# GENOME-WIDE IDENTIFICATION OF DISEASE RESISTANCE GENES IN *AEGILOPS TAUSCHII* COSS. (POACEAE)

**Ethan J. Andersen, Samantha R. Shaw, and Madhav P. Nepal***
Department of Biology & Microbiology
South Dakota State University
Brookings, SD 57007
*Corresponding author email: Madhav.Nepal@sdstate.edu

## ABSTRACT

Identifying disease resistance genes (R-genes) and revealing their functions are important for understanding a plant's defense against pathogens. *Aegilops tauschii*, the contributor of wheat's D-genome, has a recently available complete genome sequence, and genome-wide identification of R-genes in this plant would give insight into the evolution of wheat resistance genes. The main objectives of this project were to identify CNL (Coiled-coil, Nucleotide-binding site, and Leucine-rich region) R-genes within the *A. tauschii* genome, and elucidate their evolutionary relationships within *Aegilops* and across the genome of two model plants—*Arabidopsis* and rice. We conducted *in silico* analyses in which known CNL genes of *Arabidopsis* and rice were used to search for their orthologs in *A. tauschii*. We identified 402 CNL resistance genes within the *A. tauschii* genome and recovered three clades (A, B, and C) of *A. tauschii* CNL genes of which CNL C is the largest clade, a single member represents clade A, and clade D is entirely absent. Each of these clades was characterized by a consistent motif structure. The number of exons varied from 1 to 28 with an average number of 4.5. The majority of CNL genes were inferred to have originated by tandem duplications, and the historical gene duplication events perhaps diversified the members in response to a unique pathogen pressure. Identification of *Aegilops* R-genes would help us understand the evolution of R-genes, particularly those located in the D-genome of wheat, and has a potential implication in creating a durable R-gene in *Aegilops*, wheat, and other crop species in future.

Keywords

Disease resistance, NBS-LRR, R-genes, CNL genes, D-genome of wheat, *Aegilops tauschii,* bioinformatics

## INTRODUCTION

Plant defense against pathogens involves complex signaling pathways that trigger resistance responses (Jones and Dangl 2006). Such responses typically lead to a hypersensitive response, but can also include the production of anti-pathogen

chemicals or cell wall fortification (Hammond-Kosack and Jones 1996). Hypersensitive response, in particular, is a general response that involves the programmed cell death of a section of tissue that has been infected by a pathogen to quarantine the affected area (Hammond-Kosack and Jones 1996).

Disease resistance genes, or R-genes, encode proteins that are involved in the detection of pathogen attacks and activation of subsequent downstream plant response signaling. The R-genes occur as multigene families, and multiple models have been proposed to describe their mechanism of action. The Gene-for-Gene Model describes plants having specific dominant resistance genes that counter corresponding pathogen avirulence genes in an evolutionary arms-race (Flor 1971). Introducing more molecular details, the Guard Model describes resistance genes bound to plant proteins and are activated when that protein is cleaved by a pathogen protein (Van Der Biezen and Jones 1998; Shao et al. 2003), while the Zig-Zag Model describes the pathogen evolving new avirulence genes that evade plant basal immunity (Jones and Dangl 2006). Recently R-genes have been classified into eight specific groups (Gururani *et al.* 2012). Among them, the overwhelming majority of the R-genes fall under the NBS-LRR type, the largest class of R-genes (Meyers et al. 2003; Meyers et al. 2005). The NBS-LRR genes can be categorized into two major types based upon whether they start with a Toll Interleukin Receptor (TIR-NBS-LRR or TNL; absent in monocots) or a Coiled Coil (CC-NBS-LRR or CNL; present in all plants) (Meyers et al. 2003).

Resistance genes evolve rapidly due to the high selection pressure put onto the plant population by a pathogen load (Bergelson et al. 2001) that causes faster gene diversification (Michelmore and Meyers 1998). This diversification is caused primarily by gene recombination and transposable elements' activities (McGrann et al. 2014). Their loss is also possible by deficient duplications and the loss of lineages, as evidenced in cucumber and watermelon genomes that contain many fewer resistance genes (Lin et al. 2013). In addition, the evolution of R-genes occurs through a trade-off between physical, chemical, and molecular defenses in response to coevolving pathogens (Hammond-Kosack and Jones 1996).

The increasing availability of complete genome sequences of plants at various taxonomic levels allows us to carry out comparative analyses for identification of R-genes and for understanding the evolutionary processes involved. CNL R-genes have been identified for various plant species such as papaya (6; Porter et al. 2009), cucumber (18; Wan et al. 2013), rice (159, 149; Zhou et al. 2004; Benson 2014), *Arabidopsis* (55; Meyers et al. 2003), poplar (119; Kohler et al. 2008), *Medicago* (177; Ameline-Torregrosa et al. 2008), soybean (188, Benson 2014; Nepal and Benson 2015), potato (370; Lozano et al. 2012), and are yet to be identified in *Aegilops tauschii* Coss. (Poaceae), the D-genome contributor of bread wheat (*Triticum aestivum* L.). *A. tauschii* underwent hybridization with *Triticum turgidum* several thousand years ago, forming bread wheat (Jia et al. 2013). The objectives of this research were to identify *A. tauschii* CNL resistance genes and elucidate their evolutionary relationships within *A. tauschii* and across the genomes of *Arabidopsis* and rice, two model plant species.

## METHODS

*A. tauschii* protein sequences were searched in the Ensembl Genomes site (Kersey et al. 2014). Previously identified *Arabidopsis* CNL resistance genes (Meyers et al. 2003) were obtained from the Phytozome database (Goodstein et al. 2012). First, fifty CNL genes of *Arabidopsis* were aligned in the program ClustalW and used to construct a Hidden Markov Model to search for the entire set of *A. tauschii* protein sequences with a stringency of 0.05. The *A. tauschii* genes were uploaded into the program Geneious (Kearse et al. 2012) and annotated with InterProScan (Jones et al. 2014) to identify NBARCs with the program Pfam (pfam.sanger.ac.uk) that allowed the exclusion of sequences with TIR motifs.

The protein sequences with NBARCs were used to construct a reiterative HMM to search the *A. tauschii* proteins for species-specific CNL genes at a stringency of 0.001. A total of 810 genes were identified through first HMM at a stringency of 0.05. Of these genes, 711 were determined to contain NBARCs through domain annotation with InterProScan. The reiterative HMM identified 779 genes and after removing gene duplicates, 711 of these 779 genes were determined to contain NBARCs, of which only 544 genes contained both NBARC and "DiseaseResist" domains. The NBARCs of these genes were then uploaded to MEME suite to perform MEME analysis (Bailey and Elkan 1994) and annotate the three characteristic domains of the CNL genes, i.e. P-loop, Kinase-2, and GLPL motifs. All genes containing these three motifs were aligned using ClustalW integrated in the program MEGA 6.0 (Tamura et al. 2011). *Arabidopsis* as well as rice sequences were also imported into MEGA 6.0 to make two phylogenetic trees (100 bootstrap replicates using the JTT+G Model for both trees) to look for evolutionary relationships between the genes. Exon structure was also determined using exon information and scaffold location data from the Ensembl Genomes site. Gene exon coordinates were used in the program Fancygene v1.4 to visualize the exon-intron structure.

## RESULTS AND DISCUSSION

Of the 33,928 *A. tauschii* protein sequences analyzed, 402 genes (1.2% of the genome) were identified as CNL genes. All of these genes had P-loop, Kinase-2, and GLPL motifs, the characteristic domains of the CNL genes. Phylogenetic relationships of the identified CNL genes along with their orthologs in *Arabidopsis* and in rice are shown in Figure 1 and 2, respectively. The CNL genes were nested in three clades (A, B and C). The clade D found in *Arabidopsis* and other dicot species was completely absent. The CNL-A clade was severely reduced to one member in the *A. tauschii* genome, whereas *Arabidopsis* has six CNL-A members. While *A. tauschii* has a substantially larger genome than rice, the number of coding genes for *A. tauschii* and rice are quite similar, at 33,929 and 35,679 genes, respectively (Zhou et al. 2004; Jia et al. 2013). The CNL gene-content in the two genomes is not highly divergent, despite a huge difference in genome size between the two species. Table 1 shows that the number of CNL genes does

not necessarily correlate with genome size (G-value paradox; Michelmore et al. 2013). With the larger genome size (2.7Gb), however, the *A. tauschii* genome contains a higher number of CNL genes. The rice genome (420 Mb) contains approximately 150 CNL genes (Zhou et al. 2004). All CNL clade information for the 402 identified genes is summarized in Table 2.

This study confirmed through MEME analysis (Figure 3) the presence of characteristic motifs (P-loop, Kinase-2, and GLPL) in all 402 CNL genes in *Aegilops*. The motif compositions presented here are similar to that in *Arabidopsis* (Meyers et al. 1999) and corresponded to the phylogenetic clustering represented in the phylogenetic tree (Figure 1, 3). For instance, Motif 8 (CPxxL) was common in the CNL-C4 clade but only in a few genes in the rest of CNL-C (Clades CNL-C1, CNL-C2, and CNL-C3). Since only the most prevalent motifs were labeled,

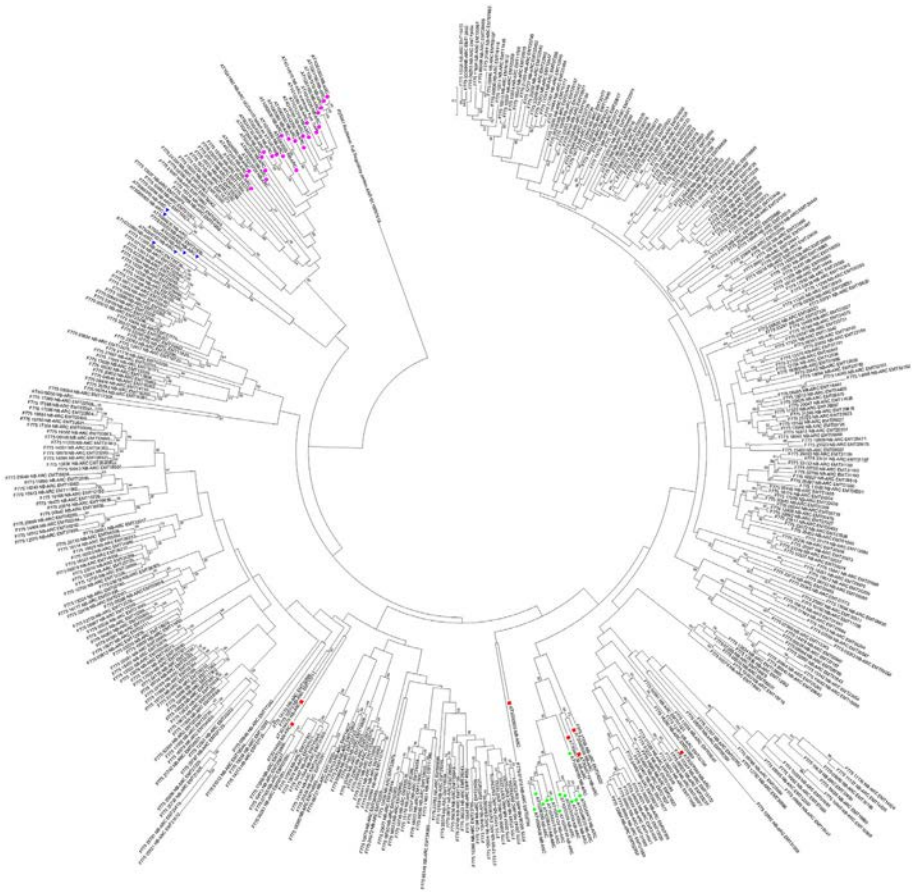

***Figure 1. Phylogenetic analysis of the CNL genes of A. tauschii and their orthologs in A. thaliana. The tree was constructed using the JTT+G model with 100 bootstrap replicates. CNL clades A, B, C, and D are shown with blue, pink, red, and green symbols, respectively. A high resolution readable TIF copy of this figure is available from the corresponding author. It can also be downloaded from the author's lab website at https://www.sdstate.edu/biomicro/people/faculty/madhav-nepal/nepal-lab.cfm.***

and few CNL-A and CNL-B genes were present, it is likely that motifs were present but not described by the MEME analysis.

Since *A. tauschii* genes have not been mapped onto their chromosomes, gene clustering analysis was not performed in the present study. It is highly likely that the genes exist in many clusters throughout the genome (Meyers et al. 2003), particularly in the extrapericentromeric regions of the chromosomes as documented in soybean (Benson 2014; Nepal and Benson 2015). Further analyses of NBS-LRR disease resistance gene clustering will need to be conducted once this information becomes available. Also not available yet are the alternate transcripts for each of the genes. This is evident because the number of protein sequences available is equal to the number of coding genes within the genome. In other genomes, such as the barley genome, many more protein sequences exist that give information on alternative splicing amongst the resistance genes. Alternative splicing would increase the possible resistance gene proteins, which would be highly useful while facing a quickly evolving pathogen. While information

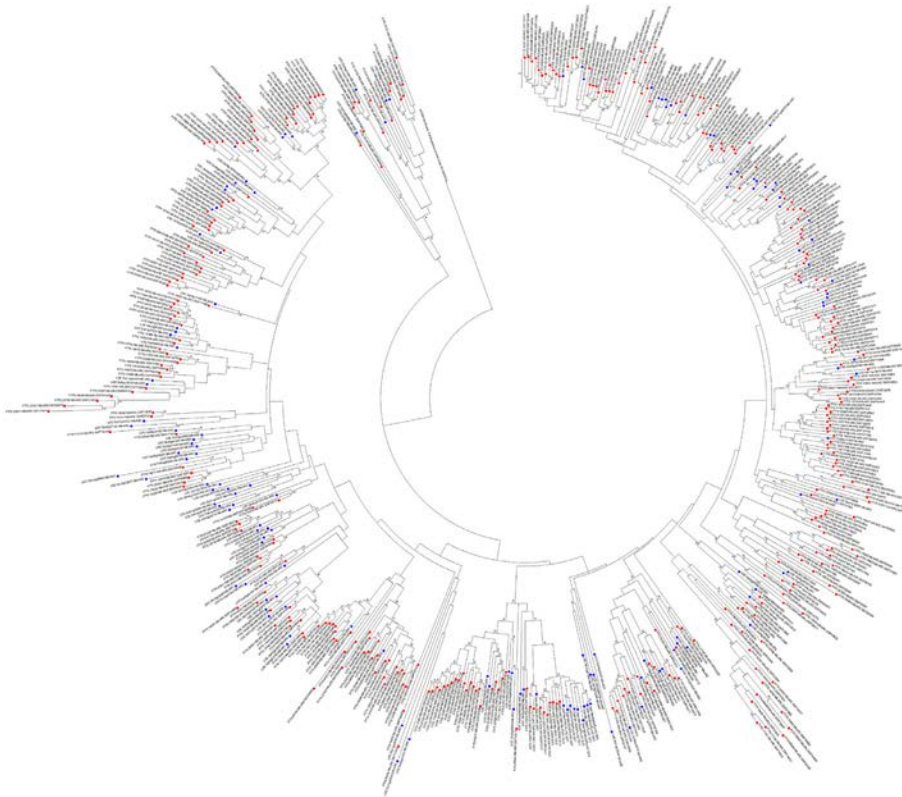

***Figure 2. Phylogenetic analysis of the CNL genes of A. tauschii and their orthologs in rice. The tree was constructed using the JTT+G model with 100 bootstrap replicates. A. tauschii and rice genes are shown with red and blue symbols, respectively. A high resolution readable TIF copy of this figure is available from the corresponding author. It can also be downloaded from the author's lab website at https://www.sdstate.edu/biomicro/people/faculty/madhav-nepal/nepal-lab.cfm.***

***Table 1. Genome size and CNL gene content of selected plant species. This table was modified from Marone et al (Marone et al. 2013). Genome size and CNL gene references are both listed in the references column.***

| Species | Genome Size | Number of CNL genes | Reference |
|---|---|---|---|
| *Aegilops tauschii* | 4.4 Gb | 402 | Jia et al. 2013 |
| *Glycine max* | 1.115 Gb | 188 | Schmutz et al. 2010; Benson 2014; Nepal and Benson 2015 |
| *Solanum tuberosum* | 844 Mb | 370 | Consortium 2011; Lozano et al. 2012 |
| *Phaseolus vulgaris* | 587 Mb | 94 | Benson 2014; Schmutz et al. 2014 |
| *Vitis vinifera* | 487 Mb | 203 | Jaillon et al. 2007; Yang et al. 2008 |
| *Populus trichocarpa* | 423 Mb | 119 | Tuskan et al. 2006; Kohler et al. 2008 |
| *Oryza sativa* | 420 Mb | 159, 149 | Zhou et al. 2004, Goff et al. 2002; Benson 2014 |
| *Medicago truncatula* | 375 Mb | 177 | Ameline-Torregrosa et al. 2008; Young et al. 2011 |
| *Carica papaya* | 372 Mb | 6 | Ming et al. 2008; Porter et al. 2009 |
| *Brassica rapa* | 284 Mb | 30 | Mun et al. 2009; Wang et al. 2011 |
| *Brachypodium distachyon* | 272 Mb | 102 | Vogel et al. 2010; Tan and Wu 2012 |
| *Cucumis sativus* | 244 Mb | 18 | Huang et al. 2009; Wan et al. 2013 |
| *Arabidopsis lyrata* | 207 Mb | 21 | Guo et al. 2011; Hu et al. 2011 |
| *Arabidopsis thaliana* | 125 Mb | 55 | Initiative 2000; Meyers et al. 2003 |

on alternate splicing is not available for the *Aegilops* CNL genes, exon/intron information is available (**Figure 4**). The average exon content of 4.45 exons per gene is higher than previously found in *Arabidopsis* and CNL-C genes in soybean (Benson 2014; Nepal and Benson 2015). The number of exons varied from 1 (F775_00002) to 28 (F775_52438). Thirty five CNL genes had one exon, 77 had two, 83 had three, 58 had four, 44 had five, 28 had six, 23 had seven, 9 had eight, 14 had nine, 10 had ten, six had 11, seven had 12, three had 13, two had 14, one had 22 and one gene had 28 exons (**Figure 5**). With the multitude of genes with many exons, it can be hypothesized that alternate splicing has a large impact on the protein structure of the resistance genes, since multiple exons allow for a higher number of combinations during splicing (Tan et al. 2007). Alternative splicing has been shown to play an important role in resistance gene expression in *Arabidopsis* (Dinesh-Kumar and Baker 2000; Tan et al. 2007).

Phylogenetic analysis of *A. tauschii* CNL genes shows an expansion of the CNL-C group and a slight reduction of CNL-B members relative to *Arabidopsis* (Figure 1). There is a severe reduction of the CNL-A clade to a single member. These results in the *Aegilops* genome are consistent with the CNL genes in rice, another monocot species (Benson 2014; Nepal and Benson 2015). There was low interspecific nesting indicating the lower prevalence of segmental duplica-

***Table 2. List of all Aegilops tauschii CNL genes according to clade.***

| *Aegilops* gene | Clade | *Aegilops* gene | Clade | *Aegilops* gene | Clade | *Aegilops* gene | Clade | *Aegilops* gene | Clade |
|---|---|---|---|---|---|---|---|---|---|
| F775_00002 | C3 | F775_09061 | C2 | F775_12934 | C2 | F775_18513 | C4 | F775_25567 | C4 |
| F775_00003 | C3 | F775_09164 | C4 | F775_12982 | C3 | F775_18529 | C2 | F775_25587 | C4 |
| F775_00009 | C4 | F775_09200 | C3 | F775_13024 | C4 | F775_18533 | C2 | F775_25618 | C2 |
| F775_00012 | C1 | F775_09247 | C4 | F775_13028 | C1 | F775_18542 | C4 | F775_25651 | C2 |
| F775_00020 | C2 | F775_09300 | C4 | F775_13037 | B | F775_18596 | C4 | F775_25666 | C2 |
| F775_00028 | C4 | F775_09360 | C4 | F775_13161 | C3 | F775_18633 | C2 | F775_25677 | C2 |
| F775_00089 | C2 | F775_09379 | C2 | F775_13322 | C2 | F775_18678 | C2 | F775_25696 | C2 |
| F775_00261 | C2 | F775_09385 | C2 | F775_13548 | C4 | F775_18692 | C4 | F775_25697 | C4 |
| F775_00279 | C3 | F775_09416 | C1 | F775_13556 | C3 | F775_18745 | C2 | F775_25723 | C4 |
| F775_00445 | C4 | F775_09429 | C4 | F775_13570 | C4 | F775_18750 | C2 | F775_25735 | C2 |
| F775_00504 | C4 | F775_09721 | C2 | F775_13594 | C4 | F775_18752 | C4 | F775_25748 | C2 |
| F775_00542 | C2 | F775_09754 | C1 | F775_13630 | C2 | F775_19013 | C2 | F775_25761 | C2 |
| F775_00546 | C2 | F775_09801 | C4 | F775_13836 | C2 | F775_19082 | C4 | F775_25787 | C2 |
| F775_00591 | C3 | F775_09834 | C1 | F775_13864 | C4 | F775_19119 | C4 | F775_25792 | C1 |
| F775_00649 | C1 | F775_09885 | C3 | F775_13876 | C2 | F775_19175 | C4 | F775_25799 | C2 |
| F775_01012 | C2 | F775_09936 | C2 | F775_13917 | C1 | F775_19216 | C4 | F775_25826 | C2 |
| F775_01226 | C3 | F775_09937 | C4 | F775_13926 | C1 | F775_19299 | C4 | F775_25860 | C2 |
| F775_01227 | C3 | F775_10024 | C3 | F775_13948 | C3 | F775_19382 | C2 | F775_26631 | C4 |
| F775_01584 | C4 | F775_10028 | C3 | F775_13994 | C3 | F775_19398 | C4 | F775_29542 | C4 |
| F775_01810 | C | F775_10030 | C1 | F775_14051 | C2 | F775_19512 | C4 | F775_31118 | C1 |
| F775_02378 | C1 | F775_10069 | C4 | F775_14065 | C4 | F775_19584 | C2 | F775_31260 | C4 |
| F775_02380 | B | F775_10122 | C4 | F775_14066 | C4 | F775_19672 | C2 | F775_32992 | C4 |
| F775_02497 | C3 | F775_10192 | C4 | F775_14094 | C2 | F775_19733 | C2 | F775_33053 | C4 |
| F775_02559 | C4 | F775_10336 | C1 | F775_14117 | C2 | F775_19734 | C2 | F775_33066 | C4 |
| F775_02729 | C2 | F775_10337 | C1 | F775_14170 | C2 | F775_19740 | C4 | F775_33089 | C4 |
| F775_02795 | C4 | F775_10338 | C2 | F775_14195 | C2 | F775_19750 | C2 | F775_33131 | C4 |
| F775_02796 | C4 | F775_10342 | C4 | F775_14213 | C2 | F775_19781 | C4 | F775_33132 | C4 |
| F775_03255 | B | F775_10343 | C4 | F775_14243 | C4 | F775_19900 | C2 | F775_33159 | C2 |
| F775_03276 | C2 | F775_10347 | C3 | F775_14254 | B | F775_19909 | C4 | F775_33179 | C4 |
| F775_03594 | B | F775_10367 | C4 | F775_14260 | C3 | F775_19928 | C2 | F775_33181 | C4 |
| F775_03781 | C4 | F775_10383 | C3 | F775_14262 | C3 | F775_20047 | C4 | F775_33215 | C4 |
| F775_03812 | C2 | F775_10389 | C1 | F775_14451 | C4 | F775_20078 | C4 | F775_33238 | C4 |
| F775_03909 | C2 | F775_10409 | C2 | F775_14478 | C4 | F775_20098 | C4 | F775_33239 | C4 |
| F775_04060 | C2 | F775_10413 | C2 | F775_14484 | C4 | F775_20113 | B | F775_33246 | C4 |
| F775_04135 | C2 | F775_10432 | C3 | F775_14498 | C2 | F775_20140 | C4 | F775_33249 | C4 |
| F775_04483 | C3 | F775_10464 | C1 | F775_14564 | C1 | F775_20226 | C4 | F775_33281 | C4 |
| F775_04549 | C1 | F775_10470 | C2 | F775_15013 | C1 | F775_20252 | C4 | F775_52103 | C2 |
| F775_04571 | C3 | F775_10485 | C4 | F775_15035 | B | F775_20381 | C2 | F775_52265 | C1 |
| F775_04590 | C2 | F775_10487 | C2 | F775_15095 | C4 | F775_20428 | C4 | F775_52271 | C4 |
| F775_04976 | C4 | F775_10498 | C2 | F775_15179 | C2 | F775_20439 | C4 | F775_52304 | C2 |
| F775_04978 | C3 | F775_10499 | C2 | F775_15186 | C4 | F775_20802 | C1 | F775_52483 | C4 |
| F775_04989 | C3 | F775_10519 | C4 | F775_15197 | C2 | F775_20828 | C4 | F775_52537 | C4 |
| F775_04991 | C3 | F775_10548 | C2 | F775_15224 | C2 | F775_20864 | C4 | | |
| F775_05010 | C4 | F775_10570 | C2 | F775_15316 | B | F775_20893 | C4 | | |
| F775_05050 | C3 | F775_10673 | C2 | F775_15432 | C2 | F775_20916 | C2 | | |
| F775_05085 | C3 | F775_10845 | C3 | F775_15460 | C4 | F775_20940 | C4 | | |
| F775_05094 | C1 | F775_10913 | C4 | F775_15674 | C4 | F775_20943 | C4 | | |
| F775_05363 | C4 | F775_10943 | C2 | F775_15677 | C2 | F775_21097 | B | | |
| F775_05510 | C3 | F775_10988 | B | F775_15785 | B | F775_21138 | C4 | | |
| F775_05536 | C2 | F775_10989 | B | F775_15841 | C4 | F775_21246 | C4 | | |
| F775_05818 | C4 | F775_11003 | C4 | F775_15860 | C4 | F775_21278 | C4 | | |
| F775_05820 | C4 | F775_11136 | C3 | F775_15890 | C2 | F775_21387 | C4 | | |
| F775_05946 | C4 | F775_11137 | C3 | F775_15918 | C2 | F775_21401 | C4 | | |
| F775_06146 | C2 | F775_11205 | C2 | F775_15949 | C4 | F775_21420 | C4 | | |
| F775_06149 | C2 | F775_11229 | C4 | F775_16114 | C2 | F775_21616 | C4 | | |
| F775_06253 | C4 | F775_11298 | C4 | F775_16168 | C2 | F775_21742 | C2 | | |
| F775_06279 | C4 | F775_11345 | C4 | F775_16243 | C2 | F775_21780 | C4 | | |
| F775_06285 | C4 | F775_11368 | C4 | F775_16266 | C2 | F775_21795 | C | | |
| F775_06326 | C3 | F775_11385 | C1 | F775_16271 | C2 | F775_21811 | C4 | | |
| F775_06411 | C4 | F775_11502 | C2 | F775_16379 | C4 | F775_21857 | C1 | | |
| F775_06721 | C2 | F775_11544 | C2 | F775_16385 | C2 | F775_22010 | C2 | | |
| F775_06827 | C4 | F775_11560 | C4 | F775_16579 | C2 | F775_22133 | C2 | | |
| F775_06830 | C4 | F775_11646 | C2 | F775_16654 | C4 | F775_22416 | C2 | | |
| F775_06989 | C1 | F775_11651 | C2 | F775_16715 | B | F775_22559 | C4 | | |
| F775_07053 | A | F775_11684 | C4 | F775_16721 | B | F775_22763 | C1 | | |
| F775_07156 | C1 | F775_11767 | C3 | F775_16774 | C4 | F775_22887 | C4 | | |
| F775_07165 | C2 | F775_11868 | C2 | F775_16813 | C4 | F775_22902 | C2 | | |
| F775_07193 | C2 | F775_11909 | C2 | F775_16814 | C4 | F775_22957 | C4 | | |
| F775_07248 | C2 | F775_11949 | C1 | F775_16933 | C2 | F775_23072 | C4 | | |
| F775_07285 | C3 | F775_12011 | C4 | F775_16964 | C4 | F775_23165 | C2 | | |
| F775_07399 | C4 | F775_12159 | C2 | F775_17304 | C2 | F775_23513 | C2 | | |
| F775_07702 | C4 | F775_12184 | C2 | F775_17322 | C4 | F775_23649 | C2 | | |
| F775_07703 | C1 | F775_12189 | C2 | F775_17331 | C4 | F775_23709 | C1 | | |
| F775_07864 | C2 | F775_12361 | C2 | F775_17386 | C2 | F775_23714 | C4 | | |
| F775_07949 | C4 | F775_12408 | C4 | F775_17388 | C2 | F775_23783 | C2 | | |
| F775_08064 | C4 | F775_12507 | C4 | F775_17389 | C2 | F775_23909 | C4 | | |
| F775_08223 | C2 | F775_12570 | C2 | F775_17599 | C4 | F775_24339 | B | | |
| F775_08252 | C3 | F775_12676 | C2 | F775_17741 | C4 | F775_24349 | C2 | | |
| F775_08345 | C4 | F775_12681 | C2 | F775_17804 | C1 | F775_24493 | C4 | | |
| F775_08380 | C2 | F775_12700 | C2 | F775_17853 | C4 | F775_24972 | C2 | | |
| F775_08523 | C3 | F775_12704 | C4 | F775_17929 | C2 | F775_25023 | C4 | | |
| F775_08534 | C3 | F775_12720 | C1 | F775_17949 | C4 | F775_25345 | C4 | | |
| F775_08544 | C4 | F775_12725 | C2 | F775_17950 | C4 | F775_25375 | C2 | | |
| F775_08623 | C4 | F775_12737 | C4 | F775_17959 | C2 | F775_25411 | C2 | | |
| F775_08715 | C1 | F775_12747 | C2 | F775_18040 | C4 | F775_25442 | C4 | | |
| F775_08722 | C4 | F775_12769 | C4 | F775_18213 | C4 | F775_25448 | C4 | | |
| F775_08786 | C3 | F775_12792 | C3 | F775_18251 | C4 | F775_25453 | C4 | | |
| F775_08856 | C2 | F775_12834 | C3 | F775_18334 | C2 | F775_25512 | C1 | | |
| F775_08907 | C2 | F775_12836 | C3 | F775_18423 | C2 | F775_25531 | C2 | | |
| F775_08994 | C4 | F775_12859 | C4 | F775_18512 | C2 | F775_25543 | C2 | | |

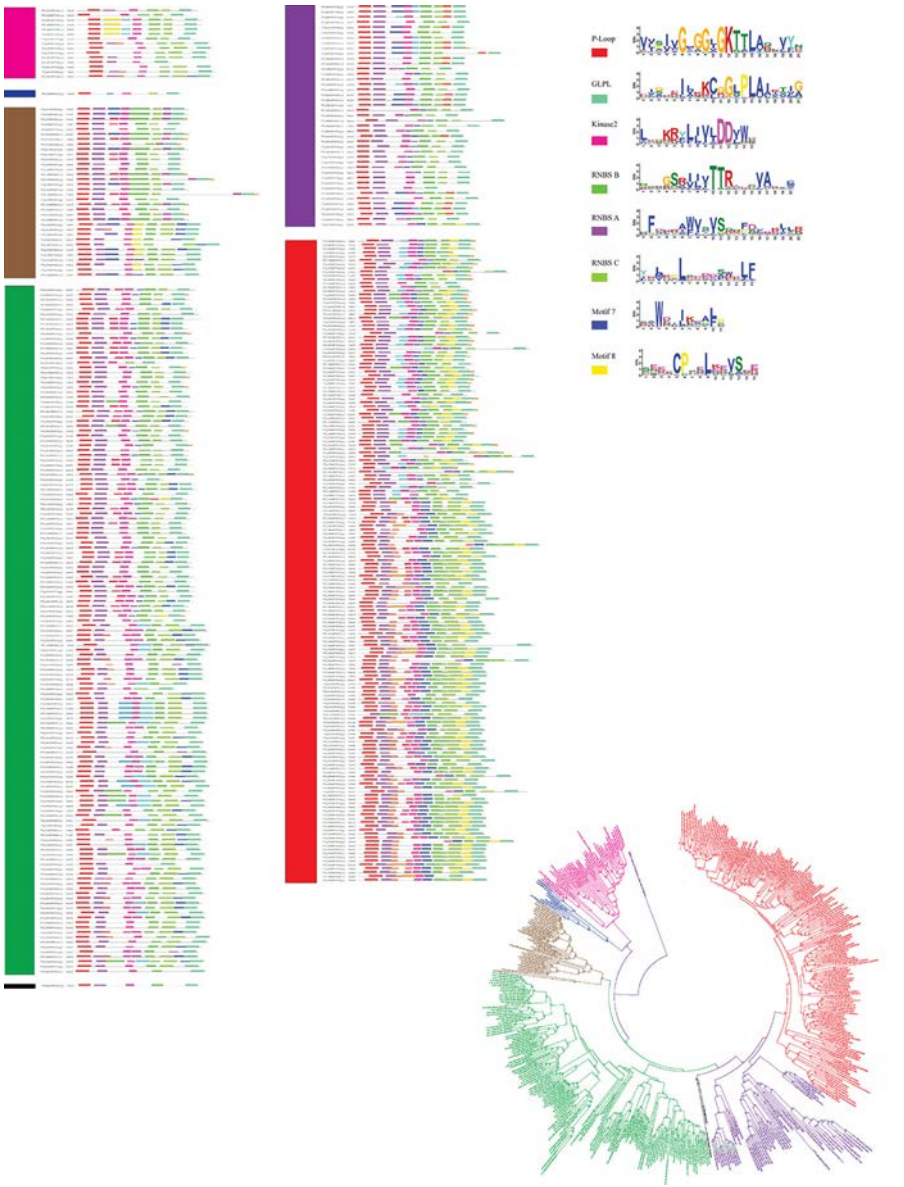


*Figure 3. MEME analysis of the 402 A. tauschii genes. The block diagrams show the characteristics three motifs used to identify CNL genes (P-Loop, Kinase-2, and GLPL) along with other highly prevalent motifs, split according to clade as shown by the tree (lower right) color-coded to represent the domain compositions in Figure 1. CNL-B, A, C1, C2, C3, and C4 are colored pink, blue, brown, green purple, and red, respectively. A high resolution readable TIF copy of this figure is available from the corresponding author. It can also be downloaded from the author's lab website at https://www.sdstate.edu/biomicro/people/faculty/madhav-nepal/nepal-lab.cfm.*

***Figure 4. Exon content of the 402 A. tauschii genes showing splice locations between exons (gray bars) and introns (dashed lines). Genes are listed by accession. A high resolution readable TIF copy of this figure is available from the corresponding author. It can also be downloaded from the author's lab website at https://www.sdstate.edu/biomicro/people/faculty/madhav-nepal/nepal-lab.cfm.***

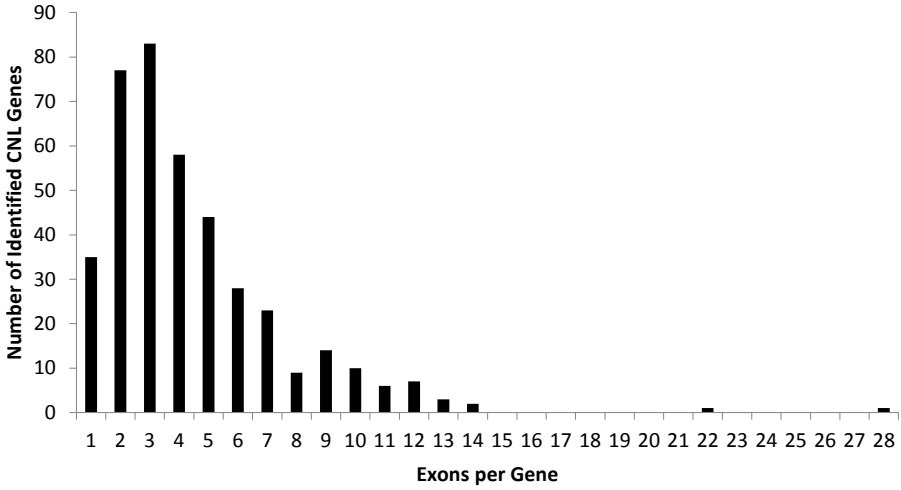


***Figure 5. Number of CNL genes with specific number of exons in A. tauschii***

tions. Since chromosome location and gene clustering information were not available, instances of tandem versus segmental duplications could not be determined with a high degree of certainty. Genes that are nested together within a clade and occurring within the same gene clusters are likely to have originated through tandem duplications. The current study presented several instances of tandem duplication: for example, because F775_14065 and F775_14066, are sister members (Figure 1), and subsequently accessioned, it is highly likely that they originated by tandem duplication. Other examples of tandem duplications include F775_11136 and F775_11137, F775_02795 and F775_02796, F775_10498 and F775_10499, F775_10336 and F775_10337, and three genes F775_17386, F775_17388 and F775_17389.

Orthologs of some *A. tauschii* CNL genes have been previously characterized. For example, RPM1 of *Arabidopsis thaliana* is involved in the resistance response to *Pseudomonas syringae* (Mackey et al. 2002). As shown in Figure 1, the *Arabidopsis* RPM1 ortholog in *Aegilops* has three paralogs (F775_10347, F775_14260, and F775_13161) indicating an expansion of this particular gene. It could be hypothesized that *A. tauschii* evolved the three genes in response to diversifying *P. syringae* strains or similar pathogens since the split of common ancestors of *Arabidopsis* and *A. tauschii*. The diversification of RPM1 orthologs in *Aegilops* might have resulted from the selection pressure imposed by different pathogens in *A. taushii*'s life history. Figure 2 shows expansions of several *Aegilops* CNL genes: for example, eleven *A. tauschii* paralogs (F775_10913, 12507, 12011, 05946, 06830, 13024, 33089, 06253, 11684, 09360, and 21401) are related to rice gene LOC_Os08g10260. This shows that *A. tauschii* might have evolved as many as 11 genes in response to the same pathogen as in rice, perhaps diversifying in the *Aegilops* niche.

Due to the growing problem of Ug99 stem rust in wheat production of East Africa and the Middle-East, the CNL resistance gene SR33 has been identified as

a possible solution (Periyannan et al. 2013). Our result determines that accession F775_10122 represents the SR33 gene in *Aegilops*, which could be the gene of interest for developing a durable resistance in wheat. Other genes (F775_13548, F775_16813, and F775_18040) closely related to SR33 might contain valuable traits as well. Further investigation of these genes, along with the splice variants of F775_10122 is warranted if SR33 proves to be useful in agricultural production. *In silico* analyses of R-genes such as presented here are integral stepping-stones toward the use of these identified genes as weapons against evolving pathogens. While further investigation of gene expression data and genomic composition is important for understanding functional characterization, the present study provides information on the diversity and evolutionary history of the CNL genes in *A. tauschi* genome, and has a potential implication in future wheat crop improvement with durable resistant genes.

## ACKNOWLEDGEMENTS

This project was supported by the Undergraduate Research Support Fund from the Department of Biology and Microbiology at South Dakota State University, and USDA-NIFA Hatch Project Fund to M. Nepal. Co-author Samantha Shaw was enrolled in M. Nepal's section of BIOL 498 (Undergraduate Research and Scholarship) course in spring 2015. The authors would like to thank Dr. Shaukat Ali and Dr. Yajun Wu for their valuable feedback on the manuscript.

## LITERATURE CITED

Ameline-Torregrosa, C., B. B. Wang, M.S. O'Bleness, S. Deshpande, H. Zhu, B. Roe, N.D. Young, and S.B. Cannon. 2008. Identification and characterization of nucleotide-binding site-leucine-rich repeat genes in the model plant Medicago truncatula. Plant Physiology 146:5-21.

Initiative. 2000. Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature 408:796.

Bailey, T.L., and C. Elkan. 1994. Fitting a mixture model by expectation maximization to discover motifs in bipolymers. UCSD Technical Report CS94-351, Department of Computer Science and Engineering, University of California, San Diego. 14 pp.

Benson, B.V. 2014. Disease Resistance Genes and their Evolutionary History in Six Plant Species. M.S. Thesis. South Dakota State University, Brookings, SD.

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

Dinesh-Kumar, S.P., and B.J. Baker. 2000. Alternatively spliced N resistance gene transcripts: their possible role in tobacco mosaic virus resistance. Proceedings of the National Academy of Sciences 97:1908-1913.

Flor, H.H. 1971. Current status of the gene-for-gene concept. Annual Review of Phytopathology 9:275-296.

Goff, S.A., D. Ricke, T-H. Lan, G. Presting, R. Wang, M. Dunn, J. Glazebrook, A. Sessions, P. Oeller, H. Varma, et al. 2002. A draft sequence of the rice genome (Oryza sativa L. ssp. japonica). Science 296:92-100.

Goodstein, D.M., S. Shu, R. Howson, R. Neupane, R.D. Hayes, J. Fazo, T. Mitros, W. Dirks, U. Hellsten, N. Putnam, and D.S. Rokhsar. 2012. Phytozome: a comparative platform for green plant genomics. Nucleic Acids Research 40:D1178-D1186.

Guo, Y-L., J. Fitz, K. Schneeberger, S. Ossowski, J. Cao, and D. Weigel. 2011. Genome-wide comparison of nucleotide-binding site-leucine-rich repeat-encoding genes in Arabidopsis. Plant Physiology 157:757-769.

Gururani, M.A., J. Venkatesh, C.P. Upadhyaya, A. Nookaraju, S.K. Pandey, and S.W. Park. 2012. Plant disease resistance genes: current status and future directions. Physiological and Molecular Plant Pathology 78:51-65.

Hammond-Kosack, K.E. and J.D. Jones. 1996. Resistance gene-dependent plant defense responses. The Plant Cell 8:1773.

Hu, T.T., P. Pattyn, E.G. Bakker, J. Cao, J-F. Cheng, R.M. Clark, N. Fahlgren, J.A. Fawcett, J. Grimwood, H. Gundlach, et al. 2011. The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. Nature Genetics 43:476-481.

Huang, S., R. Li, Z. Zhang, L. Li, X. Gu, W. Fan, W.J. Lucas, X. Wang, B. Xie, P. Ni et al. 2009. The genome of the cucumber, Cucumis sativus L. Nature Genetics 41:1275-1281.

Jaillon, O., J-M. Aury, B. Noel, A. Policriti, C. Clepet, A. Casagrande, N. Choisne, S. Aubourg, N. Vitulo, C. Jubin, et al. 2007. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature 449:463-467.

Jia, J., S. Zhao, X. Kong, Y. Li, G. Zhao, W. He, R. Appels, M. Pfeifer, Y. Tao, X. Zhang, et al. 2013. Aegilops tauschii draft genome sequence reveals a gene repertoire for wheat adaptation. Nature 496:91-95.

Jones, J.D., and J.L. Dangl. 2006. The plant immune system. Nature 444:323-329.

Jones, P., D. Binns, H-Y. Chang, M. Fraser, W. Li, C. McAnulla, H. McWilliam, J. Maslen, A. Mitchell, G. Nuka, et al. 2014. InterProScan 5: genome-scale protein function classification. Bioinformatics 9:1236-1240.

Kearse, M., R. Moir, A. Wilson, S. Stones-Havas, M. Cheung, S. Sturrock, S. Buxton, A. Cooper, S. Markowitz, C. Duran, T. Thierer, B. Ashton, P. Meintjes, and A. Drummond. 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28:1647-1649.

Kersey, P.J., J.E. Allen, M. Christensen, P. Davis, L.J. Falin, C. Grabmueller, D.S.T. Hughes, J. Humphrey, A. Kerhornou, J. Khobova, et al. 2014. Ensembl Genomes 2013: scaling up access to genome-wide data. Nucleic Acids Research 42:D546-D552.

Kohler, A., C. Rinaldi, S. Duplessis, M. Baucher, D. Geelen, F. Duchaussoy, B.C. Meyers, W. Boerjan, and F. Martin. 2008. Genome-wide identification of NBS resistance genes in Populus trichocarpa. Plant Molecular Biology 66:619-636.

Lin, X., Y. Zhang, H. Kuang, and J. Chen. 2013. Frequent loss of lineages and deficient duplications accounted for low copy number of disease resistance genes in Cucurbitaceae. BMC Genomics 14:335.

Lozano, R., O. Ponce, M. Ramirez, N. Mostajo, and G. Orjeda. 2012. Genome-wide identification and mapping of NBS-encoding resistance genes in Solanum tuberosum group phureja. PLoS One 7:0034775.

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

Marone, D., M.A. Russo, G. Laidò, A.M. De Leonardis, and A.M. Mastrangelo. 2013. Plant nucleotide binding site–leucine-rich repeat (NBS-LRR) genes: active guardians in host defense responses. International Journal of Molecular Sciences 14:7302-7326.

McGrann, G.R.D., A. Stavrinides, J. Russell, M.M. Corbitt, A. Booth, L. Chartrain, W.T.B. Thomas, and J.K.M. Brown. 2014. A trade off between mlo resistance to powdery mildew and increased susceptibility of barley to a newly important disease, Ramularia leaf spot. Journal of Experimental Botany ert452.

Meyers, B.C., A.W. Dickerman, R.W. Michelmore, S. Sivaramakrishnan, B.W. Sobral, and N.D. Young. 1999. Plant disease resistance genes encode members of an ancient and diverse protein family within the nucleotide-binding superfamily. The Plant Journal 20:317-332.

Meyers, B.C., S. Kaushik, and R.S. Nandety. 2005. Evolving disease resistance genes. Current Opinion in Plant Biology 8:129-134.

Meyers, B.C., A. Kozik, A. Griego, H. Kuang, R.W. Michelmore. 2003. Genome-wide analysis of NBS-LRR–encoding genes in Arabidopsis. The Plant Cell Online 15:809-834.

Michelmore, R. W., M. Christopoulou, and K.S. Caldwell. 2013. Impacts of resistance gene genetics, function, and evolution on a durable future. Annual Review of Phytopathology 51:291-319.

Michelmore, R.W. and B.C. Meyers. 1998. Clusters of resistance genes in plants evolve by divergent selection and a birth-and-death process. Genome Research 8:1113-1130.

Ming, R., S. Hou, Y. Feng, Q. Yu, A. Dionne-Laporte, J.H. Saw, P. Senin, W. Wang, B.V. Ly, K.L. Lewis et al. 2008. The draft genome of the transgenic tropical fruit tree papaya (Carica papaya Linnaeus). Nature 452:991-996.

Mun, J-H., H-J. Yu, S. Park, and B-S. Park. 2009. Genome-wide identification of NBS-encoding resistance genes in Brassica rapa. Molecular Genetics and Genomics 282:617-631.

Nepal, M.P., and B.V. Benson. 2015. CNL Disease Resistance Genes in Soybean and Their Evolutionary Divergence. Evolutionary Bioinformatics Online 11:49-63.

Periyannan, S., J. Moore, M. Ayliffe, U. Bansal, X. Wang, L. Huang, K. Deal, M. Luo, X. Kong, H. Bariana, R. Mago, R. McIntosh, P. Dodds, J. Dvorak, and E. Lagudah. 2013. The gene Sr33, an ortholog of barley Mla genes, encodes resistance to wheat stem rust race Ug99. Science 341:786-788.

Porter, B.W., M. Paidi, R. Ming, M. Alam, W.T. Nishijima, and Y.J. Zhu. 2009. Genome-wide analysis of Carica papaya reveals a small NBS resistance gene family. Molecular Genetics and Genomics 281:609-626.

Consortium. 2011. Potato Genome Sequencing Consortium. Genome sequence and analysis of the tuber crop potato. Nature 475:189-195.

Schmutz, J., S.B. Cannon, J. Schlueter, J. Ma, T. Mitros, W. Nelson, D.L. Hyten, Q. Song, J.J. Thelen, J. Cheng, et al. 2010. Genome sequence of the palaeopolyploid soybean. Nature 463:178-183.

Schmutz, J., P.E. McClean, S. Mamidi, G.A. Wu, S.B. Cannon, J. Grimwood, J. Jenkins, S. Shu, Q. Song, C. Chavarro, et al. 2014. A reference genome for common bean and genome-wide analysis of dual domestications. Nature Genetics 46:707-713.

Shao, F., C. Golstein, J. Ade, M. Stoutemyer, J.E. Dixon, and R.W. Innes. 2003. Cleavage of Arabidopsis PBS1 by a bacterial type III effector. Science 301:1230-1233.

Tamura, K., D. Peterson, N. Peterson, G. Stecher, M. Nei, and S. Kumar. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Molecular Biology and Evolution 28:2731-2739.

Tan, S., and S. Wu. 2012. Genome wide analysis of nucleotide-binding site disease resistance genes in Brachypodium distachyon. Comparative and Functional Genomics 2012:418208.

Tan, X., B.C. Meyers, A. Kozik, M.A. West, M. Morgante, D.A. St Clair, A.F. Bent, and R.W. Michelmore. 2007. Global expression analysis of nucleotide binding site-leucine rich repeat-encoding and related genes in Arabidopsis. BMC Plant Biology 7:56.

Tuskan, G.A., S. Difazio, S. Jansson, J. Bohlmann, I. Grigoriev, U. Hellsten, N. Putnam, S. Ralph, S. Rombauts, A. Salamov, et al. 2006. The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). Science 313:1596-1604.

Van Der Biezen, E.A., and J.D.G. Jones. 1998. Plant disease-resistance proteins and the gene-for-gene concept. Trends in Biochemical Sciences 23:454-456.

Vogel, J.P., D.F. Garvin, T.C. Mockler, J. Schmutz, D. Rokhsar, M.W. Bevan, K. Barry, S. Lucas, M. Harmon-Smith, K. Lail, et al. 2010. Genome sequencing and analysis of the model grass Brachypodium distachyon. Nature 463:763-768.

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

Wang, X., H. Wang, J. Wang, R. Sun, J. Wu, S. Liu, Y. Bai, J-H. Mun, I. Bancroft, F. Cheng, et al. 2011. The genome of the mesopolyploid crop species Brassica rapa. Nature Genetics 43:1035-1039.

Yang, S., X. Zhang, J.X. Yue, D. Tian, and J.Q. Chen. 2008. Recent duplications dominate NBS-encoding gene expansion in two woody species. Molecular Genetics and Genomics 280:187-198.

Young, N.D., F. Debellé, G.E. Oldroyd, R. Geurts, S.B. Cannon, M.K. Udvardi, V.A. Benedito, K.F. Mayer, J. Gouzy, H. Schoof, et al. 2011. The Medicago genome provides insight into the evolution of rhizobial symbioses. Nature 480:520-524.

Zhou, T., Y. Wang, J.Q. Chen, H. Araki, Z. Jing, K. Jiang, J. Shen, and D. Tian. 2004. Genome-wide identification of NBS genes in japonica rice reveals significant expansion of divergent non-TIR NBS-LRR genes. Molecular Genetics and Genomics 271:402-415.