1	1	-	1	1	0	1	0
2	Hi24f04	(ab,cd)	1	0	1	1	0	-	0	0	0	1	0	1	0	1

LG	Locus	Classification	11	12	16	19	51	52	130	276	320	327	330	335	344	353
2	CH03d10	(ab,cd)	1	0	-	1	0	0	0	0	0	0	0	1	0	1
2	CH02a04y	(ab,cd)	1	0	1	1	0	0	0	0	0	0	0	1	0	1
2	CH02c061	(ab,cd)	1	0	1	1	0	0	0	0	0	0	0	1	0	1
2	CH05e03	(ab,cd)	1	0	1	1	0	1	0	0	0	0	0	1	0	1
3	CH03g07	(ab,cd)	1	0	0	0	1	0	1	-	1	-	0	1	1	1
3	Hi03d06	(ab,cd)	-	0	0	0	1	-	1	0	-	0	0	1	0	1
3	SA-A310	(ab,cd)	-	0	0	0	0	1	1	0	1	0	0	1	0	1
3	Hi04c10x	(ab,cd)	1	0	0	0	0	1	-	0	1	0	0	1	0	-
3	SA-A293		1	0	0	0	0	1	1	0	0	0	0	1	0	0
3	CH03g12y	(ab,cd)	1	1	0	0	0	1	1	0	0	0	0	1	0	0
4	Hi23g02	(ab,cd)	0	0	1	0	0	0	1	1	0	-	1	0	1	1
4	SA-A417	(ab,cd)	1	0	1	0	0	-	1	1	0	-	1	0	-	1
4	SA-A615		0	0	1	0	0	0	1	1	-	0	1	0	1	1
4	CH01d03	(ab,cd)	1	0	1	0	0	0	1	1	0	0	1	0	-	1
4	CH02h11a	(ab,cd)	1	0	1	0	1	0	1	1	0	0	1	0	-	-
4	CH01b09b	(ab,cd)	1	0	1	0	0	0	1	1	0	0	1	0	1	1
4	CH05d02	(ab,cd)	1	0	1	0	0	0	1	1	0	0	1	0	0	0
4	CH02c02b	(ab,cd)	1	0	0	1	1	0	1	1	0	0	1	0	0	0
5	Hi22f12	(ab,cd)	1	0	0	0	1	0	0	0	1	0	0	0	0	1
5	SA-A335	(ab,cd)	1	0	0	0	1	0	0	-	1	0	-	0	-	-
5	SA-A340	(ab,cd)	1	-	0	-	1	-	-	-	1	0	0	0	-	1
5	CH03a09	(ab,cd)	1	0	0	0	1	0	0	0	1	0	0	0	0	1
5	CH05e06	(ab,cd)	1	1	0	-	1	0	0	-	-	0	0	0	0	-
5	SA-A750		1	0	0	0	1	0	0	0	-	0	0	0	0	1
5	SA-A279	(ab,cd)	0	0	0	0	1	0	0	0	1	0	0	0	0	1
5	SA-A401	(ab,cd)	1	0	0	0	1	1	0	0	1	0	0	0	1	1
5	Hi04d02	(ab,cd)	0	0	1	1	1	1	0	0	1	0	0	0	1	1
5	CH03e03	(ab,cd)	0	0	1	1	1	1	0	0	1	0	0	0	1	0
5	CH05f06	(ab,cd)	0	0	-	1	1	1	0	0	1	0	0	0	1	0
6	23g4	(ab,cd)	0	0	-	1	1	0	1	0	1	1	1	0	1	0
6	Hi04d10	(ab,cd)	0	0	1	1	1	-	0	1	1	1	1	0	0	0

LG	Locus	Classification	11	12	16	19	51	52	130	276	320	327	330	335	344	353
6	SA-A307	(ab,cd)	-	1	1	-	-	0	0	1	-	1	1	0	-	-
6	CH05a05	(ab,cd)	0	1	1	0	0	0	0	0	1	1	1	0	0	0
7	CN444794-SSR	(ab,cd)	1	0	0	0	0	-	1	0	-	1	1	1	0	1
7	SA-A718	(ab,cd)	1	0	0	0	0	0	1	-	1	1	1	1	0	1
7	Hi03a10	(ab,cd)	0	1	0	0	0	1	1	1	0	1	1	1	1	0
7	SA-A233	(ab,cd)	-	1	-	-	-	1	0	1	0	1	1	1	0	0
7	SA-A425		-	1	-	0	0	1	0	1	-	0	1	0	0	0
7	SA-A332			-	1	0	0	1	0	0	-	0	1	0	0	0
8	Hi04b12	(ab,cd)	1	0	0	0	0	0	-	1	1	1	1	0	0	-
8	CH05a02y	(ab,cd)	1	0	0	1	-	1	1	0	1	1	1	-	-	0
8	CH01c06	(ab,cd)	1	0	0	0	0	0	0	0	0	1	1	0	0	0
8	CH01e121	(ab,cd)	1	0	0	-	0	0	0	0	0	-	1	-	0	0
8	Hi20b03	(ab,cd)	1	0	0	0	0	0	0	0	0	1	1	0	-	0
8	SA-A599	(ab,cd)	0	1	1	1	1	0	1	0	1	1	1	1	0	0
8	CH01h101	(ab,cd)	0	1	0	1	1	0	1	0	0	1	-	1	0	0
9	CH01f03b	(ab,cd)	0	1	1	1	1	1	1	0	0	0	0	1	1	1
9	NZmsEB116209	(ab,cd)	0	1	1	1	1	1	1	0	0	-	-	-	1	1
9	SA-A383	(ab,cd)	0	1	1	1	-	-	1	0	0	0	0	1	1	-
9	Hi23d06	(ab,cd)	0	1	1	-	1	1	1	0	0	0	0	1	1	1
9	Hi23d02	(ab,cd)	0	1	1	-	1	1	1	0	0	0	0	1	1	1
9	Hi05e07	(ab,cd)	0	1	1	1	1	1	1	0	0	0	0	1	1	1
9	Hi04a05	(ab,cd)	0	1	1	1	1	1	1	0	0	0	0	0	0	1
9	Hi01d01	(ab,cd)	-	1	1	1	1	-	1	1	-	1	0	-	0	1
9	SA-A334	(ab,cd)	-	1	-	-	-	1	1	-	-	0	0	0	0	0
10	CH02b07	(ab,cd)	0	1	-	0	0	1	0	0	0	1	0	0	0	1
10	SA-A659b		-	1	-	0	0	1	0	0	0	1	0	0	1	1
10	SA-A659a		-	1	-	0	0	1	0	1	0	1	0	0	1	1
10	SA-A345		0	1	-	0	0	1	0	1	0	1	0	0	1	1
10	CH01f12	(ab,cd)	0	1	1	0	0	1	0	1	0	1	0	1	1	1
10	CH02c11	(ab,cd)	0	1	1	0	0	1	0	0	1	1	0	1	-	1
10	CH04g09	(ab,cd)	0	1	1	1	0	1	1	0	1	1	0	1	1	1

LG	Locus	Classification	11	12	16	19	51	52	130	276	320	327	330	335	344	353
10	SA-A253	(ab,cd)	-	1	0	0	0	1	0	-	0	0	0	-	-	-
10	SA-A326	(ab,cd)	0	1	1	-	0	-	1	-	-	0	0	0	-	1
10	SA-A462		-	1	1	1	0	1	1	-	0	0	1	0	0	1
10	Hi08h12	(ab,cd)	0	1	1	1	0	-	-	0	0	0	1	0	0	-
10	COLa	(ab,cd)	0	1	0	1	0	1	1	0	0	0	-	-	1	1
11	CH02d12	(ab,cd)	0	1	-	-	0	1	1	0	0	1	-	1	-	0
11	CH02d08	(ab,cd)	1	1	0	0	0	0	1	0	0	1	0	1	0	0
11	CH04a12	(ab,cd)	-	-	0	0	0	-	1	0	0	1	0	1	0	0
11	SA-A726	(ab,cd)	1	1	-	-	0	1	1	1	0	1	0	1	0	0
11	Hi02a09	(ab,cd)	1	1	-	-	0	1	1	1	0	1	0	1	0	0
11	CH05c02	(ab,cd)	1	1	0	0	0	1	1	1	0	1	0	1	0	0
11	SA-A504		1	1	-	0	0	1	1	1	0	1	0	1	0	0
11	CH04g07	(ab,cd)	1	1	1	0	0	1	1	0	0	1	0	1	0	1
11	SA-A485		0	1	-	0	0	1	1	1	-	1	0	1	0	1
11	SA-A756	(ab,cd)	1	1	1	0	0	1	1	-	1	-	0	1	-	1
11	CH02g01		1	1	0	0	0	0	0	-	0	0	0	1	0	0
12	SA-A219	(ab,cd)	1	1	0	1	0	1	0	1	0	1	1	0	-	0
12	28f4	(ab,cd)	1	1	1	1	0	1	0	1	1	1	1	0	-	0
12	CH01g121	(ab,cd)	1	0	1	1	0	1	0	1	1	0	1	0	0	0
12	CH04d02	(ab,cd)	1	0	1	1	0	1	0	1	1	0	1	0	-	0
12	SA-A656		0	0	1	1	0	1	0	1	1	0	1	0	0	0
12	CH05d11	(ab,cd)	1	0	1	1	0	1	0	1	1	0	1	0	0	0
12	CH01d09	(ab,cd)	1	0	1	1	0	1	0	-	-	0	1	0	1	0
12	CH01f021	(ab,cd)	1	0	1	1	0	1	-	1	1	0	1	0	1	0
12	CH02h11b	(ab,cd)	1	0	1	1	0	1	0	1	1	0	1	0	1	0
13	CH05h05	(ab,cd)	1	1	0	0	0	1	1	-	-	0	1	1	0	0
13	SA-A392	(ab,cd)	1	1	0	0	0	1	1	0	0	0	1	-	0	0
13	NH009b		1	1	0	1	1	1	1	0	0	0	1	0	1	1
13	SA-A193		0	0	0	1	1	1	1	0	0	0	1	0	1	1
13	SA-A440		1	0	0	1	1	1	1	-	0	0	-	-	1	-
13	CH03h03		1	0	0	1	1	1	1	1	0	0	1	0	1	1
13	SA-A631		0	0	0	0	1	1	1	0	0	0	1	1	1	0

LG	Locus	Classification	11	12	16	19	51	52	130	276	320	327	330	335	344	353
14	NZmsEB146613	(ab,cd)	0	0	0	0	1	1	1	0	0	0	1	1	1	0
14	CH03d08	(ab,cd)	0	0	0	0	0	1	1	0	0	0	0	1	1	0
14	CH01g05	(ab,cd)	0	0	0	0	0	1	1	0	0	0	0	1	1	0
14	CH05g11	(ab,cd)	0	0	0	0	0	1	0	0	0	0	0	1	1	0
14	Hi02d11	(ab,cd)	0	0	-	-	0	1	0	1	0	0	0	1	1	0
14	SA-A222	(ab,cd)	0	0	0	0	0	1	0	0	-	0	0	-	1	0
14	SA-A523	(ab,cd)	0	0	0	0	0	1	0	1	0	0	0	1	1	0
14	SA-A754	(ab,cd)	0	0	0	-	0	1	0	1	0	0	0	1	1	0
14	Hi01c09	(ab,cd)	0	0	0	-	0	1	0	1	0	0	0	1	1	0
14	SA-A652		0	0	-	0	0	1	0	1	-	0	0	1	1	0
14	SA-A324		0	1	-	0	0	1	0	1	0	0	0	1	1	0
14	Hi03a03y	(ab,cd)	0	1	0	0	0	-	0	0	-	0	0	1	1	0
15	SA-A663		0	1	-	0	0	1	0	0	0	0	1	1	0	1
15	CH01d08	(ab,cd)	0	1	0	0	0	1	0	0	0	0	1	1	1	1
15	SA-A186	(ab,cd)	0	-	0	0	0	-	0	0	0	0	-	1	0	-
15	Hi03g06	(ab,cd)	0	1	0	0	0	1	0	0	0	0	1	1	1	1
15	SA-A238	(ab,cd)	0	1	0	-	0	1	0	0	-	0	-	1	-	-
15	SA-A516		-	1	0	0	0	1	0	0	0	0	1	1	1	1
15	SA-A716		0	1	0	0	0	1	0	0	-	0	1	1	1	1
15	CH03b10	(ab,cd)	1	1	0	0	1	0	0	1	1	0	1	0	1	1
15	SA-A478		0	1	0	1	1	0	0	0	1	0	1	1	0	1
15	SA-A244	(ab,cd)	0	1	0	0	-	0	0	0	1	-	1	1	0	1
15	SA-A535	(ab,cd)	0	1	0	0	1	0	0	-	1	0	1	1	0	1
15	SA-A320	(ab,cd)	0	1	0	-	0	0	0	-	1	0	1	1	0	-
15	Hi23g12	(ab,cd)	0	1	0	0	0	0	-	1	1	0	1	1	1	-
15	SA-A354		-	1	-	0	0	0	0	0	1	0	1	1	1	1
15	SA-A344	(ab,cd)	-	1	0	0	0	0	0	-	1	0	1	1	1	1
15	SA-A419	(ab,cd)	1	1	1	1	1	0	0	0	1	0	1	1	1	1
15	CH02c09	(ab,cd)	-	1	0	0	0	0	0	1	1	-	1	1	1	1
16	NZmsCO905522	(ab,cd)	1	0	1	0	0	1	0	0	0	1	1	1	1	0
16	CH05e04	(ab,cd)	1	0	1	0	-	1	0	-	0	1	1	1	1	0
16	SA-A601	(ab,cd)	0	1	1	0	-	1	1	0	0	1	0	1	1	1
16	CH05a04	(ab,cd)	0	1	1	0	1	1	1	0	0	1	0	1	1	1
16	SA-A298		0	1	-	0	1	1	1	0	0	0	0	1	0	1

LG	Locus	Classification	11	12	16	19	51	52	130	276	320	327	330	335	344	353
17	CH05g03	(ab,cd)	0	0	-	1	0	0	0	1	1	0	1	0	1	0
17	CH02g04	(ab,cd)	0	0	1	1	0	0	0	1	1	0	1	1	1	1
17	CH04c06	(ab,cd)	0	0	1	1	0	0	0	0	1	0	1	1	0	-
17	CH01h011	(ab,cd)	0	0	1	1	0	0	0	0	1	0	1	1	0	1
17	SA-A736	(ab,cd)	0	1	1	1	1	0	0	0	1	0	1	1	0	1
17	SA-A413		0	1	-	1	1	0	0	0	-	0	1	1	0	1
17	SA-A236	(ab,cd)	0	1	1	0	1	0	0	1	1	0	1	1	0	1
17	Hi07h02	(ab,cd)	0	1	1	0	1	-	0	-	1	0	1	1	0	1
17	SA-A234	(ab,cd)	0	1	1	1	1	1	0	0	1	0	1	1	0	-