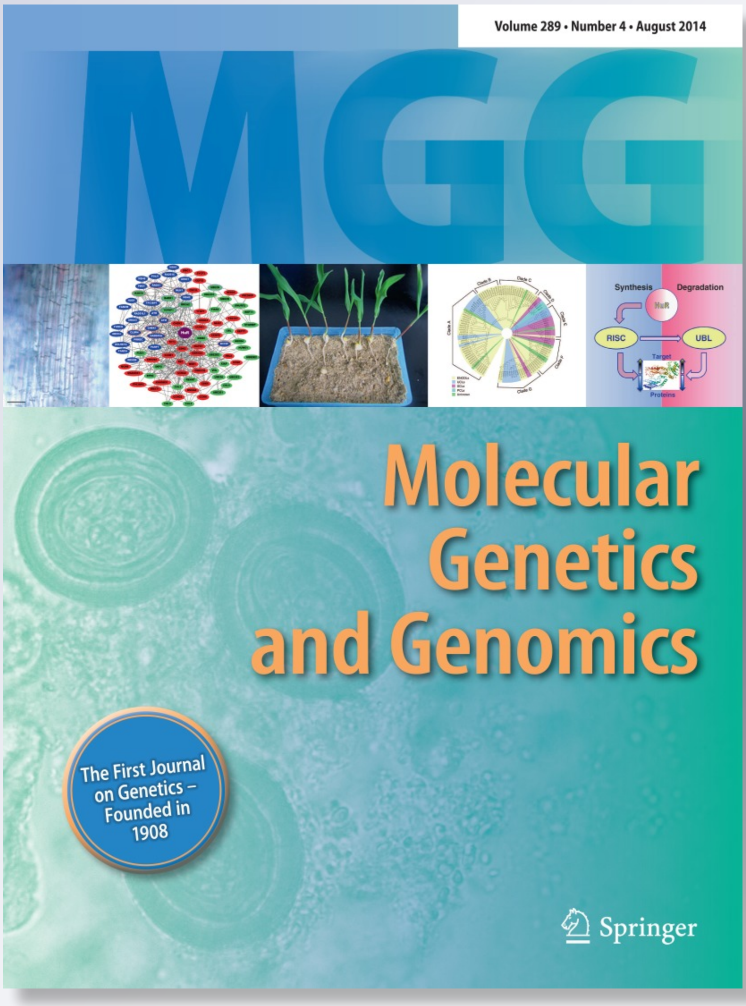
# *Characterization of novel wheat NBS domain-containing sequences and their utilization,* in silico, *for genome-scale* R-*gene mining*

**Dhia Bouktila, Yosra Habachi-Houimli, Yosra Khalfallah, Maha Mezghani-Khemakhem, Mohamed Makni & Hanem Makni**

**Molecular Genetics and Genomics**

ISSN 1617-4615
Volume 289
Number 4

Mol Genet Genomics (2014)
289:599-613
DOI 10.1007/s00438-014-0834-4


Volume 289 • Number 4 • August 2014
MGG
Molecular Genetics and Genomics
The First Journal on Genetics – Founded in 1908
Springer

Mol Genet Genomics (2014) 289:599–613
DOI 10.1007/s00438-014-0834-4

ORIGINAL PAPER

# Characterization of novel wheat NBS domain-containing sequences and their utilization, *in silico*, for genome-scale *R*-gene mining

**Dhia Bouktila · Yosra Habachi-Houimli · Yosra Khalfallah · Maha Mezghani-Khemakhem · Mohamed Makni · Hanem Makni**

Received: 6 August 2013 / Accepted: 21 February 2014 / Published online: 18 March 2014

**Abstract** In crop improvement, the isolation, cloning and transfer of disease resistance genes (*R*-genes) is an ultimate goal usually starting from tentative *R*-gene analogs (RGAs) that are identified on the basis of their structure. For bread wheat, recent advances in genome sequencing are supporting the efforts of wheat geneticists worldwide. Among wheat *R*-genes, nucleotide-binding site (NBS)-encoding ones represent a major class. In this study, we have used a polymerase chain reaction-based approach to amplify and clone NBS-type RGAs from a bread wheat cultivar, 'Salambo 80.' Four novel complete ORF sequences showing similarities to previously reported *R*-genes/RGAs were used for *in silico* analyses. In a first step, where analyses were focused on the NBS domain, these sequences were phylogenetically assigned to two distinct groups: a first group close to leaf rust *Lr21* resistance proteins; and a second one similar to cyst nematode resistance proteins. In a second step, sequences were used as initial seeds to walk up and downstream the NBS domain. This procedure enabled identifying 8 loci ranging in size between 2,115 and 7,653 bp. *Ab initio* gene prediction identified 8 gene models, among which two had complete ORFs. While GenBank survey confirmed the belonging of sequences to two groups, subsequent characterization using IWGSC genomic and proteomic data showed that the 8 gene models, reported in this study, were unique and their loci matched scaffolds on chromosome arms 1AS, 1BS, 4BS and 1DS. The gene model located on 1DS is a pseudo-*Lr21* that was shown to have an NBS-LRR domain structure, while the potential association of the RGAs, here reported, is discussed. This study has produced novel *R*-gene-like loci and models in the wheat genome and provides the first steps toward further elucidation of their role in wheat disease resistance.

**Keywords** *Triticum aestivum* · Nucleotide-binding site (NBS) · Plant · Disease · *R*-gene · Genome

Communicated by S. Hohmann.

Nucleotide sequence data reported are available in the GenBank database under the accession numbers JX566982 to JX566985.

D. Bouktila (✉) · Y. Habachi-Houimli · Y. Khalfallah · M. Mezghani-Khemakhem · M. Makni · H. Makni
Unité de Recherche Génomique des Insectes Ravageurs des Cultures d'intérêt agronomique (GIRC, UR11ES10), Faculté des Sciences de Tunis, Université de Tunis El-Manar, El-Manar, 2092 Tunis, Tunisia
e-mail: dhia_bouktila2000@yahoo.fr

D. Bouktila
Institut Supérieur de Biotechnologie de Béja (ISBB), Université de Jendouba, 9000 Béja, Tunisia

H. Makni
Institut Supérieur de l'Animation pour la Jeunesse et la Culture (ISAJC), Université de Tunis, Bir-El-Bey, Tunisia

## Introduction

In their struggle against attacks of viruses, bacteria, fungi, protozoa, nematodes and insects, plants have evolved a wide range of defense mechanisms. While some of these resistance strategies rely on simple physical or chemical barriers, modern concepts in plant immunity focus on the role and evolution of plant protein receptors corresponding to specific pathogen effectors. To explain this interaction, at least three models are currently widely endorsed. The first, called *gene-for-gene* model (Flor 1971), involves the direct effect of a plant receptor that recognizes a specific pathogen effector. The second, an extension of the gene-for-gene model, called the *guard* model (Jones and Dangl 2006), postulates that the resistance response does not occur simply as a consequence of direct recognition of the pathogen effector by its target protein, but a further cooperation is required between the target protein and some host's intracellular receptors. The third, a recent modification of the guard model, called the *decoy* model (van der Hoorn and Kamoun 2008), argues that specific host proteins, called *decoys*, bind the pathogen effectors and act as mediators in interactions with resistance proteins (*R*-proteins). If a host plant does not possess appropriate receptors, pathogen effectors will induce the suppression of defense mechanisms, which results in effector-triggered susceptibility (ETS). Conversely, if a host has suitable receptors, pathogen effectors will be the starting point of a relevant defense response, referred to as effector-triggered immunity (ETI) (DeYoung and Innes 2006; de Wit 2007). As a consequence of this, plant–pathogen interactions are similar to an endless arms race where the host (through efficient resistant proteins) and the pathogen (through virulent effectors) exert alternately selective pressure on each other (Hein et al. 2009). From the viewpoint of the plant breeder/geneticist, the rapid molecular changes involved in pathogen adaptation impose a race against time to permanently identify new resistant sources and accelerate transfer of efficient genes into commercial cultivars, through molecular and genetic engineering technologies.

In the current context marked by the advent of genomic era and emerging functional information, ever increasing amounts of sequence data related to resistance genes (*R*-genes) are being produced globally and stored in databases. To date, 112 *R*-genes have been cloned and manually curated from numerous crop and model plant species, belonging to both mono- and dicotyledons (http://www.prgdb.org; accessed January 16, 2014). These *R*-genes are currently grouped into five well-studied classes based on the presence of specific domains (Sanseverino and Ercolano 2012). The first class is the CNL one that comprises *R*-genes encoding proteins with at least a coiled-coil (CC) domain, a nucleotide-binding site and a leucine-rich repeat (CC–NBS–LRR); the TNL class includes those with a Toll–interleukin receptor-like domain, a nucleotide-binding site and a leucine-rich repeat (TIR–NBS–LRR); the RLP class, acronym for receptor-like protein, groups those with a receptor serine–threonine kinase-like domain and a leucine-rich repeat (ser/thr-LRR), the RLK class contains those with a kinase domain and a leucine-rich repeat (KIN–LRR); and the KIN class groups proteins containing only a kinase domain. In addition to these well-studied *R*-classes, a sixth class (class Others-*R*) was defined to include many other resistance proteins that have been discovered, but do not fall within the previous classes and whose functional mechanisms are also usually different. For several reasons, this classification is far from being conclusive. Indeed, with the recently discovered additional domains [e.g., WRKY (Deslandes et al. 2003)]; newly described domain combinations [e.g., TIR–CC–NBS–LRR (Kohler et al. 2008)]; and the identification of sequences formed by single domains [e.g., NBS (Sanseverino and Ercolano 2012)], new questions are raised concerning the possible involvement of new classes in the resistance process. While *R*-protein classes RLK and RLP contain a transmembrane domain and act mainly as pathogen pattern recognition receptors (PRRs), this is not the case of classes TNL and CNL, lacking clear membrane anchor domains; and operating mainly as cytoplasmic receptors, directly or indirectly recognizing pathogen effectors introduced into the host cells; or (in some cases) acting in signal transduction pathways downstream PRR receptors (Glowacki et al. 2011). Most of the identified *R*-proteins belong to these two classes making together the NBS–LRR protein family.

NBS–LRR proteins play their role due to the recruitment of a number of domains displaying various functions. The two main domains, NBS and LRR, seem to be the most crucial in the pathogen recognition process and the activation of signal transduction in response to pathogen attack. The central NBS domain is homologous to the nucleotide-binding site (NBS) of ATPases, GTPases and various other nucleotide-binding proteins (Traut 1994) and includes several highly conserved and strictly ordered motifs, such as the P-loop, kinase-2 and GLPL motifs (Tan and Wu 2012). The C-terminal LRR domain consists of multiple copies of an imperfect leucine-rich-repeat sequence (Du Preez 2005). This LRR domain is highly variable, devoted essentially to protein–protein interactions (Jones and Jones 1997). The N-terminal TIR domain of TNLs shows homology to domains found in both the Toll receptor of *Drosophila* and the mammalian interleukin receptor (Whitham et al. 1994). The N-terminal domain of CNLs is predicted to form a CC structure (Pan et al. 2000). Comparative genomic analyses have indicated that plant genomes can encode several hundreds of NBS–LRR proteins and that there is a great heterogeneity in terms of number and distribution of the classes CNL/TNL. To date, a large number of NBS-encoding sequences have been isolated from various plant species through genome-wide analyses: for example, from about 50 in *Carica papaya* (Porter et al. 2009) and *Cucumis sativus* (Wan et al. 2013) to 653 in *Oryza sativa* L. spp. Indica (Shang et al. 2009).

In monocotyledons, no TNL genes have been isolated so far; although in rice, several non-NBS–LRR proteins were identified that encoded a TIR domain (Bai et al. 2002). The first reported NBS–LRR sequence from a monocot genome was identified by Lagudah et al. (1997), who used a molecular marker co-segregating with resistance gene *Cre3*, conferring wheat resistance to the Australian pathotype of the cereal cyst nematode (CCN) (*Heterodera* *avenae*). This (non-coding) marker was used for probing wheat genomic and cDNA libraries, leading to the molecular sequencing of *Cre3*. Later, Seah et al. (1998) used the *Cre3* sequence, previously identified by Lagudah et al. (1997) to design specific primers. This polymerase chain reaction (PCR)-based approach yielded two new wheat and three new barley NBS–LRR resistance gene analog (RGA) sequences. Frick et al. (1998) identified a RAPD marker co-segregating with the stripe rust resistance gene *Yr10*. Sequencing that marker revealed that it was homologous to the NBS sequence of the *L6* flax rust resistance gene. Later, Maleki et al. (2003) utilized the *Cre3* and *Yr10* NBS sequences available, to design degenerate primer sets for amplification of wheat NBS segments spanning from the P-loop to the GLPLA motives. They obtained only two novel NBS clones using this approach. Yet, when they used a reverse primer designed to the FMYHAL motif, 22 amino acids upstream GLPLA, they could obtain six additional novel NBS sequences. Dilbirligi and Gill (2003) and Dilbirligi et al. (2004) have reported many RGA sequences in wheat, and physical map positions of some RGAs were determined. Bozkurt et al. (2007) used degenerate primers designed by Leister et al. (1998) to identify RGAs from *Triticum aestivum* L. and its wild wheat relatives *T. monococcum* and *T. dicoccoides*. Moreover, several RGAs with a resistance potential, called candidate *R*-genes, were identified (Lagudah et al. 1997; Leister et al. 1998; Seah et al. 1998; Spielmeyer et al. 1998; Collins et al. 1999; Mago et al. 1999; Deng et al. 2000; Srichumpa et al. 2005; Gennaro et al. 2009; Loutre et al. 2009).

For several years, the very large size and polyploidy complexity of the bread wheat genome have been substantial barriers to genome analysis. However, recently, Brenchley et al. (2012) have reported the sequencing of this large 17-gigabase-pair hexaploid genome using 454 pyrosequencing technology and comparison with the sequences of diploid ancestral and progenitor genomes. Between 94,000 and 96,000 genes were identified, and two-thirds were assigned to the three component genomes (A, B and D) of hexaploid wheat. The wheat variety Chinese Spring (CS42) was selected for sequencing because of its wide use in genome studies (Gill et al. 2004). Purified nuclear DNA was sequenced to generate 220 million reads (85 Gb of sequence), corresponding to approximately fivefold coverage on the basis of an estimated genome size of 17 Gb. Although these assemblies are fragmentary, they form a powerful framework for identifying genes, accelerating further genome sequencing and facilitating genome-scale analyses (Brenchley et al. 2012).

The efforts we have been deploying throughout the last decade in studying pest resistance in Tunisian cereal germplasm through several approaches (i.e., regular field tests, infestation surveys, Marker-Assisted Selection MAS, etc…) have resulted in the description of number of resistant sources to insects such as Hessian fly *Mayetiola destructor* Say (Bouktila et al. 2005, 2006; Makni et al. 2011) or greenbug *Schizaphis graminum* Rondani (Bouktila et al. 2012; Kharrat et al. 2012). Worldwide, more than 100 monogenic and polygenic arthropod plant resistance gene loci have been characterized, to date, by molecular mapping, and several are in use via MAS in breeding lines (Smith and Clement. 2012). Among these, only a couple of *R*-genes have been cloned: *Mi-1.2* conferring resistance of Tomato to the aphid *Macrosiphum Euphorbiae* (Rossi et al. 1998) and *Vat* conferring resistance of melon to the aphid *Aphis gossypii* (Boissot et al. 2010). Both these genes were found to be members of the CC–NBS–LRR *R*-gene class. Therefore, multiplying efforts to identify novel NBS sequences from phenotypically resistant cultivars will be an important step toward further *R*-gene cloning. Because NBS–LRR genes tend to occur in complex, rapidly evolving clusters often containing distantly related members (Wei et al. 1999), the use of such cultivars known to be resistant to specific stresses may result in the identification of *R*-genes/RGAs in association with different stresses. The two major aims of our study were firstly to isolate and characterize NBS–LRR class RGAs from 'Salambo 80,' a cultivar previously studied from both agronomical and molecular sides for its resistance to the Hessian fly (Bouktila et al. 2005, 2006) and secondly to develop a genome-scale data mining approach, where the identified RGAs would serve as initial seeds to identify tentative *T. aestivum R*-gene-like sequences.

## Materials and methods

### Plant material

The hexaploïd bread wheat 'Salambo 80' (Pato//Corre Camminos/IniaCM1021-7MB-14BJ-4BJ-0BJ) was used as DNA source for PCR amplification. This cultivar was tested in 2005, for phenotypic response, to the Hessian fly *M. destructor* and showed a highly resistant behavior (Bouktila et al. 2005).

### Cloning and sequencing of PCR products

DNA isolation was carried out, by the CTAB method, as previously reported by Doyle and Doyle (1987). Primers used in this study (Table 1) were synthesized from the sequences described by Maleki et al. (2003). Primers NBSfor1-4 were designed in the sense direction, corresponding to the amino acid sequence SGSGKSTL found in the P-loop of the wheat *Cre3* (AF052641; Lagudah et al. 1997) and other *R*-genes. Because the hydrophobic motif GLPLA is not common

**Table 1** Degenerate primers used to amplify resistance–gene analogs in *Triticum aestivum* L. cv. 'Salambo 80' (Maleki et al. 2003)

| Consensus motif nucleotide-binding site | Primer | Oligonucleotide sequence[a] (5′ → 3′) |
|---|---|---|
| P-loop SGSGKSTL | NBSfor1 | GGIGGIGTIGGIAAIACIAC |
| | NBSfor2 | GGIGTIWSIGGIWSIGGIAA |
| | NBSfor3 | GGNGGNGTHGGIAAGACNAC |
| | NBSfor4 | GGNTYNGGIAARACWACIC |
| FMYHAL motif, 22 amino acids upstream of the putative GLPLA motif | NBSrev1 | ARIGCTARIGGIARICC |

[a] IUPAC Standard Codes of mixed bases. *I* Iosine

in cereals (Maleki et al. 2003), the unique antisense primer NBSrev1 was designed to the FMYHAL motif, 22 amino acids upstream of the putative GLPLA motif. Therefore, four primer combinations of sense and anti-sense primers differing only by sense primer were used: I: NBSfor1/NBSrev1, II: NBSfor2/NBSrev1, III: NBSfor3/NBSrev1 and IV: NBSfor4/NBSrev1. PCR was performed in a total volume of 25 mL, containing 50 ng of genomic DNA, 1× PCR buffer, 2 mmol/L $MgCl_2$, 0.1 mmol/L dNTPs, 0.25 mmol/L of each primer and 1 U Taq polymerase (Promega). Amplifications were carried out in a 2,720 thermocycler (Applied Biosystems). Cycling conditions consisted of an initial denaturation step at 94 °C for 1 min, followed by 35 amplification cycles at 94 °C for 1 min, 52 °C for 1 min and 72 °C for 1 min 30 s. The amplification products were visualized on 1 % agarose gel with ethidium bromide staining. Three PCR products of approximately 450 bp, amplified separately by primer combinations I, III and IV, were excised from the gels, purified using the QIAquick Gel Extraction kit (Qiagen, Hilden, Germany), cloned into the pGEM-T vector (Promega, Madison, USA) and transformed into competent *Escherichia coli* DH5α cells, in accordance with the supplier's protocol. Ten colonies to be sequenced, per primer combination, were picked randomly among recombinant (white) colonies. Plasmids were purified from colonies, using the kit PureLink™ Quick Plasmid Miniprep (Invitrogen), according to fabricant recommendations. Inserts were sequenced using T7 and SP6 primers of the vector. The sequencing reactions were performed according to the recommendations of the manufacturer, using the BigDye labeling sequencing reaction mixture (PerkinElmer), and sequences were read on an ABI Prism-310 Genetic Analyzer. Out of 30 clones sequenced, 12 found to be non-redundant were analyzed *in silico*. These clones were labeled 'SALn' (SAL: initials of 'Salambo 80'; n: a number).

## Sequence homology searching

Sequences were manually edited using BioEdit version 7.0.5.3 software (Hall 1999). Homologies of the sequences with available information from plants and other living systems were searched among all non-redundant protein and CDS translations (nr database) using Blastx 2.2.27 (Altschul et al. 1997), at the web page of the National Center of Biotechnology Information (NCBI; http://blast.ncbi.nlm.nih.gov/Blast.cgi).

## Multiple sequence alignment and phylogenetic inference methods

Deduced amino acid sequences showing homology with previously reported *R*-genes/RGAs were compared with 5 plant *R*-proteins sequences: *Lr21* of *T. aestivum* (ACO53397; Huang et al. 2009), rust resistance protein *Lr10* of *T. aestivum* (AAQ01784; Feuillet et al. 2003), CCN resistance protein of *T. aestivum* (ABY28270; Zhai et al. 2008), stripe rust resistance protein *Yr10* of *T. aestivum* (AF149114) and leaf rust resistance protein *Lr21 of Aegilops tauschii* (AAP74647; Huang et al. 2003). The final sequence set was manually cropped to the P-loop–FMHYAL region. Two distinct strategies were adopted for multiple sequence alignment and phylogenetic analysis. In the first, sequences were initially submitted to multiple expectation maximization for motif elicitation (MEME) program (Bailey and Elkan 1994; http://meme.nbcr.net/meme/) to extract conserved motifs, which were subsequently submitted to BLOCKS processor, to perform a blocks-based multiple sequence alignment constructed from the most highly conserved regions of proteins, avoiding misaligned ones. The generated tree is a 100 bootstrap neighbor-joining that was visualized by TreeView 1.6.6 (http://taxonomy.zoology.gla.ac.uk/rod/treeview.html). In the second strategy based on hidden markov model (HMM), sequences were submitted to Clustal Omega software version 1.1.0, accessed from the European Bioinformatics Institute (EBI) server (http://www.ebi.ac.uk/). This software uses seeded guide trees and HMM profile–profile techniques to generate alignments. The obtained alignment was downloaded in clustal format and used as input in DNAMAN 5.2.2, to perform a 1,000 bootstrap maximum likelihood tree with midpoint rooting, using the protein model of Dayhoff (Dayhoff et al. 1978).

## *In silico* mining of resistance gene analogs using the wheat genome draft assembly

The Cereals Data Base Web site (CerealsDB; http://www.cerealsdb.uk.net/) (Wilkinson et al. 2012) resulted from collaboration between the Universities of Bristol (UK) and Liverpool (UK) together with the John Innes Centre in Norwich (UK). The site which is maintained by members of the Functional Genomics Group at the University of Bristol provides a range of facilities for the study of the wheat genome. Besides access to molecular-marker datasets for wheat, such as SNPs and DArTs, CerealsDB contains a link to the draft genome sequence for the wheat variety Chinese Spring. The draft assembly of the gene-rich regions of the Chinese Spring genome or the raw sequence reads can be searched using BLAST, and a drop down field is available for setting the *e* value cutoff.

Each of the 'SALn' sequences, identified in the present study, was blasted against the draft assembly of the gene-rich regions of the Chinese Spring genome, using the 'BLAST' option available in CerealDB (http://www.cerealsdb.uk.net/CerealsDB/Documents/DOC_search_reads.php), with an *e* value cutoff of $10^{-5}$. The algorithm used was Megablast 2.2.23 (Zhang et al. 2000), and the searched database was the gene-rich region database (5xReference.fas), counting 5,321,847 sequences and 3,800,325,216 total letters. Contigs that produced significant (better than threshold) alignments, with a coverage rate encompassing the query full span, were used for annotation and gene prediction. Because the current version of the wheat genome (GenBank: CALP000000000.1; Brenchley et al. 2012) contains fragmentary contigs that have not been yet completely assembled nor assigned to chromosome locations, the contigs matching 'SALn' sequences were further assembled, using cap3 program (Huang and Madan 1999) accessible from CerealDB database, together with the wheat genome raw reads database (genome5.fas; 219,372,774 sequences; 85,117,182,478 total letters), in order to maximize viewing of the whole genomic region flanking the NBS domain.

The assembly procedure was as follows: the selected contig is submitted to a first round blast against the wheat raw reads. All matches are assembled using cap3 (Huang and Madan 1999) and a consensus contig is then generated and used for a new blast round against the database. Blast and assembling were repeated iteratively, enabling to walk up and downstream the initial contig, until it became no longer possible to extend the search further. Final 'extra-contigs' obtained were given the label 'CSn' (CS: initials of Chinese Spring; n: a number) and aligned to the initial seed sequences (SALn), using Megablast 2.2.28 (Zhang et al. 2000), from the NCBI site, with the parameter 'align two sequences.'

### Gene modeling and characterization of the identified gene models

Final extra-contigs, obtained through the gene mining procedure, were submitted to ab initio gene prediction, using the web version of FGENESH (http://www.softberry.com) (Salamov and Solovyev 2000), with parameters for monocot plants. In a first step toward their functional assignment, FGENESH-predicted proteins were blasted against NCBI non-redundant protein collection (nr), using Blastp 2.2.28 (Altschul et al. 1997), to search for homologous proteins. For refined characterization, we downloaded the November 2013 version of high-confidence (HCS) chromosome arm-assigned gene models of the International Whet Genome Sequencing Consortium (IWGSC) sequence assembly, including splice variants, from IWGSC repository (http://wheat-urgi.versailles.inra.fr/Seq-Repository/Genes-annotations) and used them as reference for comparison of our gene models. For this purpose, we used CD-HIT-2D algorithm of the CD-HIT clustering program (http://weizhong-lab.ucsd.edu/cdhit_suite/cgi-bin/index.cgi?cmd=Server%20home). Moreover, mapping of extra-contigs (CSn) was performed, using Blasn against IWGSC genome assembly scaffolds (http://tgac-browser.tgac.ac.uk/iwgsc_css/blast.jsp), to determine their most probable genomic location, at chromosome, sub-genome, arm and scaffold levels. Finally, when similarity with a functional gene product could be ascertained, we used FGENESH+ (http://www.softberry.com) to improve prediction accuracy based on similarity and checked InterProScan for domain conservation using Geneious 6.1.5 (http://www.geneious.com).

## Results

### PCR amplification and molecular cloning of RGA sequences

The PCR amplification conditions were successfully optimized using the degenerate primers indicated in M&M. Visual analysis of PCR products revealed the absence of consistent amplification products with primer combination II, while products ranging from ~200 to ~1,200 bp were obtained by primer combinations I, III and IV. Three fragments of approximately 450 bp (one per combination I, III and IV) were isolated and cloned. Out of 30 colonies sequenced (10 per ligation reaction), 12 were found to be non-redundant. Among these, seven (SAL6-12) were obtained with primer combination III, four (SAL1-4) with primer combination IV and one (SAL5) with primer combination I.

### Sequence homology with previously identified RGAs

Only five clones (SAL1, SAL7, SAL8, SAL10 and SAL11) showed similarities to cloned *R*-genes. Clone SAL1 showed the best homology with a putative disease resistance protein from *A. tauschii* (EMT21329.1). Secondary hits included

**Table 2** Blastx analysis of 12 clones isolated from 'Salambo 80,' obtained after amplification by degenerate primers corresponding to conserved motifs in the nucleotide-binding site (NBS) domain

| Primer combination | Clone | Blastx best hits (*e* value) (% similarity) | RF | QC |
|---|---|---|---|---|
| IV | SAL1 | EMT21329.1: Putative disease resistance protein RGA4 [*Aegilops tauschii*] (2$^{e}$-61) (87 %) | +1 | 100 % |
| | | AAM69850.1: NBS–LRR class RGA [*Aegilops tauschii*] (6e-60) (87 %) | | |
| | | ADK32523.1: wheat leaf rust resistance *Lr21* [*Triticum aestivum*] (8e-59) (87 %) | | |
| | | ACO53397.1: *Lr21* [*Triticum aestivum*] (1e-58) (87 %) | | |
| | | AAP74647.1: *Lr21* [*Aegilops tauschii*] (1e-58) (87 %) | | |
| | SAL2 | No significant similarity found | – | – |
| | SAL3 | No significant similarity found | – | – |
| | SAL4 | No significant similarity found | – | – |
| I | SAL5 | No significant similarity found | – | – |
| III | SAL6 | No significant similarity found | – | – |
| | SAL7 | EMT21329.1: Putative disease resistance protein RGA4 [*Aegilops tauschii*] (2$^{e}$-65) (83 %) | +2 | 99 % |
| | | ADK32523.1: wheat leaf rust resistance *Lr21* [*Triticum aestivum*] (1e-64) (85 %) | | |
| | | AAP74647.1 : *Lr21* [*Aegilops tauschii*] (1e-64) (85 %) | | |
| | | ACO53397.1 : *Lr21* [*Triticum aestivum*] (1e-64) (85 %) | | |
| | | AAM69850.1: NBS-LRR class RGA [*Aegilops tauschii*] (1e-63) (83 %) | | |
| | SAL8 | EMS60444.1: Putative disease resistance RPP13-like protein 1 [*Triticum urartu*] (4 10-40) (68 %) | +1 | 100 % |
| | | AFN70866.1 : NBS-type RGA protein, partial [*Bambusa multiplex*] (2e-37) (64 %) | | |
| | | EMT05289.1\|hypothetical protein F775_52304 [*Aegilops tauschii*] (6 10-37) (66 %) | | |
| | | CBY91999.1: NBS-LRR-resistant protein [*Saccharum* hybrid cultivar LCP 85-384] (1e-36) (61 %) | | |
| | | gb\|AEV59558.1\|RGA3, partial [*Triticum aestivum*] (3 10-36) (64 %) | | |
| | | gb\|EMT32739.1\|Putative disease resistance RPP13-like protein 1 [*Aegilops tauschii*] (6 10-36) (67 %) | | |
| | | gb\|AEV59557.1\| RGA1 [*Triticum aestivum*] (10-35) (67 %) | | |
| | | AAC05834.2: Cyst nematode resistance gene candidate-like protein [*Aegilops tauschii*] (1e-35) (67 %) | | |
| | | ABY28270.1: Cereal cyst nematode resistance protein [*Triticum aestivum*] (1e-35) (67 %) | | |
| | SAL9 | No significant similarity found | – | – |
| | SAL10 | EMT21329.1: Putative disease resistance protein RGA4 [*Aegilops tauschii*] (2$^{e}$-61) (88 %) | +1 | 98 % |
| | | AAM69850.1: NBS-LRR class RGA [*Aegilops tauschii*] (3e-58) (88 %) | | |
| | | ADK32523.1: wheat leaf rust resistance *Lr21* [*Triticum aestivum*] (3e-57) (88 %) | | |
| | | ACO53397.1: *Lr21* [*Triticum aestivum*] (3e-57) (88 %) | | |
| | | AAP74647.1: *Lr21* [*Aegilops tauschii*] (3e-57) (88 %) | | |
| | SAL11 | CBY91999.1: NBS-LRR resistant protein [*Saccharum* hybrid cultivar LCP 85-384] **(4e-12) (62 %)** | +3 | 73 % |
| | SAL12 | No significant similarity found | – | – |

*RF* reading frame, *QC* query coverage

NBS–LRR RGA from *A. tauschii* (AAM69850.1) and leaf rust resistance *Lr21* from *T. aestivum* (ADK32523.1, ACO53397.1) and *A. tauschii* (AAP74647). For all these five matches, alignment covered full query length (100 %), with a similarity rate of 87 % (Table 2). Clones SAL7 and SAL10 aligned to the same best five matches as SAL1. Clone SAL8 showed sequence similarity to 7 NBS-type RGA proteins from *T. urartu* (EMS60444.1), *Bambusa multiplex* (AFN70866.1) *A. tauschii* (EMT05289.1, EMT32739.1), *Saccharum* (CBY91999.1) and *T. aestivum* (AEV59558.1, AEV59557.1) in addition to two function-assigned sequences: a Cyst nematode resistance gene candidate-like protein from *A. tauschii* (AAC05834.2) and the CCN resistance protein from *T. aestivum* (ABY28270.1). For all hits matching SAL8, it was noted that amino acid similarity rates were lower than 70 %, being comprised in the interval 61–68 %. Finally, clone SAL11 showed best alignment with NBS–LRR-resistant protein from *Saccharum* hybrid cultivar LCP 85–384 (CBY91999), but the alignment covered only 73 % of SAL11, from the P-loop GGVGKTT until the FSNLDTLQMKLE region, eight amino acids upstream the Kinase-2 motif (LLV-LDDV). Based on Blastx results, it can be concluded that primer combination III was the most efficient in amplifying

**Table 3** Clones showing similarity to previously described RGAs and sequence full-length open reading frames (ORFs), among 12 clones sequenced using three degenerate primer combinations

| Clone (size) | GenBank nucleotide accession no. | Potential ORF (frame) | GenBank protein ID |
|---|---|---|---|
| SAL1 (435 bp) | JX566982.1 | 145 aa (+1) | AFV73210.1 |
| SAL7 (434 bp) | JX566983.1 | 144 aa (+2) | AFV73211.1 |
| SAL8 (450 bp) | JX566984.1 | 150 aa (+1) | AFV73212.1 |
| SAL10 (433 bp) | JX566985.1 | 144 aa (+1) | AFV73213.1 |

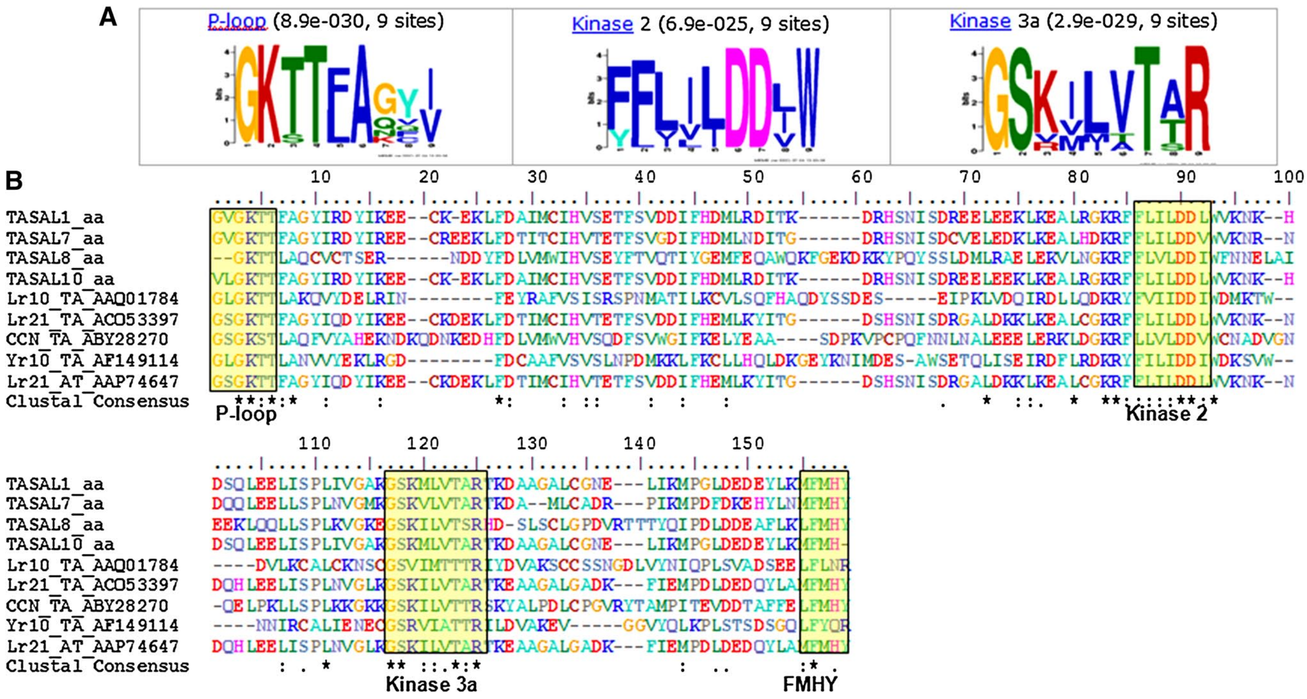


**Fig. 1** MEME discovered motifs with corresponding *e* values and logos (**a**); and HMM-based alignment of amino acid sequences of the 4 in-frame uninterrupted resistance gene analogs found in this study, with other available *R*-genes, from *Triticum aestivum* (AAQ01784, ACO53397, ABY28270 and AF149114) and *Aegilops tauschii* (AAP74647) (**b**). *Dashes* indicate the gaps introduced by Clustal Omega. Major motif signatures of the NBS domain are in *shaded boxes*. Consensus positions are indicated in the last *line*

RGA sequences, with four sequences showing homologies in the database.

## Sequence alignments and phylogenetic analyses

Clones SAL1, SAL7, SAL8 and SAL10, containing uninterrupted ORFs across the whole nucleotide span, were considered as potential coding RGAs and were registered in GenBank database under accession numbers JX566982–JX566985 (Table 3). These clones were used for the phylogenetic analysis, while SAL11 that was frame-shifted and contained a stop in position 260 was discarded. Clones SAL1, SAL7, SAL8 and SAL10 were, therefore, translated into polypeptides and aligned with the NBS domains of five plant *R*-genes indicated in M&M. Although conservation of motifs could be noticed directly in the P-loop, kinase-2, kinase-3a and FMYHAL domains, it was further confirmed by MEME analysis (Fig. 1). Phylogenetic analyses based on both neighbor joining and maximum likelihood methods were consistent in that they clearly revealed that SAL1, SAL7 and SAL10 were close to *Lr21* proteins ACO53397 and AAP74647, while SAL8 was relatively close to CCN-R (ABY28270) (Fig. 2).

## Identified RGAs served as initial seeds for *R*-gene mining from the wheat genome draft

Each of SAL1, 7, 8 and 10 was, separately, used as initial seed, to perform Megablast (*e* value = $10^{-5}$) against the gene-rich region database (5xReference.fas). In this initial Blastn round, SAL1 and SAL10 matched the same 13 contigs; SAL7 matched 7 contigs; and SAL8 matched only 2 contigs (data not shown). Among 13 contigs presenting sequence match

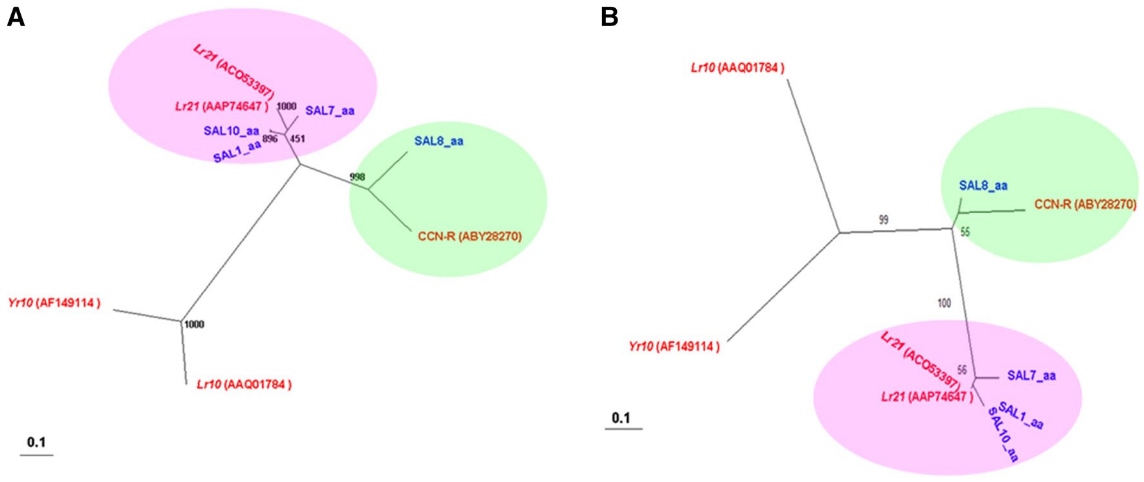


**Fig. 2** Maximum likelihood (**a**) and neighbor-joining (**b**) protein trees, derived from comparing predicted amino acid sequences of 4 NBS RGAs cloned from 'Salambo 80' with NBS domains of 4 resistance genes from wheat (*Lr21*, *Lr10*, *CCN* and *Yr10*) and one from *Aegilops tauschii* (*Lr21*). Per thousand (**a**) or percent (**b**) occurrences during bootstrap analysis (1,000 and 100 cycles, respectively) are indicated at nodes. *Ellipses* refer to major clusters inferred from both trees: a first cluster including SAL1, SAL7 and SAL10 that were close to *Lr21*; and a second cluster including SAL8 and CCN-R

with SAL1/SAL10, only 7 were considered interesting, as they matched the query sequence along its whole span (complete query coverage). With each of SAL7 and SAL8, only one matching contig presented complete query coverage (Table 4). Each of these contigs was assembled with wheat raw reads from database genome5.fas, using cap3 program in order to maximize their lengths. The procedure resulted in longer contigs that were labeled CSn. Tenting to further assemble extra-contigs CS1 to CS8, using cap3 sequence assembly program (Huang and Madan 1999; http://pbil.univ-lyon1.fr/cap3.php), revealed that they did not overlap. Pairwise nucleotide sequence alignments were performed between initial seed sequences (SALn) and the corresponding extra-contigs that were generated by Blast and assembly. Similarities were produced between plus/plus or plus/minus strands, with *e* values ranging between 0.0 and 1e-111. Dot plot similarity matrices were generated and are shown in Fig. 3.

### *Ab initio* gene prediction

FGENESH ab initio gene prediction determined one potential gene per sequence for CS1 to CS8 (Fig. 4). These potential genes were named by adding the extension −1.1 to the locus name (CSn extra-contig). The number of exons was variable, ranging from one to seven, with an average number of 3.75. Putative genes were predicted in both direct and reverse chains. Among eight predicted genes, only two (CS3-1.1 and CS6-1.1) had a complete ORF structure extending from ATG to Stop codon; three were truncated in 3′ end of coding chain (CS1-1.1, CS2-1.1 and CS8-1.1) and three truncated in 5′ end (CS4-1.1, CS5-1.1 and CS7-1.1).

### GenBank survey defined at least two groups within predicted gene models

Hypothetical gene products (either complete or partial) CS1-1.1 through CS8-1.1 were individually blasted against NCBI non-redundant protein collection (nr), using Blastp 2.2.28 (Altschul et al. 1997). Results indicated that these gene models could be divided into two major groups based on homology with GenBank contents: (1) The first group included gene models predicted from extra-contigs CS1 to CS7 that matched similar hits in the GenBank protein database. All seven queries CS1-1.1 to CS7-1.1 matched at least one *Lr21* protein variant within the top hits. These *Lr21* variants were: *Lr21* [*T. aestivum*] (ACO53397.1), *Lr21* [*A. tauschii*] (AAP74647.1) and wheat leaf rust resistance *Lr21* [*T. aestivum*] (ADK32523.1) (Table 5). (2) The second group was represented by the unique partial gene model CS8-1.1 (252 aa) that matched several *Triticeae* and non-*Triticeae* sequences, corresponding either to undefined RGAs or to *R*-protein-like sequences from *Arabidopsis thaliana* and *A. tauschii*. For CS8-1.1, all similarities were lower than 70 % for the top listed matches (Table 5), strongly suggesting that CS8-1.1 would correspond to an RGA of unknown function.

### Mapping to chromosome arms and refined characterization of gene models using IWGSC sequence data

CD-HIT clustering with IWGSC gene models revealed that all gene models predicted in the present study were unique at the similarity threshold of 90 %. However, when DNA extra-contigs CS1–CS8 were blasted against IWGSC wheat scaffolds, significant hits were obtained across all analyzed loci. High

**Table 4** *In silico* identification of extra-contigs by blast of the initial seed sequences (SALn) against the wheat gene-rich regions database, followed by multi-round blast/assembly of the produced contigs against the wheat raw reads

| Seed sequence | Hit contig[a] | Accession no.[b] | Alignment details | Iterations[c] | Resulting extra-contigs |
|---|---|---|---|---|---|
| SAL1/SAL10 | >contig45974<br>(2,953 letters) | CALP010045974 | Expect = 0.0; strand = plus/plus | 4 | CS1<br>4,136 letters |
| | >contig03136<br>(5,341 letters) | CALP01005341 | Expect = e−120; strand = plus/plus | 8 | CS2<br>7,653 letters |
| | >contig26570<br>(3,390 letters) | CALP010026570 | Expect = e−110; strand = plus/plus | 6 | CS3<br>5,642 letters |
| | >contig576781<br>(1,307 letters) | CALP0100576781 | Expect = e−107; strand = plus/plus | 5 | CS4<br>3,551 letters |
| | >contig339665<br>(1,605 letters) | CALP0100339665 | Expect = 7e−98; strand = plus/minus | 1 | CS5<br>2,115 letters |
| | >contig15877<br>(3,812 letters) | CALP010015877 | Expect = 3e-97; strand = plus/minus | 7 | CS6<br>7,022 letters |
| | >contig06947<br>(4,561 letters) | CALP010006947 | Expect = 8e−73; strand = plus/plus | 1 | CS7<br>4,959 letters |
| SAL7 | >contig45974<br>(2,953 letters) | CALP010045974 | Expect = 0.0; Strand = Plus/Plus | 4 | CS1<br>4,136 letters |
| SAL8 | >contig194172<br>(1,955 letters) | CALP0100194172 | Expect = e−136; strand = plus/plus | 4 | CS8<br>2,873 letters |

[a] Contigs with *e* value better than threshold ($10^{-5}$) and with full query coverage, matching the seed sequence, after blastn against gene-rich regions database (5xReference.fas), (CerealsDB; http://www.cerealsdb.uk.net/). The contig labels correspond to those used in database (5xReference.fas) of CerealsDB

[b] NCBI GenBank accession number of the selected hit contig

[c] Number of successive blastn iterations (*e* value $10^{-10}$) against raw reads and cap3 assembly

Blastn scores were comprised between 1.396e+04 (for CS2) and 3,494 (for CS5), probably because of its comparatively short size. All *e* values were 0.0, and query coverage rates varied between 66 % (for CS4) and 100 % (for CS2, CS3, CS5) (Table 6). Five loci (CS1, 2, 4, 5 and 7) matched scaffolds on chromosome arm 1BS; whereas CS6 locus was mapped on 1AS, CS3 on 1DS and CS8 on 4BS. Thus, even though GenBank investigation showed that gene models predicted from extra-contigs CS1–CS7 were all, to a certain extent, similar to *Lr21*, only CS3 would be structurally and functionally associated with this gene occurring at the distal end of chromosome 1DS of *T. aestivum* (Huang et al. 2003). Because only few *R*-genes from cereals have been cloned to date and are accessible through databases for comparison, the hypothetical functions of gene models CS1, CS2 and CS4–CS8, identified in this study, remain unsolved. For CS3, a more accurate 960 aa gene model was triggered by FGENESH+, based on similarity with *Lr21*, and was designated CS3-1.2 (Fig. 5a). Interproscan annotation of this hypothetical protein evidenced the presence of NBS and LRR domains (Fig. 5b).

## Discussion

In this study, we have used a PCR-based technique to isolate four novel clones containing NBS-like sequences from the bread wheat 'Salambo 80.' These RGAs have been isolated by means of conserved motifs used by Maleki et al. (2003) that differ from those previously used in wheat by Seah et al. (1998), Chen et al. (1998) and Spielmeyer et al. (1998). The amplified region spans from the P-loop to the FMYHAL motif, 22 amino acids upstream the GLPLA one, which is not common in cereals (Maleki et al. 2003). Among four primer combinations used, three have produced amplicons in the size interval 300–500 bp, corresponding to the NBS coding region size. Three of these PCR products were selected for the subsequent cloning. Size-based selection of PCR products is the same procedure that was used by Maleki et al. (2003), but differ from that adopted by Bozkurt et al. (2007), who have cloned PCR products of varying sizes from *T. aestivum*, *T. monococcum* and *T. dicoccoides*. Bozkurt et al. (2007) cloned fragments sized more than 500 bp, to enable identifying intronic regions within NBS putative ORF sequences. In our study, out of three primer combinations generating PCR products, primer combination III was the most successful in amplifying NBS sequences, with four sequences showing homologies in the GenBank nr database. This could be explained by the fact that primer NBSfor3 was less stringent than -forI and -forIV, as it contained 4 undefined nucleotide base sites (3 N and 1 H). This observation explains 99 % of the sequence difference between primers NBSfor3 and -for1.

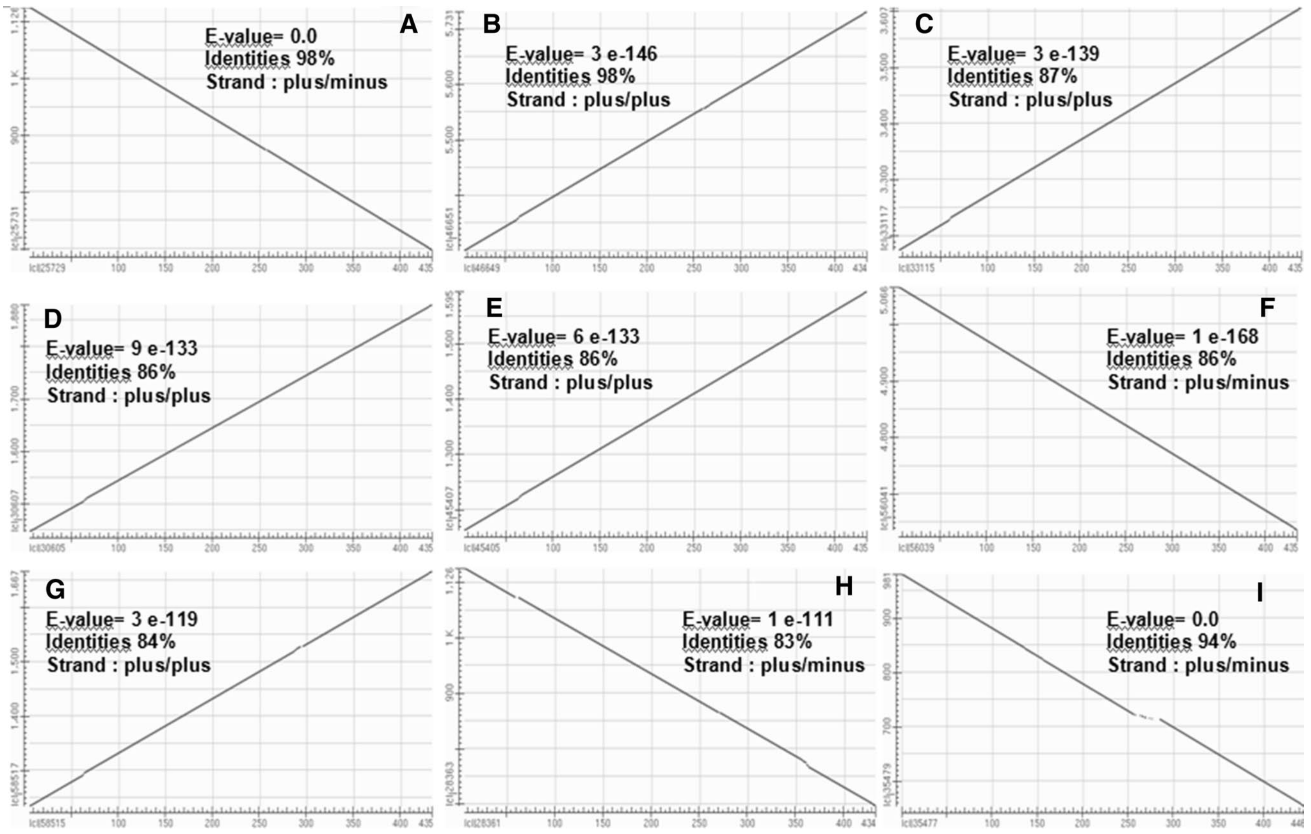


**Fig. 3** *Dot matrix* views generated by BLASTN megablast, illustrating similarities in local alignments between seed sequences (SALn) and extra-contigs assembled using the wheat draft genome (CSn). **a** SAL1 versus CS1, **b** SAL1 versus CS2, **c** SAL1 versus CS3, **d** SAL1 versus CS4, **e** SAL1 versus CS5, **f** SAL1 versus CS6, **g** SAL1 versus CS7, **h** SAL7 versus CS1, **i** SAL8 versus CS8. CSn are extra-contigs identified in the *Triticum aestivum* Chinese spring genome, based on a blast/assembly iterative procedure, starting from SALn sequence

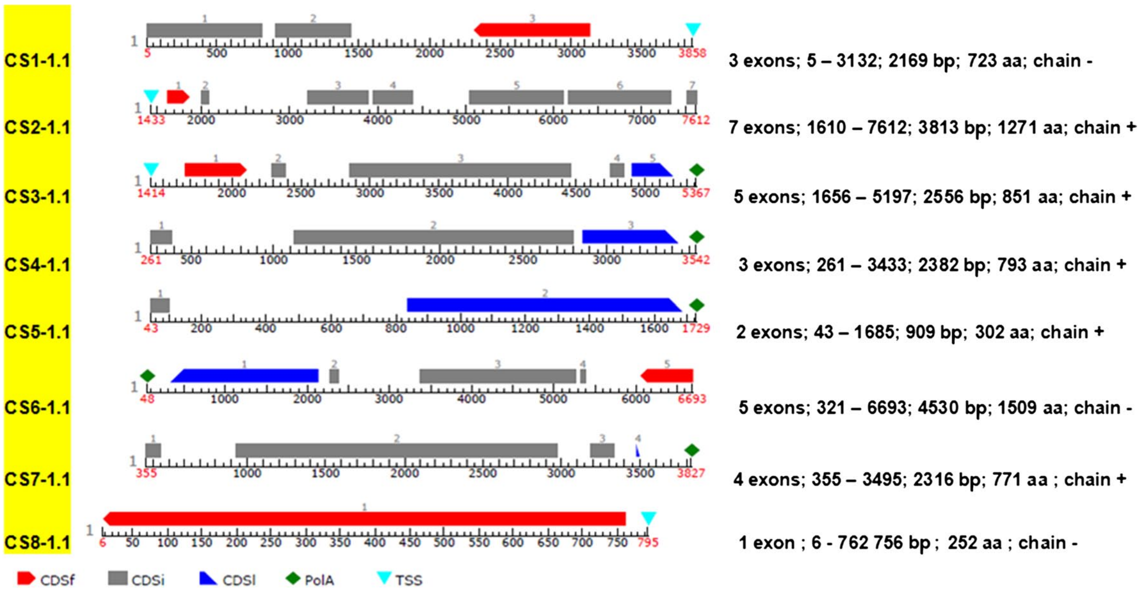

CS1-1.1
3 exons; 5 – 3132; 2169 bp; 723 aa; chain -
CS2-1.1
7 exons; 1610 – 7612; 3813 bp; 1271 aa; chain +
CS3-1.1
5 exons; 1656 – 5197; 2556 bp; 851 aa; chain +
CS4-1.1
3 exons; 261 – 3433; 2382 bp; 793 aa; chain +
CS5-1.1
2 exons; 43 – 1685; 909 bp; 302 aa; chain +
CS6-1.1
5 exons; 321 – 6693; 4530 bp; 1509 aa; chain -
CS7-1.1
4 exons; 355 – 3495; 2316 bp; 771 aa ; chain +
CS8-1.1
1 exon ; 6 - 762 756 bp ; 252 aa ; chain -
CDSf
CDSi
CDSl
PolA
TSS

**Fig. 4** Complete and partial gene models predicted ab initio from extra-contigs CS1–CS8, using FGENESH. *CDSf* first coding segment (starting with start codon), *CDSi* internal coding segment (internal exon), *CDSl* last coding segment (ending with stop codon), *TSS* position of transcription start (TATA-box position and score), *PolA* polyadenylation signal sequence (AATAAA)

**Table 5** Summary of Blastp results of predicted gene products, CS1-1.1 through CS8-1.1, against NCBI non-redundant protein collection (nr)

| | Gene model | Top matches[a] | GenBank accession | *e* value | Similarity (%) |
|---|---|---|---|---|---|
| Group 1: Lr-21 like | CS1-1.1 | Putative disease resistance protein RGA4 [*Aegilops tauschii*] | EMT21329.1 | 0.0 | 72 |
| | | ***Lr21* [*Aegilops tauschii*]** | AAP74647.1 | 0.0 | 70 |
| | | **Wheat leaf rust resistance *Lr21* [*Triticum aestivum*]** | ADK32523.1 | 0.0 | 70 |
| | | ***Lr21* [*Triticum aestivum*]** | ACO53397.1 | 0.0 | 70 |
| | CS2-1.1 | **Wheat leaf rust resistance *Lr21* [*Triticum aestivum*]** | ADK32523.1 | 0.0 | 62 |
| | | ***Lr21* [*Aegilops tauschii*]** | AAP74647.1 | 0.0 | 62 |
| | | ***Lr21* [*Triticum aestivum*]** | ACO53397.1 | 0.0 | 62 |
| | CS3-1.1 | ***Lr21* [*Triticum aestivum*]** | ACO53397.1 | 0.0 | 90 |
| | CS4-1.1 | NBS-LRR class RGA [*Aegilops tauschii*] | AAM69850.1 | 0.0 | 82 |
| | | Putative disease resistance protein RGA4 [*Aegilops tauschii*] | EMT21329.1 | 0.0 | 83 |
| | | **Wheat leaf rust resistance *Lr21* [*Triticum aestivum*]** | ADK32523.1 | 0.0 | 77 |
| | | ***Lr21* [*Aegilops tauschii*]** | AAP74647.1 | 0.0 | 77 |
| | | ***Lr21* [*Triticum aestivum*]** | ACO53397.1 | 0.0 | 77 |
| | CS5-1.1 | Putative disease resistance protein RGA4 [*Aegilops tauschii*] | EMT21329.1 | 2e-151 | 85 |
| | | ***Lr21* [*Triticum aestivum*]** | ACO53397.1 | 3e-134 | 81 |
| | CS6-1.1 | Putative disease resistance protein RGA1 [*Aegilops tauschii*] | EMT21925.1 | 0.0 | 89 |
| | | Putative disease resistance protein RGA4 [*Aegilops tauschii*] | EMT21329.1 | 0.0 | 70 |
| | | ***Lr21* [*Aegilops tauschii*]** | AAP74647.1 | 0.0 | 73 |
| | | **Wheat leaf rust resistance *Lr21* [*Triticum aestivum*]** | ADK32523.1 | 0.0 | 73 |
| | | ***Lr21* [*Triticum aestivum*]** | ACO53397.1 | 0.0 | 73 |
| | CS7-1.1 | NBS-LRR class RGA [*Aegilops tauschii*] | AAM69850.1 | 0.0 | 68 |
| | | Putative disease resistance protein RGA4 [*Aegilops tauschii*] | EMT21329.1 | 0.0 | 73 |
| | | Putative disease resistance protein RGA1 [*Aegilops tauschii*] | EMT21925.1 | 0.0 | 66 |
| | | ***Lr21* [*Aegilops tauschii*]** | AAP74647.1 | 0.0 | 73 |
| | | **Wheat leaf rust resistance *Lr21* [*Triticum aestivum*]** | ADK32523.1 | 0.0 | 73 |
| | | ***Lr21* [*Triticum aestivum*]** | ACO53397.1 | 0.0 | 73 |
| Group 2: no similarity with functional products | CS8-1.1 | Putative disease resistance RPP13-like protein 1 [*Aegilops tauschii*] | EMT23222.1 | 1e−79 | 69 |
| | | DW-RGA2 protein [*Triticum durum*] | CAD12796.1 | 3e−79 | 68 |
| | | hypothetical protein OsI_36936 [*Oryza sativa* Indica Group] | EEC68588.1 | 4e−78 | 67 |
| | | RGA3, partial [*Triticum aestivum*] | AEV59558.1 | 6e−78 | 68 |
| | | NB-ARC domain-containing protein, expressed [*Oryza sativa* Japonica Group] | ABA95210.1 | 6e−77 | 67 |

[a] Hits corresponding to Lr21 protein variants are written in bold

**Table 6** Results of Blastn between loci CS1–CS8, identified in this study, and IWGSC genome survey scaffolds, accessed through TGAC's bread wheat BLAST server (http://tgac-browser.tgac.ac.uk/iwgsc_css/blast.jsp)

| Query locus (nucleotide size) | Subject id | % Identity | Alignment length | Mismatches | q.start | q.end | s.start | s.end | *e* value | Bit score |
|---|---|---|---|---|---|---|---|---|---|---|
| CS1 (4,136) | IWGSC_CSS_1BS_scaff_3456702 | 99.83 | 3,624 | 3 | 513 | 4,136 | 7,312 | 3,692 | 0.0 | 6,656 |
| CS2 (7,653) | IWGSC_CSS_1BS_scaff_3433871 | 99.60 | 7,664 | 5 | 1 | 7,653 | 5,422 | 13,070 | 0.0 | 1.396e+04 |
| CS3 (5,642) | IWGSC_CSS_1DS_scaff_1915820 | 99.75 | 5,648 | 0 | 1 | 5,642 | 6,917 | 1,278 | 0.0 | 1.034e+04 |
| CS4 (3,551) | IWGSC_CSS_1BS_scaff_3450711 | 98.80 | 2,341 | 20 | 177 | 2,516 | 2,334 | 1 | 0.0 | 4,161 |
| CS5 (2,115) | IWGSC_CSS_1BS_scaff_3450711 | 96.65 | 2,120 | 42 | 1 | 2,115 | 2,224 | 129 | 0.0 | 3,494 |
| CS6 (7,022) | IWGSC_CSS_1AS_scaff_3255097 | 96.02 | 6,987 | 248 | 45 | 7,022 | 8,615 | 1,650 | 0.0 | 1.134e+04 |
| CS7 (4,959) | IWGSC_CSS_1BS_scaff_1321696 | 99.92 | 4,855 | 0 | 1 | 4,852 | 4,854 | 1 | 0.0 | 8,940 |
| CS8 (2,873) | IWGSC_CSS_4BS_scaff_4887885 | 97.63 | 2,869 | 57 | 1 | 2,866 | 3,025 | 5,885 | 0.0 | 4,911 |

q.start and q.end designate the query range covered by alignment; s.start and s.end designate the subject range covered by alignment

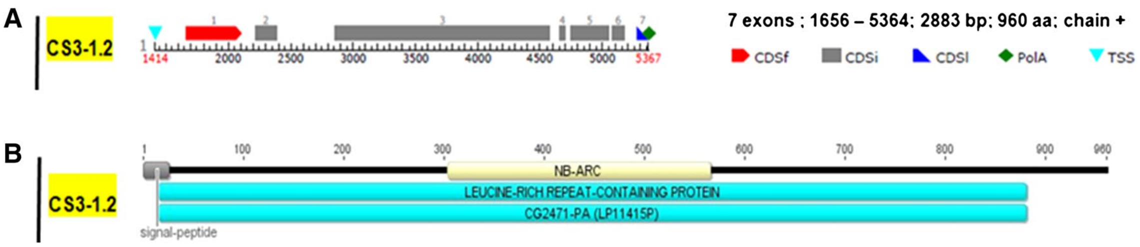


**Fig. 5** Exon/intron and domain structures of CS3-1.2, an *Lr21*-like gene model predicted in this study. **a** Exon/intron structure. *CDSf* first (starting with start codon), *CDSi* internal (internal exon), *CDSl* last coding segment (ending with stop codon), *TSS* position of transcription start (TATA-box position and score). **b** InterProScan annotation of conserved domains

Among 12 non-redundant clones analyzed, in our study, and blasted against GenBank amino acid database, only five had homologs among other plant RGAs. The amplification of the remaining sequences could be explained either by non-specific amplification or by the integration of transposable elements and viral sequences into *R*-genes, generating pseudogenes (Bozkurt et al. 2007). Among the identified RGAs, one (SAL11) was removed from the subsequent analysis because it contained an interrupted ORF. For the four uninterrupted-ORF RGAs reported in our study, Blastx hits were not perfectly aligned, signifying that the clones were unique. The similarity rates between queries and top hits ranged between 61 and 88 %. Each of SAL1, SAL7 and SAL10 was aligned to resistance proteins AAP74647.1, ADK32523.1 and ACO53397.1, all representing variants of the leaf rust resistance *Lr21* protein in *T. aestivum* and/or *A. tauschii*. In contrast, SAL8 matched two sequences of known function: ABY28270.1 and AAC05834.2, both corresponding to CCN resistance protein. But, interestingly, identity rates and *e* values were relatively weak (Table 2). Therefore, we think that SAL8 association with CCN resistance lacks strong evidence, making possible the fact that this RGA would be involved in a different resistance function.

Sequence comparisons were made with five cereal NBS–LRR genes: *T. aestivum Lr21* (ACO53397), *T. aestivum Lr10* (AAQ01784), *T. aestivum CCN-R* (ABY28270), *T. aestivum Yr10* (AF149114) and *A. tauschii Lr21* (AAP74647). As 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 (Du Preez 2005), this often complicates accurate alignment of multiple sequences in regions stretching between conserved motifs, which in turn negatively impacts motif alignment. Thus, sequence alignment was performed using two alignment strategies, allowing amplifying informative positions, even at high levels of sequence divergence. These two techniques were block-based and HMM-based alignments. The first technique enables improving actual phylogenetic topology by keeping only the less ambiguous aligned blocks within protein sequence alignments (Talavera and Castresana 2007). HMM-based alignments are also much faster than pair-wise methods (Thompson et al. 1994) and are very accurate and particularly adequate for aligning members of a protein family. Both methods generated similar topologies in the phylogenetic analysis, revealing that SAL1, SAL7 and SAL10 were most similar to *Lr21*, whereas SAL8 was clustered with CCN-R.

In 2012, the International Wheat Genome Sequencing Consortium (IWGSC http://www.wheatgenome.org/) has developed a first draft survey sequence of wheat chromosome arms. Gene models were developed and 93 % of them were assigned at chromosomal and sub-genome levels, revealing new insights on duplication, evolution and transcriptional activity of the genome (Rogers 2014). In addition to the IWGSC survey, the UK 454 survey of bread wheat sequence and genome analysis has been recently published (Brenchley et al. 2012), and a number of associated conceptual databases and web interfaces have been established to browse the wheat genome for useful information, such as physical mapping, genetic markers development, comparative genomics and annotation purposes. In the present study, we have used the CerealsDB Web site (http://www.cerealsdb.uk.net), in order analyze the genomic regions encompassing the RGAs identified in our study. Such a procedure based on the exploration of the RGA-delimited genomic regions aims to tentatively identify *R*-genes among these regions, using *in silico* analysis. This approach could help to validate their possible role in disease resistance. Eight extra-contigs (CS1–CS8), ranging in size between 2,115 and 7,653 bp, were identified in the genome of *T. aesti*vum strain 'Chinese Spring,' based on RGAs SAL1, SAL7, SAL8 and SAL10, first identified in vitro, then used as initial queries, *in silico*. *Ab initio* gene prediction determined eight potential genes, among which two (CS3-1.1 and CS6-1.1) had a complete ORF structure. Blastp against GenBank proteins and CDS revealed the existence of at least two groups within gene models: a first group including *Lr21*-like gene models and a second group represented by the unique partial gene model CS8-1.1 (252 aa) that was not similar to any functionally assigned sequence. These results give support to the phylogenetic analysis, suggesting that the four RGAs identified in this study belong to at least two probable distinct families. Subsequently, chromosome arm-mapping of the 8 loci CS1–CS8, using Blastn against IWGSC scaffolds, shed more light on their potential associations. Indeed, only CS3 locus was found to be located on chromosome arm 1DS of *T. aestivum* where *Lr21* gene conferring resistance to the fungus *Puccinia triticina* (Eriks) is known to occur (Huang et al. 2003). The remaining loci identified in our study were found on chromosome arms 1AS, 1BS and 4BS, and therefore could be involved in resistance to different pathogens and pests. For instance, genes *H5*, *H9*, *H10* and *H11* that confer to wheat resistance to the Hessian fly form a linkage group on wheat chromosome arm 1AS (Liu et al. 2005), and a molecular marker associated with *H11* was previously identified in 'Salambo 80,' the cultivar used in the present study (Bouktila et al. 2006).

Even though the *Lr21* gene has a single copy locus, it is characterized by an extensive allelic diversity. Huang et al. (2009) demonstrated that *Lr21* locus presents a wide spectrum of accumulated variations including single-nucleotide polymorphisms (SNPs), non-synonymous substitutions and indels. However, all *Lr21* variants different from the wild allele (1.36 kb, 1,080 aa) were found not to confer resistance to *P. triticina* at seedling stage and were, therefore, considered as pseudogenes (Huang et al. 2009). Based on these reports, it can be inferred that the gene model CS3-1.2, identified *in silico* in our study, would probably represent an *Lr21* pseudogene and would be paralogous to the *Lr21* functional allele locus. Huang et al. (2009) reported that the non-functional paralogs (*lr21*) were NBS–LRR sequences with at least 80 % identity to *Lr21* in the NBS region and 50 % identity in the rest of the gene. These reports are compatible with our findings as SAL1 and SAL10 matched *Lr21* with, respectively, 87 and 88 % of similarity (Table 2), while the full gene model that was derived from them (CS3-1.1) matches *Lr21* with 90 % of similarity (Table 5). The probable, expressed *Lr21* pseudogene CS3-1.2 could regulate the coding gene expression by competing for microRNA binding (Poliseno et al. 2010) and/or may be preventing the degradation of the homologous functional *R*-gene by the local silencing system (Lozano et al. 2012).

In conclusion, four novel wheat NBS domain-containing sequences were identified in the present study. Using them to blast the wheat genome led to the identification of 8 RGA models, among which we identified an NBS-LRR class *Lr21* pseudogene and assigned all of them to mapped scaffolds. Herein, reported results will provide a genomic framework for further isolation of candidate NBS-encoding genes in wheat and, hopefully, contribute to studies of wheat disease *R*-genes, especially considering the limited number of cloned *R*-genes in cereals and the current draft status of *T. aestivum* genome.

**Acknowledgments** We gratefully acknowledge Abdennour Sébéi (Centre Régional de Recherches sur les Grandes Cultures, Béja, Tunisia) for providing bread wheat cultivar 'Salambo 80'; Jacques-Deric Rouault (Laboratoire Evolution, Génome et Spéciation, CNRS, Gif sur Yvette, France) for providing training on data mining techniques; Paul A. Wilkinson (School of Biological Sciences, University of Bristol, UK) for providing help about CerealsDB 2.0.; and Abdelkader Aïnouche (UMR-CNRS Ecobio, Université de Rennes-1, France) for reading the manuscript. This study was financially supported by the Tunisian Ministry of Higher Education and Scientific Research.

**Conflict of interest** None.

## References

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucl Acids Res 25:3389–3402

Bai J, Pennill LA, Ning J, Lee SW, Ramalingam J, Webb CA, Zhao B, Sun Q, Nelson JC, Leach JE, Hulbert SH (2002) Diversity in nucleotide binding site-leucine-rich repeat genes in cereals. Genome Res 12:1871–1884

Bailey TL, Elkan C (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In: Proceedings of the second international conference on intelligent systems for molecular biology. AAAI Press, Menlo Park, pp 28–36

Boissot N, Thomas S, Sauvion N, Marchal C, Pavis C, Dogimont C (2010) Mapping and validation of QTLs for resistance to aphids and whiteflies in melon. Theor Appl Genet 121:9–20

Bouktila D, Mezghani M, Marrakchi M, Makni H (2005) Identification of wheat sources resistant to Hessian fly, *Mayetiola destructor* (Diptera: Cecidomyiidae), in Tunisia. Int J Agric Biol 7:799–803

Bouktila D, Mezghani M, Marrakchi M, Makni H (2006) Characterization of wheat random amplified polymorphic DNA markers associated with the *H11* hessian fly resistance gene. J Integr Plant Biol 48:958–964

Bouktila D, Kharrat I, Mezghani-Khemakhem M, Makni H, Makni M (2012) Preliminary identification of sources of resistance to the greenbug, *Schizaphis graminum* Rondani (*Hemiptera: Aphididae*) among a collection of Tunisian bread wheat lines. Rom Agric Res 29:115–120

Bozkurt O, Hakki EE, Akkaya MS (2007) Isolation and sequence analysis of wheat NBS–LRR type disease resistance gene analogs using degenerate PCR primers. Biochem Genet 45:469–486

Brenchley R, Spannagl M, Pfeifer M, Barker GLA, D'Amore R, Allen AM, McKenzie N, Kramer M, Kerhornou A, Bolser D, Kay S, Waite D, Trick M, Bancroft I, Gu Y, Huo N, Luo MC, Sehga S, Gill B, Kianian S, Anderson O, Kersey P, Dvorak J, McCombie WR, Hall A, Mayer KFX, Edwards KJ, Bevan MW, Hall N (2012) Analysis of the breadwheat genome using whole-genome shotgun sequencing. Nature 491:705–710

Chen XM, Line RF, Leung H (1998) Genome scanning for resistance-gene analogs in rice, barley, and wheat by high-resolution electrophoresis. Theor Appl Genet 97:345–355

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

Dayhoff M, Schwartz R, Orcutt B (1978) A model of evolutionary change in protein. Atlas Protein Seq Struct 5:345–352

de Wit PJGM (2007) How plants recognize pathogens and defend themselves. Cell Mol Life Sci 64:2726–2732

Deng Z, Huang S, Ling P, Chen C, Yu C, Weber CA, Moore GA, Gmitter FG Jr (2000) Cloning and characterization of NBS–LRR class resistance-gene candidate sequences in citrus. Theor Appl Genet 101:814–822

Deslandes L, Olivier J, Peeters N, Feng DX, Khounlotham M, Boucher C, Somssich I, Genin S, Marco Y (2003) Physical interaction between RRS1-R, a protein conferring resistance to bacterial wilt, and PopP2, a type III effector targeted to the plant nucleus. Proc Natl Acad Sci USA 100:8024–8029

DeYoung BJ, Innes RW (2006) Plant NBS–LRR proteins in pathogen sensing and host defense. Nat Immunol 7:1243–1249

Dilbirligi M, Gill KS (2003) Identification and analysis of expressed resistance gene sequences in wheat. Plant Mol Biol 53:771–787

Dilbirligi M, Erayman M, Sandhu D, Sidhu D, Gill KS (2004) Identification of wheat chromosomal regions containing expressed resistance genes. Genetics 166:461–481

Doyle JJ, Doyle JL (1987) A rapid DNA isolation procedure for small quantities of fresh leaves tissue. Phytochem Bull 19:11–15

Du Preez FB (2005) Tracking nucleotide-binding-site-leucine-rich-repeat resistance gene analogues in the wheat genome complex. Dissertation, Faculty of Natural and Agricultural Sciences, Department of Genetics, University of Pretoria (South Africa)

Feuillet C, Travella S, Stein N, Albar L, Nublat A, Keller B (2003) Map-based isolation of the leaf rust disease resistance gene *Lr10* from the hexaploid wheat (*Triticum aestivum* L.) genome. Proc Natl Acad Sci USA 100:15253–15258

Flor HH (1971) Current status of the gene-for-gene concept. Ann Rev Phytopathol 9:275–298

Frick MM, Huel R, NykiForuk CL, Conner RL, Kusyk A, Laroche A (1998) Molecular characterization of a wheat stripe rust resistance gene in Moro wheat. In: Proceedings of the 9th international wheat genetics symposium, Saskatoon, Canada. University Extension Press, University of Saskatchewan, pp 181–182

Gennaro A, Koebner RM, Ceoloni C (2009) A candidate for Lr19, an exotic gene conditioning leaf rust resistance in wheat. Funct Integr Genomics 9:325–334

Gill BS, Appels R, Botha-Oberholster AM, Buell CR, Bennetzen JL, Chalhoub B, Chumley F, Dvořák J, Iwanaga M, Keller B, Li W, McCombie WR, Ogihara Y, Quetier F, Sasaki T (2004) A workshop report on wheat genome sequencing: International Genome Research on Wheat Consortium. Genetics 168:1087–1096

Glowacki S, Macioszek VK, Kononowicz AK (2011) *R*-proteins as fundamentals of plant innate immunity. Cell Mol Biol Lett 16:1–24

Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucl Acids Symp Ser 41:95–98

Hein I, Gilroy EM, Armstrong MR, Birch PR (2009) The zig-zag-zig in oomycete-plant interactions. Mol Plant Pathol 4:547–562

Huang X, Madan A (1999) CAP3: a DNA sequence assembly program. Genome Res 9:868–877

Huang L, Brooks SA, Li W, Fellers JP, Trick HN, Gill BS (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, Brooks S, Li W, Fellers J, Nelson JC, Gill B (2009) Evolution of new disease specificity at a simple resistance locus in a crop-weed complex: reconstitution of the *Lr21* gene in wheat. Genetics 182:595–602

Jones JDG, Dangl J (2006) The plant immune system. Nature 444:323–329

Jones DA, Jones JDG (1997) The role of leucine-rich repeat proteins in plant defences. Adv Bot Res 24:90–167

Kharrat I, Bouktila D, Mezghani-Khemakhem M, Makni H, Makni M (2012) Biotype characterization and genetic diversity of the greenbug, *Schizaphis graminum* (Hemiptera: Aphididae), in north Tunisia. Rev Colomb Entomol 38:87–90

Kohler A, Rinaldi C, Duplessis S, Baucher M, Geelen D, Duchaussoy F, Meyers BC, Boerjan W, Martin F (2008) Genome-wide identification of NBS resistance genes in *Populus trichocarpa*. Plant Mol Biol 66:619–636

Lagudah ES, Moullet O, 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

Leister D, Kurth J, Laurie DA, Yano M, Sasaki T, Devos K, Graner A, Schulze-Lefert P (1998) Rapid reorganization of resistance gene homologues in cereal genomes. Proc Nat Acad Sci 95:370–375

Liu XM, Reese JC, Wilde GE, Fritz AK, Gill BS, Chen M (2005) Hessian fly-resistance genes H9, H10, and H11 are mapped to the distal region of wheat chromosome 1AS. Theor Appl Genet 10:1473–1480

Loutre C, Wicker T, Travella S, Galli P, Scofield S, Fahima T, Feuillet C, Keller B (2009) Two different CC–NBS–LRR genes are required for Lr10-mediated leaf rust resistance in tetraploid and hexaploid wheat. Plant J 60:1043–1054

Lozano R, Ponce O, Ramirez M, Mostajo N, Orjeda G (2012) Genome-wide identification and mapping of NBS-encoding resistance genes in *Solanum tuberosum* group Phureja. PLoS One 7:e34775. doi:10.1371/journal.pone.0034775

Mago R, Nair S, Mohan M (1999) Resistance gene analogues from rice: cloning, sequencing and mapping. Theor Appl Genet 99:50–57

Makni H, Bouktila D, Mezghani M, Makni M (2011) Hessian fly, *Mayetiola destructor* (say), populations in the North of Tunisia: virulence, yield loss assessment and phonological data. Chil J Agric Res 71:401–405

Maleki L, Faris JD, Bowden RL, Gill BS, Fellers JP (2003) Physical and genetic mapping of wheat kinase analogs and NBS–LRR resistance gene analogs. Crop Sci 43:660–670

Pan QL, Wendel J, Fluhr R (2000) Divergent evolution of plant NBS–LRR resistance gene homologues in dicot and cereal genomes. J Mol Evol 50:203–213

Poliseno L, Salmena L, Zhang J, Carver B, Haveman WJ, Pandolfi PP (2010) A coding independent function of gene and pseudogene mRNAs regulates tumour biology. Nature 465:1033–1038

Porter BW, Paidi M, Ming R, Alam M, Nishijima WT, Zhu YJ (2009) Genome-wide analysis of *Carica papaya* reveals a small NBS resistance gene family. Mol Genet Genomics 281:609–626

Rogers J (2014) The IWGSC survey sequencing initiative. International Plant and Animal Genome Conference XXII, San Diego, California (USA), 10–15 Jan 2014. (https://pag.confex.com/pag/xxii/webprogram/Session2168.html)

Rossi M, Goggin FL, Milligan SB, Klaoshian I, Ullman DE, Williamson VM (1998) The nematode resistance gene *Mi* of tomato confers resistance against the potato aphid. Proc Natl Acad Sci USA 95:9750–9754

Salamov A, Solovyev V (2000) *Ab initio* gene finding in *Drosophila* genomic DNA. Genome Res 10:516–522

Sanseverino W, Ercolano MR (2012) *In silico* approach to predict candidate R proteins and to define their domain architecture. BMC Res Notes 5:678. doi:10.1186/1756-0500-5-678

Seah S, Sivasithamparam K, Karakousis K, Lagudah ES (1998) Cloning and characterization of a family of disease resistance gene analogs from wheat and barley. Theor Appl Genet 97:937–945

Shang J, Tao Y, Chen X, Zou Y, Lei C, Wang J, Li X, Zhao X, Zhang M, Lu Z, Xu J, Cheng Z, Wan J, Zhu L (2009) Identification of a new rice blast resistance gene, *Pid3*, by genome wide comparison of paired nucleotide-binding site leucine-rich repeat genes and their pseudogene alleles between the two sequenced rice genomes. Genetics 182:1303–1311

Smith CM, Clement SL (2012) Molecular bases of plant resistance to arthropods. Ann Rev Entomol 57:309–328

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

Srichumpa P, Brunner S, Keller B, Yahiaoui N (2005) Allelic series of four powdery mildew resistance genes at the *Pm3* locus in hexaploid bread wheat. Plant Physiol 139:2885–2895

Talavera G, Castresana J (2007) Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. J Syst Biol 56:564–577

Tan S, Wu S (2012) Genome wide analysis of nucleotide-binding site disease resistance genes in *Brachypodium distachyon*. Comp Funct Genomics 2012:418208. doi:10.1155/2012/418208

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucl Acids Res 22:4673–4680

Traut TW (1994) The functions and consensus motifs of nine types of peptide segments that form different types of nucleotide binding sites. Eur J Biochem 222:9–19

van der Hoorn RAL, Kamoun S (2008) From guard to decoy: a new model for perception of plant pathogen effectors. Plant Cell 20:2009–2017

Wan H, Yuan W, Bo K, Shen J, Pang X, Chen J (2013) Genome-wide analysis of NBS-encoding disease resistance in *Cucumis sativus* and phylogenetic study of NBS-encoding genes in Cucurbitaceae crops. BMC Genom 14:109

Wei F, Gobelman-Werner K, Morroll SM, Kurth J, Mao L, Wing R, Leister D, Schulze-Lefert P, Wise RP (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

Whitham S, Dineshkumar SP, Choi D, Hehl R, Corr C, Baker B (1994) The product of the tobacco mosaic-virus resistance gene N: similarity to toll and the interleukin-1 receptor. Cell 78:1101–1115

Wilkinson PA, Winfield MO, Barker GLA, Allen AM, Burridge A, Coghill JA, Burridge A, Edwards KJ (2012) CerealsDB 2.0: an integrated resource for plant breeders and scientists. BMC Bioinform 13:219

Zhai XG, Zhao T, Liu YH, Long H, Deng GB, Pan ZF, Yu MQ (2008) Characterization and expression profiling of a novel cereal cyst nematode resistance gene analog in wheat. Mol Biol (NY) 42:960–965

Zhang Z, Schwartz S, Wagner L, Miller W (2000) A greedy algorithm for aligning DNA sequences. J Comput Biol 7:203–214