IN SILICO ANALYSIS OF CIS-REGULATORY ELEMENTS OF DISEASE RESISTANCE GENES ACROSS SIX PLANT SPECIES

Benjamin V. Benson[1], Madhav P. Nepal[1*], Lukas G. Davison[1], Pukar B. Duwadi[1], and Brian G. Moore[2]
[1]Department of Biology & Microbiology
South Dakota State University
Brookings, SD 57007
[2]University Networking Systems and Services
South Dakota State University
Brookings, SD 57007
*Corresponding author email: Madhav.Nepal@sdstate.edu

ABSTRACT

Plant proteins encoded by disease resistance genes (R-genes) are involved in detecting pathogen-attack and subsequently in activating defense mechanisms. Coiled coil Nucleotide binding site Leucine rich region (CNL) type of R-genes are present in all plant species. In this study, we aimed to identify Cis-acting Regulatory Elements (CREs) 2kb upstream of CNL disease resistance genes and elucidate their distribution and diversity across six plant species. Over 900 identified CNL genes from six plant species—*Oryza sativa*, *Glycine max*, *Populus trichocarpa*, *Medicago truncatula*, *Phaseolus vulgaris*, and *Arabidopsis thaliana* were searched using 469 reference CREs available at the PLACE database. Using in-house Perl scripts, we parsed the sequence data to yield 327 of 469 CREs in the described region. Eight of the CREs were common to all genes, including the most-abundant DOFCOREZM, which appeared 27,619 times. Thirteen CREs had a frequency greater than 10 per gene. The only monocot included in this study, *Oryza sativa,* had a significantly lower number of CREs than dicot species. Previous studies have failed to identify a promoter that is universally present in all transcribed plant genes, but the present study identified eight CREs that appeared in the 2kb upstream region of all CNL genes sampled. Since the CREs are involved in initiating transcriptional processes, their identification would have future implication in developing durable resistance genes that are transcribed in predictable ways and that are maintained during the natural processes of reproduction of the plant, thus being useful in crop improvement.

Keywords

Cis-Acting Regulatory Element (CRE), Legumes, Coiled-Coil Nucleotide Binding site Leucine rich region (CNL), Resistance gene (R-gene)

INTRODUCTION

Plants have evolved a sophisticated multi-layered defense system against their pathogens (Hammond-Kosack and Jones 1996). The first layer of plant defense is at least one external physical barrier, such as a waxy cuticle, bark, or trichomes (Metraux et al. 2014). The next layer is the production of chemical compounds that act on fungi or bacteria, or sometimes inhibit the growth of other neighboring plants (Baetz and Martinoia 2014). If a pathogen succeeds in overcoming these two layers of defense, it must then deal with the proteins encoded by one or many resistance genes. The disease resistance gene (R-gene) proteins are classified into eight major groups, of which the largest group comprises the Nucleotide Binding Site-Leucine Rich Repeat (NBS-LRR) proteins (Liu et al. 2014). NBS-LRR proteins are divided into two subfamilies based on the domain structure at the C-terminus: genes with Toll Interleukin Receptor (TIR) are called TIR-NBS-LRR or TNL and those with putative Coiled-Coil (CC) domain are called CC-NBS-LRR or CNL (Marone et al. 2013).

R-genes and their mode of action against pathogen proteins were first proposed by Harold Flor (Flor 1971) much before many of the modern molecular techniques became available. Current wisdom holds that plant NBS-LRR proteins are triggered by a pathogen's elicitors. The sensitized host cells then send a systemic signal to activate defense responses (Gao et al. 2013a). Plant-pathogen interactions may follow the "*Gene-for-Gene Model"* as proposed by Flor (1971), which describes resistance as a function of an individual R-gene that encodes a resistance protein for a single pathogenic elicitor (Liu et al. 2014). Alternatively, NBS-LRR proteins may serve as guards of certain proteins within the signaling pathway of the plant that are targets for pathogen elicitors. When these proteins are disrupted, the guard proteins send their signal as described in the "*Guard Model*" (Jones and Dangl 2006). Regardless of the model followed, plant defense responses to pathogen infections occur through the induction of a large number of host genes. Some directly encode anti-microbial compounds, while others encode proteins with regulatory function in the defense signaling pathways (Rushton and Somssich 1998). The host genes are induced through recognition of Cis-regulatory element (CRE)-binding sites for transcription factors, such as the WRKY zinc–finger motif (Ülker and Somssich 2004). The WRKY binding elements are involved in plant defense response triggered by pathogen elicitors or by salicylic acid when directly applied to the plant (Dong et al. 2003). For example, WRKY binding elements were found to interact with both biotrophic bacterium *Pseudomonas syringae* as well as necrotrophic fungi *Botrytis cinera* and *Alternaria brassicicola* (Zheng et al. 2006) in *Arabidopsis thaliana*. The CREs are located in the non-coding regions of the DNA conserved along with the orthologous genes across species through evolutionary pressure (Kumari and Ware 2013). These elements are involved in recruiting transcription factors and thus have functional significance (Baxter et al. 2012).

Empirical studies have shown that regulation of gene expression serves as a source of evolutionary change (Bryson and Vogel 1965; Britten and Davidson 1969; King and Wilson 1975) and CREs are believed to influence phenotypic

divergence (King and Wilson 1975; Stern and Orgogozo 2008). Among the CREs, enhancers are likely to regulate the phenotypic divergence (Brown and Feder 2005), are typically located upstream, downstream or in introns (Kleinjan and van Heyningen 2005), and their genomic locations are often conserved between species (Hare et al. 2008; Cande et al. 2009; Kalay and Wittkopp 2010). A clear understanding of the evolutionary divergence of the CREs requires study of allele-specific expression (Cowles et al 2002;Wittkopp et al. 2004), determination of functionally divergent sites, and their interaction with and among trans-acting elements (Wittkopp and Kalay 2012; Gao et al. 2013b). Previous studies in *Medicago truncatula* showed that there were four over represented regulatory WBOX cassettes associated with the WRKY transcription factors, CBF and DRE boxes, and GCC motif associated with ERF-type transcription factors (Ameline-Torregrosa et al. 2008). Similar studies on the representation of the WRKY transcriptions factors DNA binding elements in *Arabidopsis* and in grape (*Vitis vinifera*) (Marchive et al. 2007; Zheng et al. 2007) have been conducted, while other CREs are yet to be explored. The main objectives of this study were to identify CREs in 2kb upstream of CNL type of R-genes and elucidate their distribution and diversity across six plant species. The resulting identification of the CNL genes and their CREs across species will allow an understanding of the diversity, distribution and evolutionary relationships of these genes.

METHODS

We gathered from Phytozome.net 55 previously-identified CNL genes of *Arabidopsis thaliana* (hereafter AT; Meyers et al. 2003) as reference sequences, and mined the genome sequences of five plant species (*Glycine max, Medicago truncatula, Oryza sativa, Phaseolus vulgaris* and *Populus trichocarpa*, hereafter called GM, OS, PV and PT, respectively. AT CNL gene sequences were used to build a Hidden Markov Model (HMM) profile similar to those employed in *Arabidopsis* (Meyers et al. 2003) and in *Medicago* (Ameline-Torregrosa et al. 2008). Phylogenetic analyses of the NB-ARC (NBS) sequences were performed using *Streptomyces* protein sequence P25941 as outgroup. Phylogenetic analysis was performed in the program MEGA5.2 using Maximum Likelihood method with the best fit model JTT+G. Branch support was estimated for 100 bootstrap replicates. The two thousand base pairs (2kb) upstream region was searched for the CRE regions, and the identification protocol was similar to *Medicago* (Ameline-Torregrosa et al. 2008). SIGNALSCAN program available at PLACE database (Higo et al. 1999) was used for the identification of the CRE regions. Custom Perl scripts were used to parse the output files from the PLACE database. Using the data from PLACE, we estimated the number and abundance of identified CREs across 913 plant CNL genes. One-way ANOVA was conducted to test the statistical difference between the number of CREs in the genomes of monocot and dicot species.

RESULTS AND DISCUSSION

Using *in silico* analysis, we identified 912 genes that included 149, 188, 194, 235 and 94 CNL genes in rice, soybean, poplar, alfalfa, and common bean, respectively (Figure 1). As shown in the figure, these genes were nested into four clades (CNL-A, CNL-B, CNL-C and CNL-D) consistent to those described in *Arabidopsis*. Clade A had two unresolved subgroups each with moderate BS support and Clade B was moderately supported (BS 75%). A previous study (Meyers et al. 2003) described a proportionally-similar number of genes in each group in *Arabidopsis*. Clade A contained between 0% (OS) and10.4% (AT) of the total CNL genes identified within each species. Clade B was moderately supported (BS 75%) containing 2% (OS) and 50% (AT) of the CNL genes within each species. A previous study (Meyers et al. 2003) described a proportionally similar number of genes in each group in *Arabidopsis*. Among the four clades, CNL-C was the most expanded clade. The *Arabidopsis* genome contained only 14.5% CNL-C genes while the other genomes contained a much higher proportion (69.3% [PT] - 97.9% [OS]). The CNL-C clade contained multiple clusters of genes with low clade support, indicating rapidly-evolving genes. The expansion of clade C perhaps provides a source for new CNL sequences while concurrently reducing the risk of auto-activation of the resistance response through the reduction of gene expression. This is consistent with the results reported in *Arabidopsis* (Meyers et al. 2003) and *Medicago* (Ameline-Torregrosa et al. 2008). CNL-D was the least expanded clade, with a strong statistical support (BS 89%). We found that the number of CNL R-genes in GM was closer to the predicted 1.05 gene copy retention after duplication than the 3.1 gene copy number of the overall genome (Ashfield et al. 2012). Despite possessing nearly double the genome size, GM (1.1 Gb genome size and 188 CNL gene) had 1.25 times the number of CNL genes when compared to *Medicago* (500 Mb genome size and 235 CNL genes), suggesting little adherence to the assumption that more duplications or a larger genome would allow for more resistance genes. What affected these genome duplication retention rates, and to what extent were they modified by the effects of selection pressures from pathogens and auto-activation in soybeans are intriguing questions to be investigated in the future. Addressing these questions requires an understanding of how these genes are regulated in the genomes. Below we report our results on Cis-Regulatory Elements (CREs) of the CNL-genes across six plant species.

Among the 469 CREs investigated in the 2kb upstream region of the 913 CNL genes, 327 CREs (~70%) were found across the six plant genomes surveyed. Among these CREs, 253, 264, 271, 281, 283, and 292 CREs were present in AT, PV, GM, MT, PT and OS, respectively. Eight of these CREs (ARR1AT, CAATBOX1, CACTFTPPCA1, DOFCOREZM, GATABOX, GT-1CONSENSUS, POLLEN1LELAT52 and WRKY710S) were found common to all CNL genes, including the most-frequent, DOFCOREZM (DNA binding one finger zinc of *Zea maize*), which appeared a total of 27,619 times. Further analysis of multiple plant genomes at various taxonomic groups is warranted to test a previous claim (Juven-Gershon et al. 2006) that plant genomes lack a

universal regulatory element. One of the most commonly occurring CREs in the present study, DOFCOREMZ is the core site required for binding of DOF proteins in maize (*Zea mays*). The DOF proteins are DNA binding proteins with only one zinc finger and are unique to plants (Higo et al. 1999). These single zinc finger domains are related to the WRKY binding site and part of the regulatory network that was previously described (Marchive et al. 2007; Zheng et al. 2007). In the present study, there were six WBOX elements identified: WBBOXPCWRKY1, WBOXATNPR1, WBOXGACAD1A, WBOXHVISO1, WBOXNTCHN48, and WBOXNTERF3. One of these elements appeared in every sequence, with many having three to four different elements appearing in the same sequence. These WBOX elements were found to be involved in R-gene regulation in grape and *Arabidopsis* (Marchive et al. 2007; Zheng et al. 2007). The conservation of the WBOX motif indicates its importance in gene regulation, and the search for less conserved elements involved in gene regulation should continue (Wittkopp and Kalay 2012).

The average occurrence of 13 most common CREs in the present study is shown in Table 1. Evolutionary conservation of these CREs across plant species must have regulatory roles on the gene sequences when they are present. Eighty-one of the 327 CREs were present once per sequence in which they appeared. Further study of the positioning would reveal their functional significance. The average number of CREs in rice (a monocot) was significantly lower ($P < 0.0001$) than that in dicot species (Figure 2). These results correspond to the number of Core Promoters Elements (CPE), a subset of CREs reported previously (Kumari and Ware 2013), where the authors have shown that monocot core promoters

Table 1. ***The most commonly-occurring CREs. These 13 CREs appeared on average greater than 10 times per gene.***

CRE ID	Total Appearances in all 913 Genes	Total Genes the CRE Region Appears	Average Number per Gene
ARR1AT	19,857	913	21.75
CAATBOX1	22,877	913	25.06
CACTFTPPCA1	25,749	913	28.2
DOFCOREZM	27,619	913	30.25
GATABOX	15,478	913	16.95
GT1CONSENSUS	18,124	913	19.85
POLLEN1LELAT52	12,074	913	13.23
WRKY710S	10,515	913	11.52
GTGANTG10	11,922	912	13.07
ROOTMOTIFTAPOX1	16,589	910	18.23
EBOXBNNAPA	13,052	908	14.37
MYCCONSENSUSAT	13,052	908	14.37
TATABOX5	9,223	905	10.19

had lower DNA free energy than dicot core promoters. The free energy of the DNA sequences of these genes may be associated with GC content which is reported to be less in dicot genomes than that in monocots genomes (Serres-Giardi et al. 2012). Further investigation across multiple dicot and monocot genomes representing major taxonomic groups is required to confirm if this trend holds true as well as to understand the functional correlation.

One of the major caveats of the *in silico* analysis such as presented in this paper is the detection of false positives. For example, in soybean when the start position of the sequences that contained –300CORE (a CRE that has regulatory function when found close to 300 nucleotides before the transcriptional start site) was visualized, we found that the starting position of only 5 of the 23 sequences were found within 50 bases of the reported -300 nucleotide starting position from the transcriptional start site. The detection of potential false positives during any CRE prediction/ identification process may be alleviated by using more rigorous prediction methods that take into account the presence of other genes within the upstream region, by looking at the distance of the CRE from the TSS (Transcriptional Start Site) and by analyzing DNA free energy profiles (Kumari and Ware 2013). These factors are likely to give some insight

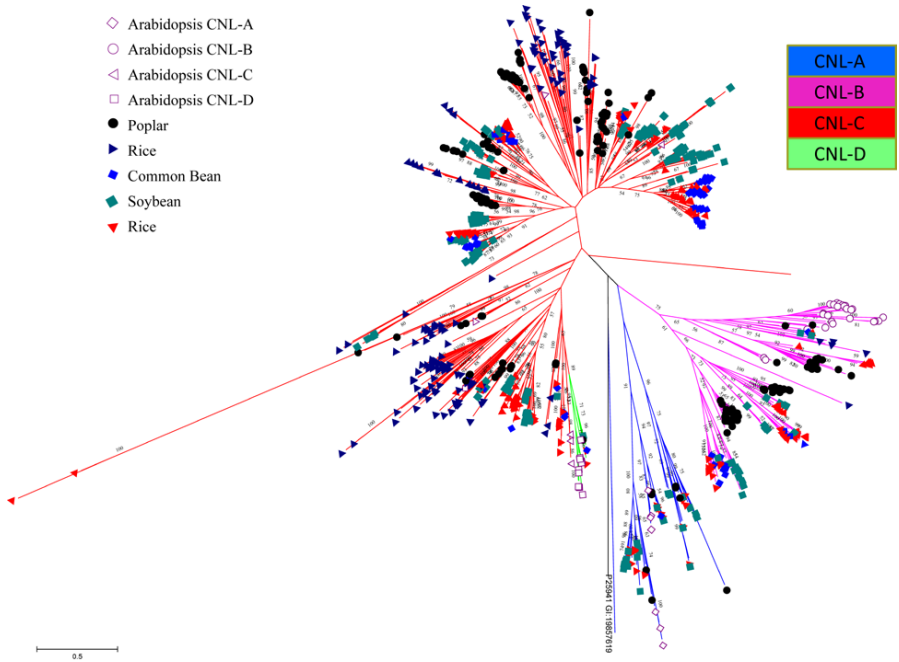

Figure 1. ***Phylogenetic analysis of the CNL genes from six species:*** *A. thaliana O. sativa, M. truncatula, P. vulgaris, P. trichocarpa, and G. max.* ***A Maximum Likelihood tree was constructed using the program MEGA 5. Branch support was estimated using the bootstrap method for 100 replicates.*** *Streptomyces* ***(GBI:19857619) was used as an outgroup. The species are color-coded. Hollow shapes were used to identify the*** *Arabidopsis* ***sequences previously assigned the CNL identifiers. The CNL clades are also color-coded: CNL-A, CNL-B, CNL-C and CNL-D in blue, purple, red and green, respectively.***

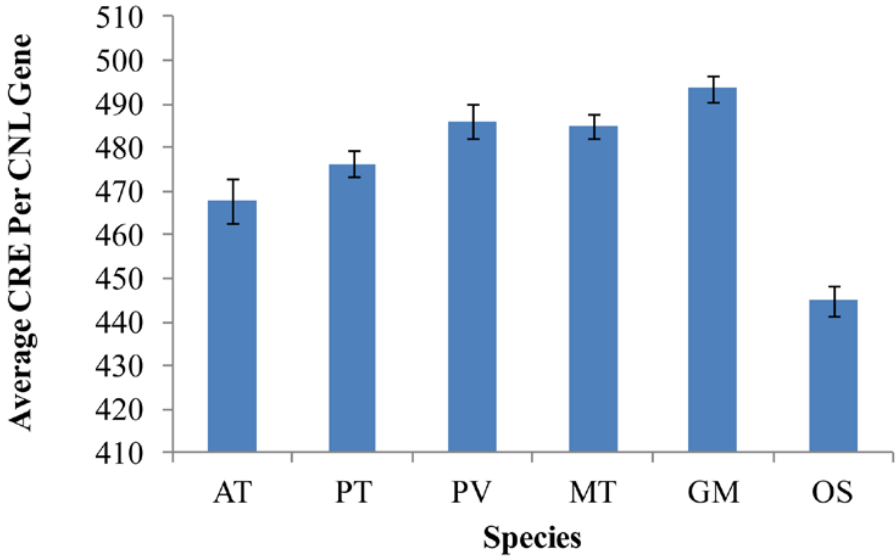

Figure 2. ***Average number of CREs per CNL gene across six species. The species from left to right are AT =*** *Arabidopsis thaliana,* ***PT =*** *Populus trichocarpa,* ***PV =*** *Phaseolus vulgaris,* ***MT =*** *Medicago trunatula,* ***GM*** *= Glycine max, and* ***OS*** *= Oryza sativa.*

as to why the monocot and dicot genomes differ beyond the specific nucleotide biases. Identification of CNL R-genes and insights into their regulatory elements presented in this project would have future implication in developing durable resistance genes that are transcribed in predictable ways and maintained during the natural processes of reproduction of the plant thus being useful in crop improvement.

ACKNOWLEDGEMENTS

This project was supported by South Dakota Agricultural Experiment Station (SDAES), Undergraduate Research Support fund from the Department of Biology and Microbiology at South Dakota State University, South Dakota Soybean Research and Promotional Council and USDA-NIFA Hatch Project Fund to M. Nepal. Co-author Lukas Davison was enrolled in M. Nepal's section of BIOL 498 (Undergraduate Research and Scholarship) course in fall 2013.

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.

Ashfield, T., A.N. Egan, B.E. Pfeil, N.W. Chen, R. Podicheti, M.B. Ratnaparkhe, C. Ameline-Torregrosa, R. Denny, S. Cannon, and J.J. Doyle. 2012. Evolution of a complex disease resistance gene cluster in diploid Phaseolus and tetraploid Glycine. Plant Physiology 159:336-354.

Baetz, U., and E. Martinoia. 2014. Root exudates: the hidden part of plant defense. Trends in Plant Science 19:90-98.

Baxter, L., A. Jironkin, R. Hickman, J. Moore, C. Barrington, P. Krusche, N.P. Dyer, V. Buchanan-Wollaston, A. Tiskin, and J. Beynon. 2012. Conserved noncoding sequences highlight shared components of regulatory networks in dicotyledonous plants. The Plant Cell Online 24:3949-3965.

Britten, R.J., and E.H. Davidson. 1969. Gene Regulation for Higher Cells: A Theory. Science 165:349-357.

Brown, R.P., and M.E. Feder. 2005. Reverse transcriptional profiling: noncorrespondence of transcript level variation and proximal promoter polymorphism. BMC Genomics 6:110.

Bryson, V., and H.J. Vogel. 1965. Evolving genes and proteins. Science 147:68-71.

Cande, J., Y. Goltsev, and M.S. Levine. 2009. Conservation of enhancer location in divergent insects. Proceedings of the National Academy of Sciences 106:14414-14419.

Cowles, C.R., J.N. Hirschhorn, D. Altshuler, and E.S. Lander. 2002. Detection of regulatory variation in mouse genes. Nature Genetics 32:432-437.

Dong, J., C. Chen, and Z. Chen. 2003. Expression profiles of the Arabidopsis WRKY gene superfamily during plant defense response. Plant Molecular Biology 51:21-37.

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

Gao, X., X. Chen, W. Lin, S. Chen, D. Lu, Y. Niu, L. Li, C. Cheng, M. McCormack and J. Sheen. 2013a. Bifurcation of *Arabidopsis* NLR immune signaling via Ca2+-dependent protein kinases. PLoS Pathogens 9:e1003127.

Gao, Z., R. Zhao, and J. Ruan. 2013b. A genome-wide cis-regulatory element discovery method based on promoter sequences and gene co-expression networks. BMC Genomics 14:S4.

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

Hare, E.E., B.K. Peterson, V.N. Iyer, R. Meier, and M.B. Eisen. 2008. Sepsid even-skipped enhancers are functionally conserved in Drosophila despite lack of sequence conservation. Plos Genetics 4:e1000106.

Higo, K.,Y. Ugawa, M. Iwamoto, and T. Korenaga. 1999. Plant cis-acting regulatory DNA elements (PLACE) database: 1999. Nucleic Acids Research 27:297-300.

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

Kalay, G., and P.J. Wittkopp. 2010. Nomadic enhancers: tissue-specific cis-regulatory elements of yellow have divergent genomic positions among Drosophila species. Plos Genetics 6(11):e1001222

King, M.C., and A.C. Wilson. 1975. Evolution at two levels in humans and chimpanzees. Science 188:107-116.

Kleinjan, D.A. and , V. van Heyningen. 2005. Long-range control of gene expression: emerging mechanisms and disruption in disease. The American Journal of Human Genetics 76:8-32.

Kumari, S., and D. Ware. 2013. Genome-Wide Computational Prediction and Analysis of Core Promoter Elements across Plant Monocots and Dicots. Plos One 8:e79011.

Liu, W., L. Triplett, J. Liu, J.E. Leach, and G.L. Wang. 2014. Novel Insights into Rice Innate Immunity against Bacterial and Fungal Pathogens. Annual Review of Phytopathology 52:DOI 10.1146/annurev-phyto-102313-045926

Marchive, C., R. Mzid, L. Deluc, F. Barrieu, J. Pirrello, A.Gauthier, M.F. Corio-Costet, F. Regad, B. Cailleteau, and S. Hamdi. 2007. Isolation and characterization of a Vitis vinifera transcription factor, VvWRKY1, and its effect on responses to fungal pathogens in transgenic tobacco plants. Journal of Experimental Botany 58:1999-2010.

Metraux, J., M.Serrano, M. Torres, F. Coluccia, and F. L'Haridon. 2014. The cuticle and plant defense to pathogens. Frontiers in Plant Science 5:274.

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

Rushton, P.J., and I.E. Somssich. 1998. Transcriptional control of plant genes responsive to pathogens. Current Opinion in Plant Biology 1:311-315.

Stern, D.L., and V. Orgogozo. 2008. The loci of evolution: how predictable is genetic evolution? Evolution 62:2155-2177.

Ülker, B., and I.E. Somssich. 2004. WRKY transcription factors: from DNA binding towards biological function. Current Opinion in Plant Biology 7:491-498.

Wittkopp, P.J., B.K. Haerum, and A.G. Clark. 2004. Evolutionary changes in cis and trans gene regulation. Nature 430:85-88.

Wittkopp, P.J., and G. Kalay. 2012. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. Nature Reviews Genetics 13:59-69.

Zheng, Z., S.L. Mosher, B. Fan, D.F. Klessig, and Z. Chen. 2007. Functional analysis of Arabidopsis WRKY25 transcription factor in plant defense against Pseudomonas syringae. BMC Plant Biology 7:2. DOI: 10.1186/1471-2229-7-2.

Zheng, Z., S.A. Qamar, Z. Chen, and T. Mengiste. 2006. *Arabidopsis* WRKY33 transcription factor is required for resistance to necrotrophic fungal pathogens. The Plant Journal 48:592-605.