UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Tracking nucleotide-binding-site-leucine-rich-repeat resistance gene analogues in the wheat genome complex

by

Franco B du Preez

Submitted in fulfillment of the requirements for the degree

Magister Scientiae

in the

Faculty of Natural and Agricultural Sciences.

Department of Genetics

University of Pretoria

Pretoria

May 2005

Prof A-M. Botha-Oberholster (Supervisor)

Dr. A.A. Myburg (Co-supervisor)

Declaration

I, the undersigned, hereby declare that the dissertation submitted herewith for the degree M.Sc. to the University of Pretoria, contains my own independent work and has not been submitted for any other degree at any other university.

_____

Franco du Preez

May 2005

Dedication

I dedicate this study to my parents, without whom's support this study would not have been possible.

Acknowledgements

I hereby acknowledge the parties aiding me in completion of this study:

First and foremost, my promoter and study leader, Prof Anna-Maria Botha-Oberholster.

My co-promoter Dr Alexander Myburg.

Dr Fourie Joubert for supervision regarding bioinformatics aspects.

The Cereal Genomics Research group.

The Genetics Department, the Forestry and Agricultural Biotechnology Institute and the Bioinformatics Institute at the University of Pretoria for access to laboratories, equipment and computational facilities.

All fellow students for their insightfull input and discussions and their aid in the laboratory, in particular Dr Eddie Venter (SCRI) and Dirk Swanevelder.

The National Research Foundation and Winter Cereal Trust for their financial assistance.

My parents for their love and support through the duration of this study.

Contents

List of Figures

List of Tables

List of Abbreviations

ATPase	Adenosine Triphosphate Phosphatase
Avr	Avirulence
BLAST	Basic Local Alignment Search Tool
CARD	Caspase Recruitment Domain
CC	Coiled Coil
cDNA	Complimentary DNA
CNL	CC-NBS-LRR
DNA	Deoxyribo Nucleic Acid
EDS1	Enhanced Disease Susceptibility 1
EM	Expectation Maximizer
EMBOSS	European Molecular Biology Open Source Suite
EST	Expressed Sequence Tag
E-value	Expectation value
FTP	File Transfer Protocol
GI	Gene Index
GNBP1	Gram-negative binding protein 1
GTPase	Guanosine Triphosphate Phosphatase
HMM	Hidden Markov Model
HR	Hyper Sensitive Response
IRAK	Interleukin Receptor-Associated Kinase
ITEC	International *Triticeae* EST cooperative
Iκβ	Inhibitory factor κβ
Ka	Nonsynonymous substitution rate
Ks	Synonymous substitution rate
LAM	Local Area Multicomputer
LPS	Lipopolysacharide
LRR	Leucine-Rich Repeat
LTA	Lipoteichoic Acid
LZ	Leucine Zipper
MEME	Multiple EM for Motif Elicitation
ML	Maximum Likelihood
MPI	Message Passing Interface
NBS	Nucleotide-Binding Site
NCBI	National Center for Biotechnology Information
NF-κβ	Nuclear Factor κβ

NIK	Nuclear Factor κβ-inducing kinase
NOD	Nucleotide-Binding Oligomerization Domain
NPR1	Non-expresser of PR genes 1
PAM	Point Accepted Mutations
PAMP	Pathogen Associated Molecular Pattern
PCD	Programmed Cell Death
PCR	Polymerase Chain Reaction
Pfam	Protein Families
PGRP-SA	Peptidoglycan Recognition Protein SA
PHI-BLAST	Position Hit Iterated BLAST
P-loop	Phosphate binding loop
PR	pathogenesis related proteins
PRI	porcine ribonuclease inhibitor
PSI-BLAST	Position Specific Iterated BLAST
***R* gene**	Resistance gene
RGA	Resistance gene analogue
RLK	Receptor Like Kinase
RNA	Ribo Nucleic Acid
RNBS	Resistance Nucleotide-Binding Site
ROI	Reactive Oxygen Intermediate
ROS	Reactive Oxygen Species
RT-PCR	Reverse Transcriptase PCR
SAR	Systemic Acquired Resistance
TIGR	The Institute for Genomic Research
TIR	*Toll* Interleukin Receptor homology domain
TLR	Toll-like receptor
TMV	Tobacco Mosaic Virus
TNL	TIR-NBS-LRR
TRAF	Tumor necrosis Factor Associated Factor

Chapter 1

Introduction

1.1 Nucleotide-binding-site-leucine-rich-repeat genes

The key role that wheat has played in the history of modern civilization as well as its centrality in current food security issues is widely acknowledged. Wheat has been with man from the dawn of modern civilization. Its origins can be traced back to a geographic region known as the Fertile Crescent, which extends from Israel, Jordan, Lebanon and western Syria into southeastern Turkey and along the Tigris and Euphrates rivers into Iran and Iraq (Lev-Yadun *et al.*, 2000). Up to ninety per cent of the current world population is dependent on wheat-derived products for daily sustenance (Moore *et al.*, 1993). Due to the limited and deteriorating agricultural resources currently available, an increased yield per unit area is the only viable solution for increasing global yield in accordance with the demands of a rapidly growing global populace (Gregory and Ingram, 2000).

Several factors are of importance in determining wheat yield, including climatic conditions, agricultural practice, and the outbreak of pests and diseases (Gurr *et al.*, 1992). Using conventional breeding methods, many wheat cultivars resistant to the biotic and abiotic stresses associated with these factors have been produced in the past. With the emergence of modern molecular biology, it is now possible to directly modify the genetic constitution of selected cultivars for increased resistance to these stress factors (Melchers and Stuiver, 2000). Major advances in our current understanding of the molecular biology of these responses have allowed genetic engineering of plants to withstand a multitude of unfavourable conditions including insect infestation, herbicide application, disease outbreaks and environmental extremes (reviewed in Dunwell, 2000).

Both agricultural practice and the genetic constitution of crop species elevate the risk and associated damages incurred by disease outbreaks (Gurr *et al.*, 1992). Due to the serious bottlenecks imposed by domestication and selective breeding, crop species generally contain little genetic variability (Nei *et al.*, 1975; Marayuma and Fuerst, 1985) and show high vulnerability to new epidemics. The high densities at which these organisms are produced commercially also facilitate rapid spreading of infectious agents. The Irish potato famine of the 1840's heeds one to seriously consider plant-pathogen interactions when discussing global food security issues (Goodwin *et al.*, 1994).

Investigations into plant-pathogen interactions have provided us with several models underlying the genetic basis of host resistance (reviewed in Hammond-Kosack and Jones, 1997). In the past decade, tens of resistance genes have been isolated from numerous crop and

model plant species. When classifying the encoded disease resistance proteins based on their domain structure, only a few classes are formed. The majority of these cloned resistance genes are members of the nucleotide-binding-site-leucine-rich-repeat (NBS-LRR) gene family (Martin, 2003). To date no NBS-LRR genes have been assigned functions outside the scope of resistance induction (Belkhadir *et al.*, 2004).

With the advent of the genomic era, ever increasing amounts of sequence data is generated globally and stored in public databases (The *Arabidopsis* Genome Initiative, 2000; Yu *et al.*, 2002). Due to the massive size of the hexaploid wheat genome (Arumuganathan and Earle, 1991) and limitations in current genome sequencing technologies, a draft of the complete genome sequence will not be produced for a few years to come (Gill *et al.*, 2004). Since the vast majority of the wheat genome appears to consist of non-coding sequences, transcript sequencing has been taken as an alternative route to molecular characterization of the wheat genome. A multi-project effort including the International *Triticeae* Expressed Sequence Tag (EST) Cooperative (ITEC) (http://wheat.pw.usda.gov/genome) and the NSF-supported US wheat EST project (http://wheat.pw.usda.gov/NSF) has recently generated over 500 000 ESTs for wheat and 300 000 ESTs for barley.

1.2 Thesis questions and layout

In the present study I aimed to characterize the domain structure, diversity and evolution of the CC-NBS-LRR (CNL) gene family in cereal species of the *Triticeae* tribe, in context of current models of the evolution of this multigene family in other plant taxa. my first objective to this end was to establish a comprehensive dataset of publically available sequences for NBS domains of the NBS-LRR gene family. Using this dataset I aimed to characterize firstly conserved motifs in the NBS domains, to determine whether they represent the CNL families characterized in other plant species, and to consider any evidence for TIR-NBS-LRR (TNL) type NBS domains. I further aimed to study the relationship of *Triticeae* NBS-LRRs clades with functional CNL *R* genes by performing a number of phylogenetic analyses on the union of these two datasets. I also aimed at characterizing the evolution of the gene family at the hand of existing models of multi-gene, and more specifically, *R* gene evolution.

Models of multigene family evolution (Otto and Yong, 2002), built around classic population genetics predict that loci where overdominant selection is possible, are likely to produce the majority of fixed gene duplications observed in natural populations, where new specificities are generated as alleles at a single locus prior to duplication via unequal recombination in a

heterozygote as opposed to previous applications of the birth-and-death model where duplication precedes divergence (Michelmore and Meyers, 1998). Considering that numerous NBS-LRR loci with alleles encoding multiple specificities are well known (Ellis *et al.*, 1999; Wei *et al.*, 2002), either balancing or overdominant selection is most likely operating across these loci, and in the context of this model, I aimed to study two duplication events, for which this model predicts different outcomes: paralogous gene duplications (functional divergence) and allopolyploidy mediated homeologous gene duplications (mutation to pseudogene). In order to study the evolutionary fate of these duplications, I evaluated basic parameters of gene family evolution, including nonsynonymous to synonymous substitution rate (Ka:Ks) ratios and gene conversion rates. I aimed to obtain and study the evolution of NBS-LRR sequences resulting from recent paralogous expansions from the results of my planned phylogenetic analysis, while identifying homeologous NBS-LRR sequences for the A (*Triticum urartu*), B (*Aegilops speltoides*) and D (*Aegilops tauschii*) genomes of hexaploid wheat by PCR using specific primer sets targetted to two previously mapped NBS-LRR sequences, namely *go35* (Lagudah *et al.*, 1997) and *KSU945* (Maleki *et al.,* 2003).

1.3 Research outputs

The following research outputs were generated in collaboration with my promoters and colleagues during the period of my MSc study:

Lacock, L., van Niekerk, C., Loots, S., **Du Preez, F.B**., and Botha, A-M. (2003) Functional and comparative analysis of expressed sequences from *Diuraphis noxia* infested wheat obtained utilizing the Nucleotide Binding Site conserved motif. *African Journal of Biotechnology* 2:4:75-81.

Du Preez, F.B., Myburg, A.A., and Botha, A-M. (2004) Tracking NBS-LRR Gene Family Members in the *Triticeae* Complex. *18th Congress of the South African Genetics Society*. 4-7 April 2004 University of StellenBosch p34.

Du Preez, F.B, and Botha, A-M. (2002). Utilizing degenerate PCR technology to obtain nucleotide binding site-Leucine Rich Repeat (NBS-LRR) proteins from wheat (*Triticum aestivum* L.em Thell.) linked to pest resistance in plants. *4th Plant Breeding Symposium*, Gordons Bay South Africa, 11-14 March 2002.

Botha, A-M., Lacock, L., Van Niekerk, C., Loots, S., and **Du Preez, F.B**. (2001). Evaluation of Expressed Sequence Tags (ESTs) from Russian Wheat Aphid induced wheat cDNA libraries. *The Quadrennial joint annual meetings of the American Society of Plant Biology and the Canadian Society of Plant Physiologists,* Providence, Rhode Island, USA, 21-25 July 2001.

Botha, A-M., Lacock, L., Van Niekerk, C., Loots, S. and **Du Preez, F.B.** (2001) Comparative analysis of expressed sequence tags from wheat. *Biologie Cellulaire*, INRA-Versailles, France, December 2001.

Botha, A-M., Lacock, L., Van Niekerk, C., Matsioloko, M. T., **Du Preez, F. B.,** Myburg, A. A., Kunert, K. and Cullis, C. A. (2003). Gene expression profiling during Diuraphis noxia infestation of *Triticum aestivum* cv. 'Tugela DN' using microarrays. *Proceedings of the 10th international Wheat Genetics Symposium, Paestum, Italy, (1-6 September 2003).* 1:334-338.

Botha A-M., Lacock, L., Van Niekerk, C., Loots, S., and **Du Preez, F.B.** (2002) A study into plant-insect interactions using cDNA microarrays : the Russian wheat aphid-wheat model. *Planteforsk,* Å

Botha, A-M., Lacock, L., Van Niekerk, C., Matsioloko, M.T., **Du Preez, F.B.** Loots, S., Venter, E. Kunert, K.J., and Cullis C.A. (2005). Is Photosynthetic Transcriptional regulation in Triticum aestivum L. cv. 'TugelaDN' a contributing factor for tolerance to Diuraphis noxia (Homoptera: Aphididae)? *Plant Cell Reports* (in press).

1.4 References

Arumuganathan, K., and Earle, E.D. (1991) Nuclear DNA content of some important plant species. *Plant Molecular Biology Reports* 9:208–218.

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* 4:391-399.

Dunwell, J.M. (2000) Transgenic approaches to crop improvement. *Journal of Experimental Botany* 51:487-496.

Ellis, J.G., Lawrence, G.J., Luck, J.E., and Dodds, P.N. (1999) Identification of regions in alleles of the flax rust resistance gene *L* that determine differences in gene-for-gene specificity. *Plant Cell* 11:495-506.

Gill, B.S., Appels, R., Botha-Oberholster, A-M., Buell, C.R., Bennetzen, J.L., Chalhoub, B., Chumley, F., Dvorak, J., Iwanaga, M., Keller, B., Li, W., McCombie, W.R., Ogihara, Y., Quetier, F. and Sasaki, T. (2004) A workshop report on wheat genome sequencing: International Genome Research on Wheat Consortium. *Genetics* 168:1087-1096.

Goodwin, S.B., Cohen, B.A., and Fry, W.E. (1994) Panglobal Distribution of a Single Clonal Lineage of the Irish Potato Famine Fungus. *Proceedings of the National Academy of Sciences USA* 91:11591-11595.

Gregory, P.J., and Ingram, J.S.I. (2000) Global change and food and forest production: future scientific challenges. *Agriculture, Ecosystems and Environment* 82:3–14.

Gurr, S.J., McPherson, M.J., and Bowles, D.J. (1992) Molecular Plant Pathology. A Practical Approach. *Oxford University Press*.

Hammond-Kosack, K.E., and Jones, J.D.G. (1997) Plant disease resistance genes. *Annual Review of Plant physiology and Plant Molecular Biology* 48:575-607.

Lagudah, E.S., Moullet, O., and Appels, R. (1997) Map based cloning of a gene sequence encoding a nucleotide binding domain and a leucine-rich repeat region at the Cre3 nematode resistance locus of wheat. *Genome* 40:659-665.

Lev-Yadun, S., Gopher, A., and Abbo, S. (2000) The cradle of agriculture. *Science* 288:1602-2603.

Maleki, L., Faris, J.D., Bowden, R.L., Gill, B.S., and Fellers, J.P. (2003) Physical and genetic mapping of wheat kinase analogs and NBS-LRR resistance gene analogues. *Crop Science* 43:660-670.

Martin, G.B., Bogdanove, A.J., and Sessa, G. (2003) Understanding the functions of plant disease resistance proteins. *Annual Review of Plant Biology* 54:23-61.

Maruyama, T., and Fuerst, P.A. (1985) Population bottlenecks and non equilibrium models in population genetics. II. Number of alleles in a small population that was formed by a recent bottleneck. *Genetics* 111:675-689.

Melchers, L.S., and Stuiver, M.H. (2000) Novel genes for disease resistance breeding. *Plant Biotechnology* 3:147-152.

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* 8:1113–1130.

Moore, G., Gale, M.D., Kurata, N., and Flavell, R.B. (1993). Molecular Analysis of small Grain cereal genomes : Current status and prospects. *Biotechnology* 11:584-589.

Nei, M., Maruyama, T., and Chakraborty, R. (1975) The bottleneck effect and genetic variability in populations. *Evolution* 29:1-10.

Otto, S.P., and Yong, P. (2002) The evolution of gene duplicates. *Advanced* Gene*tics*. 46:451-83.

The *Arabidopsis* Genome Initiative (2000) Analysis of the genome of the flowering plant *Arabidopsis thaliana. Nature*. 408: 796-815.

Wei, F., Wing, R.A., and Wise R.P. (2002). Genome Dynamics and Evolution of the *Mla* (Powdery Mildew) Resistance Locus in Barley. *Plant Cell* 14:8:1903–1917.

Yu, J., Hu, S., Wang, J., Wong, G.K., Li, S., Liu, B., Deng, Y., Dai, L., Zhou, Y., Zhang, X., Cao, M., Liu, J., Sun, J., Tang, J., Chen, Y., Huang, X., Lin, W., Ye, C., Tong, W., Cong, L., Geng, J., Han, Y., Li, L., Li, W., Hu, G., Huang, X., Li, W., Li, J., Liu, Z., Li, L., Liu, J., Qi, Q., Liu, J., Li, L., Li, T., Wang, X., Lu, H., Wu, T., Zhu, M., Ni, P., Han, H., Dong, W., Ren, X., Feng, X., Cui, P., Li, X., Wang, H., Xu, X., Zhai, W., Xu, Z., Zhang, J., He, S., Zhang, J., Xu, J., Zhang, K., Zheng, X., Dong, J., Zeng, W., Tao, L., Ye, J., Tan, J., Ren, X., Chen, X., He, J., Liu, D., Tian, W., Tian, C., Xia, H., Bao, Q., Li, G., Gao, H., Cao, T., Wang, J., Zhao, W., Li, P., Chen, W., Wang, X., Zhang, Y., Hu, J., Wang, J., Liu, S., Yang, J., Zhang, G., Xiong, Y., Li, Z., Mao, L., Zhou, C., Zhu, Z., Chen, R., Hao, B., Zheng, W., Chen, S., Guo,

W., Li, G., Liu, S., Tao, M., Wang, J., Zhu, L., Yuan, L., and Yang, H. (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science* 296:5565:79-92.

Chapter 2

Literature Review

2.1 Introduction

2.1.1 Background

Since plants form the carbon fixating foundation of all terrestrial ecologies, it is not surprising that a vast range of organisms have evolved life styles that exploit the nutrient rich environment provided by plant tissue. In order to deal with this onslaught, plants have acquired multiple levels of defense strategies. A multitude of preformed defense mechanisms be it physical such as bark, the cuticle or the cell wall, or chemical, such as the myriad of anti-microbial compounds present in plant tissues, prevent all but the best adapted phytopathogens from gaining a foothold in plant tissue (Dennis *et al.*, 1997; Cassab and Varner, 1998). In cases where all these preformed defenses fail, plants can induce a localized defense response. This localized response includes extreme measures such as the production of reactive oxygen intermediates (ROIs) and ultimately programmed cell death (PCD), creating an environment that very few pathogens can survive.

The major weakness of the localized response is that it has to be induced, unlike the preformed chemical and physical defenses. Since induction is dependent on successful detection of the invading phytopathogen, the plant needs a sophisticated surveillance system. Thus, much like specialized animal pathogens, plant pathogens have to continually evolve new ways of avoiding detection while host surveillance systems have to evolve new recognition specificities. Both recognition and induction are molecular processes mediated by specific protein interactions, ultimately allowing the cell's transcriptome to react to environmental changes such as pathogen challenge.

The first step toward unraveling the molecular basis of host-phytopathogen interactions was made by H.H. Flor in the 1940's. Flor studied the host-pathogen interaction between flax (*Linum usitatissimum*) and the fungal rust (*Melamspora lini*). His initial observation was that the majority of flax resistance genes were inherited as single dominant genes. Furthermore, each resistance gene was only capable of providing resistance against specific rust isolates. Studying the inheritance patterns of the rust's ability to elicit a defense response led Flor to his so-called "gene-for-gene" hypothesis. This states that a complementary resistance and avirulence gene must be present in the host and pathogen respectively, in order to facilitate induction of the host's resistance response, leading to an incompatible interaction. In the

absence of either or both these genes, the pathogen is not detected and no localized response is induced in the host, resulting in a compatible interaction (Flor, 1971).

Because the ancestors of modern crop plants have centers of origin and have co-evolved with pathogens, many breeding projects were initiated (Lepik, 1970) to introgress resistance genes from wild crop relatives into modern cultivars. This resulted in the development of several "gene-for-gene" model systems where susceptible/resistant near isogenic lines (NILs) became available. The *Pto* mediated resistance of tomato to *Pseudomonas syringae* pv *tomato* strains harboring the *avr-Pto* gene, was one of these systems that had been well studied at the time (Martin *et al.*, 1991).

With the advent of modern molecular biology, the first race specific resistance genes (*R* genes) were isolated using these model systems as well as the model plant, *Arabidopsis thaliana*. A positional cloning approach yielded the first isolated *R* gene – *Pto* from tomato (Martin *et al.*, 1993). *Arabidopsis* researches soon followed by isolating *RPS2*, and later *RPM1, RPS4* and *RPS5,* each providing resistance to strains of *Pseudomonas syringae* bearing complementary avirulence genes (Kunkel, 1996).

2.1.2 Classes of Phytopathogenic organisms

2.1.2.1 Viruses

Incapable of breaching the cuticle or cell-wall, plant viruses are dependant on biological vectors such as insects, nematodes, fungi and pollen (Gurr, 1992). Several *R* genes that provide plants with resistance to plant viruses have been isolated in the last decade (Whitham *et al*., 1994; Bendahmane *et al*., 1999; Cooley *et al*., 2000). One of the earliest and best-studied plant-virus interactions is that between tobacco and the tobacco mosaic virus (TMV). The single dominant *N R* gene provides resistance to TMV and other tobamoviruses (Tobias *et al*., 1982). The avirulence factor encoded by the virus is the helicase domain of the 126 kDa replicase protein (Erickson *et al.*, 1999).

2.1.2.2 Bacteria

Only a small number of bacterial genera contain true phytopathogens, capable of causing disease on otherwise healthy plants. Prominent Gram-negative genera include *Agrobacterium* (galls and cankers), *Erwinia* (soft rot), *Pseudomonas* (leaf spot) and *Xanthomonas* (leaf spot). Unlike fungi, bacteria are incapable of penetrating healthy plant cells and must enter through stomata, hydathodes or wounds (including those induced by insect feeding). The site of entry

strongly determines tissue localization, with stomatal entrants being restricted to the intercellular leaf spaces, hydathodal entrants to the vascular system and phloem-feeder associated bacteria to the phloem. In general, a phytopathogenic bacterial strain is only capable of infecting a single plant species or a few closely related plant species. *Pseudomonas* syringae and *Xanthmonas campestris* provide an exception to this rule, being capable of infecting a wide range of plant families (Gurr, 1992). The *Arabidopsis–Pseudomonas* syringae interaction is currently one of the best studied host-pathogen interactions and several *R-Avr*-gene pairs involved have been molecularly characterized (Kunkel, 1996).

2.1.2.3 Fungi

Phytopathogenic fungi have evolved diverse strategies for gaining access to plant tissues and individual plant cells. Modified hyphal structures are commonly employed for initial penetration. Specialized feeding structures known as haustoria are passed through the cell wall of a targeted plant cell where they provide a large surface area for extracting nutrients (Mendgen *et al*., 1996). Depending on the specific feeding strategy employed, phytopathogenic fungi can be divided into three broad classes.

The first class consists of the necrotrophs, which kill host cells prior to feeding on them. Killing of host cells is achieved by the introduction of toxins targeted to specific host components. The virulence genes encoding these toxins follow dominant inheritance patterns in pathogen populations due to their functional role. Host resistance on the other hand is usually inherited recessively in plant populations, because resistance frequently occurs due to the loss or modification of toxin targeted host components. In some cases however, the resistance genes encode enzymes capable of active detoxification. Such resistance genes follow a dominant inheritance pattern. *Hm1* from maize is a typical example of one such resistance gene, since it detoxifies the HC-toxin produced by the leaf spot fungus *Cochliobolus carbonum.* (Johal and Briggs, 1992; Walton, 1996).

The second class, known as biotrophs derive their nutrients from living host cells and usually have a very narrow host range due to the complex relationship that they have with host cells. Biotrophs have to subvert both the host's metabolism in order to favor their own growth as well as the host's defense responses in order to stay undetected (Agrios, 1998). As a result, mutation of the virulence genes involved in these processes and/or the evolution of new host genes capable of detecting pathogen borne gene products lead to an incompatible reaction in which the host's defense responses are activated. Biotroph virulence genes are thus usually inherited in a dominant fashion due to the functional role that they fulfill during pathogenesis.

Downey and powdery mildews, rusts and smuts, adopt this colonization strategy with the flax rust (*Melamspora lini*) – flax (*Linum usitatissimum*) interaction being the classical example studied by H.H. Flor (1971).

The third group of phytopathogenic fungi consists of the hemi-biotrophs. These pathogens often have a broader host range than biotrophic pathogens and differ in their feeding strategy by killing host cells during the later stages of infection (Agrios, 1998). The genetic basis of resistance to these organisms is similar to that of biotrophic pathogens. The infection cycle of the oomycete *Phytophthora infestans*, the causative agent of potato late blight, follows a typical hemi-biotrophic strategy, with a short biotrophic phase followed by devastating necrotrophy (Gurr *et al.*, 1992).

2.1.2.4 Nematodes

Phytopathogenic nematodes follow either an endo- or ectoparasitic lifestyle, most species colonizing root tissue. Nematodes feed on plant cells using stylets, and most economically important species modify some plant cells to form specialized feeding sites capable of supporting reproducing females. Economically important nematode species include the cyst and root-knot nematodes (Gurr, 1992). The *R* gene *Mi* from tomato provides resistance to both the root-knot nematode *Meloidogyne incognita* and the potato aphid *Macrosiphum euphorbiae* (Rossi *et al.*, 1998).

2.1.3 Avirulence genes

The functional role of avirulence genes in a compatible host-pathogen interaction is still elusive. *Avr* genes encode diverse and often unrelated products. Mutation studies have indicated a role in virulence for some avirulence genes (Jamir, 2004). Since gram-negative phytopathogenic bacteria have been found to secrete their avirulence products directly into the host cytoplasm via a conserved type III secretion system, it appears that avirulence genes form part of the bacterium's virulence functions (Galan, 1999). Fungal avirulence genes are less well understood and no common secretion system has as yet been characterized. It has recently been shown that the *L5, L6* and *L7 R*-genes of flax (*Linum usitatissimum*) recognize their avirulence complements (encoded by flax rust, *Melamspora lini*) intracellularly and that these avirulence genes are expressed specifically in flax rust haustoria. This strongly indicates that at least some fungal pathogens transport their *avirulence* factors into host cytoplasm as is known for gram-negative bacterial phytopathogens (Dodds, 2004).

2.1.4 Induced Defence Mechanisms

2.1.4.1 Localized Response

The hypersensitive response (HR) is an important component of the local defense response and follows minutes after a pathogenic invasion is detected, typically by the interaction between a complementary *R* and *Avr* gene pair (Goodman, 1994). Some of the first effects of the HR include the production of reactive oxygen species (ROS's) such as H_2O_2 and superoxide radicals ($^{\cdot}O_2$). This "oxidative burst" causes damage to both the host cell and the invading pathogen, and is involved in downstream signalling events and cell wall reinforcement (Bolwell, 1999). A spreading front of cell death becomes visible during later stages of the HR, which limits the invading pathogen's access to nutrients (Stakman, 1915; Goodman and Novacky, 1994; Heath, 2000; Shirasu and Schulze-Lefert, 2000).

2.1.4.2 Systemic Response

Long-range defense-signals travel beyond the HR, inducing a systemic acquired resistance (SAR) throughout plant organs, providing long-term elevation of disease resistance mechanisms. Initial observations indicated salicylic acid as the long range messenger, although grafting experiments suggest that salicylic acid is only required for effecting induction of the SAR response (Ryals *et al.*, 1996). Proteins induced specifically by the SAR response are known collectively as pathogenesis related proteins (PR-proteins). The SAR facilitates production of more than 300 structurally distinct low molecular weight compounds, which play crucial roles in preventing subsequent infection. These compounds are known collectively as phytoalexins and their exact role in disease signaling as well as their anti-microbial properties have not yet been fully characterized (Greenberg ,1996).

Recent studies suggest that the SAR response differentiates between necrotrophic and biotrophic infections by inducing specific defense measures for each (Traw *et al.*, 2003). Biotrophic infections in *Arabidopsis* would for example trigger upregulation of *PR*-1, *PR*-2 and *PR*-5 whereas necrotrophic infection would not (Kunkel and Brooks, 2002). Insect feeding on the other hand, causes specific induction of proteinase inhibitors (Fitandsef *et al,*. 1999) and glucosinolates (Bennet and Wallsgrove, 1994). Two plant hormones, namely salicylic acid and jasmonic acid are central to effecting this specificity. Jasmonic acid is rapidly accumilated in tissue suffering herbivore damage (Reymond *et al*., 2000) or necrotrophic infections (Penninckx *et al.,* 1996*)* whereas salicylic acid accumulates during biotrophic infections (Ton *et al*., 2002). The salicylic- and jasmonic acid SAR pathways also

show significant overlap with regards to signaling components, and many of their defense components are shared, such as the upregulation of peroxidase and exochitinase transcription (Davis *et al.,* 2002).

2.1.5 Classes of *R* genes

2.1.5.1 Background

The isolation and sequencing of disease resistance genes from various plant-pathogen interaction models has greatly increased our understanding of the biochemical basis for induced plant innate defense responses. More than 40 resistance genes have been isolated, (Table 2.1) and these show very interesting relationships at the DNA and amino acid level (Meyers *et al.*, 1999, Martin *et al.*, 2003). Based on the modular domains present, the resistance genes isolated thus far can be subdivided into five classes (Figure 2.1), some of which share functional domains and the signaling network of the HR.

2.1.5.2 Classification

Enzymes involved in detoxification

As discussed in the section on necrotrophic pathogens, this class is represented by the *Hm1* gene of maize (Johal and Briggs, 1992), which actively detoxifies the HC toxin produced by the leaf spot fungus *Cochliobolus carbonum.*

Intracellular serine-threonine protein kinases

A second class is represented by the *Pto* and *PBS1* genes of tomato and *Arabidopsis* respectively. *Pto* provides resistance to an extracellular bacterial pathogen *Pseudomonas syringae (*Martin *et al*., 1993). *Pto* functions as an intracellular serine-threonine protein kinase, and binds directly to it's *Avr*-ligand, *Avr-Pto* which is delivered to the host's cytoplasm via a *P.syringae's* type III secretion system (Scofield *et al.*, 1996). *Pto* function has since been shown to be dependent on the presence of another resistance gene, *Prf* (Rathjen *et al*, 1999), which belongs to the nucleotide-binding-site-leucine-rich-repeat class of resistance genes, which will be discussed next. A similar system has been characterized for the triad consisting of *PBS1* (kinase), *RPS5* (CC-NBS-LRR) and *AvrPphB* (Avirulence protein) (Swiderski and Innes, 2001; Warren *et al.*, 1998).

Nucleotide-Binding-Site-Leucine-Rich-Repeat proteins

The majority of isolated *R* genes are all grouped in the third class of resistance genes called the NBS-LRR genes (Nucleotide-Binding-Site-Leucine-Rich-Repeat). NBS-LRR gene products are located in the cytoplasm of plant cells and typically contain three distinct functional domains (Traut, 1994). NBS-LRR genes can be assigned to one of two sub-groups based on the identity of the N-terminal domain. The N-terminal domain of the first subclass, called the TIR-NBS-LRR, shows homology to domains found in both the Toll receptor of *Drosophila* and the mammalian Inter-Leukin receptor. This domain is known as the TIR-domain (Toll-Interleukin Receptor homology domain) (O'Neill, 2000). In the second subclass, called the CC-NBS-LRR, the N-terminal domain is predicted to form a coiled-coil (CC) structure (Pan *et al.*, 2000). The central domain, called the nucleotide-binding site (NBS), is homologous to the nucleotide-binding site of ATPases, GTPases and various other nucleotide binding proteins (Saraste *et al.*, 1990; Traut, 1994). The C-terminal domain is known as the leucine-rich-repeat (LRR) domain and consists of multiple copies of an imperfect leucine-rich-repeat sequence (Bai et al, 2002). NBS-LRR genes are currently thought to encode cytoplasmic receptors, capable of detecting the presence of pathogen borne avirulence proteins in the host's cytoplasm.

The CC-NBS-LRR class of plant resistance genes includes *RPS2* (Resistance to *Pseudomonas syringae*) from *Arabidopsis* (Bent *et al.*, 1994; Mindrios *et al.*, 1994), which provides resistance against the extracellular bacterial pathogen, *Pseudomonas syringae* (See Table 2.1 for more examples). The *N*-gene from tobacco (Witham *et al.*, 1994), which belongs to the TIR-subgroup of NBS-LR*R* gene products, provides resistance against the tobacco mosaic virus (TMV). Other TIR-NBS-LR*R* genes include *L6* (Lawrence *et al.*, 1995) from flax and *RPP5* (Jones *et al.*, 1994) from *Arabidopsis*. An interesting observation at this point is that no TIR-NBS-LRR genes have to date been isolated from monocotyledons although both classes are present in dicotyledons (Greenberg, 1996).

Transmembrane leucine-rich-repeat receptor-like proteins

The fourth class of resistance genes includes trans-membrane proteins with extracellular LRRs, a transmembrane region and a short cytoplasmic region (Figure 2.1). *Cf-9, Cf-2, Cf-4* and *Cf-5* from tomato all have this domain architecture (Jones *et al.*, 1994; Dixon *et al.*, 1996; Thomas *et al.*, 1997). Each *R* gene provides resistance to isolates of the extracellular biotrophic fungus *Cladosporium fulvum* harboring a specific *Avr* gene (Van den Ackerveken *et al.*, 1992; Joosten *et al.*, 1994; Dixon *et al.*, 1996).

Transmembrane leucine-rich-repeat receptor kinases

The fifth class is represented by the domain architecture found in *Xa-21* from rice (Figure 2.1), which contains in addition to the membrane spanning region and extracellular LRRs of the *Cf* proteins, a cytoplasmic serine/threonine kinase domain. *Xa-21* provides resistance against *Xanthomonas oryzae*, which is an extracellular bacterial pathogen of rice (Song *et al.*, 1995; He *et al.*, 2000).

2.2 The NBS-LRR gene family

2.2.1 Domain Strucure

2.2.1.1 Background

As already described, the third class of *R* genes constitutes the majority of *R* genes isolated thus far (Meyers *et al.*, 1999). Preliminary interpretations of NBS-LRR structure and localization as well as inheritance patterns suggest that these genes could act directly as the primary receptors for pathogen derived ligands (including *Avr* gene products), which subsequently initiate the plant's inducible hypersensitive response. Mutation studies and domain swap experiments however, show deviations from this theory and the field of innate immunity in plants is currently very actively researched (Ellis *et al.*, 1999). Before moving on to current models of *R* gene function, an overview of the key domains found in *R* genes and their functional significance in the innate immune responses of animal models is given, and relationships to plant innate immunity is highlighted. It is important to note at this stage that the field of innate immunity has been resolved to much higher detail in animal models than in plant models and that the signaling networks involved in inducible defense responses share homology over a large spectrum of eukaryotes.

2.2.1.2 The TIR Domain

The TIR domain is found in three different classes of animal genes. The members of all three classes play a central role in the innate immune response by mediating pathogen detection.

Toll-like Receptors

Toll from *Drosophila* is a typical member of the first class, which contains the Toll-like receptors (TLR's) . The Toll-protein forms the central component of a trans-membrane receptor complex that mediates both dorso-ventral axis-formation during *Drosophila* embryogenesis, as well as antimicrobial responses in the adult insect (Anderson, 2000).

TLRs typically have an extracellular N-terminal LRR, a single membrane-spanning region and an intracellular TIR domain. TLR's are also abundant in vertebrates and several have been studied in the past decade (Poltorak *et al.*, 2000; Underhill and Ozinsky, 2002).

Interleukin Receptors

The domain structure of interleukin receptors differ from that of TLR's by having three extracellular N-terminal immunoglobulin-like domains instead of the LRR's found in TLR's. Interleukins are polypeptide-cytokines (intercellular messenger molecules), which are central to the coordination of immune and inflammatory responses (Munford and Hall, 1986).

Signal Transducing Adapter Proteins

The remaining class of animal TIR-containing proteins includes signal transducing adapter proteins such as human MyD88, for which homologues have been characterised from other vertebrate and invertebrate genomes. MyD88 conducts the recognition signal from a TLR complex via a TIR-TIR homodomian interaction to lower level signaling components via another homodomain interaction involving it's N-terminal death domain (Fitzgerald et al., 2001).

TIR-mediated signaling in animal innate immunity

Pathogen Associated Molecular Patterns (PAMPs) are typically components of microbial cell walls such as the lipopolysacharides (LPS) found in Gram-negative bacterial cell walls, lipoteichoic acids (LTA) from Gram-positive bacteria and the mannans produced by fungi. Many different TLRs are found in animal genomes and each type forms part of a receptor complex that responds to different PAMPs. The human *TLR4* receptor for example, is responsible for recognizing the presence of LPS by binding a plasma protein-LPS complex. Binding of this ligand allows the assembly of a receptor associated complex at the cytoplasmic interface of the TLR (Figure 2.2; Poltorak *et al.*, 2000). As already mentioned, TLRs are transmembrane proteins. They occur in the cell membrane as multimers and are

associated with several other proteins to form trans-membrane receptor complexes capable of detecting a variety of PAMPs.

In this receptor complex, the cytoplasmic TIR domain of the TLR molecule interacts via a homo-domain interaction with the C-terminal TIR domain of the MyD88 adapter protein. MyD88 possesses an additional N-terminal death domain, which in turn interacts with the N-terminal death domain of a protein present in the receptor-associated complex called IRAK (Interleukin Receptor-Associated Kinase). This interaction activates IRAKs C-terminal kinase domain, which phosphorylates a protein called TRAF-6 (Tumor Necrosis Factor Associated Factor 6). TRAF-6 is currently thought to link the receptor-associated complex to a signalosome containing Iκβ (Inhibitory factor κβ), Nf-κβ (a transcription factor strongly bound by IkB), NF-κβ-inducing-kinase (NIK) and other accessory proteins (Underhill and Ozinsky, 2002; Kopp and Medzhitov, 1999).

The association of this signaling complex with the TLR receptor complex results in the phosphorylation of I*κβ* (inhibitory factor *κβ*). This allows one of the substrate recognition sub-units of ubiquitin ligase to ubiquinate I*κβ*. This targets I*κβ* for destruction by the proteasome, freeing NF-*κβ* for translocation to the cell nucleus. NF-*κβ* (Nuclear factor *κβ*) is a transcription factor and binds to specific promoter elements in order to alter gene expression patterns in response to a pathogenic invasion as detected by the presence of LPS (Hatada *et al.*, 2002).

A homologous signaling network exists in *Drosophila* downstream of the *Toll* receptor where a gene called *cactus* is an I*κβ* homologue whereas *Dif* is an NF-*κβ* homologue. *Drospohila* has another *NF-κβ* homologue in *Relish*, which controls the expression of genes involved in dorso-ventral axis formation during embryogenesis as mentioned earlier. In the case of the Toll receptor, the ligand is not bacterial LPS but a proteolytically processed version of the spaëtzle protein, which is encoded by the *Drosophila* genome. During embryogenesis dorso-ventral axis-formation is determined by spaëtzle cleavage as performed by an endogenous serine protease cascade. During adulthood spaëtzle cleavage is the result of an as yet uncharacterized protease cascade involving the persephone serine-protease, which is activated under conditions of fungal infection (Hultmark, 2003). Gram-positive bacteria are recognized by peptidoglycan recognition protein SA (PGRP-SA) in conjunction with Gram-negative binding protein 1 (GNBP1) (Gobert *et al.*, 2003).

Candidate TIR signaling plant proteins

In plants, TIR-NBS-LRR, TIR-NBS and TIR-X genes have been identified (Meyers *et al.*, 2002). Many NBS-LRR transcripts can undergo alternative splicing, yielding transcripts containing the TIR, NBS and a short C terminal domain absent in full-length NBS-LRR transcripts. The proteins encoded by these transcripts could potentially fulfill the same role as the MyD88 protein of *Drosophila* via homodomain interactions with the TIR domains of TIR-NBS-LRR proteins (Meyers *et al.*, 2002). It is interesting to note that shorter splicoforms occur only during the early stages of pathogen detection. In the case of resistance to tobacco mosaic virus (TMV) in *Nicotiana*, the resistance gene *N* can only confer resistance if it contains an intron allowing it to undergo alternative splicing (Jordan *et al*., 2002).

2.2.1.3 The Coiled-Coil Domain

Initially NBS-LRR genes lacking the carboxy terminal TIR-domain were thought to posses a leucine zipper (LZ) domain. Recent analysis of many non-TIR NBS-LR*R* genes however, has shown that a coiled-coil (CC) domain is present at the C-terminus (Pan *et al.*, 2000). Coiled-coil domains have a 7-residue-repeat primary structure and a tertiary structure consisting of between two and five helices. The helices found in the CC-domain interface through two hydrophobic amino acids. Leucine zippers are classified as members of this wider class of structural elements and contain repetitive leucine residues, spaced at regular intervals corresponding to the number of residues in a helix revolution, causing the long alpha-helices in LZs to be amphiphylic. LZs are free to interact with other such helices, which are often found in interacting dimers (Lupas, 1966; Bai *et al.*, 2002).

2.2.1.4 The Nucleotide-Binding-Site Domain

The NBS domain consists of two sub-domains. The N-terminal sub-domain is known as the NB-domain and contains the consensus kinase1a, kinase2 and kinase3a motifs shared by many known nucleotide-binding proteins (Moffet *et al*., 2002; Hammond-Kosack and Jones, 1997). The kinase1a motif is also known as the P-loop and binds the phosphate group of the bound nucleotide. The consensus for this motif is GXGXXG(R/K)V. The *RPS2 R* gene from *Arabidopsis* is an NBS-LRR *R* gene and contains the kinase1a motif GPGGVGKT. Most kinase1a regions from other plant NBS-LRR proteins closely match the *RPS2* kinase-1a motif. The kinase2 motif coordinates the metal-ion-binding required for phospho-transfer reactions. A highly conserved argenine is present, which intercts with the purine base of ATP or GTP in other nucleotide binding proteins (Traut, 1994). The C-terminal sub-domain is

referred to as the ARC (Apoptosis Activating Factor 1 (*Apaf-1*), *R* gene products and Cell Death abnormality (*CED-4*)) sub-domain and is a conserved region in plant *R* genes and NBS-containing proteins involved in animal innate immunity and apoptosis (Traut, 1994). Protein modeling studies suggest that the ARC sub-domain mediates oligomerization of CED-4 and Apaf-1 (Jarozsewski *et al.*, 2000). This critical step in initiating apoptosis in animal cells is discussed in further detail below.

The exact role of the NBS domain remains unknown in the context of plant innate immune responses. More is known however of its functions in animal innate immune responses. Proteins of the NOD (nucleotide-binding oligomerization domain) family, such as *Apaf-1* (mammals) and *Ced-4* (*C. elegans*) are intracellularly located and involved in activating apoptosis. These proteins contain three domains: an N-terminal effector-binding domain (among others the CARD (Caspase Recruitment Domain)), a central NBS domain and a C-terminal ligand recognition domain (LRRs in NOD-1 and NOD-2). Apaf-1 is known to bind dATP following disruption of its CARD-WD40 intra-domain interaction by the presence of cytochrome C (released by disintegrated mitochondria). Nucleotide binding further opens up Apaf-1 confirmation so as to allow Apaf-1 oligomerization into a scaffold upon which procaspase-9 molecules aggregate via CARD-homodomain interactions. The resulting structure is referred to as an apoptosome and initiates apoptosis in mammalian cells (Hu *et al* 1998; Srinivasula *et al.,* 1998, Saleh *et al.,* 1999; Zou *et al.*, 1999). Vertebrate NOD-1 and NOD-2 are both involved in intracellular PAMP recognition and appear to be the closest functionally related counterparts of NBS-LRR resistance genes of plants (Inoharam *et al.*, 2002; Royet and Reichhart, 2003).

2.2.1.5 The Leucine-Rich-Repeat Domain

R gene LRR domains match the cytoplasmic LRR sequence LxxLxxLxxLxLxx(N/C/T) x(x)LxxIPxx) and have an average repeat unit length of 23 amino acids. LRR's are implicated in various protein-protein interactions (Kobe and Deisenhofer, 1995). The LRR region of Toll from *Drosophila* for example binds its ligand, Spaëtzle after proteolytic processing (Hultmark, 2003). The LRR domain of mammalian TLR-4 is known to bind the opsonin formed by extracellular lipopolysaccharide receptor CD14 and LPS (Underhill and Ozinsky, 2002).

A very well studied example of an LRR mediated protein-protein interaction is that which occurs between porcine ribonuclease and its inhibitor, porcine ribonuclease inhibitor (PRI). The inhibitor contains an LRR domain and X-ray crystallography has revealed the three

dimensional structure of the ribonuclease-inhibitor complex. The LRR domain of PRI consists of alternating α-helices and β-sheets, forming a horseshoe shaped LRR domain. The concave side of this horseshoe (consisting of β-sheets) faces the solvent and forms a binding pocket where specific residues mediate binding of the protein ligand, in this case porcine ribonuclease (Kobe and Deisenhofer, 1995; Papageorgiou *et al.*, 1997). An LRR repeat consists typically of around 20-29 residues and 11 of these residues follow the conserved pattern LxxLxLxxN/CxL, which is the region of the LRR surrounding the β-sheet. The structures for many other LRR containing proteins have also been determined, although none that are in the same class of LRR's as that of plant *R* genes. Based on the consistency of the LRR motif structures determined thus far, it is expected that plant LRR's would also have similar structural properties (Kobe and Kajava, 2001).

Comparisons between the *R* genes isolated to date and their homo/paralogues have revealed that the amino acid residues that are predicted to be exposed in the β–sheet of the LRR domain are under diversifying selection. This fits the hypothesis that the specificity of *Avr-R* gene interactions are determined by the LRR domain. Comparisons of 11 different alleles of the flax *L* gene (TIR-NBS-LRR) which provide 10 different resistance specificities against flax rust, revealed that polymorphisms occurred all over the coding region, but at a higher density in the LRR region. L6 and L11 differed only in the LRR (33 amino-acid substitutions) and had different resistance specificities, indicating that the specificity of L6 and L11 is determined by the LRR domain. However, L6 and L7, which differ only in the TIR domain have different resistance specificities indicating that the LRR domain is not always the sole determinant of *R* gene specificity (Ellis *et al.*, 1999).

The rapid co-evolution of hosts and pathogens prompts us to ask which evolutionary forces are responsible for shaping new *R* gene specificities. Gene conversion and unequal recombination have already been identified as major factors involved in generating variability in the LRR region (McDowell *et al.*, 1998; Noel *et al.*, 1999; Ellis *et al.*, 1999). It is also interesting to note that around 1% of the coding capacity found in Arabidopsis can be ascribed to NBS-LRR genes. This is similar to the percentage taken up by the immunoglobulin genes encoded by mammalian genomes (Meyers *et al.*, 1999).

2.2.2 Inter-Domain interactions

Interactions between the different domains of NBS-LRR proteins are critical to *R* gene function and recent studies on the *Rx*-gene of potato (TIR-NBS-LRR) have shown that Rx can

function as two separate polypeptides (Moffett *et al.*, 2002). It is possible to reconstitute Rx function by the co-expression of either the TIR and NBS-LRR polypeptides or the TIR-NBS and LRR polypeptides. It was also shown that these domains co-immunoprecipitated when expressed as separate polypeptides, providing evidence for strong inter-domain interactions.

2.2.3 Downstream components

The detection of pathogen ingress by plant *R* genes culminates in a signaling cascade that initially activate the hypersensitive response at the site of infection and ultimately up-regulates the transcription rate of pathogenesis related (PR) proteins such as glucanases and chitinases, which damage fungal cell walls (Fritig *et al.*, 1998; Broglie *et al.*, 1991).

The signaling pathways leading to the elevation of PR gene expression is currently an intensely studied field although still rather incomplete. *R* genes are thought to be situated at the very start of the HR-signaling cascade. It is known that the two classes of NBS-LR*R* genes require different signaling components up to the point where their signaling pathways converge. It is known for example that TIR-NBS-LR*R* genes require the *EDS1* (Enhanced Disease Susceptibility 1) gene, while some CC-NBS-LRR genes require the *NDR1* gene in order to mediate a successful HR (Parker *et al.*, 2000).

NPR1 (non-expresser of PR genes) from Arabidopsis is required for TIR-NBS-LRR mediated resistance (*RPP5*) to *Peronospora parasitica* and appears to be a homolog of I*κβ*, which inhibits nuclear translocation of the transcription factor NF-κβ, and subsequent activation of the NF-*κβ* pathway in the animal innate immunity response (Rairdan and Delaney, 2002). Downstream of *NPR1* the HR signaling pathway ends with the elevated expression of *PR* genes (Hammond-Kosack and Jones, 1997).

2.3 Recognition Models

The ability of the animal innate immune system to detect many different PAMPs essential to specific classes of pathogenic organisms is also shared by plant cells. Various conserved components of pathogens are detected by plant cells including lipopolysaccharides and flagellin proteins (Van Wees *et al.*, 1997). Furthermore, many of the enzymatic breakdown products generated by pathogen activity are also recognized by conserved detection systems. The cloned *FLS2* gene of *Arabidopsis* (encoding an LRR containing Receptor Like Kinase protein (LRR-RLK)) provides a clear example of such a PAMP detecting protein, being

capable of detecting the presence of flagellin which is harbored by all phytopathogenic bacteria (Gómez-Gómez and Boller, 2002).

Specialized phytopathogens are capable of preventing plant cells from initiating a defense response after PAMP detection. This is accomplished by translocating virulence factors into the cytoplasm where they interfere with the defence signaling of host cells. Cytoplasmic *R* genes however, are capable of detecting the presence of virulence factors either directly (receptor-ligand-model: *Pi-ta* and *Avr-Pita*) or indirectly by detecting modification of host components by virulence proteins.

The latter case is referred to as the guard hypothesis and recently elegant examples of this defense approach have been characterized, such as the interaction of *Arabidopsis* and *P.syringae* where the action of two NBS-LRR and three avirulence proteins converge on a single host protein, RIN4. Three virulence proteins from *P. syringae* target RIN4: AvrB and AvrRpm1 cause the phosphorylation of RIN4, which is detected by Rpm1 (Mackey *et al.,* 2002), while AvrRpt2 causes degradation of RIN4, which activates RPS2 (Mackey *et al.*, 2003). Both *R* gene products are known to form complexes with native RIN4 (Axtell and Staskawicz, 2003). Mackey and co-workers (2003) also reported that the expression level of RIN4 is limited by the level of Rpm1, but not vice versa, suggesting that the level of the guard protein is matched to that of the guarded host protein. The guard hypothesis also provides a direct explanation as to why some *R* genes have multiple unrelated avirulence partners, as discussed for Rpm1 above.

The interaction mode of an Avr-R protein pair is expected to be a major determinant of the mode of evolution for the two genes involved. Directly interacting receptor-ligand pairs can be expected to co-evolve rapidly, while evolution of guard proteins detecting the results of avirulence protein action, preclude further evolution of an avirulence gene with respect to its current function. Limited information on the mode of interaction for *R*-genes isolated to date however hampers further investigation of this hypothesis.

2.4 Genomic organization of NBS-LRR loci

2.4.1 General aspects

A prominent aspect in the genomic organization of the NBS-LRR gene family is that members tend to occur in localized clusters. These clusters often contain distantly related

members. The *Mla* locus of barley for example, contains three families of NBS-LRR *R* genes interspersed over a region of 240 kb (Wei *et al.*, 1999) and provides resistance against multiple strains of powdery mildew. The clustered organization of the NBS-LRR gene family in plant genomes is thought to facilitate the formation of new *R* gene specificities via unequal recombination and inter- and intra-gene conversion.

Not all *R* genes of the NBS-LRR family are found in clusters; *RPS2* from *Arabidopsis* occurs as a singleton with two ancient haplotypes and homologues across a wide range of other plant species (Caicedo *et al.*, 1999). It seems as though *R* genes occurring in complex clusters are rapidly evolving to detect co-evolving pathogen virulence genes while others are ancient and have been retained as single copies, which evolve very slowly.

2.4.2 The NBS-LRR family in fully sequenced genomes

2.4.2.1 Arabidopsis thaliana

Meyers and co-workers recently published a detailed study on the *A. thaliana* NBS-LR*R* gene family (Meyers *et al.*, 2003), estimating the number of intact NBS-LR*R* genes at 149. Phylogenetic analysis of the sequences identified showed a clear distinction between the NBS-LRR subfamilies (TIR and CC) as expected, but also revealed that sequences from both the TIR and CC sub-family could be partitioned into smaller clades. Beside the NBS-LRR family, TIR-X and TIR-NBS genes were also found although no functional role has yet been assigned to genes with this domain configuration in plants.

2.4.2.2 Oryza sativa

The amount of NBS-LRR sequences in the draft rice genome-sequence was estimated at over 600, which is approximately four times the amount annotated in the *Arabidopsis* genome. In agreement with previous observations, no members of the TIR-NBS-LRR sub-family were found among cereal NBS-LRR sequences, although gene sequences encoding TIR-NBS and other domain configurations (TIR-X), have representatives in the rice genome (Bai *et al.*, 2002). A recent study by Monosi *et al.* (2004) has lowered the estimate for rice NBS-LRRs to 500 with 100 pseudogenes. Interestingly, 20% of NBS-LRRs sequenced from cDNAs were also pseudogenes, indicating that many of these pseudogenes are still expressed, despite large deletions or inframe stop codons.

2.5 Evolution of the NBS-LRR gene family

2.5.1 Origin

Studies on the NBS-LRR families of a diverse range of plant taxa highlight several important features. Firstly, both the TIR and CC sequence families are highly diverse, with the latter exhibiting the highest diversity. The CC-family contains at least four ancient clades spanned by multiple plant families. Two of the four clades contain both eudicot and gymnosperm sequences, while three of the four contain sequences from both monocot and dicot species. This indicates that some duplication in the NBS-LRR family predate even the Angiosperm-Gymnosperm divergence (Cannon *et al*., 2002).

2.5.2 Models of Evolution

Several models have been invoked for explaining the evolution of multigene families, the two major modes being the concerted (Irwin and Wilson, 1990) and birth-and-death models (Nei *et al.*, 1997). The birth-and-death model proposed for evolution of the human MHC complex has been adapted to the evolution of plant *R* gene families, such as the NBS-LRR family by Michelmore and Meyers (1998). In short, little gene conversion occurs among paralogues in clusters, while unequal recombination facilitates contraction and expansion of paralogue arrays. Gene conversion and unequal crossovers do not homogenize members in a haplotype such that orthology between the genes in different haplotypes is lost, in agreement with the closer relationship observed between othologues than paralogues for different *R* gene haplotypes as seen in the *Pto* (Salmeron *et al*., 1996), *Dm* (Shen *et al*., 2002) and *Cf* loci (Dixon *et al*., 1998). Divergent selection follows gene duplication to create new specificities. Detailed studies on the NBS-LRR family in *Arabidopsis* (Baumgarten *et al.*, 2003) indicates that NBS-LRR family evolution in *Arabidopsis* involves detectable levels of gene conversion among paralogues, with duplications occurring mainly within restricted chromosomal segments, via unequal crossovers. Ectopic duplications/translocations in *Arabidopsis* were found to be rare events (< 5% of duplications) and to involve chromosomal segments, maintaining synteny during translocation (Baumgarten *et al.*, 2003). The picture in grass genomes is probably quite different owing to the massive difference in genome size, brought about by the large number of retrotransposon sequences present. Retrotransposons can shape *R* gene clusters by driving duplication and ectopic translocation, possibly via active transposition and unequal recombination mediated by transposon similarity tracts. The tight

colinearity, which is a hallmark of grass genomes, is often violated by members of the NBS-LRR family (Leister *et al.*, 1998). This argues against chromosome segment duplication being responsible for ectopic NBS-LRR translocation as seen in *Arabidopsis,* and for a more specific transfer mechanism.

The fate of recently duplicated genes is central in any discussion concerning multi-gene family evolution. Theoretically, the possible outcomes of gene duplication can be placed in four categories: 1.) The gene can lose its function and evolve neutrally as a pseudogene. 2.) If the gene in question has separable functions encoded by discrete domains, it is possible that mutational inactivation of separate domains in each of the gene copies can make both copies essential. This would then be followed by specialization of each gene, resulting in sub-functionalization. 3.) The gene could adopt a new independent function (neofunctionalization) and would have the opportunity to further diverge under positive selection pressure. Neofunctionalization can also occur by changes in transcription response. 4.) Both copies can be retained in the genome if they provide a selective advantage due to elevated levels of expression (Otto and Yong, 2002). The probabilities of each possible outcome of gene duplication as categorized above are clearly highly dependent on the characteristics of the specific gene being duplicated. Genes that are expressed in very high quantities are likely to benefit from duplication events and prominent examples include the tRNA and rRNA genes (Ohno, 1970), which have high copy numbers. Genes with many separable functions encoded by discrete domains would be more likely to evolve sub-functionalization. The probability of a single gene duplicate evolving new functionality is thought to be miniscule. Recent studies on the fate of duplicated genes (Wagner 1998; Otto and Yong, 2002; Blanc and Wolfe, 2004) have focused largely on the fate of ancient duplications resulting from polyploidization. These studies found high estimates for the probability of duplicate genes evolving novel functions – in the region of 50% (Wagner, 1998). Since genes duplicated by polyploidization may form part of metabolic networks (gene dosage effect) and are often subunits of multimeric proteins where mutations can produce dominant negative phenotypes, the evolutionary fate of numerous genes is often bound collectively. Hence gene duplications produced by polyploidy are likely to experience purifying selection, which drastically slows gene loss, maintaining functionality and allowing subsequent functional divergence. A recent study by Blanc and Wolfe (2004) found that the gene duplicates retained from ancient polyploidization events in the *Arabidopsis* genome were statistically biased for specific functional categories. Interestingly *R* genes were found to be preferentially lost, as their basic attributes (Low transcription rate, relative dosage insensitivity, dominant phenotype, no structural role in

multimeric protein complexes) would most likely not place new duplicates under purifying selection following duplication.

The potential for evolving independent function is central to determining the size and expansion rate of a particular gene family. Genes acting in recognition of exogenous entities usually has the greatest potential in this regard as exemplified by the large gene families of plant and animal immune systems (Nei *et al.*, 1997) and the large olfactory receptor gene family found in animals (Niimura and Nei, 2003). Another factor influencing multi-gene family evolution, which can easily be overlooked, is the population dynamics involved in the fixation of new gene duplications. Under basic population genetics assumptions, the chance of fixing a selectively neutral gene duplication is 1/2N (N being the effective population size) and this occurs at exactly the rate of neutral mutation irrespective of population size. Otto and Yong (2002) have shown that for loci where heterozygote advantage is present, the majority of gene duplications reaching fixation would be those yielding permanent heterozygosity in a tightly linked haplotype. Examples include the spread of duplications providing pesticide resistance (Lenormand *et al.*, 1998) in *Culex pipiens* and the independent evolution of color vision in a single New World monkey species (Jacobs and Degan, 2001). In short, this model has functional divergence occurring between different alleles at a single locus prior to their fixation via duplication in a single haplotype (most likely by unequal crossing over in a heterozygote), acting as a permanent heterozygote that can spread rapidly through a population by overdominant selection. This model further predicts a much lower fraction of pseudogenes as opposed to models where functionaly redundant duplicates approach or reach fixation in a population prior to functional divergence or loss.

2.5.3 Comparison of the NBS-LRR family across modern plant species

2.5.3.1 Synteny

A comparative study of the NBS-LRR gene family of the *Solanaceae* revealed that *R* gene clusters often occurred in syntenic positions between the tomato and potato genomes. Null alleles were also observed for two members of the *Lycopersicon* genus, *L. esculentum* and *L. pennelli* (Pan *et al.*, 2000).

The rapid rate of evolution associated with *R* gene families causes the syntenic relationships of *R* gene clusters to vanish quickly as progressively distant taxa are compared. Very little synteny can be observed for example when comparing the *R*-gene clusters known for *Arabidopsis* with those of rice. The same applies to comparisons within the monocotyledons

where the syntenic relationship of *R* gene clusters is unclear when comparing rice and barley, whereas in dicotyledons, comparisons between potato and tomato still yield observable *R* gene cluster synteny (Pan *et al*., 2000b). The higher synteny observed for the *Solanaceae* as compared to the *Poaceae*, might be due to the more ancient origin of members of the *Poaceae* (46 and 40 Mya, respectively) (Pan *et al*., 2000a) or possibly due to a different mode of ectopic translocation such as retrotransposition as opposed to chromosome segement duplication (Baumgarten *et al.*, 2003).

2.5.3.2 Intron positions

Out of twenty dicotyledon NBS-LRR *R* genes, only the *RPP8/Hrt* gene from *A. thaliana* contains an intron in the NBS-domain. In contrast, three of the characterized cereal NBS-LRR *R* genes have introns in their NBS-domains: *Mla1* (Zou *et al.*, 1999), *Pi-ta* (Bryan *et al.*, 2000) and *Pib* (Wang *et al.*, 1999). The *Bai et al.* study (2002) investigated intron positions for some full-length cereal CC-NBS-LRR genes by cloning and sequencing corresponding cDNAs. The most common intron position was found at the N-terminal side of the kinase-2 motif, and is estimated to occur in roughly a quarter of rice NBS-LR*R* genes. Sequences with this intron position were also found to possess related NBS-domain sequences.

2.6 Hexaploid wheat and its diploid genome donors

2.6.1 Karyotype

Due to the immense size of the wheat genome (16 000 Mb), very little is known about *R* gene content and localization (Arumuganathan and Earle, 1991). Modern hexaploid wheat (AABBDD) arose around 8 000 years ago due to early agricultural practices. Two major allopolyploidization events ocurred during this process. Initially tetraploid durum wheat (*Triticum turgidum*) containing both the A (*Triticum urartu*) and B *(*likely *Aegilops speltoides*) (Zhang *et al.*, 2002) genomes arose. During a subsequent allopolyploidization event, the *Triticum turgidum* (AB) genome was combined with the D genome of *Aegilops tauschii,* resulting in the modern hexaploid wheat (*Triticum aestivum*) with the genome designation AABBDD (Kihara, 1944; McFadden and Sears, 1946; Lagudah *et al.,* 1991). The A, B and D genomes are closely related and all three contain seven co-linear chromosomes (Gill and Raupp, 1987). Current estimates for A, B, and D genome divergence range between 2.5 and 4.5 million years ago (Huang *et al.*, 2002a; Huang *et al.*, 2002b) Due to the

evolutionary bottlenecks encountered during allopoliploidization, hexaploid wheat posses little genetic diversity, making it a difficult subject for molecular mapping and breeding programs.

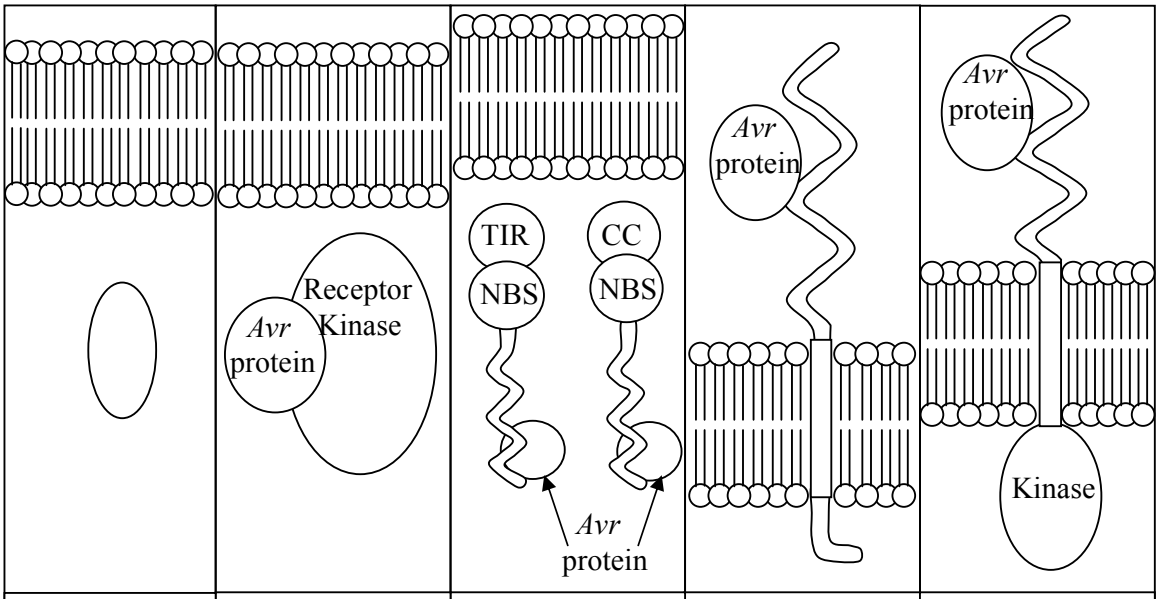
2.6.2 Taxonomy

The *Poaceae* family consist of as much as 10 000 grass species (Huang *et al.*, 2002a; Huang *et al.*, 2002b), which radiated into four major subfamilies approx. 50-80 Mya including the *Pooideae* subfamily, from which radiated the *Triticeae*, *Poeae* and *Avenae* tribes around 35 Mya (Figure 2.3). Current estimates for the divergence times of barley, rye and wheat are around 11Mya for barley and 7Mya for wheat and rye (Huang *et al.*, 2002a; Huang *et al.*, 2002b).

2.6.3 Resources

A wealth of plant DNA sequence data has been generated over the last decade, especially after completion of the *Arabidopsis thaliana* and *Oryza sativa* genome projects (The *Arabidopsis* Genome Initiative, 2000; Delseny, 2003).

Studies of resistance gene families can however provide great entry points for breeding new lines resistant to the latest pathogen outbreaks. RGA sequences can often be converted into useful markers, even more so than other defense related genes (DR), which are inherited in a more quantitative fashion (Ramalingam *et al.*, 2003). NBS-LRR-like RGA markers have previously been found to co-localize with known resistance loci in cereal species (Lagudah *et al.*, 1997; Leister *et al.*, 1998; Seah *et al.*, 1998; Collins *et al.,* 1999; De Majnik *et al.*, 2003).



1.) Enzymes responsible for detoxification	2.) Kinases interacting with Avr genes and initialising defence signalling cascades in association with genes of the NBS-LRR class	3.) Cytoplasmic receptors (NBS-LRR genes)	4.) Transmembrane receptors	5.) Transmembrane receptor kinases
Hm1 (Maize)	*Pto* (Tomato)	*N* (Tobacco) *RPS2* (*Arabidopsis thaliana*)	*Cf2,Cf9* (Tomato)	*Xa-21* (Rice)

Figure 2.1 Schematic representation of the five classes of characterized *R* gene encoded proteins.

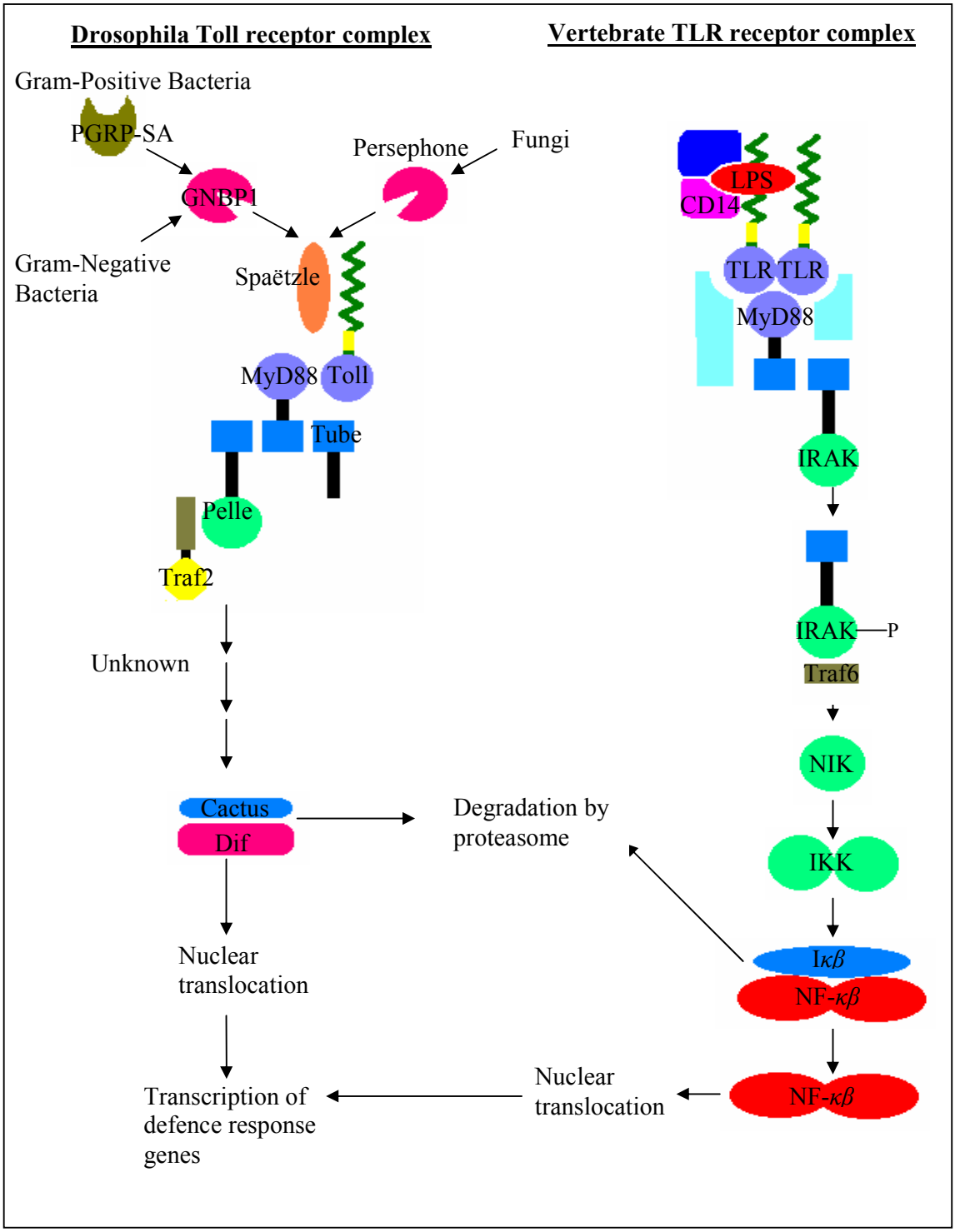


Figure 2.2 Schematic illustration of the homologies present in signal transduction pathways of *Drosophila* and vertebrate innate immune responses as mediated by *Toll* receptor complexes (Underhill and Ozinsky, 2002; Kopp and Medzhitov, 1999).

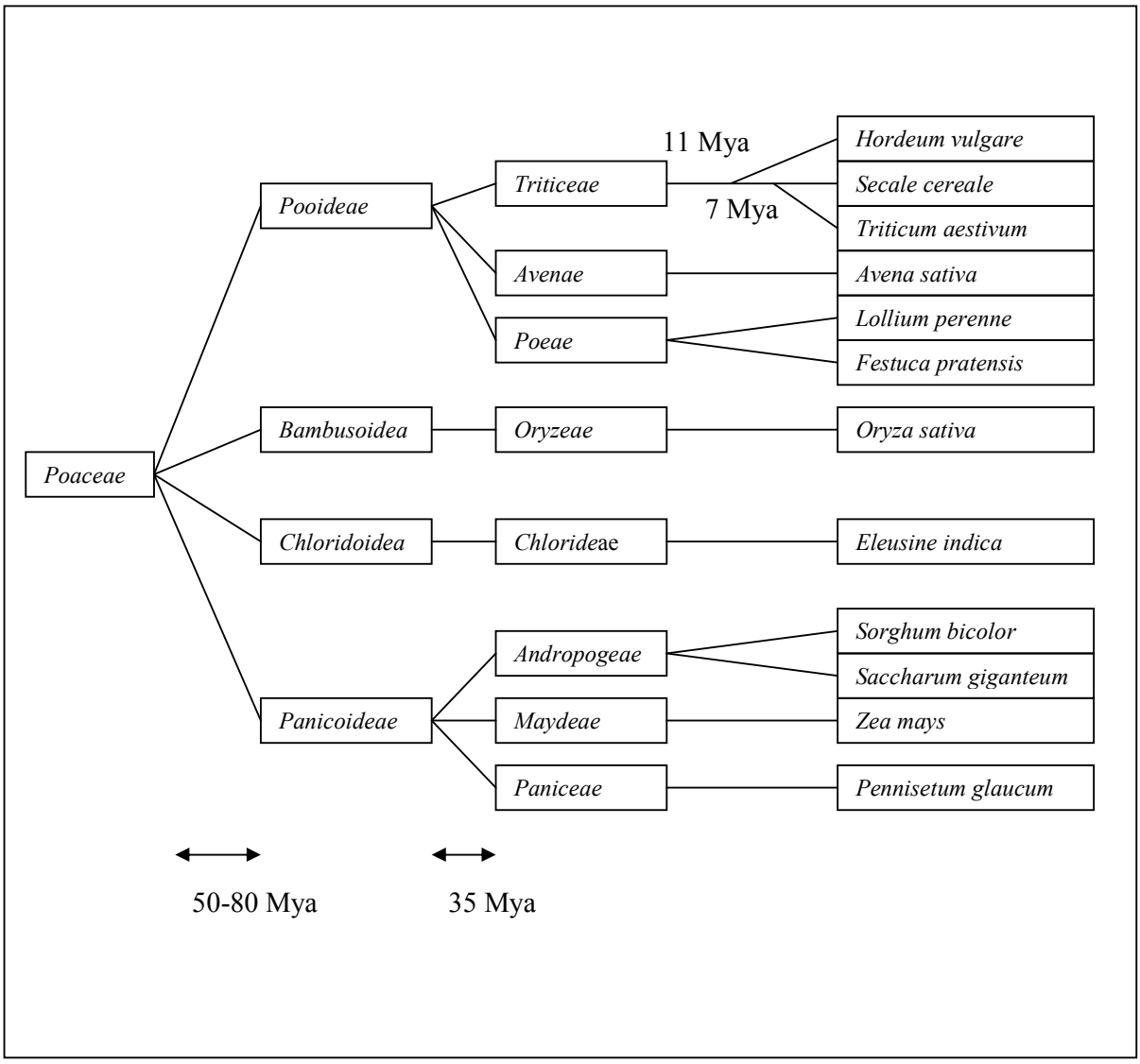


Figure 2.3 Time frame of major evolutionary events inferred for taxonomic units in the *Poaceae* family (Huang *et al.*, 2002a; Huang *et al.*, 2002b).

Table 2.1 Disease resistance genes isolated to date.

Resistance Gene Class	***R* gene**	**Plant species**	**Pathogen**	**Avirulence factor**	**References**
1.) Enzymatic detoxification	*Hm1*	Maize	*Cochliobolus carbonum*		Multani *et al.*, 1998
2.) Serine-threonine kinases	*Pto*	Tomato (Requires *Prf*)	*Pseudomonas syringae*	*AvrPto, AvrPtoB*	Martin *et al.*, 1993
	Pbs1	*Arabidopsis thaliana* (Requires *Rps5*)	*Pseudomonas syringae*	*AvrPphB*	Swiderski and Innes, 2001
3.a) CC-NBS-LRR	*BS2*	Pepper	*Xanthomonas campestris*	*AvrBs2*	Tai *et al.*, 1999
	Gpa2	Potato	*Globodera pallida*		Van der Vossen *et al.*, 2000
	Hero	Potato	*Globodera rostochiensis* *Globodera pallida*		Ernst *et al.*, 2002
	HRT	*Arabidopsis thaliana*	*Turnip Crinkle Virus*	Coat Protein	Cooley *et al.*, 2000
	I2	Tomato	*Fusarium oxysporum*		Ori *et al.*, 1997
	Lr10	*Triticum aestivum*	*Puccinia triticina*		Feuillet *et al.*, 2003
	Lr21	*Aegilops tauschii*	*Puccinia triticina*		Huang *et al.*, 2003

Table 2.1(continued) Disease resistance genes isolated to date.

Resistance Gene Class	*R* gene	Plant species	Pathogen	Avirulence factor	References
3.a) CC-NBS-LRR	*Mla*	Barley	*Erysiphe graminis*		Wei *et al.*, 2002
	Pib	Rice	*Magnaporthe grisea*		Wang *et al.*, 1999
	Pi-ta	Rice	*Magnaporthe grisea*	*Avr-Pita*	Bryan *et al.*, 2000
	Pm3b	*Triticum aestivum*	*Blumeria graminis f. sp. tritici*	*AvrPm3b*	Yahiaoui et *al.*, 2004
	Prf (Needs Pto)	*Lycopersicon esculentum*	*Pseudomonas syringae*	*AvrPto, AvrPtoB*	Salmeron et *al.*, 1996
	R1	Potato	*Phytophtora infestans*		Ballvora *et al.*, 2002
	Rp1	Maize	*Puccinia sorghi*		Collins *et al.*, 1999
	RPI	*Solanum bulbocastanum*	*Phytophtora infestans*		Van Der Vossen *et al.*, 2003
	RPM1	*Arabidopsis thaliana*	*Pseudomonas syringae*	*AvrRpm1, AvrB*	Grant *et al.*, 1995
	RPP13	*Arabidopsis thaliana*	*Peronospora parasitica*		Bittner-Eddy *et al.*, 2000
	RPP8	*Arabidopsis thaliana*	*Peronospora parasitica*		Cooley *et al.*, 2000
	RPS2	*Arabidopsis thaliana*	*Pseudomonas syringae*	*AvrRpt2*	Mindrios *et al.*, 1994
	RPS5 (Needs PBS1)	*Arabidopsis thaliana*	*Pseudomonas syringae*	*AvrPphB*	Warren *et al.*, 1998
	Rx	Potato	*Potato virus X*	Coat protein	Bendahmane *et al.*, 1999

Table 2.1(continued) Disease resistance genes isolated to date.

Resistance Gene Class	*R* gene	Plant species	Pathogen	Avirulence factor	References
3a.) CC-NBS-LRR	*Sw-5*	Tomato	*Tomato Spotted Wilt Virus*		Brommonschenkel *et al.*, 2000
	Tm-2	*Lycopersicon esculentum*	*Tomato mosaic virus*		Lanfermeijer *et al.*, 2003
	Xa1	Rice	*Xanthomonas oryzae*		Yoshimura *et al.*, 1998
3.b) TIR-NBS-LRR	*L*	Flax	*Melamspora lini*		Ellis *et al.*, 1999
	M	Flax	*Melamspora lini*		Anderson *et al.*, 1997
	N	Tobacco	*Tobacco Mosaic Virus*	Helicase	Witham *et al.*, 1994
	P	Flax	*Melamspora lini*		Dodds *et al.*, 2001
	RPP1	*Arabidopsis thaliana*	*Peronospora parasitica*		Botella *et al.*, 1994
	RPP4	*Arabidopsis thaliana*	*Peronospora parasitica*		Van der Biezen *et al.*, 2002
	RPP5	*Arabidopsis thaliana*	*Peronospora parasitica*		Noel *et al.*, 1999
	RPS4	*Arabidopsis thaliana*	*Pseudomonas syringae*	*Avr-Rps4*	Gassmann *et al.*, 1999
4.) Receptors (no kinase)	*Cf-2*	Tomato	*Cladosporium fulvum*	*Avr2*	Dixon *et al.*, 1996
	Cf-4	Tomato	*Cladosporium fulvum*	*Avr4*	Parniske *et al.*, 1997
	Cf-5	Tomato	*Cladosporium fulvum*		Dixon *et al.*, 1998

Table 2.1(continued) Disease resistance genes isolated to date.

Resistance Gene Class	***R* gene**	**Plant species**	**Pathogen**	**Avirulence factor**	**References**
4.) Receptors (no kinase)	*Cf-9*	Tomato	*Cladosporium fulvum*	*Avr9*	Parniske *et al.*, 1997
5.) Receptors (kinase)	*Xa21*	Rice	*Xanthomonas oryzae*		Song *et al.*, 1995
6.) Other: LRR+putative transmembrane region	*HS1-pro-1*	Sugar beet	*Heterodera schachtii*		Cai *et al.*, 1997
G-protein-coupled receptor	*Mlo*	Barley	*Blumeria graminis*		Kim *et al.*, 2002
Receptor kinase-like protein with two tandem protein kinase domains	*Rpg1*	Barley	*Puccinia graminis*		Brueggeman *et al.*, 2002
Activates SA-dependant HR	*RPW8*	*Arabidopsis thaliana*	*Erisyphe chicoracearum*		Xiao *et al.*, 2001
TIR-NBS-LRR with WRKY transcription factor domain.	*RRS1-R*	*Arabidopsis thaliana*	*Ralstonia solanacearum*		Deslandes *et al.*, 2002
Jacalin repeats – restricts longrange movement	*RTM1*	*Arabidopsis thaliana*	*Tobacco Etch Virus*		Chisholm *et al.*, 2000
N-terminal heatshock protein homology, large C-terminal repeats	*RTM2*	*Arabidopsis thaliana*	*Tobacco Etch Virus*		Witham *et al.*, 2000
Cell-surface glycoproteins with receptor-mediated endocytosis-like signals and leucine zippers.	*Ve1e,* *Ve2e*	Tomato	*Verticillium alboatrum*		Kawchuck *et al.*, 2001

2.7 References

Agrios, G.N. (1998) Plant Pathology. *London : Academic. 3rd ed*.

Anderson, K.V. (2000) Toll signalling pathways in the innate immune response. *Current Opinion in Immunology* 12:13-19.

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. *Plant Cell* 9:4:641-651.

Arumuganathan, K., and Earle, E.D. (1991) Nuclear DNA content of some important plant species. *Plant Molecular Biology Reports* 9: 208-218.

Axtell, M.J., and Staskawicz, B.J. (2003) Initiation of *RPS2* specified disease resistance in *Arabidopsis* is coupled to AvrRpt2-directed elimination of RIN4. *Cell* 112:369-377.

Bai, J., Pennill, L.A., Ning, J., Lee, S.W., Ramalingam, J., Webb, C.A., Zhao, B., Sun, Q., Nelson, J.C., Leach, J.E. and Hulbert, S.H. (2002) Diversity in Nucleotide-Binding Leucine-Rich-Repeat genes in Cereals. *Genome Research* 12:1871-1884.

Ballvora, A., Ercolano, M.R., Weiss, J., Meksem, K., Bormann, C.A., Oberhagemann, P., Salamini, F. and Gebhardt, C. (2002) The *R1* gene for potato resistance to late blight (*Phytophthora infestans*) belongs to the leucine zipper-NBS-LRR class of plant resistance genes. *Plant Journal* 30:3:361-371.

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

Bendahmane, A., Kanyuka, K. and Baulcombe, D.C. (1999) The *Rx* gene from potato controls separate virus resistance and cell death responses. *Plant Cell* 11:781-791.

Bennett, R.N., and Wallsgrove, R.M. (1994) Secondary metabolites in plant defence mechanisms. *New Phytologist* 127:617-633.

Bent, A.F., Kunkel, B.N., Dahlbeck, D., Brown, K.L., Schmidt, R., Giraudat, J., Leung, J., and Staskawicz, B.J. (1994) *RPS2* of *Arabidopsis thaliana*: A Leucine-Rich Repeat Class of Plant Disease resistance Genes. *Science* 265:1856-1860.

Bittner-Eddy, P.D., Crute, I.R., Holub, E.B., and Beynon, J.L. (2000) *RPP13* is a simple locus in *Arabidopsis thaliana* for alleles that specify downy mildew resistance to different avirulence determinants in *Peronospora parasitica. Plant Journal* 21:2:177-188.

Blanc, G., and Wolfe, K.H. (2004) Functional Divergence of Duplicate Genes Formed by Polyploidy during Arabidopsis Evolution. *The Plant Cell* 16:1679-1691.

Bolwell, G.P. (1999) Role of active oxgen species and NO in plant defence responses. *Current Opinion In Plant Biology* 2:287-294.

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. (1994) Three genes of the *Arabidopsis RPP1* complex resistance locus recognize distinct *Peronospora parasitica* avirulence determinants. *Plant Cell* 10:11:1847-1860.

Broglie, K., Chet, I., Holliday, M., Cressman, R., Biddle, P., Knowlton, S., Mauvais, C.J. and Broglie, R. (1991) Transgenic plants with enhanced resistance to the fungal pathogen *Rhizoctanio solani*. *Science* 254:1194-1197.

Brommonschenkel, S.H., Frary, A. and Tanksley, S.D. (2000) The broad-spectrum tospovirus resistance gene *Sw-5* of tomato is a homolog of the root-knot nematode resistance gene *Mi*. *Molecular Plant Microbe Interactions*. 13:10:1130-1138.

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*. *Plant Cell* 12:11:2033-46.

Brueggeman, R., Rostoks, N., Kudrna, D., Kilian, A., Han, F., Chen, J., Druka, A., Steffenson, B., and Kleinhofs, A. (2002) The barley stem rust-resistance gene *Rpg1* is a novel disease-resistance gene with homology to receptor kinases. *Proceedings of the National Academy of Sciences USA* 99:14:9328-9333.

Cannon, S.B., Zhu, H., Baumgarten, A.M., Spangler, R., May, G., Cook, D.R., Young, N.D. (2002) Diversity, distribution, and ancient taxonomic relationships within the TIR and non-TIR NBS-LRR resistance gene subfamilies. *Journal of Molecular Evolution* 54:548-562.

Cai, D., Kleine, M., Kifle, S., Harloff, H.-J., Sandal, N.N., Marcker, K.A., Klein-Lankhorst, R.M., Salentijn, E.M.J., Lange, W., Stiekema, W.J., Wyss, U., Grundler, F.M.W., and Jung, C. (1997) Positional cloning of a gene for nematode resistance in sugar beet. *Science* 275:5301:832-834.

Caicedo, A.L., Schaal, B.A., and Kunkel, B.N. (1999) Diversity and molecular evolution of the *RPS2* resistance gene in *Arabidopsis thaliana*. *Proceedings of the National Academy of Sciences USA* 96:302-306.

Cassab, G.I., and Varner, J.E. (1998) Cell wall proteins. *Annual Review of Plant Physiology and Plant Molecular Biology* 39:321-353.

Chisholm, S.T., Mahajan, S.K., Whitham, S.A., Yamamoto, M.L., and Carrington, J.C. (2000) Cloning of the Arabidopsis *RTM1* gene, which controls restriction of long-distance movement of tobacco etch virus. *Proceedings of the National Acadamy of Sciences USA* 97:1:489-494.

Collins, N., Drake, J., Ayliffe, M., Sun, Q., Ellis, J., Hulbert, S. and Pryor, T. (1999) Molecular characterization of the maize *Rp1-D* rust resistance haplotype and its mutants. *Plant Cell* 11:7:1365-1376.

Cooley, M.B., Pathirana, S., Wu, H.-J., Kachroo, P., and Klessig, D.F. (2000) Members of the *Arabidopsis HRT/RPP8* family of resistance genes confer resistance to both viral and oomycete pathogens. *Plant Cell* 12:663-676.

Davis, J.M., Wu, H., Cooke, J.E.K., Reed, J.M., Luce, K.S. and Michler, C.H. (2002) Pathogen challenge, salicylic acid, and jasmonic acid regulate expression of chitinase gene homologs in pine. *Molecular Plant-Microbe Interactions* 15:380-387.

Delseny, M. (2003) Towards an accurate sequence of the rice genome. *Current Opinion In Plant Biology* 6:2:101-105.

De Majnik, J., Ogbonnaya, F.C., Moullet, O., and Lagudah, E. (2003) The *Cre1* and *Cre3* Nematode resistance genes are located at homeologous loci in the wheat genome. *Molecular Plant Microbe Interactions* 16:12:1129-1134.

Dennis, D.T., Turpin, D.H., Lefebvre, D.D. and Layzell, D.B. (1997) Plant metabolism. *Addison Wesley Longman, Harlow*.

Dixon, M.S., Hatzixanthis, K., Jones, D.A., Harrison, K. and Jones, J.D. (1998) The tomato *Cf-5* disease resistance gene and six homologs show pronounced allelic variation in leucine-rich repeat copy number. *Plant Cell* 10:11:1915-1925.

Dixon, M.S., Jones, D.A., Kedie, J.S., Thomas, C.M., Harrison, K. and Jones, J.D.G. (1996) The tomato *Cf-2* disease resistance locus comprises two functional genes encoding leucine-rich repeat proteins. *Cell* 84:451-459.

Dodds, P.N., Lawrence, G.J., Catanzariti, A.M., Ayliffe, M.A., and Ellis, J.G. (2004) The *Melampsora lini AvrL567* Avirulence Genes Are Expressed in Haustoria and Their Products Are Recognized inside Plant Cells. *Plant Cell* 16:3:755-768.

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. *Plant Cell* 1:163-178.

Ellis, J.G., Lawrence, G.J., Luck, J.E., and Dodds, P.N. (1999) Identification of regions in alleles of the flax rust resistance gene *L* that determine differences in gene-for-gene specificity. *Plant Cell* 11:3:495-506.

Erickson, F.L., Dinesh-Kumar, S.P., Holzberg, S., Ustach, C.V., Dutton, M., Handley, V., Corr, C., and Baker, B.J. (1999) Interactions between tobacco mosaic virus and the tobacco *N* gene. *Philosophical Transactions of the Royal Society of London B* 354:653-658.

Ernst, K., Kumar, A., Kriseleit, D., Kloos, D.U., Phillips, M.S., and Ganal, M.W. (2002) The broad-spectrum potato cyst nematode resistance gene (Hero) from tomato is the only member of a large gene family of NBS-LRR genes with an unusual amino acid repeat in the LRR region. *Plant Journal* 31:2:127-136.

Feuillet C., Travella S., Stein N., Albar L., Nublat A., and Keller B. (2003) Map-based isolation of the leaf rust disease resistance gene *Lr10* from the hexaploid wheat (*Triticum aestivum L.*) genome. *Proceedings of the National Academy of Sciences USA* 100:25:15253-15258.

Fidantsef, A.L., Stout, M.J., Thaler, J.S., Duffey, S.S., and Bostock, R.M. (1999) Signal interactions in pathogen and insect attack: expression of lipoxygenase, proteinase inhibitor II, and pathogenesis-related protein *P4* in the tomato, *Lycopersicon esculentum. Physiological and Molecular Plant Pathology* 54:97-114.

Fitzgerald, K.A., Palsson-McDermott, E.M., Bowie, A.G., Jefferies, C.A., Mansell, A.S., Brady, G., Brint, E., Dunne, A., Gray, P., Harte, M.T., McMurray, D., Smith, D.E., Sims, J.E., Bird, T.A. and O'Neill, L.A.J. (2001) Mal(Myd88-Adapter-Like) is required for Toll-like receptor-4 signal transduction. *Nature* 413:78-83.

Flor, H.H. (1971) Current status of the gene-for gene concept. *Annual Review of Phytopathology* 9:275-298.

Fritig, B., Geitz, H.T., and Legrand, M. (1998) Antimicrobial proteins in induced plant defense. *Current Opinion Immunology* 10:16-22.

Deslandes, L., Olivier, J., Theulieres, F., Hirsch, J., Feng, D.X., Bittner-Eddy, P., Beynon, J., and Marco, Y. (2002) Resistance to *Ralstonia solanacearum* in *Arabidopsis thaliana* is

conferred by the recessive *RRS1-R* gene, a member of a novel family of resistance genes. *Proceedings of the National Academy of Sciences USA* 19:99:4:2404-2409.

Gassmann, W., Hinsch, M.E., and Staskawicz B.J. (1999) The *Arabidopsis RPS4* bacterial-resistance gene is a member of the TIR-NBS-LRR family of disease-resistance genes. *Plant Journal* 20:3:265-277.

Galan, J.E., and Collmer, A. (1999) Type III secretion machines: Bacterial devices for Protein delivery into host cells. *Science* 284:1322-1329.

Gill, B.S., and Raupp, W.J. (1987) Direct genetic transfers from *Aegilops squarrosa L.* to hexaploid wheat. *Crop Science* 27:445-450.

Gobert, V., Gottar, M., Matskevich, A.A., Rutschmann, S., Royet, J., Belvin, M., Hoffmann, J.A., and Ferrandon, D. (2003) Dual activation of the *Drosophila* Toll pathway by two pattern recognition receptors. Science 302:5653:2126-2130.

Gómez-Gómez, L., and Boller, T. (2002) Flagellin perception: A paradigm for innate immunity. *Trends in Plant Science* 7:251-256.

Goodman, R.N., and Novacky, A.J (1994) The Hypersensitive Reaction in plant to Pathogens: A Resistance Phenomenon. *American Phytopathology Society, St Paul.*

Grant, M.R., Godiard, L., Straube, E., Ashfield, T., Lewald, J., Sattler, A., Innes, R.W., and Dangl, J.L. (1995) Structure of the *Arabidopsis RPM1* gene enabling dual specificity disease resistance. *Science* 269:5225:843-846.

Greenberg, J.T. (1996) Programmed Cell Death: a way of life for plants. *Proceedings of the National Academy of Sciences USA* 93:12094-12097.

Gurr, S.J., McPherson, M.J., and Bowles, D.J. Molecular Plant Pathology. A Practical Approach. (1992) *Oxford University Press.*

Hammond-Kosack, K.E., and Jones, J.D.G. (1997) Plant disease resistance genes. *Annual Review of Plant Physiology and Plant Molecular Biology*. 48:575-607.

Hatada, E.N., Krappmann, D., and Scheidereit, C. (2002) *NF-kB* and the innate immune response. *Current opinion in immunology*. 12:52-58.

He, Z., Wang, Z-Y., Li, J., Zhu, Q., Lamb, C., Ronald, R., Chory, J. (2000) Perception of *Brassinosteroids* by the extracellular domain of the receptor kinase BRI1. *Science* 288:2360-2363.

Heath, M.C. (2000) Hypersensitive response-related death. *Plant Molecular Biology* 44:321-344.

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

Huang, S., Sirikhachornkit, A., Faris, J.D., Su, X., Gill, B.S., Haselkorn, R. and Gornicki, P. (2002a) Phylogenetic analysis of the acetyl-CoA carboxylase and 3-phosphoglycerate kinase loci in wheat and other grasses. *Plant Molecular Biology* 48:5-6:805-20.

Huang, S., Sirikhachornkit, A., Su X., Faris, J.D., Gill, B.S., Haselkorn, R., and Gornicki, P. (2002b) Genes encoding plastid acetyl-CoA carboxylase and 3-phosphoglycerate kinase of the *Triticum/Aegilops* complex and the evolutionary history of polyploid wheat. *Proceedings of the National Academy of Sciences USA* 99: 8133-8138.

Huang, L., Brooks, S.A., Li, W., Fellers, J.P., Trick, H.N., and Gill, B.S. (2003) Map-Based Cloning of Leaf Rust Resistance gene *Lr21* from the large and polyploid genome of Bread Wheat. *Genetics* 164:655-664.

Hultmark, D. (2003) Drosophila immunity: Paths and Patterns. *Current Opinion in Immunology* 15:1-8.

Inoharam, N., Ogura, Y., and Nunez, G. (2002) Nods: a Family of cytosolic proteins that regulate the host response to pathogens. *Current Opinion in Microbiology* 5:76-80.

Irwin, D.M., and Wilson, A.C. (1990) Concerted evolution of ruminant stomach lysozymes. Characterization of lysozyme cDNA clones from sheep and deer. *Journal Biological Chemistry* 265:9:4944-4952.

Jacobs, G.H., and Deegan, J.F. 2nd. (2001) Photopigments and colour vision in New World monkeys from the family *Atelidae*. *Proceedings of the Royal Society London Series B Biological Sciences* 268:1468:695-702.

Jamir, Y., Guo, M., Oh, H.S., Petnicki-Ocwieja, T., Chen, S., Tang, X., Dickman, M.B., Collmer, A., and Alfano, J.R. (2004) Identification of Pseudomonas syringae type III effectors that can suppress programmed cell death in plants and yeast. *Plant Journal* 37:4:554-565.

Jaroszewski, L., Rychlewski, L., Reed, J.C., and Godzik, A. (2000) ATP-Activated Oligomerization as a Mechanism for Apoptosis Regulation: Fold and Mechanism Prediction for CED-4. *Proteins: Structure, Function and Genetics* 39:197-203.

Johal, G.S., and Briggs, S.P. (1992) Reductase activity encoded by the *HM1*-disease resistance gene in maize. *Science* 258:985-987.

Jones, D.A., Thomas, C.M., Hammond-Kosack, K.E., Balint-Kurti, P.J., and Jones, J.D.G. (1994) Isolation of the tomato *Cf-9* gene for resistance to *Cladosporium fulvum* by transposon tagging. *Science* 266:789-793.

Joosten, M.H.A.J., Cozijnsen, T.J., and De Wit, P.J.G.M. (1994) Host resistance to a fungal tomato pathogen lost by a single base-pair change in an avirulence gene. *Nature* 367:384-386.

Jordan, T., Schornack, S., and Lahaye, T. (2002) Alternative splicing of transcripts encoding Toll-like plant resistance proteins-what's the functional relevance to innate immunity? *Trends in Plant Science* 7:9:392-398.

Kawchuk, L.M., Hachey, J., Lynch, D.R., Kulcsar, F., van Rooijen, G., Waterer, D.R., Robertson, A., Kokko, E., Byers, R., Howard, R.J., Fischer, R., and Pruefer, D. (2001) Tomato *Ve* disease resistance genes encode cell surface-like receptors. *Proceedings of the National Academy of Sciences USA* 98:11:6511-6515.

Kihara, H. (1944) Discovery of the DD-analyser, one of the ancestors of vulgare wheats. *Agriculture and Horticulture* 19:889-890.

Kim, M.C., Panstruga, R., Elliott, C., Muller, J., Devoto, A., Yoon, H.W., Park, H.C., Cho, M.J., and Schulze-Lefert, P. (2002) Calmodulin interacts with MLO protein to regulate defence against mildew in barley. *Nature* 416:6879:447-451.

Kobe, B., and Deisenhofer, J. (1995) The Leucine-rich repeat: a versatile binding motif. *Trends in Biochemical Sciences* 19:415-421.

Kobe, B., and Kajava, A. (2001) The Leucine-Rich repeat motif as a protein recognition motif. *Current Opinion in Structural Biology* 11:725-732.

Kopp, E.B., and Medzhitov, R. (1999) The Toll-receptor family and control of innate immunity. *Current Opinion in Immunology* 11:13-18.

Kunkel, B.N. (1996) A useful weed put to work: genetic analysis of disease resistance in *Arabidopsis thaliana*. *Trends in* Gene*tics* 12:63-69.

Kunkel, B.N., and Brooks, D.M. (2002) Cross-talk between signaling pathways in pathogen defense. *Current Opinion in Plant Biology* 5:325-331.

Lagudah, E.S., Appels, R., Brown, A.H.D. and McNeill, D. (1991) The molecular genetic analysis of *Triticum tauschii*, the D-genome donor to hexaploid wheat. *Genome* 34:362-374.

Lagudah, E.S., Moullet, O., and Appels, R. (1997) Map based cloning of a gene sequence encoding a nucleotide binding domain and a leucine-rich repeat region at the *Cre3* nematode resistance locus of wheat. *Genome* 40: 659-665.

Lanfermeijer, F.C., Dijkhuis, J., Sturre, M.J.G., de Haan, P., and Hille, J. (2003) Cloning and characterization of the durable tomato mosaic virus resistance gene *Tm-22* from *Lycopersicon esculentum*. *Plant Molecular Biology* 52:5:1037-1049.

Lawrence, J.G., 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. Plant Cell* 7:1195-1206.

Leister, D., Kurth, J., Laurie, D.A., Yano, M., Sasaki, T., Devos, K., Graner, A., and Schulze-Lefert. (1998) Rapid reorganization of resistance gene homologues in cereal genomes. *Proceedings of the National Academy of Sciences USA* 95:370-375.

Lenormand, T., Guillemaud, T., Bourguet, D. and Raymond, M. (1998) Appearance and sweep of a gene duplication: Adaptive response and potential for new functions in the mosquito *Culex pipiens*. *Evolution* 52:1705-1712.

Lepik, E.E. (1970) Gene centers of plants as sources of resistance. *Annual Review of Phytopathology* 8:323-344.

Lupas, A. (1966) Coiled-coils: new structures and new functions. *Trends in Biochemical Sciences* 21:375-382.

Mackey, D., Belkhader, Y., Alonso, J.M., Ecker, J.R., and Dangl, J.L. (2003) *Arabidopsis* RIN4 is a target of the type III virulence effector AvrRpt2 and modulates RPS2 mediated resistance. *Cell* 112:379-389.

Mackey, D., Holt, B.F., Wiig, A., and Dangl, J.L. (2002) RIN4 interacts with Pseudomonas syringae type III effector molecules and is required for *RPM1* mediated resistance in *Arabidopsis*. *Cell* 108:743-754.

Martin, G.B., Williams, G.K., and Tanksley, S.D. (1991) Rapid identification of markers linked to a Pseudomonas resistance gene in tomato by using random primers and near-isogenic lines. *Proceedings of the National Academy of Sciences* 88:2336-2340.

Martin, G.B., Brommonschenkel, S., Chunwongse, J., Frary, A., Ganal, M.W., Spivery, R., Wu, T., Earl, E.D., and Tanksley, S.D. (1993) Map-based cloning of a protein kinase conferring disease resistance in tomato. *Science* 262:1432-1436.

Martin, G.B., Bogdanove, A.J., and Sessa, G. (2003) Understanding the functions of plant disease resistance proteins. *Annual Review of Plant Biology* 54:23-61.

McDowell, J.M., Dhandaydham, M., Long, T.A., Aarts, M.G.M., Goff, S., Holub, E.B., and Dangl, J.L. (1998) Intragenic recombination and diversifying selection contributes to the evolution of Downey-mildew resistance at the *RPP8*-locus of *Arabidopsis*. *Plant Cell* 10:1861-1874.

McFadden, E.S. and Sears, E.R. (1946). The origin of *Tritcum spelta* and its free-threshing hexaploid relatives. *Journal of Heredeity* 37:81-89.

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

Meyers, B.C., Dicerkmann, A.W., Michelmore, R.W., Sivaramakrishnan, S., Sorbal, B.W. and Young, N.D. (1999) Plant Disease resistance genes encode members of an ancient and diverse protein family within the nucleotide-binding super family. *Plant Journal* 20:317-322.

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* 15:809-834.

Meyers, B.C., Morgante, M., and Michelmore, R.W. (2002) TIR-X and TIR-NBS proteins: two new families related to disease resistance TIR-NBS-LRR proteins encoded in *Arabidopsis* and other plant genomes. *The Plant Journal* 32:77-92.

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* 8:1113–1130.

Mindrios, M., Katagiri, F., Yu, G-L., and Ausabel, F.M. (1994) The *A.thaliana* disease resistance gene *rps2* encodes a protein containing a nucleotide-binding site and Leucine-Rich repeats. *Cell* 78:1089-1099.

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.

Monosi, B., Wisser, R.J., Pennill, L., and Hulbert, S.H. (2004) Full-genome analysis of resistance gene homologues in rice. *Theoretical and Applied* Gene*tics* 109:1434-1447.

Multani, D.S., Meeley, R.B., Paterson, A.H., Gray, J., Briggs, S.P., and Johal, G.S. (1998) Plant-pathogen microevolution: molecular basis for the origin of a fungal disease in maize. *Proceedings of the National Acadamy of Sciences USA* 95:4:1686-1691.

Munford, R.S., and Hall, C.L. (1986) Detoxification of bacterial lipopolysaccharides (endotoxins) by a human neutrophil enzyme. *Science* 234:203-205.

Nei, M., Gu, X., and Sitnikova, T. (1997) Evolution by the birth-and-death process in multigene families of the vertebrate immune system. *Proceedings of the National Academy of Sciences USA* 94:15:7799-806.

Niimura, Y., and Nei, M. (2003) Evolution of olfactory recepto*R* genes in the human genome. *Proceedings of the National Acadamy of Sciences USA* 100:21:12235-12240.

Noel, L., Moores, T.L., van der Biezen, E.A., Parniske, M., Daniels, M.J., Parker, J.E. and Jones, J.D.G. (1999) Pronounced intraspecific haplotype divergence at the *RPP5* complex disease resistance locus of *Arabidopsis* locus. *Plant Cell* 11:2099-2112.

O'Neill, L. (2000) The Toll/interleukin-1 receptor domain: a molecular switch for inflammation and host defence. *Biochemical Society Transactions* 28:5:557-563.

Ohno, S. (1970) Evolution by gene duplication, *George Allen and Unwin, London*.

Ori, N., Eshed, Y., Paran, I., Presting, G., Aviv, D., Tanksley, S., Zamir, D., and Fluhr, R. (1997) The *I2C* family from the wilt disease resistance locus *I2* belongs to the nucleotide binding, leucine-rich repeat superfamily of plant resistance genes. *Plant Cell* 9:4:521-532.

Otto, S.P., and Yong, P. (2002) The evolution of gene duplicates. *Advanced Genetics* 46:451-83.

Papageorgiou, A.C., Shapiro, R., and Acharya, K.R. (1997) Molecular recognition of human angiogenin by placental ribonuclease inhibitor - an X-ray crystallographic study at 2.0 Å resolution. *The EMBO Journal* 16:5162-5177.

Parker, J., Feys, B.J., van der Biezen, E.A., Noel, L., Aarts, N., Austin, M.J., Botella, M.A., Frost, L.N., Daniels, M.J., and Jones, J.D.G. (2000) Unravelling *R* gene-mediated disease resistance pathways in *Arabidopsis*. *Molecular Plant Pathology* 1:1:17-24.

Parniske, M., Hammond-Kosack, K.E., Golstein, C., Thomas, C.M., Jones, D.A., Harrison, K., Wulff, B.B., and Jones, J.D. (1997) Novel disease resistance specificities result from sequence exchange between tandemly repeated genes at the *Cf-4/9* locus of tomato. *Cell* 91:6: 821-832.

Pan, Q.L., Wendel, J., and Fluhr, R. (2000a) Divergent evolution of plant NBS-LRR resistance gene homologues in dicot and cereal geneomes. *Journal Molecular Evolution* 50:203-213.

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*. *Genetics* 155:309-322.

Penninckx, I.A., Eggermont, K., Terras, F.R.G., Thomma, B.G., De Samblanx, G.W., Buchala, A., Métraux, A.P., Manners, J.M., and Broekaert, W.F. (1996) Pathogen-induced systemic activation of a plant defensin gene in *Arabidopsis* follows a salicylic acid-independent pathway. *Plant Cell* 8:2309-2323.

Poltorak, A., Ricciardi-Castagnoli, P., Citterio, S., and Beutler, B. (2000) Physical contact between lipopolysacharide and Toll-like receptor 4 revealed by genetic complementation. *Proceedings of the National Academy of Sciences USA* 97:5:2163-2167.

Rairdan, G.J., and Delaney, T.P. (2002) Role of Salicylic Acid and *NIM1/NPR1* in Race-Specific Resistance in *Arabidopsis*. *Genetics* 161:803-811.

Rathjen, J.P., Chang, J.H., Staskawicz, B.J. and Michelmore, R.W. (1999) Constitutively active *Pto* induces a Prf-dependant hypersensitive response in the absence of avrPto. *The EMBO Journal* 18:12:3232-3240.

Ramalingham, J., Vera Cruz, C.M., Kukreja, K., Chittoor, J.-M., Wu, J.-L., Lee, S.W., Baraoidan, M., George, M.L., Cohen, M.B., Hulbert, S.H., Leach, J.E., and Leung, H. (2003) Candidate Defense Genes from Rice, Barley, and Maize and Their Association with Qualitative and Quantitative Resistance in Rice. *Molecular Plant Microbe Interactions* 16:1:14-24.

Reymond, P., Weber, H., Damond, M., and Farmer, E.E. (2000) Differential gene expression in response to mechanical wounding and insect feeding in *Arabidopsis*. *Plant Cell* 12: 707-719.

Rossi, M., Goggin, F.L., Milligan, S.B., Kaloshain, I., Ullman, D.E. and Williamson, V.M. (1998) The nematode resistance gene *Mi* of tomato confers resistance against the potato aphid. *Proceedings of the National Academy of Sciences* 95:9750-9754.

Royet, J., and Reichhart, J-M. (2003) Detection of peptidoglycans by NOD proteins. *Trends in Cell Biology.* 13:12:610-614.

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* 8:1809-1819.

Saleh, A., Srinivasula, S.M., Acharya, S., Fishel, R., Alnemri, E.S. (1999) Cytochrome c and dATP-mediated oligomerization of Apaf-1 is a prerequisite for procaspase-9 activation. *Journal of Biological Chemistry* 274:17941-17945.

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* 12:86:1:123-33.

Saraste, M., Sibbald, P.R., and Wittinghofer, A. (1990) The P-loop - a common motif in ATP- and GTP-binding proteins. *Trends in Biochemical Scienes* 15:430-434.

Scofield, S.R., Tobias, C.M., Rathjen, J.P., Chang, J.H., Lavelle, D.T., Michelmore, R.W., and Staskawicz, B.J. (1996) Molecular Basis of Gene-fo*R* gene Specificity in Bacterial Speck Disease of Tomato. *Science* 274:5295:2063-2065.

Seah, S., Sivasithamparam, K., Karakousis, K., and Lagudah, E.S. (1998) Cloning and characterization of a family of disease resistance gene analogs from wheat and barley. *Theoretical and Applied Genetics* 97:937-945.

Shen, K.A., Chin, D.B., Arroyo-Garcia, R., Ochoa, O.E., Lavelle, D.O., Wroblewski, T., Meyers, B.C., and Michelmore, R.W. (2002) *Dm3* is one member of a large constitutively expressed family of nucleotide binding site-leucine-rich repeat encoding genes. *Molecular Plant Microbe Interactactions* 15:3:251-61.

Shirasu, K., and Schulze-Lefert, P. (2000) Regulators of cell death in disease resistance. *Plant Molecular Biology* 44:371-385.

Song, W-Y., Wang, G-L., Chen, L-L., Kim, H-S., Pi, L-Y., Holsten, T., Gardner, J., Wang, B., Zhai, W-X., Zhu, L-H., Fauquet, C., and Ronald, P. (1995) A Receptor-Kinase Like Protein Encoded by the Rice Disease Resistance Gene, *Xa-21*. *Science* 270:1804-1806.

Sirinivasula, S.M., Ahmad, M., Fernandes-Alnemri, T. and Alnemri, E.S. (1998) Autoactivation of procaspase-9 by Apaf-1-mediated oligomerization. *Molecular Cell* 1:949-957.

Stakman, E.C. (1915) Relationship between *Puccinia graminis* and plants highly resistant to its attack. *Journal of Agricultural Research* 4:3:193-199.

Swiderski, M.R., and Innes, R.W. (2001) The *Arabidopsis PBS1* resistance gene encodes a member of a novel protein kinase subfamily. *Plant Journal* 2001 26:1:101-12.

Tai, T.H., Dahlbeck, D., Clark, E.T., Gajiwala, P., Pasion, R., Whalen, M.C., Stall, R.E., and Staskawicz, B.J. (1999) Expression of the *Bs2* peppe*R* gene confers resistance to bacterial spot disease in tomato. *Proceedings National Academy of Sciences USA* 96:24:14153-14158.

Tobias, I., Rast, B., and Maat, D.Z. (1982) Tobamoviruses of pepper, eggplant, and tobacco: comparative host reactions and serological relationships. *Netherlands Journal of Plant Pathology* 88:257-268.

Ton, J., De Vos, M., Robben, C., Buchala, A., Métraux, J-P., Van Loon, L-C. and Pieterse, M.J. (2002) Characterization of *Arabidopsis* enhanced disease susceptibility mutants that are affected in systemically induced resistance. *Plant Journal* 29:11-21.

The Arabidopsis Genome Initiative (2000) Analysis of the genome of the flowering plant *Arabidopsis thaliana*. *Nature* 408: 796-815.

Thomas, C.M., Jones, D.A., Parniscke, M., Harrison, K., Balint-Kurti, P.J., Hatzixanthis, K., and Jones, J.D.G. (1997) Characterization of the tomato *Cf-4* gene for resistance to *Cladosporium fulvum* identifies sequences that determines recognitional specifity in *Cf-4* and *Cf-9*. *Plant Cell* 9:2209-2224.

Traut, T.W. (1994) The functions and consensus motifs of nine types of peptide segments that form different types of nucleotide binding sites. *European Journal of Biochemistry* 222:9-19.

Traw, M.B., Kim, J., Enright, S., Cipollini, D.F., and Bergelson, J. (2003) Negative cross-talk between salicylate- and jasmonate-mediated pathways in the Wassilewskija ecotype of *Arabidopsis thaliana*. *Molecular Ecology* 12:5:1125-1135.

Underhill, D.M., and Ozinsky, A. (2002) Toll-like receptors: Key mediators of microbe detection. *Current Opinion in Immunology* 14:103-110.

Van den Ackerveken, G.F.J.M., Van Kan, J.A.L., and De Wit, P.J.G.M. (1992) Molecular analysis of the avirulence gene *avr9* of the fungal tomato pathogen *Cladosporium fulvum* fully supports the gene-fo*R* gene hypothesis. *Plant Journal* 2:359-366.

Van Der Biezen, E.A., Freddie, C.T., Kahn, K., Parker, J.E., and Jones, J.D. (2002) *Arabidopsis RPP4* is a member of the *RPP5* multigene family of TIR-NB-LR*R* genes and confers downy mildew resistance through multiple signalling components. *Plant Journal* 29:4:439-51.

Van Der Vossen, E.A., van der Voort, J.N., Kanyuka, K., Bendahmane, A., Sandbrink, H., Baulcombe, D.C., Bakker, J., Stiekema, W.J., and Klein-Lankhorst, R.M. (2000) Homologues

of a single resistance-gene cluster in potato confer resistance to distinct pathogens: a virus and a nematode. *Plant Journal* 23:5:567-76.

Van Der Vossen, E.A., Sikkema, A., Hekkert, B.L., Gros, J., Stevens, P., Muskens, M., Wouters, D., Pereira, A., Stiekema, W., and Allefs, S. (2003) An ancient *R* gene from the wild potato species *Solanum bulbocastanum* confers broad-spectrum resistance to *Phytophthora* infestans in cultivated potato and tomato. *Plant Journal* 36:6:867-882.

Van Wees, S.C., Pieterse, C.M., Trijssenaar, A., Van 't Westende, Y.A., Hartog, F., and Van Loon, L.C. (1997) Differential induction of systemic resistance in *Arabidopsis* by biocontrol bacteria. *Molecular Plant Microbe Interactions* 10:6:716-724.

Walton, J.D. (1996) Host-selective toxins: agents of compatibility. *Plant Cell* 8:1723-1733.

Wagner, A. (1998) The fate of duplicated genes: loss or new function? *BioEssays* 20:785-788.

Wang, Z.X., Yano, M., Yamanouchi, U., Iwamoto, M., Monna, L., Hayasaka, H., Katayose, Y., and Sasaki, T. (1999) The *Pib* gene for rice blast resistance belongs to the nucleotide binding and leucine-rich repeat class of plant disease resistance genes. *Plant Journal* 19:1:55-64.

Warren, R.F., Henk, A., Mowery, P., Holub, E., and Innes, R.W. (1998) A mutation within the Leucine-Rich Repeat Domain of the Arabidopsis disease resistance gene *RPS5* partially supresses multiple bacterial and Downey Mildew resistance genes. *The Plant Cell* 10:1439-1452.

Wei, F., Gobelman-Werner, K., Morroll, S.M., Kurth, J., Mao, L., Wing, R., Leister, D., Schulze-Lefert, P., and Wise, R.P. (1999) The *Mla* (Powdery Mildew) Resistance Cluster is Associated with Three NBS-LRR families and Suppressed recombination Within a 240-kb DNA interval on Chromosome 5S(1HS) of Barley. *Genetics* 153:1929-1948.

Wei, F., Wing, R.A., and Wise R.P. (2002). Genome Dynamics and Evolution of the *Mla* (Powdery Mildew) Resistance Locus in Barley. *Plant Cell* 14:8:1903–1917.

Witham, 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 Interleukin-1 Receptor. *Cell* 78:1101-1115.

Whitham, S.A., Anderberg, R.J., Chisholm, S.T., and Carrington, J.C. (2000) *Arabidopsis RTM2* gene is necessary for specific restriction of tobacco etch virus and encodes an unusual small heat shock-like protein. *Plant Cell* 12:4:569-582.

Xiao, S., Ellwood, S., Calis, O., Patrick, E., Li, T., Coleman, M., and Turner, J.G. (2001) Broad-spectrum mildew resistance in *Arabidopsis thaliana* mediated by *RPW8*. *Science* 291:5501:118-120.

Yahiaoui, N., Srichumpa, P., Dudler, R., and Keller, B. (2004) Genome analysis at different ploidy levels allows cloning of the powdery mildew resistance gene *Pm3b* from hexaploid wheat. *Plant Journal* 37:4:528-538.

Yoshimura, S., Yamanouchi, U., Katayose, Y., Toki, S., Wang, Z.X., Kono, I., Kurata, N., Yano, M., Iwata, N., and Sasaki, T. (1998) Expression of *Xa1*, a bacterial blight-resistance gene in rice, is induced by bacterial inoculation. *Proceedings of the National Academy of Sciences USA* 17:95:4:1663-8.

Zhang, W., Qu, L.J., Gu, H., Gao, W., Liu, M., Chen, J., and Chen, Z. (2002) Studies on the origin and evolution of tetraploid wheats based on the internal transcribed spacer (ITS) sequences of nuclear ribosomal DNA. *Theoretical and Applied Genetics* 104:(6-7):1099-1106.

Zou, H., Li, Y., Liu, X., and Wang, X. (1999) An APAF-1.Cytochrome c multimeric complex is a functional apoptosome that activates procaspase-9 activation. *Journal of Biological Chemistry* 274:11549-11556.

Chapter 3

Bioinformatic and phylogenetic analysis of *Triticeae* NBS-LRR homologues

3.1 Introduction

3.1.1 NBS-LRR members isolated from the wheat genome complex

The isolation and sequencing of disease resistance genes from a variety of plant-pathogen interaction models has greatly increased our understanding of inducible plant defense responses. A large fraction of these interactions are best described by the "gene-for-gene" model introduced by H.H. Flor (1971), which states that the elicitation of a successful disease resistance response in the plant host requires a functional resistance gene (*R* gene) in the host and its specific avirulence (*Avr)* gene complement in the phytopathogen. More than forty of these *R* genes have been isolated to date (Table 2.1) and they show very interesting relationships at the DNA and amino acid level (Meyers *et al.*, 1999, Martin *et al.*, 2003). Based on the modular domains present, resistance genes can be subdivided into five classes (Figure 2.1), some of which share functional domains and the signaling network of the hypersensitive response (HR). The five classes are based on domain structure and function and consist of detoxifying enzymes, intracellular serine-threonine protein kinases, nucleotide-binding-site-leucine-rich repeat proteins (NBS-LRRs), transmembrane LRR receptor-like proteins and transmembrane LRR receptor kinases (Hammond-Kosack and Jones, 1997). The majority of cloned *R* genes belong to the NBS-LRR class, which can be subdivided into two subfamilies based on the identity of the N-terminal domain, which carries either a coiled coil (CC) or a Toll-Interleukin Receptor homology (TIR) domain (Hammond-Kosack and Jones, 1997). More than 70% of NBS-LRR *R* genes (Table 2.1) belong to the CC-NBS-LRR (CNL) subfamily, which appears to be the only one present in the genomes of grasses and possibly all monocotyledonous plants (Meyers *et al.*, 2002).

Angiosperm genomes harbor large numbers of NBS-LRR genes, making out roughly 1% of the genes encoded by the *Arabidopsis thaliana* and *Oryza sativa* genomes, respectively (Monosi *et al.,* 2004). Members generally occur in clusters, containing closely related paralogous sequences, but in some cases multiple divergent NBS-LRR groups occur in a single cluster (Hammond-Kosack and Jones, 1997). A substantial number of NBS-LRR *R* genes occur as singletons, such as *RPS2* (Mindrios *et al.*, 1994)*, RPM1* (Grant *et al.*, 1995) and *RPS5* (Warren *et al.*, 1998). This complex genomic distribution is also characterisitic of other large multi-gene families in eukaryotic genomes, such as the olfactory receptor and adaptive immune system genes in vertebrate genomes (Flajnik and Kasahara, 2001; Niimura

and Nei, 2003). Two opposing, but not mutually exclusive models currently serve as heuristics for describing the evolution of multi-gene families. The older of the two is the concerted evolution model (Irwin and Wilson, 1990), which predicts high rates of intergenic gene conversion between members of a gene cluster, causing individual gene clusters to diverge as units from each other, with paralogues in a gene cluster becoming closer related than orthologues following a speciation event. In contrast, the birth-and-death model (Nei, 1997) predicts low rates of intragenic gene conversion, with orthology maintained for the individual loci in clusters. New specificities are "born" when duplicated loci, freed from selective constraint adopt new functionality. The latter model captures many of the characteristics of NBS-LRR family evolution as seen in previous studies in a diverse range of plant taxa (Michelmore and Meyers, 1998).

Numerous *R* genes have been isolated in the last decade from the genomes of both model and crop plants (See Table 2.1). The first *R* genes in the genome of bread wheat, namely *Lr21, Lr10* and *Pm3b* were isolated by Huang *et al. (*2003), Feuillet *et al.* (2003) and Yahiaoui *et al.* (2004) respectively. All three genes belong to the CNL family and in the present study I aimed to investigate the structure and evolution of the members of this gene family in the allohexaploid genome of wheat (*Triticum aestivum*) and its close relatives in the *Triticeae* tribe. To date only a tiny fraction of the expected number of NBS-LRR-like sequences have been obtained from the wheat genome (Maleki *et al.*, 2003). In 1997, Lagudah and co-workers isolated the first NBS-LRR sequence from a monocot genome using a molecular marker co-segregating with the *Cre3* gene, which provides resistance against the Australian pathotype of the cereal cyst nematode (*Heterodera avenae*). This gene was originally introgressed into hexaploid wheat from *Aegilops tauschii*. The marker, *CsE20*, which did not contain a coding region, was initially used for probing a wheat genomic library, from which a pseudogenic NBS-LRR sequence was obtained. This genomic sequence was used in turn for probing a root cDNA library, from which a clone (designated *CD2*) was obtained. *CD2* possessed the domain structures typically found in NBS-LRR genes and showed high homology to the NBS-LRR pseudogene probe sequence. *CD2* and the genomic pseudogene fragment were found to co-segregate with the resistance phenotype and mapped to the distal 0.06 cM fragment of chromosome 2DL.

In somewhat similar fashion, Frick and co-workers (1998) found a 1100 kb RAPD fragment co-segregating with the stripe rust resistance gene *Yr10* (*Puccinia striiformis*) located on the short arm of chromosome 1B. The sequence of the RAPD fragment was determined and found to be homologous to the NBS sequence of the *L6* flax rust resistance gene. Seah and

co-workers (1998) used the *Cre3* sequence identified by Lagudah and co-workers (1997) to design specific primers based on the Kinase-2 conserved motif of the NBS-domain. This approach yielded two new wheat NBS-LRR RGA-sequence fragments and three new barley NBS-LRR RGA-fragments. Spielmeyer and co-workers (1998) used the RGAs identified in the above-mentioned studies and generated additional cereal RGA sequences for maize, rice and barley using specific and degenerate PCR approaches. In addition, Southern blotting was used to identify five wheat RGA-like segments from a seedling cDNA library using an RGA probe from barley (*Hv1LRR*), which was in turn derived from a barley genomic library screened with a *Cre3* subclone (Lagudah *et al.*, 1997). The obtained sequences were pooled with sequences identified previously, yielding five wheat, eight barley, four maize and two rice sequences, which were all mapped onto the wheat genome via Restriction Fragment Length Polymorphism (RFLP) analysis using the ITMI (International *Triticeae* Mapping Initiative; http://www.scri.sari.ac.uk/ITMI/ default.html) mapping population. As could be expected from the homeology of the three wheat sub-genomes, many RGAs mapped to homeologous locations across all three sub-genomes. A clear clustering pattern was evident since many individual Resistance Gene Analogue (RGA) probes mapped to the same chromosomal locations, often close to known resistance loci.

Spielmeyer and co-workers (2000) used the earlier mentioned *Yr10* linked NBS-LRR fragment identified by Frick and co-workers (1998) to detected homologous NBS-LRR RGAs on chromosome 1DS of wheat. The detected RGAs co-localized with a known leaf rust (*Puccinia triticana*) resistance gene, *Lr21* on chromosome 1DS. An RFLP marker *KSUD14,* also segregating with *Lr21* was sequenced by Huang and Gill in 2001. The DNA sequence showed similar domain structure to that obtained for the *Cre3* sequence from wheat, containing motifs indicative of the NBS domain of plant *R* genes. The *Lr21 R* gene as mentioned earlier was itself recently cloned by Huang *et al.* (2003) using a diploid-polyploid shuttle mapping strategy (Huang *et al.*, 2003), with genetic mapping performed in hexaploid wheat and library screening in a large-insert-library constructed from *Aegilops tauschii* genomic DNA. The *Lr21 R* gene was found to encode a CNL protein.

Scherrer and co-workers (2002) used sequence data (211 Kb) from a *T. monococcum* BAC-contig to describe two NBS-LRR RGAs in a region homologous to the *Lr10* leaf rust resistance locus in wheat (chromosome 1AS). The RGAs were used subsequently to identify their counterparts in a wheat genomic library. The wheat versions mapped to the same location on chromosome 1A as in *T. monococcum* and detected an additional sequence on chromosome 1D. PCR and hybridization analysis indicated two conserved haplotypes of

approximately 200kb in screened wheat populations, spanning the *Lr10* rust resistance locus. As mentioned earlier, the *Lr10 R* gene was cloned in 2003 by Feuillet *et al*. also using a shuttle-mapping strategy, using a *Triticum monococcum* large-insert library and a similar strategy was employed by Yahiaoui *et al.* in 2004 to clone the *Pm3b R* gene that confers resistance to *Blumeria graminis* from hexaploid wheat. Like *Lr21* and *Lr10, Pm3b* also encode CNL proteins.

Maleki *et al.* (2003) utilized the *Cre3*, *Ksud14* (Huang and Gill, 2001) and *Yr10* NBS sequences available to design degenerate primer sets for amplification of wheat NBS segments spanning from the P-loop to the GLPLAL region. They obtained only two novel NBS clones using this approach. Using a reverse primer 22 amino acids short of the GLPLAL motif, they obtained an additional 6 novel NBS-LRR RGA sequences (designated *KSU940-947*).

A number of studies, including some of the above mentioned generated numerous CNL sequences for barley. The studies of Madsen *et al.* (2003) and Rostoks *et al.* (2002) yielded a large number of closely related expressed NBS-LRR sequences, some of which were mapped in the barley genome, revealing their clustered organization and association with previously characterized resistance loci.

NBS-LRR sequences have in addition been obtained through the efforts of the ITEC (International *Triticeae* EST cooperative) initiative, which has generated over 500 000 ESTs for wheat and over 300 000 ESTs for barley. This excellent source of transcriptional data should be utilized to its full capacity for obtaining and characterizing new expressed NBS-LRR gene sequences for wheat and barley, where ultimately resistance-breeding programs can benefit from this effort.

In the present study I aimed to characterize the domain structure, diversity and evolution of the CNL gene family in cereal species of the *Triticeae* tribe, in context of current models of the evolution of this multigene family in other plant taxa. My first objective to this end was to establish a comprehensive dataset of publically available sequences for NBS domains of the NBS-LRR gene family. Using this dataset I aimed to characterize firstly conserved motifs in the NBS domains, to determine whether they represent the CNL families characterized in other plant species, and to consider any evidence for TIR-NBS-LRR (TNL) type NBS domains. I further aimed to study the relationship of *Triticeae* NBS-LRRs clades with functional CNL *R* genes by performing a number of phylogenetic analyses on the union of these two datasets. I also aimed at characterizing the evolution of the gene family at the hand of existing models of multi-gene, and more specificallty, *R* gene evolution.

Models of multigene family evolution (Otto and Yong, 2002), built around classic population genetics predict that loci where overdominant selection is possible, are likely to produce the majority of fixed gene duplications observed in natural populations, where new specificities are generated as alleles at a single locus prior to duplication via unequal recombintaion in a heterozygote as opposed to previous applications of the birth-and-death model where duplication precedes divergence (Michelmore and Meyers, 1998). Considering that numerous NBS-LRR loci with alleles encoding multiple specificities are well known (Ellis *et al.*, 1999; Wei *et al.*, 2002), either balancing or overdominant selection is most likely operating across these loci, and in the context of this model, I aimed to study two duplication events, for which this model predicts different outcomes: paragolous gene duplications (functional divergence) and allopolyploidy mediated homeologous gene duplications (mutation to pseudogene). In order to study the evolutionary fate of these duplications, I evaluated basic parameters of gene family evolution, including nonsynonymous to synonymous substitution rate (Ka:Ks) ratios and gene conversion rates. I aimed to obtain and study the evolution of NBS-LRR sequences resulting from recent paralogous expansions from the results of my planned phylogenetic analysis, while identifying homeologous NBS-LRR sequences for the A (*Triticum urartu*), B (*Aegilops speltoides*) and D (*Aegilops tauschii*) genomes of hexaploid wheat by PCR using specific primer sets targetted to two previously mapped NBS-LRR sequences, namely *go35* (Lagudah *et al.*, 1997) and *KSU945* (Maleki *et al.,*. 2003).

3.2 Materials and Methods

3.2.1 Plant Materials

Triticum turgidum (AABB), *Triticum urartu* (likely AA donor), *Aegilops speltoides* (likely BB donor) and *Aegilops tauschii* (DD) seed (accessions PI 221425, PI 428317, PI 499261 and TA 1649, respectively) was obtained from the Germplasm Bank at Kansas State University. Hexaploid wheat (AABBDD) Tugela*Dn1* (SA1684/Tugela*5) seed was obtained from the Small Grain Institute, Bethlehem, South Africa. Seeds were planted in well-drained potting soil and kept in a controlled environment due to high ambient temperatures. A simple day-night cycle of 12 hours was implemented and the temperature kept at 16°C.

3.2.2 Methods

3.2.2.1 Database-mining

BLAST searching (Altschul *et al*., 1990) is commonly used to interrogate large DNA and protein sequence databases. Although slightly less sensitive than exhaustive local and global alignment algorithms such as Needleman-Wunsch (Needleman and Wunsch, 1970) and Smith-Waterman (Smith and Waterman, 1981), the BLAST algorithm is more than an order of magnitude faster. Since BLAST performs a heuristic search, it is prone to missing long weak alignments, which are easily picked up by algorithms that do exhaustive searches (Baxevanis and Ouellette, 2001). The shear size of public molecular sequence databases, limits search methods largely to variations of the BLAST algorithm (blastn, blastp, megablast, tblastx, tblastn, blastx, PHI-BLAST (Pattern Hit Iterated) and PSI-BLAST (Position Specific Iterated)).

Beside its speed, another distinct advantage associated with BLAST searching is that a statistical framework exists for measuring the significance of a given hit. The expectation value (E-value) gives the expected number of hits with the same or higher significance when entering a random sequence of the same information content as the query sequence into a database of the size querried. E-values can thus be larger than one, although one would typically be interested in homologies with E-values several orders of magnitude lower for inferring orthology (Karlin and Altschul, 1993).

Hidden Markov Models (HMMs) were initially developed and applied in complex pattern recognition problems such as voice and speech recognition (Rabiner, 1989). A Hidden Markov Model contains statistical parameters in the form of two matrices, one describing the transitions between a different number of hidden states, and the other describing the emition probablilities for each hidden state. In addition, the distribution of initial hidden state frequencies is required. Algorithms for determining the most likely state sequences for a given observation and for determining the probability of emitting this observation rely on dynamic programming principles and are very fast and efficient once the model has been properly built and calibrated. The HMM models used for describing sequence features are typically reduced to a subset known as profile HMMs (Eddy, 1998) which are restricted in their transition patterns for hidden states, allowing for insertion, deletion and match states. With the recent emergence of large amounts of genomic and transcriptional data, HMMs are also becoming a standard tool in detecting biologically relevant signals in sequence data,

being superior to Position Specific Scoring Matrices (PSSMs) for detecting distant homology and having a rigid statistical underpinning (Delorenzi and Speed, 2002).

PSI-BLAST searches

Previously annotated NBS-LRR sequences were retrieved from the Genbank database (All non-redundant Genbank CDS translations, RefSeq Proteins, PDB, SwissProt, PIR and PRF) at NCBI for members of the *Triticeae* tribe using Position Specific Iterated BLAST (PSI-BLAST) searches (Altschul *et al.,* 1990). The amino acid sequence of the NBS domain of *Lr21,* the first wheat *R* gene to have been characterized at the molecular level (Huang *et al.,* 2003) was used as the initial seed for building an initial PSSM. Using the full amino acid sequence was less effective since proteins containing the more variable domains, such as the LRR caused spurious hits. Searching was repeated until the result set converged for an E-value cutoff of 10^{-7}. The BLOSUM-62 matrix was used, and has been shown to be among the best at detecting weaker protein similarities, as is the case when searching for distant NBS-LRR homologues (Henikoff and Henikoff, 1992). All sequences were subsequently cropped to the region spanning from the P-loop up to the GLPL region (core-NBS) via scripted paired alignment to the seed sequence, using Perl scripting (http://www.perl.com) and custom EMBOSS (European Molecular Biology Open Source Suite, http://www.hgmp.mrc.ac.uk/Software/EMBOSS) library extensions for global alignment. The resulting dataset was aligned using T-Coffee (Notredame *et al.*, 2000; http://igs-server.cnrs-mrs.fr/~cnotred/ Projects_home_page/t_coffee_home_page.html) and sequences lacking any of the conserved motifs between the P-loop and GLPL region were removed.

HMM searches

A profile HMM search was trained to detect putative member sequences among the Gene Indices (GIs) maintained by The Institute for Genomic Research (TIGR; http://www.tigr.org). The TIGR Expressed Sequence Tag (EST) based Gene Indices (GIs) are generated by a clustering process where sequences representing a single transcript are grouped in a single cluster. The generation of consensus sequences for each cluster greatly enhances the utility of the vast number of EST sequences publicly available. Clusters with a single member, called singletons, are assigned unique gene indices and form their own consensus. The EST set used as the starting point for clustering contains among others 540 000 wheat and 340 000 barley ESTs recently generated through the efforts of the International *Triticeae* EST Cooperative (ITEC, http://wheat.pw.usda.gov/genome), the USDA-ARS Center for Bioinformatics and

Comparative Genomics at Cornell University (http://www.ars.usda.gov) and the U.S. Wheat Genome Project (http://www.ars.usda.gov/ NSF).

In order to perform the HMM search, the wheat and barley tentative clusters (TCs) and singletons were downloaded from TIGR via File Transfer Protocol (FTP) in FASTA format and translated in all six reading frames using the EMBOSS toolkit. A profile HMM for the NBS domain of the *Triticeae* was trained on the T-Coffee alignment of the PSI-BLAST dataset. This model was used to scan through the barley and wheat translations at an E-value (Expectation value) threshold of 10, with E-values computed based on the size of the Pfam-database (Protein families, http://wustl.pfam.edu). Search results were parsed and combined into a single FASTA format file and cropped to the region spanning from the P-loop up to the GLPL region and non-redundantly merged (for both accession and sequence) using Perl scripting language, EMBOSS and locally developed EMBOSS extensions. All profile HMM training and searches were conducted using the HMMer package (Durbin *et al.*, 2000; http://hmmer.wustl.edu).

Dataset reduction

Since many of the sequence pairs in the final alignment were near identical, the dataset was reduced by filtering out all sequences showing more than 95% identity. This was accomplished by Perl scripted looping with distance calculations performed by a modified version of the Needle program of the PHYLIP package (Felsenstein, 1989) which implements the Needleman-Wunsch global alignment algorithm (Needleman and Wunsch, 1970). The Needle program was modified to ignore terminal gap overhangs in computing percentage amino-acid identity when comparing sequences of different lengths.

Multiple sequence alignment

Members of the NBS-LRR gene family often exhibit amino acid identity as low as 30%, and only the residues in core motifs of the domain are strongly conserved (Meyers *et al.*, 1999; Cannon *et al.*, 2003). This complicates accurate alignment of multiple sequences in regions stretching between conserved motifs, which in turn negatively impacts motif alignment. Sequence alignment was thus performed using the profile HMM models built for database mining, in order to improve the alignment of conserved motifs hidden in more variable regions. HMM based alignments are also much faster than pair-wise methods such as T-Coffee and ClustalW (Thompson *et al.*, 1994) once the profile HMM model has been built and are very accurate. Full length or fragmentary sequences can also be added to existing

alignments with ease (Eddy, 1995; 1998). Multiple sequence alignments were manually edited mainly for removing large indel regions using the BioEdit sequence alignment editor (http://www.mbio.ncsu.edu/BioEdit/bioedit.html), since indel regions can create large biases in phylogenetics results. Gap positions were also flagged as uninformative characters.

3.2.2.2 Phylogenetic inference

The immense size and diversity of the NBS-LRR family provided a particularly challenging dataset for phylogenetic methods. In order to cross-validate results, various basic phylogenetic approaches were applied. All indel (insertion-deletion) and unreliably aligned regions in multiple sequence alignments were removed prior to application of distance, parsimony and maximum-likelihood methods.

Distance methods

The alignment obtained by data mining was bootstrapped to 1000 replicates. Pair wise distance matrices were generated using the PAM (Point Accepted Mutations) series of scoring matrices (Dayhoff, 1983) via the Protpars program (Felsenstein, 1989), which was executed in parallel across a 64 node Linux cluster using C programming language Message Passing Interface (MPI) controlled execution of the Protdist program.

The PAM001 matrix consists of scores for mutations from a given amino acid to each possible amino acid. The score is derived from alignments of proteins that differ by mutation of 1% of their amino acids residues. Multiplying the PAM001 matrix by itself multiple times, produces matrices for scoring progressively divergent alignments, like the PAM040, PAM120 and PAM250 matrices. Using the appropriate matrix for the level of sequence divergence optimizes the average information content score for matched pairs and this technique is used extensively in multiple alignment programs where alignments of various divergence levels are scored (Dayhoff *et al.*, 1983).

Distance trees were constructed from the distance matrices generated using the neighbor-joining approach as implemented in the Neighbor program of the PHYLIP package (Felsenstein, 1989), and a consensus tree indicating bootstrap support was generated for the 1000 distance trees using the Consense program of the PHYLIP package. The branches of the final bootstrapped topology were augmented with maximum-likelihood distance estimates as computed by the TREE-PUZZLE program (Schmidt *et al.*, 2002; http://www.tree-puzzle.de/). The BLOSUM62 distance matrix (Henikoff and Henikoff, 1992) was chosen along with accurate parameter estimation and eight Gamma distributed rate categories allowing for

heterogeneous substitution rates across sites, since the substitution pattern of accepted mutations along positions in the NBS domain differs greatly along sites.

Maximum Likelihood

Phylogenetics approaches based on probability aims at ranking alternative tree topologies based on either their prior or posterior probabilities under a specific sequence evolution model (Durbin *et al.*, 2000). The prior probability or likelihood for a specific tree topology is the probability of observing the alignment dataset given the tree topology in question. The posterior probability (Bayesian approach) is the probability of observing the tree in question when considering the alignment dataset being tested evolving under the specific sequence model chosen (Durbin *et al.*, 2000). Beside tree topologies, maximum likelihood methods also consider branch lengths (representing evolutionary time) when searching for the most probable tree topologies. This forces the use of heuristic methods to find one of the more probable trees in the countless topologies possible for large datasets. When computing either the prior or posterior probabilities, amino-acid substitution matrices are used and heterogeneity in mutation rate is accommodated in the form of gamma-distributed rates (Schmidt *et al.*, 2002).

Maximum likelihood methods were performed using the TREE-PUZZLE program (Schmidt *et al.*, 2002). TREE-PUZZLE follows three separate steps in order to construct a maximum likelihood tree from an alignment, rendering a tree with estimates of statistical support for each branch pattern. During the first step, trees for all possible combinations of four sequences or quartets are evaluated. For each quartet, the three possible topologies are evaluated and weighted according to their posterior probabilities. The quartets with high probability values are subsequently maintained in a set of supported quartets. During the second step, known as quartet puzzling, a single quartet is used as the starting point for randomly adding the remaining sequences in a manner corresponding best to the supported quartet topologies. This yields a specified number of intermediate trees, which are combined into a single consensus tree in the final step, for which branch patterns with statistical support exceeding 50% are created. Finally, branch lengths and ML-values are estimated (Schmidt *et al.*, 2002).

TREE-PUZZLE was implemented using eight gamma-distributed rate categories. As only clades with support levels above 50% are created, trees with multi-furcating nodes are generated whereas the distance and parsimony approaches both yield only bifurcating trees.

The parallelised version of the TREE-PUZZLE program was used for all maximum-likelihood methods and was executed across a 64-unit Linux cluster.

Maximum Parsimony

The maximum parsimony based phylogenetic reconstruction minimizes the total number of evolutionary events implied by the tree generated. Since the number of tree topologies generated grows rapidly, only around 12 taxa can be analyzed in a reasonable timespan using the computational capacity of modern hardware systems. Therefore, heuristic methods are implemented for larger datasets, such as stepwise addition and branch swapping algorithms (Nei and Kumar, 2000).

Parsimony analysis was performed using programs in the PHYLIP package (Felsenstein, 1989). Sequence datasets were bootstrapped using the seqboot program and execution of the protein parsimony program (protpars) was distributed in parallel over a 64-unit Linux cluster using MPI as implemented in the Local Area Multicomputer (LAM). Consensus generation was performed using the consense program with default parameters (Felsenstein, 1989).

Tree visualization

Due to the large number of leaf nodes present for the phylogenetic trees constructed, standard freeware tree visualization packages such as TreeView (http://taxonomy.zoology.gla.ac.uk/rod/treeview.html) were not viable for visualization and manipulation of the large tree-structures produced by the various phylogenetic approaches. Visualization was thus performed using a locally developed JAVA application with additional options for tree viewing such as translation, zooming and rendition of large tree-images to bitmap (BMP) image files. Support was also integrated for manual co-linearization of terminal nodes for comparing the topologies obtained for the various phylogenetic approaches. TreeJuxtaPoser is a freeware package supporting the concurrent visualization of large phylogenies. However, support for terminal node co-linearization is lacking, hampering visual comparison (Munzner *et al.*, 2003). The association of original sequence information with the node labels received as output from the various phylogenetic packages applied allowed visualization of sequences with respect to their accession numbers, taxa and descriptions. Support was also included for rendering the results of the MEME (Bailey and Elkan, 1994) motif discovery tool in association with terminal node labels.

Motif extraction

MEME (Bailey and Elkan, 1994; http://meme.sdsc.edu/meme/website/intro.html) allows extraction of conserved motifs from a set of unaligned sequences. This tool becomes very useful for visualizing motif conservation among sequence families when they are as divergent as the NBS-LRR gene family. This method has been applied successfully by Meyers and co-workers (2003) to simplify visualization of important domains/motifs and rearangements of them within NBS-LRR sequences annotated in the *Arabidopsis thaliana* genome. MAST (Motif Alignment and Search Tool) (Bailey and Gribskov, 1998; http://meme.sdsc.edu/meme/website/intro.html) is used in conjunction with MEME for visualizing hits and provides an alignment of input sequences based on the presence and location of MEME motifs in addition to performing databases searches with defined motif patterns. As with BLAST search results, MAST hits have E-values assigned, which provide an estimation of the expected number of random hits of similar significance.

Short conserved motives were extracted from the alignment used for phylogenetic tree construction using MEME. MEME output includes "machine readable" format, which allows visual association of MEME motif hits for individual sequences with their position in the phylogenetic trees developed during previous steps via the visualization package developed.

Synonymous and nonsynonymous substitution rate estimation

The ratio of nonsynonymous substitutions per nonsynonymous site (Ka) to synonymous substitutions per synonymous site (Ks) is an important indicator of the mode of evolution of a particular gene. The great majority of genes are evolving under purifying selection and have Ka:Ks ratios that are much smaller than one, since amino acid changes are far more likely to disrupt an existing protein function than enhance it, while silent mutations usually have negligible effects on fitness (Kimura, 1983). In pseudogenes, where selective constraint is completely abolished and sequence evolution is completely neutral, Ka:Ks ratios are close to one (Miyata, 1980). In a very small number of protein families, positive selection operates, resulting in sites with Ka:Ks values that are larger than one. Examples of positive selection occur predominantly in proteins that are directly involved in detection and recognition during host-pathogen co-evolution (Lee, 1995), and has been inferred previously for specific residues in the LRR of NBS-LRR genes (Mondragón-Palomino *et al.*, 2002).

Several methods exist for estimating Ka and Ks from sequence alignments. They can be grouped into three categories: 1.) Evolutionary pathway methods, 2.) methods based on Kimura's 2-parameter model and 3.) maximum-likelihood based methods with codon

substitution models. The first class of methods computes all possible pathways for a given codon transitions to partition the transition count into synonymous and nonsynonymous parts. The second method divides codons into three groups based on degeneracy and then uses approximations relating transitions and transversions to synonymous and nonsynonymous substitutions. The third method uses a model of codon substitution to determine maximum-likelihood estimates for the transition-transversion bias and the nonsynonymous-synonymous substitution ratio (Nei and Kumar, 2000).

Pamilo and Bianchi (1993) and Li (1993) independently extended Li *et al.*'s method (1985) based on Kimura's two parameter model of nucleotide substitution (Kimura, 1980). This method is more accurate under transition-transversion bias, and was implemented in this study for estimating Ka and Ks values. Ka:Ks ratios were calculated for selected clades in 3.3.2.1 by taking the average value calculated for all pairwise comparisons using MEGA version 3.0 (Kumar *et al.*, 2004).

Detection of gene conversion

Gene conversion events result in the copying of one segment of DNA onto another. Short-segment gene conversions often occur at higher frequencies than point mutations and play a prominent role in the evolution of multigene families (Guttman and Dykhuizen, 1994; Nei and Kumar, 2000). Procedures for detecting gene conversion events are based on runs (Sneath 1998), detection of changes in local estimated phylogenies (Maynard Smith and Smith, 1998) and a number of other techniques (Drouin *et al.*, 1999). In this study GENECONV (Sawyer, 1999; http://www.math.wustl.edu/~sawyer/ geneconv/) was used to detect gene conversion events in nucleotide and protein sequence alignments. GENECONV classifies polymorphisms in an alignment relative to a pair of sequences in the alignment as either concordant (same polymorphism in pair) or discordant (pair differ for polymorphism). Conversion events are then detected by finding the highest scoring fragments bound by discordant sites or sequence ends. Only fragments with P-values less than 0.05 are considered candidate gene conversion events. In this study, gene conversion was investigated for nucleotide alignments of specific clades identified in the phylogenetic analysis, see 2.2.2.

3.2.2.3 Amplification of *R* gene sequences

DNA Extraction

DNA extractions were performed according to the protocol of Edwards *et al.* (1991). Approximately 0.5 g of fresh leaf material was used for extractions and this yielded sufficient amounts of genomic DNA to allow direct scooping following ethanol precipitation. Extracted DNA was examined by 1% (w/v) agarose gel electrophoresis and spectrophotometry at wavelengths of 230, 260, 280 and 320 nm.

Specific Primer design

Two previously mapped NBS-LRR genes identified were targeted for allele amplification across the wheat genome donor group, namely the *go35* gene at the *Cre3* locus (Lagudah *et al.*, 1997) and the *KSU945* sequence identified by Maleki *et al.* (2003).

The *KSU945* gene hybridizes to two *HindIII* RFLP bands in wheat, one mapping to a locus on chromosome 1B and the other to chromosome 2D (Maleki *et al.*, 2003). The *Cre3* NBS-LRR locus contains a cluster of NBS-LRR genes on chromosome 2DL, including the *go35* gene, which was recently used to detect a new homologous locus on 2BL through hybridization techniques (de Majnik *et al*., 2003).

Two primer pairs were designed for amplifying diagnostic regions from the NBS domain of the two genes mentioned (Figure 3.1). Primers (Table 3.1) were chosen based on the results of Primer3.0 primer design software (Rozen and Staletsky, 2000). Both of the fragments targeted for amplification lacked intron sequences.

PCR reaction conditions for the amplification of specific sequence fragments using the primers in Table 3.1 were as follows: 20 ng/µl genomic DNA, 0.2 µM each primer, 0.16 mM dNTP's, 2 mM magnesium chloride, 1X Promega PCR buffer (without $MgCl_2$) and 1 unit (0.2 µl) of Promega *Thermus aquaticus* (Taq) DNA polimerase. Predicted T_m values for primer sets as calculated by the manufacturer were incorporated into an amplification protocol with the denaturation and extension temperatures of 95°C and 72°C, respectively. Thermocycling was performed in a Perkin Elmer PCR System (GeneAmp 9700) according to the following program: Pre-amplification denaturation step of 5 minutes, 30 cycles of amplification (0:30@95°C; 0:30@57°C; 0:30@72°C) and final extension of 5 minutes at 72°C.

Degenerate primer sets

The degenerate primer sets (Table 3.2) designed by Yu *et al*. (1996) were synthesized at Inqaba Biotec (Pretoria, South Africa). The two degenerate primers are designated NB1 and NB2. NB1 is a 512-fold degenerate 23-mer targeted to the P-loop motif amino-acid sequence GPGGVGKT and NB2 is a 128-fold degenerate 23-mer targeted to the RNBS-B motif (Resistance NBS) amino-acid sequence CKVMFTTR. NB1 and NB2 was used successfully in the original Yu *et al.* (1996) study for amplifying novel NBS-LRR sequences from the *Glycine max* genome. The nucleotide sequences for both are indicated in Table 3.2.

Degenerate PCRs were optimized using the methods of Taguchi as modified by Cobb and Clarkson (1994). The following set of parameters was found optimal for amplification of discrete bands: 50 ng/µl genomic DNA, 1µM each primer, 0.16 mM dNTPs, 2.5 mM magnesium chloride, Promega (Madison, Wisonsin, USA) PCR buffer (without MgCl2) and 1 unit (0.2µl) of Promega Taq DNA polymerase. The following thermo-cycling program was found optimal: Pre-amplification denaturation step of 5 minutes, 35 cycles of amplification (1:00@95°C; 1:30@55°C; 1:00@72°C) and a final extension of 5 minutes at 72°C.

Characterization and sequencing

The PCR amplification products generated by degenerate and specific primer sets were examined by 1% (w/v) agarose gel electrophoresis. Selected bands amplified by degenerate primer sets were extracted from agarose gels and purified with the Geneclean III kit (BIO101, Carlsbad, California, USA). Amplification products were ethanol precipitated and ligated to Promega pGEM®-T Easy vector. High efficiency (>108 cfu/µg) *E.coli* JM109 cells (Promega) were transformed with ligation mix and plated onto 70mm Luria-Bertani (LB) medium-Agar dishes (15g/L agar, 10g/L tryptone, 5g/L yeast extract and 5g/L NaCl at pH 7.0) containing 80µg/ml X-gal (5-bromo-4-chloro-3-indolyl-*β*-D-galactosidase), 0.5mM IPTG (Isopropyl-*β*-D-thiogalacto-pyranoside) and 80µg/ml ampicilin for blue-white and ampicilin-resistance based selection of transformants.

Inserts in white colonies were examined by PCR using small amounts of toothpick-transferred colony material as source of plasmid DNA (Güssow and Clackson, 1989). *Sp6* (5' cat acg att tag gtg aca cta tag 3') and *T7* (5' taa tac gac tca cta tag gg 3') promoter targeted primers were used for amplifying cloned fragments from both orientations of the vector's cloning cassette. Ethidium bromide-stained agarose gels were photographed using a BioRad Versadoc imaging system (Rosebank, South Africa). Band sizes were scored using the BioRad Quantity One software accompanying the BioRad Versadoc imaging system.

Colony PCR products from clones containing fragments of the required sizes were sequenced using BigDye™ Dye Terminators v3.0 (Applied Biosystems, Warrington, UK). Sequencing gels were run on an ABI™ 3100 automated sequencer (Applied Biosystems, Warrington, UK). Base calling was performed using the Perkin Elmer Genescan program and curated by hand where necessary.

3.3 Results

3.3.1 Database mining

To obtain a comprehensive dataset of *Triticeae* NBS-LRR sequences for phylogenetic analysis, sequence data was mined from protein annotations in Genbank and additional data gathered from EST data. The gene indices compiled by TIGR were used as source of EST data. For this study, analysis was restricted to the NBS domain since it shows the highest degree of motif conservation, which greatly facilitates database mining, multiple sequence alignment and distinction of TIR-NBS-LRR and CC-NBS-LRR genes (Meyers *et al.,* 1999; Pan *et al.*, 2000b).

3.3.1.1 PSI-BLAST search

For the *Triticeae* tribe, a PSI-BLAST search was performed on the Genbank database at NCBI (*http://www.ncbi.nlm.nih.gov*) using the *Lr21 R* gene of *Aegilops tauschii* as initial query. This was done firstly to obtain all NBS-LRR protein sequences for the *Triticeae* in Genbank and secondly for providing an alignment of the *Triticeae* NBS-LRR family upon which the HMM model can be built for refined searching of the wheat EST data curated by TIGR. The PSI-BLAST search was performed using only the NBS-domain of the *Lr21 R* gene, and reached convergence at 164 hits, after four iterations at an E-value threshold of 10^{-7}. Only 106 of the 164 hits obtained spanned the P-LOOP and GLPL motifs, and all hits were cropped to this region as described in the methods section. Eleven of the 106 sequences clearly lacked one or more of the motifs found in the NBS-region of NBS-LRR *R* genes (Meyers *et al.*, 1999; Meyers *et al.*, 2002; Meyers *et al.*, 2003). The remaining 95 sequences were derived from the species *Aegilops tauschii, Triticum aestivum, Triticum turgidum, Hordeum vulgare, Aegilops ventricosa, Triticum monococcum, Thinopyrum intermedium* and *Secale strictum*.

Considering that the rice genome at 350 Mb (Pers. Commun. Prof. Jan A Leach, CSU, 2005) in size contains in excess of 600 NBS-LRR genes (Bai *et a.l,* 2002), these 95 sequences represent only a small subset of the complete gene family projected for species of the *Triticeae* tribe, showing that despite extensive transcript sequencing efforts, the majority of NBS-LRR gene family members have not yet been detected for wheat or barley. This result can be expected if genes of the NBS-LRR family in the *Triticeae* are expressed constitutively at very low levels, as would be the case for receptor proteins located at the very start of signal transduction pathways (Shen *et al.*, 2002). To maximally extend this dataset, an initial alignment was generated for training an HMM model to be used for detecting NBS-LRR homologues in wheat EST data.

3.3.1.2 Hidden Markov Model searching of EST databases

The wheat and barley gene indices compiled by TIGR (version: Release 7.0 August 20, 2003) were downloaded as FASTA format flat files. A total of 109 509 wheat gene indices were compiled from 315 276 individual ESTs and consisted of 68 161 singletons and 41 348 tentative cluster consensus sequences. The 41 190 barley gene indices (version Release 7.0 August 20, 2003) were compiled from 343 206 individual ESTs and consisted of 27 930 singletons and 21 955 tentative cluster consensus sequences. Current estimates on transcriptome coverage for wheat are in the range of 60% (Li *et al.*, 2004) with transcripts assigned to 32 881 gene clusters (each representing the A, B and D genome homeologs). For barley, EST data has been clustered and assigned to 22 000 genes representing the barley transcriptome for microarray analysis (Close *et al.*, 2004). The significantly greater extent of barley EST clustering observed in the TIGR gene indices appear to be due to unlinked EST clusters originating from single wheat transcripts, rather than differences in transcriptome coverage or sampling error.

A 178-position profile HMM was trained and calibrated for the 89 sequence T-coffee alignment of the final PSI-BLAST dataset using the HMMer toolkit. This HMM was used for searching translations of the EST data in all six reading frames, and the resulting ORFs were searched using the PSI-BLAST based HMM model. Altogether 70 sequences were detected for the wheat dataset and 97 hits for the barley dataset. Of these hits, 22 spanned the P-loop and GLPL motifs for wheat and 67 for barley. The significantly higher number of barley hits was best explained in this case by a relatively higher presence of rare transcripts in the barley cDNA libraries used for EST generation, which would be the case if a higher fraction of barley cDNA libraries were normalized. Alignment of the obtained sequence hits using the

Triticeae HMM model, confirmed that all major motifs identified in previous studies (Meyers *et al.*, 1999; Meyers *et al.*, 2002; Meyers *et al.*, 2003) were present.

Both the PSI-BLAST and EST datasets were non-redundantly merged for accession, sequence and sub-sequence duplication using custom Perl scripts controlling execution of EMBOSS tools. Of the 22 wheat EST sequences, six were not presented in the PSI-BLAST results, while 25 of the 67 barley EST sequences were not duplicated in the PSI-BLAST results. The profile HMM search thus contributed significantly to the datamining set, adding 6 wheat and 25 barley transcript sequences to the final 155-sequence dataset.

Preliminary alignments of the 155 sequence dataset revealed a number of sequences having near-identical allelic or orthologous sequences. Regarding the low level of average sequence identity between NBS sequences (around 30%); the dataset was further reduced by removing all sequences with representatives within a 5% amino-acid difference range (considered probable alleles in the Bai *et al.*, 2002 study). This reduced the 155 sequence dataset to 92 sequences, leaving 5 wheat EST sequences, 22 barley EST sequences and 60 PSI-BLAST obtained sequences. The sequences eliminated at this stage were stored and all multiple taxon sequence sets collapsed to single nodes saved for later addition to phylogenetic trees. At this point, a pairwise distance matrix was generated for the dataset to assess the level of sequence divergence. For the 92 Triticeae sequence dataset (Appendices A and B), the average pairwise distance was 177 PAM units (Point Accepted Mutations per 100 residues). Seventeen sequences representing the diversity found in the 92 Triticeae sequences obtained for phylogenetic analysis are shown in Figure 3.2. Closely related *R* gene relatives are included for reference.

3.3.2 Motif analysis

Motifs were detected in the *Triticeae* sequences using MEME for comparison to NBS motifs of CNL and TNL genes from other plant taxa and for validating the presence of key motifs identified in other studies. MEME was set to detect nine motifs, leaving the possibility of detecting motifs additional to the six major motifs described (Feuillet *et al.*, 2003; Meyers *et al.*, 2003) in the NBS region. The unaligned dataset was analyzed and results obtained in both the standard HTML format and as motif summaries with coordinates for co-visualization with the phylogenetic trees to be generated. The consensus MEME motifs obtained for the *Triticeae* dataset are given in Table 3.3 along with their counterparts in *Arabidopsis* as identified by Meyers *et al.* (2003). Distribution of the motifs over individual sequences is

illustrated in Figure 3.1 and Figure 3.3-Figure 3.5, where it can be seen that the internal motifs (RNBS-A, kinase2, RNBS-B and RNBS-C) were present in individual *Triticeae* sequences. Thus the sequences obtained via database mining had all the motifs present in functional NBS domains, in the expected order and as seen in the diagrammatic representations in Figure 3.3, with the same inter-motif spacing as cloned CC-NBS-LRR *R* genes.

Meyers *et al.* (2003) and Pan *et al.* (2000b) found major differences between the motif structure of NBS domains for CC-NBS-LRR and TIR-NBS-LRR NBS genes. Pan *et al.* found the P-loop, Kinase2, RNBS-B and GLPL motifs to be similar between these two classes of genes, while the RNBS-A, RNBS-D and especially the RNBS-C motif differed considerably. The same pattern was seen in the *Triticeae* data, with the P-loop, kinase2, GLPL and RNBS-B motifs all sharing a substantial number of conserved residues with both the TNL and CNL NBS motifs of *Arabidopsis.* The motifs listed in Table 3.3, that differ to the largest extent between CNL and TNL genes (Meyers *et al.*, 2003), namely the RNBS-A and RNBS-C motifs, were similar to their CNL counterparts in *Arabidopsis*, but shared virtually no residues with their TNL counterparts. Except for the RNBS-B motif, all other *Triticeae* motifs in Table 3.3 were more similar to their CNL than TNL counterparts in *Arabidopsis*. Based on this observation and also the study of Bai *et al*. (2000b), no TNL NBS domains appear to be present in the *Triticeae* dataset. This is in agreement with previous studies concerning the NBS-LRR gene family in cereals and other monocots. The presence of TIR-NBS-LRR genes in gymnosperms (Meyers *et al.*, 2002; Morgante *et al.*, 2002) and apparent absence in the *Triticeae* and other monocots could be explained by a loss of TIR-NBS-LRR genes from the common ancestor of modern monocotyledons. Interestingly TIR-X and TIR-NBS genes are still present at low levels in grass genomes (Meyers *et al.*, 2002), with these genes expected to interact through homodomain interactions with TIR-NBS-LRR genes like the MyD88 protein in animal systems as discussed under 2.1.2.2.

Two alternative versions of the RNBS-A motif were detected by MEME and are listed in Table 3.3. The visual representation of motif occurrences in Figure 3.3-Figure 3.5 shows that the alternate RNBS-A motif did not form monophyletic clades as might be expected. An additional motif was also detected between the kinase2 and RNBS-B motifs in two alternate forms (Table 3.3). The visualizations in Figure 3.3-Figure 3.5 show that these alternate forms occur in separate sequence clades, although not forming a strict monophyletic group across all three phylogenetic trees. The substitution patterns within these motifs might be interdependent yielding similar motif patterns in separate clades through homoplasy.

Structural constraints on the NBS region are also apparent in the variation of intermotif distances as illustrated in Figure 3.3-Figure 3.5. The highest variability was seen in the longest spacer region stretching between the RNBS-A and Kinase2 motifs, probably due to the greater number of functional configurations for this longer peptide.

3.3.3 Phylogenetic analysis

In order to compare the *Triticeae* NBS-LRR gene family with the 25 characterized CNL *R* gene sequences, phylogenetic analysis was performed on a non-redundant combination of the *Triticeae* dataset and all 25 CNL *R* gene sequences (Appendix A). Few *Triticeae* sequences spanned the full NBS-domain. To include the maximum number of fragmentary sequences while retaining the motifs sufficient for distinguishing TNL and CNL genes (Meyers *et al.,* 2003), phylogenetic analysis was limited to the core NBS fragment spanning the P-loop and GLPL motifs. The NBS domain of the *Apaf-1a* gene for *Rattus norvegicus* was added as outgroup, yielding a final dataset of 118 sequences. This dataset was aligned using the *Triticeae* CNL HMM built for EST datamining and three basic phylogenetic methods namely distance, parsimony and maximum likelihood were performed. This allowed comparison of clade support across multiple approaches for this highly divergent dataset.

The multi-furcating ML tree obtained using the TREE-PUZZLE program is shown in Figure 3.3 and the distance and parsimony trees generated by PHYLIP programs protdist and protpars are shown in Figure 3.4-Figure 3.5 respectively. The three phylogenies were co-linearized for their terminal node order using the java-based viewer developed for generating Figure 3.3- Figure 3.5. This allowed visual comparison of all three phylogenies and fast comparison of clade support across the three methods employed.

Statistical support for branching patterns of the ML phylogeny in Figure 3.3 was calculated via quartet puzzling as percentage of occurrence in intermediate trees. TREE-PUZZLE generates by default only branching patterns with statistical support of 50% and higher. For the distance and parsimony trees, statistical support for each branch is given as the number of supporting bootstrap topologies out of a thousand. Branch lengths for all three trees were scaled to evolutionary distance in PAM units using TREE-PUZZLE.

The ML tree in Figure 3.3 contains a total of nine singletons and twenty-seven gene clades. Clades were labeled on this tree, since all branch patterns generated by TREE-PUZZLE carry minimum statistical support of 50% (as computed by quartet puzzling). The larger clusters and those containing *R* genes were labeled alphabetically from A to P on the ML tree and

indicated on the distance and parsimony trees (Figure 3.3 and Figure 3.5). Table 3.4 shows the number of species and statistical support for each clade across the three phylogenetic approaches implemented. The majority of clades were supported by at least two of the three phylogenetic methods employed (Table 3.4). The low levels of statistical support obtained for deeper branching patterns was mostly due to the limited length of sequence alignment available for phylogenetic analysis, the large number of sequences in the dataset and limited alignability of inter-motif regions. Sequence conversion and recombination was also assessed as discussed below, but did not appear to contribute significantly to the unresolvability of deep branches.

Substitution rates varied significantly between the different clades (see scale in Figure 3.3-Figure 3.5), in similar fashion to that seen in the larger multi taxon trees produced in the study of Cannon *et al* (2002). This appears to be an inherent problem when generating phylogenies for large datasets (order of one hundred) using limited lengths of sequence alignment, since recent studies (Pan *et al.*, 2000b; Bai *et al.*, 2002; Meyers *et al.*, 2003) utilizing the full length of the NBS domain for smaller datasets show more consistency in substitution rates. The average distance per sequence for the ML, distance and parsimony trees was 52, 39 and 42 PAM units, respectively. As one would expect for a more restricted taxonomic group, this is lower than the 63.3 PAM units per sequence calculated for the multi-taxon CC-NBS-LRR family tree produced in the study of Cannon *et al.* (2002), but similar to the 40.9 PAM per sequence calculated for the *Arabidopsis* CNL tree produced in the same study.

3.3.3.1 Clades containing functional homologues

Of the 25 isolated CC-NBS-LRR *R* genes (indicated in red in Figure 3.3-Figure 3.5), thirteen occur in ten of the *Triticeae* clades, while the remaining twelve form a single clade in both distance and parsimony analysis. Eight of these thirteen overlapping *R* genes are from grass genomes and the remaining five from dicotyledonous species [Clades H and I contain *RPI* (*Solanum bulbocastanum*) and *I2* (*Lycopersicon esculentum*) respectively, whilst clades J and N contain *RPS2* (also *RPS5*) and *RPM1* from *Arabidopsis thaliana* respectively]. Considering that only a small fraction of the expected number of *Triticeae* NBS-LRR genes were available for this analysis, the degree of overlap between the dicot CNL *R* genes and the *Triticeae* NBS-LRR family is striking, suggesting that the NBS-LRR family in the *Triticeae* also functions mainly in resistance. It should also be noted that four of the five dicot *R* genes mentioned are considered ancient; *RPS2*, *RPS5* and *RPM1* exist as singletons with conserved

alternate haplotypes (Grant *et al.*, 1995; Caicedo *et al.*, 1999; Tian *et al.*, 2002), while the *RPI* locus contains four paralogues with high synonymous divergence and few amino-acid substitutions (Van der Vossen *et al.*, 2003). Interestingly, both *RPS2* and *RPM1* from *Arabidopsis* are known to guard a single host protein, *RIN4* (closest grass homologue: 37% identity in *Oryzae sativa*) against modification as discussed under heading 2.3. It might thus be that these ancient *R* genes have a conserved guard function across a wide range of plant taxa, as opposed to newer rapidly evolving clusters.

Four of the 25 CNL *R* genes included originate from the *Triticeae*: *Lr21* from *Aegilops tauschii*, *Lr10* and *Pm3b* from *Triticum aestivum* and *Mla* from *Hordeum vulgare*. The *Lr10* locus also exists as a singleton with a balanced polymorphism for a functional allele and null allele (Scherrer *et al.*, 2002). *Lr10* was grouped in the ML tree with a single orthologue from *Triticum monococcum* in clade L, consistent with it being located on chromosome 1AS as singleton, the A genome being closely related to that of *Triticum monococcum* (Kimber and Sears, 1987). The *Lr21* gene also occurs as a singleton (Huang *et al.*, 2003) on 1DS, and has a proximally located paralogue that is closely related as seen in clade E. The barley *Mla* locus is more complex with three families of NBS-LRR sequences present and the closest barley paralogue in clade K originating from this locus. As opposed to *Lr21* and *Lr10*, *Mla* contains multiple alleles conferring specificity to various *Blumeria graminis* isolates. The close orthology of the barley *Mla* gene with a wheat sequence as indicated in clade K, shows that barley can be a good model for the NBS-LRR family of wheat, whereas the rice *R* genes *Pib*, *Pi-ta* and *Xa1* have very distant orthologues in the *Triticeae*.

The clade composed of E and F contains in addition to *Lr21*, the *go35* CC-NBS-LRR sequence for the *Cre3* nematode resistance locus (chromosome 2DL) isolated by Lagudah *et al*. (1997) and the *KSU945* (gi17940787) sequence mapped to chromosome 2D by Maleki *et al.* (2003). *KSU945* and *go35* were selected for amplification across the diploid genome donors of wheat as well as their polyploid derivatives as discussed in section 3.4.1. No closely related barley homologues for the sequences in clades E and F have been isolated, although close relatives for *A. ventricosa* and *T. turgidum* are known.

3.3.3.2 Evolution of recently diverged paralogue clades

The *Triticeae* NBS-LRR phylograms in Figure 3.3-Figure 3.5 contain three clades (B, C and G) with numerous barley sequences that have diverged recently. Datasets including all sequences removed at the 5% identity cropping step (2.3.1.4) were generated for these clades, yielding 46 coding sequences for clades B and C and 12 coding sequences and 3 pseudogenes

for clade G. Sequences in clades B and C were published and mapped by Madsen *et al.* (2003) and those in clade G by Rostoks *et al.* (2002). Mapping positions are as indicated in Figure 3.3. The pair of clade B sequences on chromosome 5H, the pair of clade C sequences on chromosome 2H and the pair of sequences on chromsome 7H are found clustered in regions spanning around 5 cM. Clades B, C and G have diverged to different degrees with average pairwise PAM distances of 0.20, 0.36 and 0.47 respectively and this correlates well with the number of chromosomal locations that members have been mapped to, with clade B sequences mapping to one chromosome, clade C to two and clade G to three. These young paralogous sequence groups were investigated for their mode of evolution by examining nonsynonymous to synonymous substitution rates and by testing for tracts of gene-conversion.

Nonsynonymous to synonymous substitution ratios

The average pairwise synonymous to non-synonymous substitution rates for clades B, C and G were determined using the method of Li (1993) and Pamilo and Bianchi (1993) as implemented in MEGA version 3.0 (Kumar *et al.*, 2004). Average pairwise Ka:Ks ratios obtained were 0.33, 0.36 and 0.32 for clades B, C and G respectively. The low Ka:Ks ratios indicate large deviations from neutrality, with all three sequence clades evolving under purifying selection. No sequence pairs had Ka:Ks ratios close to, or larger than one. All of the sequences in clades B and C were obtained by genomic PCR, and virtually all of those tested for expression were detected by RT-PCR (Madsen *et al.,* 2003). Taken together, both the Ka:Ks ratios and active transcription status indicate that the sequences in these new clades are not pseudogenes.

Gene conversion and unequal recombination

Gene conversion was detected in the nucleotide sequence alignments of the NBS-domains for clades B, C and G using the program GENECONV (Sawyer, 1999) with various values for the mismatch parameter. The only statistically significant tracts of gene conversion detected were limited to short stretches occurring in highly conserved motifs. These short tracts (7 to 11 nucleotides) appear to be due purely to homoplasy between highly divergent sequences evolving under the same selective constraints.

Contrasting models of NBS-LRR evolution have been invoked for cereal and *Arabidopsis* NBS-LRR gene families. Ectopic translocation of NBS-LRR genes in *Arabidopsis* appears to have arisen mainly due to duplication of chromosome segments (Baumgarten *et al.*, 2003),

while in cereals, retrotransposition and retrotransposon-mediated ectopic recombination might have had a greater influence (Leister *et al.*, 1998). Considering that gene conversion is readily detectable between NBS-LRR loci in the *Arabidopsis* genome, even between sequences located on different chromosomes (Baumgarten *et al.*, 2003), the absence of gene conversion in this large sample of closely related paralogous sequence groups is striking and might indicate another distinction between cereal and *Arabidopsis* NBS-LRR evolution.

3.3.4 Sequence amplification

3.3.4.1 Allele sequencing

To examine the fate of recently duplicated NBS-LRR genes in the *Triticeae*, two NBS-LRR genes designated *go35* and *KSU945* were chosen for amplification across the diploid and polyploid species of the wheat complex. The *go35* gene from the *Cre3* locus (Resistance to Cereal Cyst Nematode) of *Aegilops tauschii* has previously been mapped to homeologous positions on the long arms of chromosomes 2B and 2D, and the *KSU945* sequence to chromosomes 1B and 2D.

For the section of the *Cre3* region spanning the kinase2 to GLPL motif, PCR bands of the expected size (460 bp) were obtained from *Aegilops speltoides* (BB), *Aegilops tauschii* (DD), *Triticum turgidum* (AABB) and hexaploid wheat (AABBDD) (Figure 3.6 and Figure 3.7), but was lacking in *Triticum urartu* (AA), from which was amplified only a single 1100 bp fragment (Table 3.5). The sequences for the 460 bp bands were determined and blasted against Genbank, identifying all as potential *go35* alleles (>98% identity). Failure to amplify the PCR bands close to the expected size range for the *go35* primer sets from the *T. urartu* genome agrees with the mapping locations (chromosomes 2DL and 2BL) detected by cross hybridization approaches (de Majnik *et al.*, 2003). This suggests that the *go35* homeoloci on chromosomes 2L are new clusters originating in the *Aegilops* genus, or have otherwise been lost from an ancestor of the *Triticum urarutu* (AA) genome after the divergence of *Triticum* and *Aegilops* genera. The *A. tauschii* sequence was found to be 100% identical to its published counterpart. The sequence obtained from *A. speltoides* was 98.8% identical to the *A. tauschii* version, and originates most likely from the homeolocus on chromosome 2B (de Majnik *et al.*, 2003). A total of five nucleotide substitutions were detected between the *A. speltoides* and *A. tauschii* sequences, all of which were silent, indicating that the gene is most likely evolving under purifying selection. The sequences derived from tetraploid (*T. turgidum*) and hexaploid wheat (*T. aestivum*) contained four of the five silent substitutions

seen in the *A. speltoides* sequence, as well a single non-synonymous substitution in the GLPL motif (Figure 3.8; Figure 3.10), mutating the motif from LK**GSPL**AART to LK**ESPL**AART. This mutation is very likely to have functional consequences and may indicate diminished selection pressure at this locus in polyploid wheat, although alternative explanations exist for this substitution pattern, including the bottleneck-effect brought about by the two polyploidization events of wheat. The lack of a *go35* homologue in the A genome however, shows that it did not buffer this locus in *T. turgidum* so as to relax selective constraint by duplicating function.

BLAST searching was performed to further augment the *go35* sequences obtained for polyploid wheats. The TIGR gene index for the wheat *go35* gene was based on a cluster composed of 5 sequences. Four of the five sequences spanned parts of the core NBS domain, one being the full length coding sequence from *Aegilops tauschii*. The three wheat NBS fragments spanning the NBS were obtained from wheat cultivars Xinong88 and Chinese Spring. Using the B-genome *go35* sequence and the *A. tauschii go35* sequence, the three sequences were classified as derivatives of either the B or D genome versions. Two of the three genes were derived from Xinong88 and one from Chinese Spring. The Chinese Spring derived gene and one of the two Xinong88 derived genes are most probably B-genome derived and contain four of the five silent mutations seen in the *A. speltoides go35* version, but both lack the nonsynonymous substitution seen in the GLPL motif of the *T. turgidum* and *T. aestivum* versions. This might be due to introgressed segments in the Xinong88 and Chinese Spring cultivars, since *T. turgidum* and Tugela *Dn1 T. aestivum* contained this mutation. The remaining Xinong88 sequence differed by one silent and one nonsynonymous substitution from the *A. tauschii go35* sequence, mutating the **GSKILVTTR** RNBS-B motif to **GSKIPVTTR**, and is likely derived from the D-genome.

The fragment sizes indicated in Table 3.5 for the *KSU945* primer pair shows that fragments of roughly the expected size (351 bp) were obtained from *Triticum urartu* (AA), *Aegilops speltoides* (BB) and *Triticum aestivum* (AABBDD). Sequence analysis confirmed that the fragments were indeed closely related to the published *KSU945* sequence (Figure 3.9 and Figure 3.11), but that pairwise sequence identity was around 90%, indicating that the amplification products are most likely paralogues at different chromosomal loci. Average pairwise Ka:Ks ratios between the three sequences and the published *KSU945* sequence for *T. aestivum* was 0.55, showing that some of the genes in this group are evolving under purifying selection. Maleki *et al.* (2003) mapped the *KSU945* gene to chromosome 1B and 2D using hybridization approaches. In this study, no bands of the expected size range were obtained

from the *A. tauschii* genome, possibly due to differences in the primer binding sites. This also appears to be the case for the *Triticum turgidum* (AABB) genome. Alternative explanations include homeologous recombination and reproducible selective sequence eliminations as have been described for the synthetic allopolyploids (Ozkan *et al.*, 2001).

3.3.4.2 Degenerate PCR

Degenerate PCR was performed to further extend the NBS-LRR sequences obtained by data-mining. Using the degenerate primer set designed by Yu *et al.* (1996), a distinct band at around 500 base pairs was obtained (Figure 3.12). This band was larger than expected for the primer combination (expected size around 340bp), but was isolated and glassmilk purified, as no bands closer to the expected size range were detected. The DNA fragments obtained were subsequently cloned into the pGEM®-T Easy Vector. Colony PCRs were performed on the resulting positive clones and band-sizes in the region of 500bp were observed. Sequencing was performed on a number of the variant band sizes. The majority of clones sequenced represented the *WIS-2-1A* retrotransposon sequence from *Triticum aestivum,* but a single clone showed homology to a putative rice NBS-LRR gene (Appendix C).

3.4 Discussion

3.4.1 Main findings

The combined iterative data-mining approach implemented in my study effectively expanded my search result set, but the number of NBS-LRR sequences obtained was much lower than the projected amount considering the wheat transcriptome coverage achieved by EST sequencing projects. I interpret this as evidence supporting low basal transciption levels as expected for *R* genes (Shen *et al.*, 2002), situated at the very start of signal amplification cascades.

My motif analysis showed that all key motifs of the CNL core-NBS domain was present in my dataset, with no evidence for TNL type sequences as previously observed for monocot taxa. I also found significant overlap between the *Triticeae* CNL members and CNL *R* genes from other taxa in my phylogenetics results. I tested three recently diverged clades of paralogous NBS-core sequences for barley that I identified in my phylogenetics analysis for gene-conversion events, but detected none. This is in contradiction to *Arabidopsis,* where

even ectopic gene conversion events have been detected previously. My Ka:Ks ratio tests for comparing the evolution of recent paralogous and homeologous duplications showed that the NBS-core domain of the three barley paralogue clades examined were under strong purifying selection in contrast to my results for the core-NBS domain of the wheat *go35* CNL gene. Here I identified four different nonsynonymous substitutions in polyploid wheats, whereas only synonymous differences were seen between the sequences obtained from two diploid ancestors of wheat, *A. tauschii* and *A. speltoides*. I consider this as evidence supporting a divergence-before-duplication model of *R* gene evolution.

3.4.2 Iterative data-mining approach detected a low number of CNL genes considering of total family size

In this study, a comprehensive set of *Triticeae* NBS-LRR gene family members was mined from public sequence databases. Wheat and barley transcripts constituted the majority of NBS-LRR sequences for the *Triticeae* tribe, mainly due to the large EST sequencing efforts that have been initiated for these two crop species (the International *Triticeae* EST Cooperative (ITEC) (http://wheat.pw.usda.gov/ genome), the USDA-ARS Center for Bioinformatics and Comparative Genomics at Cornell University (http://www.ars.usda.gov) and the U.S. Wheat Genome Project (http://www.ars.usda. gov/NSF)). Considering that the 350 Mbp rice genome (Pers. Commun. Prof Jan A Leach, CSU, 2005) contains around six hundred CC-NBS-LRR genes (Bai *et al.*, 2002), data-mining results show that only a small fraction of those expected for wheat (16 000 Mbp genome) have been sequenced. With current estimates of wheat transcriptome coverage in the range of 60% (Li *et al.*, 2004), the fraction of NBS-LRR genes mined was disproportionately small. This is in agreement with the idea that NBS-LRR genes are expressed at low basal levels (Shen *et al.*, 2002) in accordance with their function as receptors inducing signal-transduction cascades.

The PSI-BLAST searches at Genbank were sufficient for generating a base set of annotated NBS-LRR proteins for the *Triticeae*. The HMM built from this base set retrieved a significant number of non-redundant entries from translations of the TIGR gene indices, showing that EST clustering added significant value to wheat and barley EST data and that building the HMM model for harvesting NBS-LRR sequences enhanced detection of distant homologues. Some previous studies (Bai *et al.*, 2002; Meyers *et al.*, 2002; Monosi *et al.,* 2004) utilized only BLAST-based searches or combined them with HMM searches using the Pfam (Protein families database of alignments and HMMs) NB-ARC HMM (pfam00931.11). Since the Pfam NB-ARC domain is based on only 9 NBS regions derived from dicot *R* genes and

vertebrate outgroup sequences, the *Triticeae* specific HMM used in this study should enhance the sensitivity of TIGR gene index searches.

3.4.3 Motif analysis indicate typical CNL NBS-core for *Triticeae* NBS-LRRs

Motif analysis revealed that all motifs previously characterized in the core NBS-domain of numerous plant taxa were present in the sequences obtained by database mining. It has previously been shown that NBS-LRR genes of the CNL and TNL subclasses can be distinguished by the motif variants present in their NBS domains (Pan *et al.*, 2000b). Comparison of the motifs generated for the *Triticeae* dataset with those generated for the full compliment of *Arabidopsis* CNL and TNL genes by Meyers *et al.* in 2003 (Table 3.3) show that except for the RNBS-B motif, all *Triticeae* motifs were more similar to their CNL than TNL counterparts in *Arabidopsis*. Some *Arabidopsis* CNL type *R* genes, such as *RPP13* posses RNBS-B motifs more similar to the *Triticeae* RNBS-B motif. No explanation was apparent for the selective divergence of the RNBS-B motifs of the majority of *Arabidopsis* CNL sequences. Since all remaining motifs showed much higher similarity to their CNL counterparts in *Arabidopsis,* all NBS domains in the dataset appear to belong to CNL genes, in accordance with previous studies of this gene family in monocotyledons (Leister *et al.*, 1998; Pan *et al.*, 2000a; Bai *et al*., 2002).

3.4.4 Phylogenetic analysis reveals significant overlap with functional CNL *R* genes

Three phylogenetic methods were applied to the *Triticeae* NBS-LRR dataset in combination with the set of all characterized CNL *R* genes; distance, parsimony and maximum likelihood. Although low bootstrap values were obtained for many deeper branches due to high divergence, limited alignment length and challenging alignment of inter-motif regions, major clade structures identified with at least 50% statistical support in the maximum likelihood tree were present in the parsimony and distance trees once their terminal node orders were colinearized, using the custom phylogeny visualization program developed for this study. Previous studies where longer stretches of the NBS domain were available for alignment and where smaller datasets were used, reported higher bootstrap support for deep branching patterns (Monosi *et al.*, 2004). In addition to colinearization, the viewer facilitated visualization of large phylogenies (order of one hundred terminal nodes) in association with

motif detection results. Two motifs were detected in alternate forms, but these variants were not strictly monophyletic, as might be expected. Intra-motif dependencies with regards to substitution patterns might explain the independent generation of similar motifs in separate lineages, or alternatively the motif variants might be monophyletic in a better-resolved phylogeny. Previous studies on NBS-LRR evolution have not reported on patterns of motif distribution within phylogenies of the CNL or TNL subfamilies.

Ten of the twenty-seven sequence clades on the maximum likelihood tree were supported by functional *R* gene members, of which five were of dicot origin. The significant overlap seen between this small fraction of the *Triticeae* NBS-LRR gene family and the CNL *R* genes isolated from various taxa strongly implicates a role in resistance for this family in the *Triticeae*. In addition, the first three *R* genes have recently been characterized for wheat, namely *Lr21* (Huang *et al.*, 2003), *Lr10* (Feuillet *et al.,* 2003) and *Pm3b* (Yahiaoui *et al.*, 2004). All three of these *R* genes are members of the CNL subfamily. Currently, no examples of NBS-LRR genes performing functions different from that of typical *R* genes have been characterized in plants (Belkhadir, 2004).

3.4.5 Absence of gene conversion events support different model for *Triticeae* NBS-LRR evolution

Two important parameters were assessed to investigate the evolutionary dynamics of three recently diverged barley sequence clades in the dataset, namely gene conversion and nonsynonymous to synonymous substitution ratios. Despite the inclusion of numerous closely related paralogous or allelic sequences (>95% identity) in the analysis (Rostoks *et al.*, 2002; Madsen *et al.*, 2003), no evidence of gene conversion was detected; although interlocus and even a low rate of ectopic gene conversion have been detected with strong statistical significance in *Arabidopsis* and other dicot taxa (Baumgarten *et al.*, 2003). A low rate of gene conversion was also found in grass *R* genes in the study of Zhang *et al.* (2001), and is compatible with the birth-and-death model proposed by Nei (1997) for the MHC loci of vertebrates and adapted by Michelmore and Meyers (1998) to the evolution of the NBS-LRR family in plants. Evolutionary patterns of the NBS-LRR gene family in the cereals thus appear to differ from that of *Arabidopsis* and dicotyledons by diminished interlocus gene conversion operating in NBS-LRR clusters. The tight colinearity generally observed for grass genomes (Gale and Devos, 1998) also argues against the model of ectopic translocation by segmental duplication as proposed for *Arabidopsis.* The large fraction of retrotransposons making out grass genomes, is likely to dominate ectopic translocation events, and might also

be an important factor in unequal recombination events that expand or contract NBS-LRR existing clusters.

3.4.6 Ka:Ks ratios for NBS-LRR loci investigated differ for paralogues and homeologues, supporting a divergence-before-duplication model for NBS-LRR gene family expansion

Previous studies on substitution rates in CNL and TNL genes have detected positive selection for selected surface-exposed residues in the LRR domain and purifying selection in the NBS and TIR domains (Mondragón-Palomino, 2002). In this study, the average pairwise Ka:Ks ratio was determined for the core-NBS domain of closely related NBS-LRR genes that were recently duplicated by two independent mechanisms: segmental duplication (paralogues) and allopolyloidy (homeologues). I determined the Ka:Ks ratio for the core NBS of CNL paralogues in the three recently diverged barley sequence clades that I assessed for gene conversion, and examined synonymous and nonsynonymous substitutions in the alleles of the *go35* (Lagudah *et al.*, 1997) and *KSU945* (Maleki *et al.*, 2003) loci, of which the *go35* locus is thought to have two homeologues across a homeolocus on based on the results of previous hybridization studies (de Majnik *et al.*, 2003).

All three barley clades were found to evolve under purifying selection with Ka:Ks ratios around 0.3 for all three clades investigated. No pairwise comparisons within clades yielded Ka:Ks ratios indicative of neutral evolution. For clades B and C, only non-pseudogenic sequences were published whereas for the twelve sequences published by Rostoks *et al.* (2002) for clade G, three were pseudogenic. Low numbers of pseudogenes have also been reported for the *Arabidopsis* genome, making out around 10% of NBS-LRR genes in the genome. Considering that NBS-LRR *R* genes function as single dominant genes, with very low rates of transcription, it would be expected that following a gene duplication event, one of the two gene copies would experience no selection pressure and evolve neutrally, most likely forming a pseudogene and not diverging in function. Recent studies into the population genetics of gene duplication have proposed that the vast majority of duplications fixed in large populations are those providing an immediate selective advantage (Otto and Yong, 2002). This is mediated by an initial divergence in function of alleles at a single locus, with subsequent duplication by unequal crossover resulting in a haplotype with permanent heterozygote advantage. Considering that NBS-LRR alleles are known to evolve independent functions at single loci (e.g. numerous *Avr* specificties are encoded by alleles at flax *L* locus

and *Arabidopsis RPP13* locus (Ellis *et al.*, 1999; Bittner-Eddy *et al.*, 2000)) and that duplications of identical alleles would provide no selective advantage as they act qualitatively, the divergence-before-duplication model appears to be a good alternative explaining the apparent lack of pseudogenes that contradict current divergence-after-duplication models (Michelmore and Meyers, 1998).

In addition to intercluster duplication and ectopic translocations, multi-gene families are also shaped by polyploidization events. In this study, the fate of duplicate NBS-LRR loci was studied in tetraploid and hexaploid wheat, which are recent allopolyploids arising 10 000 and 8 000 years ago respectively. Since the three homeologus genomes of wheat are estimated to have diverged between 2.5 and 4.5 million years ago (Huang *et al.*, 2002a; Huang *et al.*, 2002b), it was possible to distinguish genes for each subgenome in polyploid wheat based on sequence homology to versions of the genome donor species. The *go35* gene from the *Cre3* locus has previously been mapped to chromsome 2DL and hybridizes very specifically to a *Cre1* homeolocus on chromosome 2BL (de Majnik *et al.*, 2003). Functional overlap exists between *R* gene actions from both *Cre* loci, in the form of resistance to the Australian pathotype of the cereal cyst nematode (CCN). In this study a PCR approach was used to clone and sequence part of the NBS-domain for the *go35* gene from diploid and polyploid wheats. In accordance with the study of de Majnik *et al.* (2003), no homologue was detected in *Triticum urartu* by PCR while near identical sequences were obtained for *T. aestivum* (Dn1 line)*, T. turgidum*, *A. tauschii* and *A. speltoides*. Translations of the B and D genome derived *go35* sequeces are identical, with five silent mutations distinguishing the two copies, indicating strong purifying selection. Both the *go35* sequences obtained in this study from tetraploid and hexaploid wheat appear to have been derived from the B genome by sequence homology. Both contain a mutation in the leading amino acid of the conserved GLPL motif, presumably one of the first nonsynonymous substitutions fixed in this region since divergence of the *A. speltoides* and *A. tauschii* versions. It is likely that this mutation negatively affects gene function provided that the remainder of the gene is still functional, since many single amino-acid substitutions in individual NBS-LRR *R* genes are known to cause inactivation (Yahiaoui *et al.*, 2004). Since the *go35* gene is most likely absent from the A genome, the presence of this mutation in the genome of *T. turgidum* (AABB) cannot be readily explained by polyploid buffering, but might be due to relaxed selection pressure brought about by the environmental changes associated with domestication or alternatively the genetic bottlenecks introduced by polyploidization and domestication.

Of the three *go35* sequences previously published for *T. aestivum*, two sequences derived from the line Xinong88 (one B-derived, one D-derived) also carried single nonsynonymous substitutions, while a third sequence derived from Chinese Spring (B-genome derived) was identical in translation to the *A. speltoides* and *A. tauschii go35* sequences, differing from the *A. speltoides* version by a single synonymous substitution. Considering that three out of the four *go35* sequences from polyploid wheat contained unique non-synonymous substitutions relative to the *A. tauschii* and *A. speltoides* versions, relaxed selection pressure at this locus is supported rather than a bottleneck effect, although it is not possible to tell whether this was the result of environmental factors associated with domestication or whether it is evidence of polyploid buffering at a homeolocus in hexaploid wheat. It appears that the *go35* gene has either been lost in the predecessor of diploid *Triticum* species or was gained in that of the *Aegilops* genus. In conclusion, I found evidence for relaxed selection pressure on both the B and D-genome derived *go35* sequences in polyploid wheat as opposed to the complete amino-acid sequence conservation for these genes in *A. tauschii* and *A. speltoides*. This result is in accordance with previous studies on the fate of duplicated genes in polyploids, where it was found that resistance genes are preferentially lost after genome duplication events, while highly expressed genes such as rRNAs and tRNAs, and those that are dosage dependant are often retained (Blanc and Wolfe, 2004).

The *KSU945* NBS-LRR sequence was also investigated in this study, but the presence of closely related paralogues prevented specific amplification from a single locus. The sequences obtained from *T. urartu*, *A. speltoides* and *T. aestivum* were between 89% and 95% identical to the *KSU945* sequence published for wheat. Previous studies in wheat have mapped the KSU945 fragment to chromosomes 1B and 2D. Since the 89% identical homologues from the A genome of *T. urartu* should have been detected in these mapping studies, it might have been eliminated during the poliploidization events involved in forming hexaploid wheat, as selective and repeatable sequence eliminations have been shown to occur in synthetic allopolyploids (Ozkan *et al.*, 2001)

3.4.7 Degenerate PCR approach yielded single NBS-LRR homologue

The degenerate PCR approach used to amplify additional NBS-LRR homologues, yielded no distinct bands of the expected size, and only a single NBS-LRR homologue sequence was obtained from the closest distinct band seen at 500bp. The majority of sequences were derived from high copy DNA elements such as chloroplast DNA and retrotransposons.

Previous studies focused on obtaining significant numbers of novel NBS-LRR genes from the wheat genome have also yielded limited results, most likely due to its large genome size (Spielmeyer *et al.*, 1998; Spielmeyer *et al.*, 2000; Maleki *et al.*, 2003).

3.4.8 Conclusions and future perspective

I found a low number of CNL genes for wheat relative to projected amounts, in public sequence databases, showing that more than transcriptome sequencing efforts will be required to obtain a comprehensive set of CNL genes. Previous studies in this direction yielding new classes of wheat NBS-LRR sequences have typically relied on hybridization techniques, with degenerate PCR approaches achieving lower rates of success as I found in my own study. This advocates, in my experience screening of C_0t enriched genomic libraries with a range of *Triticeae* core-NBS probes, rather than using degenerate PCR-based approaches.

In my phylogenetic analysis, I identified wheat NBS sequences that are related to ancient CNL guard proteins of *A. thaliana* and a number of CNL *R* genes from other taxa. I believe that further investigations into these homologues, which could potentially participate in ancient defense strategies common to a wide range of plant taxa is warranted. In the present study I was incapable of detecting gene-conversion events in a number of recently diverged barley paralogues, but once sequenced scaffolds containing such paralogous clusters are available, allele sequencing in the barley gene pool should provide a more accurate estimate of the extent of gene conversion in *Triticeae* species.

My comparison of paralogous and homeologus Ka:Ks ratios was complicated by a lack of finely mapped/sequenced scaffolds containing CNL clusters for wheat. The availability of newly characterised wheat CNL *R* genes, now allow such comparisons for loci of known resistance specificities. This comparison can also in future studies be extended to more loci, especially for Ka:Ks ratio determination across homologous CNL loci in young autopolyploid species. In order to further address the question of whether *R* gene specialization indeed precedes duplication in general, studies can be initiated aimed at detecting newly formed duplication haplotypes of previously characterized polymorphic NBS-LRR *R* loci alleles in natural populations, which are known to be polymorphic for resistance specificities, such as the *Mla* locus in barley (Wei *et al.*, 2002).

Many outstanding questions remain in the field of *R* gene mediated resistance in plants, including the roles and components of signal transduction pathways, the basis of *R* mediated

Avr recognition and the evolution of new specificities. The generation of new specificities at a rate sufficient to keep up with the evolution of new virulence genes in pathogen populations, which have much shorter generation times, is still a mystery. Also, the evolution of polymorphic CNL loci having independent *R* gene specificities is poorly understood, and space or time effects might be important in addition to heterozygote advantage or frequency-dependant selection.

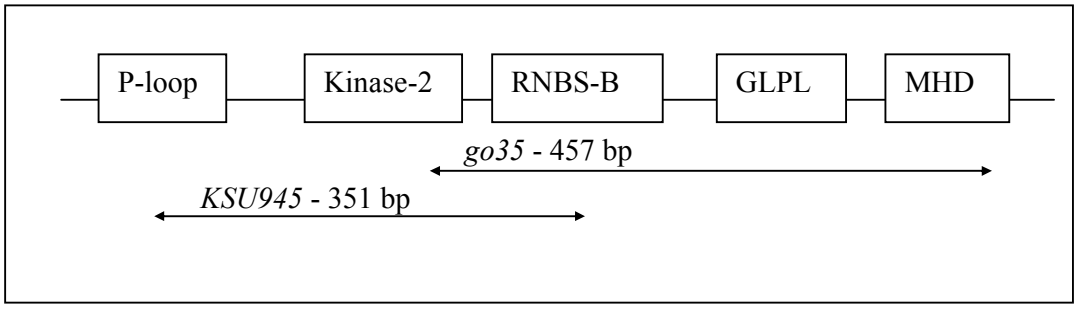


Figure 3.1 Schematic representation of the primer pairs utilized for amplification of the *KSU945* and *Cre3* NBS-LRR genes.

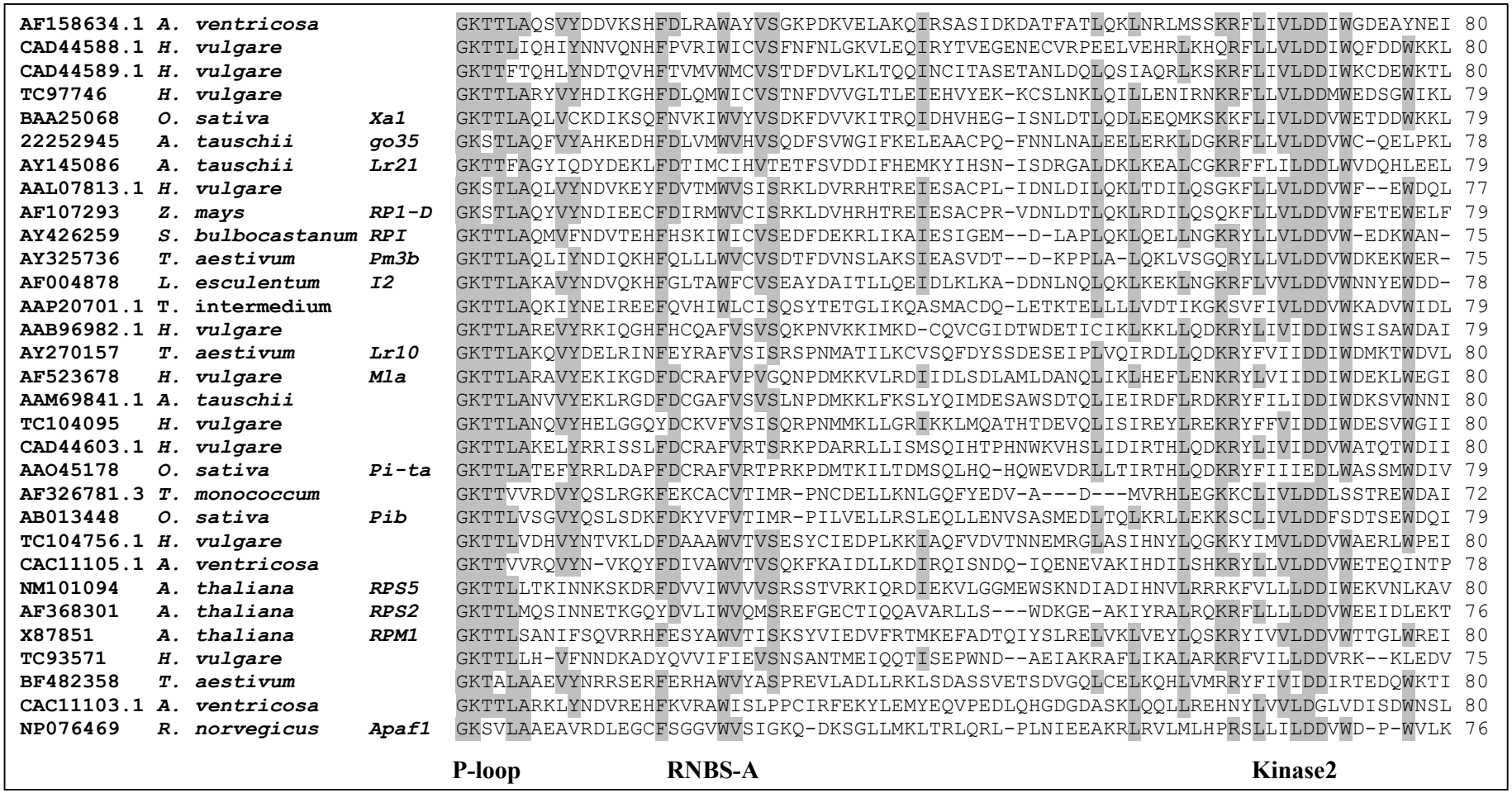


Figure 3.2 Multiple sequence alignment for translations of 17 sequences representative of the 155 sequences obtained by datamining. The closest *R* gene neighbours are included and the position of conserved motifs indicated. Major indels were removed and columns shaded for 50% amino acid conservation. The alignment was rendered using BioEdit ver 5.0.9 (Hall, 1999).

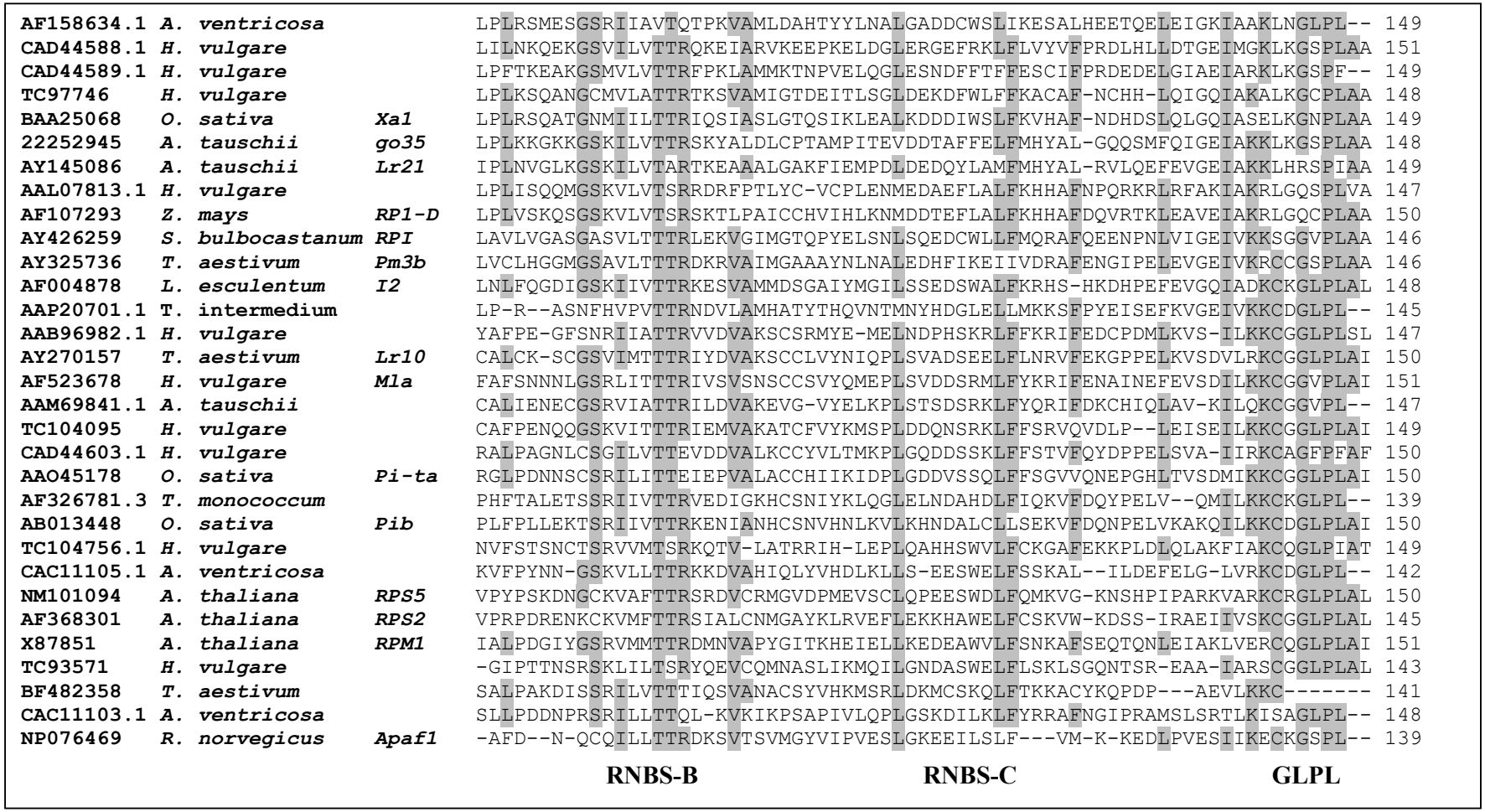
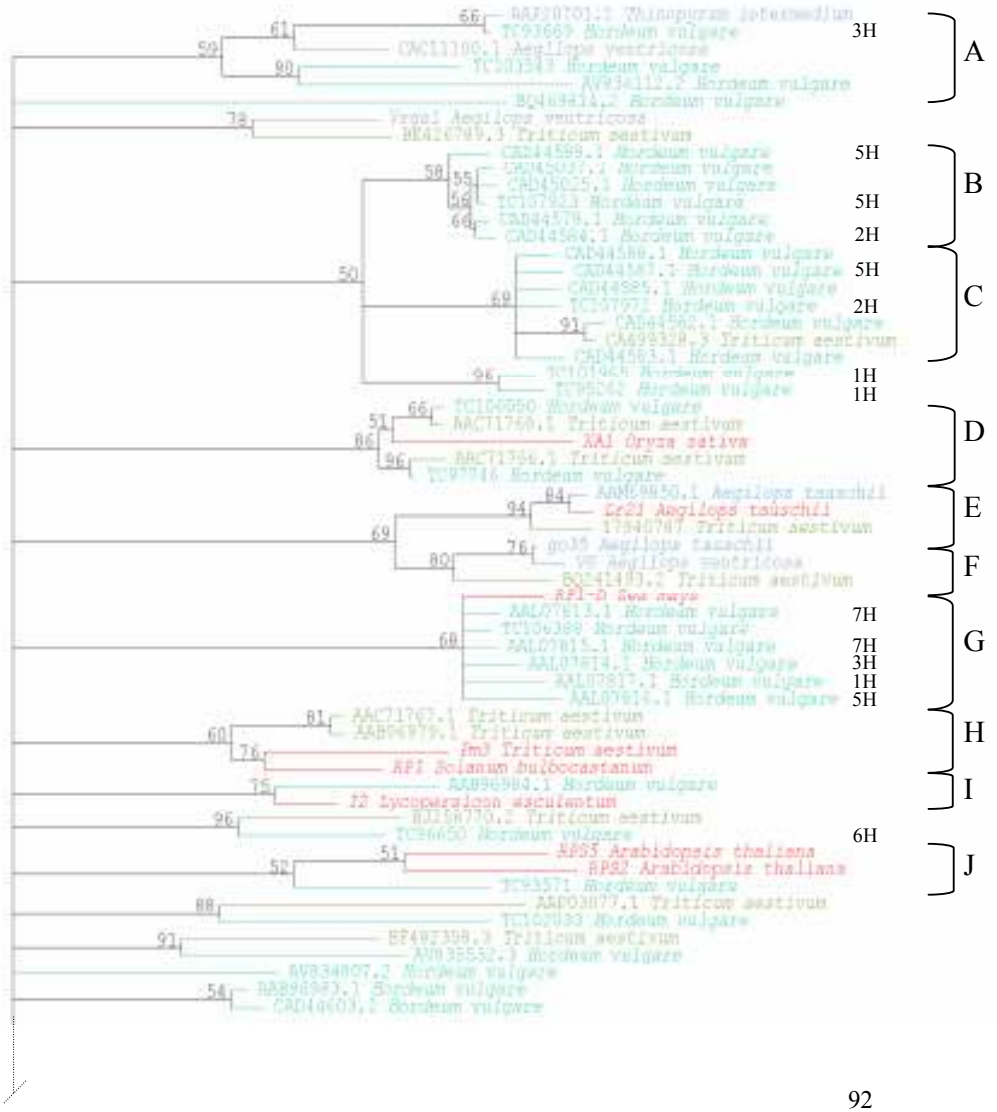
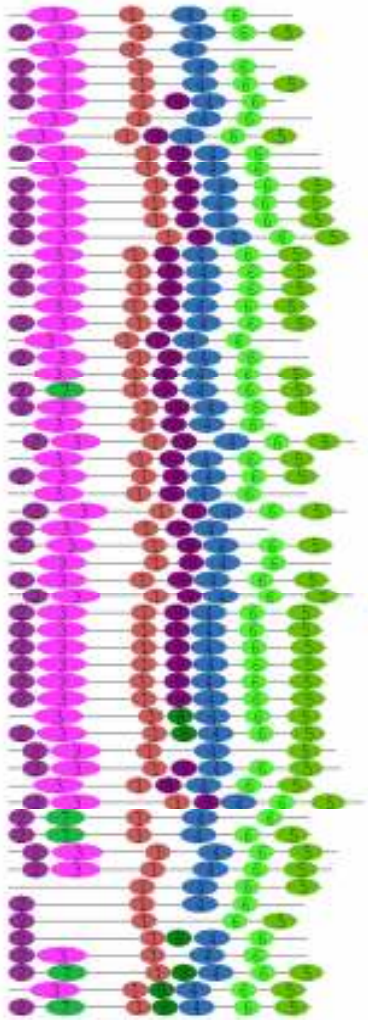


Figure 3.2 (continued) Multiple sequence alignment for translations of 17 sequences representative of the 155 sequences obtained by datamining. The closest *R* gene neighbours are included and the position of conserved motifs indicated. Major indels were removed and columns shaded for 50% amino acid conservation. The alignment was rendered using BioEdit ver 5.0.9 (Hall, 1999).


3H
A
5H
B
5H
2H
5H
C
2H
1H
1H
D
E
F
7H
G
7H
3H
1H
5H
H
I
6H
J



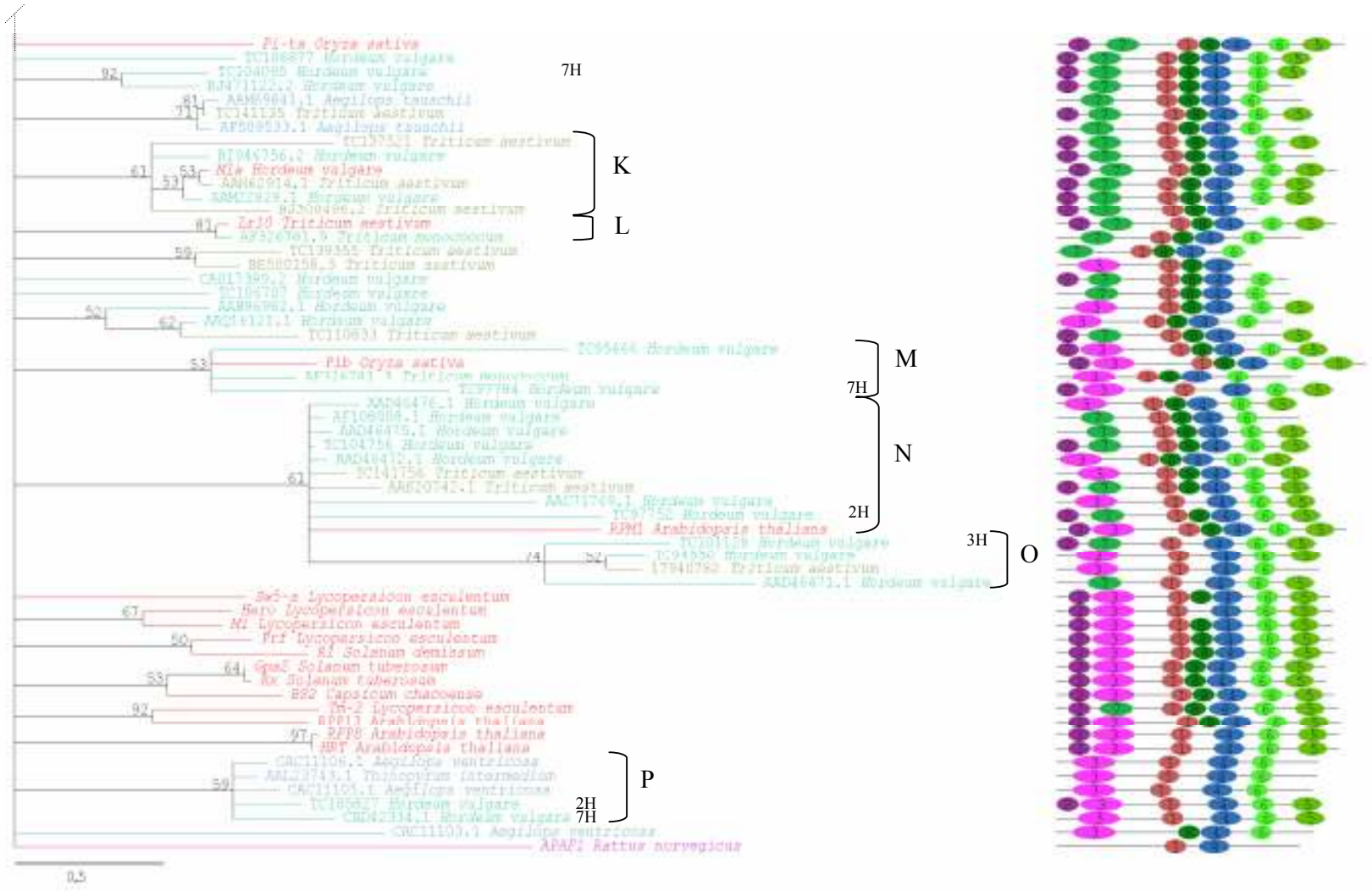
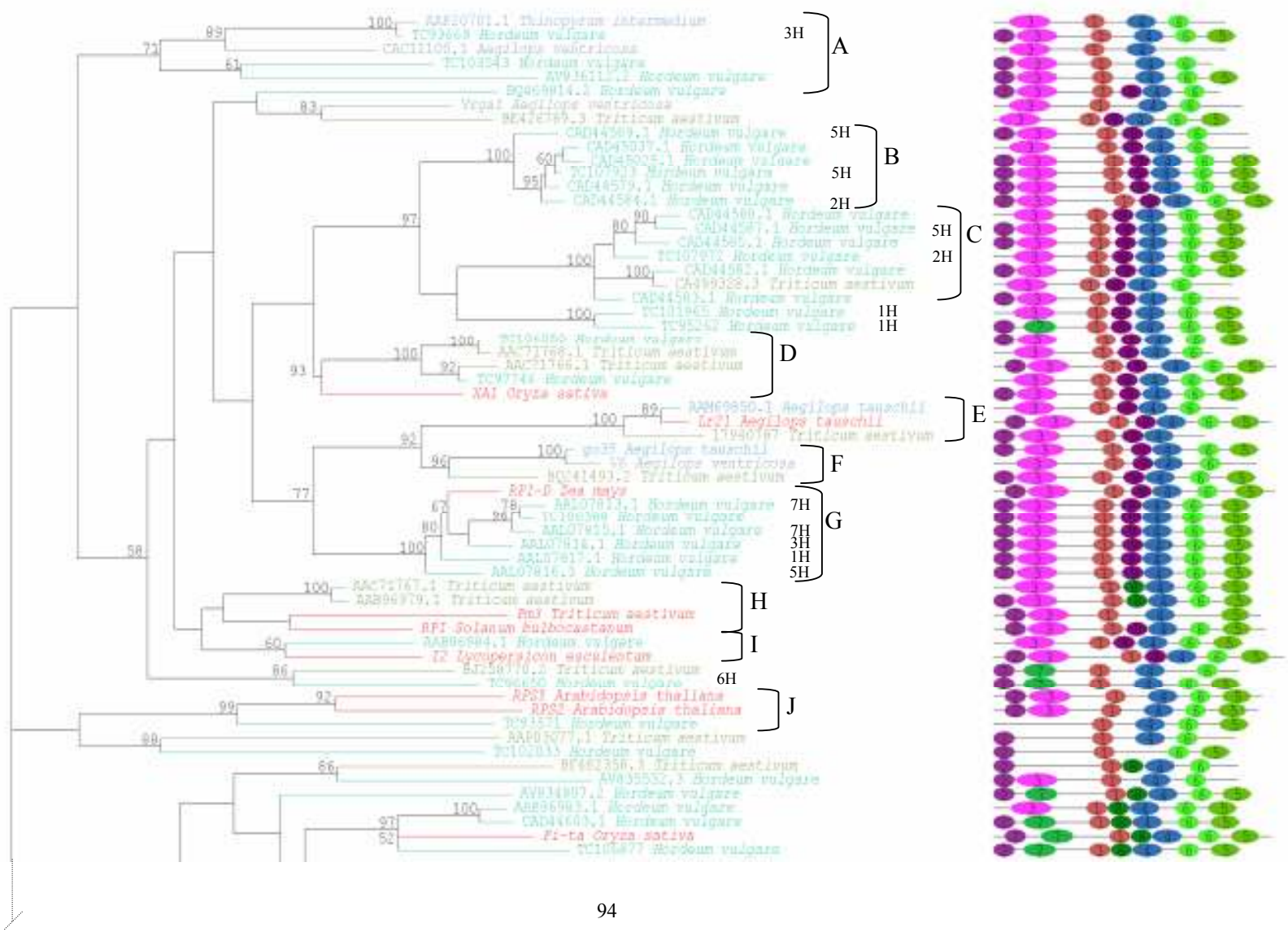


Figure 3.3 Maximum likelihood-based phylogeny reconstructed using the TREE-PUZZLE program (Scmidt *et al.*, 2002). The 118 amino-acid sequence alignment described in 3.3.3 was used. Motif structures are indicated opposite corresponding nodes (numbers correspond to motifs in Table 3.3) as detected by MEME (Bailey and Elkan, 1994). Major clade structures discussed are indicated with round braces, and barley chromosome positions indicated where known. The scale bar indicates amino-acid substitutions per site as computed by the ML implemented in TREE-PUZZLE.


3H
A
5H
5H
2H
B
5H
2H
C
1H
1H
D
E
F
7H
7H
3H
1H
5H
G
H
I
6H
J

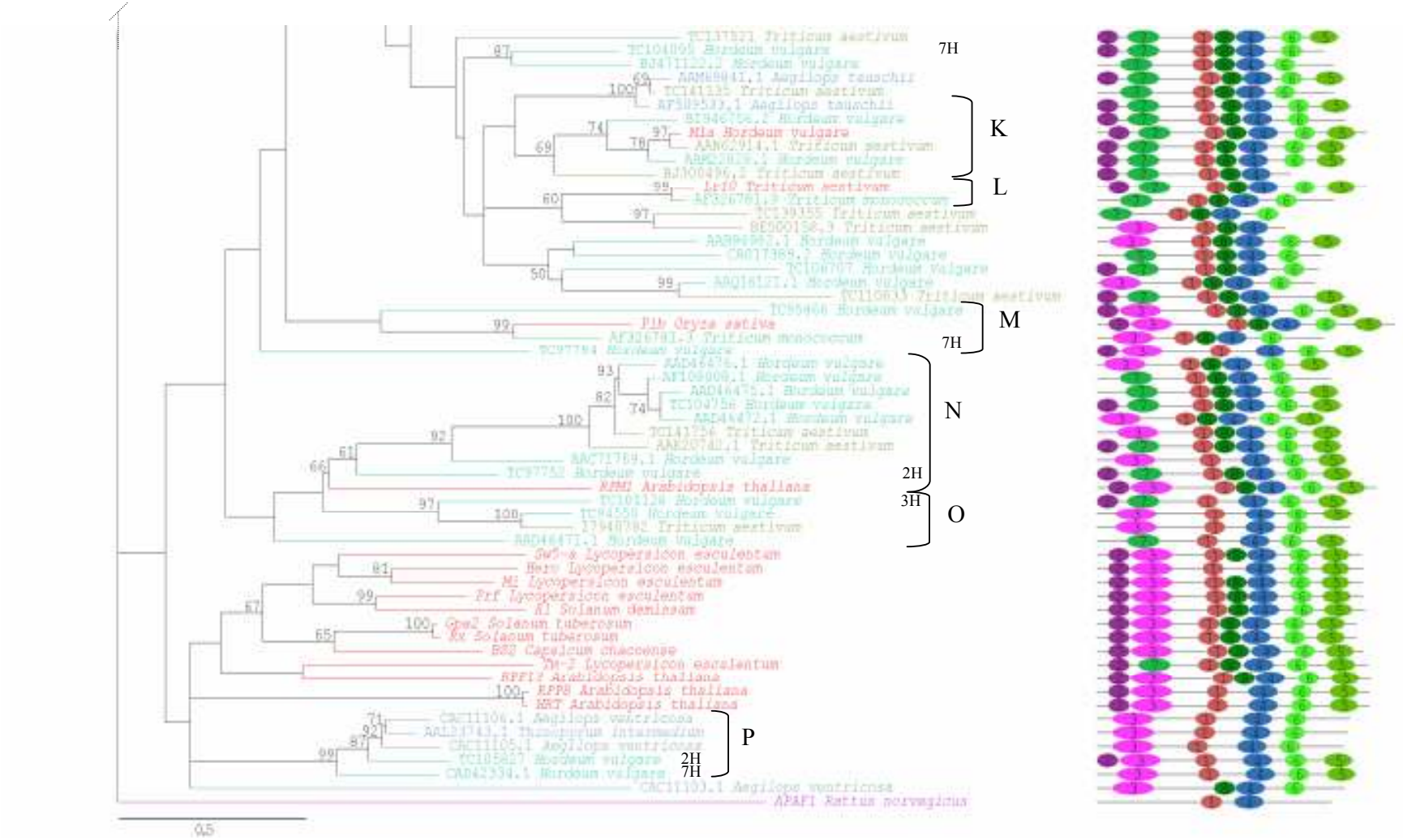
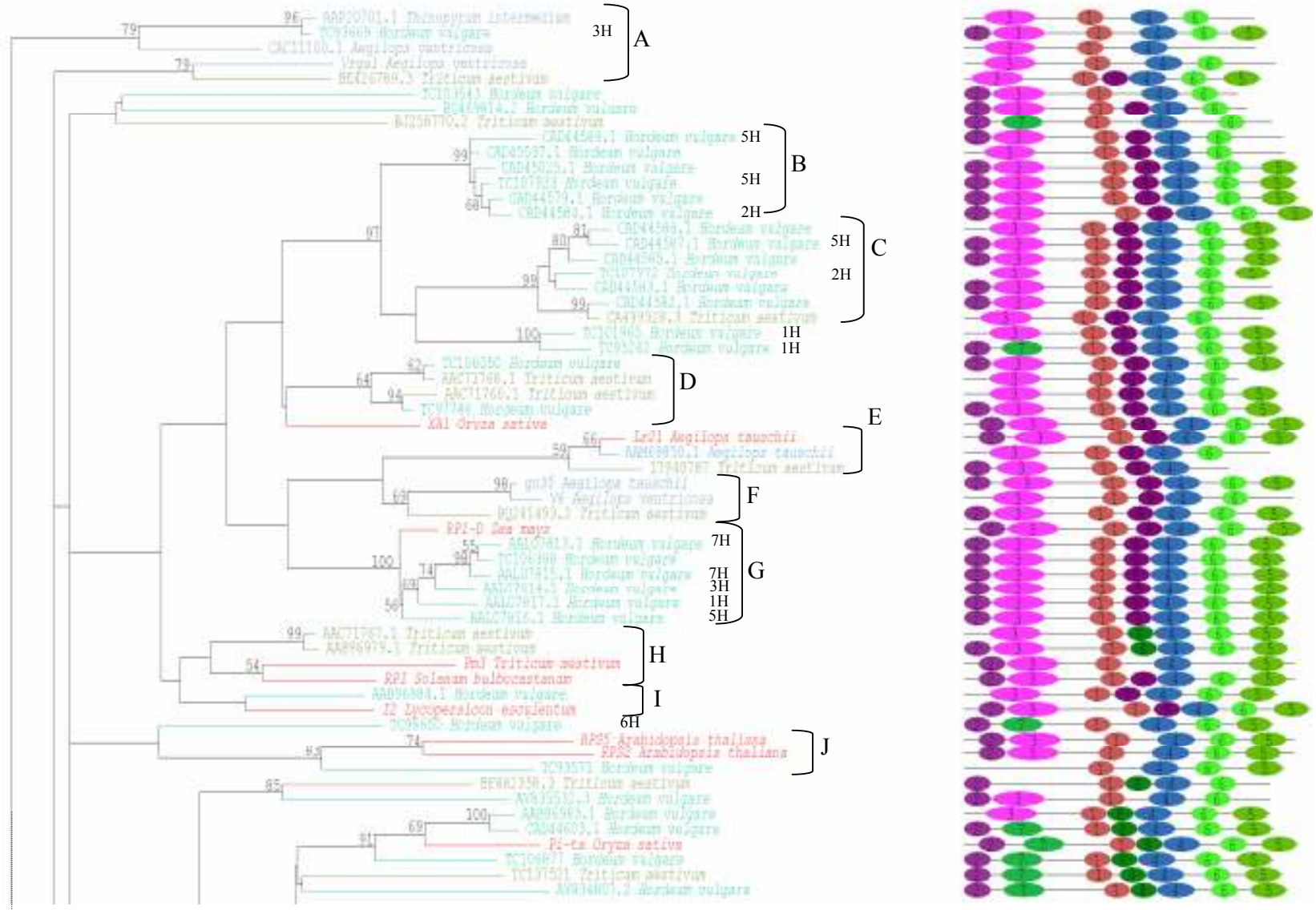


Figure 3.4 Bootstrapped distance-based phylogeny generated using the protdist program of the PHYLIP package. The 118 amino-acid sequence alignment described in 3.3.3 was used. Motif structures are indicated opposite corresponding nodes (numbers correspond to motifs in Table 3.3) as detected by MEME (Bailey and Elkan, 1994). Major clade structures discussed are indicated with round braces, and barley chromosome positions indicated where known. The scale bar indicates amino-acid substitutions per site as computed by the ML implemented in TREE-PUZZLE.


3H
A
5H
5H
2H
B
5H
2H
C
1H
1H
D
E
F
7H
7H
3H
1H
5H
G
H
I
6H
J

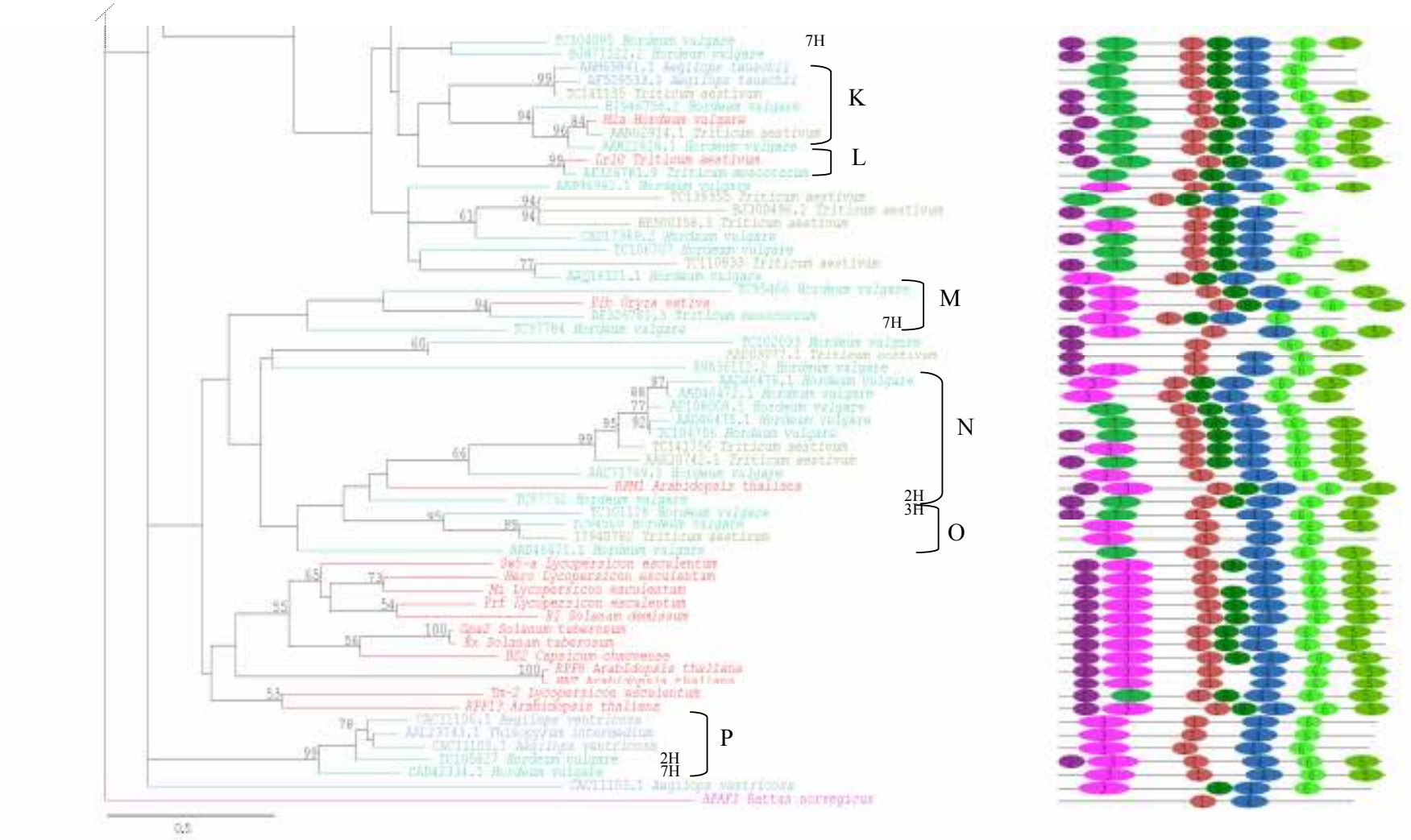


Figure 3.5 Maximum parsimony-based phylogeny reconstructed using the protpars program of the PHYLIP package (Felsenstein, 1989). The 118 amino-acid sequence alignment described in 3.3.3 was used. Motif structures are indicated opposite corresponding nodes (numbers correspond to motifs in Table 3.3) as detected by MEME (Bailey and Elkan, 1994). Major clade structures discussed are indicated with round braces, and barley chromosome positions indicated where known. The scale bar indicates amino-acid substitutions per site as computed by the ML implemented in TREE-PUZZLE.

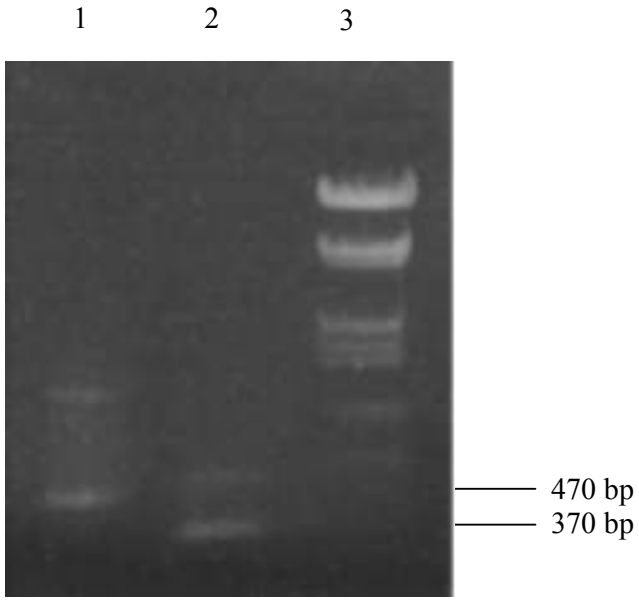


Figure 3.6 PCR bands obtained with the three specific primer sets indicated in Table 3.1. Bands are visualized on a one percent agarose gel, stained with ethidium bromide. Lane one: *go35* primer set, lane two: *KSU945* primer set and lane three: Lambda III size standard (Phage Lambda DNA restricted by *Eco*R*I* and *Hind*III). All amplifications were performed on *Triticum aestivum* (Tugela *Dn1*) genomic DNA.

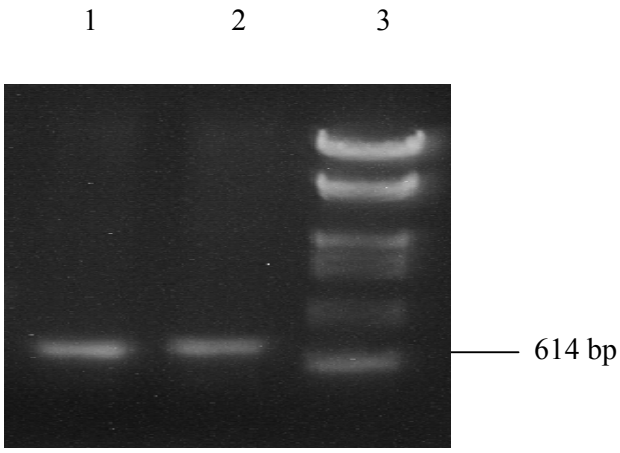


Figure 3.7 PCR bands amplified from the cloning cassette of pGEM-T Easy vector, using the *Sp6*- and *T7*-promoter targeted primer pair (the cloning cassette added 144bp). Colonies for two clones of the specific primer set *go35* are indicated on a one percent agarose gel, stained with ethidium bromide in lane one and two, and lane three contains the Lambda III size standard (Phage Lambda DNA restricted by *Eco*RI and *Hind*III). The cloned fragments indicated were amplified from *Aegilops speltoides* genomic DNA.

```
AY124651(A.tauschii)       D   AAGTTACTTTCTCCACTGAAGAAAGGAAAGATCCTAGTGACAACTCGAAGTAAATATGCACTACCGGATCTATGTCCTGGTGTGAGATATACTGCCATGC 100
        (A.tauschii)       D   AAGTTACTTTCTCCACTGAAGAAAGGAAAGATCCTAGTGACAACTCGAAGTAAATATGCACTACCGGATCTATGTCCTGGTGTGAGATATACTGCCATGC 100
AF320845(T.aestivum,X88)   D   AAGTTACTTTCTCCACTTAAGAAAGGAAAGATCCCAGTGACAACTCGAAGTAAATATGCACTACCGGATCTATGTCCTGGTGTGAGATATACTGCCATGC 100

         (A.speltoides)    B   AAGTTACTTTCTCCACTTAAGAAAGGAAAGATCCTAGTGACAACTCGAAGTAAATATGCACTACCGGATCTATGTCCTGGTGTGAGATATACTGCCATGC 100
         (T.aestivum, Dn1) B   AAGTTACTTTCTCCACTTAAGAAAGGAAAGATCCTAGTGACAACTCGAAGTAAATATGCACTACCGGATCTATGTCCTGGTGTGAGATATACTGCCATGC 100
         (T.turgidum)      B   AAGTTACTTTCTCCACTTAAGAAAGGAAAGATCCTAGTGACAACTCGAAGTAAATATGCACTACCGGATCTATGTCCTGGTGTGAGATATACTGCCATGC 100
AF052398 (T.aestivum, CS)  B   AAGTTACTTTCTCCACTTAAGAAAGGAAAGATCCTAGTGACAACTCGAAGTAAATATGCACTACCGGATCTATGTCCTGGTGTGAGATATACTGCCATGC 100
AY550176 (T.aestivum, X88) B   AAGTTACTTTCTCCACTTAAGAAAGGAAAGATCCTAGTGACAACTCGAAGTAAATATGCACTACCGGATCTATGTCCTGGTGTGAGATATACTGCCATGC 100

AY124651(A.tauschii)       D   CGATAACTGAGGTTGATGATACCGCCTTCTTTGAGTTGTTCATGCATTATGCCCTCGAAGATGGCCAAGATCAAAGCATGTTCCAGAACATTGGGGTTGA 200
        (A.tauschii)       D   CGATAACTGAGGTTGATGATACCGCCTTCTTTGAGTTGTTCATGCATTATGCCCTCGAAGATGGCCAAGATCAAAGCATGTTCCAGAACATTGGGGTTGA 200
AF320845(T.aestivum,X88)   D   CGATAACTGAGGTTGATGATACCGCCTTCTTTGAGTTGTTCATGCATTATGCCCTCGAAGATGGCCAAGATCAAAGCATGTTCCAGAACATTGGGGTTGA 200

         (A.speltoides)    B   CAATAACTGAGGTTGATGATACCGCCTTCTTTGAGCTGTTCATGCATTATGCCCTCGAAGATGGCCAAGATCAAAGCATGTTCCAGAACATTGGGGTTGA 200
         (T.aestivum, Dn1) B   CGATAACTGAGGTTGATGATACCGCCTTCTTTGAGCTGTTCATGCATTATGCCCTCGAAGATGGCCAAGATCAAAGCATGTTCCAGAACATTGGGGTTGA 200
         (T.turgidum)      B   CGATAACTGAGGTTGATGATACCGCCTTCTTTGAGCTGTTCATGCATTATGCCCTCGAAGATGGCCAAGATCAAAGCATGTTCCAGAACATTGGGGTTGA 200
AF052398 (T.aestivum, CS)  B   CGATAACTGAGGTTGATGATACCGCCTTCTTTGAGCTGTTCATGCATTATGCCCTCGAAGATGGCCAAGATCAAAGCATGTTCCAGAACATTGGGGTTGA 200
AY550176 (T.aestivum, X88) B   CGATAACTGAGGTTGATGATACCGCCTTCTTTGAGCTGTTCATGCATTATGCCCTCGAAGATGGCCAAGATCAAAGCATGTTCCAGAACATTGGGGTTGA 200

AY124651(A.tauschii)       D   GATTGCAAAAAAGCTGAAGGGGTCACCTTTAGCAGCTAGAACAGTGGGTGGAAATTTACGTCGACAGCAAGATGTTGACCATTGGAGAAGAGTCGGAGAT 300
        (A.tauschii)       D   GATTGCAAAAAAGCTGAAGGGGTCACCTTTAGCAGCTAGAACAGTGGGTGGAAATTTACGTCGACAGCAAGATGTTGACCATTGGAGAAGAGTCGGAGAT 300
AF320845(T.aestivum,X88)   D   GATTGCAAAAAAGCTGAAGGGGTCACCTTTAGCAGCTAGAACAGTGGGTGGAAATTTACGTCGACAGCAAGATGTTGACCATTGGAGAAGAGTCGGAGAT 300

         (A.speltoides)    B   GATTGCAAAAAAGCTGAAGGGGTCACCTCTAGCAGCTAGAACAGTGGGTGGAAATTTACGTCGACAGCAAGATGTTGACCATTGGAGAAGAGTCGGAGAT 300
         (T.aestivum, Dn1) B   GATTGCAAAAAAGCTGAAGGAGTCACCTCTAGCAGCTAGAACAGTGGGTGGAAATTTACGTCGACAGCAAGATGTTGACCATTGGAGAAGAGTCGGAGAT 300
         (T.turgidum)      B   GATTGCAAAAAAGCTGAAGGAGTCACCTCTAGCAGCTAGAACAGTGGGTGGAAATTTACGTCGACAGCAAGATGTTGACCATTGGAGAAGAGTCGGAGAT 300
AF052398 (T.aestivum, CS)  B   GATTGCAAAAAAGCTGAAGGGGTCACCTCTAGCAGCTAGAACAGTGGGTGGAAATTTACGTCGACAGCAAGATGTTGACCATTGGAGAAGAGTCGGAGAT 300
AY550176 (T.aestivum, X88) B   GATTGCAAAAAAGCTGAAGGGGTCACCTCTAGCAGCTAGAACAGTGGGTGGAAATTTACGTCGACAGCAAGATGTTGACCATTGGAGAAGAGTCGGAGAT 300

AY124651(A.tauschii)       D   CAAGACCTTTTCAAGGTATGGACGGGACCTCTGTGGTGGAGCTACTATCAGCTTGGTGAGCAGGCTAGGCGTTGCTTTGCTTACTGCAGTATTTTTCCTA 400
        (A.tauschii)       D   CAAGACCTTTTCAAGGTATGGACGGGACCTCTGTGGTGGAGCTACTATCAGCTTGGTGAGCAGGCTAGGCGTTGCTTTGCTTACTGCAGTATTTTTCCTA 400
AF320845(T.aestivum,X88)   D   CAAGACCTTTTCAAGGTATGGACGGGACCTCTGTGGTGGAGCTACTATCAGCTTGGTGAGCAGGCTAGGCGTTGCTTTGCTTACTGCAGTATTTTTCCTA 400

         (A.speltoides)    B   CAAGACCTTTTCAAGGTATGGACGGGACCTCTGTGGTGGAGCTACTATCAGCTAGGTGAGCAGGCTAGGCGTTGCTTTGCTTACTGCAGTATTTTTCCTA 400
         (T.aestivum, Dn1) B   CAAGACCTTTTCAAGGTATGGACGGGACCTCTGTGGTGGAGCTACTATCAGCTAGGTGAGCAGGCTAGGCGTTGCTTTGCTTACTGCAGTATTTTTCCTA 400
         (T.turgidum)      B   CAAGACCTTTTCAAGGTATGGACGGGACCTCTGTGGTGGAGCTACTATCAGCTAGGTGAGCAGGCTAGGCGTTGCTTTGCTTACTGCAGTATTTTTCCTA 400
AF052398 (T.aestivum, CS)  B   CAAGACCTTTTCAAGGTATGGACGGGACCTCTGTGGTGGAGCTACTATCAGCTAGGTGAGCAGGCTAGGCGTTGCTTTGCTTACTGCAGTATTTTTCCTA 400
AY550176 (T.aestivum, X88) B   CAAGACCTTTTCAAGGTATGGACGGGACCTCTGTGGTGGAGCTACTATCAGCTAGGTGAGCAGGCTAGGCGTTGCTTTGCTTACTGCAGTATTTTTCCTA 400

AY124651(A.tauschii)       D   GGAGACATCGCTTGTACCGT 420
        (A.tauschii)       D   GGAGACATCGCTTGTACCGT 420
AF320845(T.aestivum,X88)   D   GGAGACATCGCTTGTACCGT 420

         (A.speltoides)    B   GGAGACATCGCTTGTACCGT 420
         (T.aestivum, Dn1) B   GGAGACATCGCTTGTACCGT 420
         (T.turgidum)      B   GGAGACATCGCTTGTACCGT 420
AF052398 (T.aestivum, CS)  B   GGAGACATCGCTTGTACCGT 420
AY550176 (T.aestivum, X88) B   GGAGACATCGCTTGTACTGT 420
```

Figure 3.8 Multiple sequence alignment for nucleotide sequences obtained using the *go35* primer (Table 3.1). Cultivars are indicated: CS=Chinese Spring, X88 = Xinong88 and Dn1 = Tugela*Dn1*. Inferred genome source is indicated by B or D. Synonymous and nonsynonymous substitutions are indicated on white and black backgrounds respectively.

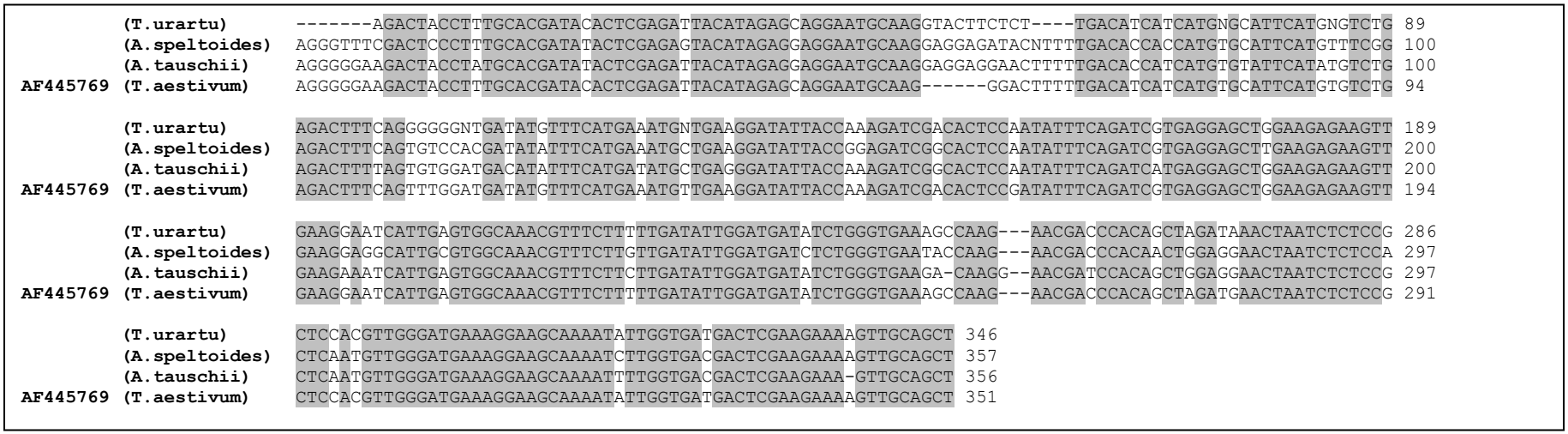


Figure 3.9 Multiple sequence alignment for nucleotide sequences obtained using the *KSU945* primer (Table 3.1).

```
AY124651(A.tauschii)       D   KLLSPLKKGKILVTTRSKYALPDLCPGVRYTAMPITEVDDTAFFELFMHYALEDGQDQSMFQNIGVEIAKKLKGSPLAARTVGGNLRRQQDVDHWRRVGD 100
        (A.tauschii)       D   KLLSPLKKGKILVTTRSKYALPDLCPGVRYTAMPITEVDDTAFFELFMHYALEDGQDQSMFQNIGVEIAKKLKGSPLAARTVGGNLRRQQDVDHWRRVGD 100
AF320845(T.aestivum,X88)   D   KLLSPLKKGKIPVTTRSKYALPDLCPGVRYTAMPITEVDDTAFFELFMHYALEDGQDQSMFQNIGVEIAKKLKGSPLAARTVGGNLRRQQDVDHWRRVGD 100

         (A.speltoides)    B   KLLSPLKKGKILVTTRSKYALPDLCPGVRYTAMPITEVDDTAFFELFMHYALEDGQDQSMFQNIGVEIAKKLKGSPLAARTVGGNLRRQQDVDHWRRVGD 100
         (T.aestivum, Dn1) B   KLLSPLKKGKILVTTRSKYALPDLCPGVRYTAMPITEVDDTAFFELFMHYALEDGQDQSMFQNIGVEIAKKLKESPLAARTVGGNLRRQQDVDHWRRVGD 100
         (T.turgidum)      B   KLLSPLKKGKILVTTRSKYALPDLCPGVRYTAMPITEVDDTAFFELFMHYALEDGQDQSMFQNIGVEIAKKLKESPLAARTVGGNLRRQQDVDHWRRVGD 100
AF052398 (T.aestivum, CS)  B   KLLSPLKKGKILVTTRSKYALPDLCPGVRYTAMPITEVDDTAFFELFMHYALEDGQDQSMFQNIGVEIAKKLKGSPLAARTVGGNLRRQQDVDHWRRVGD 100
AY550176 (T.aestivum, X88) B   KLLSPLKKGKILVTTRSKYALPDLCPGVRYTAMPITEVDDTAFFELFMHYALEDGQDQSMFQNIGVEIAKKLKGSPLAARTVGGNLRRQQDVDHWRRVGD 100

AY124651(A.tauschii)       D   QDLFKVWTGPLWWSYYQLGEQARRCFAYCSIFPRRHRLYR 140
        (A.tauschii)       D   QDLFKVWTGPLWWSYYQLGEQARRCFAYCSIFPRRHRLYR 140
AF320845(T.aestivum,X88)   D   QDLFKVWTGPLWWSYYQLGEQARRCFAYCSIFPRRHRLYR 140

         (A.speltoides)    B   QDLFKVWTGPLWWSYYQLGEQARRCFAYCSIFPRRHRLYR 140
         (T.aestivum, Dn1) B   QDLFKVWTGPLWWSYYQLGEQARRCFAYCSIFPRRHRLYR 140
         (T.turgidum)      B   QDLFKVWTGPLWWSYYQLGEQARRCFAYCSIFPRRHRLYR 140
AF052398 (T.aestivum, CS)  B   QDLFKVWTGPLWWSYYQLGEQARRCFAYCSIFPRRHRLYR 140
AY550176 (T.aestivum, X88) B   QDLFKVWTGPLWWSYYQLGEQARRCFAYCSIFPRRHRLYC 140
```

Figure 3.10 Multiple sequence alignment for translations of nucleotide sequences obtained using the *go35* primer (Table 3.1). Cultivars are indicated: CS=Chinese Spring, X88 = Xinong88 and Dn1 = Tugela*Dn1*. B or D indicates inferred genome source. Residues at the start of the alignment that were translated from nucleotides with ambiguous base calls were excluded.

```
          (T.urartu)       --XTTFARYTRDYIEQECKVLLXXDIIMXIHXSETFRGXDMFHEMXKDITKDRHSNISDREELEEKLKESLSGKRFFLILDDIWVKAKNDPQLDKLISPL 98
          (A.speltoides)   RVSTPFARYTREYIEEECKEEIXFDTTMCIHVSETFSVHDIFHEMLKDITGDRHSNISDREELEEKLKEALRGKRFLLILDDLWVNTKNDPQLEELISPL 100
          (A.tauschii)     RGKTTYARYTRDYIEEECKEEELFDTIMCIHMSETFSVDDIFHDMLRDITKDRHSNISDHEELEEKLKKSLSGKRFFLILDDIWVKXKNDPQLEELISPL 100
AF445769  (T.aestivum)     RGKTTFARYTRDYIEQECK--GLFDIIMCIHVSETFSLDDMFHEMLKDITKDRHSDISDREELEEKLKESLSGKRFFLILDDIWVKAKNDPQLDELISPL 98

          (T.urartu)       HVGMKGSKILVMTRRKVAA 117
          (A.speltoides)   NVGMKGSKILVTTRRKVAA 119
          (A.tauschii)     NVGMKGSKILVTTRRXVAA 119
AF445769  (T.aestivum)     HVGMKGSKILVMTRRKVAA 117
```

Figure 3.11 Multiple sequence alignment for translation of sequences obtained using the *KSU945* primer (Table 3.1).

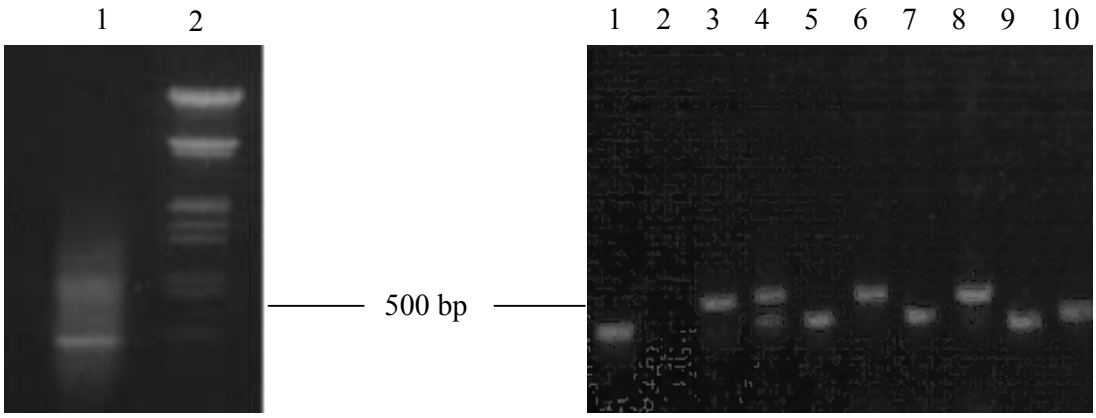


Figure 3.12 A.) Lane 1: PCR fragment smear obtained for the NB1 and NB2 (Yu *et al.*, 1996) primer combination using wheat (Tugela *Dn1*) genomic DNA. Lane 2: Lambda III molecular size marker. **B.)** Lanes 1-10: colony PCR of 10 clones. Bands were visualized on a 1% agarose gel stained with ethidium bromide.

Table 3.1 Specific primers used for amplification of the *go35* and *KSU945* genes.

Target sequence	Primer sequence	Target motif	Tm_{50}	Target length
go35	5′-cggatgttggtaaccaggag-3′	Kinase2 forward	60ºC	460 bp
	5′-tcacggtacaagcgatgtct-3′	MHD reverse	59ºC	
KSU945	5′-agggggaagactacctttgc-3′	P-loop forward	60ºC	351 bp
	5′-agctgcaacttttcttcgagtc-3′	RNBS-B reverse	60ºC	

Table 3.2 Degenerate primers NBS-F1 and NBS-R1, used for amplification of a section of the core NBS domain (Yu *et al.*, 1996). Primers are based on the consensus of the TNL *R* gene *N* (*Nicotiana glutinosa*) and the CNL *R* gene *RPS2* (*A. thaliana*).

Target sequence	Primer sequence	Target motif	Tm_{50}	Target length
NBS-F1	5′-GGAATGGGNGGNGTNGGNAARAC-3′	**P-loop:** **G**M**GGVGKT** (***N***) **G**P**GGVGKT** (***RPS2***)	63-72ºC	340bp
NBS-R1	5′-YCTAGTTGTRAYDATDAYYYTRC-3′	**RNBS-B:** SRII**TTR** (***N***) CKVMF**TTR** (***RPS2***)	50-67ºC	

Table 3.3 Summary of major motifs detected in NBS-LRR dataset using MEME (Bailey and Elkan, 1994). Residues identical to the *Triticeae* motifs in the *Arabidopsis* motifs are indicated in bold.

Motif	Consensus sequence		Annotation	E-value
1	LKGKRYLLVLDDVW	(*Triticeae* CNL)	Kinase2	$6.1e^{-886}$
	KRF**LLVLDD**I**W**	(*A thaliana* CNL)		
	RLD**K**KV**LI**V**LDDV**D	(*A thaliana* TNL)		
2	GMGGVGKTTLAQxVY	(*Triticeae* CNL)	P-loop	$8.5e^{-728}$
	VGYI**GMGGVGKTTLA**RQIF	(*A thaliana* CNL)		
	VGIW**G**PP**G**I**GKTTIA**RALF	(*A thaliana* TNL)		
3	DQRVKEHFDVRAWVCVSQxFDVxKLLKEI	(*Triticeae* CNL)	RNBS-A	$1.3e^{-750}$
	VKxG**FD**IVI**WV**V**VSQ**E**F**TLK**K**IQQD**I**LEK	(*A thaliana* CNL)		
	DYGM**K**L**H**LQEQFLSEILNQK**D**IKIxHLGV	(*A thaliana* TNL)		
4	NKGSRILVTTRIKDVAKxxCx	(*Triticeae* CNL)	RNBS-B	$1.3e^{-750}$
	N**G**CKV**L**F**TTR**SEE**V**C	(*A thaliana* CNL)		
	QLDALAGETxWFGP**GSRI**I**VTT**ED**K**	(*A thaliana* TNL)		
5	ELEEIGKKIAKKCGGLPLAA	(*Triticeae* CNL)	GLPL	$3.6e^{-552}$
	EV**AKKCGGLPLA**LKVI	(*A thaliana* CNL)		
	EV**A**xLA**GGLPL**GLKVL	(*A thaliana* TNL)		
6	LxEDDSWxLFxKRAF	(*Triticeae* CNL)	RNBS-C	$9.1e^{-419}$
	KVEC**L**TPEEA**WELF**QRKV	(*A thaliana* CNL)		
	NHIYEVxFPSxEEALQI**F**CQY**AF**GQNSPP	(*A thaliana* TNL)		
7	FDCRAFVSVSQNPDMKKLLKDIL	(*Triticeae* CNL)	RNBS-A alternative	$1.4e^{-132}$
	VKxG**FD**IVIW**VV**V**SQ**EFTL**KK**IQQ**DIL**EK	(*A thaliana* CNL)		
	DYGMKLHLQEQFL**S**EILNQ**K**DIKIxH**L**GV	(*A thaliana* TNL)		
8	DExLWDxIKCAFPDN	(*Triticeae* CNL)	N-terminal to Kinase2	$5.4e^{-122}$
9	DEDEWKKLLAAPLKKG	(*Triticeae* CNL)	N-terminal to Kinase2 alternative	$1.0e^{-111}$

Table 3.4 Summary of statistical support, number of *Triticeae* taxa and number of *R* gene members included for each clade indicated in Figure 3.3 to Figure 3.5.

Clade	Statistical support			Total members	*Triticeae* species	*R* gene members
	ML	Distance	Parsimony			
A	59%	72%	0%	5	3	0
B	58%	100%	52%	6	1	0
C	69%	27%	99%	7	2	0
D	86%	94%	34%	5	2	1[a]
E	94%	100%	59%	3	2	1
F	80%	96%	69%	3	3	0
G	60%	100%	100%	7	1	1
H	60%	46%	16%	4	2	2[a]
I	75%	61%	32%	2	1	1
J	52%	99%	83%	3	1	2[b]
K	61%	36%	28%	6	2	1
L	81%	99%	99%	2	2	1
M	53%	36%	28%	4	2	1
N	61%	61%	30%	10	2	1
O	74%	36%	36%	4	2	0
P	59%	92%	92%	5	3	0

[a] Contains one *R* gene member from a dicot species.

[b] Contains two *R* gene members from a dicot species.

Table 3.5 PCR band sizes and most significant BLAST hits to Genbank for sequenced bands. Percentage identity is indicated where homologues to the targeted genes were amplified.

Expected Fragment Sizes:	***go35* (Kinase2 to MHD)** **457 bp** *Triticum aestivum*		***KSU945* (P-loop to RNBS-B)** **351 bp** *Triticum aestivum*	
Triticum urartu (AA)	**PCR bands:** 1100 bp	**BLAST hits:** Not detected	**PCR bands:** **350 bp** 550bp	**BLAST hits:** *KSU945* (95% identity) Actin
***Aegilops speltoides* (BB)**	**PCR bands:** **530 bp** 890 bp 370 bp	**BLAST hits:** *go35* (98% identity)	**PCR bands:** **350 bp** 550 bp	**BLAST hits:** *KSU945* (91% identity)
***Aegilops tauschii* (DD)**	**PCR bands:** **450 bp**	**BLAST hits:** *go35* (100% identity)	**PCR bands:** 210 bp	**BLAST hits:** Not detected
Triticum turgidum (AABB)	**PCR bands:** **430 bp**	**BLAST hits:** *go35* (99% identity)	**PCR bands:** **440 bp**	**BLAST hits:** Chloroplast sequence
Triticum aestivum (AABBDD) Tugela Dn1	**PCR bands:** **470 bp**	**BLAST hits:** *go35* (99% identity)	**PCR bands:** **370 bp** 250 bp	**BLAST hits:** *KSU945* (89% identity)

3.5 References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990) Basic Local Alignment Search Tool. *Journal of Molecular Biology* 215:403-410.

Arumuganathan, K., and Earle, E.D. (1991) Nuclear DNA content of some important plant species. *Plant Molecular Biology Reports* 9:208–218.

Bai, J., Pennill, L.A., Ning, J., Lee, S.W., Ramalingam, J., Webb, C.A., Zhao, B., Sun, Q., Nelson, J.C., Leach, J.E., and Hulbert, S.H. (2002) Diversity in Nucleotide-bindin-Leucine-rich-repeat genes in Cereals. *Genome Research* 12:1871-1884.

Bailey, T.L., and Elkan, C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology* 28-36.

Bailey, T.L., and Gribskov, M. (1998) Combining evidence using p-values: Application to sequence homology searches. *Bioinformatics* 14:48-54.

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

Baxevanis, A.D., and Ouellette, B.F.F. (2001*) Bioinformatics: a practical guide to the analysis of genes and proteins*. 2nd Edition John Wiley & Sons, Inc.

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* 4:391-399.

Bittner-Eddy, P.D., Crute, I.R., Holub, E.B., and Beynon, J.L. (2000) *RPP13* is a simple locus in *Arabidopsis thaliana* for alleles that specify downy mildew resistance to different avirulence determinants in *Peronospora parasitica. Plant Journal* 21:2:177-188.

Blanc, G., and Wolfe, K.H. (2004) Functional Divergence of Duplicate Genes Formed by Polyploidy during *Arabidopsis* Evolution. *The Plant Cell* 16:1679-1691.

Caicedo, A.L., Schaal, B.A., and Kunkel, B.N. (1999) Diversity and molecular evolution of the *RPS2* resistance gene in Arabidopsis thaliana. *Proceedings of the National Academy of Sciences USA* 96:302-306.

Cannon, S.B., and Young, N.D. (2003) The genomic architecture of NBS-LRRs. *Plant Microbe Interactions* 2003 6th edition Stacey, G., Keen, N.T., eds. APS press, St.Paul, M.N.

Cannon, S.B., Zhu, H., Baumgarten, A.M., Spangler, R., May, G., Cook, D.R., and Young, N.D. (2002) Diversity, distribution, and ancient taxonomic relationships within the TIR and non-TIR NBS-LRR resistance gene subfamilies. *Journal of Molecular Evolution* 54:548-562.

Close, T.J., Wanamaker, S.I., Caldo, R.A., Turner, S.M., Daniel, A.A., Dickerson, J.A., Wing, R.A., Muehlbauer, G.J., Kleinhofs, A., and Wise, R.P. (2004) A New Resource for Cereal Genomics: 22K Barley GeneChip Comes of Age. *Plant Physiology* 134:960–968.

Cobb, B.D., and Clarkson, J.M. (1994) A simple procedure for optimizing the polymerase chain reaction (PCR) using modified Taguchi methods. *Nucleic Acids Research* 22:18:3801-3805.

Cortese, M.R., Fanelli, E., and De Giorgi, C. (2003) Characterization of nematode resistance gene analogs in tetraploid wheat. *Plant Science* 164:1:71-75.

Dayhoff, M.O., Barker, W.C. and Hunt, L.T. (1983) Establishing Homologies in Protein Sequences. *Methods in Enzymology* 91:524-545.

De Majnik, J., Ogbonnaya, F.C., Moullet, O., and Lagudah, E. (2003) The *Cre1* and *Cre3* Nematode resistance genes are located at homeologous loci in the wheat genome. *Molecular Plant Microbe Interactions* 16:12:1129-1134.

Delorenzi, M., and Speed, T. (2002) An HMM model for coiled-coil domains and a comparison with PSSM-based predictions. *Bioinformatics* 18:4:617-625.

Drouin, G., Prat, F., Ell, M., and Clarke, P.G.D (1999) Detecting and characterizing gene conversions between multigene family members. *Molecular Biological Evolution* 16:1369-1390.

Durbin, R., Eddy, S., Krogh, A. and Mitchison, G. (2000) *Biological sequence analysis*. Cambridge University Press.

Eddy, S.R. (1995) Multiple Alignment using Hidden Markov Models. *Proceedings of Intelligent Systems for Molecular Biology* AAAI Press 0:114-120.

Eddy, S.R. (1998) Profile Hidden Markov Models. *Bioinformatics* 14:755-763.

Edwards, K., Johnstone, C., and Thompson, C. (1991) A simple and rapid method for the preparation of plant genomic DNA for PCR analysis. *Nucleic Acids Research* 19:6:1349.

Ellis, J.G., Lawrence, G.J., Luck, J.E., and Dodds, P.N. (1999) Identification of regions in alleles of the flax rust resistance gene L that determine differences in gene-for-gene specificity. *Plant Cell* 11:495-506.

Felsenstein, J. (1989) PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics* 5:164-166.

Feuillet, C., Travella, S., Stein, N., Albar, L., Nublat, A., and Keller, B. (2003) Map-based isolation of the leaf rust disease resistance gene *Lr10* from the hexaploid wheat (*Triticum aestivum* L.) genome. *Proceedings of the National Acadamy of Sciences USA* 100:25:15253-15258.

Flajnik, M.F., and Kasahara, M. (2001) Comparative genomics of the MHC: glimpses into the evolution of the adaptive immune system. *Immunity* 15:3:351-62.

Flor, H.H. (1971) Current status of the gene-for gene concept. *Annual Review of Phytopathology* 9:275-298.

Frick, M.M., Huel, R., NykiForuk, C.L., Conner, R.L., Kusyk, A., and Laroche, A. (1998) Molecular characterization of a wheat stripe rust resistance gene in Moro wheat. *Proceedings of the 9th International Wheat Genetics Symposium* Slinkard AE (ed) 3:181-182.

Gale, M.D., and Devos, K.M. (1998) Comparative genetics in the grasses. *Proceedings of the National Academy of Sciences USA* 95:5:1971-1974.

Gill, B.S., Appels, R., Botha-Oberholster, A-M., Buell, C.R., Bennetzen, J.L., Chalhoub, B., Chumley, F., Dvorak, J., Iwanaga, M., Keller, B., Li W., McCombie, W.R., Ogihara, Y., Quetier, F., and Sasaki, T. (2004). A workshop report on wheat genome sequencing: International Genome Research on Wheat Consortium. *Genetics* 168:1087-1096.

Grant, M.R., Godiard, L., Straube, E., Ashfield, T., Lewald, J., Sattler, A., Innes, R.W., and Dangl, J.L. (1995) Structure of the *Arabidopsis RPM1* gene enabling dual specificity disease resistance. *Science* 269:5225:843-846.

Güssow, D., and Clackson, T. (1989) Direct clone characterization from plaques and colonies by the polymerase chain reaction. *Nucleic Acids Research* 17:4000.

Guttman, D.S., and Dykhuizen, D.E. (1994) Clonal divergence in *Escherichia coli* as a result of recombination, not mutation. *Science* 266:1380-1383.

Hall, T.A. (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Serials* 41:95-98.

Hammond-Kosack, K.E., and Jones, J.D.G. (1997) Plant disease resistance genes. *Annual Review of Plant Physiology and Plant Molecular Biology* 48:575-607.

Henikoff, S., and Henikoff, J.G. (1992) Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences USA* 89:10915-10919.

Huang, L., Brooks, S.A., Li, W., Fellers, J.P., Trick, H.N., and Gill, B.S. (2003) Map-Based Cloning of Leaf Rust Resistance gene *Lr21* from the large and polyploid genome of Bread Wheat. *Genetics* 164:655-664.

Huang, L., and Gill, B.S. (2001) An RGA-like marker detects all known *Lr21* leaf rust resistance gene family members in *Aegilops tauschii* and wheat. *Theoretical and Applied Genetics* 103:1007-1013.

Huang, S., Sirikhachornkit, A., Faris, J.D., Su, X., Gill, B.S., Haselkorn, R. and Gornicki, P. (2002a) Phylogenetic analysis of the acetyl-CoA carboxylase and 3-phosphoglycerate kinase loci in wheat and other grasses. *Plant Molecular Biology* 48:5-6:805-20.

Huang, S., Sirikhachornkit, A., Su X., Faris, J.D., Gill, B.S., Haselkorn, R., and Gornicki, P. (2002b) Genes encoding plastid acetyl-CoA carboxylase and 3-phosphoglycerate kinase of the *Triticum/Aegilops* complex and the evolutionary history of polyploid wheat. *Proceedings of the National Academy of Sciences USA* 99: 8133-8138.

Irwin, D.M., and Wilson, A.C. (1990) Concerted evolution of ruminant stomach lysozymes. Characterization of lysozyme cDNA clones from sheep and deer. *Journal Biological Chemistry* 265:9:4944-4952.

Kimber G., and Sears, E.R. (1987) Evolution in the genus *Triticum* and the origin of cultivated wheat. *Wheat and Wheat Improvement* Heyne, E.G. (ed.) 2nd ed *Agronomy*. ASA, CSSA, SSSA. 13:154-163.

Karlin, S., and Altschul, S.F. (1993) Applications and statistics for multiple high-scoring segments in molecular sequences. *Proceedings of the National Acadamy of Sciences USA*. 90:12:5873-5877.

Kimura, M. (1983) *The neutral theory of molecular evolution*. Cambridge University Press, Cambridge, England.

Kimura, M. (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution* 16:111–120.

Kumar S., Tamura K. and Nei M. (2004) MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. *Briefings in Bioinformatics* 5:2 (In press).

Lagudah, E.S., Moullet, O., and Appels, R. (1997) Map based cloning of a gene sequence encoding a nucleotide binding domain and a leucine-rich repeat region at the *Cre3* nematode resistance locus of wheat. *Genome* 40: 659-665.

Lee, Y.H., Ota, T., and Vacquier, V.D. (1995) Positive selection is a general phenomenon in the evolution of abalone sperm lysin. *Molecular Biological Evolution* 12:231-238.

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 Acadamy of Sciences* 95:370-375.

Li, W.-H. (1993) Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *Journal of Molecular Evolution* 36:1:96-9.

Li, W.-H., Wu, C.-I., and Luo, C.-C. (1985) A new method for estimating synonymous and non-synonymous rates of nucleotide substitutions considering the relative likelihood of nucleotide and codon changes. *Molecular Biological Evolution* 2:150–174.

Li, W., Zhang, P., Fellers, J.P., Friebe, B., and Gill, B.S. (2004) Sequence composition, organization, and evolution of the core *Triticeae* genome. *Plant Journal* 40:4:500-511.

Madsen, L.H., Colins, N.C., Rakwalska, M., Backes, G., Sandal, N., Krusell, L., Jensen, J., Waterman, E.H., Jahoor, A., Ayliffe, M., Pryor, A.J., Langridge, P., Schulze-Lefert, P., and Stougaard, J. (2003) Barley disease resistance gene analogues of the NBS-LRR class: identification and mapping. *Molecular Genetics and Genomics* 269:150-161.

Maleki, L., Faris, J.D., Bowden, R.L., Gill, B.S., and Fellers, J.P. (2003) Physical and genetic mapping of wheat kinase analogs and NBS-LRR resistance gene analogues. *Crop Science* 43:660-670.

Martin, G.B., Bogdanove, A.J., and Sessa, G. (2003) Understanding the functions of plant disease resistance proteins. *Annual Review of Plant Biology* 54:23-61.

Maynard Smith, J., and Smith, N.H. (1998) Detecting recombination from gene trees. *Molecular Biological Evolution* 15:590-599.

Meyers, B.C., Dicerkmann, A.W., Michelmore, R.W., Sivaramakrishnan, S., Sorbal, B.W., and Young, N.D. (1999) Plant Disease resistance genes encode members of an ancient and diverse protein family within the nucleotide-binding superfamily. *Plant Journal* 20:317-322.

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* 15:809-834.

Meyers, B.C., Morgante, M., and Michelmore, R.W. (2002) TIR-X and TIR-NBS proteins: two new families related to disease resistance TIR-NBS-LRR proteins encoded in *Arabidopsis* and other plant genomes. *The Plant Journal* 32:77-92.

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* 8:1113–1130.

Mindrios, M., Katagiri, F., Yu, G-L., and Ausabel, F.M. (1994) The *A.thaliana* disease resistance gene *rps2* encodes a protein containing a nucleotide-binding site and Leucine-Rich repeats. *Cell* 78:1089-1099.

Miyata, T., and Yasunaga, T. (1980) Molecular evolution of mRNA: a method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its applications. *Journal of Molecular Evolution* 16:23-36.

Mondragón-Palomino, M., Meyers, B.C., Michelmore, R.W., and Gaut, B.S. (2002) Patterns of Positive Selection in the Complete NBS-LRR Gene Family of *Arabidopsis thaliana*. *Genome Research* 12:9:1305-1315.

Monosi, B., Wisser, R.J., Pennill, L., and Hulbert, S.H. (2004) Full-genome analysis of resistance gene homologues in rice. *Theoretical and Applied* Gene*tics* 109:1434-1447.

Morgante, M., and Michelmore, R.W. (2002) TIR-X and TIR-NBS proteins: two new families related to disease resistance. TIR-NBS-LRR proteins encoded in *Arabidopsis* and other plant genomes. *The Plant Journal* 32:77-92.

Munzner, T., Guimbretiere, F., Tasiran, S., Zhang, L., and Zhou, Y. (2003) TreeJuxtaposer: Scalable Tree Comparison using Focus+Context with Guaranteed Visibility. *ACM Transactions on Graphics* 22:3:453-462.

Needleman, S.B., and Wunsch, C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology* 48:3:443-53.

Nei, M., Gu, X., and Sitnikova, T. (1997) Evolution by the birth-and-death process in multigene families of the vertebrate immune system. *Proceedings of the National Academy of Sciences USA* 94:15:7799-806.

Nei, M., and Kumar, S. (2000) *Molecular evolution and phylogenetics* Oxford University Press.

Niimura, Y., and Nei, M. (2003) Evolution of olfactory receptor genes in the human genome. *Proceedings of the National Acadamy of Sciences USA* 100:21:12235-12240.

Notredame, C., Higgins, D.G., and Heringa, J. (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology* 302:205-217.

Otto, S.P., and Yong, P. (2002) The evolution of gene duplicates. *Advanced* Gene*tics* 46:451-83.

Ozkan, H., Levy, A.A., and Feldman, M. (2001) Allopolyploidy-induced rapid genome evolution in the wheat (*Aegilops-Triticum*) group. *Plant Cell* 13:8:1735-47.

Pamilo, P., and Bianchi, N.O. (1993) Evolution of the *Zfx* and *Zfy* genes: rates and interdependence between the genes. *Molecular Biology and Evolution* 10:271-281.

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* 155:309-322.

Pan, Q.L., Wendel, J., and Fluhr, R. (2000b) Divergent evolutuion of plant NBS-LRR resistance gene homologues in dicot and cereal genomes. *Journal Molecular Evolution* 50:203-213.

Rabiner, L.R. (1989) A Tutorial on Hidden Markov Models and selected applications in speech recognition. *Proceedings of the Institute of Electrical and Electronic Engineers* 77:2:257-286.

Rostoks, N., Zale, J., Soule, J., Brueggeman, R., Druka, A., Kudrna, D., Steffenson, B., and Kleinhofs, A. (2002) A barley gene family homologous to the maize rust resistance gene *Rp1-D. Theoretical and Applied Genetics* 104:1298-1306.

Rozen, S., and Skaletsky, H.J. (2000) Primer3 on the WWW for general users and for biologist programmers. Krawetz S., Misener S. (eds) Bioinformatics Methods and Protocols: *Methods in Molecular Biology* 365-386.

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.

Scherrer, B., Keller, B., and Feuillet, C. (2002) Two haplotypes of resistance gene analogs have been conserved during evolution at the leaf rust resistance locus *Lr10* in wild and cultivated wheat. *Functional and Integrative Genomics* 2:40-50.

Schmidt, H.A., Strimmer, K., Vingron, M., and von Haeseler, A. (2002) TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* 18:3:502-504.

Seah, S., Sivasithamparam, K., Karakousis, K., and Lagudah, E.S. (1998) Cloning and characterization of a family of disease resistance gene analogs from wheat and barley. *Theoretical and Applied Genetics* 97:937-945.

Shen, K.A., Chin, D.B., Arroyo-Garcia, R., Ochoa, O.E., Lavelle, D.O., Wroblewski, T., Meyers, B.C., and Michelmore, R.W. (2002) *Dm3* is one member of a large constitutively expressed family of nucleotide binding site-leucine-rich repeat encoding genes. *Molecular Plant Microbe Interactactions* 15:3:251-261.

Sneath, P. (1998) The effect of evenly spaced constant sites on the distribution of the random division of a molecular sequence. *Bioinformatics* 14:608-616.

Smith, T.F., and Waterman, M.S. (1981) Identification of Common Molecular Subsequences. *Journal of Molecular Biology* 147:195-197.

Spielmeyer, W., Huang, L., Bariana, H., Laroche, A., Gill, B.S., and Lagudah, E.S. (2000) NBS-LRR sequence family is associated with leaf and stripe rust resistance on the end of homeologous group 1S of wheat. *Theoretical Applied Genetics* 101:1139-1144.

Spielmeyer, W., Robertson, M., Collins, N., Leister, D., Schulze-Lefert, D. Seah, S., Moullet, O., and Lagudah, E.S. (1998) A superfamily of disease resistance gene analogs is located on all homeologus chromosome groups of wheat (*Triticum aestivum*). *Genome* 41:782-788.

The Arabidopsis Genome Initiative (2000) Analysis of the genome of the flowering plant *Arabidopsis thaliana*. *Nature* 408:796-815.

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

Tian, D., Araki, H., Stahl, E., Bergelson, J., and Kreitman, M. (2002) Signature of balancing selection in *Arabidopsis*. *Proceedings of the National Academy of Sciences USA* 99:17:11525-30.

Van Der Vossen, E., Sikkema, A., Hekkert, B.L., Gros, J., Stevens, P., Muskens, M., Wouters, D., Pereira, A., Stiekema, W., and Allefs, S. (2003) An ancient *R* gene from the wild potato species *Solanum bulbocastanum* confers broad-spectrum resistance to *Phytophthora* infestans in cultivated potato and tomato. *Plant Journal* 36:6:867-882.

Warren, R.F., Henk, A., Mowery, P., Holub, E., and Innes, R.W. (1998) A mutation within the Leucine-Rich Repeat Domain of the *Arabidopsis* disease resistance gene *RPS5* partially supresses multiple bacterial and Downey Mildew resistance genes. *The Plant Cell* 10:1439-1452.

Wei, F., Wing, R.A., and Wise R.P. (2002). Genome Dynamics and Evolution of the *Mla* (Powdery Mildew) Resistance Locus in Barley. *Plant Cell* 14:8:1903–1917.

Yahiaoui, N., Srichumpa, P., Dudler, R., and Keller, B. (2004) Genome analysis at different ploidy levels allows cloning of the powdery mildew resistance gene *Pm3b* from hexaploid wheat. *Plant Journal* 37:4:528-538.

Yu, J., Hu, S., Wang, J., Wong, G.K., Li, S., Liu, B., Deng, Y., Dai, L., Zhou, Y., Zhang, X., Cao, M., Liu, J., Sun, J., Tang, J., Chen, Y., Huang, X., Lin, W., Ye, C., Tong, W., Cong, L., Geng, J., Han, Y., Li, L., Li, W., Hu, G., Huang, X., Li, W., Li, J., Liu, Z., Li, L., Liu, J., Qi, Q., Liu, J., Li, L., Li, T., Wang, X., Lu, H., Wu, T., Zhu, M., Ni, P., Han, H., Dong, W., Ren, X., Feng, X., Cui, P., Li, X., Wang, H., Xu, X., Zhai, W., Xu, Z., Zhang, J., He, S., Zhang, J., Xu, J., Zhang, K., Zheng, X., Dong, J., Zeng, W., Tao, L., Ye, J., Tan, J., Ren, X., Chen, X., He, J., Liu, D., Tian, W., Tian, C., Xia, H., Bao, Q., Li, G., Gao, H., Cao, T., Wang, J., Zhao, W., Li, P., Chen, W., Wang, X., Zhang, Y., Hu, J., Wang, J., Liu, S., Yang, J., Zhang, G., Xiong, Y., Li, Z., Mao, L., Zhou, C., Zhu, Z., Chen, R., Hao, B., Zheng, W., Chen, S., Guo, W., Li, G., Liu, S., Tao, M., Wang, J., Zhu, L., Yuan, L., and Yang, H. (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science* 296:5565:79-92.

Yu, Y.G., Buss, G., and Maroof, S.A. (1996) Isolation of a superfamily of candidate disease-resistance genes in soybean based on a conserved nucleotide-binding site. *Proceedings of the National Acadamy of Sciences USA* 93:11751-11756.

Zhang, L., Pond, S.K., and Gaut, B.S. (2001) A Survey of the Molecular Evolutionary Dynamics of Twenty-Five Multigene Families from Four Grass Taxa. *Journal of Molecular Evolution* 52:144-156.

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Tracking nucleotide-binding-site-leucine-rich-repeat resistance gene analogues in the wheat genome complex

Franco B du Preez

Supervised by **Prof A-M. Botha-Oberholster** and **Dr A.A. Myburg**

Submitted in fulfillment of the requirements for the degree ***Magister Scientiae***

Department of Genetics

University of Pretoria

Summary

Investigations into plant-pathogen interactions have provided us with several models underlying the genetic basis of host resistance in plants. In the past decade, tens of resistance genes have been isolated from numerous crop and model plant species and these form a few distinct classes when classified by domain structure, the majority being nucleotide-binding-site-leucine-rich-repeat (NBS-LRR) genes. The NBS-LRR family consists of two sub-families based on the N-terminal domain: the coiled-coil (CC) NBS-LRRs and the Toll Interleukin Receptor homology domain (TIR) NBS-LRRs. The potential of these genes for future and current agricultural breeding programs has driven a large number of studies exploring the members of these gene families in the genomes of a variety of crop species.

In the present study I focused on the NBS-LRR family in the allohexaploid wheat genome and obtained a comprehensive set of *Triticeae* NBS-LRR homologues using a combination of data-mining approaches. As starting point I detected conserved motifs in the dataset, finding all six previously characterized in the core-NBS domain of other plant NBS-LRRs. Phylogenetic analysis was performed to study relationships between the *Triticeae* NBS-LRR family and the 25 CC-NBS-LRR (CNL) *R* genes identified to date. I found the *Triticeae* CNL family to be highly divergent, containing ancient clade lineages, as seen in all angiosperm

taxa previously studied, and found a number of "ancient" dicotyl *R* genes grouped with *Triticeae* clades.

The evolution of recent NBS-LRR gene duplications in the *Triticeae* was studied at the hand of two modes of duplication - firstly individual gene duplications yielding paralogous loci and secondly gene duplication by allopolyploidy. Current models of NBS-LRR family evolution predict that functional divergence occurs after gene duplication. An alternative is that divergence takes place at allele level, followed by a locus duplication that fixes heterozygosity in a single haplotype by unequal recombination. I investigated this hypothesis by studying the evolution of gene duplicates in two different contexts – paralogous duplications in the diploid barley genome and homeologous duplications in the allohexaploid genome of wheat.

Nonsynonymous to synonymous substitution rate ratios were estimated for paralogous gene duplications in three recently diverged NBS-LRR clades. All pairwise comparisons yielded Ka:Ks ratios strongly indicative of purifying selection. Given that *R* gene mediated resistance is inherited qualitatively rather than quantitatively, I interpret this as evidence that even closely related paralogous copies (90-95% identity) should have independent recognition specificities maintained by purifying selection.

Homeologous duplications were studied in allohexaploid wheat (AABBDD) using a section of the *go35* NBS-LRR gene (2L) of the B and D diploid donor species of wheat. Numerous synonymous substitutions distinguished the B and D genome copies, with an absence of non-synonymous substitutions. In contrast, single unique nonsynonymous substitutions were found in four out of five polyploid wheat *go35* alleles, indicating that selection pressure was indeed relaxed across the homeolocus. Recent studies on polyploid genomes have shown that duplicated resistance genes are far more likely to be eliminated than highly transcribed genes such as tRNAs and rRNAs. These results are in agreement with the view that functional divergence takes place before duplication for NBS-LRR genes, as the loci duplicated by polyploidy appear not to evolve under purifying selection, as I found for the paralogous loci investigated.

Appendix

Appendix A Accesion numbers for *Triticeae* sequences collected for phylogenetic analysis.

Search Method	**Species**	**Accession**
PSI-BLAST(Genbank)	*Aegilops tauschii*	AF509533, AAM69841, AAM69850.
	Aegilops ventricosa	CAC11106, CAC11105, AF158634, CAC11100.
	Hordeum vulgare	AAB96982, BJ471122, BI946756, AAD46472, AAD46475, AF108008, AAD46476, AAC71769, AAD46471, AAB96983, AV834807, CAD44603, AAQ16121, CA017389, AAO43441, AAM22828, AV835532, BF482358, CAD42334, AAL07813, AAL07815, AAL07814, AAL07817, AAL07816, CAD44588, CAD44587, CAD44585, CAD44583, CAD44582, CAD45037, CAD45025, CAD44579, CAD44584, CAD44589, BQ469814, AAB96984.
	Thinopyrum intermedium	AAL23743, AAP20701.

Appendix A (continued) Accesion numbers for *Triticeae* sequences collected for phylogenetic analysis (continued from Appendix A).

Search Method	Species	Accession
PSI-BLAST(Genbank)	*Triticum aestivum*	BE500158, BJ300496, AAP03077, AAK20742, AAN62914, BQ241493, CA499328, AAC71768, AAC71766, BE426789, AAC71767, AAB96979, BJ258770.
	Triticum monococcum	AF326781, AF326781.
HMM search of TIGR *Triticum aestivum* Gene Indices	*Triticum aestivum*	TC139355, TC141135, TC137521, TC141756, TC110833.
HMM search of TIGR *Hordeum vulgare* Gene Indices	*Hordeum vulgare*	TC104095, TC102033, TC104756, TC97752, TC101128, TC94550, TC106877, TC106707, TC97784, TC95466, TC105827, TC107972, TC107923, TC101965, TC95262, TC106050, TC97746, TC96650, TC93669, TC103543, AV836112, TC93571.

Appendix B Accesion numbers for *R* genes used in phylogenetic analysis.

Species	**Accession**
Aegilops tauschii	AY145086 *Lr21*
Arabidopsis thaliana	AF234174 *HRT*
	X87851 *RPM1*
	AF209732 *RPP13*
	AY062514 *RPP8*
	AF368301 *RPS2*
	NM_101094 *RPS5*
Capsicum chacoense	AF202179 *BS2*
Hordeum vulgare	AF523678 *Mla*
Lycopersicon esculentum	AJ457051 *Hero*
	AF004878 *I2*
	AF091048 *Mi*
	U65391 *Prf*
	AY007366 *Sw5-a*
	AF536200 *Tm-2*

Appendix B (continued) Accesion numbers for *R* genes used in phylogenetic analysis.

Species	**Accession**
Oryza sativa	AB013448 *Pib*
	AAO45178 *Pi-ta*
	BAA25068 *XA1*
Solanum bulbocastanum	AY426259 *RPI*
Solanum demissum	AF447489 *R1*
Solanum tuberosum	AF195939 *Gpa2*
	AJ011801 *Rx*
Triticum aestivum	AY270157 *Lr10*
	AY325736 *Pm3*
Zea mays	AF107293 *RP1-D*

Appendix C 1.) Alignment result of top tblastx search hit on DNA sequence obtained using degenerate NBS primer set designed by Yu *et al.* (1996) on *Triticum aestivum.* **2.)** Nucleotide sequence.

1.) gi|15292619|gb|AAK93796.1| .NBS-LRR-like protein [*Oryza sativa* subsp. *japonica*]

Expect value = $2x10^{-21}$
Identity = 53%

```
QLTELLRRVEPIECCIYDAEKRRTKELAVNNWLGQLRDIIYDVDEILDVVRCKGSKLLPN
+L EL RR + I   + DAE RR K+ AV  WL QLRD++YDVD+I+D+ R KGS LLPN
ELEELQRRTDLIRYSLQDAEARRMKDSAVQKWLDQLRDVMYDVDDIIDLARFKGSVLLPN

YPXXXXXXXFACKGLSVSSCFCNIGSRRHVAVTTRNMS
YP        AC GLS+SSCF NI  R  VAV  R+++
YPMSSSRKSTACSGLSLSSCFSNICIRHEVAVKIRSLN
```

2.)

TTTTATCTCATCCCNTCTATGCATCCAACGCGTTGGGAGCTCTCCCATATGGTCGACCTGCAGGCGGCC
GCGAATTCACTAGTGATTGGAATGGGTGGTGTGGGAAAGACAGCTCACAGAACTGCTGCGACGAGTAG
AACCAATAGAGTGTTGTATATATGATGCTGAGAAAAGGAGGACAAAAGAGCTAGCAGTAAATAATTGGCT
TGGTCAATTGAGAGATATTATATATGATGTAGATGAAATCTTGGACGTGGTTAGATGTAAAGGAAGCAAG
CTACTGCCTAATTATCCTTCATCATCATCAAGCAAATCATTTGCATGTAAAGGCCTTTCAGTTTCCTCTTG
TTTTTGTAACATTGGGTCACGTCGTCATGTTGCTGTCACTACAAGAAATATGTCAACTAGTGACCTTCTGT
CCGTGACCCTGGAAGAATTGGTCATAGATCTATGACCATTTCAGACCAATTGGTCGAAAGCTATTCGGG
GGGCTCCAAACCCTAAACCATTACGACCATTTTGGTCAGAAAGGTCATAATTTCCTTACACGAAAAGGTC
ATAAAGCAAACAGCGCTAGTCCGCTGCCTTACTTCTAGTTGTTAACGACCAATATAGATGGTCATAGCCT
TGTAAATTGTGGTGGGTTGCTATGACTAGGCCCACCTCATCAATTTTACCCACCCCCCCATTNCA
ATCGAATTCCCGCNGGCCGNCATGGCGGG