LETTER

doi:10.1038/nature10625

The *Medicago* genome provides insight into the evolution of rhizobial symbioses

Nevin D. Young[1]*, Frédéric Debellé[2,3]*, Giles E. D. Oldroyd[4]*, Rene Geurts[5], Steven B. Cannon[6,7], Michael K. Udvardi[8], Vagner A. Benedito[9], Klaus F. X. Mayer[10], Jérôme Gouzy[2,3], Heiko Schoof[11], Yves Van de Peer[12], Sebastian Proost[12], Douglas R. Cook[13], Blake C. Meyers[14], Manuel Spannagl[10], Foo Cheung[15], Stéphane De Mita[5], Vivek Krishnakumar[15], Heidrun Gundlach[10], Shiguo Zhou[16], Joann Mudge[17], Arvind K. Bharti[17], Jeremy D. Murray[4,8], Marina A. Naoumkina[8], Benjamin Rosen[13], Kevin A. T. Silverstein[18], Haibao Tang[15], Stephane Rombauts[12], Patrick X. Zhao[8], Peng Zhou[1], Valérie Barbe[19], Philippe Bardou[2,3], Michael Bechner[16], Arnaud Bellec[20], Anne Berger[19], Hélène Bergès[20], Shelby Bidwell[15], Ton Bisseling[5,21], Nathalie Choisne[19], Arnaud Couloux[19], Roxanne Denny[1], Shweta Deshpande[22], Xinbin Dai[8], Jeff J. Doyle[23], Anne-Marie Dudez[2,3], Andrew D. Farmer[17], Stéphanie Fouteau[19], Carolien Franken[5], Chrystel Gibelin[2,3], John Gish[13], Steven Goldstein[16], Alvaro J. González[24], Pamela J. Green[14], Asis Hallab[25], Marijke Hartog[5], Axin Hua[22], Sean J. Humphray[26], Dong-Hoon Jeong[14], Yi Jing[22], Anika Jöcker[25], Steve M. Kenton[22], Dong-Jin Kim[13,27], Kathrin Klee[25], Hongshing Lai[22], Chunting Lang[5], Shaoping Lin[22], Simone L. Macmil[22], Ghislaine Magdelenat[19], Lucy Matthews[26], Jamison McCorrison[15], Erin L. Monaghan[15], Jeong-Hwan Mun[13,28], Fares Z. Najar[22], Christine Nicholson[26], Céline Noirot[29], Majesta O'Bleness[22], Charles R. Paule[1], Julie Poulain[19], Florent Prion[2,3], Baifang Qin[22], Chunmei Qu[22], Ernest F. Retzel[17], Claire Riddle[26], Erika Sallet[2,3], Sylvie Samain[19], Nicolas Samson[2,3], Iryna Sanders[22], Olivier Saurat[2,3], Claude Scarpelli[19], Thomas Schiex[29], Béatrice Segurens[19], Andrew J. Severin[7], D. Janine Sherrier[14], Ruihua Shi[22], Sarah Sims[26], Susan R. Singer[30], Senjuti Sinharoy[8], Lieven Sterck[12], Agnès Viollet[19], Bing-Bing Wang[1], Keqin Wang[22], Mingyi Wang[8], Xiaohong Wang[1], Jens Warfsmann[25], Jean Weissenbach[19], Doug D. White[22], Jim D. White[22], Graham B. Wiley[22], Patrick Wincker[19], Yanbo Xing[22], Limei Yang[22], Ziyun Yao[22], Fu Ying[22], Jixian Zhai[14], Liping Zhou[22], Antoine Zuber[2,3], Jean Dénarié[2,3], Richard A. Dixon[8], Gregory D. May[17], David C. Schwartz[16], Jane Rogers[31], Francis Quétier[19], Christopher D. Town[15] & Bruce A. Roe[22]

Legumes (Fabaceae or Leguminosae) are unique among cultivated plants for their ability to carry out endosymbiotic nitrogen fixation with rhizobial bacteria, a process that takes place in a specialized structure known as the nodule. Legumes belong to one of the two main groups of eurosids, the Fabidae, which includes most species capable of endosymbiotic nitrogen fixation[1]. Legumes comprise several evolutionary lineages derived from a common ancestor 60 million years ago (Myr ago). Papilionoids are the largest clade, dating nearly to the origin of legumes and containing most cultivated species[2]. *Medicago truncatula* is a long-established model for the study of legume biology. Here we describe the draft sequence of the *M. truncatula* euchromatin based on a recently completed BAC assembly supplemented with Illumina shotgun sequence, together capturing ~94% of all *M. truncatula* genes. A whole-genome duplication (WGD) approximately 58 Myr ago had a major role in shaping the *M. truncatula* genome and thereby contributed to the evolution of endosymbiotic nitrogen fixation. Subsequent to the WGD, the *M. truncatula* genome experienced higher levels of rearrangement than two other sequenced legumes, *Glycine max* and *Lotus japonicus*. *M. truncatula* is a close relative of alfalfa (*Medicago sativa*), a widely cultivated crop with limited genomics tools and complex autotetraploid genetics. As such, the *M. truncatula* genome sequence provides significant opportunities to expand alfalfa's genomic toolbox.

Optical mapping indicates that the eight pseudomolecules of assembly Mt3.5 span a physical distance of 375 million base pairs (Mb), and fluorescence *in situ* hybridization indicates they extend from pericentromeres almost to telomeric ends (Supplementary Figs 1 and 2). Altogether, Mt3.5 consists of 2,536 bacterial artificial chromosomes (BACs; Supplementary Tables 1 and 2) with 273 physical gaps (including centromeres, Supplementary Table 3) and 101 internal sequencing gaps. The pseudomolecules contain 246 Mb of non-redundant sequence (Supplementary Table 2) located entirely within the optical map (Supplementary Fig. 3). Another 146 unfinished BACs/BAC pools that cannot be placed on the optical map contribute 17.3 Mb. Regions not represented in pseudomolecules or unanchored BACs were captured through assembly of approximately 40× coverage Illumina sequencing, yielding 104.2 Mb of additional unique sequence. Although not directly tested, the Illumina sequence is expected to lie

[1]Departments of Plant Pathology and Plant Biology, University of Minnesota, St Paul, Minnesota 55108, USA. [2]INRA, Laboratoire des Intéractions Plantes-Microorganismes (LIPM), UMR441, BP 52627, F-31326 Castanet-Tolosan CEDEX, France. [3]CNRS, Laboratoire des Intéractions Plantes-Microorganismes (LIPM), UMR2594, BP 52627, F-31326 Castanet-Tolosan CEDEX, France. [4]Department of Disease and Stress Biology, John Innes Centre, Norwich NR4 7UH, UK. [5]Laboratory of Molecular Biology, Department of Plant Science, Wageningen University, Droevendaalsesteeg 1, 6708PB Wageningen, The Netherlands. [6]USDA-ARS Corn Insects and Crop Genetics Research Unit, Ames, Iowa 50011, USA. [7]Department of Agronomy, Iowa State University, Ames, Iowa 50011, USA. [8]Plant Biology Division, Samuel Roberts Noble Foundation, 2510 Sam Noble Parkway, Ardmore, Oklahoma 73401, USA. [9]Department of Genetics and Developmental Biology, Plant and Soil Science Division, West Virginia University, Morgantown, West Virginia 26506, USA. [10]MIPS/Institute for Bioinformatics and Systems Biology, Helmholtz Center Munich, Ingolstädter Landstrasse 1, Neuherberg, Germany. [11]University of Bonn, INRES Crop Bioinformatics, Katzenburgweg 2, 53115 Bonn, Germany. [12]Department of Plant Systems Biology, VIB, Ghent University, Technologiepark 927, B-9052 Ghent, Belgium. [13]Department of Plant Pathology, University of California, Davis, California 95616, USA. [14]Department of Plant & Soil Sciences and Delaware Biotechnology Institute, University of Delaware, Newark, Delaware 19711, USA. [15]J. Craig Venter Institute, 9704 Medical Center Drive, Rockville, Maryland 20850, USA. [16]Laboratory for Molecular and Computational Genomics, University of Wisconsin-Madison, Wisconsin 53706, USA. [17]National Center for Genome Resources, 2935 Rodeo Park Drive East, Santa Fe, New Mexico 87505, USA. [18]Masonic Cancer Center, Biostatistics and Bioinformatics Group, University of Minnesota, Minneapolis, Minnesota 55455, USA. [19]Genoscope/Centre National de Séquençage, 2 rue Gaston Crémieux, CP 5706, 91057 Evry CEDEX, France. [20]INRA, Centre National de Ressources Génomiques Végétales (CNRGV), BP 52627, F-31326 Castanet-Tolosan CEDEX, France. [21]College of Science, King Saud University, Post Office Box 2455, Riyadh 11451, Saudi Arabia. [22]Advanced Center for Genome Technology, Department of Chemistry and Biochemistry, Stephenson Research and Technology Center, University of Oklahoma, Norman, Oklahoma 73019, USA. [23]Department of Plant Biology, Cornell University, Ithaca, New York, 14853 USA. [24]Department of Computer & Information Sciences, and Delaware Biotechnology Institute, University of Delaware, Newark, Delaware 19711, USA. [25]Max Planck Institute for Plant Breeding Research, Plant Computational Biology, Carl von Linné Weg 10, 50829 Köln, Germany. [26]Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK. [27]International Institute for Tropical Agriculture, (c/o P.O. Box 30709 Nairobi, Kenya 00100), Ibadan, Nigeria. [28]National Institute of Agricultural Biotechnology, Rural Development Administration, 225 Seodun-dong, Gwonseon-gu, Suwon 441-707, South Korea. [29]INRA, Unité de Biométrie et d'Intelligence Artificielle (UBIA), UR875, BP 52627, F-31326 Castanet-Tolosan CEDEX, France. [30]Department of Biology, Carleton College, Northfield, Minnesota 55057 USA. [31]The Genome Analysis Centre, Norwich Research Park, Norwich, Norfolk NR4 7UH, UK.
*These authors contributed equally to this work.

predominantly within the boundaries of pseudomolecules (see below). On the basis of expressed sequence tag alignments, the combined data sets capture ~94% of expressed genes, providing a highly informative platform for analysing the euchromatin of *M. truncatula*, although still at the draft stage.

Altogether there are 62,388 gene loci in Mt3.5 (Supplementary Table 4 and Supplementary Fig. 4), with 14,322 gene predictions annotated as transposons. Pseudomolecules and unassigned BACs contain a total of 44,124 gene loci, 177,271 retroelement-related regions and 26,487 DNA transposons, and non-redundant Illumina assemblies contribute an additional 18,264 genes, 75,777 retrotransposon regions and 8,476 DNA transposons (Supplementary Tables 5–9) along with 1,418 organellar insertions (Supplementary Data 1). For pseudomolecules and unassigned BACs, this translates to 16.8 genes, 67.6 retrotransposons and 10.1 DNA transposons per 100 kilobases (kb). Within Illumina sequence assemblies, gene density (17.1 per 100 kb) and retrotransposon density (72.2 per 100 kb) are similar to pseudomolecules and unassigned BACs, whereas DNA transposon density is lower (8.2 per 100 kb). Similarities in gene and transposon densities between BAC and Illumina sequences support the assertion that the Illumina sequence is euchromatic, although the possibility that some Illumina assemblies come from low-copy regions within heterochromatin can not be excluded. Considering only the 47,845 genes with experimental or database support (Supplementary Table 4), the average *M. truncatula* gene is 2,211 bp in length, contains 4.0 exons, and has a coding sequence of 1,001 bp. These values are similar to those observed previously in *Arabidopsis thaliana* (2,174 bp), *Oryza sativa* (3,403 bp) and *Populus trichocarpa* (2,301 bp)[4–6].

Recent analyses of plant genomes indicate a shared whole-genome hexaploidy (WGH) preceding the rosid–asterid split at 140–150 Myr ago[7]. Duplication patterns and genomic comparisons strongly suggest an additional WGD approximately 58 Myr ago in the papilionoids[8,9]. Near the time of this WGD, papilionoids radiated into several clades, the largest of which split quickly into two subclades, the Hologalegina (including *M. truncatula* and *L. japonicus*) and the milletioids (including *G. max* and other phaseoloids) at about 54 Myr ago[2]. We therefore compared *M. truncatula* pseudomolecules with other sequenced plant genomes to learn more about shared synteny and genome duplication history.

There is significant macrosynteny among *M. truncatula*, *L. japonicus* and *G. max* (Fig. 1 and Supplementary Fig. 5a, b). Conserved blocks, sometimes as large as chromosome arms, span most euchromatin in all three genomes. A given *M. truncatula* region is typically syntenic with one other *M. truncatula* region as a result of the approximately 58-Myr-ago WGD, usually in small blocks showing degraded synteny (Fig. 2 and Supplementary Fig. 6). A given *M. truncatula* region is most similar to two *G. max* regions via speciation at about 54 Myr ago and the *Glycine* WGD at <13 Myr ago[10] and less similar to two other *G. max* regions resulting from the ~58-Myr-ago and <13-Myr-ago WGD events. A *M. truncatula* region is likewise most similar to one *L. japonicus* region via speciation at about 50 Myr ago and less similar to a second *L. japonicus* region as a result of the ~58-Myr-ago WGD. Finally, each *M. truncatula* region and its homeologue typically show similarity to three *Vitis vinifera* regions via the pre-rosid WGH. Exceptions to these patterns could be due to gene losses, gains, or rearrangements specific to the *M. truncatula* lineage, resulting in synteny being more evident between *M. truncatula* and other genomes than in self-comparisons. Indeed, self-comparisons within *M. truncatula* reveal few remnants of the legume-specific WGD (Fig. 2 and Supplementary Fig. 6). Whereas this seems paradoxical, it is probably explained by extensive gene fractionation between WGD-derived homeologues in *M. truncatula*. In Fig. 3, two short regions on Mt1 and Mt3 resulting from the ~58-Myr-ago WGD are displayed beside microsyntenic regions of *G. max* and *V. vinifera*. As expected, many genes are microsyntenic between *M. truncatula* and *G. max* (ranging from 7/19 between Mt3 and Gm14 to 10/20 between Mt1 and Gm17).

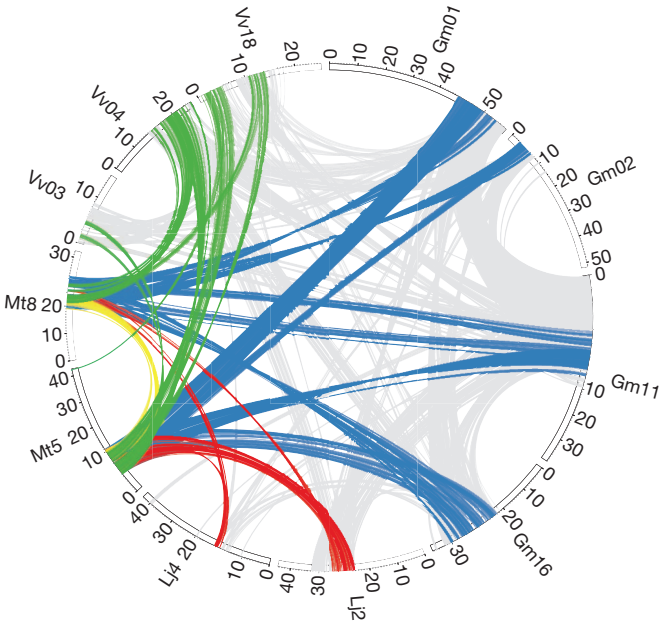

Figure 1 | Circos diagram illustrating syntenic relationships between *Medicago*, *Glycine*, *Lotus* and *Vitis*. Homologous gene pairs were identified for all pairwise comparisons between *M. truncatula*, *G. max*, *L. japonicus* and *V. vinifera* genomes. Syntenic regions associated with the ancestral WGD events were identified by visually inspection of corresponding dot-plots. The large Mt5–Mt8 synteny block (yellow) was found to have two syntenic regions in *L. japonicus* (red), four syntenic regions in *G. max* (blue) and three in *V. vinifera* (green).

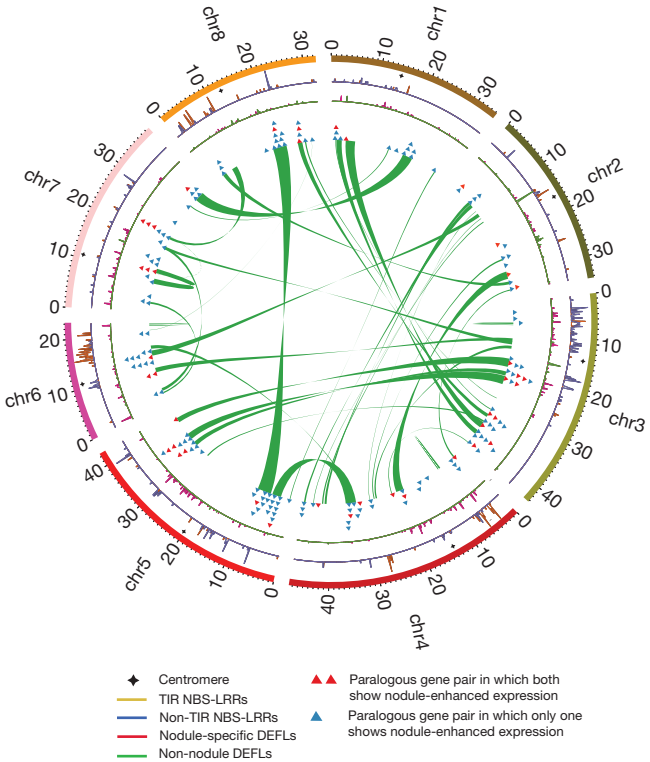

Figure 2 | Circos diagram illustrating the *Medicago* WGD and selected gene families. The 963 WGD-derived paralogous gene pairs were examined for overlap with the nodule-enhanced gene list (Supplementary Data 2). Resulting gene pairs were joined and plotted as either blue triangles (only one of the duplicates is nodule-enhanced) or red (both nodule enhanced). Gene densities of NBS-LRRs, NCRs and other defensin-like proteins are plotted against chromosome position. Density was calculated using a sliding window (100-kb window with 50-kb steps).

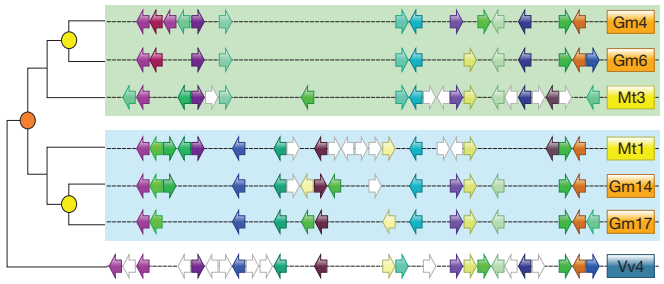

Figure 3 | Microsynteny comparison between *Medicago* homeologues and corresponding regions of *Glycine* and *Vitis*. Microsyntenic genome segments are centred around Medtr3g104510/Medtr1g015890 (Supplementary Table 10), a duplicated region derived from the ~58-Myr-ago WGD event noted in orange. The <13-Myr-ago *G. max*-specific WGD is coloured yellow. Orthologous/paralogous gene pairs are indicated through use of a common colour. White arrows represent genes with no syntenic homologue(s) in this genome region. Some of these genes may actually have a syntenic sequence in soybean but no corresponding model reported in the current annotation (http://www.phytozome.net/soybean).

Between the two *M. truncatula* homeologues, however, only 6 out of 33 genes (or collapsed gene families) are microsyntenic, with a homeologue missing from one or the other duplicate (Supplementary Table 10). Apparently, there have been many more changes, large and small, in *M. truncatula* than in *G. max* since the legume WGD. This is borne out by the fact that synteny blocks in *M. truncatula* are one-third the length of those remaining from the papilionoid WGD in *G. max* (524 kb against 1,503 kb) with the average number of homologous gene pairs per block correspondingly lower (12.4 against 31.0).

The *M. truncatula* genome also has undergone high rates of local gene duplication. The ratio of related genes within local clusters compared to all genes in families is 0.339 in *M. truncatula*, 3.1-fold higher than in *G. max* and 1.6-fold higher than in *A. thaliana* or *P. trichocarpa*. ('Local clusters' are defined as genes in a family all within 100 gene models of one another.) The excess of local gene duplications in *M. truncatula* is observed genome-wide and affects many families. There are 2.63 times as many gene families with local duplications in *M. truncatula* compared with *G. max* (2,980 against 1,131), an excess that also is seen in detailed comparisons of syntenic regions in *M. truncatula* and *G. max*. We examined 16.3 Mb of Mt05 showing synteny to two large regions of Gm01 plus homeologous blocks on Gm02, Gm09 and Gm11. In these regions, 25.8% of *M. truncatula* genes are locally duplicated compared with just 8.0% in *G. max*. Local gene duplications and losses have contributed both to synteny disruptions (Fig. 3 and Supplementary Fig. 7) and to high gene count (62,388) in *M. truncatula*—a value nearly as high as the 65,781 total gene models in *G. max* despite its additional (<13 Myr ago) WGD. Local gene duplications are evident in certain gene families, such as F-box genes, which have undergone pronounced expansions (Supplementary Fig. 8 and Supplementary Table 11). *M. truncatula* also has experienced higher rates of base substitution compared to other plant genomes (Supplementary Fig. 9). Assuming 58 Myr ago as the date of the legume WGD, then the rate of synonymous substitutions per site per year in *M. truncatula* is 1.08×10^{-8}, 1.8 times faster than estimates in other vascular plants[11]. Higher rates of mutation and greater levels of rearrangement in *M. truncatula* following the papilionoid duplication may have been driven by factors including short generation times, high selfing rates or small effective population sizes, although these characteristics are not unique to *M. truncatula*.

Legumes and actinorhizal species are capable of forming a specialized organ, the root nodule, a highly differentiated structure hosting nitrogen-fixing symbionts. Phylogenetic studies suggest that nodulation may have evolved multiple times in the Fabidae, but the observation that all nodulating species are contained within this single clade indicates that a predisposition to nodulate evolved in their common ancestor[12]. It is unknown whether nodulation with rhizobia preceded the divergence of the three legume subfamilies or evolved on multiple occassions[13]. Nevertheless, rhizobial nodulation and the 58-Myr-ago WGD are features common to most papilionoid legumes and both occurred early in the emergence of the group[2]. Given that WGDs generate genetic redundancy that potentially facilitates the emergence of novel gene functions without compromising existing ones[14], we examined the *M. truncatula* genome to ask whether the 58-Myr-ago WGD might have had a role in the evolution of rhizobial nodulation in *M. truncatula* and its relatives.

Nod factors are bacterial signalling molecules that initiate nodulation. Previous studies have shown that several of the plant components involved in the response to Nod factors also function in mycorrhizal signalling[15]. However, some Nod factor receptors and transcription factors have distinctly nodulation-specific functions. Among these nodulation-specific components, we found that the Nod factor receptor, *NFP*, and the transcription factor, *ERN1*, each have paralogues, *LYR1* and *ERN2* respectively, that trace back to the papilionoid WGD based on genome location and synonymous substitution rate values (Supplementary Fig. 10 and Supplementary Data 2). Both sets of gene pairs also show contrasting expression patterns and functional specialization. *NFP* and *ERN1* are expressed predominantly in the nodule and are known to function in nodulation[16,17], whereas *LYR1* and *ERN2* are highly expressed during mycorrhizal colonization (Supplementary Fig. 11). These observations indicate that two important nodulation-specific signalling components in *M. truncatula* might have evolved from more ancient genes originally functioning in mycorrhizal signalling and then duplicated by the 58-Myr-ago WGD. In the case of *M. truncatula NFP*/*LYR1*, this conclusion is supported by the observation that the apparent orthologue of *NFP* in the nodulating non-legume *Parasponia andersonii* functions in both nodule and mycorrhizal signalling[18]. Thus, the 58-Myr-ago WGD seems to have led to sub-functionalization of an ancestral gene participating in both interactions, resulting in two homeologous genes that each performs just one of the original functions.

To assess further the contribution of the WGD to *M. truncatula* nodulation, we analysed expression of paralogous gene pairs using RNA-seq data from six different organs (Supplementary Methods 5.1). A total of 963 WGD-derived gene pairs were found (Supplementary Data 2) with 618 pairs (1,046 genes) having RNA-seq data for one or both homeologue. We then determined the number of genes showing organ-enhanced expression (defined as genes with expression level in a single organ at least twice the level in any other) within the pseudomolecule and the WGD-derived gene sets (Supplementary Table 12). In both cases, different organs contained markedly different numbers of genes with enhanced expression (χ^2 with 5 degrees of freedom, $P = 10^{-272}$); however, the rank order among the organs was identical. Roots had the largest number of genes with enhanced expression followed by flower, nodule, leaf, seed/pod and bud. Among gene pairs with nodule-enhanced expression, both paralogues were nodule-enhanced in eight pairs, whereas just a single paralogue was nodule-enhanced in the other 43 pairs. This is consistent with nodulation pre-dating the WGD and further sub- and neo-functionalization emerging afterwards. We went on to examine transcription factors because they can act as regulators of plant growth and development. A total of 3,692 putative TF genes were discovered (Supplementary Data 3), representing 5.9% of all *M. truncatula* gene models (Supplementary Table 13). Of the 1,513 TF genes on pseudomolecules with RNA-seq data, 142 genes (9.4%) derived from the 58-Myr-ago WGD (Supplementary Fig. 12 and Supplementary Data 4), consistent with previous observations indicating greater retention of transcription factors following polyploidy[19]. Nodule-enhanced expression was significantly higher among transcription factors (92 out of 1,513 or 6.1%) than among all pseudomolecule genes (1,111 out of 23,478 or 4.7%) (χ^2 with 1 degree of freedom, $P = 0.024$) (Supplementary Table 12).

Nodule-enhanced expression was even higher in WGD-derived transcription factors (11 out of 142 or 7.7%), although this enrichment did not reach statistical significance ($P = 0.113$). As expected, *ERN1* is found within this group of WGD-retained, nodule-enhanced transcription factors.

These results show that many paralogous genes retained from the 58-Myr-ago WGD, especially signalling components and regulators, have undergone sub- or neo-functionalization, including several with specialized roles in nodulation. Nevertheless, separate phylogenetic analyses (Supplementary Methods 5.5) indicate that some nodule-related genes derive from the more ancient pre-rosid WGH, with their nodule-related functions pre-dating the 58-Myr-ago WGD (Supplementary Data 5). Taken together, these results are consistent with a model where the capacity for primitive interaction with new symbionts derived from existing mycorrhizal machinery involving genes recruited from the pre-rosid WGH. This capacity would have arisen early in the Fabidae clade and led to the appearance of nodulation in multiple lineages[13,20]. Later, the 58-Myr-ago WGD would have resulted in additional genes, including *NFP*, *ERN1* and the transcription factors described above, that went on to become specialized for nodule-related functions in the Papilionoideae.

Medicago contains additional amplified gene families, many nodulation-related and found in tandem clusters. *M. truncatula* has nine symbiotic leghaemoglobins, more than twice the number in *L. japonicus* or *G. max* (Supplementary Fig. 13). Five of these genes are located in a tight cluster on Mt5. The *M. truncatula* genome contains 593 nodule cysteine-rich peptides (NCRs) (Supplementary Data 6), a gene family restricted to *M. truncatula* and its relatives[21]. NCRs are noteworthy because they include members essential for terminal differentiation of rhizobia[22]. NCRs are tightly clustered within the *M. truncatula* genome (Fig. 2), with 75% found in clusters of up to 11 members. The *M. truncatula* genome also has 764 nucleotide-binding site and leucine-rich repeat (NBS-LRR) genes (Supplementary Data 7), more than other plant genomes that have been sequenced so far[23–25], many with nodule-specific expression (Supplementary Fig. 14). Almost 90% of NBS-LRRs occur in clusters and genome regions showing limited macrosynteny to other species, such as Mt3 and Mt6, are locations of large NBS-LRR superclusters (Fig. 2 and Supplementary Tables 14 and 15). Finally, *M. truncatula* secretes flavonoid signalling molecules to induce the *nod* genes of *Sinorhizobium meliloti*[26]. In *M. truncatula*, the corresponding biosynthetic pathway has expanded markedly, with 28 *M. truncatula* chalcone synthase genes in clusters of up to seven members compared to just four chalcone synthases in *A. thaliana*[27] (Supplementary Data 8). *M. truncatula* has ten chalcone reductases compared to none in *A. thaliana*[28] and *M. truncatula* has 11 chalcone isomerase genes, including one cluster of seven members, compared to just one representative in *A. thaliana*[29] (Supplementary Figs 15 and 16).

Analysis of the *M. truncatula* genome supports earlier studies indicating that the dramatic radiation of the legume family (at least the papilionoid subfamily) is partly attributed to the 58-Myr-ago WGD[30]. Our results indicate that the WGD early in papilionoid evolution allowed the emergence of critical components in Nod factor signalling and contributed to the complexity of rhizobial nodulation observed in this clade. As such, the WGD seems to have had a crucial role in the success of papilionoid legumes, enhancing their utility to humans.

METHODS SUMMARY

DNA sequencing. Six A17 BAC and one fosmid library were used to create Mt3.5 (Supplementary Table 1). Most were processed by Sanger paired-end sequencing of 3–6-kb shotgun libraries. Sequences were downloaded in February/March 2009 with scaffolding performed by aligning all BAC and fosmid ends against contigs and then anchored and ordered primarily by optical mapping. Separately, 25 billion base pairs (Gb) of Illumina sequence was generated using short (375 nt) inserts plus 2.1 Gb from a 5 kb mate-pair library, then assembled using CLCbio (http://www.clcbio.com) and Soap (http://soap.genomics.org.cn/).

RNA sequencing. Five tissues were used for RNA-seq analysis with ~10 million Illumina 36-bp reads per library (Supplementary Table 12). Three tissues were used for small RNA analysis with ~3 million reads per Illumina library (Supplementary Figs 17–18, Supplementary Table 16 and Supplementary Data 9).

Received 13 June; accepted 13 October 2011.
Published online 16 November 2011.

1. Wang, H. *et al.* Rosid radiation and the rapid rise of angiosperm-dominated forests. *Proc. Natl Acad. Sci. USA* **106,** 3853–3858 (2009).
2. Lavin, M., Herendeen, P. S. & Wojciechowski, M. F. Evolutionary rates analysis of Leguminosae implicates a rapid diversification of lineages during the tertiary. *Syst. Biol.* **54,** 575–594 (2005).
3. Kulikova, O. *et al.* Integration of the FISH pachytene and genetic maps of *Medicago truncatula*. *Plant J.* **27,** 49–58 (2001).
4. The Arabidopsis Genome Initiative. I. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408,** 796–815 (2000).
5. International Rice Genome Sequencing Project. The map-based sequence of the rice genome. *Nature* **436,** 793–800 (2005).
6. Tuskan, G. A. *et al.* The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* **313,** 1596–1604 (2006).
7. Tang, H. *et al.* Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. *Genome Res.* **18,** 1944–1954 (2008).
8. Pfeil, B. E., Schlueter, J. A., Shoemaker, R. C. & Doyle, J. J. Placing paleopolyploidy in relation to taxon divergence: a phylogenetic analysis in legumes using 39 gene families. *Syst. Biol.* **54,** 441–454 (2005).
9. Cannon, S. B. *et al.* Polyploidy did not predate the evolution of nodulation in all legumes. *PLoS ONE* **5,** e11630 (2010).
10. Schmutz, J. *et al.* Genome sequence of the palaeopolyploid soybean. *Nature* **463,** 178–183 (2010).
11. Lynch, M. & Conery, J. S. The evolutionary fate and consequences of duplicate genes. *Science* **290,** 1151–1155 (2000).
12. Soltis, D. E. *et al.* Chloroplast gene sequence data suggest a single origin of the predisposition for symbiotic nitrogen fixation in angiosperms. *Proc. Natl Acad. Sci. USA* **92,** 2647–2651 (1995).
13. Doyle, J. J. & Luckow, M. A. The rest of the iceberg. Legume diversity and evolution in a phylogenetic context. *Plant Physiol.* **131,** 900–910 (2003).
14. Freeling, M. & Thomas, B. C. Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. *Genome Res.* **16,** 805–814 (2006).
15. Oldroyd, G. E. & Downie, J. A. Coordinating nodule morphogenesis with rhizobial infection in legumes. *Annu. Rev. Plant Biol.* **59,** 519–546 (2008).
16. Arrighi, J. F. *et al.* The *Medicago truncatula* lysine motif-receptor-like kinase gene family includes *NFP* and new nodule-expressed genes. *Plant Physiol.* **142,** 265–279 (2006).
17. Middleton, P. H. *et al.* An ERF transcription factor in *Medicago truncatula* that is essential for Nod factor signal transduction. *Plant Cell* **19,** 1221–1234 (2007).
18. Op den Camp, R. *et al. LysM*-type mycorrhizal receptor recruited for rhizobium symbiosis in nonlegume *Parasponia*. *Science* **331,** 909–912 (2011).
19. Thomas, B. C., Pedersen, B. & Freeling, M. 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 (2006).
20. Kistner, C. & Parniske, M. Evolution of signal transduction in intracellular symbiosis. *Trends Plant Sci.* **7,** 511–518 (2002).
21. Kato, T. *et al.* Expression of genes encoding late nodulins characterized by a putative signal peptide and conserved cysteine residues is reduced in ineffective pea nodules. *Mol. Plant Microbe Interact.* **15,** 129–137 (2002).
22. Van de Velde, W. *et al.* Plant peptides govern terminal differentiation of bacteria in symbiosis. *Science* **327,** 1122–1126 (2010).
23. Meyers, B. C., Kozik, A., Griego, A., Kuang, H. & Michelmore, R. W. Genome-wide analysis of NBS-LRR-encoding genes in *Arabidopsis*. *Plant Cell* **15,** 809–834 (2003).
24. Yang, S., Zhang, X., Yue, J. X., Tian, D. & Chen, J. Q. Recent duplications dominate NBS-encoding gene expansion in two woody species. *Mol. Genet. Genomics* **280,** 187–198 (2008).
25. Zhou, T. *et al.* Genome-wide identification of NBS genes in *japonica* rice reveals significant expansion of divergent non-TIR NBS-LRR genes. *Mol. Genet. Genomics* **271,** 402–415 (2004).
26. Peters, N. K., Frost, J. W. & Long, S. R. A plant flavone, luteolin, induces expression of *Rhizobium meliloti* nodulation genes. *Science* **233,** 977–980 (1986).
27. Winkel-Shirley, B. Flavonoid biosynthesis. A colorful model for genetics, biochemistry, cell biology, and biotechnology. *Plant Physiol.* **126,** 485–493 (2001).
28. Hegnauer, R. Relevance of seed polysaccharides and flavonoids for the classification of the leguminosae: a chemotaxonomic approach. *Phytochemistry* **34,** 3–16 (1993).
29. Shirley, B. W. *et al.* Analysis of *Arabidopsis* mutants deficient in flavonoid biosynthesis. *Plant J.* **8,** 659–671 (1995).
30. Singer, S. R. *et al.* Venturing beyond beans and peas: what can we learn from *Chamaecrista*? *Plant Physiol.* **151,** 1041–1047 (2009).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements Funding support to N.D.Y., C.D.T. and B.A.R. from The Noble Foundation and NSF-PGRP 0321460, 0604966; to N.D.Y., J.M. and G.D.M. from NSF-PGRP 0820005; to C.D.T. from NSF-PGRP 0821966; to F.D., G.E.D.O., R.G., K.F.X.M., T.B., J. Denarié, F.Q. and J.R. from FP6 EU project GLIP/Grain Legumes

FOOD-CT-2004-506223; to G.E.D.O. and J.R. from BBSRC BBS/B/11524; to F.D. and F.Q. from ANR project SEQMEDIC 2006-01122; to R.G. from the Dutch Science Organization VIDI 864.06.007, ERA-PG FP-06.038A; to Y.V.d.P. from the Belgian Federal Science Policy Office IUAP P6/25, Fund for Scientific Research Flanders, Institute for the Promotion of Innovation by Science and Technology in Flanders and Ghent University (MRP N2N); to D.R.C. from NSF IOS-0531408, IOS-0605251; to D.J.S., B.C.M. and P.J.G. from USDA CSREES 2006-03567; to J. Gouzy from 'Laboratoire d'Excellence' (LABEX) TULIP (ANR-10-LABX-41). We also acknowledge technical support from the University of Minnesota Supercomputer Institute and thank Y. W. Nam for a BamHI BAC library used by Genoscope, S. Park and M. Accerbi for RNA isolation, T. Paape for statistical consulting, and M. Harrison for supplying *myc* infected and control root tissues used to make small RNA libraries.

Author Contributions Planning, coordination and writing: N.D.Y., J. Doyle, F.Q., J. Weissenbach, P.W., K.F.X.M., C.D.T., G.E.D.O., G.D.M., J. Mudge, E.F.R., R.A.D., M.K.U., F.D., J. Denarié, D.R.C., P.J.G., B.C.M., D.J.S., C.R.P., B.A.R., D.C.S., S.B.C., Y.V.d.P., R.G., T.B., J.R., S.R.S.; BAC libraries: B.S., A. Bellec, H.B., J. Gish, D.-J.K.; Mapping and assembly: V.B., N.C., S.F., G.M., S. Samain, E.L.M., F.P., N.S., O.S., A.Z., C.G., J.-H. Mun, R.D., M.B., S.Z., C.L., M.H., C.F., C. Nicholson, C.R.; sequencing: A. Berger, J.P., A.V., D.-H.J., S.D., Y.J., H.L., S.L.M., F.Z.N., B.Q., C.Q., M.O., I.S., R.S., K.W., D.D.W., G.B.W., Y.X., L.Y., Z.Y., F.Y., L.Z., S.J.H., L.M., S. Sims; annotation and bioinformatics: A.C., C.S., H.G., M. Spannagl, C. Noirot, T.S., A.J.S., S.B., F.C., V.K., J. McCorrison, H.T., A. Hallab, A.J., K.K., J. Warfsmann, A.K.B., A.D.F., V.A.B., J.D.M., M.A.N., S. Sinharoy, P.X.Z., P.B., A.-M.D., J. Gouzy, E.S., H.S., B.R., A.J.G., J.Z., B.-B.W., X.W., P.Z., K.A.T.S., A. Hua, S.M.K., S.L., J.D.W., S.G., S.P., S.R., L.S., S.D.M., M.W.

Author Information *Medicago truncatula* pseudomolecules are found at DDBJ/EMBL/GenBank as accession numbers CM001217–CM001224 and unanchored BACs as GL982851–GL982996. Illumina genome sequences are in the Short Read Archive under SRS150378, RNA-seq sequences under SRP008485, and small RNA sequences in GEO under GSM769273, GSM769274 and GSM769276. Pseudomolecule annotation and Illumina assemblies are available at ftp://ftp.jcvi.org/pub/data/m_truncatula/Mt3.5/. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to N.D.Y. (neviny@umn.edu).