Comparative genomics suggests that an ancestral polyploidy event leads to enhanced root nodule symbiosis in the Papilionoideae

Article

Accepted Version

MBE Advance Access published September 4, 2013

Article

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]

[1] State 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

[2] State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, People's Republic of China

[3] Henan Sesame Research Center, Henan Academy of Agricultural Sciences, Zhengzhou, People's Republic of China

[4] School of Agriculture, Policy and Development, University of Reading, Earley Gate, Reading RG6 6AR, UK

***Corresponding author**: E-mail: soyzhang@njau.edu.cn or soyzhang@hotmail.com

Running Title: Role of papilionoid polyploidy event

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 occurred approximately 58 million years ago. 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 signalling, 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.

Keywords: 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, non-legume 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. By 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 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 million years ago (MYA), shortly after the origin of legumes at approximately 60 MYA. 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 recognised 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 signalling 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 non-legume 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 (Dehal 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 non-legume 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. Based on 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 dataset 1 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 dataset 1 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 dataset 1 online). Because this polyploidy event occurred approximately 58 MYA 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 non-supported 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 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 polyploidy event (Gu et al. 2003; He and Zhang 2005; Wagner 2008; Van de Peer et al. 2009).

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 over-represented in the genes of interest compared with all the annotated genes of genome. Similarly, 694 GO terms were identified in *M. truncatula*. In addition, 440 of these GO terms were found in both *G. max* and *M. truncatula* (see supplementary dataset 2 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 dataset 2 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, according to the molecular mechanisms of RNS described by Oldroyd et al. (2011) and Bapaume and Reinhardt (2012). **Fig. 2** shows selected GO terms likely associated with RNS and gene expression in nodules.

Symbiotic signalling

Flavonoids are essential signal molecules that play multiple roles in legume-*rhizobium* symbiosis, e.g., the induction of the biosynthesis of the Nod factor (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 dataset 2 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 (supplementary fig. S1), CHR (supplementary fig. S2), F3H (supplementary fig. S3), FLS (supplementary fig. S4), F3'H (supplementary fig. S5), HIDH (supplementary fig. S6), I2'H (supplementary fig. S7) and IOMT (supplementary fig. S8). These results suggest that more abundant and diverse flavonoids would be synthesised as a result of the polyploidy event and that the enrichment of flavonoids might be adaptive for the complex signalling 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 **fig. 4**, the papilionoid LysM receptors closely homologous to AT3G21630 (chitin elicitor receptor kinase 1, AtCERK1), which recognises 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 **fig. 4** also indicates that these NFRs might have been evolved from an original chitin elicitor receptor kinase involved in plant defence, in a similar way to AtCERK1. Therefore, the evolution of RNS is an interesting example of the transition from resistance to cooperation 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 dataset 2 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 signalling, such as *NFR5*, *SINA4*, *ERN2*, *NSP2* and *MtHMGR1*, also have papilionoid polyploidy paralogues (**Table 2**). These increased symbiotic signalling genes suggest that the papilionoid polyploidy event might have induced the emergence of critical symbiotic genes and increased the complexity of the symbiotic signalling 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 signalling and polar auxin transport (Oldroyd et al. 2011). Enriched GO terms, such as GO:0009755 (hormone-mediated signalling pathway) and GO:0009736 (cytokinin-mediated signalling pathway), were consistently identified as significant (**fig. 2** and supplementary dataset 2 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 signalling 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 dataset 2 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 N_2 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 localisation), were identified (**fig. 2** and supplementary dataset 2 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 MsPG3 (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 paralogues (**Table 2**). Additional genes involved in the formation of ITs, such as *nsRING*, *PUB1*, *sickle* and *NIP*, also have papilionoid-WGD-derived paralogues (**Table 2**).

Nutrient exchange and transport

Inside legume nodules, rhizobial bacteroids reduce N_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 dataset 2 online). Plant glutamine synthetase (GS) is a

key enzyme that assimilates the ammonium produced by bacteroids, and MtGS1b has WGD-derived paralogues (**Table 2**, Carvalho et al. 1997). Other protein families associated with nutrient transport, such as the putative glutamine dumper family (supplementary fig. S22), the putative oligopeptide transporter family (supplementary fig. S23) and the putative amino acid permease family (supplementary fig. S24), were also amplified by this polyploidy event.

In addition, 84% of these 1,160 nodes are supported by gene synteny (supplementary dataset 1 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 dataset 2 online). 209 of these terms coincided with the 440 GO terms mentioned above (supplementary dataset 2 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. Since 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 co-retained 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 adaptation 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 unravelling of the shared or unique molecular mechanisms of papilionoid RNS. In addition, this genome-wide comparative study generated phylogenetic trees of legume genes which could also provide clues for the evolution of other legume traits, such as secondary metabolites.

Materials and Methods

Genomic data and annotation

The genomic data of four papilionoid species (*Glycine max*, *Medicago truncatula*, *Lotus japonicus* and *Cajanus cajan*) and two non-legumes (*Arabidopsis thaliana* and *Prunus persica*) were downloaded. The detailed gene annotations for the six species are shown in Supplementary table S1. 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).

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 (BRH), 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 specialisation 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 above 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 *Ks* (synonymous substitutions) values between the gene pairs with collinear relationships using the codeml program from the PAML package (Yang 2007) and the median *Ks* 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 *Ks* 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). 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. The metabolic pathways of *G. max* were downloaded from the PlantCyc database (http://www.plantcyc. org).

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). 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 non-duplicate genes expressed in a nodule are m and n, respectively; in addition, the numbers of duplicate and non-duplicate 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 fig. S25.

Supplementary Material

Supplementary figures S1-S25, supplementary datasets 1 and 2 and supplementary table S1 are available at Molecular Biology and Evolution online.

Acknowledgments and funding information

We are grateful to the International Peach Genome Initiative for the assembled and annotated draft of the peach genome, and 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 (30971848), Fundamental Research Funds for the Central

Universities (KYT201002, KYZ201202-9), the Specialised Research Fund for the Doctoral Programs of Higher Education (20100097110035, 20120097110023), 111 project (grant number B08025) and the Priority Academic Program Development of Jiangsu Higher Education Institutions.

References

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

Andriankaja A, Boisson-Dernier A, Frances L, Sauviac L, Jauneau A, Barker DG, de Carvalho-Niebel F. 2007. AP2-ERF transcription factors mediate Nod Factor–dependent MtENOD11 activation in root hairs via a novel cis-regulatory motif. Plant Cell 19: 2866–2885.

Bapaume L, Reinhardt D. 2012. How membranes shape plant symbioses: signaling and transport in nodulation and arbuscular mycorrhiza. Front Plant Sci. 3: 223.

Birchler JA, Veitia RA. 2007. The gene balance hypothesis: from classical genetics to modern genomics. Plant Cell 19: 395–402.

Birchler JA, Veitia RA. 2012. Gene balance hypothesis: Connecting issues of dosage sensitivity across biological disciplines. Proc Natl Acad Sci U S A. 109: 14746–14753.

Blanc G, Wolfe KH. 2004. Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution. Plant Cell 16: 1679–1691.

Bremer B, Bremer K, Chase M, Fay M, Reveal J, Soltis D, Solitis P, Stevens P. 2009. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III. Bot J Linn Soc. 161: 105–121.

Brunet FG, Roest Crollius H, Paris M, Aury JM, Gibert P, Jaillon O, Laudet V, Robinson-Rechavi M. 2006. Gene loss and evolutionary rates following whole-genome duplication in teleost fishes. Mol Biol Evol. 23: 1808–1816.

Cannon SB, Ilut D, Farmer AD, Maki SL, May GD, Singer SR, Doyle JJ. 2010. Polyploidy did not predate the evolution of nodulation in all legumes. PLoS ONE 5: e11630.

Carvalho H, Sunkel C, Salema R, Cullimore JV. 1997. Heteromeric assembly of the cytosolic glutamine synthetase polypeptides of *Medicago truncatula*: complementation of a *glnA Escherichia coli* mutant with a plant domain-swapped enzyme. Plant Mol Biol. 35: 623–632.

Crespi M, Frugier F. 2008. De novo organ formation from differentiated cells: root nodule organogenesis. Sci Signal. 1: re11.

Dehal P, Boore JL. 2005. Two rounds of whole genome duplication in the ancestral vertebrate. PLoS Biology 3: e314.

Doyle JJ. 2011. Phylogenetic perspectives on the origins of nodulation. Mol Plant Microbe In. 24: 1289–1295.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32: 1792–1797.

Edger PP, Pires JC. 2009. Gene and genome duplications: the impact of dosage-sensitivity on the fate of nuclear genes. Chromosome Res. 17: 699–717.

Falcon S, Gentleman R. 2007. Using GOstats to test gene lists for GO term association. Bioinformatics 23: 257–258.

Gentleman RC, Carey VJ, Bates DM, et al. (25 co-authors). 2004. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 5: R80.

Gu Z, Steinmetz, LM, Gu X, Scharfe C, Davis RW, Li WH. 2003. Role of duplicate genes in genetic robustness against null mutations. Nature 421: 63–66.

He XL, Zhang JZ. 2005. Gene complexity and gene duplicability. Curr Biol. 15: 1016–1021.

Herder GD, Yoshida S, Antolín-Llovera M, Ried MK, Parniske M. 2008. Seven in absentia proteins affect plant growth and nodulation in *Medicago truncatula*. Plant Physiol. 148: 369–382.

Kaló P, Gleason C, Edwards A, et al. (13 co-authors). 2005. Nodulation signaling in legumes requires NSP2, a member of the GRAS family of transcriptional regulators. Science 308: 1786–1789.

Ke D, Fang Q, Chen C, Zhu H, Chen T, Chang X, Yuan S, Kang H, Ma L, Hong Z, Zhang Z. 2012. The small GTPase ROP6 interacts with NFR5 and is involved in nodule formation in *Lotus japonicus*. Plant Physiol. 159: 131–143.

Kevei Z, Lougnon G, Mergaert P, Horváth GV, Kereszt A, Jayaraman D, Zaman N, Marcel F, Regulski K, Kiss GB, Kondorosi A, Endre G, Kondorosi E, Ané JM. 2007. 3-Hydroxy-3-Methylglutaryl Coenzyme A Reductase1 Interacts with NORK and Is crucial for nodulation in *Medicago truncatula*. Plant Cell 19: 3974–3989.

Kondrashov FA, Rogozin IB, Wolf YI, Koonin EV. 2002. Selection in the evolution of gene duplications. Genome Biol. 3: research0008.1–0008.9.

Li C, Li QG, Dunwell JM, Zhang YM. 2012. Divergent evolutionary pattern of starch biosynthetic pathway genes in grasses and dicots. Mol Biol Evol. 29: 3227–3236

Li Q-G, Zhang Y-M. 2012. The original and function transition of *P34*. Heredity 110: 259–266.

Li W-H, Yang J, Gu X. 2005. Expression divergence between duplicate genes. Trends Genet. 21: 602–607.

Libault M, Farmer A, Joshi T, Takahashi K, Langley RJ, Franklin LD, He J, Xu D, May G, Stacey G. 2010. An integrated transcriptome atlas of the crop model *Glycine max*, and its use in comparative analyses in plants. Plant J. 63: 86–99.

Limpens E, Franken C, Smit P, Willemse J, Bisseling T, Geurts R. 2003. LysM domain receptor kinases regulating rhizobial Nod Factor-induced infection. Science 302: 630–633.

Liu T, Liu Z, Song C, et al. (12 co-authors). 2012. Chitin-induced dimerization activates a plant immune receptor.

Science 336: 1160–1164.

Lynch M. 2007. *The origins of genome architecture*. Sinauer Associates, Inc.: Sunderland, MA.

Madsen EB, Madsen LH, Radutoiu S, et al. (11 co-authors). 2003. A receptor kinase gene of the LysM type is involved in legume perception of rhizobial signals. Nature 425: 637–640.

Makova KD, Li WH. 2003. Divergence in the spatial pattern of gene expression between human duplicate genes. Genome Res. 13: 1638–1645.

Mbengue M, Camut S, de Carvalho-Niebel F, Deslandes L, et al. (12 co-authors). 2010. The *Medicago truncatula* E3 ubiquitin ligase PUB1 interacts with the LYK3 symbiotic receptor and negatively regulates infection and nodulation. Plant Cell 22: 3474–3488.

McCarthy FM, Wang N, Magee GB, et al. (13 co-authors). 2006. AgBase: a functional genomics resource for agriculture. BMC Genomics 7: 229.

Miya A, Albert P, Shinya T, Desaki Y, Ichimura K, Shirasu K, Narusaka Y, Kawakami N, Kaku H, Shibuya N. 2007. CERK1, a LysM receptor kinase, is essential for chitin elicitor signaling in *Arabidopsis*. Proc Natl Acad Sci U S A. 104: 19613–19618.

Muñoz JA, Coronado C, Pérez-Hormaeche J, Kondorosi A, Ratet P, Palomares AJ. 1998. MsPG3, a Medicago sativa polygalacturonase gene expressed during the alfalfa–*Rhizobium* meliloti interaction. Proc Natl Acad Sci U S A. 95: 9687–9692.

Murray JD, Karas BJ, Sato S, Tabata S, Amyot L, Szczyglowski K. 2007. A cytokinin perception mutant colonized by *rhizobium* in the absence of nodule organogenesis. Science 315: 101–104.

Ohno S. 1970. Evolution by gene duplication. Springer-Verlag, New York.

Oldroyd GED, Long SR. 2003. Identification and characterization of nodulation-signaling pathway 2, a gene of *Medicago truncatula* involved in nod factor signaling. Plant Physiol. 131: 1027–1032.

Oldroyd GED, Murray JD, Poole PS, Downie JA. 2011. The rules of engagement in the legume-rhizobial symbiosis. Annu Rev Genet. 45: 119–144.

Op den Camp RH, De Mita S, Lillo A, Cao Q, Limpens E, Bisseling T, Geurts R. 2011. A phylogenetic strategy based on a legume-specific whole genome duplication yields symbiotic cytokinin type-A response regulators. Plant Physiol. 157: 2013–2022.

Pawlowski K, Sprent JI. 2008. Nitrogen-fixing actinorhizal symbioses. Springer: Dordrecht, The Netherlands.

Penmetsa RV, Uribe P, Anderson J, et al. (17 co-authors). 2008. The *Medicago truncatula* ortholog of Arabidopsis EIN2, sickle, is a negative regulator of symbiotic and pathogenic microbial associations. Plant J. 55: 580–595.

Radutoiu S, Madsen LH, Madsen EB, et al. (11 co-authors). 2003. Plant recognition of symbiotic bacteria requires two LysM receptor-like kinases. Nature 425: 585–592.

Santi C, Bogusz D, Franche C. 2013. Biological nitrogen fixation in non-legume plants. Ann Bot. 111: 743–767.

Sato S, Nakamura Y, Kaneko T, Asamizu E, Kato T, Nakao M, Sasamoto S, Watanabe A, Ono A, Kawashima K.

2008. Genome structure of the legume, *Lotus japonicus*. DNA Res. 15: 227–239.

Schmutz J, Cannon SB, Schlueter J, et al. (45 co-authors). 2010. Genome sequence of the palaeopolyploid soybean. Nature 463: 178–183.

Shimomura K, Nomura M, Tajima S, Kouchi H. 2006. LjnsRING, a novel RING finger protein, is required for symbiotic interactions between *Mesorhizobium loti* and *Lotus japonicus*. Plant Cell Physiol. 47: 1572–1581.

Singer SR, Maki SL, Farmer AD, Ilut D, May GD, Cannon SB, Doyle JJ. 2009. Venturing beyond beans and peas: what can we learn from Chamaecrista? Plant Physiol. 151: 1041–1047.

Soltis DE, Soltis PS, Morgan DR, Swensen SM, Mullin BC, Dowd JM, Martin PG. 1995. Chloroplast gene sequence data suggest a single origin of the predisposition for symbiotic nitrogen fixation in angiosperms. Proc Natl Acad Sci U S A. 92: 2647–2651.

Sprent JI. 2001. Nodulation in legumes. Royal Botanic Gardens, Kew.

Stajich JE, Block D, Boulez K, et al. (21 co-authors). 2002. The Bioperl toolkit: Perl modules for the life sciences. Genome Res. 12: 1611–1618.

Subramanian S, Stacey G, Yu O. 2007. Distinct, crucial roles of flavonoids during legume nodulation. Trends Plant Sci. 12: 282–285.

Swensen SM, Benson DR. 2008. Evolution of actinorhizal host plants and *Frankia* endosymbioses. In: Pawlowski K, Newton WE, editors. Nitrogen-fixing actinorhizal symbioses. Series: Nitrogen fixation: origins, applications, and research progress. Springer: Boston. p. 73–104.

Thomas BC, Pedersen B, Freeling M. 2006. Following tetraploidy in an *Arabidopsis* ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. Genome Res. 16: 934–946.

Tirichine L, Sandal N, Madsen LH, Radutoiu S, Albrektsen AS, Sato S, Asamizu E, Tabata S, Stougaard J. 2007. A gain-of-function mutation in a cytokinin receptor triggers spontaneous root nodule organogenesis. Science 315: 104–107.

Van de Peer Y, Maere S, Meyer A. 2009. The evolutionary significance of ancient genome duplications. Nat Rev Genet. 10: 725–732.

Wagner A. 2008. Gene duplications, robustness and evolutionary innovations. BioEssays 30: 367–373.

Wan J, Zhang XC, Neece D, Ramonell KM, Clough S, Kim SY, Stacey MG, Stacey G. 2008. A LysM receptor-like kinase plays a critical role in chitin signaling and fungal resistance in *Arabidopsis*. Plant Cell 20: 471–481.

Wang Y, Tang H, Debarry JD, Tan X, Li J, Wang X, Lee TH, Jin H, Marler B, Guo H, Kissinger JC, Paterson AH. 2012. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res. 40: e49.

Varshney RK, Chen W, Li Y, Bharti AK, Saxena RK, Schlueter JA, Donoghue MT, Azam S, Fan G, Whaley AM. 2011. Draft genome sequence of pigeonpea (*Cajanus cajan*), an orphan legume crop of resource-poor farmers.

Nat Biotechnol. 30: 83–89.

Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E. 2009. EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. Genome Res. 19: 327–335.

Yang Z. 2007. PAML4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 24: 1586–1591.

Young ND, Debellé F, Oldroyd GE, et al. (125 co-authors). 2011. The *Medicago* genome provides insight into the evolution of rhizobial symbioses. Nature 480: 520–524.

Zahran HH. 1999. Rhizobium-legume symbiosis and nitrogen fixation under severe conditions and in an arid climate. Microbiol Mol Biol Rev. 63: 968–989.

Table 1. Overrepresented GO slim terms after the papilionoid polyploidy event compared with the individual genome background

GO ID	Term	*P*-value	
		M. truncatula	*G. max*
Biological Process			
GO:0008219	cell death	1.49E-24	4.67E-21
GO:0016265	death	1.49E-24	4.67E-21
GO:0019748	secondary metabolic process	1.36E-24	1.86E-19
GO:0007165	signal transduction	2.48E-17	9.45E-17
GO:0009856	pollination	2.66E-06	3.18E-13
GO:0007154	cell communication	7.91E-11	4.39E-12
GO:0009719	response to endogenous stimulus	6.91E-16	2.11E-10
GO:0040007	growth	6.49E-09	2.66E-09
GO:0016049	cell growth	6.70E-05	1.04E-06
GO:0007610	behavior	1.62E-03	7.96E-06
GO:0009607	response to biotic stimulus	3.27E-03	3.75E-04
GO:0006810	transport	2.31E-12	5.84E-04
Molecular Function			
GO:0016740	transferase activity	1.57E-03	2.87E-04
GO:0004871	signal transducer activity	2.27E-07	5.04E-04
GO:0003700	transcription factor activity	8.65E-11	3.08E-03
Cellular Component			
GO:0005886	plasma membrane	8.80E-37	3.37E-21
GO:0016020	membrane	2.22E-15	5.60E-04

P-value: indicate the significance of overrepresented GO terms conducted by GOstat software

Table 2. Symbiosis-related genes that retained their papilionoid WGD paralogues

Gene	Accesion	Genome ID	Gene tree	Reference
Symbiotic signaling				
NFR1	CAE02591.1	chr2.CM0545.250.r2.m_lj	fig. 4	Radutoiu et al. 2003
LYK3	AAQ73159.1	Medtr5g086130_mt	fig. 4	Limpens et al. 2003
NFR5	CAE02597.1	chr2.CM0323.400.r2.d_lj	Supplementary fig. S9	Madsen et al. 2003
SINA4	ABW70162.1	Medtr3g091510_mt	Supplementary fig. S10	Herder et al. 2008
ERN2	ABW06103.2	Medtr6g029180_mt	Supplementary fig. S11	Andriankaja et al. 2007
NSP2	Q5NE24.1	Medtr3g072710_mt	Supplementary fig. S12	Oldroyd and Long 2003; Kaló et al. 2005
MtHMGR1	ABY20972.1	Medtr5g026500_mt	Supplementary fig. S13	Kevei et al. 2007
Nodule organogenesis				
LHK1	CAL18382.1	chr4.CM0042.1600.r2.m_lj	Supplementary fig. S14	Tirichine et al. 2007; Murray et al. 2007
MtRR9	AET86869.1	Medtr3g015490_mt	Supplementary fig. S15	Op den Camp et al. 2011
Rhizobial infection				
MsPG3	CAA72003.1 (*Medicago sativa*)		Supplementary fig. S16	Muñoz et al. 1998
ROP6	ADY16660.1	chr2.CM0272.860.r2.m_lj	Supplementary fig. S17	Ke et al. 2012
nsRING	BAF38781.1	chr4.CM0042.810.r2.m_lj	Supplementary fig. S18	Shimomura et al. 2006
PUB1	DAA33939.1	Medtr5g083030_mt	Supplementary fig. S19	Mbengue et al. 2010
sickle	ACD84889.1	Medtr7g101410_mt	Supplementary fig. S20	Penmetsa et al. 2008
Nutrient Exchange and transport				
MtGS1b	CAA71317.1	Medtr3g065250_mt	Supplementary fig. S21	Carvalho et al. 1997

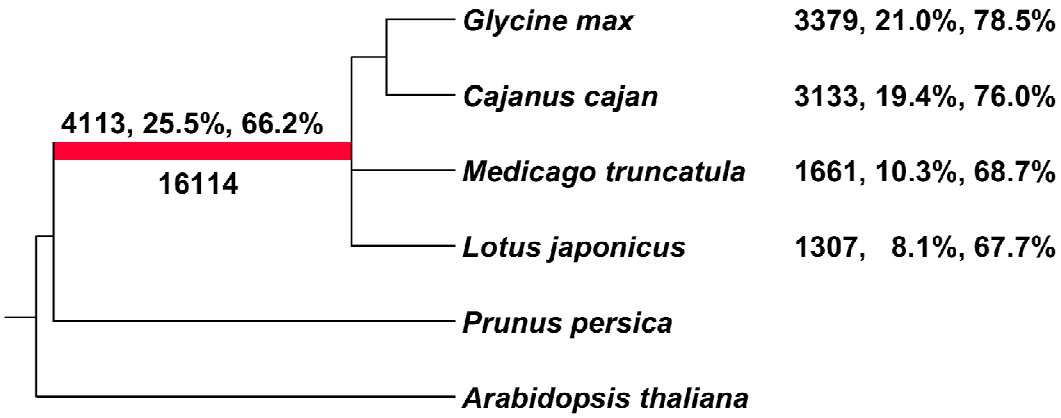

FIG. 1. Cladogram of species tree and numbers of gene retentions of papilionoid WGD-derived duplicates. The topology of the species tree is from APG III (Bremer et al. 2009), and branch lengths of this tree show no means. The bold branch indicates the papilionoid polyploidy event. The number 16,114 indicates 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 4113 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. *G. max* for example, 3379 duplicate branches are retained, which represent 21.0% of the total 16,114 papilionoid nodes, and 78.5% of the 3379 nodes are supported by gene collinearity.

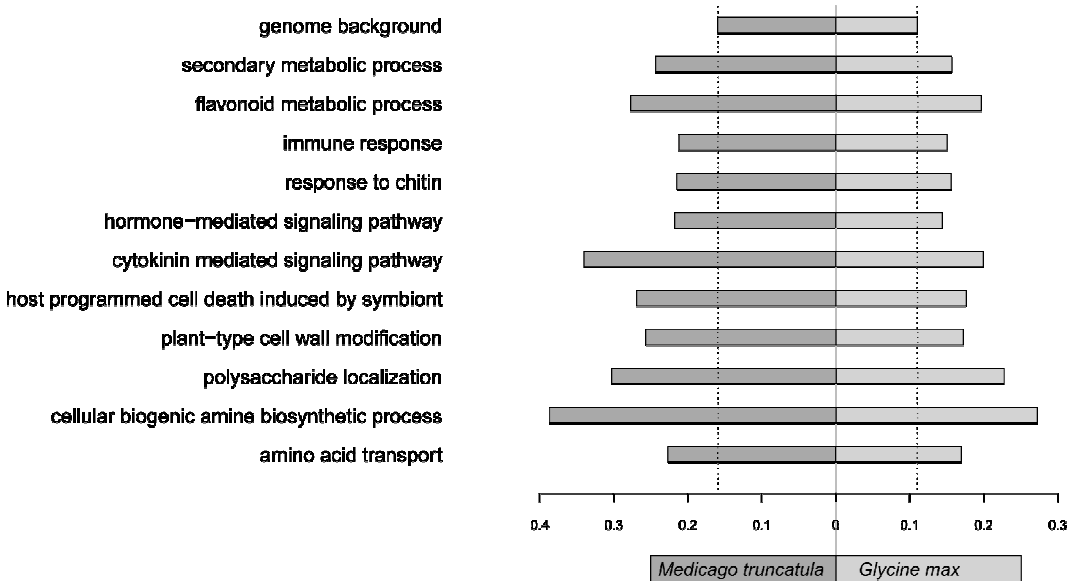

FIG. 2. Selected GO terms with excess gene duplicates expressed in nodules. The horizontal axis represents the proportion of duplicate genes expressed in nodules. For all these selected biological process, the gene duplicates expressed in nodules are in excess compared with the individual genome background which represents the proportion of gene duplicates in all genes annotated in biological process and expressed in nodules.

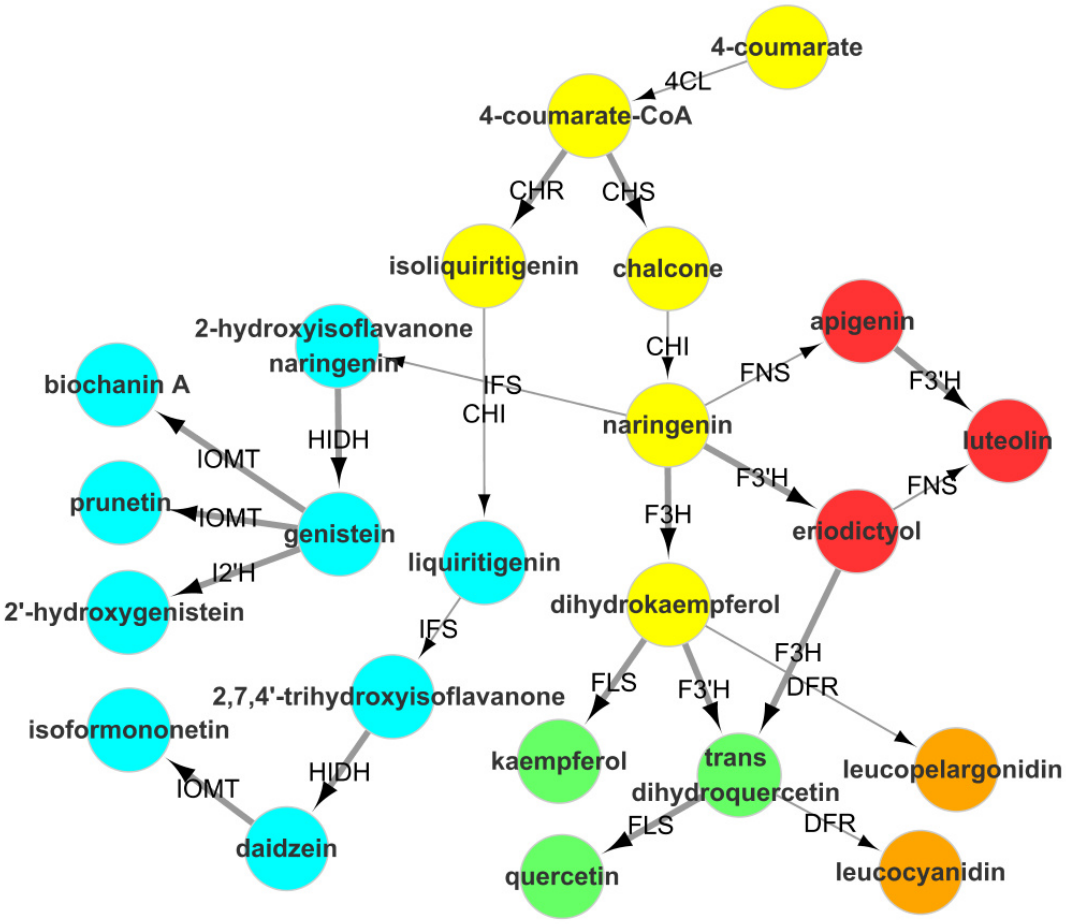

FIG. 3. Simplified schematic of flavonoid biosynthesis. This scheme was derived from SoyCyc, which was downloaded from the Plant Metabolic Network database (http://www.plantcyc.org). The pathway includes chalcone biosynthesis (yellow), flavone biosynthesis (red), flavonol biosynthesis (green), leucoanthocyanidin (flavan-3,4-diol) biosynthesis (orange) and isoflavonoid biosynthesis (cyan). The bold edges indicate those reactions that are catalysed by enzymes duplicated as a result of papilionoid polyploidy event: CHS (chalcone synthase), CHR (chalcone reductase), F3H (flavanone 3-hydroxylase), FLS (flavonol synthase), F3'H (flavonoid 3'-hydroxylase), HIDH (2-hydroxyisoflavanone dehydratase), I2'H (isoflavone-2'-hydroxylase) and IOMT (isoflavone 7-O-methyltransferase or isoflavone 4'-O-methyltransferase).

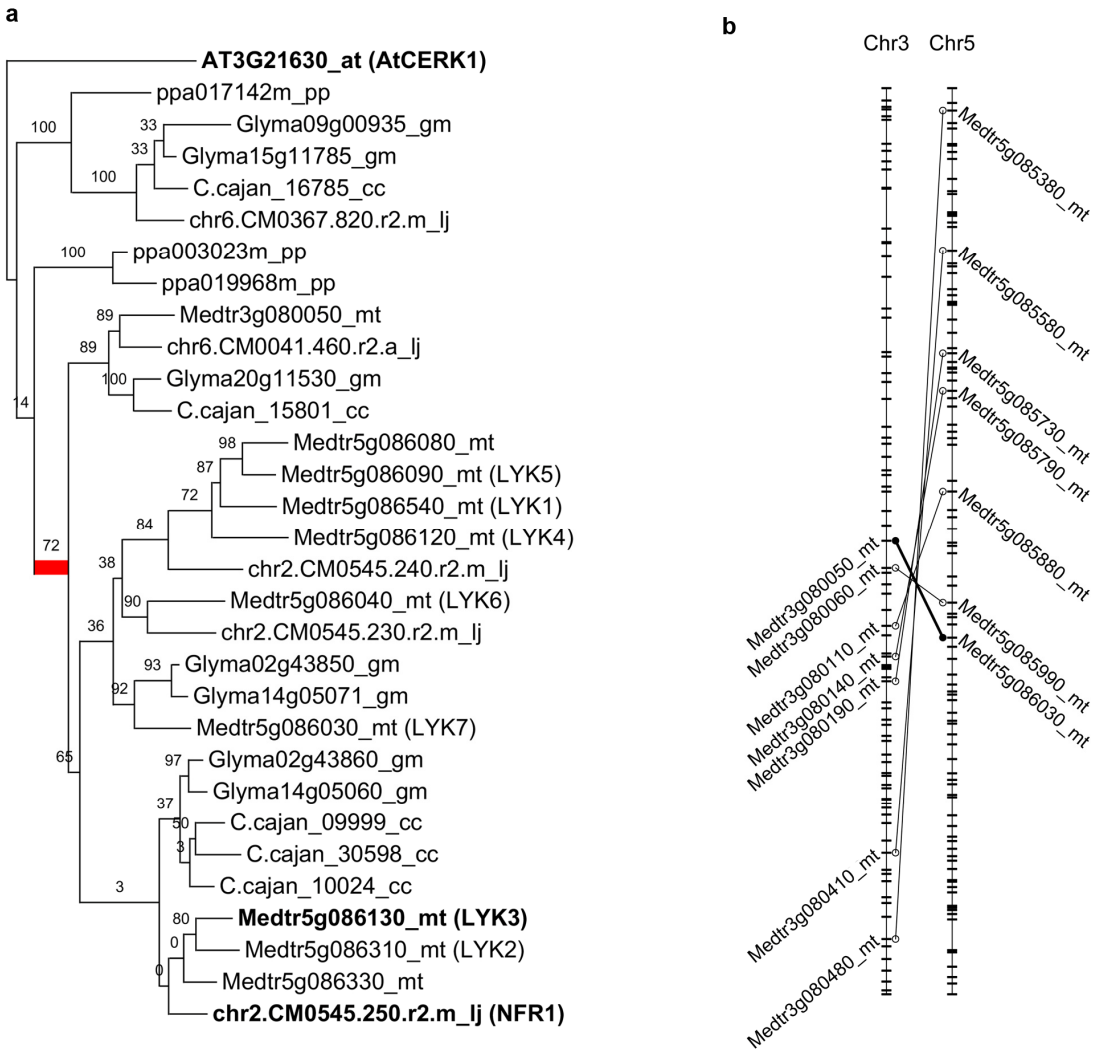

FIG. 4. Phylogenetic tree of the NF receptor LYK/NFR1 (a) and an example of gene syntenic blocks (b). (a) The phylogenetic tree was estimated with TreeBeST software, in which the numbers on the branches of the phylogenetic tree represent the bootstrap supports. The bold branch represents the papilionoid polyploidy event that is also supported by the gene synteny analysis conducted by MCScanX software. (b) This figure provides such an example from *Medicago truncatula* including a collinear gene pair of Medtr5g086030 and Medtr3g080050 with a bold line.