Cite as: M. Griesmann *et al.*, *Science*
10.1126/science.aat1743 (2018).

Phylogenomics reveals multiple losses of nitrogen-fixing root nodule symbiosis

Maximilian Griesmann,[1,2] Yue Chang,[3,4] Xin Liu,[3,4] Yue Song,[3,4] Georg Haberer,[2] Matthew B. Crook,[5] Benjamin Billault-Penneteau,[1] Dominique Lauressergues,[6] Jean Keller,[6] Leandro Imanishi,[7] Yuda Purwana Roswanjaya,[8] Wouter Kohlen,[8] Petar Pujic,[9] Kai Battenberg,[10] Nicole Alloisio,[9] Yuhu Liang,[3,4] Henk Hilhorst,[11] Marco G. Salgado,[12] Valerie Hocher,[13] Hassen Gherbi,[13] Sergio Svistoonoff,[13] Jeff J. Doyle,[14] Shixu He,[3,4] Yan Xu,[3,4] Shanyun Xu,[3,4] Jing Qu,[3,4] Qiang Gao,[3,15] Xiaodong Fang,[3,15] Yuan Fu,[3,4] Philippe Normand,[9] Alison M. Berry,[10] Luis G. Wall,[7] Jean-Michel Ané,[16,17] Katharina Pawlowski,[12] Xun Xu,[3,4] Huanming Yang,[3,18] Manuel Spannagl,[2] Klaus F. X. Mayer,[2,19] Gane Ka-Shu Wong,[3,20,21] Martin Parniske,[1]* Pierre-Marc Delaux,[6]* Shifeng Cheng[3,4]*

[1]Faculty of Biology, Genetics, LMU Munich, Großhaderner Strasse 2-4, 82152 Martinsried, Germany. [2]Plant Genome and Systems Biology, Helmholtz Center Munich–German Research Center for Environmental Health, 85764 Neuherberg, Germany. [3]BGI-Shenzhen, Shenzhen 518083, China. [4]China National GeneBank, BGI-Shenzhen, Shenzhen 518120, China. [5]Department of Microbiology, Weber State University, Ogden, UT 84408-2506, USA. [6]Laboratoire de Recherche en Sciences Végétales, Université de Toulouse, CNRS, UPS, 24 chemin de Borde Rouge, Auzeville, BP42617, 31326 Castanet Tolosan, France. [7]LBMIBS, Departamento de Ciencia y Tecnología, Universidad Nacional de Quilmes, CONICET, R. Saénz Peña 352, B1876BXD Bernal, Argentina. [8]Laboratory for Molecular Biology, Wageningen University, Droevendaalsesteeg 1, 6708 PB Wageningen, Netherlands. [9]Université Lyon 1, Université de Lyon, CNRS, Ecologie Microbienne, UMR 5557, Villeurbanne, 69622 Cedex, France. [10]Department of Plant Sciences, University of California, Davis, Davis, CA 95616, USA. [11]Laboratory for Plant Physiology, Wageningen University, Droevendaalsesteeg 1, 6708 PB Wageningen, Netherlands. [12]Department of Ecology, Environment and Plant Sciences, Stockholm University, 106 91 Stockholm, Sweden. [13]French National Research Institute for Sustainable Development (IRD), UMR LSTM (IRD–INRA–CIRAD–Université Montpellier–Supagro), Campus International de Baillarguet, TA A-82/J, 34398, Montpellier Cedex 5, France. [14]Section of Plant Breeding and Genetics, School of Integrated Plant Science, Cornell University, Ithaca, NY 14853, USA. [15]BGI Genomics, BGI-Shenzhen, Shenzhen 518083, China. [16]Department of Agronomy, University of Wisconsin–Madison, WI 53706, USA. [17]Department of Bacteriology, University of Wisconsin–Madison, WI 53706, USA. [18]James D. Watson Institute of Genome Sciences, Hangzhou 310058, China. [19]TUM School of Life Sciences Weihenstephan, Technical University of Munich, 85354 Freising, Germany. [20]Department of Biological Sciences, University of Alberta, Edmonton, AB T6G 2E9, Canada. [21]Department of Medicine, University of Alberta, Edmonton, AB T6G 2E, Canada.

***Corresponding author. Email: chengshf@genomics.cn (S.C.); pierre-marc.delaux@lrsv.ups-tlse.fr (P.-M.D.); parniske@lmu.de (M.P.)**

The root nodule symbiosis of plants with nitrogen-fixing bacteria impacts global nitrogen cycles and food production but is restricted to a subset of genera within a single clade of flowering plants. To explore the genetic basis for this scattered occurrence, we sequenced the genomes of ten plant species covering the diversity of nodule morphotypes, bacterial symbionts and infection strategies. In a genome-wide comparative analysis of a total of 37 plant species, we discovered signatures of multiple independent loss-of-function events in the indispensable symbiotic regulator *NODULE INCEPTION* (*NIN*) in ten out of 13 genomes of non-nodulating species within this clade. The discovery that multiple independent losses shaped the present day distribution of nitrogen-fixing root nodule symbiosis in plants reveals a phylogenetically wider distribution in evolutionary history and a so far underestimated selection pressure against this symbiosis.

Nitrogen is one of the main requirements for plant growth. Members of the legume family (Fabaceae, order Fabales) and of nine additional plant families benefit from symbiosis with nitrogen-fixing bacteria, either phylogenetically diverse rhizobia or *Frankia*, which are hosted inside plant cells found within nodules–specialized host-derived lateral organs typically found on roots. In this mutualistic nitrogen-fixing root nodule (NFN) symbiosis, intracellular bacteria convert atmospheric nitrogen into ammonium using the enzyme nitrogenase (*1*). This "fixed nitrogen", delivered to the host plant, is an essential building block for amino acids, DNA, RNA, tetrapyrroles such as chlorophyll and many other molecules. This symbiosis enables plant survival under nitrogen-limiting conditions. In agriculture, this independence from chemical nitrogen-fertilizer reduces costs and fossil fuel consumption imposed by the Haber–Bosch process (*2*). Since the discovery of the NFN symbiosis, with rhizobia in 1888 and with *Frankia* in 1895 (*3*, *4*), it has been unclear why it is restricted to only a limited number of flowering plant species.

A major scientific step forward in our understanding was the reorganization of the phylogenetic tree of angiosperms in 1995, which revealed that plants forming the NFN symbiosis are restricted to the Fabales, Fagales, Cucurbitales, and Rosales that together form the NFN clade (*5*, *6*). However, this re-organization also opened new questions, because only ten out of the twenty-eight plant families within the NFN clade contain plants that form nodules (referred here as "nodulating species"), and these do not form a monophyletic group; moreover, within nine of these ten families most genera do not form NFN symbiosis (*7*). In addition to this scattered distribution, a further unsolved mystery that surrounds the evolution of NFN symbiosis is its diversity at multiple levels: Legumes (Fabales) and the non-legume *Parasponia* (Rosales) (*8*) form nodules with rhizobia whereas species of the actinobacterial genus *Frankia* infect actinorhizal plants from eight plant families of the orders Fagales, Cucurbitales and Rosales (*9*). A diversity of infection mechanisms has been described (*9*), and root nodule structures display wide variations (*5*, *7*). The most parsimonious hypothesis to explain the restricted and yet scattered distribution pattern of such diverse NFN symbioses predicted a genetic change in the ancestor of the NFN clade, a predisposition event, that enabled the subsequent independent evolution of NFN symbiosis specifically and exclusively in this clade (the *multiple origin* hypothesis), along with a number of losses (*5*, *7*, *10*–*12*). Recent quantitative phylogenetic modeling studies supported scenarios with independent gains and switches between the non-nodulating and nodulating states during the evolution of the NFN clade (*11*, *13*). However, none of these analyses provide direct evidence or the molecular causes of the specific gains and losses that explain the distribution of NFN symbiosis in extant genera.

Exploring the genetic basis underlying the evolutionary dynamics of the NFN symbiosis in plants will improve our understanding of the diversity of symbiotic associations observed in extant taxa and the ecosystems they inhabit, and potentially provide keys to engineer it in crops and to predict the stability of this trait over long evolutionary times. Here, we employed a genome-wide comparison including genomic and phylogenomic methods (*14*) to address the long-standing conundrum of the evolution of NFN symbiosis and to identify the underlying genetic players (*15*, *16*).

Results

Genome sequencing in the NFN clade

Sequenced genomes of nodulating species were only available for a few agriculturally relevant legume species belonging to a single subfamily (Papilionoideae), all derived from a single predicted evolutionary origin of NFN symbiosis, with no representation either of taxa representing possible additional origins within legumes or of non-legume nodulating species widely accepted as representing multiple additional origins (*17*–*19*). Conversely, sequenced genomes of non-nodulating species were restricted to the Fagales, Rosales and Cucurbitales, and did not include non-nodulating legume taxa (table S1). To overcome this sampling bias which restricted the phylogenomic analysis, we sequenced de novo the genomes of seven nodulating species belonging to the Cucurbitales, Fagales, and Rosales and the Caesalpinioideae subfamily of the Fabaceae, representing a possible second origin of NFN symbiosis in legumes. Three non-nodulating species from the NFN clade were also sequenced, notably *Nissolia schotii*, a papilionoid legume that has lost the ability to form the NFN symbiosis, and which therefore provides insights on the genomic consequences of losing the symbiosis (Fig. 1 and fig. S1). For each species, 144 to 381 Gb of Illumina reads were obtained covering the estimated genomes at least 189-fold and up to 1,113-fold, with the resulting scaffold N50 length between 96 kb and 1.18 Mb and an average genome completeness of 96% (Fig. 1, figs. S1 and S2, and tables S2 to S29). Altogether, the genomes of species sequenced here represent six families and most known nodule anatomy types and root infection pathways, include hosts for the main classes of nodule-inducing symbiotic nitrogen-fixing alpha-proteo-, beta-proteo- and actino-bacteria, and cover 6–7 independent evolutionary origins of NFN symbiosis according to the *multiple gains* hypothesis (*5*, *7*, *10*, *11*). These ten sequenced genomes, together with 18 other genomes from the NFN clade and nine genomes from other flowering plants as outgroup (Fig. 1), were compared to detect molecular traces supporting any of the three postulated events in the evolution of NFN symbiosis: i) predisposition to evolve it, ii) multiple independent gains and iii) multiple independent losses of NFN symbiosis.

The putative predisposition event did not involve NFN clade-specific gene gains

The predisposition event postulated by Soltis *et al.* in 1995 (*5*) may be based on the acquisition of one or several genes or sequence modifications specific for the NFN clade. This acquisition would be also consistent with a *single origin* hypothesis, in which NFN symbiosis in all taxa is predicted as a homologous trait. Genes acquired during the predisposition event are expected to be specific to the NFN clade and present in all nodulating species. To search for genes following this evolutionary pattern we identified gene families across all 37 plant genomes in our dataset using the Orthofinder pipeline (*20*). For each of the resulting 29,433 gene family clusters we calculated a separate phylogeny and subsequently inferred orthologs for all genes of the reference species, *Medicago truncatula* (*16*). We selected groups for which orthologs were absent in the nine species outside of the NFN clade, but were retained in nodulating species (fig. S3). To obtain a candidate set for manual validation from the total of 29,213 orthologous groups we employed an automated filtering approach with relaxed criteria, which allowed for the absence of orthologs in a small subset of nodulating species (*16*). This step was necessary to avoid the loss of putative candidates due to missed gene models resulting from false negatives as they often occur in automated gene prediction pipelines. Our relaxed filter identified a total of 31 orthologous candidate groups (table S30). All of these candidate groups underwent an iterative manual curation including a search for missed gene models and recalculation of phylogenies and orthologs (*16*).

Not a single candidate gene was identified matching the evolutionary pattern expected for predisposition-related genes, suggesting that genes gained in the most recent common ancestor of the NFN clade have been conserved or lost irrespectively of the symbiotic state of the lineages. If the predisposition indeed occurred, this result indicates that it did not involve the acquisition of novel genes but rather the cooption of existing genes and their corresponding pathways.

Gene family dynamics is compatible with multiple gains

Multiple molecular mechanisms leading to the convergent evolution of a trait have been identified (*21*). Deep-homology (*22*), the independent recruitment of a homologous gene set for the development of non-homologous traits, has been proposed for the NFN symbiosis (*7*, *23*). Indeed, several genes initially identified for their symbiotic role in legumes were later found to also play a symbiotic role in Fagales (*24*–*27*), Rosales (*28*, *29*), and Cucurbitales (*30*). An alternative mechanism for the evolution of novelty is gene family expansion, as exemplified by the parallel diversification of the Zn-finger transcription factor family in the evolution of a dominant yeast form in fungi (*31*), or the acquisition of strigolactone perception in parasitic plants (*32*). We hypothesized that if NFN symbiosis evolved multiple times, independent expansions in the same gene family might have been involved. Given that our dataset covers 6–7 of the predicted independent gains of NFN symbiosis it allowed us to search for gene families whose evolutionary patterns are consistent and would support the *multiple gains* hypothesis (*7*). We analyzed Orthofinder clusters for copy number variation to identify gene family expansion events at each of these nodes (Fig. 2) (*16*). Multiple alternative models have been proposed for the independent gain of NFN symbiosis. In one scenario (*7*) a single gain before the radiation of the legume family has been suggested that correlates with the expansion of 33 clusters in our analysis (Fig. 2). To test whether any of these clusters was enriched with differentially expressed genes (DEGs) in nodule tissue versus root tissue, we derived for *Medicago truncatula* (Medicago) gene expression data for both conditions from the gene expression atlas (*33*) and tested for an enrichment of Medicago DEGs in each cluster (table S31). We only found one such cluster that belongs to the nitrate transporter family NRT1/PTR (*34*). However, this gene family appears to be also expanded at nodes non-related to independent gains within and outside of the NFN clade (table S31). An alternative model proposes two independent gains within the legumes: the first at the most recent common ancestor of the papilionoids and the clade to which *Castanospermum australe* belongs (21/291 enriched/expanded clusters), or at the base of the papilionoids (4/32) and the second gain at the base of the caesalpinioid/mimosoid clade (7/92). This clade could alternatively comprise two independent gains for *Chamaecrista fasciculata* (39/710) and mimosoids (37/723). Larger numbers of expanded gene families were observed for the predicted events in the Rosales, in the Fagales and in the Cucurbitales (Fig. 2 and table S31). Taken together we found 52 gene family clusters that were enriched with differentially expressed genes in Medicago nodules and expanded multiple times at proposed independent gain nodes. However, similar to the NRT1/PTR family all of the enriched clusters also expanded outside the NFN clade. Inside the NFN clade these clusters expanded beside the hypothesized gain of NFN symbiosis nodes at many additional nodes (table S31). In our survey of all independent gene family expansions, we did not identify any cluster that displayed parallel expansions, one possibility among others that would indicate convergent recruitment for the independent gains of NFN symbiosis (*7*, *16*). If NFN symbiosis indeed evolved multiple times, our genome-wide analysis revealed hundreds of clade-specific candidate genes that, together with gene co-option, may have played a role in the putative independent evolutions of this trait.

Genomic evidence for multiple losses

Testing homologies of a trait shared by multiple taxa typically involves assessment of the trait in these species, and inferring origins of the trait once homology is accepted or rejected. However, information on the origin of a trait, and thus its homology, can also be obtained from taxa lacking the trait, by distinguishing primary absence (the taxon never had the trait) from secondary loss of the trait (*23*). It has been demonstrated that genes involved in an unique biological process are lost following the loss of this trait, a process known as gene co-elimination (*35*–*37*). To test the *multiple losses* hypothesis we searched the ortholog groups calculated above for an evolutionary pattern that retained orthologs in all nodulating species but lost them in non-nodulating ones. To filter and confirm the list of candidate orthologous groups we used the same two-step process combining an automated pipeline with relaxed criteria for nodulating species followed by a careful manual curation and evaluation step with stringent criteria (presence in all nodulating species required). The automated pipeline with relaxed criteria resulted in a list of 121 candidate groups (table S32). During our manual confirmation and refinement, we rejected 31 of these candidate groups because orthologs were absent from one or more nodulating species. Another 62 candidate groups were rejected because orthologs of more than 50% of non-nodulating species were present. A weak phylogenetic signal did not allow inferring a reliable orthology for 27 candidate groups. A single gene, *NIN*

(*NODULE INCEPTION*), was confirmed in the genomes of all the nodulating species and in the genome of species outside the NFN clade, and absent from most non-nodulating ones in the NFN clade (Fig. 2 and fig. S4). Forward genetic screens in the legumes *Lotus japonicus* (*27*), *Pisum sativum* (*38*) and *Medicago truncatula* (*24*) identified *NIN* as indispensable for the two developmental aspects of the NFN symbiosis: initiation of root nodule development and the formation of the plant structure facilitating intracellular uptake of bacteria (*27*). Furthermore, RNAi-based suppression of *NIN* expression in *Casuarina glauca* (Fagaceae, Fagales) impaired nodule formation (*25*) consistent with a conservation of the role of *NIN* in NFN symbiosis in actinorhizal host plants.

Confirmation of NIN *absence by microsynteny*

Besides its presence in all the nodulating species in our dataset, synteny analysis revealed the conservation of the syntenic blocks surrounding the *NIN* locus across the NFN clade (Fig. 2). In contrast, ten out of 13 non-nodulating species of the Fabales, Fagales, Cucurbitales, and Rosales underwent partial (four species) or complete (six species) deletions of *NIN* from the conserved genomic block (Fig. 2). The legume family is divided into six subfamilies, two of which include nodulating species (*39*). According to previous estimates (*7*), *Cercis* (a member of Cercidoideae, which is sister to most or all other legumes) and *Castanospermum* (a member of a clade sister to the crown group of papilionoid legumes, in which nearly all members form the NFN symbiosis) both represent lineages in which NFN symbiosis never occurred. In contrast, *Nissolia* and its sister genus *Chaetocalyx* are non-nodulating genera nested within the nodulating crown group of papilionoids, and thus have been predicted to represent secondary loss of NFN symbiosis (*11*, *40*). Our synteny approach allowed the discovery of the complete absence of *NIN* from the genome of *Nissolia schottii* (Fig. 2). In addition, we confirmed the absence of *NIN* in three *Chaetocalyx* species (fig. S5). Since *NIN* was present in the most recent common ancestor of the NFN clade and conserved in nodulating species, the absence of this gene in the *Nissolia–Chaetocalyx* lineage represents a loss that correlates with, and is sufficient to explain, the loss of NFN symbiosis. *Cercis canadensis* harbored only a *NIN* pseudogene remnant in the genomic block, while the genome of *Castanospermum australe* completely lacked *NIN* (Fig. 2). These results demonstrate three independent losses of *NIN* in the legume family. Similarly, the synteny analyses confirmed a minimum of three independent losses in non-nodulating Rosales and two in the Cucurbitales (Fig. 2). Together, the diversity of *NIN* deletions in the non-nodulating species is indicative of at least eight independent evolutionary events that led to the loss of *NIN* function (Fig. 3). Following the *multiple-gains* hypothesis (*7*), NFN symbiosis was predicted to have evolved independently at least six times in the species space sampled here (Fig. 2). Loss of *NIN* provides an alternative model with at least eight independent losses of NFN symbiosis (Fig. 3). Thus, the current distribution of NFN symbiosis might be a combination of these two complementary and not mutually exclusive models.

Loss of RPG

In parallel with these multiple confirmed losses of *NIN*, the synteny analysis also confirmed the presence of *NIN* in three non-nodulators of the NFN clade (Fig. 2). We hypothesized that the loss of genes other than *NIN* may explain the non-nodulating state of these species and that such genes may have been missed by the specific and stringent criteria for the detection of the presence-absence patterns. In addition to *NIN*, we determined the presence/absence pattern of 21 genes that were identified by forward and reverse genetics to be critical for NFN symbiosis in legumes (Fig. 3 and table S33). Among these genes, 20 were conserved in nodulating species and most non-nodulating ones (figs. S6 to S25). By contrast, *RHIZOBIUM-DIRECTED POLAR GROWTH* (*RPG*) (*41*), which is present outside the NFN clade is missing in *Nissolia schottii* and eleven other non-nodulating species from the four orders of the NFN clade (Fig. 3). These losses were confirmed by microsynteny analyses for all non-nodulating species (fig. S26). In the *M. truncatula rpg* mutant, infection threads are still present but their structure is abnormal (*41*) indicating that *RPG*, similar to *NIN,* is required for proper infection thread progression. However, in contrast to *nin* mutants, nodules are formed on *rpg* roots (*41*). Among the nodulating species, *RPG* is absent in the papilionoid *Arachis ipaensis* (Fig. 3), a genus in which rhizobia infect nodules intercellularly (*42*). Polymorphism in *RPG* may represent an intermediate step on an evolutionary path toward the loss of this symbiosis in *Arachis*, a genus in which NFN symbiosis is described as a labile trait (*43*, *44*). Absence of *RPG* in *Arachis* also explains why the genome-wide comparative phylogenomic approach did not identify this gene, given that the pipeline required the candidate genes to be present in all nodulating species. *RPG* was one of the genes rejected for not fulfilling this criterion. Among non-nodulating species, *Juglans regia* (Fagales), *Ziziphus jujuba* and *Prunus persica* (Rosales) have lost *RPG* but retained *NIN* suggesting that additional mutations might be causative for the loss of NFN symbiosis in these species. Some of the candidate mutation targets, for example LysM receptors involved in the perception of symbiotic signals produced by nitrogen-fixing nodulating rhizobia, will be difficult to identify by phylogenomic analysis, because of the rapid evolution and expansionary dynamics of these gene families (*45*). In contrast to other symbiosis-relevant genes involved in infection, *NIN* and *RPG* are only known to have NFN symbiosis-specific functions while the mutation of other genes may have more pleiotropic effects. For example, the key signaling components *SYMRK*, *CCaMK* and *CYCLOPS* are involved in both NFN symbiosis and arbuscular mycorrhizal symbiosis, the most widespread symbiosis in land plants (*46*). Mutations in any of these three genes would also affect arbuscular mycorrhizal symbiosis. Illustrating this dual selection pressure, retention of these genes has been described in the genus *Lupinus* that lost the arbuscular mycorrhizal symbiosis but retained NFN symbiosis (*37*, *47*, *48*). Given that both *NIN* and *RPG* are present in species outside the NFN clade (Fig. 3 and figs. S4 and S6), their consistent losses in non-nodulating species suggest the shift in constraints on sequence evolution specifically in the NFN clade. While the ancestral function of both genes remains unknown, such relaxation might be mirrored by signature of relaxed or positive selection on both genes at the base of the NFN clade. We investigated the selective pressure acting on the NFN clade for these two genes using the PAML package (*49*). Results did not reveal a significant positive or relaxed selection occurring in the NFN clade that would have reflected a putative neo functionalization for NFN symbiosis-specific functions (table S34).

Discussion

In recent decades, the favored model to explain the scattered occurrence of NFN symbiosis in flowering plants predicted a single predisposition event at the base of the NFN clade followed by up to sixteen origins, even though the occurrence of multiple losses was never excluded (*7*, *11*).

Our genome-wide comparative analysis did not detect gene gains specific to the NFN clade and maintained in all nodulating species. Such genes would have been ideal candidates for either the predisposition event (in the *multiple gains* hypothesis) or the evolution of NFN symbiosis itself in the hypothesis that NFN symbiosis evolved only once in the most recent common ancestor of the NFN clade (the *single gain* hypothesis). This indicates that this step involved either fast evolving genes that were not captured by our phylogenomics pipeline or more subtle genetic changes. Evolutionary developmental genetics in plant, fungal and animal systems have revealed that even more than gains of genes, novelty often arises from the rewiring of existing gene networks via gains or losses of *cis* regulatory elements leading to the co-option of ancestral genes (*50*). A similar mechanism may have acted in the most recent common ancestor of the NFN clade.

Co-option of different, or homologous in the case of deep-homology, genetic components may lead to the convergent evolutions of non-homologous traits in multiple species (*50*). Deep-homology has been invoked for the evolution of NFN symbiosis, either during the predisposition or the following putative multiple gains (*7*). Our results support this hypothesis given that all the genes characterized for their involvement in NFN symbiosis in legumes were already present in the most recent common ancestor of the NFN clade and that we did not detect genes specific to the NFN clade that were conserved in all nodulating species (Figs. 2 and 3). For the putative multiple gains of NFN symbiosis, it cannot be excluded that gene gains were also involved in addition to co-option of ancestral pathways. We identified hundreds of such lineage-specific candidate genes (Fig. 2). However, considering that most of the predicted gains in our analysis are located in terminal taxa, that are known to accumulate orphan genes and species-specific duplications in comparative approaches, it can be anticipated that only a subset of them participated in the evolution of NFN symbiosis (in the *multiple gains* hypothesis) or in lineage-specific refinements of the trait (in the *single gain* hypothesis).

Our results validate another hypothesis: multiple independent losses of NFN symbiosis in the four orders of the NFN clade. In a classical model of evolution, if the number of losses necessary to explain the distribution of a trait in a given clade outnumbers the predicted gains, multiple gains will be favored over multiples losses to explain the distribution of the trait. In legumes, up to six gains of NFN symbiosis were predicted (*7*) while the clear losses that we identified in *Cercis canadensis*, *Castanospermum australe,* and *Nissolia schotii* now argue for a single origin before the radiation of the family (Fig. 3). Beyond legumes, the number of validated losses of NFN symbiosis is consistent with a single gain of this symbiosis in the most recent common ancestor of the NFN clade, even though it does not reject the possible occurrence of multiple gains. The recent identification of loss of NFN symbiosis in the Rosales *Trema orientalis* brings further support to this hypothesis (*51*). Besides NFN symbiosis, the single or multiple origin(s) of other traits, such as the evolution of complex multicellularity in fungi, are currently debated with the accumulating evidence of multiple-losses demonstrated by the loss of associated essential genes (*52*). Thus multiple gains and multiple losses are not mutually exclusive scenarios to explain the evolution of complex traits such as NFN symbiosis. This also suggests that reduction, similarly to the evolution of complexity, might be a major driver of the phenotypic diversity observed in extant organisms (*36*, *53*).

The fixation of loss of function alleles of *NIN* (either complete loss or pseudogenization) in non-nodulating species provides the genetic explanation for the loss of NFN symbiosis in ten species representing eight non-nodulating lineages. Fixation of such alleles requires ecological conditions in which the cost of symbiotic nitrogen-fixation, involving infection, building nodules to host bacteria, and providing carbon to feed them, outweighs the benefit to the plant. In most terrestrial habitats nitrogen is limiting (*54*), suggesting that the scale should be tipped toward the conservation of NFN symbiosis once this complex trait evolves. In nitrogen-rich habitats NFN symbiosis is known to be inhibited in legumes (*55*). Long-term fertilizer application would make NFN symbiosis-specific genes superfluous, leading to their eventual mutational inactivation and loss. In addition to this abiotic constraint, NFN symbiosis may be undermined by "cheating" bacteria that gain entry into root nodules and are fed by the plant, but do not deliver nitrogen (*56–58*). Cheaters may therefore imbalance the tradeoff between the costs and benefits of the association, as already proposed based on patterns of legume NFN symbiosis in Africa (*59*). This would result in the loss of NFN symbiosis being adaptive thus providing an ecological explanation for the occurrence of this symbiosis in a few flowering plant species. In this context, the finding that *NIN* participates in shaping the root microbiome beyond NFN symbiosis makes it a target of adaptive selection against this symbiosis (*60*).

Engineering biological nitrogen-fixation in crops remains a goal of plant synthetic biologists, with the aim to improve food production in developing countries in which the application of nitrogen-fertilizer is limited by economic and infrastructural constraints. Our results supporting the occurrence of multiple independent losses indicates that the apparent selection against NFN symbiosis must be taken into account by projects whose aim is to improve legumes and, even more, when considering the engineering of nitrogen-fixation in other crops.

Materials and Methods

Plant material and sample preparation

The origin of the plant material used for DNA or RNA extraction in this study is summarized in Data S2.

Methods for plant growth, DNA extractions and RNA extractions are described in the supplementary material online.

Genome sequencing

Figure S1 describes the overall strategy and results of the dataset production in this study.

Whole genome sequencing for the ten genomes was performed using Illumina sequencing technology (HISEq. 2000 and HISEq. 4000) at BGI-Shenzhen. Hierarchical library construction strategy was applied that typically included multiple paired-end libraries with insert sizes of 170, 250, 350, 500, and 800bp and mate-pair libraries with insert sizes of 2, 5, 10, and 20 Kb. Most of the paired-end and mate-pair libraries were prepared from large genomic fragments, typically of size 20-40 Kb, or even larger. For some species, more small-insert-size PE libraries were constructed to complement the limited mate-pair libraries. The library construction for each species is summarized in table S36. Deep genome sequencing was performed for the majority of species, with at least 110-fold coverage after a stringent data filtering and the highest sequencing depth reached 535-fold in cleaned data.

The overview statistics of data production are summarized in table S2 and fig. S1.

To minimize sequencing errors and reduce genome assembly artefacts, several quality control steps were taken to filter out low-quality sequencing reads:

Removal of N-rich reads: Reads that contained more than 10 percent of 'N's bases or polyA structure were removed;

Removal of low quality reads: Reads that had 40% of the bases are low-quality (quality scores ≤ 7) were filtered out;

Filtering of reads with ≥10nt aligned to the adapter sequences: adaptor sequences were aligned to read1 and read2 using a dynamic programming approach, if the aligned fragments from read1 or read2 were reverse complementary to each other, the pair was also removed.

Filtering of small insert size reads with insert size (170-800 bp): the overlapping length between read1 and read2 is ≥ 10bp, 10% mismatch was allowed.

Filtering of PCR duplicates: if the PE read1 and read2 were 100% identical, these reads were treated as duplicates and only one was retained.

Trimming of read end: the low-quality bases from read ends (5′-5bp, 3′-8bp) were directly trimmed.

This filtering process was carried out using an in-house Perl program. After filtering, in average 150-fold sequencing coverage were generated. For each species, clean reads were then passed to the genome assembler pipeline for de novo genome assembly.

The statistic of clean data are also summarized in table S2.

Genome assembly

To optimize the strategy for genome assembly, a genome survey is necessary to estimate the genome complexity. Some genomes are abundant in repetitive content and/or maintain a high heterozygosity rate through genome kmer-analysis. A k-mer refers to a continuous sequence with k base pairs, typically extracted from the reads (thus shorter than the read length, e.g., 17 bases per k-mer). If an 'ideal' sequencing data set is produced from randomly whole genome shotgun process without sequencing errors or coverage bias, the start positions of reads along the genome will follow Poisson distribution (*61*). Supposing that the read length is far shorter than the genome size, the k-mer can be regarded as randomly generated from the genome and their occurrence (sequencing depth) also is expected to be Poisson distributed (fig. S3A).

Based on this assumption, the genome size can be estimated as (*62*):

$$\text{Genome size} = \frac{\text{k-mer number}}{\text{Average sequencing depth}}$$

For a 'normal' diploid genome, the k-mer frequency produced from adequate reads would follow Poisson distribution. For genomes which are either repeat-rich or highly heterozygous, an additional peak either indicative of highly repetitive content (typically two-fold depths of the main peak) or of highly heterozygosity (*63*) (typically half depths of the main peak) is expected next to the 'main peak' (indicative of the normal diploid genome) from the frequency distribution, or some even more complex scenarios either caused by the unexpected non-canonical genomic characteristics (e.g., degree of heterozygosity, complexity of the size and distribution of repetitive content, etc.) coupled with the use of different k-mer size.

K-mer statistics and distributions are presented in fig. S2 and tables S6 to S25.

During de novo genome assembly, we tried different k–mers (from 23-mer to 33-mer) to construct contigs and the best k-mer (with the largest contig N50 length) was selected for the final run. Due to differences in genome complexities between species, multiple genome assemblers were applied to achieve the optimal assembly result. As described in fig. S1, SOAPdenovo2 (version 2.04) (*64*) was the most frequently used assembler and Platanus (version 1.2.4) (*65*) for highly heterozygous genomes. After several rounds of assembly evaluations regarding contig contiguity and genome completeness, the best assemblies (largest contig N50 and highest BUSCO gene mapping rate) were selected for the downstream gap-closing step by Gapcloser (version 1.2) (*64*). For the assembly of *Discaria trinervis* genome, we employed the Celera assembler CA8.3rc1. Because the Celera assembler (*66*) is sensitive to excessive coverage, library sizes were down-sampled to equal sized batches with a total coverage of approximately 50-fold. In a subsequent step the entire sequence information was used to generate scaffolds and close gaps with SSpace (*67*) and Gapcloser, respectively.

Table S3 indicates the assembly strategy for each genome.

The statistics of genome assemblies are summarized in table S3 and the details for each species are presented in tables S6 to S25.

Genome assembly evaluation

The assembly evaluations for all of the genomes are provided in table S4.

Basically, mapping of the 1,440 ultra-conserved core eukaryotic genes from the BUSCO (*68*) data set, resulted in >90% of the core eukaryote genes recovered for the majority of the genome assemblies. Taken together, these results indicate good genome assembly qualities for most of the newly sequenced species in this study, especially with respect to the genic regions (Fig. 1).

Genome annotation

A schematic workflow for genome annotation is given in fig. S1.

Repeat identification

Identification of transposable elements (TEs) was carried out by RepeatMasker (version 4-0-5) (*69*). A custom repeat library was constructed for each species by careful self-training. To construct the repeat custom library, we first collected the miniature inverted repeat transposable elements (MITEs) from many closely-related species, created a lineage-specific custom library by MITE-hunter (*70*) with default parameters. For the prediction of long terminal repeats (LTR), we used LTRharvest (*71*) integrated in Genometools (version 1.5.8) (*72*), defining LTR in the length of 1.5 kb to 25 kb, with two terminal repeats ranging from 100 bp to 6000 bp with ≥ 99% similarity. Elements with intact PPT (poly purine tract) or PBS (primer binding site) were necessary to define LTR, which were identified by LTRdigest (*71*) using an eukaryotic tRNA library (http://gtrnadb.ucsc.edu/), while elements without appropriate PPT or PBS location were removed. In order to remove false positives such as local gene clusters and tandem local repeats, 50 bp flanking sequences on both sides of the LTRs of each candidate element were aligned using MUSCLE (*73*) with default parameters; if the identity ≥ 60%, the LTR element was considered as a false positive and removed. LTR elements nested with other inserted, but unrelated components were also removed. Exemplars were built using a cutoff of 80% identity in 90% of element length from an all vs. all BLASTn search. Terminal repeat retrotransposon in miniature (TRIM) libraries, with length of 70 bp to 500 kb, were built following a similar prediction strategy. Furthermore, the genomic sequence was masked to run RepeatModeler (version 1-0-8) (*69*) to extensively de novo predict repetitive sequences for each species. The MITE, LTR and TRIM repetitive sequence libraries were integrated together to make a complete and non-redundant custom library. This custom repeat library was taken as the input for RepeatMasker to identify and classify transposable elements genome-wide for each species.

Gene annotation

Repeat elements were masked for each genome assembly before gene model prediction. Protein-coding genes were identified using the MAKER-P pipeline (version 2.31) (*74*) with two rounds of iterations. To obtain an optimal gene prediction, series of trainings was performed. First, for genomes that have RNA samples sequenced, a set of transcripts was generated by a genome-guided approach using Trinity and then mapped back to the genome using PASA (version 2.0.2) (*75*). This process generated a set of complete gene models from each genome assembly, and thus obtained real gene characteristics (size and number of exons/introns per gene, distribution of genes, features of splicing sites, etc.) by Augustus (*76*). Genemark-ES (version 4.21) (*77*) was self-trained with default parameters. SNAP (*78*) was trained using RNA- or protein-based gene models from the first iteration of MAKER-P pipeline. For RNA-seq aided gene annotation, RNA clean reads were assembled into inchworms using Trinity (*79*). For some species, transcriptome/ESTs data were obtained from NCBI or 1KP database if available (https://sites.google.com/a/ualberta.ca/onekp/). An optimal core protein set was collected from several closely-related species for homolog-based gene prediction for each species. For example, gene models from the model plants like *Arabidopsis thaliana*, *Oryza sativa*, as well as from some well-annotated legumes like Medicago and *Glycine max*. Default parameters were used to run MAKER-P with all integrated annotation sources and to produce the final set of gene models for each species. Number of gene models for each species is summarized in table S26 and detailed statistics are summarized in table S28. BUSCO evaluation suggests complete and reliable gene annotation for all newly-sequenced genomes (tables S4 and S5).

HMMER-based engineer InterproScan (version 5.11) (*80*) was used to predict gene function from several functional databases. The motifs and domains of genes were determined by searching against protein databases. An integrated gene functional annotation is summarized in table S29 for all species.

Transcriptome sequencing

RNA samples were sequenced for seven species to assist gene prediction in this study. The overview of total RNA samples with various tissues is summarized in table S37. For each RNA sample, a pair-end library with insert size of ~200bp was constructed following the manufacturer protocol. Libraries were barcoded and pooled together as input to the Illumina Hiseq 4000 platform for sequencing. All of the RNA samples were sequenced in-depth, with an average of 6 Gb sequences per sample, to ensure a complete coverage for each transcriptome.

Genome-wide comparative phylogenomic analysis

Clustering of gene families

Clustering of gene families are based on predicted proteomes of the gene annotation of 37 species (table S1). To generate a set of non-redundant representative sequences, we removed multiple isoforms of a gene applying a cd-hit clustering using an identity threshold of 99.5% (*81*). Subsequently, homologs were identified with an all versus all blastp (v.2.2.30+) search of the 37 species preoteomes. For each query-subject-pair we summed up the aligned sequence of all its HSPs (blast high scoring pair) ignoring overlaps and compared it with the sequence length of both subject and query. We removed all query-subject-pairs from the blast tables for which the alignment coverage was less than 40% of either the total query or subject sequence length. According to Yang and Smith (*82*) this hit fraction filter step with the used cutoff of 40% significantly improves phylogenetic trees and orthology inference in the subsequent steps. Based on these modified blast tables we clustered the remaining homologs to gene families with OrthoFinder (inflation parameter of 1.3) (*20*). The resulting 29,433 gene families were used as a starting point for the genome-wide phylogeny-based ortholog presence/absence analysis and candidate confirmation. We also calculated gene family clusters without applying the 40% hit fraction filter and used these 23,869 gene family clusters for the analysis of gene copy number variation. Figure S3 provides an overview both for the individual steps and the complete pipeline.

Genome-wide phylogeny-based ortholog presence/absence analysis

For each OrthoFinder gene family cluster a separate phylogeny was calculated with FastME (*83*). As suggested by Yang and Smith (*82*), unreliable or wrongly resolved super-long branches were removed from each family tree by the following way: for each tree, average length and standard deviation of terminal branches were calculated and branches longer than the mean of the average terminal branch length plus three-fold standard deviation were removed. A python script was written to root pruned trees with farthest-oldest outgroup method (implemented in the Python package ETE 3 (*84*)) and ortholog/paralog relationships were inferred with the species overlap algorithm (implemented in the Python package ETE 3) (*84*). We then searched lists of orthologs of each gene of the reference species, the well-characterized nodulating legume Medicago, in total 29,213 ortholog lists, for the presence or absence of orthologs in all remaining 36 species of our dataset. The following criteria were applied to identify candidates:

a) Orthologs present in at least 66% of the nodulating species (10 of 15),

b) Orthologs present in a fraction of nodulating species from each order of the NFN clade (at least 7 of 10 nodulating Fabales, at least 1 of 2 nodulating Rosales, at least 1 of 2 nodulating Fagales, 1 of 1 nodulating Cucurbitales),

c) (Only for predisposition hypothesis) orthologs absent from ALL outgroup species outside of the NFN clade.

d) (Only for *multiple losses* hypothesis) orthologs absent from more than 50% of the non-nodulating species (at least 7 of 13).

Instead of filtering for the presence of orthologs of all nodulating species (10/10 Fabales, 2/2 Rosales, 2/2 Fagales and 1/1 Cucurbitales) we used relatively relaxed criteria (as defined in a and b) for nodulating species to avoid missed potential candidates due to erroneous gene annotations (e.g., false negatives from gene annotation pipelines). Each of the candidates resulting from these criteria (predisposition hypothesis: 31 candidates; *multiple losses* hypothesis: 121 candidates) underwent a refined candidate analysis described below with stricter criteria for nodulating species.

Besides the candidates of the genome-wide presence/absence analysis of phylogeny-based orthologs, we collected 22 candidate genes (table S33) that have been reported to be involved in root nodule symbiosis mostly from the model legume organisms Medicago and *Lotus japonicus*. For each OrthoFinder gene family cluster containing one of these genes we calculated maximum likelihood gene family trees (RaxML v.8.2.4 Model: CATWAG). Based on the topology of the gene family tree and its protein alignment (MAFFT v.7.222 L-INS-I (*85*), trimming: BMGE gap-rate cut-off 80%) (*86*) we manually selected the subtree that contained the gene of interest (orthogroup). Subsequently we realigned (MAFFT v.7.222 L-INS-i, trimming: BMGE gap-rate cutoff 20%) and recalculated the phylogenetic tree (RaXML v.8.2.4 Model: GAMMAJTT, 200 bootstraps) (*87*) of the orthologous group to improve the quality of the subtree. Trees were rooted manually. Starting from the gene of interest and traversing to the root of the tree we marked all nodes as duplication or speciation events. If the two subclades of a node shared genes that were originating from the same species, this node was interpreted as a duplication, otherwise as a speciation event. Based on speciation nodes we inferred the orthologs of the gene of interest. At duplication nodes all genes in the subclade lacking the query gene were inferred as paralogs. In case of a speciation node all genes belonging to both subclades of that particular node were inferred as orthologs of the query gene unless they were annotated as paralogs at a previous node. All orthologs with incomplete gene models were removed as long as they had paralogs among the species they were derived from keeping at least one ortholog per species. We defined gene models as incomplete/fragmented, if more than 20% of the conserved amino acid sequence was absent. As conserved amino acid sequences we used the trimmed alignments (BMGE 20% gap-rate cutoff). To avoid false conclusions for missing orthologs, we retested a potential absence of genes by a homolog search. In such a case and in case of still remaining incomplete gene models we searched the complete genome sequence of the corresponding species (tblastn v2.2.30+, default parameters) with the closest homolog from the gene tree for regions containing potential gene loci of putative orthologs. These regions were then used to predict gene models (fgenesh+) (*88*). The resulting gene models were included in the set of sequences of the orthologous group and another round of alignment and tree calculation was performed (same settings and tools as last round). These identified sequences complementing the species gene annotation are provided as a separate fasta-formatted file in Data S1. The resulting tree was then used for a final round of ortholog inference so that the final set of orthologous genes is the result of an iterative process with constant improvement of gene models and phylogenies. If a complete gene model was not detected, the fragmented model was used and the ortholog was annotated as 'fragmented' for the respective species. Fragmented models were only annotated as complete, if the different fragments merged to a complete model and the fragmentation could be explained by the fragmentation of genomic scaffolds.

For the pre-filtered candidates from the genome-wide phylogeny-based ortholog presence/absence analysis we used more strict criteria than in the fully automated approach that led to the pre-filtered candidates. We only kept such candidates from the genome-wide phylogeny-based automated presence/absence pipeline for which orthologs of all nodulating species were present and orthologs in more than 50% of non-nodulating species (7 of 13) were absent. Orthologs absent from the orthofinder output were independently searched by microsynteny to exclude the possibility that the ortholog could not be found because the syntenic region of the ortholog was not in the corresponding genome assembly (for more detail see "synteny analysis" section). The presence, absence and fragmentation of orthologs for each of the 22 selected known symbiosis genes and each species is summarized in Fig. 3. The diversification times shown in the chronogram are based on estimates from Bell *et al.* (Outgroup, BEAST, 36 minimum age constraints treated as lognormal distributions) (*89*), Xi *et al.* (Malpighiales, BEAST, uncorrelated lognormal model) (*90*, *91*) and Li *et al.* (NFN clade, r8s, 1008 taxatree) (*13*). Phylogenetic trees for all candidate genes are provided in figs. S4 and S6 to S25.

Analysis of gene copy number variation

OrthoFinder gene family clusters (see above "1. Gene family clusters") were used to identify nodes in the species tree of our dataset, where gene family expansions or contractions in fast evolving gene families occurred. The number of genes for each species of each cluster were counted and analyzed employing the software tool CAFE (*92*). Following instructions given in the CAFE manual we removed gene family clusters with strong outliers in gene copy number. Therefore, we excluded 193 gene family clusters from the analysis for which the difference between maximum copy number and median copy number was greater than or equal to 50 copies which meet the elbow criterion to identify the optimal number of clusters in a clustering problem (fig. S27). To avoid overestimation of gene family contractions, we only used gene family clusters that contained orthologs from at least 28 species. This enabled us to analyze gene families that lost all genes in 0 up to 9 species. We chose this cut-off, because of the 9 outgroup species in the dataset. In the extreme case that all these 9 outgroups have a gene count of 0 we could still analyze gene families originating from the last common ancestor of the NFN clade. After applying this cut-off a total of 10,237 gene families were kept for the gene family evolution analysis (automatic λ and μ estimation, significance level for fast evolving families 5%). The results of the analysis are shown in Fig. 2.

From the Medicago Gene Expression Atlas we obtained all differentially expressed probesets that were either upregulated with a fold change of 2 or downregulated with a fold change of 0.5 in root nodule tissue of different age (7, 10 14 and 28 dpi) compared to untreated root tissue . We associated all of these 18,131 regulated probesets to 17,521 Medicago v4.0 gene IDs using the mapping file provided by MtGEA. For each of the 14 hypothesized independent gain of NFN symbiosis nodes (Fig. 2, blue boxes) we extracted all gene family clusters that showed expansions at these nodes. For each of these clusters we counted the number of transcriptionally regulated und not regulated Medicago genes ignoring all genes that could not be mapped to probesets. To test whether a gene family cluster was enriched for Medicago genes differentially expressed in nodulating versus mock control roots we performed Fisher's exact test on a significance level of 5%.

Synteny analysis

Genome-wide syntenic and collinear blocks were identified across the 37 selected genomes in this study. First, all vs. all Blastp (E-value ≤ $1e^{-10}$) was performed on the translated protein sequences of the 37 set of annotated gene models, resulting in a database of protein similarity. We then used the Multiple Collinearity Scan (Mcscan toolkit version 1.1, 2016, ≥ 5 homologous gene pairs/block) to identify conserved collinear blocks between the 37 genomes, creating a syntenic/collinear block database across all of the 37 species. In order to find all of the homologous syntenic blocks of interest, we first used Medicago genome as reference, searching and locating the target genes along the syntenic blocks with the flanking genes surrounding up-/down-100 kb genomic regions as well as the counterparts from different genomes. For any other given genome, the optimal collinear block (the highest score if multiple duplicated blocks were found) was defined according to conservation of gene content (the largest number of orthologous gene pairs) and consistency of gene order. If a corresponding ortholog was present in the collinear block from other aligned genome, this was called scenario-1 (indicating synteny supports gene presence and consistent with the genome-wide gene family ortholog prediction), if the target orthologous gene was absent from the collinear blocks of the given genome, this was called scenario-2 (synteny supports gene absence). After this first round was finished, to complete any missing genes/blocks due to weak alignment signals, we classified these identified orthologous genes as well as the corresponding collinear blocks, and repeat the searching and locating process between genomes according to their evolutionary proximity in phylogenetic position (using the most closely-related genome as query). By this process, we updated scenario-1 and scenario-2 as described above. In addition, for some species, no collinear block was identified around the possible target gene, we manually re-visited the protein similarity database, and searched the genomic regions flanking the Medicago gene to confirm the gene presence or absence supported by synteny. Finally, if no synteny was identified, but the candidate ortholog gene was predicted from the genome-wide gene family analysis, we called this as scenario-3 (gene presence without synteny support, which indicates possible gene translocation and synteny erosion).

Detecting selection pressure on gene trees

For *NIN* and *RPG*, protein sequences were aligned using MAFFT v7.380 (*85*). The protein alignment served as matrix for codon alignment performed using the Perl script pal2nal v14 with the –nogap option enabled to remove all gapped positions. Codon alignments were then subjected to a maximum likelihood analysis using IQ-TREE v1.6.1 (*93*) with 10,000 Ultrafast bootstraps replicates (*94*). The best-fitted evolutionary model was previously investigated using ModelFinder (*95*). The unrooted tree obtained was controlled to fit with the evolutionary frame of species and the NFN clade labeled as the foreground branch that was tested for being under positive selection. This latter was investigated using the branch-site model A implemented in the codeml module from the PAML package v4.9 g (*49*). An alternative hypothesis (NFN clade may have proportion of sites under positive selection) was compared to the null hypothesis (NFN clade may have different proportion of sites under neutral selection compared to the other clades). For the null model, the parameters were set as follows: "model=2, NSites=2, fix_kappa=0, fix_omega=1 and omega=1" while the parameters for the alternative model were: "model=2, NSites=2, fix_kappa=0, fix_omega=0 and omega=1.5". The two hypotheses were compared using the likelihood ratio test based on a chi2 distribution with one degree of freedom. If the alternative hypothesis was validated, codon sites likely to fall under positive selection were identified using the Bayes Empirical Bayes procedure (*49*).

PCR-validation of the absence of NIN *in non-nodulating legumes*

A nested PCR approach was undertaken using primers designed on an alignment of *NIN* gDNA from Medicago and *Mimosa pudica*. These primers are designed to amplify ~120 bp on Exon 4, ~170 bp on Exon 5 and the intron in between. Size of this intron ranges from 133bp in Medicago to 397 bp in *Mimosa pudica*. For the first PCR we used the degenerated primer pair NIN-Fwd-3 5′- GGAGAAAGTCMGGCGASAA and NIN-Rev-3 5′- GRAARCTGGCATAGAATGA. The Nested PCR was run with 0.2 μl of the PCR reaction and primers NIN-Fwd-2 5′- CGAACCAAGGCTGAGAAGAC and NIN-Rev-2 5′-ATCTGTATGGCACCCTCTGC. The first PCR run was as follow: 94°C 30s, 45°C 1 min, 72°C 1 min for 35 cycles; and for PCR two: 94°C 30s, 50°C 1 min, 72°C 1 min for 35 cycles. For all species the PCR was run on >2 samples. In addition, a PCR on the 28S was run to confirm the quality of the samples. All PCRs were run using a GoTaq DNA polymerase.

REFERENCES AND NOTES

1. P. C. Dos Santos, Z. Fang, S. W. Mason, J. C. Setubal, R. Dixon, Distribution of nitrogen fixation and nitrogenase-like sequences amongst microbial genomes. *BMC Genomics* **13**, 162 (2012). doi:10.1186/1471-2164-13-162 Medline
2. P. H. Beatty, A. G. Good, Future prospects for cereals that fix nitrogen. *Science* **333**, 416–417 (2011). doi:10.1126/science.1209467 Medline
3. L. Hiltner, Über die Bedeutung der Wurzelknöllchen von *Alnus glutinosa* für die Stickstoffernährung dieser Pflanze. *Landw. Versuchsstat* **46**, 153–161 (1895).
4. H. Hellriegel, H. Wilfarth, *Untersuchungen über die Stickstoffnahrung der Gramineen und Leguminosen* (Buchdruckerei der Post Kayssler, Berlin, 1888).
5. D. E. Soltis, P. S. Soltis, D. R. Morgan, S. M. Swensen, B. C. Mullin, J. M. Dowd, P. G. Martin, 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 (1995). doi:10.1073/pnas.92.7.2647 Medline
6. C. Kistner, M. Parniske, Evolution of signal transduction in intracellular symbiosis. *Trends Plant Sci.* **7**, 511–518 (2002). doi:10.1016/S1360-1385(02)02356-7 Medline
7. J. J. Doyle, Phylogenetic perspectives on the origins of nodulation. *Mol. Plant Microbe Interact.* **24**, 1289–1295 (2011). doi:10.1094/MPMI-05-11-0114 Medline
8. J. I. Sprent, J. Ardley, E. K. James, Biogeography of nodulated legumes and their nitrogen-fixing symbionts. *New Phytol.* **215**, 40–56 (2017). doi:10.1111/nph.14474 Medline
9. K. Pawlowski, K. N. Demchenko, The diversity of actinorhizal symbiosis. *Protoplasma* **249**, 967–979 (2012). doi:10.1007/s00709-012-0388-4 Medline
10. S. M. Swensen, The evolution of actinorhizal symbioses: Evidence for multiple origins of the symbiotic association. *Am. J. Bot.* **83**, 1503–1512 (1996). doi:10.1002/j.1537-2197.1996.tb13943.x
11. G. D. A. Werner, W. K. Cornwell, J. I. Sprent, J. Kattge, E. T. Kiers, A single evolutionary innovation drives the deep evolution of symbiotic N_2-fixation in angiosperms. *Nat. Commun.* **5**, 4087 (2014). doi:10.1038/ncomms5087 Medline
12. S. M. Swensen, D. R. Benson, in *Nitrogen-fixing Actinorhizal Symbioses*, K. Pawlowski, W. E. Newton, Eds. (Springer Netherlands, 2008), pp. 73–104.
13. H. L. Li, W. Wang, P. E. Mortimer, R.-Q. Li, D.-Z. Li, K. D. Hyde, J.-C. Xu, D. E. Soltis, Z.-D. Chen, Large-scale phylogenetic analyses reveal multiple gains of actinorhizal nitrogen-fixing symbioses in angiosperms associated with climate change. *Sci. Rep.* **5**, 14023 (2015). doi:10.1038/srep14023 Medline
14. J. A. Eisen, C. M. Fraser, Phylogenomics: Intersection of evolution and genomics. *Science* **300**, 1706–1707 (2003). doi:10.1126/science.1086292 Medline
15. P. M. Delaux, G. Radhakrishnan, G. Oldroyd, Tracing the evolutionary path to nitrogen-fixing crops. *Curr. Opin. Plant Biol.* **26**, 95–99 (2015). doi:10.1016/j.pbi.2015.06.003 Medline
16. Materials and methods are available as supplementary materials.
17. N. D. Young, F. Debellé, G. E. D. Oldroyd, R. Geurts, S. B. Cannon, M. K. Udvardi, V. A. Benedito, K. F. X. Mayer, J. Gouzy, H. Schoof, Y. Van de Peer, S. Proost, D. R. Cook, B. C. Meyers, M. Spannagl, F. Cheung, S. De Mita, V. Krishnakumar, H. Gundlach, S. Zhou, J. Mudge, A. K. Bharti, J. D. Murray, M. A. Naoumkina, B. Rosen, K. A. T. Silverstein, H. Tang, S. Rombauts, P. X. Zhao, P. Zhou, V. Barbe, P. Bardou, M. Bechner, A. Bellec, A. Berger, H. Bergès, S. Bidwell, T. Bisseling, N. Choisne, A. Couloux, R. Denny, S. Deshpande, X. Dai, J. J. Doyle, A.-M. Dudez, A. D. Farmer, S. Fouteau, C. Franken, C. Gibelin, J. Gish, S. Goldstein, A. J. González, P. J. Green, A. Hallab, M. Hartog, A. Hua, S. J. Humphray, D.-H. Jeong, Y. Jing, A. Jöcker, S. M. Kenton, D.-J. Kim, K. Klee, H. Lai, C. Lang, S. Lin, S. L. Macmil, G. Magdelenat, L. Matthews, J. McCorrison, E. L. Monaghan, J.-H. Mun, F. Z. Najar, C. Nicholson, C. Noirot, M. O'Bleness, C. R. Paule, J. Poulain, F. Prion, B. Qin, C. Qu, E. F. Retzel, C. Riddle, E. Sallet, S. Samain, N. Samson, I. Sanders, O. Saurat, C. Scarpelli, T. Schiex, B. Segurens, A. J. Severin, D. J. Sherrier, R. Shi, S. Sims, S. R. Singer, S. Sinharoy, L. Sterck, A. Viollet, B.-B. Wang, K. Wang, M. Wang, X. Wang, J. Warfsmann, J. Weissenbach, D. D. White, J. D. White, G. B. Wiley, P. Wincker, Y. Xing, L. Yang, Z. Yao, F. Ying, J. Zhai, L. Zhou, A. Zuber, J. Dénarié, R. A. Dixon, G. D. May, D. C. Schwartz, J. Rogers, F. Quétier, C. D. Town, B. A. Roe, The *Medicago* genome provides insight into the evolution of rhizobial symbioses. *Nature* **480**, 520–524 (2011). doi:10.1038/nature10625 Medline
18. J. Schmutz, P. E. McClean, S. Mamidi, G. A. Wu, S. B. Cannon, J. Grimwood, J. Jenkins, S. Shu, Q. Song, C. Chavarro, M. Torres-Torres, V. Geffroy, S. M. Moghaddam, D. Gao, B. Abernathy, K. Barry, M. Blair, M. A. Brick, M. Chovatia, P. Gepts, D. M. Goodstein, M. Gonzales, U. Hellsten, D. L. Hyten, G. Jia, J. D. Kelly, D. Kudrna, R. Lee, M. M. S. Richard, P. N. Miklas, J. M. Osorno, J. Rodrigues, V. Thareau, C. A. Urrea, M. Wang, Y. Yu, M. Zhang, R. A. Wing, P. B. Cregan, D. S. Rokhsar, S. A. Jackson, A reference genome for common bean and genome-wide analysis of dual domestications. *Nat. Genet.* **46**, 707–713 (2014). doi:10.1038/ng.3008 Medline
19. J. Schmutz, S. B. Cannon, J. Schlueter, J. Ma, T. Mitros, W. Nelson, D. L. Hyten, Q. Song, J. J. Thelen, J. Cheng, D. Xu, U. Hellsten, G. D. May, Y. Yu, T. Sakurai, T. Umezawa, M. K. Bhattacharyya, D. Sandhu, B. Valliyodan, E. Lindquist, M. Peto, D. Grant, S. Shu, D. Goodstein, K. Barry, M. Futrell-Griggs, B. Abernathy, J. Du, Z. Tian, L. Zhu, N. Gill, T. Joshi, M. Libault, A. Sethuraman, X.-C. Zhang, K. Shinozaki, H. T. Nguyen, R. A. Wing, P. Cregan, J. Specht, J. Grimwood, D. Rokhsar, G. Stacey, R. C. Shoemaker, S. A. Jackson, Genome sequence of the palaeopolyploid soybean. *Nature* **463**, 178–183 (2010). doi:10.1038/nature08670 Medline
20. D. M. Emms, S. Kelly, OrthoFinder: Solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* **16**, 157 (2015). doi:10.1186/s13059-015-0721-2 Medline
21. D. L. Stern, The genetic causes of convergent evolution. *Nat. Rev. Genet.* **14**, 751–764 (2013). doi:10.1038/nrg3483 Medline
22. N. Shubin, C. Tabin, S. Carroll, Deep homology and the origins of evolutionary novelty. *Nature* **457**, 818–823 (2009). doi:10.1038/nature07891 Medline
23. J. J. Doyle, Chasing unicorns: Nodulation origins and the paradox of novelty. *Am. J. Bot.* **103**, 1865–1868 (2016). doi:10.3732/ajb.1600260 Medline
24. J. F. Marsh, A. Rakocevic, R. M. Mitra, L. Brocard, J. Sun, A. Eschstruth, S. R. Long, M. Schultze, P. Ratet, G. E. D. Oldroyd, *Medicago truncatula NIN* is essential for rhizobial-independent nodule organogenesis induced by autoactive calcium/calmodulin-dependent protein kinase. *Plant Physiol.* **144**, 324–335 (2007). doi:10.1104/pp.106.093021 Medline
25. F. Clavijo, I. Diedhiou, V. Vaissayre, L. Brottier, J. Acolatse, D. Moukouanga, A. Crabos, F. Auguy, C. Franche, H. Gherbi, A. Champion, V. Hocher, D. Barker, D. Bogusz, L. S. Tisa, S. Svistoonoff, The *Casuarina NIN* gene is transcriptionally activated throughout *Frankia* root infection as well as in response to bacterial diffusible signals. *New Phytol.* **208**, 887–903 (2015). doi:10.1111/nph.13506 Medline
26. H. Gherbi, K. Markmann, S. Svistoonoff, J. Estevan, D. Autran, G. Giczey, F. Auguy, B. Péret, L. Laplaze, C. Franche, M. Parniske, D. Bogusz, SymRK defines a common genetic basis for plant root endosymbioses with arbuscular mycorrhiza fungi, rhizobia, and *Frankia*bacteria. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 4928–4932 (2008). doi:10.1073/pnas.0710618105 Medline

27. L. Schauser, A. Roussis, J. Stiller, J. Stougaard, A plant regulator controlling development of symbiotic root nodules. *Nature* **402**, 191–195 (1999). doi:10.1038/46058 Medline
28. A. van Zeijl, T. A. K. Wardhani, M. Seifi Kalhor, L. Rutten, F. Bu, M. Hartog, S. Linders, E. E. Fedorova, T. Bisseling, W. Kohlen, R. Geurts, CRISPR/Cas9-mediated mutagenesis of four putative symbiosis genes of the tropical tree *Parasponia andersonii* reveals novel phenotypes. *Front. Plant Sci.* **9**, 284 (2018). doi:10.3389/fpls.2018.00284 Medline
29. S. Svistoonoff, F. M. Benabdoun, M. Nambiar-Veetil, L. Imanishi, V. Vaissayre, S. Cesari, N. Diagne, V. Hocher, F. de Billy, J. Bonneau, L. Wall, N. Ykhlef, C. Rosenberg, D. Bogusz, C. Franche, H. Gherbi, The independent acquisition of plant root nitrogen-fixing symbiosis in Fabids recruited the same genetic pathway for nodule organogenesis. *PLOS ONE* **8**, e64515 (2013). doi:10.1371/journal.pone.0064515 Medline
30. K. Markmann, G. Giczey, M. Parniske, Functional adaptation of a plant receptor-kinase paved the way for the evolution of intracellular root symbioses with bacteria. *PLOS Biol.* **6**, e68 (2008). doi:10.1371/journal.pbio.0060068 Medline
31. L. G. Nagy, R. A. Ohm, G. M. Kovács, D. Floudas, R. Riley, A. Gácser, M. Sipiczki, J. M. Davis, S. L. Doty, G. S. de Hoog, B. F. Lang, J. W. Spatafora, F. M. Martin, I. V. Grigoriev, D. S. Hibbett, Latent homology and convergent regulatory evolution underlies the repeated emergence of yeasts. *Nat. Commun.* **5**, 4471 (2014). doi:10.1038/ncomms5471 Medline
32. C. E. Conn, R. Bythell-Douglas, D. Neumann, S. Yoshida, B. Whittington, J. H. Westwood, K. Shirasu, C. S. Bond, K. A. Dyer, D. C. Nelson, Convergent evolution of strigolactone perception enabled host detection in parasitic plants. *Science* **349**, 540–543 (2015). doi:10.1126/science.aab1140 Medline
33. J. He, V. A. Benedito, M. Wang, J. D. Murray, P. X. Zhao, Y. Tang, M. K. Udvardi, The *Medicago truncatula* gene expression atlas web server. *BMC Bioinformatics* **10**, 441 (2009). doi:10.1186/1471-2105-10-441 Medline
34. G. Criscuolo, V. T. Valkov, A. Parlati, L. M. Alves, M. Chiurazzi, Molecular characterization of the *Lotus japonicus* NRT1(PTR) and NRT2 families. *Plant Cell Environ.* **35**, 1567–1581 (2012). doi:10.1111/j.1365-3040.2012.02510.x Medline
35. L. Aravind, H. Watanabe, D. J. Lipman, E. V. Koonin, Lineage-specific loss and divergence of functionally linked genes in eukaryotes. *Proc. Natl. Acad. Sci. U.S.A.* **97**, 11319–11324 (2000). doi:10.1073/pnas.200346997 Medline
36. R. Albalat, C. Cañestro, Evolution by gene loss. *Nat. Rev. Genet.* **17**, 379–391 (2016). doi:10.1038/nrg.2016.39 Medline
37. P. M. Delaux, K. Varala, P. P. Edger, G. M. Coruzzi, J. C. Pires, J.-M. Ané, Comparative phylogenomics uncovers the impact of symbiotic associations on host genome evolution. *PLOS Genet.* **10**, e1004487 (2014). doi:10.1371/journal.pgen.1004487 Medline
38. A. Y. Borisov, L. H. Madsen, V. E. Tsyganov, Y. Umehara, V. A. Voroshilova, A. O. Batagov, N. Sandal, A. Mortensen, L. Schauser, N. Ellis, I. A. Tikhonovich, J. Stougaard, The *Sym35* gene required for root nodule development in pea is an ortholog of *Nin* from *Lotus japonicus*. *Plant Physiol.* **131**, 1009–1017 (2003). doi:10.1104/pp.102.016071 Medline
39. N. Azani, M. Babineau, C. D. Bailey, H. Banks, A. R. Barbosa, R. B. Pinto, J. S. Boatwright, L. M. Borges, G. K. Brown, A. Bruneau, E. Candido, D. Cardoso, K.-F. Chung, R. P. Clark, A. S. Conceição, M. Crisp, P. Cubas, A. Delgado-Salinas, K. G. Dexter, J. J. Doyle, J. Duminil, A. N. Egan, M. De La Estrella, M. J. Falcão, D. A. Filatov, A. P. Fortuna-Perez, R. H. Fortunato, E. Gagnon, P. Gasson, J. G. Rando, A. M. G. Azevedo Tozzi, B. Gunn, D. Harris, E. Haston, J. A. Hawkins, P. S. Herendeen, C. E. Hughes, J. R. V. Iganci, F. Javadi, S. A. Kanu, S. Kazempour-Osaloo, G. C. Kite, B. B. Klitgaard, F. J. Kochanovski, E. J. M. Koenen, L. Kovar, M. Lavin, M. Roux, G. P. Lewis, H. C. de Lima, M. C. López-Roberts, B. Mackinder, V. H. Maia, V. Malécot, V. F. Mansano, B. Marazzi, S. Mattapha, J. T. Miller, C. Mitsuyuki, T. Moura, D. J. Murphy, M. Nageswara-Rao, B. Nevado, D. Neves, D. I. Ojeda, R. T. Pennington, D. E. Prado, G. Prenner, L. P. de Queiroz, G. Ramos, F. L. Ranzato Filardi, P. G. Ribeiro, M. L. Rico-Arce, M. J. Sanderson, J. Santos-Silva, W. M. B. São-Mateus, M. J. S. Silva, M. F. Simon, C. Sinou, C. Snak, É. R. de Souza, J. Sprent, K. P. Steele, J. E. Steier, R. Steeves, C. H. Stirton, S. Tagane, B. M. Torke, H. Toyama, D. T. Cruz, M. Vatanparast, J. J. Wieringa, M. Wink, M. F. Wojciechowski, T. Yahara, T. Yi, E. Zimmerman, A new subfamily classification of the Leguminosae based on a taxonomically comprehensive phylogeny – The Legume Phylogeny Working Group (LPWG). *Taxon* **66**, 44–77 (2017). doi:10.12705/661.3
40. J. I. Sprent, Evolving ideas of legume evolution and diversity: A taxonomic perspective on the occurrence of nodulation. *New Phytol.* **174**, 11–25 (2007). doi:10.1111/j.1469-8137.2007.02015.x Medline
41. J.-F. Arrighi, O. Godfroy, F. de Billy, O. Saurat, A. Jauneau, C. Gough, The *RPG* gene of *Medicago truncatula* controls *Rhizobium*-directed polar growth during infection. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 9817–9822 (2008). doi:10.1073/pnas.0710273105 Medline
42. M. R. Chandler, Some observations on infection of *Arachis hypogaea* L. by *Rhizobium*. *J. Exp. Bot.* **29**, 749–755 (1978). doi:10.1093/jxb/29.3.749
43. Z. Peng, F. Liu, L. Wang, H. Zhou, D. Paudel, L. Tan, J. Maku, M. Gallo, J. Wang, Transcriptome profiles reveal gene regulation of peanut (*Arachis hypogaea* L.) nodulation. *Sci. Rep.* **7**, 40066 (2017). Medline
44. D. W. Gorbet, J. C. Burton, A non-nodulating peanut. *Crop Sci.* **19**, 727–728 (1979). doi:10.2135/cropsci1979.0011183X001900050045x
45. S. De Mita, A. Streng, T. Bisseling, R. Geurts, Evolution of a symbiotic receptor through gene duplications in the legume-rhizobium mutualism. *New Phytol.* **201**, 961–972 (2014). doi:10.1111/nph.12549 Medline
46. M. Parniske, Arbuscular mycorrhiza: The mother of plant root endosymbioses. *Nat. Rev. Microbiol.* **6**, 763–775 (2008). doi:10.1038/nrmicro1987 Medline
47. P. Favre, L. Bapaume, E. Bossolini, M. Delorenzi, L. Falquet, D. Reinhardt, A novel bioinformatics pipeline to discover genes related to arbuscular mycorrhizal symbiosis based on their evolutionary conservation pattern among higher plants. *BMC Plant Biol.* **14**, 333 (2014). doi:10.1186/s12870-014-0333-0 Medline
48. A. Bravo, T. York, N. Pumplin, L. A. Mueller, M. J. Harrison, Genes conserved for arbuscular mycorrhizal symbiosis identified through phylogenomics. *Nat. Plants* **2**, 15208 (2016). doi:10.1038/nplants.2015.208 Medline
49. Z. Yang, PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007). doi:10.1093/molbev/msm088 Medline
50. S. B. Carroll, Evo-devo and an expanding evolutionary synthesis: A genetic theory of morphological evolution. *Cell* **134**, 25–36 (2008). doi:10.1016/j.cell.2008.06.030 Medline
51. R. van Velzen, R. Holmer, F. Bu, L. Rutten, A. van Zeijl, W. Liu, L. Santuari, Q. Cao, T. Sharma, D. Shen, Y. Roswanjaya, T. A. K. Wardhani, M. S. Kalhor, J. Jansen, J. van den Hoogen, B. Güngör, M. Hartog, J. Hontelez, J. Verver, W. C. Yang, E. Schijlen, R. Repin, M. Schilthuizen, M. E. Schranz, R. Heidstra, K. Miyata, E. Fedorova, W. Kohlen, T. Bisseling, S. Smit, R. Geurts, Comparative genomics of the nonlegume *Parasponia* reveals insights into evolution of nitrogen-fixing rhizobium symbioses. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E4700–E4709 (2018). Medline
52. L. G. Nagy, Evolution: Complex multicellular life with 5,500 genes. *Curr. Biol.* **27**, R609–R612 (2017). doi:10.1016/j.cub.2017.04.032 Medline
53. M. A. O'Malley, J. G. Wideman, I. Ruiz-Trillo, Losing complexity: The role of simplification in macroevolution. *Trends Ecol. Evol.* **31**, 608–621 (2016). doi:10.1016/j.tree.2016.04.004 Medline
54. D. S. LeBauer, K. K. Treseder, Nitrogen limitation of net primary productivity in terrestrial ecosystems is globally distributed. *Ecology* **89**, 371–379 (2008). doi:10.1890/06-2057.1 Medline
55. J. Streeter, P. P. Wong, Inhibition of legume nodule formation and N_2-fixation by nitrate. *Crit. Rev. Plant Sci.* **7**, 1–23 (1988). doi:10.1080/07352688809382257
56. E. T. Kiers, R. A. Rousseau, S. A. West, R. F. Denison, Host sanctions and the legume-rhizobium mutualism. *Nature* **425**, 78–81 (2003). doi:10.1038/nature01931 Medline
57. H. Fujita, S. Aoki, M. Kawaguchi, Evolutionary dynamics of nitrogen fixation in the legume-rhizobia symbiosis. *PLOS ONE* **9**, e93670 (2014). doi:10.1371/journal.pone.0093670 Medline
58. L. Carro, P. Pujic, M. E. Trujillo, P. Normand, *Micromonospora* is a normal occupant of actinorhizal nodules. *J. Biosci.* **38**, 685–693 (2013). doi:10.1007/s12038-013-9359-y Medline
59. J. I. Sprent, *Legume Nodulation: A Global Perspective* (Wiley, 2009).
60. R. Zgadzaj, R. Garrido-Oter, D. B. Jensen, A. Koprivova, P. Schulze-Lefert, S. Radutoiu, Root nodule symbiosis in *Lotus japonicus* drives the establishment of distinctive rhizosphere, root, and nodule bacterial communities. *Proc. Natl. Acad. Sci. U.S.A.* **113**, E7996–E8005 (2016). doi:10.1073/pnas.1616564113 Medline
61. G. Marçais, C. Kingsford, A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770 (2011).

doi:10.1093/bioinformatics/btr011 Medline

62. E. Veeckman, T. Ruttink, K. Vandepoele, Are we there yet? Reliably estimating the completeness of plant genome sequences. *Plant Cell* **28**, 1759–1768 (2016). doi:10.1105/tpc.16.00349 Medline

63. R. Arratia, D. Martin, G. Reinert, M. S. Waterman, Poisson process approximation for sequence repeats, and sequencing by hybridization. *J. Comput. Biol.* **3**, 425–463 (1996). doi:10.1089/cmb.1996.3.425 Medline

64. R. Luo, B. Liu, Y. Xie, Z. Li, W. Huang, J. Yuan, G. He, Y. Chen, Q. Pan, Y. Liu, J. Tang, G. Wu, H. Zhang, Y. Shi, Y. Liu, C. Yu, B. Wang, Y. Lu, C. Han, D. W. Cheung, S.-M. Yiu, S. Peng, Z. Xiaoqian, G. Liu, X. Liao, Y. Li, H. Yang, J. Wang, T.-W. Lam, J. Wang, SOAPdenovo2: An empirically improved memory-efficient short-read de novo assembler. *Gigascience* **1**, 18 (2012). doi:10.1186/2047-217X-1-18 Medline

65. R. Kajitani, K. Toshimoto, H. Noguchi, A. Toyoda, Y. Ogura, M. Okuno, M. Yabana, M. Harada, E. Nagayasu, H. Maruyama, Y. Kohara, A. Fujiyama, T. Hayashi, T. Itoh, Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Res.* **24**, 1384–1395 (2014). doi:10.1101/gr.170720.113 Medline

66. E. W. Myers, G. G. Sutton, A. L. Delcher, I. M. Dew, D. P. Fasulo, M. J. Flanigan, S. A. Kravitz, C. M. Mobarry, K. H. Reinert, K. A. Remington, E. L. Anson, R. A. Bolanos, H. H. Chou, C. M. Jordan, A. L. Halpern, S. Lonardi, E. M. Beasley, R. C. Brandon, L. Chen, P. J. Dunn, Z. Lai, Y. Liang, D. R. Nusskern, M. Zhan, Q. Zhang, X. Zheng, G. M. Rubin, M. D. Adams, J. C. Venter, A whole-genome assembly of *Drosophila*. *Science* **287**, 2196–2204 (2000). doi:10.1126/science.287.5461.2196 Medline

67. M. Boetzer, C. V. Henkel, H. J. Jansen, D. Butler, W. Pirovano, Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* **27**, 578–579 (2011). doi:10.1093/bioinformatics/btq683 Medline

68. F. A. Simão, R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva, E. M. Zdobnov, BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015). doi:10.1093/bioinformatics/btv351 Medline

69. A. Smit, R. Hubley, P. Green, RepeatMasker Open−4.0, 2013–2015; www.repeatmasker.org.

70. Y. Han, S. R. Wessler, MITE-Hunter: A program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic Acids Res.* **38**, e199 (2010). doi:10.1093/nar/gkq862 Medline

71. D. Ellinghaus, S. Kurtz, U. Willhoeft, LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics* **9**, 18 (2008). doi:10.1186/1471-2105-9-18 Medline

72. G. Gremme, S. Steinbiss, S. Kurtz, GenomeTools: A comprehensive software library for efficient processing of structured genome annotations. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **10**, 645–656 (2013). doi:10.1109/TCBB.2013.68 Medline

73. R. C. Edgar, MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004). doi:10.1093/nar/gkh340 Medline

74. M. S. Campbell, M. Law, C. Holt, J. C. Stein, G. D. Moghe, D. E. Hufnagel, J. Lei, R. Achawanantakun, D. Jiao, C. J. Lawrence, D. Ware, S.-H. Shiu, K. L. Childs, Y. Sun, N. Jiang, M. Yandell, MAKER-P: A tool kit for the rapid creation, management, and quality control of plant genome annotations. *Plant Physiol.* **164**, 513–524 (2014). doi:10.1104/pp.113.230144 Medline

75. B. J. Haas, A. L. Delcher, S. M. Mount, J. R. Wortman, R. K. Smith Jr., L. I. Hannick, R. Maiti, C. M. Ronning, D. B. Rusch, C. D. Town, S. L. Salzberg, O. White, Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654–5666 (2003). doi:10.1093/nar/gkg770 Medline

76. M. Stanke, R. Steinkamp, S. Waack, B. Morgenstern, AUGUSTUS: A web server for gene finding in eukaryotes. *Nucleic Acids Res.* **32**, W309–W312 (2004). doi:10.1093/nar/gkh379 Medline

77. A. V. Lukashin, M. Borodovsky, GeneMark.hmm: New solutions for gene finding. *Nucleic Acids Res.* **26**, 1107–1115 (1998). doi:10.1093/nar/26.4.1107 Medline

78. I. Korf, Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59 (2004). doi:10.1186/1471-2105-5-59 Medline

79. M. G. Grabherr, B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. di Palma, B. W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman, A. Regev, Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011). doi:10.1038/nbt.1883 Medline

80. E. M. Zdobnov, R. Apweiler, InterProScan—an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* **17**, 847–848 (2001). doi:10.1093/bioinformatics/17.9.847 Medline

81. W. Li, A. Godzik, Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006). doi:10.1093/bioinformatics/btl158 Medline

82. Y. Yang, S. A. Smith, Orthology inference in nonmodel organisms using transcriptomes and low-coverage genomes: Improving accuracy and matrix occupancy for phylogenomics. *Mol. Biol. Evol.* **31**, 3081–3092 (2014). doi:10.1093/molbev/msu245 Medline

83. V. Lefort, R. Desper, O. Gascuel, FastME 2.0: A comprehensive, accurate, and fast distance-based phylogeny inference program. *Mol. Biol. Evol.* **32**, 2798–2800 (2015). doi:10.1093/molbev/msv150 Medline

84. J. Huerta-Cepas, F. Serra, P. Bork, ETE 3: Reconstruction, analysis, and visualization of phylogenomic data. *Mol. Biol. Evol.* **33**, 1635–1638 (2016). doi:10.1093/molbev/msw046 Medline

85. K. Katoh, D. M. Standley, MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013). doi:10.1093/molbev/mst010 Medline

86. G. Tan, M. Muffato, C. Ledergerber, J. Herrero, N. Goldman, M. Gil, C. Dessimoz, Current methods for automated filtering of multiple sequence alignments frequently worsen single-gene phylogenetic inference. *Syst. Biol.* **64**, 778–791 (2015). doi:10.1093/sysbio/syv033 Medline

87. A. Stamatakis, RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014). doi:10.1093/bioinformatics/btu033 Medline

88. V. Solovyev, in *Handbook of Statistical Genetics*, D. J. Balding, M. Bishop, C. Cannings, Eds. (Wiley, 2008), pp. 97–159.

89. C. D. Bell, D. E. Soltis, P. S. Soltis, The age and diversification of the angiosperms re-revisited. *Am. J. Bot.* **97**, 1296–1303 (2010). doi:10.3732/ajb.0900346 Medline

90. Z. Xi, B. R. Ruhfel, H. Schaefer, A. M. Amorim, M. Sugumaran, K. J. Wurdack, P. K. Endress, M. L. Matthews, P. F. Stevens, S. Mathews, C. C. Davis, Phylogenomics and a posteriori data partitioning resolve the Cretaceous angiosperm radiation Malpighiales. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 17519–17524 (2012). doi:10.1073/pnas.1205818109 Medline

91. M. T. J. Johnson, E. J. Carpenter, Z. Tian, R. Bruskiewich, J. N. Burris, C. T. Carrigan, M. W. Chase, N. D. Clarke, S. Covshoff, C. W. Depamphilis, P. P. Edger, F. Goh, S. Graham, S. Greiner, J. M. Hibberd, I. Jordon-Thaden, T. M. Kutchan, J. Leebens-Mack, M. Melkonian, N. Miles, H. Myburg, J. Patterson, J. C. Pires, P. Ralph, M. Rolf, R. F. Sage, D. Soltis, P. Soltis, D. Stevenson, C. N. Stewart Jr., B. Surek, C. J. M. Thomsen, J. C. Villarreal, X. Wu, Y. Zhang, M. K. Deyholos, G. K.-S. Wong, Evaluating methods for isolating total RNA and predicting the success of sequencing phylogenetically diverse plant transcriptomes. *PLOS ONE* **7**, e50226 (2012). doi:10.1371/journal.pone.0050226 Medline

92. M. V. Han, G. W. C. Thomas, J. Lugo-Martinez, M. W. Hahn, Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. *Mol. Biol. Evol.* **30**, 1987–1997 (2013). doi:10.1093/molbev/mst100 Medline

93. L.-T. Nguyen, H. A. Schmidt, A. von Haeseler, B. Q. Minh, IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015). doi:10.1093/molbev/msu300 Medline

94. D. T. Hoang, L. S. Vinh, T. Flouri, A. Stamatakis, A. von Haeseler, B. Q. Minh, MPBoot: Fast phylogenetic maximum parsimony tree inference and bootstrap approximation. *BMC Evol. Biol.* **18**, 11 (2018). doi:10.1186/s12862-018-1131-3 Medline

95. S. Kalyaanamoorthy, B. Q. Minh, T. K. F. Wong, A. von Haeseler, L. S. Jermiin, ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017). doi:10.1038/nmeth.4285 Medline

96. M. Griesmann, Y. Chang, X. Liu, Y. Song, G. Haberer, M. B. Crook, B. Billault-Penneteau, D. Lauressergues, L. Imanishi, Y. P. Roswanjaya, W. Kohlen, P. Pujic, K. Battenberg, N. Alloisio, W. Sun, H. Hilhorst, M. G. Salgado, V. Hocher, H. Gherbi,

S. Svistoonoff, J. J. Doyle, S. He, Y. Xu, W. Xian, Y. Fu, P. Normand, A. M. Berry, L. G. Wall, J. M. Ané, K. Pawlowski, X. Xu, H. Yang, M. Spannagl, K. F. X. Mayer, G. K. Wong, M. Parniske, P. M. Delaux, S. Cheng, Data supporting multiple independent losses of nitrogen-fixing root nodule symbiosis. GigaDB (2018); doi:10.5524/100300

97. V. Tolieng, B. Prasirtsak, J. Sitdhipol, N. Thongchul, S. Tanasupawat, Identification and lactic acid production of bacteria isolated from soils and tree barks. *Malays. J. Microbiol.* **13**, 100–108 (2017).

98. C. Valverde, L. G. Wall, Time course of nodule development in the *Discaria trinervis* (Rhamnaceae) – *Frankia* symbiosis. *New Phytol.* **141**, 345–354 (1999). doi:10.1046/j.1469-8137.1999.00345.x

99. G. Fahraeus, The infection of clover root hairs by nodule bacteria studied by a simple glass slide technique. *J. Gen. Microbiol.* **16**, 374–381 (1957). Medline

100. S. L. Dellaporta, J. Wood, J. B. Hicks, A plant DNA minipreparation: Version II. *Plant Mol. Biol. Report.* **1**, 19–21 (1983). doi:10.1007/BF02712670

101. A. Healey, A. Furtado, T. Cooper, R. J. Henry, Protocol: A simple method for extracting next-generation sequencing quality genomic DNA from recalcitrant plant species. *Plant Methods* **10**, 21 (2014). doi:10.1186/1746-4811-10-21 Medline

102. A. Ribeiro, A. D. L. Akkermans, A. van Kammen, T. Bisseling, K. Pawlowski, A nodule-specific gene encoding a subtilisin-like protease is expressed in early stages of actinorhizal nodule development. *Plant Cell* **7**, 785–794 (1995). doi:10.1105/tpc.7.6.785 Medline

103. J. M. Chirgwin, A. E. Przybyla, R. J. MacDonald, W. J. Rutter, Isolation of biologically active ribonucleic acid from sources enriched in ribonuclease. *Biochemistry* **18**, 5294–5299 (1979). doi:10.1021/bi00591a005 Medline

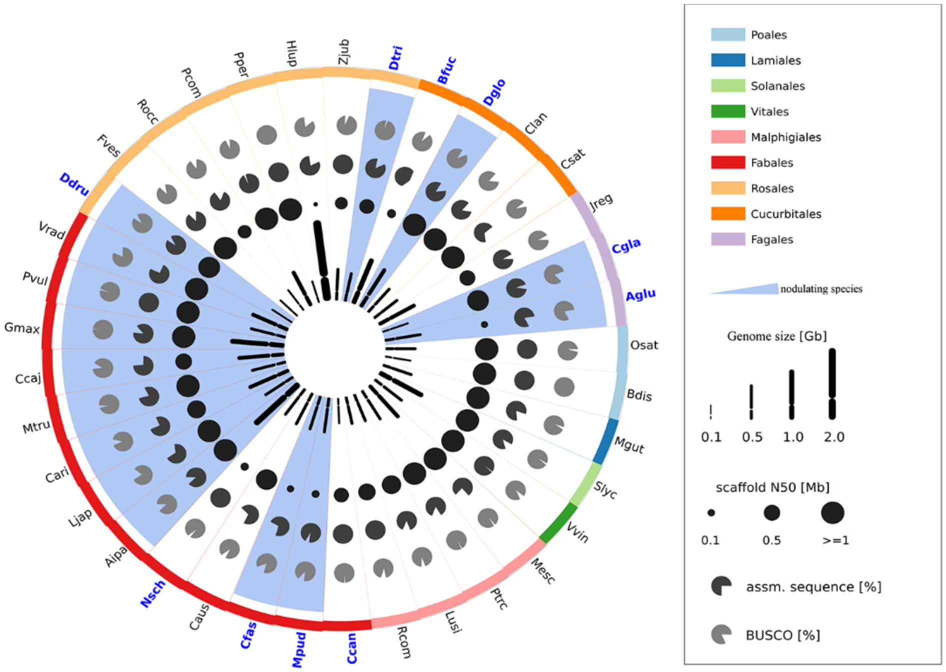
ACKNOWLEDGMENTS

The authors thank Janet Sprent and Gijsbert Werner for initial discussions, Randall Hagen and Susan Singer for *Chamaecrista fasciculata* leaves, Steven Cannon for *Cercis canadensis* leaves, Nathalie Séjalon-Delmas (Toulouse Botanical Garden), the International Center for Tropical Agriculture (CIAT) and USDA ARS-GRIN for seeds and Daniel Lang for bioinformatics advice. **Funding:** This work was partly performed in the LRSV, which belongs to the TULIP LABEX (ANR-10-LABX-41) and was supported by the Shenzhen Municipal Government of China to SC (NO. JCYJ20150529150409546) and YC (NO. JCYJ20150831201643396), by grant from the Swedish Research Council Vetenskapsrådet (VR 2012-03061) to KP, by the German Ministry of Education and Research grant 031A536 "de.NBI" and the Deutsche Forschungsgemeinschaft Project SFB924 to KFXM and MP, by grants from the National Science Foundation (NSF#1237936 and #1546742) to JMA, by grants from the French National Research Agency (ANR-12-BSV7-0007-02) and the United States Department of Agriculture (USDA; NIFA 2015-67014-22849) and ECOS-SUD (A13B03) to VH, HG and SS, by grant from the National University of Quilmes, Argentina (PUNQ 1411/15) to LGW and by USDA-NIFA590 CA-D-PLS-2173-H to AMB. WK is supported by NWO-VENI (863.15.010) and LGW and LI are members of CONICET (Argentina). GKSW is funded by Alberta Innovates Technology Futures (AITF) Innovates Centres of Research Excellence (iCORE). MP acknowledges the funding by the ERC Advanced grant 340904 (EvolvingNodules) which supported the work of MG and BPP and the sabbatical stay of KP at the LMU Munich. **Author contributions:** KP, MP, PMD and JMA defined the species to sequence and designed the project. KP coordinated the work on actinorhizal species, supervised MGS, isolated DNA for *Datisca glomerata* and *Begonia fuchsioides*, and extracted RNA of *B. fuchsioides*. MGS prepared RNA of *Datisca glomerata*. PMD coordinated the consortium, provided DNA and RNA samples for *Nissolia schottii*, conducted phylogenetic analyses of target genes and supervised DL and JK. DL extracted DNA for and conducted experiment presented in fig. S5. JK conducted the PAML analysis on *NIN* and *RPG*. JMA supervised experiments on most legumes. MBC extracted DNA and RNA of *Chamaecrista fasciculata, Mimosa pudica and Cercis canadensis*. WK and YPR extracted DNA samples for *Castanospermum australe, Sophora flavescens* and the *Chamaecrista* species used in fig. S5. LGW collected *Discaria trinervis* plant material and supervised LI. LI prepared DNA and RNA samples for *Discaria trinervis*. PN supervised experiments on *Alnus glutinosa*. NA and PP collected *Alnus glutinosa* material, designed extraction protocols and extracted RNA (NA) and DNA (PP). HG, VH and SS collected *Casuarina glauca* material, supervised DNA extraction and extracted RNA for *Casuarina glauca*. KFXM and MS coordinated genome assembly and annotation. MS and GH supervised MG. GH performed genome assembly, gene annotation, phylogenetic analyses, and prepared Fig. 1. MP initiated the consortium, supervised MG and BBP and the work on *Dryas drummondii*. MG conducted most of the phylogenetic analyses and prepared Fig. 2, Fig. 3 and related Supplementary texts and files. BBP collected material and extracted DNA and RNA from *Dryas drummondii*. SC and GKSW designed the sequencing strategy, coordinated and supervised the bioinformatics and genomic analysis, including genome sequencing, assembly and annotation at BGI-Shenzhen with HY, XL and XX. XL performed genome analyses. YC performed samples management and data quality control, and developed custom bioinformatics pipelines for genome assembly, annotation and evaluation with SX. SX managed the super computer clusters to perform assembly and annotation. YS identified orthologs for Fig. 3. And performed most of the microsynteny analyses. YL ran RNA data quality control, assembly, alignment and annotation. SH finalized the synteny analysis and prepared Fig. 3. YX, JQ and YF performed genome assembly, repeat and gene annotation respectively for *Mimosa pudica, Alnus glutinosa and Cercis canadensis* (YX) *Nissolia schottii* and *Discaria trinervis* (JG) and *Casuarina glauca, Chamaecrista fasciculata* (YF). HH provided access to the unpublished genome of *Castanospermum australe*. JJD provided comments and suggestions on the possible evolutionary scenarios and contributed to writing the manuscript. SC wrote the Supplementary texts with contributed sections from all co-authors. MP and PMD wrote the manuscript with inputs and comments from co-authors. **Competing interests:** The authors declare no competing interest. **Data and materials availability:** All raw reads have been deposited in the Sequence Read Archive (SRA) under project PRJNA374770, accession codes are summarized in supplementary material. All files have been deposited in GigaDB (*96*). All material used for genome sequencing is available upon request to the responsible author as presented in Data S2. All material used for genome sequencing is available upon request to PN for *Alnus glutinosa*, to KP for *Datisca glomerata* and *Begonia fuchioides*, to SS, HG or VH for *Casuarina glauca*, to JMA or MC for *Cercis canadensis*, *Chamaecrista fasciculata* and *Mimosa pudica*, to LW for *Discaria trivernis*, to MP for *Dryas drummondii* and to PMD for *Nissolia schottii* (also listed in Data S2).

SUPPLEMENTARY MATERIALS

www.sciencemag.org/cgi/content/full/science.aat1743/DC1
Materials and Methods
Figs. S1 to S27
Tables S1 to S37
References (*97*–*103*)
Data S1 and S2

12 February 2018; accepted 16 May 2018
Published online 24 May 2018
10.1126/science.aat1743



Osat	*Oryza sativa*	**Ccan**	***Cercis canadensis***	Ccaj	*Cajanus cajan*	Hlup	*Humulus lupulus*
Bdis	*Brachypodium distachyon*	**Mpud**	***Mimosa pudica***	Gmax	*Glycine max*	Zjub	*Ziziphus jujuba*
Mgut	*Mimulus guttatus*	**Cfas**	***Chamaecrista fasciculata***	Pvul	*Phaseolus vulgaris*	**Dtri**	***Discaria trinervis***
Slyc	*Solanum lycopersicum*	Caus	*Castanospermum australe*	Vrad	*Vigna radiata*	**Bfuc**	***Begonia fuchsoides***
Vvin	*Vitis vinifera*	**Nsch**	***Nissolia schottii***	**Ddru**	***Dryas drummondii***	**Dglo**	***Datisca glomerata***
Mesc	*Manihot esculenta*	Aipa	*Arachis ipaensis*	Fves	*Fragaria vesca*	Clan	*Citrullus lanatus*
Ptrc	*Populus trichocarpa*	Ljap	*Lotus japonicus*	Rocc	*Rubus occidentalis*	Jreg	*Juglans regia*
Lusi	*Linum usitatissimum*	Cari	*Cicer arietinum*	Pcom	*Pyrus communis*	Csat	*Cucumis sativus*
Rcom	*Ricinus communis*	Mtru	*Medicago truncatula*	Pper	*Prunus persica*	**Cgla**	***Casuarina glauca***
						Aglu	***Alnus glutinosa***

Fig. 1. Genome features of species used in this study. Genome statistics are shown as pictograms for species used in this study, with nodulating species highlighted by blue sectors. Species names are shown as four letter abbreviations at the outer circle, with their taxonomic order color coded shown at the top right legend. Newly sequenced species are in bold blue letters. The next two circles show as pie charts the proportion of complete BUSCO genes detected in the genome assembly (light grey) and the percentage of assembled sequence relative to the estimated genome size (dark grey), respectively. Scaffold N50 values are depicted as bubble charts (black) capped to a maximal N50 of 1 Mb to reduce graphical biases by finished genomes assembled to pseudo-chromosomes. Note that even for assemblies in this study with low contiguity, the BUSCO results suggest that the gene space has been well covered (tables S4 and S5). The innermost circle represents the genome size by proportional chromosome pictograms.

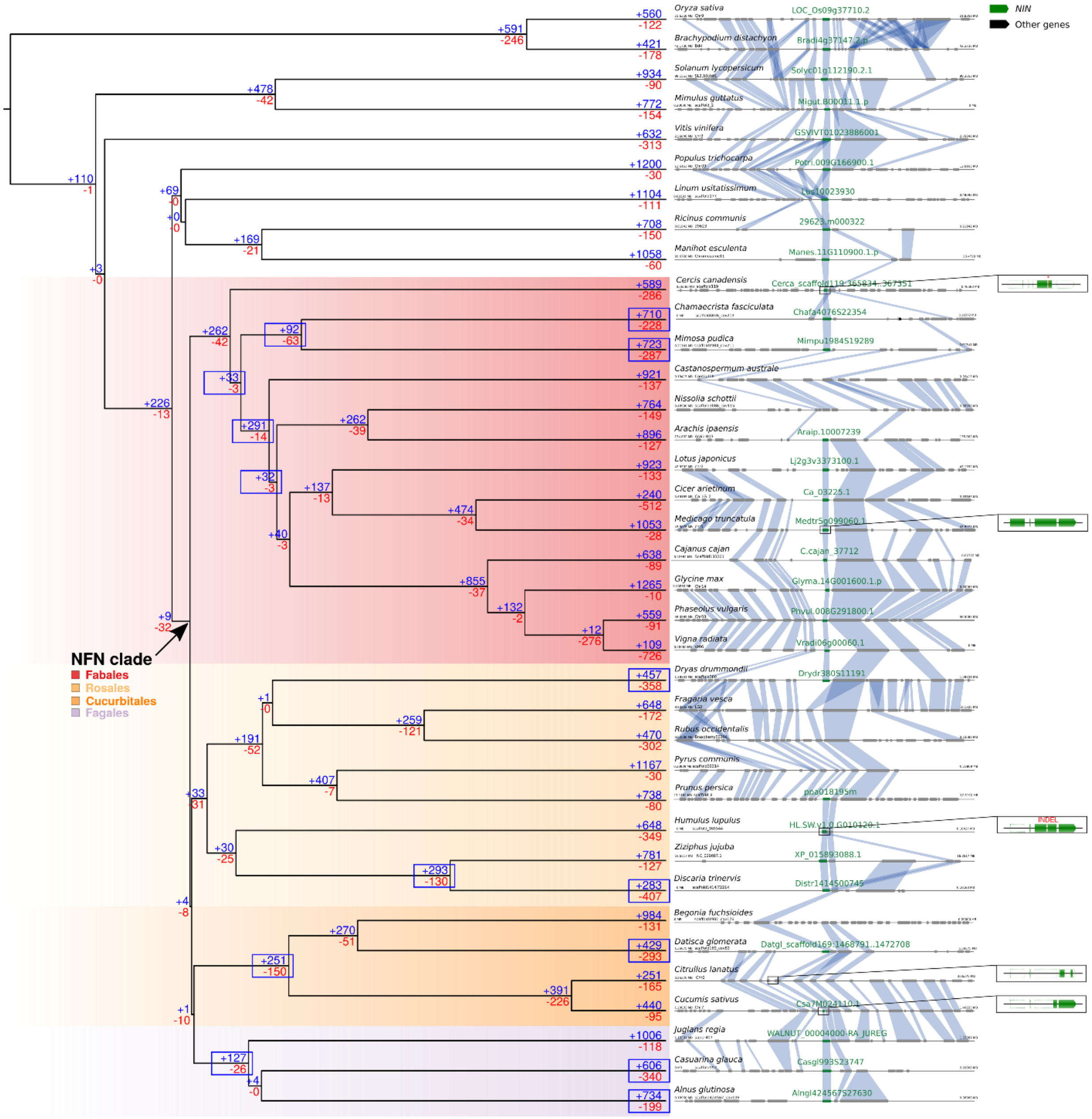


Fig. 2. Gene family expansions and contractions in the NFN clade. In the left panel, a species phylogeny of the used dataset is depicted. For each node, the number of gene family showing expansions (+, blue) and contractions (-, red) are given. Blue boxes point out nodes hypothesized to be positions of independent gain events of the NFN symbiosis including all suggested alternative models in the given dataset (*7*). For example, among the Fabales the dataset could comprise one, two or three independent gains. A black arrow marks the base of the NFN clade. The right panel depicts syntenic relationships of the *NIN* region. *NIN* genes are colored in green. *NIN* gene IDs are shown above the gene symbol. The synteny analysis upholds orthologous relationships drawn from the phylogenetic analysis and supports the absence of *NIN* in several species by verifying the existence of contiguous *NIN* regions without *NIN* genes. Enlarged gene models are only shown for fragmented *NIN* genes in comparison to the full *Medicago truncatula NIN* gene. Blocks with no and green filling represent parts absent and present, respectively, when compared to Medicago *NIN*. Insertions/deletions and premature stop codons (*) are symbolized by a vertical red line.

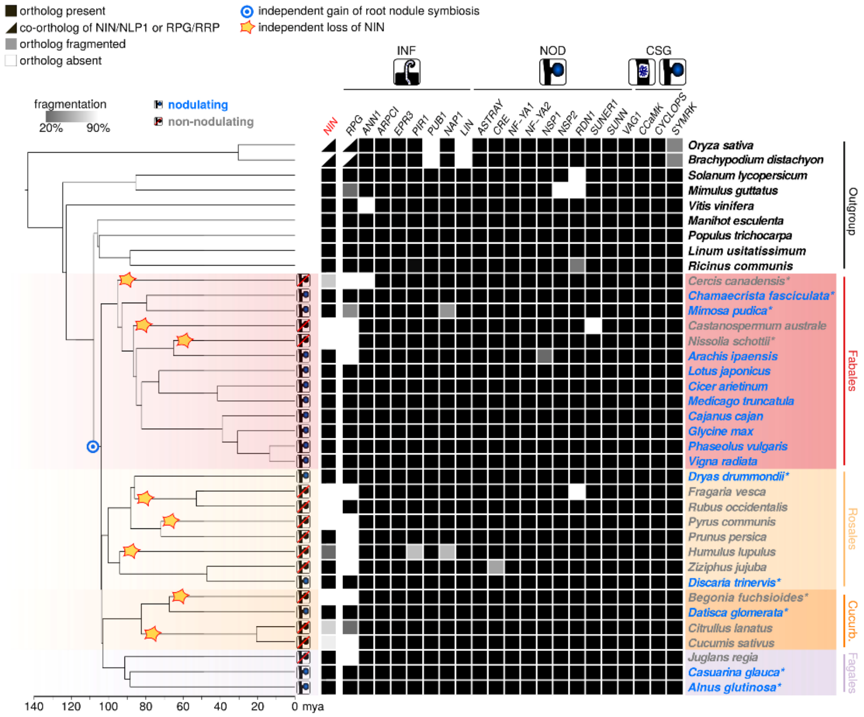


Fig. 3. Phylogenetic pattern of NFN symbiosis-related genes. The chronogram contains nodulating (blue text) and non-nodulating species (grey) from all four orders of the NFN clade (blue dot), to which NFN symbiosis is limited. Nine species outside the NFN clade are included as outgroup at the top. Absence and presence of entire or fragmented copies of 21 symbiosis genes are indicated by white, black and grey boxes, respectively. Stars indicate independent losses of *NIN*. The independent loss or fragmentation of *NIN* correlates with the absence of nodules after the emergence of the NFN clade. *RPG* is lost or fragmented in even more non-nodulating species than *NIN*, but also in the nodulating species *Arachis ipaensis* and *Mimosa pudica*. Asterisk: Sequenced for this study. Cucurb. Cucurbitales. INF: genes required for infection. NOD: genes involved in nodule organogenesis and regulation. CSG: genes required for both NFN symbiosis and arbuscular mycorrhiza symbiosis.

Science

Phylogenomics reveals multiple losses of nitrogen-fixing root nodule symbiosis

Maximilian Griesmann, Yue Chang, Xin Liu, Yue Song, Georg Haberer, Matthew B. Crook, Benjamin Billault-Penneteau, Dominique Lauressergues, Jean Keller, Leandro Imanishi, Yuda Purwana Roswanjaya, Wouter Kohlen, Petar Pujic, Kai Battenberg, Nicole Alloisio, Yuhu Liang, Henk Hilhorst, Marco G. Salgado, Valerie Hocher, Hassen Gherbi, Sergio Svistoonoff, Jeff J. Doyle, Shixu He, Yan Xu, Shanyun Xu, Jing Qu, Qiang Gao, Xiaodong Fang, Yuan Fu, Philippe Normand, Alison M. Berry, Luis G. Wall, Jean-Michel Ané, Katharina Pawlowski, Xun Xu, Huanming Yang, Manuel Spannagl, Klaus F. X. Mayer, Gane Ka-Shu Wong, Martin Parniske, Pierre-Marc Delaux and Shifeng Cheng

published online May 24, 2018

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005. 2017 © The Authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original U.S. Government Works. The title *Science* is a registered trademark of AAAS.