En masse CLONING OF EXPRESSED DISEASE RESISTANCE GENES OF WHEAT (*Triticum aestivum*) USING RNA DIFFERENTIAL DISPLAY VIA DEGENERATE PRIMERS AND DATA MINING METHODS

by

Muharrem Dilbirligi

A DISSERTATION

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfillment of Requirements

For the Degree of Doctor of Philosophy

Major: Agronomy

(Molecular Plant breeding and Genetics)

Under the Supervision of Professor Dr. K. S. Gill

Lincoln, NE

August, 2003

En masse CLONING OF DISEASE RESISTANCE GENES OF WHEAT (*Triticum aestivum*) USING RNA DIFFERENTIAL DISPLAY VIA DEGENERATE PRIMERS AND DATA MINING METHODS

Muharrem Dilbirligi, Ph. D.

University of Nebraska, 2003

Adviser: Kulvinder S. Gill

Several pests causing approximately 25% yield loss translating to about 70 billion dollar/years attack wheat plant. Genetic control via manipulating resistance genes is the most effective and economical method of controlling plant pests using two conventional approaches, map-based cloning and transposon tagging. No transposon-based system has so far been optimized to clone wheat genes. Wheat has a larger genome, of which 95-99% is non-transcribing-nested cluster of retrotransposons and duplicated genes. This complex genome organization of wheat poses serious hindrance to the map based cloning approach. The gene cloning approaches that target only the expressed portion of the genome and mapping approaches that physically localize genes into only gene containing regions are desirable for wheat.

Forty-six resistance genes conferring resistance to various types of pests have been cloned from 12 plant species. Irrespective of the host or the pest type, most resistance genes share a strong protein sequence similarity especially for domains and motifs. Also, expression profile of the majority of cloned resistance genes are unique as they are rarely and constitutively expressed and expression levels do not dramatically

change during infection. The objective of our study was to identify expressed resistance genes and physical localize them. Using modified RNA fingerprinting and data mining approaches, we totally identified 220 expressed resistance gene candidates out of the 728 analyzed. Of these 220, 125 sequences structurally resembled the known resistance gene types, as in addition to 25 to 87% protein sequence similarity, the protein sequence, order and distribution of domains and motifs unique to the cloned resistance genes, were the same. Among the remaining 95, 17 were resistance and pathogen related, 21 were the new class of nucleotide binding kinases, 21 were probable kinases, and 36 were p-loop containing unknown sequences. About 76% (167/220) were also rare transcripts including 73 novel sequences.

Irrespective of the cloning methods used, 184 out of 220 (36 unknown were eliminated) expressed resistance gene candidates were attempted to map physically using aneuploids and 339 deletion lines. Of the 184 sequences, 87 were NB/LRR type, 16 were receptor-like kinase (*Xa21* type), 13 were protein kinase (*Pto* type), 7 were *Hm1*, 2 were $Hs1^{pro-1}$, 17 were pathogen related and 42 were novel nucleotide binding/kinase-encoding type of sequences. Physical mapping localized 310 resistance loci in wheat genome detected by the 121 candidate resistance gene sequences. We combined the physical mapping information from the three homeologous groups to generate high-resolution consensus physical maps as the gene distribution and synteny is conserved among three wheat genomes. A total of 121 probe detected 151 loci, of which 143 were clustered in 26 smaller chromosomal regions encompassing ~16% of the genome. Five major resistance gene clusters were observed to contain loci corresponding to 67 sequences (55% of the total). Wheat genome contains 269 morphologically characterized resistance

genes. Two hundred twenty nine of them were genetically studied. Construction of consensus genetic maps using individual 137 genetic linkage maps allowed us to localize 110 (80 single gene inherited and 30 QTL type) morphologically characterized wheat resistance genes on the consensus genetic maps. Comparison of consensus genetic maps with consensus physical maps, 80 single-gene-inherited and 10 QTL-type wheat resistance genes were physically located in 20 small physical chromosomal regions. About 93% (85 of 90) of the wheat resistance genes were localized into 18 smaller regions where 131 of the candidate resistance gene loci mapped. About 75% of both the morphologically characterized resistance genes and the identified candidate sequences mapped to the distal 20% of the chromosomal regions.

ACKNOWLEDGEMENTS

I would like to thank to the Turkish Government for giving me the opportunity to pursue my graduate program among wonderful people in an excellent educational environment. I also would like to thank to the Agronomy and Horticulture department for employing me as a Graduate Research Assistant during my Ph. D. I am sincerely thankful to Dr. Kulvinder S. Gill for serving my major advisor. Without his invaluable suggestions, encouragement, guidance and friendship, I would not have stayed here for five years. I wish to express my gratitude to my committee members, Dr. P. Stephen Baenziger, Dr. Paul Staswick and Dr. John Foster, for their priceless support and patience throughout my graduate program. I must acknowledge my colleagues and friends (Alphabetical order): *Harvinder Bennypaul, Svetlana Bondereva, Dr. Mustafa Erayman, Jasdeep Mutti, Dr. Maroof Shah, Dr. Devinder Sandhu, Deepak Sidhu*, and all past and present graduate students for their companionship and support. My thanks also extend to our present secretaries *Sherry* and *Marlene*. I would like to thank faculty and staff of the Agronomy and Horticulture Department for their smiling faces and helps. I am thankful to my mother and father, brothers, In-laws for their encouragements and support from thousands of miles away. Last on the list but first in my heart, to my wife, **Esin**. Thanks to you for always being there when I needed; and to our daughter **Elif Nur** and **Ayse Sema**, my source of enjoyment, for inspiring the energy to accomplish my goals. Thank you all.

TABLE OF CONTENTS

FOREWORD

This dissertation is written as two manuscripts in the format required for publication in PLANT JOURNAL (for chapter 1) and GENETICS (for chapter 2) journal.

CHAPTER 1

IDENTIFICATION AND ANALYSIS OF EXPRESSED RESISTANCE GENE SEQUENCES IN WHEAT

Summary

Forty-six resistance (*R*) genes conferring resistance to various types of pests have been cloned from 12 plant species. Irrespective of the host or the pest type, most *R* genes share a strong protein sequence similarity especially for domains and motifs. The objective of this study was to identify expressed *R* genes of wheat, the fraction of which is expected to be very low in the genome. Using modified RNA fingerprinting and data mining approaches, we identified 220 expressed *R*-gene candidates out of the 728 analyzed. Of these 220, 125 sequences structurally resembled the known *R* genes as, in addition to 25 to 87% protein sequence similarity, the sequence, order, and distribution of domains and motifs unique to the know *R* genes were the same. Among the remaining 95, 17 were probable *R* related, 21 were a new class of nucleotide binding kinases, 21 were probable kinases, and 36 were p-loop containing unknown sequences. About 76% (167/220) were rare including 73 novel sequences. Three new *R*-gene specific motifs were also identified. Physical mapping of 164 best *R*-gene candidates on 339 deletion lines localized 121 mappable *R*-gene candidates to 26 small chromosomal regions encompassing ~16% of the genome. About 90 of the 110 phenotypically characterized wheat *R* genes also mapped in these 20 regions of 26 small chromosomal regions.

Introduction

Wheat (*Triticum aestivum* L. em Thell, 2n = 42, AABBDD) is attacked by a variety of pests including insects, pathogens and nematodes. On an average these pests cause 20-37% yield loss worldwide translating to about $70 billion/year (http://pseru.ars.usda.gov; Pimental et al., 1997). Genetic manipulation of resistant genes is an efficient, economical, and well-tested method of controlling wheat pests. About 229 genes conferring resistance to various pests have been identified in wheat or its wild relatives (McIntosh et al., 2000). Of the 110 that are well studied, 80 *R* genes show single-gene inheritance (3:1 ratio in F_2) and seem to have the gene-for-gene interaction with the pests (McIntosh et al., 2000; http://wheat.pw.usda.gov/ggpages/maps.shtml). This is a significant observation considering that wheat is an allo-hexaploid containing three homoeologous genomes with a very similar gene synteny and colinearity (Kihara, 1944; Sears, 1954). Inheritance of the remaining 30 genes is complex and the resistance seems to be manifested in a non-specific manner.

In crop plants, map-based cloning and transposon tagging are the two most successful methods of cloning genes with a distinct phenotype. No transposon-based system has so far been optimized to clone wheat genes. The wheat genome has approximately 16-million kb/haploid cell, and about 95-99% of it is non-transcribed, mainly consisting of nested clusters of retrotransposons and duplicated genes (Arumuganathan and Earle, 1991; SanMiguel et al., 1996; Panstruga et al., 1998; Wendel, 2000; Wicker et al., 2001; Sandhu and Gill, 2002a). The remaining 1-5% contains genes that are present in clusters of varying size and gene densities (Gill et al., 1996b; Gill et al., 1996a; Sandhu and Gill, 2002b). The difference in the gene cluster size and gene

density is due to variable amounts of non-transcribed repeated DNA interspersing genes. The gene clusters are interspersed by larger yet variable sized blocks of DNA. This complex organization of the wheat genome poses serious hindrance to the map based cloning approach. Because of the complexity and the size of the wheat genome, no functionally proven *R* gene has so far been cloned. Therefore, the gene cloning approaches that target the expressed portion of the genome are particularly desirable for wheat.

Hitherto 46 *R* genes conferring resistance against various insects, nematodes, bacteria, fungi and viruses attacking 12 plant species have been cloned (Table 1). Most of the cloned *R* genes exhibit structural similarity for DNA and putative protein sequence (Hammond-Kosack and Jones, 1997; Meyers et al., 1999; Mondragon-Palomino et al., 2002). Most of these genes follow the typical gene-for-gene hypothesis and seem to encode for components of signal transduction pathway (Michelmore and Meyers, 1998). The exceptions are *HM1* in maize (*Zea mays*) that detoxifies the fungal toxin by NADPH reductase, and *MLO* in barley (*Hordeum vulgare*) and *RPW8* in *Arabidopsis thaliana* that show broad-spectrum resistance to the corresponding fungal pathogens (Johal and Briggs, 1992; Buschges et al., 1997; Xiao et al., 2001). The structurally conserved cloned *R* genes contain one or more of the four major protein domains, namely nucleotide binding (NB) site, receptor-like transmembrane kinase (RLK), cytoplasmic protein kinase (PK) and leucine rich repeat (LRR). The NB domain is also present in apoptosis regulatory proteins such as *APF-1* and *CED-4* that have been characterized in humans and other eukaryotes (Van der Biezen and Jones, 1998; Noel et al., 1999).

The cloned *R* genes can be grouped into four distinct classes based on the domains present (Table 1). The most abundant with 31 of the 46 *R* genes is the NB and LRR (NB/LRR) containing class. This class can further be divided into two sub-classes differentiated by the presence of either a Tool-Interleukin Receptor-like (TIR) or a coiled coil (CC) domain at the N-terminal (Meyers et al., 1999; Mondragon-Palomino et al., 2002). The TIR domain has only been reported in dicots. The second class of *R* genes contains LRR along with some infrequent but conserved regions. The third class contains the LRR and the PK domains (RLK). The fourth class of *R* genes contains only PK but lacks both LRR and NB domains. Significant structural and functional differences may exist among *R* genes within a class. For instance, LRR present in $HS1^{pro-1}$ gene is intercellular compared to *CF2, CF4, CF5, CF9, XA21, VE1 and VE2* where it is extracellular (Dixon et al., 1996; Cai et al., 1997; Kawchuk et al., 2001). Similarly, there are two tandemly arrayed PK domains in *RPG1* compared to only one in *XA21*, *PTO* and *PBS1* (Martin et al., 1993; Song et al., 1995; Swiderski and Innes, 2001; Brueggeman et al., 2002).

The four major domains present in plant *R* genes contain one or more motifs with highly conserved amino acid sequences. The LRR domain contains 9 to 41 imperfect repeats, each about 25 amino acids long with a consensus amino acid sequence of xx(L)x(L)xxxx (Cooley et al., 2000). The PK domain of both *PTO* and *XA21* contains up to 25 amino acids long motifs, where the first three (DFG) and the last two (PE) residues are highly conserved. An internal threonine (T) and a serine (S) residue are essential for autophosphorylation and thus are conserved (Liu et al., 2002). The NB domain is present in *R* genes as well as several other kinases such as ATP/GTP binding proteins. It

contains three motifs; a kinase-1a (p-loop), kinase-2, and a putative kinase-3a (Traut, 1994; Tameling et al., 2002). These motifs in *R* genes have a consensus sequence of GxxGxGK(T/S)T, LxxxDDVW and GxxxxTxR for p-loop, kinase-2, and the putative kinase-3a, respectively, that is remarkably different from that present in other NB-encoding proteins (Hammond-Kosack and Jones, 1997; Meyers et al., 1999). Other motifs present in the NB domain of NB/LRR-type *R* genes are GLPL, RNBS-D and MHD (Meyers et al., 1999). The sequences interspersing these motifs and domains can be very different even among homologs of an *R* gene (Michelmore and Meyers, 1998; Pan et al., 2000).

Genomic DNA sequence analysis revealed that there are 166 putative NB/LRR genes in *Arabidopsis thaliana* and about 600 in rice (*Oryza sativa*) (TAGI, 2000; Goff et al., 2002; Richly et al., 2002). It is however unknown how many of these genes are functional. Analyses of NB/LRR-containing genomic sequences from various crop plants suggested that only a small fraction of these are functional (Chin et al., 2001; Sun et al., 2001; Shen et al., 2002). With the objective to clone resistance genes *en masse*, primers complimentary to the conserved parts of the motifs have been used in various crop plants to amplify genomic DNA followed by selection for fragment size conserved among *R* genes (Leister et al., 1996; Yu et al., 1996; Collins et al., 1998). Transcripts for only nine of the 173 resistant gene analogs (RGAs) from six crop species have been observed as ESTs (http://www.ncbi.nlm.nih.gov/entrez; Table 6). These data suggest that only a small fraction of the *R*-gene-like genomic sequences of crop plants is functional, and this fraction is expected to be even smaller in crop plants with larger genomes.

Therefore, the objective of this study was to isolate and characterize expressed fraction of the wheat *R* genes.

Results

Modified RNA fingerprinting

Ten different primer combinations —p-loop degenerate primer for the 5' and GLPL or one of the nine 'T' primers for the 3' end of the genes— were used to amplify cDNA (Table 2, Methods). Upon size separation on a 5% polyacrylamide-urea gel, each p-loop/'T' primer combination generated 80 to 100 bands (Figure 1A). The p-loop/GLPL primer set generated 76 bands. The size of the fragment bands ranged between ~150 and 1300 bp except for the p-loop/GLPL primer combination, where the size of the smallest band was ~300 bp. Presence of similar band patterns suggested that various primer combinations amplified the same fragment bands. Of the total 900, only about 220 fragment bands were unique (Figure 1A). These bands were excised from the gel in 160 pieces because some of the pieces contained two to seven bands that were too close to be individually excised from the gel. Of the 160 excised gel pieces, 104 contained 144 fragment bands of the p-loop/'T' primers and 56 contained 76 of the p-loop/GLPL generated bands. The DNA from each gel piece was eluted, re-amplified, and cloned (Figure 1B). Two clones corresponding to each gel piece were sequenced. Additional clones were sequenced if two clones from a sample were different. A total of 385 clones from the 160 gel pieces were sequenced and analyzed further.

"ContigExpress" analysis using 385 sequences resulted in 121 unique contigs of which 64 corresponded to the p-loop/'T' primers and 57 to the p-loop/GLPL

combination. The longest sequence of each contig was used for further analysis and all sequences were confirmed to contain a p-loop (kinase-1a) site. Individual 'Gap' (GCG) analysis of these sequences showed that 121 sequences shared 21% (*unl115 to unl175*) to about 90% (*unl184* to *unl185, unl201* to *unl202 and unl208, and unl139* to *unl160*) sequence similarity. Only about 27% of the sequences were closely related and sequence similarity for the remaining 73% was less than 50%.

Putative protein sequences of these clones were compared with the available sequence database using BLASTX searches (Altschul et al., 1997; www.ncbi.nlm.nih.gov/BLAST). Only 48 of the 121 sequences showed partial or overall homology to the known or putatively annotated genes. Nineteen of these showed more than 80% sequence similarity and 29 showed from 35 to 80%, with E values of E $\leq 10^{-40}$ to E $\leq 10^{-8}$, respectively. Of these 48, nine were wheat homologous for *PTO* (*unl176*), *XA21* (*unl57*) and NB/LRR-type cloned *R* genes (e.g. *unl174, unl177*), seven were disease-resistance related, five were other NB encoding (ATP and GTP binding) and five were disease-resistance, defense-response or stress-related kinases (Table 3). Twenty-two sequences were homologous (BLASTX E values E $\leq 10^{-29}$ to E $\leq 10^{-57}$) to genes controlling cell structural and metabolic activities.

Motif analysis of unknown sequences

The 'MotifSearch' (GCG) was performed for the 73 sequences for which no significant match was identified in the database in order to identify putative functional motifs and domains. All sequences had a conserved p-loop (kinase-1a) motif and 38 also had both predicted kinase-2 and kinase-3a motifs (Table 4). Of these 38, 17 contained an NB domain highly similar to NB/LRR-type *R* genes, of which seven lack either a kinase-

2 or a kinase-3a motif, with varying distances between motifs (sequences with 'unl' prefix in Figure 4a). Motif residues among these 17 sequences were in a consensus of (L/V) (L/V) (L/V) (L/I/D) D (D/I/V/L) for kinase-2 and (E/G/F/V) (T/S/G/Q) x (T/Y) (T/S) R for kinase-3a along with GVGKTT for p-loop. Further, the 'BestFit' analysis in GCG showed that the sequence homology between cloned NB/LRR-type *R* genes and these 17 fragments ranged from 21 % (between *RPS2* and *Xunl194*) to 31 % (between *PIB* and *Xunl204*) in nucleotide level. The remaining 21 of the 38 sequences were probable kinases that contain an invariant aspartate (D) in kinase-2 and an arginine (R) in kinase-3a. Of the remaining 35 sequences, 21 had either a kinase-2 or a kinase-3a along with the p-loop motif as found in NB or kinase-encoding proteins. The remaining 14 showed no motif other than the p-loop thus, were not analyzed further. In summary, 43 of the 121 sequences structurally resemble *R* genes including 26 NB/LRR, PK and RLK type, 12 were *R* related kinases (Ca^{++} dependent, ser/thr, protein kinases and etc.) and five were probable *R*-related sequences such as ATP and GTP-binding proteins. Additional 42 sequences contained putative kinase-2, kinase-3a and p-loop (Figure 2).

Data mining

In order to identify wheat ESTs structurally resembling the known *R* genes, sequence information of 22 *R* genes representing the five known classes —NB/LRR (18 genes), PK (*PTO)*, RLK *(XA21)*, *HM1* and $HS1^{pro-1}$— was used for data mining as described in the material and methods section. All four methods —(*i*) a consensus sequence comparison (*ii*) domain search, (iii) individual full-length comparison, and (iv) consensus sequences of single and multiple motif comparison— were used for the

NB/LRR-type *R* genes. Only individual full-length and domain comparisons, however, were performed for the other *R*-gene classes because of the lack of known motifs.

The first three methods of data mining identified 78, 221 and 344 wheat ESTs, respectively with E values ranging from E $\leq 10^{-1}$ to E $\leq 10^{-121}$. The fourth method however failed to identify any EST, probably due to insufficient size of the query sequences for the BLAST search. The number of ESTs identified by the consensus sequence comparisons is significantly less than the other two methods, probably because of large gaps in the query sequences corresponding to the variable regions among various cloned *R* genes. The individual full-length search method was the most successful as it identified maximum number of EST including the ones identified by the consensus sequence and the domain search.

The ESTs selected by various methods were analyzed further in order to identify sequences with structural similarity to the known resistance genes. Homology with the known *R* genes, TBLASTN E value, presence, order and size of the interspersing gaps of the motifs, and a few other structural criteria were used to select *R*-gene-like wheat ESTs. Homologs for *HM1*, *HS1*$^{pro-1}$, *PTO* and *XA21* were selected at >50% sequence homology with TBLASTN E value of $\leq 10^{-20}$. The motif search was not used as no significantly conserved motif has been observed for these *R* genes. The NB/LRR-type *R* genes in monocots usually contain four distinct domains: CC, NB, NB-spanning, and LRR. The CC and the LRR domains lack any conserved motifs whereas the NB and the NB-spanning regions have kinase-1a (p-loop), kinase-2, kinase-3a within the NB domain, and GLPL, RNBS-D, and MHD within the NB-spanning region. Among the cloned *R* genes,

the NB and NB-spanning domains are more conserved compared to the CC and the LRR domains. In addition to the sequence, size of the regions interspersing the motifs is also conserved. Size and the putative protein sequences of the identified ESTs were compared with that of the known *R* genes especially for the motifs and the interspersing regions. The ESTs containing less than two of the above-mentioned motifs were not selected for further analysis. Considering that the *R* genes share domains and motifs with other proteins, only the ESTs showing more than 40% protein sequence similarity ($E \leq 10^{-10}$ TBLASTN) with one or more of the *R* genes were selected. The selected ESTs were assembled into contigs using 'ContigExpress' (default values) and CLUSTALW (gap opening penalty 3.0, gap extension penalty 1.0). The longest sequence from each of the contigs was used for further analysis.

The full-length sequence comparisons identified 243 NB/LRR-type ESTs and 101 homologous to other types of *R* genes with E values ranging from $E \leq 10^{-1}$ to $E \leq 10^{-117}$. Further analysis (as mentioned earlier) of these EST sequences selected 112, 21, 27, 14 and two ESTs corresponding to NB/LRR, *PTO*, *XA21*, *HM1* and *HS1^{pro-1}* types of *R* genes, respectively. The selected 176 ESTs were then used as query sequences for BLASTX search in the non-redundant (nr) protein database. Only the ESTs where the best hit was a cloned *R* gene were selected. All 64 ESTs corresponding to the *PTO*, *XA21*, *HM1*, and *HS1^{pro-1}* selected as the best hits were one of these genes with a significant E value range from $E \leq 10^{-27}$ (between *BE415090* and *HS1-$^{pro-1}$* gene) to $E \leq 10^{-67}$ (between *BF482703* and *PTO* gene). Of the 112 NB/LRR-like ESTs, the cloned *R* genes were the best hits for 99 with E values ranging from $E \leq 10^{-10}$ (between *BE444145* and *MLA6* gene) to $E \leq 10^{-94}$ (between *BE499523* and *VRGA1* gene). Of the remaining 13

ESTs, seven showed strong homology to ABC transporter proteins (from E $\leq 10^{-20}$ to E $\leq 10^{-128}$), and two to unknown proteins (from E $\leq 10^{-23}$ to E $\leq 10^{-45}$). The remaining four ESTs showed weak homology to the known *R* genes (E $\leq 10^{-3}$ to E $\leq 10^{-5}$) but a higher level of homology to unknown proteins. These 13 ESTs were not selected for further analysis. The selected 163 ESTs were analyzed with 'VecScreen' 'ORF Finder' 'BLASTP' programs (www.ncbi.nlm.nih.gov) to confirm annotation of the selected ESTs. These ESTs were assembled into 99 contigs and the longest EST sequences of the contigs were used for further analysis. As a result, 11 *PTO*-like ESTs (*BE403520, BE417607, BE442802, BE442854, BE492937, BE637850, BE637867, BF201229, BF482232, BF482703, BG604519*), 12 *XA21*-like ESTs (*AW448377, BE213652, BE403187, BE403270, BE405531, BE443325, BE443579, BE445978, BE492332, BE518042, BF473126, BF473851*), seven *HM1*-like ESTs (*BE213346, BE591764, BE499505, BE517781, BE518335, BE585541, BE607075*) and two *HS1*$^{pro-1}$-like ESTs (*BE415090, BE604247*) along with 67 NB/LRR-like ESTs were selected.

Structural analysis of the putative R-gene sequences

Although the domains and motifs present in the NB/LRR-type *R* genes are also present in other proteins, the sequence, order, and gap length among various motifs is unique to *R* genes. We therefore compared the detailed structure of the NB/LRR-type identified *R* genes with that of the know *R* genes. Based on the available structural information of the known *R* genes, we constructed a hypothetical monocot *R* gene to which the identified wheat sequences were compared (Figure 3). For each identified sequence, the *R* gene showing the highest level of homology (best hit) is given on the left

(Figure 3). Extent of sequence homology ranged from 40% (between *BE425456* and *XA1*, *BE418575* and *RPP13*, *BE605005* and *CRE3*, *BM137173* and *MI*) to 87% (*BE499523* and *VRGA1*). Homology for 31 sequences was about 40 to 50%, for 23 was 50 to 60% and for 13 was 60 to 87%. The sequence homology was mainly due to the conserved motifs and domains, whereas the interspersing sequences were usually very different. As expected, the number of 'best hits' for the identified wheat sequences was the highest with barley and wheat genes: 15 for *MLA6* and nine for *CRE3*. Similar numbers for *PIB*, *RPR1* and *XA1* genes of rice, and *RP1-D* of maize ranged from four to six. There were 14 wheat sequences for which the 'best hit' was one of the *A. thaliana R* genes, *RPP13, RPP8, RPS2, RPM1* and *RPP5*. There was only one sequence for which the 'best hit' was the dicot *R* genes *MI* and *L6*.

A total of 87 NB/LRR-like wheat sequences were cloned and identified (Table 5). Of these, 34 containing two or more NB motifs were aligned with cloned 15 NB/LRR-type *R* genes (Figure 4a). Location and order of the NB domain motifs in 27 wheat sequences was the same as found in the *R* genes. Seven sequences lacked either kinase-2 or kinase-3a motifs (Figure 4a). As shown by the MEME and CLUSTALW analysis, the sequence of kinase-1a (p-loop) and hydrophobic (GLPL) motifs in the identified wheat sequences was the same as found in the *R* genes. The consensus sequence for these motifs was GGLGKTTL and GLPLAI, respectively (Figure 3 and 4A). The kinase-2 motif was also well conserved with a consensus sequence of LVVLDDVW, where two aspartate amino acids (DD) in the motif residues were almost always invariant. As seen in the *R* genes, the putative kinase-3a motif was variable in the identified sequences also

with a consensus sequence of GSRVIVTTR. The arginine (R) residue was however invariant, as found in cloned *R* genes also (Figure 3 and 4A).

Since the NB domain of the NB/LRR-type *R* genes is the most conserved and perhaps functionally most important, we concentrated on the structural comparison of this domain between the cloned *R* genes and the wheat sequences. In addition to the conserved motif sequences, the size of regions interspersing the motifs was also conserved (Figure 4a). The range of gap in the cloned *R* genes was 43 to 64 aa between kinase-1a and kinase-2, 13 to 16 between kinase-2 and kinase-3a, seven between RNBS-D and WAEGF, 31 to 44 between WAEGF and MHD, 65 to 110 between MHD and LRVLDL, and 9 to 15 between LRVLDL and LRYL motifs. In 32 of the 34 NB/LRR-type the wheat sequences the size range for the interspersing region among these motifs was the same as observed for the cloned *R* genes (Figure 4a and B).

Comparison of the cloned *R* genes with the identified wheat sequences revealed three new motifs specific for *R* genes. A motif with a consensus sequence of WIAEGF was located between RNBS-D and MHD, always 7 amino acids from RNBS-D (Figure 4b). The second motif was positioned in the core of the third LRR and had a consensus sequence of LRVLDL (Figure 4b). The fourth LRR also had a conserved region with a consensus of LRYL; however, this was not as prominent as LRVLDL motif. The combined "FindPattern" analysis (GCG) with RNBS-D, WIAEGF and LRVLDL (with no ambiguity) motifs identified only the CC/NB/LRR-type *R* genes suggested that these motifs are unique to this class of *R* genes. However, these motifs were individually present in some other proteins as well. Some RLK and unknown proteins, including plant, virus, bacteria, and human, contained the LRVLDL, while photosystem II

chlorophyll a-binding protein and photosystem II p680 chlorophyll —an apoprotein from monocots— contained the WIAEGF motif.

Mapping and physical analysis of the sequences

Of the 220 fragments, 164 corresponding to the best *R*-gene candidates were physically mapped using wheat nullisomic-tetrasomic, ditelosomic, and 339 deletion lines. Of these, 43 were not mapped as a smear pattern was observed upon gel-blot DNA hybridization. The remaining 121 fragments detected 664 RFLP bands corresponding to 310 loci present on all 21 wheat chromosomes. Each fragment band was localized to the smallest possible region on a consensus physical map using the strategy described earlier (Gill and Gill, 1994; Gill et al., 1996b; Gill et al., 1996a); the results are given in Figure 5. Most of the loci were clustered in 26 chromosomal regions. Size of these regions totaled to about 16% of the wheat genome (Dilbirligi et al., 2003). The number of loci mapping to a region ranged from 2 to 29. About 50% of the loci mapped to five regions (red colored; Figure 5). About 70% of the loci mapped to 7% of the distal chromosomal regions.

To date, 229 *R* loci conferring resistance to various wheat diseases have been phenotypically characterized, of which 110 have been genetically mapped (McIntosh et al., 2000; http\\wheat.pw.usda.org). Of these, 90 mapped to the 26 *R*-gene-candidate-containing regions (appendix Figure A1). The proportion of the known *R* genes in each of the regions matched with that of the *R*-gene candidates (Figure 5).

Discussion

Forty-six resistance (*R*) genes conferring resistance to various types of pests have been cloned from 12 plant species. The mechanism by which the plants recognize and respond to pest attack seems to be conserved among plants as most *R* genes share a strong structural similarity irrespective of the host or the pest type. It should therefore be possible to exploit this structural similarity to identify and isolate *R* genes *en masse* from any crop plant. Amplifying genomic DNA using primers for the conserved domains and motifs and selecting for the interspersing fragment size typical of *R* genes, was used to clone RGAs from various plants (Leister et al., 1996; Yu et al., 1996; Kanazin et al., 1996; SanMiguel et al., 1996; Aarts et al., 1998; Collins et al., 1998; Seah et al., 1998; Shen et al., 1998). This approach was not so successful probably because there are many pseudogenes with structural similarities to the *R* genes. Furthermore, most of the functional *R* genes have been shown to be single or few copies that will be out-competed by multiple copy non-functional RGA during random cloning and sequencing of the amplified fragments. So far about 800 RGAs have been cloned from 20 plant species (http://www.ncbi.nlm.nih.gov/entrez). Only one of these has been confirmed to correspond to a functional *R* gene (*DM3*) (Shen et al., 2002). Comparison of the RGAs from six major plants each with an average number of 200,000 ESTs revealed that only about 5% of the RGAs are expressed (Table 6). The remaining are either non-functional or are rare transcripts.