Comparative Genomics Suggests That an Ancestral Polyploidy Event Leads to Enhanced Root Nodule Symbiosis in the Papilionoideae

Qi-Gang Li,1,2 Li Zhang,1 Chun Li,1,3 Jim M. Dunwell,4 and Yuan-Ming Zhang*,1
1State Key Laboratory of Crop Genetics and Germplasm Enhancement, Department of Crop Genetics and Breeding, College of Agriculture, Nanjing Agricultural University, Nanjing, People’s Republic of China
2State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, People’s Republic of China
3Henan Sesame Research Center, Henan Academy of Agricultural Sciences, Zhengzhou, People’s Republic of China
4School of Agriculture, Policy and Development, University of Reading, Reading, United Kingdom
*Corresponding author: E-mail: soyzhang@njau.edu.cn.
Associate editor: Shizhong Xu

Abstract

Root nodule symbiosis (RNS) is one of the most efficient biological systems for nitrogen fixation and it occurs in 90% of genera in the Papilionoideae, the largest subfamily of legumes. Most papilionoid species show evidence of a polyploidy event that occurred approximately 58 Ma. Although polyploidy is considered to be an important evolutionary force in plants, the role of this papilionoid polyploidy event, especially its association with RNS, is not understood. In this study, we explored this role using an integrated comparative genomic approach and conducted gene expression comparisons and gene ontology enrichment analyses. The results show the following: 1) Approximately a quarter of the papilionoid-polyploidy–derived duplicate genes are retained; 2) there is a striking divergence in the level of expression of gene duplicate pairs derived from the polyploidy event; and 3) the retained duplicates are frequently involved in the processes crucial for RNS establishment, such as symbiotic signaling, nodule organogenesis, rhizobial infection, and nutrient exchange and transport. Thus, we conclude that the papilionoid polyploidy event might have further refined RNS and induced a more robust and enhanced symbiotic system. This conclusion partly explains the widespread occurrence of the Papilionoideae.

Key words: legume, nitrogen fixation, papilionoid, polyploidy, root nodule symbiosis.

Introduction

Root nodule symbiosis (RNS) or nodulation is one of the most productive systems for biological nitrogen fixation. Such RNS is confined to a single large clade termed “the N2-fixing clade” (NFC) (Soltis et al. 1995). Although the legume family, the third largest family of flowering plants, is dominated by nodulators, nonlegume families in which nodulation is universal or widespread are mostly small (Swensen and Benson 2008). Within the legume family, the Papilionoideae is the largest and most widely distributed subfamily. It includes most of the cultivated plants and model legume species (Gepts et al. 2005; Pawlowski and Sprent 2008), and 90% of genera belonging to this subfamily exhibit nodulation, whereas only approximately 5% of Caesalpinioideae genera show nodulation. In contrast, nodulation is nearly ubiquitous in the relatively small Mimosoideae subfamily (Sprent 2001). These different distribution patterns suggest that RNS established in the Papilionoideae may be more stable than those in other lineages in NFC.

A whole-genome duplication (WGD) or polyploidy event is shared by most of the papilionoid lineages, except some early splitting papilionoid lineages, and occurred approximately 58 Ma, shortly after the origin of legumes at approximately 60 Ma. This WGD event was identified and confirmed by genomic analyses of four legume species: Glycine max (Schmutz et al. 2010), Medicago truncatula (Young et al. 2011), Lotus japonicus (Sato et al. 2008), and Cajanus cajan (Varshney et al. 2011). Although polyploidy has long been recognized as an important evolutionary force, for example, in species radiation, organ innovation, and complex innovations in cellular networks (Ohno 1970; Lynch 2007; Edger and Pires 2009; Huminiecki and Conant 2012; Li et al. 2012), the roles of the papilionoid polyploidy event have not been extensively studied, especially in the context of the ubiquitous distribution of RNS in the Papilionoideae.

In an investigation of the papilionoid polyploidy event, Singer et al. (2009) showed that it could be associated with major species radiations of legumes (particularly the papilionoid subfamily), although this study suggested a series of alternative hypotheses. Op den Camp et al. (2011) applied a phylogenetic strategy to scan the genes in the cytokinin phosphorelay pathway and found that two papilionoid-WGD–derived type A cytokinin response regulators, MtRR9
and MtRR11, in *M. truncatula* are recruited during nodulation. In addition, Young et al. (2011) indicated that the papilionoid polyploidy event might have facilitated the emergence of critical components of nod factor (NF) signaling and contributed to the complexity of rhizobial nodulation found in the Papilionoideae. Although these previous studies have suggested the important roles of this polyploidy event, most of them are based on studies of single genes, pathways, or genomes, and very little insight has been gained from genome-wide and cross-species comparative studies. Recently, the increasing number of completely sequenced legume and nonlegume genomes has provided a rich opportunity for a comparative genomics study.

The first step for a comparative genomics analysis of the papilionoid WGD event is to identify gene duplicates derived from this event. Currently, there are two widely used approaches for comparative genomics studies: a synteny-based method and a phylogenetic approach (Deshal and Boore 2005; Thomas et al. 2006; Vilella et al. 2009; Young et al. 2011; Wang et al. 2012; Li and Zhang 2013). Of these two methods, the synteny method can give a view of genome structure variation, and the phylogenetic approach provides not only a group of homologies but also their phylogenetic relationships. However, the identification of interesting lineages in the phylogenetic approach can be affected by incorrect topologies in the structure of gene trees. Therefore, it would be beneficial to integrate the two aforementioned approaches.

In this study, we utilized an integrated comparative genomics approach using the completely sequenced genomes of four papilionoid species and two nonlegume species to identify the papilionoid WGD-derived gene lineages. We also investigated the retention patterns of these lineages. Gene expression comparison and GO enrichment analysis were conducted to identify the papilionoid WGD-enhanced biological processes and/or pathways likely associated with the evolution of RNS in Papilionoideae. On the basis of these results, we were able to infer the potential roles of the papilionoid polyploidy event in the evolution of RNS.

**Results and Discussion**

**Differential Gene Loss and Retention**

After clustering, multiple alignment, and phylogenetic reconstruction, 11,485 phylogenetic trees were constructed. An analysis of these trees predicted that 16,114 interior nodes are ancestral to papilionoid genes (fig. 1). In two branches or lineages doubled by a papilionoid WGD node, one of duplicate branches is said to be lost if all genes in that lineage are absent; otherwise, both duplicate branches are said to be retained. As a result, 25.5% of papilionoid WGD nodes retained both duplicate branches (fig. 1 and supplementary data set S1, Supplementary Material online), and 74.5% of ancestral nodes lost one of duplicate branches (fig. 1). Similarly, in a particular papilionoid species, *G. max*, for example, if both duplicate branches with genes in *G. max* are observed, we concluded that this species retained both gene duplicates derived from the papilionoid WGD. The retention percentages calculated for *G. max*, *C. cajan*, *M. truncatula*, and *L. japonicus* are 21%, 19.4%, 10.3%, and 8.1%, respectively (fig. 1 and supplementary data set S1, Supplementary Material online), which suggests that the rates of gene retention differ across legume lineages. This is borne out by the similarities between the closely related *G. max* and *C. cajan*, where numbers are different from those of *L. japonicus* and *M. truncatula*. Another explanation is that the varying conditions of the genome assemblies may affect conclusions about gene retention in this study.

To utilize the gene synteny results efficiently, we used the findings as evidence for the predicted papilionoid WGD nodes. If a pair of duplicate genes predicted in this study locates in a syntenic block, we say that the predicted papilionoid WGD node is supported by gene synteny. As a result, 66.2% of the aforementioned 4,113 predicted papilionoid WGD nodes are supported by gene collinearity; in individual species, 78.5% of the predicted papilionoid WGD nodes shared in *G. max* are supported, and the corresponding fraction is 68.7% in *M. truncatula*, 76% in *C. cajan*, and 67.7% in *L. japonicus* (fig. 1 and supplementary data set S1, Supplementary Material online). Because this polyploidy event occurred approximately 58 Ma and the gene order would have been destroyed over time, some of the true papilionoid WGD nodes are not supported by gene synteny. Thus, we manually checked the nonsupported papilionoid WGD nodes and found that many of them have reasonable topological structures compared with the species tree. As a result, the noise caused by incorrect gene topology should be minor.

In this study, we mainly investigated papilionoid-specific and -shared characteristics. Thus, we focused on retained

![Cladogram of species tree and numbers of gene retentions of papilionoid WGD-derived duplicates. The topology of the species tree is from APG III (Bremer et al. 2009) and branch lengths of this tree show no means. The bold branch indicates the papilionoid polyploidy event. The number 16,114 indicates the total number of predicted ancestral papilionoid nodes. In two branches or lineages doubled by a papilionoid WGD node, both duplicate branches are considered to be retained if both branches are observed with papilionoid genes. Thus, the three numbers on the bold branch indicate that 4,113 papilionoid WGD nodes retained both duplicate branches, which represent 25.5% of 16,114 papilionoid nodes and 66.2% of these are supported by gene collinearity. Other sets of three successive numbers represent similar meanings in individual legume species. *Glycine max*, for example, 3,379 duplicate branches are retained, which represent 21% of the total 16,114 papilionoid nodes, and 78.5% of the 3,379 nodes are supported by gene collinearity.](https://academic.oup.com/mbe/article-abstract/30/12/2602/1012894/doi)
duplicates that co-exist throughout the entire papilionoid subfamily. Note that 1,160 papilionoid WGD nodes were shared by three species (G. max, M. truncatula, and C. cajan) and that 395 papilionoid WGD nodes were shared by the four studied species. Owing to the fact that the L. japonicus genome contains a large number of gaps and that its gene annotation is incomplete, the 1,160 nodes were used as the representative papilionoid WGD nodes.

Expression Divergence between Duplicate Genes
In each of the two lineages derived from the papilionoid polyploidy event, the gene with the highest expression value was selected as the representative duplicate gene. Two genes, one from each duplicate lineage, form a pair of duplicate genes. As a result, 1,160 gene pairs were individually obtained from G. max and M. truncatula. If one gene of a pair is expressed in a particular tissue and another is not, the gene pair is said to have diverged expression in that tissue (Makova and Li 2003). As a result, 27.0% and 32.2% gene pairs in G. max and M. truncatula, respectively, have diverged in expression in at least two tissues studied, and the two percentages increase to 45.0% and 44.0%, respectively, in at least one tissue. The expression divergence of gene duplicates implies an increase in regulatory gene complexity and robustness fuelled by the expression divergence of gene duplicates implies an increase.

Functionally Preferential Retention of Duplicates
The genes of interest were obtained from the lineages after the 1,160 papilionoid WGD nodes were identified, which resulted in the identification of 5,239 and 3,722 genes in G. max and M. truncatula, respectively. Through a GO enrichment analysis of the G. max genome, we identified 726 GO terms for biological processes, molecular functions, and cellular components significantly overrepresented in the genes of interest compared with all the annotated genes of genome. Similarly, 694 GO terms were identified in M. truncatula. 440 of these GO terms were found in both G. max and M. truncatula (supplementary data set S2, Supplementary Material online), and these shared GO terms included 17 GO slim terms (table 1). Like observations made in bacteria (Kondrashov et al. 2002), teleost fishes (Brunet et al. 2006), and Arabidopsis (Blanc and Wolfe 2004; Thomas et al. 2006), transcription factors, signal proteins, and membrane proteins are preferentially retained after duplication (table 1). Because genes associated with transcriptional regulation and signal transduction are frequently dosage sensitive, the differential retention of duplicates after the papilionoid polyploidy event may also follow the gene-dosage balance hypothesis (Birchler and Veitia 2007, 2012).

We further used the gene expression data to investigate whether a significantly higher number of duplicates with a particular function are recruited in nodules. Among the aforementioned 440 shared GO terms, 362 and 372 GO terms show nodule significance (P < 0.01) in M. truncatula and G. max, respectively. Of these GO terms, 334 are shared between the two species (supplementary data set S2, Supplementary Material online). These findings suggest that many nodulation-related functions were enhanced by the polyploidy event. More importantly, we investigated the possible roles of this polyploidy event for RNS establishment.

Table 1. Overrepresented GO Slim Terms after the Papilionoid Polyploidy Event Compared with the Individual Genome Background.

<table>
<thead>
<tr>
<th>GO ID</th>
<th>Term</th>
<th>P Value</th>
<th>Glycine max</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Biological process</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0008219</td>
<td>Cell death</td>
<td>1.49E-24</td>
<td>4.67E-21</td>
</tr>
<tr>
<td>GO:0016265</td>
<td>Death</td>
<td>1.49E-24</td>
<td>4.67E-21</td>
</tr>
<tr>
<td>GO:0019748</td>
<td>Secondary metabolic process</td>
<td>1.36E-24</td>
<td>1.86E-19</td>
</tr>
<tr>
<td>GO:0007165</td>
<td>Signal transduction</td>
<td>2.48E-17</td>
<td>9.45E-17</td>
</tr>
<tr>
<td>GO:0009856</td>
<td>Pollination</td>
<td>2.66E-06</td>
<td>3.18E-13</td>
</tr>
<tr>
<td>GO:0007154</td>
<td>Cell communication</td>
<td>7.91E-11</td>
<td>4.39E-12</td>
</tr>
<tr>
<td>GO:0009719</td>
<td>Response to endogenous stimulus</td>
<td>6.91E-16</td>
<td>2.11E-10</td>
</tr>
<tr>
<td>GO:0040007</td>
<td>Growth</td>
<td>6.49E-09</td>
<td>2.66E-09</td>
</tr>
<tr>
<td>GO:0016049</td>
<td>Cell growth</td>
<td>6.70E-05</td>
<td>1.04E-06</td>
</tr>
<tr>
<td>GO:0007610</td>
<td>Behavior</td>
<td>1.62E-03</td>
<td>7.96E-06</td>
</tr>
<tr>
<td>GO:0009607</td>
<td>Response to biotic stimulus</td>
<td>3.27E-03</td>
<td>3.75E-04</td>
</tr>
<tr>
<td>GO:0006810</td>
<td>Transport</td>
<td>2.31E-12</td>
<td>5.84E-04</td>
</tr>
<tr>
<td></td>
<td>Molecular function</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0016740</td>
<td>Transferase activity</td>
<td>1.57E-03</td>
<td>2.87E-04</td>
</tr>
<tr>
<td>GO:0004871</td>
<td>Signal transducer activity</td>
<td>2.27E-07</td>
<td>5.04E-04</td>
</tr>
<tr>
<td>GO:0003700</td>
<td>Transcription factor activity</td>
<td>8.65E-11</td>
<td>3.08E-03</td>
</tr>
<tr>
<td></td>
<td>Cellular component</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0005886</td>
<td>Plasma membrane</td>
<td>8.80E-37</td>
<td>3.37E-21</td>
</tr>
<tr>
<td>GO:0016020</td>
<td>Membrane</td>
<td>2.22E-15</td>
<td>5.60E-04</td>
</tr>
</tbody>
</table>

Note.—P value indicates the significance of overrepresented GO terms conducted by GOstat software.
according to the molecular mechanisms of RNS described by Oldroyd et al. (2011) and Bapaume and Reinhardt (2012). Figure 2 shows selected GO terms likely associated with RNS and gene expression in nodules.

**Symbiotic Signaling**

Flavonoids are essential signal molecules that play multiple roles in legume–*Rhizobium* symbiosis, for example, the induction of the biosynthesis of the NF and the regulation of auxin transport (Subramanian et al. 2007). Enriched terms, such as GO:0019748 (secondary metabolic process) and GO:0009812 (flavonoid metabolic process), have a significantly higher number of gene duplicates expressed in nodules (fig. 2 and supplementary data set S2, Supplementary Material online). Through an analysis of the enzymes involved in the flavonoid biosynthesis, we found that at least eight types of putative enzymes (indicated by bold arrows in fig. 3) were duplicated and retained after the polyploidy event; these enzymes include CHS (chalcone synthase; supplementary fig. S1, Supplementary Material online), CHR (chalcone reductase; supplementary fig. S2, Supplementary Material online), F3H (flavanone 3-hydroxylase; supplementary fig. S3, Supplementary Material online), FLS (flavonol synthase; supplementary fig. S4, Supplementary Material online), F3'H (flavonoid 3'-hydroxylase; supplementary fig. S5, Supplementary Material online), HIDH (2-hydroxyisoflavone dehydratase; supplementary fig. S6, Supplementary Material online), I2'H (isoflavone-2'-hydroxylase; supplementary fig. S7, Supplementary Material online), and IOMT (isoflavone 7-O-methyltransferase or isoflavone 4'-O-methyltransferase; supplementary fig. S8, Supplementary Material online). These results suggest that more abundant and diverse flavonoids would be synthesized as a result of the polyploidy event and that the enrichment of flavonoids might be adaptive for the complex signaling required for legume–*Rhizobium* symbiosis.

NF receptors (NFRs) have an extracellular domain, which contains two to three lysine motif (LysM) repeats and an intracellular kinase domain. As shown in figure 4, the papilionoid LysM receptors closely homologous to AT3G21630 (chitin elicitor receptor kinase 1, AtCERK1), which recognizes chitin elicitor and activates immune responses (Miya et al. 2007; Wan et al. 2008; Liu et al. 2012), were duplicated as a result of the papilionoid polyploidy event. In addition, one of the WGD duplicates induces a tandem duplication shared by the four papilionoid species studied. These two gene duplications significantly amplify the gene family of LysM receptors in the Papilionoideae. NFR1 (Radutoiu et al. 2003) and LYK3 (Limpens et al. 2003) have been shown to be NFRs. The phylogenetic tree in figure 4 also indicates that these NFRs might have been evolved from an original chitin elicitor receptor kinase involved in plant defense, in a similar way to AtCERK1. Therefore, the evolution of RNS is an interesting example of the transition from resistance to co-operation observed in plants. Furthermore, the enriched GO terms associated with plant immunity, such as GO:0006955 (immune response) and GO:0010200 (response to chitin), have a significant excess of gene duplicates expressed in nodules (fig. 2 and supplementary data set S2, Supplementary Material online). These findings suggest that RNS is highly associated with plant immunity and that an increased number of immune gene duplicates were recruited for nodulation. Other genes involved in symbiotic signaling, such as NFR5, SINA4, ERN2, NSP2, and MtHMGR1, also have papilionoid polyploidy.
paralogs (table 2). These increased symbiotic signaling genes suggest that the papilionoid polyploidy event might have induced the emergence of critical symbiotic genes and increased the complexity of the symbiotic signaling pathway.

**Nodule Organogenesis**

NF recognition at the root surface activates nodule organogenesis, which requires the regulation of plant hormones, particularly cytokinin and auxin (Crespi and Frugier 2008). Nodule development is regulated by cytokinin signaling and polar auxin transport (Oldroyd et al. 2011). Enriched GO terms, such as GO:0009755 (hormone-mediated signaling pathway) and GO:0009736 (cytokinin-mediated signaling pathway), were consistently identified as significant (fig. 2 and supplementary data set S2, Supplementary Material online). The function gain or loss of the cytokinin receptor LHK1 would trigger spontaneous nodule organogenesis (Tirichine et al. 2007) or block its formation (Murray et al. 2007). With the exception of *M. truncatula*, the other three legume species under study retained the papilionoid-WGD–derived duplicates paralogous to LHK1 (table 2). Another example in the cytokinin signaling pathway is a pair of type A cytokinin response regulators (MtRR9 and MtRR11) in *M. truncatula* (table 2); Op den Camp et al. (2011) suggested that MtRR9 is involved in nodulation. In addition, GO:0034050 (host-programmed cell death induced by the symbiont) has a significant excess of gene duplicates expressed in nodules (fig. 2 and supplementary data set S2, Supplementary Material online), which suggests that an increased number of duplicates in the control of cell death were recruited for nodulation.

**Rhizobial Infection**

The perception of NF by root hairs activates the formation of infection threads (ITs), which induce the movement of rhizobial bacteria into the nodule cell for N2 fixation. The initiation of ITs requires the loosening and reconfiguration of the localized root hair cell wall, and the formation of ITs is a type of polar growth. Enriched GO terms, such as GO:0009827 (plant-type cell wall modification) and GO:0033037 (polysaccharide localization), were identified (fig. 2 and supplementary data set S2, Supplementary Material online), which suggests that an increased number of duplicates were recruited for the formation of ITs. For example, polygalacturonase (PG) is an enzyme associated with plant cell wall degradation. Duplicates closely homologous to MsSPG3 (Muñoz et al. 1998) of *M. sativa* were retained (table 2). ROP6, which is a Rho-like small
GTPase from *L. japonicus*, controls the growth of ITs (Ke et al. 2012), and has WGD paralogs (table 2). Additional genes involved in the formation of ITs, such as nsRING, PUB1, sickle, and NIP, also have papilionoid-WGD–derived paralogs (table 2).

### Nutrient Exchange and Transport

Inside legume nodules, rhizobial bacteroids reduce N\textsubscript{2} to ammonium, which is then secreted to the host in exchange for carbon and energy sources and then exchanged for the transport of needed active materials. Consistently, enriched terms, such as GO:0006576 (cellular biogenic amine metabolic process), GO:0043090 (amino acid import), and GO:0006865 (amino acid transport), were identified (fig. 2 and supplementary data set S2, Supplementary Material online). Plant glutamine synthetase (GS) is a key enzyme that assimilates the ammonium produced by bacteroids, and MtGS1b has WGD-derived paralogs (table 2; Carvalho et al. 1997). Other protein families associated with nutrient transport, such as the putative glutamine dumper family (supplementary fig. S22, Supplementary Material online), the putative oligopeptide transporter family (supplementary fig. S23, Supplementary Material online), and the putative amino acid permease family (supplementary fig. S24, Supplementary Material online), were also amplified by this polyploidy event.

In addition, 84% of these 1,160 nodes are supported by gene synteny (supplementary data set S1, Supplementary Material online). Plant glutamine synthetase (GS) is a key enzyme that assimilates the ammonium produced by bacteroids, and MtGS1b has WGD-derived paralogs (table 2; Carvalho et al. 1997). Other protein families associated with nutrient transport, such as the putative glutamine dumper family (supplementary fig. S22, Supplementary Material online), the putative oligopeptide transporter family (supplementary fig. S23, Supplementary Material online), and the putative amino acid permease family (supplementary fig. S24, Supplementary Material online), were also amplified by this polyploidy event.

In addition, 84% of these 1,160 nodes are supported by gene synteny (supplementary data set S1, Supplementary Material online).
As a supplement and comparison, we also conducted GO enrichment analysis and nodule recruitment expression analysis on this set of data in the same manner and obtained 463 GO terms (supplementary data set S2, Supplementary Material online). 209 of these terms coincided with the 440 GO terms mentioned earlier (supplementary data set S2, Supplementary Material online). With the exception of GO:0019748 (secondary metabolic process) and GO:0009812 (flavonoid metabolic process), the other selected GO terms (fig. 2) are included in both results. Thus, the two results are consistent. As a high proportion (84%) of these 1,160 nodes are supported by gene synteny, and some of the true WGD nodes are also not supported as their gene syntenic blocks were destroyed over time, we consider that the results of these 1,160 papilionoid WGD nodes provide a more comprehensive assessment of the role of the papilionoid polyploidy event.

### Roles of Papilionoid Polyploidy Event in the Evolution of Nodulation

Cannon et al. (2010) suggested that the polyploidy event did not predate the evolution of nodulation in all legumes, and thus a logical consequence, as suggested by Doyle (2011), is that the papilionoid WGD could have provided genes for modifying and refining the symbiosis, if legume nodulation is homologous. Consistently, we observed a large number of retained duplicates, and gene expression divergence, suggesting that the polyploidy event is able to facilitate the formation of a more complex and diversified papilionoid RNS. More importantly, we found that many gene duplicates or functions crucial for RNS were preferentially coretained in the three papilionoid species; this suggests that this polyploidy event could have provided genes for further enhancing the symbiosis. Whether caused by multiple origins or independent loss of the RNS in the NFC (Doyle 2011), the distribution pattern, and the fact that RNS is ubiquitous and numerous in Papilionoideae, whilst being mostly rare elsewhere in NFC, could suggest that the initial or ancestral RNS may be unstable; if this were not so, then RNS should be widespread in the entire NFC. Note that this polyploidy event increased genetic complexity and robustness (Gu et al. 2003; He and Zhang 2005; Wagner 2008; Van de Peer et al. 2009) and enhanced the papilionoid RNS, so that most of the papilionoid species are able to establish a more stable symbiotic relationship with rhizobial bacteria, and that the benefits from the symbiosis further facilitate the papilionoid species to adapt to various...
environments (Zahran 1999; Santi et al. 2013), eventually leading to the wide distribution of the Papilionoideae.

In addition, it is interesting to note that RNS is also ubiquitous in the Mimosoideae, suggesting that there has been a different way to generate an effective RNS, without the benefit of the WGD. However, we also need to note that the Mimosoideae is a relatively small subfamily with 80 genera and 3,200 species compared with the Papilionoideae with 470 genera and 14,000 species, and that the Papilionoideae is more widely distributed. WGD, as an important evolutionary force, can quickly provide abundant raw materials for organ innovation or adaption to diverse environments; however, the corresponding evolution events driven by single gene mutation or duplication could cost more time and are even less likely. Thus, granted that the Mimosoideae could have an effective RNS, the papilionoid RNS is widely distributed and dominant in nodulating species within the NFC.

In this study, we identified the potential functional groups of the papilionoid WGD duplicates important for the establishment of the papilionoid characteristics, such as RNS. Experimentally, some of the genes have been proven to be involved in nodulation. Therefore, our study provides rich and systemic clues to the unraveling of the shared or unique involved in nodulation. Therefore, our study provides rich and systemic mechanisms of the papilionoid WGD duplicates important for the establishment of the papilionoid characteristics, such as RNS.

Materials and Methods

Genomic Data and Annotation

The genomic data of four papilionoid species (G. max, M. truncatula, L. japonicus and C. cajan) and two nonlegumes (Arabidopsis thaliana and Prunus persica) (Verde et al. 2013) were downloaded. The detailed gene annotations for the six species are shown in supplementary table S1, Supplementary Material online. The longest complete coding sequence (CDS) of each gene was chosen, and repeat sequences of all of the selected CDSs were masked using RepeatMasker (http://repeatmasker.org, last accessed September 23, 2013).

Gene Synteny

We used BLASTP version 2.2.26+ (Altschul et al. 1997) and MCScanX programs (Wang et al. 2012) to perform the gene synteny analysis. BLASTP provided the inputs for MCScanX. To generate more reasonable results with MCScanX, the BLASTP hits were restricted to the top 5 and to E values less than 1E−10. In this study, two sets of BLASTP results were generated for two different purposes. For gene clustering, each protein was queried using BLASTP against a database that includes all of the proteins from the six species. For the evaluation of the papilionoid WGD nodes, each papilionoid protein was queried against its self-database.

Phylogenetic Reconstruction

The phylogenetic reconstruction approach was derived from the method developed by Vilella et al. (2009). However, we modified this method by integrating collinear relationships between genes into the gene clustering step. Briefly, in the graph construction, the edges between the nodes (proteins) were retained if they satisfied any of the following three conditions: a best reciprocal, a Blast score ratio (BSR) of at least 0.33, and a collinear gene pair (the difference). The connected components were then extracted from the graph using single-linkage clustering. Each connected component represents a cluster or a gene family. Using the MUSCLE version 3.7 program (Edgar 2004), a protein alignment was conducted for each cluster. TreeBeST version 1.92 was used to build the gene trees and classify the nodes as either specialization or duplication. For example, a node tagged as “papilionoid” and “duplication” by TreeBeST means that the specified duplication event was shared by all papilionoids in the study. The definition of BSR and other details of the earlier mentioned steps can be found in the report published by Vilella et al. (2009).

Identification and Evaluation of Papilionoid-Specific WGD Nodes

For each gene tree, the identification of papilionoid-specific WGD nodes began with the last common ancestor node of all papilionoid genes, and passed down until the first papilionoid nodes were found. The first non-tandem duplication papilionoid nodes are considered to be papilionoid-specific WGD nodes, and the other first papilionoid nodes are considered to be WGD-loss nodes in which one of the WGD-derived branches was lost. We used both the gene rank and the genomic coordinate to discriminate a WGD node from a tandem duplication node. All of the gene pairs split by the duplication node were analyzed to determine whether at least one gene pair met one of the following two restrictions: a gene rank distance between the two paired genes of less than 50 or a genomic distance of less than 200 kb. If any gene pair met one of these two restrictions, the node was considered to be a tandem duplication node; however, if any gene pair did not meet either of the two restrictions, it was classified as a WGD node.

The results of the gene collinearity analysis were also used to evaluate the accuracy of the identified papilionoid WGD nodes. We first calculated the pairwise $K_s$ (synonymous substitutions) values between the gene pairs with collinear relationships using the codeml program from the PAML package (Yang 2007) and the median $K_s$ values for each paralogous collinear block in the four legumes. For a given papilionoid WGD node, if at least one gene pair of all of the gene pairs split by the node is located in a collinear paralogous block from any of the four papilionoid species with median $K_s$ values between 0.2 and 1.2 (these blocks were considered to be papilionoid-WGD–derived paralogous blocks), then, the papilionoid WGD node is directly supported by gene synteny.

GO Annotation and GO Enrichment Analysis

The GO annotations of the M. truncatula genes, including molecular function, molecular location, and biological
process, were conducted using the online tool Goanna (McCarthy et al. 2006; http://agbase.msstate.edu/cgi-bin/tools/GOanna.cgi, last accessed September 23, 2013). The GO enrichment analysis was performed using GOstats with a threshold P value of less than 0.01 (Falcon and Gentleman 2007). The GO slims, which are a subset of GO terms for a broad overview of the ontology content, were downloaded from http://www.geneontology.org/GO_slims/goslim_plant.obo (last accessed September 23, 2013). The metabolic pathways of G. max were downloaded from the PlantCyc database (http://www.plantcyc.org, last accessed September 23, 2013).

Gene Expression Data and Nodule Enrichment Analysis

The transcriptome data of M. truncatula (Young et al. 2011) were downloaded from the SRA database (http://www.ncbi.nlm.nih.gov/sra, last accessed September 23, 2013). These data included six tissues: root (SRS265483), flower (SRS265482), bud (SRS265481), blade (SRS265480), seed (SRS265479), and nodule (SRS265478). The accession numbers for the corresponding six tissues of G. max are SRS024744, SRS024739, SRS024738, SRS024741, SRS024740, and SRS024742, respectively (Libault et al. 2010). The numbers of all duplicate and all nonduplicate genes expressed in a nodule are m and n, respectively; in addition, the numbers of duplicate and nonduplicate genes expressed in a nodule and have a GO term of interest are q and k, respectively. We then used a hypergeometric distribution to compute the probability that the number of duplicates expressed in a nodule is not less than q.

The complete computational pipeline was coded in either Perl using the BioPerl version 1.60 software (Stajich et al. 2002) or the R programming language in the Bioconductor version 2.15 platform (Gentleman et al. 2004). The flow diagram of the pipeline is shown in supplementary figure S25, Supplementary Material online.

Supplementary Material

Supplementary figures S1–S25, data sets S1 and S2, and table S1 are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors are grateful to Dr Hugo Zheng at Department of Biology, McGill University for help with improvements to the English text. This work was supported by the National Natural Science Foundation of China (grant 30971848), Fundamental Research Funds for the Central Universities (grants KYYT201002 and KYZZ201202-9), the Specialized Research Fund for the Doctoral Programs of Higher Education (grants 20100097110035 and 20120097110023), 111 project (grant B08025), and the Priority Academic Program Development of Jiangsu Higher Education Institutions.

References


