RESOURCE

Patterns of divergence of a large family of nodule cysteine-rich peptides in accessions of Medicago truncatula

Sumitha Nallu1,*, Kevin A. T. Silverstein1,†, Peng Zhou2, Nevin D. Young2 and Kathryn A. VandenBosch1,§

1Department of Plant Biology, University of Minnesota, 250 Biological Sciences, 1445 Gortner Avenue, Saint Paul, MN 55108, USA and
2Department of Plant Pathology, University of Minnesota, St. Paul, MN 55108, USA

Received 24 November 2013; revised 20 February 2014; accepted 4 March 2014; published online 17 March 2014.
*For correspondence (e-mail sumitha@uchicago.edu).
†Present address: Ecology and Evolution, Erman Biological Center, University of Chicago, Chicago, IL 60637, USA.
‡Present address: Supercomputing Institute for Advanced Computational Research, University of Minnesota, Minneapolis, MN 55455, USA.
§Present address: College of Agricultural and Life Sciences, 1450 Linden Drive, Madison, WI 53706, USA.

SUMMARY

The nodule cysteine-rich (NCR) groups of defensin-like (DEFL) genes are one of the largest gene families expressed in the nodules of some legume plants. They have only been observed in the inverted repeat loss clade (IRLC) of legumes, which includes the model legume Medicago truncatula. NCRs are reported to play an important role in plant–microbe interactions. To understand their diversity we analyzed their expression and sequence polymorphisms among four accessions of M. truncatula. A significant expression and nucleotide variation was observed among the genes. We then used 26 accessions to estimate the selection pressures shaping evolution among the accessions by calculating the nucleotide diversity at non-synonymous and synonymous sites in the coding region. The mature peptides of the orthologous NCRs had signatures of both purifying and diversifying selection pressures, unlike the seed DEFLs, which predominantly exhibited purifying selection. The expression, sequence variation and apparent diversifying selection in NCRs within the Medicago species indicates rapid and recent evolution, and suggests that this family of genes is actively evolving to adapt to different environments and is acquiring new functions.

Keywords: NCRs, selection pressures on NCRs, Medicago truncatula accessions, expression polymorphism in NCRs, diversifying selection in NCRs.

INTRODUCTION

Recent advances in genomic data collection and analysis have facilitated new prospects for understanding plant function and evolution. Medicago truncatula was chosen as a model species for the study of legume symbiotic nitrogen fixation because of its relatively small, diploid genome, rapid generation time and self-pollination (Cook, 1999). Through an international collaborative effort, multiple tools and resources, such as the publication of the complete genome and the release of gene model Mt4.0v.1 (http://jcvi.org/medicago/index.php), are currently available for molecular, genetic and evolutionary studies in M. truncatula (Branca et al., 2011; Young et al., 2011).

The availability of sequence information facilitates the study of large and diverse gene families. A large diverse family of cysteine-rich proteins was discovered in the nodules of M. truncatula (Fedorova et al., 2002; Mergaert et al., 2003; Graham et al., 2004), and these were subsequently found to be expressed in other developmental stages, including seeds, of legumes and other dicots (Silverstein et al., 2005, 2007). Because of the similarity of the gene sequence and structure of these cysteine-rich proteins to defensins, this gene family has been called defensin-like genes (DEFLs; Graham et al., 2004; Silverstein et al., 2005, 2007). Using the Affymetrix M. truncatula Genome Array, and later a more comprehensive DEFL array, the majority of the DEFLs were found to be expressed in the nodules of M. truncatula (Benedito et al., 2008; Tesfaye et al., 2013). The 566 nodule cysteine-rich (NCR) groups of DEFL genes that are expressed in nodules have unique expression and regulatory patterns (Nallu et al., 2013). Using a custom
DEFL array for expression profiling across various nodule developmental stages, and in response to different rhizobial mutants, we observed the NCR expression levels to be dependent on the volume of rhizobia present in the nodule. We also reported the presence of unique conserved regulatory elements in their 1–Kb upstream regions.

To date, NCRs have been found only in legumes belonging to the inverted repeat loss clade (IRLC) within the subfamily Papilionoideae, and appear to be absent from Lotus japonicus and Glycine max, which occur in other clades within the Papilionoideae (Graham et al., 2004; Alunni et al., 2007), and which are estimated to have diverged from the IRLC approximately 39 million years ago (Cannon et al., 2009). When rhizobia are released into the symbiosome compartment in determinate nodules of L. japonicus and G. max, the rhizobia retain the cell and genome size of free-living rhizobia, whereas in M. truncatula and related indeterminate nodule-forming legumes, the rhizobia terminally differentiate into enlarged bacteroids with an amplified genome (Mergaert et al., 2005). The correlation between NCR expression in nodules and bacteroid differentiation appears to be causally related. Recently it was demonstrated that when one of the M. truncatula NCRs was expressed in L. japonicus, terminal bacteroid differentiation resulted (Van de Velde et al., 2010).

As known NCRs are restricted to a single clade within legumes and paralogs show evidence of diversification (Graham et al., 2004; Alunni et al., 2007), we initiated a study to examine the expression and sequence polymorphism of NCRs within the accessions of M. truncatula. In the last decade, studies have addressed DNA polymorphisms within species using DNA markers like microsatellites, single-nucleotide polymorphisms (SNPs) and single-feature polymorphisms (SFPs) in Arabidopsis thaliana (Zhu et al., 2003; Schmid et al., 2005; Borevitz et al., 2007) and M. truncatula (Ronfort et al., 2006); however, identifying sequence variation is only the first step in studying the evolutionary patterns of gene families. The availability of technology in recent years has prompted researchers to correlate the DNA variation with variation in the transcriptome and metabolome.

Environmental effects, sequence polymorphism in coding region, cis and trans regulatory elements, and transcription factor binding sites can cause gene expression variation among accessions. There are various reports correlating cis- and trans-acting promoter polymorphisms with expression differences (Cowles et al., 2002; Rockman and Wray, 2002; Caicedo et al., 2004). Nucleotide variation in the binding sites of microRNAs (miRNAs) and small interfering RNAs (siRNAs), or in the 3′ untranslated region (3′–UTR) of the gene, can also affect gene expression levels (Mignone et al., 2002; Wilkie et al., 2003). Nucleotide polymorphism in coding regions leads to synonymous and non-synonymous substitutions in amino acids. There have been various studies on the intraspecific variation in the loci belonging to the nucleotide-binding site leucine-rich repeat (NBS-LRR) gene family in A. thaliana (Tian et al., 2002; Mauricio et al., 2003; Rose et al., 2004). Nucleotide diversity measurements in different regions of the NBS-LRR loci have been helpful to understand the selection pressures and evolutionary mechanisms of this large family of defense genes.

In this study, we explored the expression and sequence polymorphisms of NCRs among four accessions of M. truncatula. We also used two different rhizobial strains to assess the NCR expression differences caused by host-symbiont specificity. The NCRs displayed significant expression differences among the accessions, but had subtle differences in gene expression because of strain difference. The sequence variation was estimated in terms of SFPs and SNPs. The pattern of DNA variation among the accessions correlated with the pattern observed in expression polymorphism. In addition to polymorphisms in the coding region that may lead to expression differences, there was variation observed in the upstream 1–kb putative promoter, downstream 1–kb putative 3′–UTR and in a few introns. When examined closely some of the upstream SNPs mapped to the earlier reported cis-acting elements in upstream regions of NCRs. By estimating the nucleotide diversity of NCRs among 26 accessions and comparing them with the seed DEFLs, we observed that the mature peptide of NCRs has signatures of both diversifying and purifying selection in M. truncatula.

RESULTS

NCR expression is polymorphic among different accessions of M. truncatula

To characterize the difference in expression levels of NCRs, we selected four different accessions of M. truncatula: the reference genome Jemalong A17, plus DZA315–16, F83005–5 and R108–C3. They were inoculated with strains of two rhizobial species: Sinorhizobium meliloti Sm1021 (Meade et al., 1982) or Sinorhizobium medicae A321 (Rome et al., 1996) to evaluate the expression differences caused by host-symbiont interaction. Strain S. medicae A321 was also chosen on the basis of an earlier report that S. meliloti Sm1021 is a poor nitrogen-fixing partner for A17 (Terpolilli et al., 2008). A two-factorial analysis of variance (two-way ANOVA) indicated expression differences in NCRs as an effect of accession, accession × strain interaction and a subtle effect of strain alone (Figure 1). Using the variance shrinkage method (Cui et al., 2005a,b), we further calculated differentially expressed genes in the accession × strain effect, and found that the differentially expressed genes had subtle fold changes (all very close to a fold change of two), supporting a conclusion that a difference in strains used for inoculation contributes minimally to expression differences among NCRs.

© 2014 The Authors.
The Plant Journal published by Society for Experimental Biology and John Wiley & Sons Ltd.,
The Plant Journal, (2014), 78, 697–705
To calculate the number of NCRs that were differentially expressed among *M. truncatula* under each of the inoculation regimes, we performed the variance shrinkage analysis on different pairs of accessions (Table S1). R108–C3 showed the highest number of differentially expressed genes in comparison with A17 under both inoculation regimes, followed by F83005–5 and the least differentially expressed genes are seen in DZA315–16 (Figure 2).

**Coding sequence polymorphisms among orthologous NCRs**

We used the 25-bp oligonucleotide features of the custom DEFL chip to estimate sequence polymorphisms in NCR coding sequences in pairwise comparisons among the four accessions. A common set of genes with SFPs was extracted from the analysis of the two inoculation regimes for each accession pair. As seen in the case of expression polymorphism patterns, R108–C3 exhibited the highest sequence polymorphism, followed by F83005–5 with the lowest sequence polymorphism seen in DZA315–16, in comparison with A17 (Table 1). From the current *M. truncatula* hapmap project, we were able to extract information on coding region SNPs for 318 NCRs for DZA315–16, F83005–5 or R108–C3 in reference to the A17 genome. The pattern of sequence polymorphism seen with SNP analysis was consistent with the pattern observed by SFP detection and expression polymorphism. R108–C3 again exhibited the highest occurrence of polymorphisms in comparison with A17 (Figure 3). A random set of 20 NCRs, with SFPs from A17 versus the DZA315–16 accession, were selected to confirm the SFP method of detection. We were able to confirm 68% of SFP calls by comparing them with the available SNP data (Table S2); however, the 32% of SFPs that could be not validated may not represent the true false discovery rate (FDR). The *Medicago* hapmap project is ongoing and there are a few missing sequence gaps, especially in highly duplicated genes like NCRs. Hence, with the availability of updated sequence information, there is a possibility of validating a higher percentage of SFPs. In addition, even with the aforementioned limitations, our FDR is lower than the previously reported 40% FDR (Rostoks et al., 2005; Xu et al., 2009).

**Figure 1.** Number of differentially expressing NCRs with different effects. The number of differentially expressed NCRs calculated using a two-way ANOVA at \( P \leq 0.01 \) with the effect of accession, accession \( \times \) strain interaction and a subtle effect of strain alone are represented as a Venn diagram.

**Figure 2.** Differential expressions of NCRs among *Medicago truncatula* accessions. Numbers indicate the NCRs that are differentially expressed in reference to A17 when inoculated with (a) *Sinorhizobium medicae* or (b) *Sinorhizobium meliloti*. The differentially expressed NCRs were estimated using the variance shrinkage method with \( P \leq 0.01 \) and a fold change cut-off of 2.0.

**Figure 3.** Number of single-nucleotide polymorphisms (SNPs) observed in different accessions. The Venn diagram represents the SNPs in the coding regions for 318 NCRs extracted from the three accessions: DZA315–16, F83005–5 or R108–C3, in reference to the A17 genome.

© 2014 The Authors, *The Plant Journal* published by Society for Experimental Biology and John Wiley & Sons Ltd., *The Plant Journal*, (2014), 78, 697–705

**Table 1.** NCRs with single-feature polymorphisms observed between pairs of accessions

|        | A17 | DZA315–16 | F83005–5 |
|--------|-----|-----------|----------|
| DZA315–16 | 202 |           |          |
| F83005–5   | 220 | 156       |          |
| R108–C3    | 388 | 418       | 411      |

The data indicates the set of NCRs containing SFPs that were common between the two inoculation regimes for each of the pairwise comparison of *Medicago truncatula* accessions.
Sequence polymorphisms in cis-regulatory regions of NCRs

The expression polymorphisms observed among the pairwise comparisons of the accessions could be true expression-level polymorphisms (ELPs; Doerge, 2002), or could result from sequence polymorphisms that cause a probe mismatch between the accessions used and the probe sequence on the DEFL array, which would result in a signal difference that inaccurately predicts expression differences. We categorized the NCRs into four groups: (i) differentially expressed genes with SFPs; (ii) differentially expressed genes with no SFPs; (iii) genes with SFPs but not differentially expressed; and (iv) genes with no SFPs and no differential expression for pairwise comparisons of accessions to A17, when inoculated with two different strains (Table 2). We further filtered the group of differentially expressed genes with no SFPs using the SNP data. The results from the DEFL array should be the most accurate representation of gene expression for the group of NCRs that lack evidence of sequence polymorphisms in their coding regions.

For NCRs that are differentially expressed but have no SFPs or SNPs in their coding regions, we mapped the occurrence of SNPs onto their 1-kb upstream, downstream regions from the start and stop translational sites of the NCRs, respectively, and also included the intronic regions if present in the genes. This set of differentially expressed genes exhibited SNPs and indels in other regions of the NCRs that are required for nodule-specific expression (Nallu et al., 2013). Some of the SNPs in the upstream regions of the NCRs were mapped onto the cis-acting regulatory elements, suggesting that these nucleotide changes in the regulatory motifs might be one of the reasons for the observed differential expression of NCRs (Figure 4; Table S3).

Nucleotide diversity of NCRs in M. truncatula

The DEFLs were classified as either NCRs or seed DEFLs based on expression patterns, following the criteria described in the Experimental procedures. To determine the evolutionary pressures acting on NCRs, the total nucleotide diversity (\(d_H\)) and the average nucleotide diversity (\(d\)) were estimated at synonymous and nonsynonymous sites for 26 accessions of M. truncatula. The 26 accessions were sampled from geographically distinct populations and were deeply sequenced, with an average coverage of \(-15X\) (Table S4). As the Medicago hapmap project is still in progress the SNP data, especially for highly duplicated genes like DEFLs, are incomplete (Branca et al., 2011). We obtained sequence information for a total of 263 NCRs and 43 seed DEFLs, providing a significant representative set of genes for each group. The \(\Theta_{\text{syn}}\), \(\pi_{\text{syn}}\), and \(\Theta_{\text{non}}\), \(\pi_{\text{non}}\) were calculated for the entire coding region, and separately for regions that code for the signal and mature peptides of the NCRs and seed DEFLs. From the nonsyn/syn ratios observed in different regions of NCRs and seed DEFLs, the signal peptides of both the NCRs and seed DEFLs were conserved, with more than half displaying no substitutions and the majority of the other half exhibiting purifying selection (Figure 5a). Mature polypeptides of NCRs exhibit evidence of both purifying and diversifying selection, whereas the mature polypeptides of seed DEFLs are mostly under purifying selection (Figure 5b). To test the statistical significance of the differences in rates of substitutions among different regions in NCRs and seed DEFLs, we designated all genes with nonsyn/syn > 1 as 1, nonsyn = syn as 0 and nonsyn/syn < 1 as -1, and employed the Mann–Whitney U-test. We could not use the raw nonsyn/syn values for analysis because many of the genes had no synonymous substitutions, and in such cases nonsyn/
syn would give an undefined number. The Mann–Whitney U-tests indicated a significant difference within the full coding regions and the mature polypeptides of NCRs versus seed DEFLs. There was no significant difference in substitution rates observed between the signal peptides of NCRs versus seed DEFLs, or between the signal peptides versus mature polypeptides of the seed DEFLs. Among NCRs, the mature polypeptides were significantly different from the signal peptides in their selection pressures (Table 3).

**DISCUSSION**

**NCRs have similar trends in expression and sequence polymorphism in *M. truncatula***

The four accessions used in this study were selected because they were from different ecological zones (Ronfort et al., 2006). A17 is the reference genome for the *M. truncatula* sequencing project (http://www.medicago.org/genome), R108 has been the common choice for generating *Medicago* insertion mutants, and both DZA315-16 and F83005-5, along with A17, have been used as parents for recombinant inbred lines (RILs). Additionally, all four accessions were resequenced under the Medicago HapMap Project (http://www.medicagohapmap.org). We found a significant number of differentially expressed NCRs among the four accessions of *M. truncatula*. Potentially, differences in NCR gene expression patterns in *M. truncatula* accessions could result from differences in host/endophyte specificity, resulting in global differences in nodule effectiveness and nodulin expression, rather

| Table 3 Mann–Whitney P values of NCRs and seed DEFLs |
|------------------------------------------------------|
|                                                       |
| Coding region                                         |
| NCRs-seed DEFLs                                       |
| NCRs-NCRs                                            |
| Seed DEFLs-seed                                       |
| Coding region                                         | 0.00045*** | – | – |
| Signal peptide                                       | 0.61       | – | – |
| Mature peptide                                       | 0.0023***  | – | – |
| Signal-mature peptide                                | –          | 0.00087*** | 0.38 |

The nucleotide diversity (θ and π) was calculated in the predicted signal peptide, mature peptide and total coding regions of the NCRs and seed DEFLs. The data represent *P* values from the Mann–Whitney *U*-test, employed to calculate the statistical significance of diversity in each region among the NCRs, seed DEFLs and NCRs-seed DEFLs. **Significance at *P* ≤ 0.05, two-tailed test. It is observed that the total coding and mature peptide of NCRs is significantly different from seed DEFLs in their values.

Project (http://www.medicagohapmap.org). We found a significant number of differentially expressed NCRs among the four accessions of *M. truncatula*.
than specific differences in NCR regulation. To address this possibility, we inoculated each ecotype with two different strains of rhizobia to evaluate the difference in expression patterns of NCRs in response to different symbiotic partners. *Sinorhizobium meliloti* 1021 has been a common strain of study for symbiosis in *M. truncatula*, as a result of the extensive analysis of symbiotic genes in this genetic background and the available genome sequence (Galibert et al., 2001). *Sinorhizobium meliloti* 1021 has been described as an inefficient partner in comparison with the other gene families among the four gene families that have a significant higher sequence variation in *M. truncatula* 26 that observed in expression polymorphism. This trend was similar to that observed in *NCR* expression variation in the four accessions, which is reported to show high compatibility with *M. truncatula*. The interaction between *S. medicae* A321 and *M. truncatula* results in the formation of deeply colored nodules with an increased dry mass, and increased chlorophyll content and biomass of the plant: all indicators of an efficient nitrogen fixation interaction (Sugawara et al., 2013).

Inoculation with these two strains resulted in minimal differences in the expression patterns of NCRs, suggesting that these strains, which are expected to differ in nitrogen fixing abilities, are not a significant factor in *NCR* induction. In support of this possibility, we have previously seen in Nallu et al. (2013) that the *NCR* expression levels differ little between nodules infected by wild-type Sm1021 and nodules infected with a nifH mutant that is deficient in fixing nitrogen (Hirs, et al., 1983).

In addition to expression polymorphisms among the accessions, we hypothesized that the NCRs would also demonstrate sequence polymorphism among the accessions. We used two approaches to evaluate *NCR* sequence polymorphism. Single-feature polymorphism (SFP) is a rapid and cost-effective technique for scanning genomic polymorphism, especially in species for which the genomic sequence is unknown or under construction (Hazen and Kay, 2003). First, we used the hybridization results from our analysis of *NCR* expression using the custom *DEFL* array to scan for SFPs among the four accessions. Second, we used the available genome sequence from the Medicago HapMap project for the four accessions. Among the four accessions, R108-C3 had the highest sequence variation in comparison with A17 under both inoculation regimes, followed by F83005–5 and the least sequence variation was seen with DZA315-16. This trend was similar to that observed in expression polymorphism.

The recent whole-genome nucleotide diversity study of 26 *M. truncatula* accessions shows that the NCRs are one among the four gene families that have a significant higher divergence in comparison with the other gene families (Branca et al., 2011). From the phylogenetic tree constructed by Branca et al. (2011), based on the chloroplast genomes of the 26 *M. truncatula* accessions, it is observed that the reference genome A17 (HM101) is closest to DZA315-16 (HM005), followed by F83005–5 (HM006) and R108-C3 (HM029), which is identified as a different subspecies and placed into a different clade from *M. truncatula*. These reports are similar to the trends in expression and sequence variation of NCRs seen in our study.

**Polymorphism is observed in regulatory elements of NCRs**

Expression variation within a species may correlate with nucleotide heterozygosity (Ronald et al., 2005; Klibelstein et al., 2006); however, in addition to expression variation resulting from true ELPS (Doerge, 2002), sequence differences between the target cDNA from various accessions and the probe set on the array can result in a probe mismatch. In the latter case, array hybridization may not be an accurate report of expression differences among accessions. We defined a subset of differentially expressed NCRs that had no evidence of sequence polymorphism (no SFPs and SNPs) in their coding regions; these NCRs were likely to be true ELPS and were used to identify putative regulatory differences. Our report of the presence of SNPs in the cis-acting regulatory elements in the 1–kb upstream region of NCRs (Nallu et al., 2013) is in agreement with studies showing the importance of cis elements in the regulation of the expression of genes (Cowles et al., 2002; Rockman and Wray, 2002), although additional research is required to establish the effect of SNPs in these conserved regulatory elements on the expression patterns of NCRs.

We also observed substantial levels of nucleotide polymorphism in the 1 kb downstream region, which is a putative 3′–UTR of NCRs. The 3′–UTR contains binding sites for miRNAs and AU-rich elements, which affect the mRNA stability and thus influence gene expression (Wilkie et al., 2003). A previous report identified 100 novel miRNA in roots and nodules of *M. truncatula*, and two of the NCRs were found to be putative targets for a few miRNAs (Lelandais-Briere et al., 2009). We predict that in addition to the upstream regulatory elements found in NCRs there are additional binding sites found in the 3′–UTR regions of NCRs, which could not be predicted by the methods used in our earlier report on the regulation of NCRs (Nallu et al., 2013).

**NCRs differ from seed DEFLs in evolutionary rates**

DEFLs are closely related to the diverse defensin family of genes (Graham et al., 2004). The evolutionary patterns within the paralogous subgroups of β-defensins in mammals indicate patterns of purifying or diversifying selection (Maxwell et al., 2003; Semple et al., 2003, 2006). Another functionally related family, the NBS-LRRs, also show similar trends in selective pressures within the paralogous subgroups (Bauerman et al., 2003; Meyers et al., 2003) as well as orthologous groups in *A. thaliana* (Bakker et al., 2006). We observed that equal proportions of orthologous
groups of NCRs exhibit patterns of diversifying or purifying selection in their coding regions, which is in contrast to nearly 73% of seed DEFLs exhibiting purifying selection. Further analysis of the mature peptide regions of NCRs and seed DEFLs had similar results as observed with the entire coding region; however, the signal peptides of both the NCR and seed DEFLs were under purifying selection. Previously, it has been reported that paralogous clusters of DEFLs in A. thaliana (Silverstein et al., 2005) and M. truncatula (Alunni et al., 2007) show patterns of diversifying or purifying selection in the mature peptide region of the gene, whereas the signal peptide is predominantly under purifying selection. NCRs are targeted into the secretory pathway by their signal peptides (Mergaert et al., 2003), and were later found to be targeted specifically to the bacterial membrane and cytosol (Van de Velde et al., 2010). The absence of DNF1–1, a signal peptidase expressed in M. truncatula nodules (Wang et al., 2010), in the dnf1–1 mutant disrupted the localization of the NCRs, indicating the importance of signal peptide in proper post-translational processing and subcellular localization of the NCR peptides. The presence of purifying selection in signal peptides among orthologous NCRs and seed DEFLs indicates that the signal peptide is the most conserved part of DEFLs, irrespective of the species.

The differences in the selection pressures exhibited by the mature peptides of NCRs and seed DEFLs may reflect their different times of origin and diverse path of gene duplications. Seed DEFLs are seen in different clades of legumes and non-legumes (Graham et al., 2004; Silverstein et al., 2007), whereas NCRs are reported only in the IRLC of the legume family (Mergaert et al., 2003; Graham et al., 2004). Because the seed DEFLs have a more ancient lineage and more conserved sequences, their functions may have become fixed before the diversification of the legumes. NCRs have been reported only in the legumes with swollen bacteroids, a trait where the rhizobia are differentiated into swollen and branched nitrogen-fixing forms. They are absent in legumes with non-swollen bacteroids. It has been previously reported that the swollen bacteroid is a derived trait and non-swollen bacteroid is an ancestral trait (Oono et al., 2010). This suggests that NCRs, although extremely numerous in the M. truncatula genome, are a recently derived innovation in legumes. Our analysis of orthologous NCRs in 26 accessions of M. truncatula indicates that many loci are under purifying selection, whereas a large proportion of others exhibit patterns of diversifying selection, which may result in new functions by introducing non-synonymous substitutions into the coding region for the mature polypeptide. There are different ways a gene pair may evolve following duplication, such as through partitioning the original function between the two paralogs (subfunctionalization) or by acquisition of a new function by one copy while the other retains the original function (neofunctionalization). It has been reported that subfunctionalization and neofunctionalization can coexist, a condition referred to as subneofunctionalization (He and Zhang, 2005; Rastogi and Liberles, 2005). We predict from our observations of purifying and diversifying selection signatures in the mature peptide regions of orthologous NCRs that these genes may be a good example of subneofunctionalization. The set of orthologous NCRs under purifying selection probably share the same function (subfunctionalization), and the other set under diversifying selection probably acquired new functions (neofunctionalization), although additional functional studies on NCRs are necessary to establish this.

### EXPERIMENTAL PROCEDURES

#### Plant material and growth conditions

Seeds of M. truncatula accessions Jemalong A17 (hereafter called A17), DZA315–16, F83005–5 and R108–C3 were sterilized and germinated as described previously for A17 (Lohar et al., 2006). Then 1-day-old germinated seedlings were grown on sterile plates (245-mm × 245-mm dishes; Corning, http://www.corning.com) containing buffered nodulation medium (BNM; Ehhardt et al., 1992), at pH 6.5, and 1.2% agar (plant tissue culture grade; Sigma-Aldrich, http://www.sigmaaldrich.com). The radicles were placed on moist, sterile germination paper on the surface of the medium and covered with a sterile black cotton cloth (Cotton Club Black, http://www.sigmaaldrich.com). Plates were maintained vertically in a growth chamber under the following conditions: 16-h day/8-h night (200–300 µmol m⁻² s⁻¹); 25–21°C; and 50% relative humidity. At 5 days after planting, plants were inoculated with 100 µL/root of a washed suspension of S. meliloti Sm1021 (Meade et al., 1982) or S. medicae A321 (Rome et al., 1996) in sterile water (OD₆₀₀ = 0.05). Approximately 5-cm-long nodule-bearing root segments were harvested from inoculated plants at 14 days post-inoculation (dpi) by carefully measuring the root segments among all accessions. Any excess of root was trimmed in order to maintain equal portions of root tissue. The root segments from all treatments were collected from tissue biological replicates, with a pool of multiple plants in each replicate, were frozen immediately in liquid nitrogen and stored at −80°C for subsequent RNA extraction. Total RNA extraction, probe preparation and array hybridization was performed as described in Tesfaye et al. (2013).

#### Microarray data analysis

Data were normalized across the different nodule treatments using SBO normalization (Sato et al., 2007; Tesfaye et al., 2013). For all further analysis, only probes with at least one present call among the treatments were considered as NCRs. The accessions, accession-strain interaction and strain effects on NCR expression were calculated using a two-way ANOVA (P ≤ 0.01) using GeneSpring GX11 (Agilent, http://www.agilent.com).

Genes that were differentially expressed among the four M. truncatula accessions, and amongst or within an accession between inoculation treatments, were identified using the Empirical Bayes for variance shrinkage (Cui et al., 2005a,b) within the limma package distributed with R/Bioconductor (Gentleman et al., 2004). Both a twofold change cut-off and Storey’s false discovery rate (FDR) correction (P ≤ 0.01) were applied.
All microarray data in this study have been deposited in the Gene Expression Omnibus under accession number GSE34804 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE34804).

**Single-feature polymorphism detection**

The probe affinity shape power (PASP) method (Xu et al., 2009) was used to detect SFPs from the probes of NCRs. Briefly, the normalized signal intensities were used to calculate the perfect match (PM) affinities for each probe and the expression index of the probe set was based on the affinity model (Li and Wong, 2001; Hubbell et al., 2002; Irizarry et al., 2003; Cui et al., 2005a,b). SFP weight scores were calculated as described in Xu et al. (2009). A more stringent cut-off was applied by filtering off all the SFPs with weight scores lower than 2.5, and finally only the common SFP calls from the two strains between the pairs of accessions was reported.

**Single-nucleotide polymorphism detection**

The coding sequences of all DEFLs represented on the MtDEFL array were mapped onto Medicago genome assembly Mt3.5 using BLAT (Kent, 2002) with a 90% identity threshold. Using the mapped coordinates, SNPs were extracted in the coding region, 1-kb upstream region, 1-kb downstream region and introns of NCR genes for accessions DZA315 using BLAT (Kent, 2002) with a 90% identity threshold. Using the mapped coordinates, SNPs were extracted in the coding region, different regions in NCRs were mapped coordinates, SNPs were extracted in the coding region, and to map the SNPs to cis-regulatory elements in the differentially expressed NCRs.

**Estimation of NCR nucleotide diversity**

All of the mapped DEFLs (as explained in SNP detection) were categorized as NCRs and seed DEFLs if they had a present call in at least one treatment across the different nodule treatments (Nallu et al., 2013) and seed treatments (Tesfaye et al., 2013). The complete coding sequences of the NCRs and seed DEFLs from 26 accessions were recovered from the A17 reference sequence (Mt3.5) using SNP and indel data (http://www.medicagohapmap.org), using Perl scripts. Perl scripts were also used for validation of a few random SFPs against the reference genome A17 Jemalong (http://www.medicagohapmap.org). The sequences were aligned using Perl scripts. A total of 263 NCRs and 43 seed DEFLs with at least 75% sequence information were used to estimate the total (θ) and average (π) nucleotide diversity. The coding sequences of NCRs and seed DEFLs were divided into signal peptide and mature peptide regions by SIGNALP 3.0 (Bendtsen et al., 2004). The θ and π values were estimated using the Nei–Gojobori method (Nei and Gojobori, 1986) from DsUPF 5.10 (Librado and Rozas, 2009) and π codes. To test the statistical significance of the differences in θ values among different regions in NCRs and seed DEFLs, all the genes with θsyn/θsyn > 1 were assigned as 1, θnuc = θsyn were assigned as 0 and θnuc/θsyn < 1 were assigned as −1, and examined by Mann–Whitney U-test. All the steps were repeated for testing the statistical significance of the π values.

**ACKNOWLEDGEMENTS**

We would like to thank Fumikazu Katagiri for help with the statistical analysis of expression data, Wayne Wenzhong Xu for the assistance with SFP analysis, and Peter Tiffin and Timothy Paape for helpful suggestions. Part of this work was carried out using the Minnesota Supercomputing Institute for computational infrastructure and systems support, and funded by NSF grant awards OB-0516811 and IOS-1237993.

**SUPPORTING INFORMATION**

Additional Supporting Information may be found in the online version of this article.

**Table S1.** The mean expression values of NCRs in different accessions when inoculated with the two rhizobial strains.

**Table S2.** Comparison of randomly chosen SFPs with genome sequence.

**Table S3.** SNPs in upstream cis motifs of NCRs.

**Table S4.** Twenty-six accessions used for nucleotide diversity estimates.

**REFERENCES**

Alunni, B., Kevei, Z., Redondo-Nieto, M., Kondorosi, A., Margaert, P. and Kondorosi, E. (2007) Genomic organization and evolutionary insights on GRP and NCR genes, two large nodule-specific gene families in medicago truncatula. Mol. Plant Microbe Interact. 20(9), 1138–1148.

Bakker, E.G., Stahl, E.A., Toomajian, C., Nordborg, M., Kreitman, M. and Bergelson, J. (2006) Distribution of genetic variation within and among local populations of arabidopsis thaliana over its species range. Mol. Ecol. 15(6), 1405–1418.

Baumgarten, A., Cannon, S., Spangler, R. and May, G. (2003) Genome-level evolution of resistance genes in arabidopsis thaliana. Genetics, 165(1), 309–319.

Bendtsen, J.D., Nielsen, H., von Heijne, G. and Brunak, S. (2004) Improved prediction of signal peptides: SignalP 3.0. J. Mol. Biol. 340(4), 783–795.

Bendito, V.A., Torres-Jerez, I., Murray, J.D. et al. (2008) A gene expression atlas of the model legume medicago truncatula. Plant J. 55(3), 504–513.

Borevitz, J.O., Hazen, S.P., Michael, T.P. et al. (2007) Genome-wide patterns of single-feature polymorphism in arabidopsis thaliana. Proc. Natl Acad. Sci. USA 104(29), 12057–12062.

Branca, A., Pape, T.D., Zhou, P. et al. (2011) Whole-genome nucleotide diversity, recombination, and linkage disequilibrium in the model legume Medicago truncatula. Proc. Natl Acad. Sci. USA 108(42), E884–E870.

Caicedo, A.L., Stinchcombe, J.R., Olsen, K.M., Schmitt, J. and Purugganan, M.D. (2004) Epistatic interaction between arabidopsis FRI and FLC flowering time genes generates a latitudinal cline in a life history trait. Proc. Natl Acad. Sci. USA 101(14), 4767–4771.

Cannon, S.B., May, G.D. and Jackson, S.A. (2009) Three sequenced legume genomes and many crop species: Rich opportunities for translational genomics. Plant Physiol. 151(3), 970–977.

Cook, D.R. (1999) Medicago truncatula—a model in the making. Curr. Opin. Plant Biol. 2(4), 301–304.

Coulten, C.R., Hirschhorn, J.N., Altschuler, D. and Lander, E.S. (2002) Detection of regulatory variation in mouse genes. Nat. Genet. 32(3), 432–437.

Cui, X., Hwang, J.T., Giu, J., Blades, N.J. and Churchill, G.A. (2005a) Improved statistical tests for differential gene expression by shrinking variance components estimates. Biostatistics, 6(1), 59–75.

Cui, X., Xu, J., Ashgar, R., Condamine, P., Svensson, J.T., Wanamaker, S., Stein, N., Roose, M. and Close, T.J. (2005b) Detecting single-feature polymorphisms using oligonucleotide arrays and robustified projection pursuit. Bioinformatics, 21(20), 3852–3858.

Doerge, R.W. (2002) Mapping and analysis of quantitative trait loci in experimental populations. Nat. Rev. Genet. 3(1), 43–52.

Ehrhardt, D.W., Atkinson, E.M. and Long, S.R. (1992) Depolarization of alfalfa root hair membrane potential by rhizobium melliloti nod factors. Science, 256(5059), 998–1000.

Fedorova, M., van de Mortel, J., Matsumoto, P.A., Cho, J., Town, C.D., Van-denBosch, K.A., Gantt, J.S. and Vance, C.P. (2002) Genome-wide identification of nodule-specific transcripts in the model legume medicago truncatula. Plant Physiol. 130(2), 519–537.

Galibert, F., Finan, T.M., Long, S.R. et al. (2001) The composite genome of the legume symbiont sinorhizobium melliloti. Science, 293(5530), 668–672.

Gentleman, R.C., Carey, V.J., Bates, D.M. et al. (2004) Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 5(10), R80.
Graham, M.A., Silverstein, K.A., Cannon, S.B. and Vandenbergoch, K.A. (2004) Computational identification and characterization of novel genes from legumes. *Plant Physiol.* 135(3), 1179–1197.

Hazen, S.P. and Kay, S.A. (2003) Gene arrays are not just for measuring gene expression. *Trends Plant Sci.* 8(9), 413–416.

He, X. and Zhang, J. (2005) Rapid subfunctionalization accompanied by prolonged and substantial neo-functionalization in duplicate gene evolution. *Genetics*, 169(2), 1157–1164.

Hirsch, A.M., Bang, M. and Ausubel, F.M. (1983) Ultrastructural analysis of ineffective alalfa nodules formed by nifTv mutants of rhizobium meliloti. *J. Bacteriol.* 155(1), 367–380.

Hubbell, E., Liu, W.M. and Mei, R. (2002) Robust estimators for expression analysis. *Bioinformatics*, 18(12), 1585–1592.

Irizarry, R.A., Bolstad, B.M., Collin, F., Cope, L.M., Hobbs, B. and Speed, T.P. (2003) Summaries of affymetrix GeneChip probe level data. *Nucleic Acids Res.* 31(4), e15.

Kent, W.J. (2002) BLAT-the BLAST-like alignment tool. *Genome Res.* 12(4), 656–664.

Kliebenstein, D.J., West, M.A., van Leeuwen, H., Loudet, O., Hobbs, B. and Speed, T.P. (2003) Gene arrays are not just for measuring gene expression. *Plant Physiol.* 37, 1025–1035.

Mauricio, R., Stahl, E.A., Korves, T., Tian, D., Kreitman, M. and Bergelson, J. (2003) Comparative genomewide analysis of NBS-LRR-encoding genes in Arabidopsis. *Genome Biol.* 4, REVIEWS0004.

Nallu, S., Silverstein, K.A., Samac, D.A., Bucciarelli, B., Vance, C.P. and Van Belik, D. (2000) Rapid subfunctionalization by 0- and 3-UTR-binding factors. *Trends Biochem. Sci.* 28(4), 182–188.

Oono, R., Schmitt, I., Sprent, J.I. and Denison, R.F. (2010) Multiple evolutionary origins of legume traits leading to extreme rhizobial differentiation. *New Phytol.* 187(2), 508–520.

Rockman, M.V. and Wiray, G.A. (2002) Abundant raw material for cis-regulatory evolution in humans. *Mol. Biol. Evol.* 19(11), 1991–2004.