Comparison of RWP-RK transcription factors between Arabidopsis and legumes uncovers a relationship between nitrogen signaling and nodulation

Zhihua Wu
South-Central University for Nationalities
Hong Liu
South-Central University for Nationalities
Wen Huang
South-Central University for Nationalities
Lisha Yi
South-Central University for Nationalities
Xiufang Li
South-Central University for Nationalities
Erdai Qin
South-Central University for Nationalities
Yarui Li
South-Central University for Nationalities
Tiange Yang
South-Central University for Nationalities
Yuxiang Zhang
South-Central University for Nationalities
Can An
South-Central University for Nationalities
Yuqin Gan
South-Central University for Nationalities
Xinrong Wang
South-Central University for Nationalities
Jing Wang
Chinese Academy of Agricultural Sciences
Rui Qin (qinrui@scuec.edu.cn)
South-Central University for Nationalities https://orcid.org/0000-0003-4093-7044

Research article

Keywords: RWP-RK, nitrogen-fixation clade, coexpression network, nitrate signaling, nodulation

Posted Date: November 27th, 2019

DOI: https://doi.org/10.21203/rs.2.17800/v1

Abstract

Background Nitrogen, as a constituent of amino acids and nucleic acids, is an essential macronutrient for all living organisms. The nitrogen-fixation clade (NFC) is a clade, consisting of Fabales, Fagales, Cucurbitales, and Rosales, where all nodulating plants have been originated. The plant-specific RWP-RK family of transcription factors are involved in nitrate responses and play specific roles in nodule inception. In the present study, by investigation of RWP-RKs at genome-wide level and comparative coexpression networks, the roles of RWP-RKs involved in nitrate response and nodulation were analyzed to reveal evolution of RWP-RKs and a relationship between nitrogen signaling and nodulation.

Results Here, we systematically investigated 292 RWP-RKs from 26 species of legumes and non-legumes of NFC by whole-genomic analysis and characterized their evolutionary relationships, protein motifs, and gene structures. We compared RWP-RK networks from *Arabidopsis thaliana* under N-starvation and N-supplementation conditions, as well as transcriptome atlases from *Phaseolus vulgaris* and *Glycine max*. This revealed that N starvation, which is essential for nodulation, alters the connectivity of RWP-RKs to other genes, including symbiosis-related genes. Meanwhile, appropriately low concentrations of nitrates stimulate nodulation by regulating RWP-RK expression in *P. vulgaris*.

Conclusions Our comparative evolutionary analysis of RWP-RKs between *A. thaliana* and legumes revealed the evolutionary features and the relationship between the nitrate signaling pathway in a model organism and nodulation in legumes.

Background

Much effort has been made to improve N use efficiency (NUE) in crop plants to allow high-yielding crops to be grown with low N input without significant yield losses [1]. NUE can be improved via the interaction of plants with microorganisms, such as rhizobia. When legumes interact with rhizobia, newly formed root nodules (referred to hereafter as nodules) use nitrogen from the atmosphere via symbiotic nitrogen fixation [2]. Many studies have explored the mechanism of symbiotic nitrogen fixation with the goal of engineering plants that can directly fix nitrogen for crop improvement [3].

RWP-RKs, which contain a conserved RWP-RK DNA binding motif, are a class of transcription factors (TFs) that control N uptake efficiency and N utilization by sensing nitrate signals. These plant-specific TFs are grouped into two subfamilies: Nodule inception (NIN)-like proteins (NLPs) and RWP-RK-domain proteins (RKDs). In addition to sensing nitrate signals in non-nodule-forming Arabidopsis thaliana [4] and rice (Oryza sativa) [5], RWP-RKs have specifically evolved with the TF NIN (which is involved in nodule inception) in nodule-forming plants, such as Medicago truncatula [6, 7]. Besides the common DNA binding RWP-RK domain, NLPs contain an additional domain known as PB1 (Phox and Bem 1) at their C-termini that allows interactions with additional proteins [8]. For example, NLP1 functions through PB1-mediated interactions with NIN, leading to the suppression of its target gene, CRE1, in Medicago [6]. In A.

thaliana, NLPs play a central role in nitrate signaling by binding to the nitrate-responsive cis-elements in their target genes [9].

NIN, an RWP-RK family member, is essential for nodulation in legumes and non-legumes of the NFC [7, 10]. Nodule formation is dependent on the perception of limited N levels by the plant [11, 12]. The nitrate signaling pathway and nodulation appear to be involved in controlling processes regulated by RWP-RKs. Some NFC members lack NIN and do not form nodules. To date, the RWP-RK family in the NFC has not been subjected to detailed analysis. The recent sequencing of multiple genomes of legumes and non-legumes capable or incapable of undergoing nodulation has paved the way for comparisons of the RWP-RK family across species [7].

A. thaliana as an ideal model dicot, our understanding of the roles of RWP-RKs in nitrate signaling in A. thaliana could facilitate the improvement of NUE in crops via evolutionary and comparative analysis [13, 14]. The construction of coexpression network based on correlated gene expression patterns is a valuable tool for revealing sets of genes that function in specific biological processes[15]. Therefore, combined coexpression network and gene evolution analysis of both model plants and non-model crops should provide new insights into the relationship between the nitrate signaling pathway and symbiosis mediated by RWP-RK family members.

In this study, we constructed a phylogenetic tree of 26 plant species based on genome-wide gene duplication events to explore the evolutionary features of the NFC. We identified 292 RWP-RK family members in the 26 species covering five orders (Fabales, Rosales, Cucurbitales, Fagales, and Brassicales), including 25 species of four orders belonging to the NFC, using A. thaliana (Brassicales, non-NFC) as the outgroup (Table S1). We analyzed the physicochemical properties, genomic positions, gene structures, protein motifs, and phylogenetic relationships of these genes. We then constructed comparative coexpression networks of A. thaliana under N-starvation and N-supplementation conditions, as well as coexpression networks of transcriptome atlases from Phaseolus vulgaris and Glycine max containing nodules, to uncover the relationships among RWP-RKs involved in nitrate responses and symbiosis. Quantitative RT-PCR analysis of P. vulgaris under nitrogen-free, low-nitrogen, and high-nitrogen conditions shed light on the diverse roles of these RWP-RKs in nitrate signaling and nodulation.

Results

Phylogeny of NFCs and distribution of RWP-RKs

To learn about the distribution of RWP-RKs across the NFC, based on gene duplication events, we built a phylogenetic tree with high bootstrap values for 26 species from three clades, defined as clade I (Brassicales), clade II (Fagales, Cucurbitales, and Rosales), and clade III (Fabales), with *A. thaliana* as the outgroup. Of the 25 species from the NFC, 21 form nodules, whereas the four remaining species lack the capacity for nodulation **(Fig. 1A)**. A total of 292 RWP-RK domains and 156 PB1 domains has been identified in these analyzed species **(Table S1)**. RWP-RKs appear to be randomly distributed in each order, indicating that genome-wide gene duplication events did not contribute to the expansion of RWP-RKs in

specific orders. Moreover, it appears that the RWP-RKs did not expand in nodulated plants **(Fig. 1A and Fig. S1)**. For example, *Begonia fuchsioides* (which lacks nodules) contains the second-highest number (22) of RWP-RKs after *G. max*, and *A. thaliana* contains 14 RWP-RKs, more than many nodule-forming plants. These results suggest that the expansion of RWP-RKs involved in the nitrate signaling pathway and nodule inception might represent an adaptive response of plants to diverse nitrogen conditions. Compared to RKDs, NLPs (such as NIN, which is specifically required for nodulation in the NFC) contain an additional PB1 domain, which increases their interaction with other proteins [33, 34]. Despite the similar median values for each domain number between nodule-forming and non-nodule-forming plants, the numbers of RWP-RK and PB1 domains are more variable in non-nodulating than in nodulating plants, except for the outlier, large number of RWP-RK in *G. max* (possibly because of its two genome-wide duplications) **(Fig. 1B).** We speculate that non-nodulating plants require more RWP-RK and PB1 domains to improve NUE due to their inability to fix nitrogen from atmospheric inorganic nitrogen.

We also analyzed the consensus motifs of both the RWP-RK and PB1 domains within each species. The consensus motifs could not be analyzed in some species with few instances of these domains, for example for both the RWP-RK and PB1 domains in *Lupinus angustifolius* and for PB1 in *Vigna angularis* and *Lotus japonicus* **(Table S1)**. A few species have large insertions in their RWP-RK domains, such as *A. thaliana* AT2G43500 (AtNLP8) and *Discaria trinervis* Distr2169S18482. In addition to conserved motifs commonly existed in almost RWP-RKs, similar insertions also occur in PB1 domains to some extent **(Fig. S2)**.

Analysis of RWP-RK and PB1 domains in the NFC

Despite the important roles of RWP-RKs in both nodulating and non-nodulating plants [7, 9, 10], the features of these proteins within the NFC, such as protein motifs, gene structures, and evolutionary relationships, have not been investigated in detail. We therefore analyzed the consensus domains of RWP-RK and PB1 in nodule-forming and non-nodule-forming plants separately. These domains are strongly conserved but contain some modified insertions at different positions **(Fig. 2A)**. Alignment of the domains revealed that the 49^{th} site (K, Lys) and 63^{th} site (R, Arg) are conserved across all RWP-RKs **(Fig. S3)**, suggesting that these sites are vital for RWP-RK activity. Except for *Arachis ipaensis* Araip.KR88K, *B. fuchsioides* Begfu255S10103, *B. fuchsioides* Begfu91828S44750, which have lost the RWP-RK domain, all species contain the conserved RWPxRK signature, with some modifications, such as HWP-RK (HWPHRK of Araip.YWB61, Fabales), NWP-RK (NWPHRK of C.cajan_36762, Fabales), WWP-RK (WWPYRK of Datgl206S24145, Cucurbitales), HWP-RK (HWPSRK of Datgl229S25120, Cucurbitales), KWP-RQ (KWPHRQ of Glyma.04G054800.1.p, Fabales), and KWP-RK (KWPQRK of Glyma.06G054900.1.p) **(Fig. S3)**.

Interestingly, almost all of these modifications occurred in the first amino acid of the conserved RWP-RK motif in nodule-forming Fabales and Cucurbitales plants. Family members with losses or modifications

of RWP-RK might be nonfunctional or might have divergent function. Some proteins (Araip.16Y1B, Araip.6QW2Y, Araip.G0MMI, Vang0010ss03040.1, Vang01g00130.1, Vradi06g00060.1) containing only PB1 domains but lacking RWP-RK domains might have different origins or play different roles and were therefore excluded from our analysis. The 4^{th} K, 13^{th} R, 62^{th} D, and 75^{th} D sites of the PB1 domains of RWP-RKs are conserved across nodulating and non-nodulating plants **(Fig. S4)**.

To explore the possible origins of all the identified RWP-RKs among the 26 species, we grouped the RWP-RKs into different orthogroups. An orthogroup is defined as a set of genes that descended from a single gene in the last common ancestor of the analyzed species [17]. Except for one *A. thaliana* RWP-RK (AT4G38340) and six *Arachis duranensis* RWP-RKs (Aradu.I1BME, Aradu.TG0QF, Araip.377BK, Araip.6M6N8, Araip.YB35N, Araip.YWB61), which lack orthogroups (and are possibly orphan genes), the 285 remaining RWP-RKs clustered into 11 orthogroups containing 1–69 RWP-RKs genes, suggesting that these genes have diverse origins **(Fig. 2B and Table S3)**.

We predicted the length, isoelectric point, and subcellular localization of each RWP-RK using various databases [24, 35]. The isoelectric points of the 292 RWP-RKs range from 4.73 for Drydr146S16094 to 10.6 for Aradu.I1BME, and the molecular weights range from 7.14 kDa for Aradu.I1BME to 156.01 kDa for Aradu.G4SB3. Most proteins were predicted to localize to the nucleus **(Table S4)**.

Phylogeny and characteristics of the RWP-RKs

To investigate the phylogenetic relationships among RWP-RK family members in the 26 species, we constructed a phylogenetic tree of 292 RWP-RKs via the neighbor-joining method with 1000 bootstrap values. To limit issues related to high divergence between proteins, we selected the RWP-RK domains with 30 additional amino acids upstream and downstream of the domains for alignment and phylogenetic analysis. The tree formed six clades based on the relationships of the NLPs and RKDs in *A. thaliana*. The NLP subfamily clustered into a single clade with three subclades, NLP-1 (AtNLP1, AtNLP5), NLP-2 (AtNLP6, AtNLP7), and NLP-3 (AtNLP8, AtNLP9), containing all NLPs with PB1 domains. The RKD subfamily includes RKD-1 (AtRKD4, AtRKD5), RKD-2 (AtRKD1, AtRKD2, AtRKD3), and RKD-3 **(Fig. 3A)**. The NLP subfamily also includes several non-NLPs without PB1 domains (Glyma.06G000400.1, vigan.Vang06g08410.1, Araip.YWB61, Araip.YWB61, vigan.Vang02g05230.1, Araip.377BK, and Araip.5C6JK), perhaps due to partial gene duplications of the NLP genes [8].

To learn more about the features of each clade of RWP-RKs in the phylogenetic tree, we investigated their gene structures using GSDS 2.0 [36] and predicted additional conserved motifs using MEME 5.0.1 [25]. Due to the lack of exon information in the annotated gff3 file of *Trifolium subterraneum* (with 13 RWP-RKs), we omitted those data and used the 279 remaining RWP-RKs for analysis. The average number of exons in the NLP subfamily and RKD subfamily is 4.9 (ranging from 2–13) and 4.0 (1–9), respectively. The number of exons in each subclade of NLP subfamily members is 4.1 for NLP-1 (2–11), 5.3 for NLP-2 (3–8), and 5.6 for NLP-3 (4–13). The number of exons in each subclade of RKDs is 3.7 for RKD-1 (1–6),

5.6 for RKD-2 (2–9), and 3.5 for RKD-3 (3–5) **(Fig. 3B and Table S5)**. Therefore, the average number of exons is higher in the NLPs than in the RKDs.

Of the 50 consensus motifs predicted by MEME, in addition to the three common motifs (including known RWP-RK and PB1 motifs) in both the RKDs and NLPs, the NLP subfamily contains 39 enriched motifs, whereas the RKD subfamily contains only 8 enriched motifs. Each subclade contains some unique motifs, pointing to the diverse roles of RWP-RKs, such as motif #24 in NLP-1, motif #18 in NLP-2, and motif #23 in NLP-3. Based on predictions from NetPhos 3.1 Server [37], motif #24 contains a predicted phosphorylation site (Y) with a score of 0.89; motif #18 contains a predicted phosphorylation site (S) with a score of 1.00; and motif #23 contains multiple predicted phosphorylation sites (S) with high scores. In addition, members of RKD-3 (only in the NFC) contain several unique motifs, including motif #34, motif #37, and motif #41 **(Fig. S5, Fig. S6 and Table S6)**. These unique motifs in each subclade may increase the number of specific interactions with other proteins.

Statistical analysis of intron phases, that is whether the intron disrupts a codon, revealed that RWP-RKs contain the most phase 0 introns, which cause no disruption of a codon), followed by phase 2 and phase 1, which disrupt the codon between bases 2 and 3 or bases 1 and 2, respectively. However, RKDs contain more phase 1 introns than NLPs with phase 1 introns **(Fig. 3C)**. Overall, the results of exon number and intron phase analyses were consistent with the phylogenetic analysis, but there were some exceptions. For example, *Cerca58S27147* has more short additional exons than other members of the NLP-1 clade but lacks additional protein motifs, suggesting that additional exons of this gene might play a regulatory role in *Cercis canadensis* **(Fig. S6)**.

Comparison of RWP-RKs in non-nodulating vs. nodulating plants

a. *thaliana* has many advantages to study interactions between diazotrophic bacteria and dicots [14], and therefore the knowledge of *A. thaliana* is helpful to translate biological knowledge from model organisms to crops. To explore the relationship of RWP-RKs in nodulating and non-nodulating plants, we select *A. thaliana* (non-nodulation), *G. max* (nodulation), and *P. vulgaris* (nodulation) for subsequently comparative analysis. We classified the RWP-RKs from *A. thaliana* (14), *G. max* (28), and *P. vulgaris* (12) into three clades: unique RKDs in *G. max* and *P. vulgaris* (Clade I), RKDs in all three species (Clade II, including AtRKD1–AtRKD5), and NLPs in all three species (Clade III, including AtNLP1–AtNLP9). Analysis of gene structure revealed that Clade III genes contain many more exons than those of clade I, which is consistent with the finding that additional protein motifs (such as the important PB1 domain) are essential for the functioning of NLPs (**Fig. 4A, Fig. S6 and Table S6**). In addition to motif 1 and motif 3 (motif RWPxRK), which are present in all three clades, clade II contains a unique motif (#10), indicating that this motif plays a unique role in the clade. Based on the phylogenetic analysis and the presence of protein motifs in different positions, clade III was classified into three subclades: subclades IIIa (including AtNLP6 and AtNLP7), IIIb (including AtNLP8 and AtNLP9), and IIIc (including AtNLP1–AtNLP5). Almost all members of subclade IIIa contain a

unique motif (#21) in their centers and motif #25 and motif #18 at their C termini; subclade IIIb members contain a unique motif (#23) at their C termini; and subclade IIIc members contain a unique motif (#24) at their N termini **(Fig. 4A and Fig. S5)**. The additions and deletions in genes at different positions within clade III might have given rise to different functions, such as different cellular localizations and protein interactions.

Despite the important roles of nitrates in plant growth, N limitation, including N starvation or low N, is essential for nodulation in legumes [11, 12]. To explore the relationship between N limitation and nodulation, we integrated time-series transcriptome datasets from *A. thaliana*, including root samples treated with KCl (defined as N starvation) and KNO_3 (defined as N supplementation) [38], as well as expression atlases from *G. max* and *P. vulgaris* [39, 40]. In *A. thaliana*, all *AtRKDs* are not expressed in roots, whereas *AtNLPs* are differentially regulated. Specifically, *AtNLP1* and *AtNLP3* are downregulated and other *AtNLPs* are upregulated, indicating that the different NLPs have different effects on plant responses to N starvation **(Fig. 4B)**.

NLP genes are expressed in different tissues, with some specifically expressed in the nodules of *G. max* and *P. vulgaris* (*Glyma.02G311000*, *Glyma.06G000400*, *Glyma.04G000600*, *Glyma.14G001600*, *Phvul.008G291800*, and *Phvul.009G115800*). These nodule-specific genes clustered together and are closely related to *AtNLP*1 and *AtNLP*2 within subclade IIIc **(Fig. 4A, C, D)**. However, *AtNLP1* and *AtNLP2* showed opposite fold changes in response to N starvation, indicating that regulatory elements have rapidly evolved. Taking the evolutionary relationships and expression patterns of model dicot and legume crops together, we conclude that compared to subclades IIIa and IIIb, the acquisition of additional protein motifs in subclade IIIc has provided the prerequisite for nodule inception and that the further evolution of regulatory elements within IIIc will be crucial for the development of this process.

The connected genes of *AtNLPs* in coexpression networks provide the prerequisite for nodulation

Our results demonstrate that *AtNLPs* are widely involved in plant responses to N starvation. Weighted correlation network analysis (WGCNA) is a powerful tool for dividing genes with correlated expression patterns into different modules with biological significance [15, 27, 41]. To investigate the possible relationship between N starvation and nodulation, we explored differences in the modules containing *AtNLPs* under N starvation vs. N supplementation. We used WGCNA to construct a weighted network from N-starvation datasets with rlog normalization **(Fig. S7)** implemented in the DESeq2 package [31]. Expression correlation (rho = 0.99) is more conserved than connectivity correlation (rho = 0.84) under N starvation **(Fig. S8)**, suggesting that changes in connected neighbors play more important roles than changes in expression levels under N starvation. *AtNLPs* were present in five of the 12 modules **(Fig. 5A–E)**. Different modules showed different expression patterns across a series of time points, indicating that their component genes play different roles under N starvation.

GO enrichment analysis of genes in the biological process category revealed no commonly enriched biological process in the five modules **(Fig. 5F and Table S7)**. In the purple module (with *AtNLP2*, *AtNLP4* and *AtNLP9*), in which genes were continuously upregulated after 20 min under the treatment of N starvation, we detected a high proportion of unique GO terms associated with transport (20/54), such as “calcium ion transport” (GO:0006816) and “calcium ion transmembrane transport” (GO:0070588), which may be related to “calcium spiking”, a symbiotic signaling event. Genes in the blue module (with *AtNLP3*) were relatively upregulated at the early stage (10–15 min of treatment). We discovered uniquely enriched GO terms in the blue module, such as “endocytosis” (GO:0006897). The unique GO terms in the green module (with *AtNLP6* and *AtNLP7*) included many important terms, such as “cell wall organization or biogenesis” (GO:0071554), “cell wall organization” (GO:0071555), “plant-type cell wall organization or biogenesis”(GO:0071669), “cell wall modification” (GO:0042545), “root hair elongation” (GO:0048767), “root hair cell development” (GO:0080147), and “auxin-activated signaling pathway”(GO:0009734). Interestingly, the blue and green modules were both enriched in “symbiosis, encompassing mutualism through parasitism” (GO:0044403), with 36 and 24 genes, respectively **(Table S7)**. Although genes in both modules were highly expressed only at 10–15 min under the treatment of N starvation, the genes enriched for the GO term “symbiosis, encompassing mutualism through parasitism” did not overlap, highlighting the independent roles of genes regulated by AtNLP3, AtNLP6, and AtNLP7 within each module. Also, this result suggested that response to N starvation at early stage may be essential for symbiosis.

The connectivity value of *AtNLP3* within the blue module was 0.32 under N-starvation conditions and 0.08 under N-supplementation conditions. *AtNLP3* was positively correlated with various pathogen/virus-related genes, such as *AT5G03210* (*ATDIP2*, resistance to virus), *AT5G08790* (*ANAC081*, may function as a repressor of pathogenesis-related proteins), *AT1G32400* (*TOM2A*, viral replication complex formation and maintenance), *AT5G48160* (*OBE2*, transport of virus in host), *AT5G10270* (CDKC;1, response to virus), *AT5G42950* (*EXA1*, defense response to virus), *AT3G11650* (NHL2,defense response to virus), *AT5G23570* (*ATSGS3*, defense response to virus), *AT2G25620* (*ATDBP1*, regulation of defense response to virus), *AT2G23350* (*PAB4*, viral process), *AT1G60800* (*AtNIK3*, defense response), and *AT3G04720* (*AtPR4*, defense response to bacterium).

Genes positively correlated with *AtNLP6/AtNLP7* also included several pathogen/virus-related genes, such as *AT3G11660* (*NHL1*, defense response to virus), *AT5G06320* (*NHL3*, defense response to bacterium), *AT2G35980* (*ATNHL10*, defense response to virus), *AT4G13350* (*NIG*, response to virus), *AT1G70690* (*HWI1*, defense response to bacterium), *AT3G60240* (*CUM2*, response to virus), and *AT5G04430* (*BTR1*, regulation by virus of viral protein levels in host cell). Correlation analysis showed that *AtNLP3*, *AtNLP6*, and *AtNLP7* were more highly correlated with genes in GO categories GO:0044403 (symbiosis, encompassing mutualism through parasitism) and GO:0009267 (cell response to starvation) under N-starvation than under N-supplementation conditions **(Fig. 6A,B,C)**. In addition, *AtNLP3* was downregulated under N starvation compared to N supplementation **(Fig. 5B)**. The downregulation of *AtNLP3*, along with the downregulation of immune response genes, may be beneficial for rhizobial invasion via reducing the immune response. The expression levels of *AtNLP6* and *AtNLP7*, which have

high sequence similarity, changed little (fold change < 2) between N-starvation and N-supplementation conditions. AtNLP7 is a major regulatory element among NLP proteins [42]. Despite their weak upregulation, *AtNLP6* and *AtNLP7* might function primarily at the protein level rather than the transcriptional level, as previously reported [42].

We constructed a protein association network of AtNLP3, AtNLP6, and AtNLP7 to genes using STRING 11.0 **(Fig. 6C,D)**. While these proteins were connected to genes from GO term GO:0044403 (symbiosis, encompassing mutualism through parasitism), there was no significant association between these genes and other proteins. Additional experiments should be carried out to investigate the interactions of these highly connected genes, such as *AT4G13350* (response to virus, connectivity value of 0.80 for AtNLP6 and 0.90 for AtNLP7 under N starvation, 0.26 for AtNLP6 and 0.35 for AtNLP7 under N supplementation). A set of genes from GO:0009267 (cell response to starvation), including *AT5G45380* (*ATDUR3*), *AT4G35090* (*CAT2*), *AT3G05630* (*PDLZ2*), *AT1G20620* (*ATCAT3*), and *AT1G20630* (*CAT1*), were also enriched in GO:0006995 (cellular response to nitrogen starvation). Moreover, AtNLP6 and AtNLP7 were both associated with AT1G13300 (NIGT1/HRS1), which is regulated by AtNLP7 [43]. Although AtNLP6 and AtNLP7 might play redundant roles [44], greater differences in correlation were observed between AtNLP6 and NIGT1/HRS1 than between AtNLP6 and NIGT1/HRS1 under N-starvation and N-supplementation conditions. *NIGT1/HRS1* is induced by NO_3^- [43]. The higher correlation between AtNLP6/AtNLP7 and NIGT1/HRS1 and the higher connectivity of AtNLP6 under N starvation (0.62 under N starvation vs. 0.16 under N supplementation), and AtNLP7 (0.47 vs. 0.11) and NIGT1/HRS1 (0.41 vs. 0.06) in the green module **(Table S8),** indicate that these genes strongly interact under N starvation conditions.

Most NLPs (*AtNLP2– AtNLP7*), especially *AtNLP5* from the turquoise module, which were significantly upregulated under N starvation **(Fig. 4B)**, showed higher correlations to more genes under N starvation than under N supplementation (absolute pcc ≥ 0.8) **(Fig. S9)**. *AtNLP5* (in the turquoise module) was significantly upregulated after 20 min of N starvation, whereas *AtNLP3* (in the blue module) was significantly downregulated **(Fig. 5C,E)**; these opposite expression patterns indicate that AtNLPs play diverse roles under N starvation. *AtNLP1*, *AtNLP8*, and *AtNLP9* without strong correlation to any other genes (absolute pcc≥ 0.8) under N supplementation (defined as lonely expressed genes) showed differential expression patterns under N starvation vs. N supplementation. The expression patterns of these genes were not significantly correlated to those of other genes under N starvation, indicating that they have lost many connected neighbors.

Among the genes most closely related to nodule-specific *NLPs* of *G. max* and *P. vulgaris*, *AtNLP1* and *AtNLP2* had opposite expression patterns. *AtNLP2* has many connected neighbors, whereas *AtNLP1* has completely lost highly connected neighbors **(Fig. 4 and Fig. S9)**. *AtNLP2* reflects the evolutionary imprint of nodule-specific *NLP* genes. Taking the expression patterns of *AtNLP6*/*AtNLP7* together, these results suggest that the differential connectivities of the *NLPs*, along with their differential expression patterns, may be essential for nodulation under N starvation. The association of genes with *NLP* genes under N

starvation, which are related biological processes of symbiosis, cell cycle, cell wall organization or biogenesis, and calcium ion transport, might have paved the way for nodulation in the NFC.

Genes regulated under N starvation and nodulation are connected to multiple transcription factors, transcriptional regulators, and protein kinases

TFs, transcriptional regulators (TRs), and protein kinases (PKs) are important classes of regulatory proteins associated with numerous aspects of plant growth and development, as well as biotic and abiotic stress responses [45]. To further explore the relationships of NLPs to other TFs, TRs, and PKs under the N-starvation conditions, N supplementation in *A. thaliana*, and in transcriptome atlaes of *G. max* and *P. vulgaris*, we analyzed sub-networks of the NLPs connected to these regulatory proteins with absolute pcc ≥ 0.8. Many more TFs, TRs, and PKs were highly correlated to NLPs under N starvation (1,371 with 674 TFs, 199 TRs, and 498 PKs, with 397,777 edges) vs. N supplementation (288 with 166 TFs, 29 TRs, and 93 PKs, with 7,298 edges), indicating that N starvation strongly influences the TFs, TRs, and PKs connected to NLPs at the transcriptional level **(Fig. 7A,B and Table S10)**.

The highly correlated TFs, TRs, and PKs were enriched in both the same and unique GO terms. For example, of the 577 TFs highly correlated only to AtNLPs under N starvation (pcc ≥ 0.8), 502 were enriched in “regulation of nitrogen compound metabolic process” (GO:0051171). In addition, 109 of 182 TRs were enriched in “regulation of nitrogen compound metabolic process” (GO:0051171) **(Fig. 7C,D and Table S11)**. The highly correlated PKs were enriched in “protein phosphorylation” (GO:0006468), “cell communication” (GO:0007154), and “cell surface receptor signaling pathway” (GO:0007166), indicating that N starvation influences plant cell–cell signaling pathways mediated by NLPs **(Fig. 7E and Table S11)**.

To further explore the possible roles of RWP-RKs in nodule-forming plants, we constructed networks for *G. max* and *P. vulgaris*. *P. vulgaris* contains 9 RWP-RKs, including 7 NLPs and 2 RKDs, which are highly correlated to 130 regulators, with 1,639 edges. *G. max* contains 16 RWP-RKs, including 14 NLPs and 2 RKDs, which are highly correlated to 270 regulators, with 4,113 edges **(Fig. 8A,B and Table S12)**. Recent whole-genomic duplication event within *G.max* affecting RWP-RKs have increased the number of edges of RWP-RKs connected to gene regulators compared to *P. vulgaris*. Among these RWP-RKs, *Phvul.009G115800* and *Phvul.008G291800* in *P. vulgaris* and *Glyma.04G000600*, *Glyma.02G311000*, *Glyma.14G001600*, and *Glyma.06G000400* in *G. max*, which are *NIN* (nodule inception) genes [6], were highly expressed in separate nodules and are located in separate modules, pointing to their overall upregulation in nodule tissue **(Fig. 8C, D and Table S12)**. These nodule-specific NLPs clustered together in clade IIIc.

Two other genes in *G. max*, *Glyma.13G346300* and *Glyma.12G050100*, which are closely related to *AtNLP8* and *AtNLP9* within clade IIIb, were highly expressed in nodules. Interestingly, *Phvul.005G15510* in clade IIIb was expressed at higher levels in ineffective N-fixation nodules than in effective N-fixation

nodules. This observation suggests that the regulation of NLPs from clade IIIb is also influenced by rhizobia without the capacity for N fixation and that NLPs from clade IIIc are critical for N-fixing nodules.

We also detected divergence within IIIc: *Phvul.009G01120*, together with *Glyma.04G017400* and *Glyma.06G017800*, which are closely related to *AtNLP4* and *AtNLP5*, also showed the highest expression levels in ineffective N-fixation nodules. Only genes in the small clade closely related to *AtNLP1* and *AtNLP2* were specifically upregulated in effective N-fixation nodules in both *G. max* and *P. vulgaris* **(Fig. 4)**. These findings reflect the functional divergence of NLPs with regard to nodulation, that is, the existence of nodulation with and without N fixation.

Effect of nitrogen on nodulation via the regulation of *NLPs* in *P. vulgaris*

Not only are NLPs involved in the nitrate signaling pathway, but they also influence nodule inception [6]. Whether nitrate influences nodulation via the regulation of NLPs mediating nitrate signaling in *P. vulgaris* has been unclear. In plants inoculated with *Rhizobium tropici* treated with different concentrations of nitrate, 10 mM nitrate inhibited nodulation in both early and mature nodules. Mature nodules from plants treated with 5 mM nitrate appeared to be browner, and plant height was taller, than for plants treated with either 0 mM or 10 mM nitrate **(Fig. 9A, B, C and Fig. S10)**.

Expression analysis of six *PvNLP* genes in early roots and root-nodule mixtures under different concentrations of nitrates indicated that *Phvul.004G114100*, *Phvul.008G291800*, and *Phvul.011G052100* showed higher expression at 5 mM nitrate (low-nitrogen conditions) with inoculation than they did under either 0 mM nitrate (nitrogen-free conditions) or 10 mM nitrate (high-nitrogen conditions) regardless of inoculation **(Fig. 10A)**. When the roots were treated with different concentration of nitrates without inoculation, *Phvul.009G011200* and *Phvul.009G115800* showed highest expression under low-nitrogen as compared to both nitrogen-free and high-nitrogen conditions, whereas the other genes showed gradual inhibition with increasing nitrogen concentration. Finally, *Phvul.008G291800*, *Phvul.009G011200*, and *Phvul.011G052100* were significantly inhibited under high-nitrogen conditions **(Fig. 10B)**. In mature nodules, *Phvul.007G071900* was significantly upregulated under low-nitrogen vs. nitrogen-free conditions, whereas other genes were downregulated or varied in expression pattern **(Fig. 10C)**.

Discussion

Nitrogen (N) is an essential element for crop growth, productivity, and grain quality. N is also a central component of DNA, RNA, proteins, and various metabolic compounds [1]. To biosynthesize the precursors of these biomacromolecules, plants must integrate inorganic carbon (CO_2) and nitrogen (NO_3^-/NH_4^+); however, compared to CO_2, NO_3^-/NH_4^+ is more difficult to obtain from the environment. Therefore, plant growth is generally nitrogen limited in both natural and agricultural environments [46]. The extensive use of synthetic fertilizers in developed countries is expensive and environmentally damaging. By contrast, in developing countries, the lack of fertilizer causes low crop yields, leading to hunger and malnutrition. NLPs of RWP-RKs play important roles in nitrate signaling [4, 9, 47]. Meanwhile, the NFC is a clade

originated at 100 million years ago, in which NIN (a member of RWP-RKs) is reported to be essential for nodulation [7, 48]. Therefore, comparative research on RWP-RKs in *A. thaliana* and NFC which involve in both nitrogen signaling and nodulation is crucial for improving sustainable agriculture and human health now and in the future.

Distribution and features of RWP-RKs in NFC

In contrast to animals, the success of angiosperms is partially attributed to innovations caused by gene or whole-genome duplications [49]. An accurate phylogenetic tree of species is required for evolutionary comparisons of RWP-RKs. It is impossible to deduce a phylogenetic tree from one-to-one corresponding orthologues in that few of these orthogroups in 26 species contain just one orthologue from each species. Therefore, we used a newly developed method for phylogenetic tree construction based on the analysis of gene duplications for all genes [50]. In the phylogenetic tree (**Fig. 1A**), species from separate orders clustered together, regardless of their ability to undergo nodulation. These results suggest that the genome-wide duplication of genes is not the direct driver of the evolution of nodulation across the NFC. This is also supported by the finding that the evolutionary relationship between polyploidy and nodulation is not sufficient to make a species able to form nodules [51]. The presence of RWP-RKs with multiple origins **(Fig. 1B)** might help these species adapt their growth and metabolism in response to fluctuations in nitrogen availability in different habitats, as legumes are exceptionally ecologically diverse [52, 53].

In both RKDs and NLPs, there are conservative and non-conservative amino acid sites (**Fig.2A**, **Fig.S3 and Fig.S4**), which appears to be not relevant to status of nodulation across the species of NFC. However, the 49^{th} site (K, Lys) and 63^{th} site (R, Arg) are conserved across all RWK-RDs, suggesting that these sites are vital for RWP-RK activity. Also, completely conserved sites (4^{th} K, 13^{th} R, 62^{th} D and 75^{th} D) in PB1 domain is conserved in animals, fungi, amoebas, and plants indicating in diverse biological processes beside plant-specific nitrate signaling and nodulation [54]. In addition, variation of motifs of both the RWP-RK and PB1 domains caused by insertion/deletion may provide novel regulatory function for different RWP-RK members. For example, AtNLP8 with insertion in the RWP-RK domain functions as a master regulator of nitrate-promoted seed germination and is activated by nitrate via a mechanism different from AtNLP7 without insertion [4]. Whether this insertion within RWP-RK of AtNLP8 is involved in this specific mechanism (as is the case for AtNLP6) remains to be explored. Variations in the domain lengths of RWP-RK and PB1 and specific amino acid sites (resulting in noncanonical domains) might mediate noncanonical interactions during various biological events [54]. The diversity of RWP-RKs within species indicates that these proteins play diverse roles in plant responses to nitrates.

Phylogeny of all RWP-RKs from 26 species based on protein sequences showed six subclades into two clades (RKDs and NLPs) (**Fig.3A**). The variations in exon number, leading to more protein motifs in NLPs than in those in RKDs (**Fig.3B**), highlight the functional novelty of the RWP-RK subclades, as these

variations are usually accompanied by the additions/deletions of functional motifs. Meanwhile, high numbers of exons increase the chances for alternative splicing of precursor mRNAs, which is important for gene regulation [55]. In addition, analysis of intron phases of the six subclades (**Fig.3C**) suggested that generally number of exons with phase 0 introns for RWP-RKs is higher than those with phase 0 and phase 1 introns. There is an excess of phase 0 introns also found in other cases, supported by prediction of the exon theory of genes [56]. However, the RKDs has more exons with phase 1 introns than NLPs with phase 1 introns, which indicates different origins of intron-exon structures for RKDs and NLPs, respectively [57].

Effects of N-starvation on nodulation mediated by NLPs

Over the past few decades, the model dicot *A. thaliana* has become an excellent model for studying symbiosis in non-nodulating plants [14]. The most important nodule-bearing legume crops such as *G. max* and *P. vulgaris* have not been studied in as much detail as *A. thaliana*. RWP-RKs play important roles in both nitrate responses and nodule inception, and they interact with each other to coordinate nitrate signaling and nodulation [6]. Therefore, we felt that a comparison of RWP-RK expression and homology in *A. thaliana* vs. legume crops (*G. max* and *P. vulgaris*) would allow gene functional analyses in model organisms to be applied to nodule-forming crops. Such studies might also facilitate the transfer of the nitrogen-fixation trait into non-nodulating plants to improve NUE [3, 58]. Interestingly, *AtRKD1– AtRKD4* are highly expressed in reproductive organs, and *AtRKD5* has pleiotropic effects on phytohormone pathways, highlighting the regulatory importance of *AtRKDs* in female gametophyte development [59, 60]. Finally, almost RKDs are absent from the separated transcriptome atlas of *G. max* and *P. vulgaris* due to the lack of gametophyte tissues **(Fig. 4C, D)**. Together, our results indicate that the RKDs and NLPs show significant functional diversity in many biological processes.

At the post-transcriptional level, AtNLP7 localizes to the cytoplasm after N starvation, whereas it relocates to the nucleus after N supplementation [42]. Here, we determined that *AtNLP7* was downregulated at 30 min and upregulated at 45 min under N-starvation conditions, and its direct targets, *AtNLP1* and *AtNLP3*, were expressed at minimal levels at 30 min of N-starvation treatment. However, other direct targets of AtNLP7, i.e., *AtNLP2* and *AtNLP5*, were upregulated under N-starvation conditions. These findings suggest that in addition to having its own cellular localization regulated at the protein level, AtNLP7 in turn functions at the transcriptional level to regulate other *AtNLPs*.

Modules including different AtNLPs provide differential prerequisites for nodulation. GO term of “endocytosis” (GO:0006897) in the blue module results in the formation of organelle-like structures known as “symbiosomes” during early rhizobial invasion. The term “response to cytokinin” (GO:0009735) in the blue module is also essential for the initiation of nodulation. Auxin signaling is required for rhizobial infection in *Medicago truncatula*. The cell wall starts to weaken when the growing infection thread is close to the base of the root hair. Therefore, unique genes associated with the term “cell-wall

modifications” are also essential for nodule initiation. The GO term “cell cycle” (GO:0051726) in the purple module, also is related to the rhizobial infection [61], was uniquely enriched in the turquoise module. These unique GO terms involved in specific biological processes reflect the diverse roles played by *AtNLPs* under N-starvation conditions and provide the prerequisite for nodulation.

Simulation of low nitrogen on nodulation mediated by NLPs

It is reported that high nitrate repressed nodulation while nodulation only occurred under low nitrates or free nitrates[11, 12, 62]. In our experiment, we found the effects of nitrate on nodulation were possibly mediated by regulation of NLPs in *P. vulgaris*. Higher efficiency of nitrogen-fixation under 5 mM nitrates than that under 0 mM suggests that appropriately low concentrations of nitrates simulate nodulation by increasing nitrogenase activity or nodule number in *P. vulgaris*. The stimulation of nodulation by low nitrogen treatment has also been observed in *G. max* [63]. In *Lotus japonicus*, the NLP NITRATE UNRESPONSIVE SYMBIOSIS 1 (NRSYM1) is a key regulator of the nitrate-induced control of root nodule symbiosis [47]. The phylogenetic tree of NLPs from *A. thaliana* and *P. vulgaris*, as well as NRSYM1 from *L. japonicus*, indicates that Phvul.007G071900 is closely related to NRSYM1 **(Fig. S11)**. This finding, combined with the uniquely high expression pattern of *Phvul.007G071900* in mature nodules under low-nitrogen conditions **(Fig. 10C)**, indicates that *Phvul.007G071900* is involved in integrating nitrate signaling and nodule symbiosis in *P. vulgaris*.

Finally, we determined that Phvul.007G071900 and NRSYM1 are closely related to AtNLP6 and AtNLP7 within clade IIIa **(Fig. 4A and Fig. S11)**. This finding points to the functional similarity of these proteins from clade IIIa in integrating the nitrate-signaling pathway and nodulation. This notion is supported by the observation that transformation with *AtNLP6* or *AtNLP7* partially rescues the nodulation phenotype of the *L. japonicus nrsym1* mutant [47]. Indeed, AtNLP6 or AtNLP7 might be involved in integrating symbiosis and nitrate signaling under N-starvation. Compared to nitrogen-free conditions, we propose that the differential upregulation of genes under low-nitrogen conditions partially explains the stimulatory effects of low-nitrogen treatment on nodulation, which might occur via the upregulation of specific NLPs in *P. vulgaris*.

Conclusions

Our analysis of and comparative evolution between *A. thaliana* (an ideal model dicot) vs. NFC uncovered the evolutionary features (phylogeny, exon structure and protein motif) of RWP-RK family in detail. Coexpression network analysis of *A. thaliana* under both the conditions of N-starvation and N-

supplement revealed differential response of *AtNLP*s, and multiple biological processes correlated to *AtNLP*s under N-starvation pave the way for nodulation within the NFC, which mediated a link between the nitrate signaling pathway and symbiosis. This helped translate biological knowledge from a model organism (*A. thaliana*) to nodulating legumes [58]. Further network analysis in nodulated *G.max* and *P. vulgaris* suggested the differential response of *NLP*s during nodulation programs under different concentrations of nitrates. These results will provide new insights into the relationship between nitrates signaling and nodulation and be helpful to improve NUE from soil by the genetic improvement of *NLP*s.

Methods

Inference of species tree based on gene duplication events

The complete protein sequences of 26 plant species with available whole-genome sequences were integrated from multiple databases **(Table S1)**. To avoid false annotation as much as possible, only the proteins with complete coding sequences encoded by standard genetic codes were retained. All-against-all BLASTP (BLAST 2.7.1+) was conducted for complete proteins from 26 plant species. STRIDE [16] implemented in OrthoFinder-2.2.7 was used for phylogenetic analysis with *A. thaliana* as the outgroup based on deduced gene duplication events. Orthogroups descended from a single gene in the last common ancestor were deduced using OrthoFinder-2.2.7 [17].

Identification and characterization of RWP-RK genes

To identify a complete set of the RWP-RK family in the 26 species as much as possible, 131 RWP-RKs from nine taxa covering the plant kingdom were retrieved from PlantTFDB 4.0 [18] for BLASTP (BLAST 2.7.1+) [19] against proteins of the 26 species with at least e-value of 1e-5 and identity of 50%. The species (from nine taxa) included *Micromonas* sp. RCC299, *Klebsormidium flaccidum*, *Marchantia polymorpha*, *Physcomitrella patens*, *Selaginella moellendorffii*, *Picea abies*, *Oryza sativa* ssp. *indica*, *Arabidopsis thaliana*, and *Amborella trichopoda*. The hidden Markov model (HMM) profile of the RWP-RK family (PF02042) was extracted from the Pfam 32.0 database [20] used for searching using HMMER 3.2 [21]. After integration of the results from BLASTP and HMMER and removal of redundancy, the results were further checked using NCBI CDD [22] and the SMRT database [23]. Basic bioinformatics analysis of various features of the proteins including the molecular weight (MW), isoelectric point (pI), and length were performed using ExPASy (http://www.expasy.ch/tools/pi_tool.html), and the subcellular localizations of the proteins were predicted using BUSCA [24]. Additional conserved motifs were identified using the MEME 5.0.1 package [25] with the parameters minimum motif width, 6; maximum motif width, 50; and maximum number of motifs, 50.

Phylogenetic analysis of RWP-RKs

All identified RWP-RKs were aligned using ClustalW implemented in MEGA X. The best model for the aligned matrix was evaluated with ProteinModelSelection.pl (https://cme.h-its.org/exelixis/web/software/raxml/), and the Jones-Taylor-Thornton (JTT) model was selected for phylogenetic analysis. The phylogenetic tree was built with MEGA X using the neighbor-joining method with 1000 bootstrap values [26].

Coexpression network analysis

Coexpression networks were constructed for *A. thaliana* (GSE97500), *P. vulgaris* (SRX695931), and *G. max* (SRX017401 and SoyBase) using WGCNA_1.64-1 [27]. In brief, the primitive aligned read counts for *A. thaliana* were downloaded for subsequent normalization. For *P. vulgaris* and *G. max*, the primitive reads were downloaded and treated as follows: controlled for read quality with Trimmomatic 0.36 [28]; aligned with HISAT2 2.1.0 [29]; converted to read counts with featureCounts 1.5.3 [30]; and normalized with *regularized logarithm* (rlog), *variance-stabilizing transformation* (VST), and $\log_2$ RPKM using DESeq2 and edgeR [31, 32]. The rlog normalized expressed matrix was transformed into an adjacency matrix based on the simulation of soft threshold for network construction.

Plant materials, growth, and treatment conditions

P. vulgaris (cultivar Tianmadidou) seeds were purchased from Shijiazhuang Xianfeng Seed Industry Co., Ltd, and sterilized with 95% ethanol for 1 minute (min), followed by 0.1% HgCl for 15 min, and transferred to 0.8% water agar for germination at 23°C, 45% relative humidity in the dark for 48 h. The germinatedseedlings were transferred to pots containing a 1:1 mixture of vermiculite and perlite and grown under a 16 h/8 h photoperiod at 23°C and 45% relative humidity. When the first true leaf emerged, 2 mL *Rhizobium tropici* CIAT899 (OD_{600} = 0.2) purchased from Culture Collection of China Agricultural University (CCBAU) was used to infect the *P. vulgaris* roots, and nutrient solution was added every 4–6 days. The plants were grown in nitrate-free Fahraeus medium (FM) supplemented with 0 mM (nitrogen free), 5 mM (low nitrogen), and 10 mM KNO_3 (high nitrogen) under inoculated or non-inoculated conditions. Because early nodules are too small to sample, the early nodules were defined as root-nodule mixtures at 2 cm below the stem, which were obtained on the 7^{th} day after inoculation, while mature nodules were sampled on the 21^{st} day after inoculation. Samples of the corresponding roots were obtained from non-inoculated plants treated with the same concentrations of KNO_3 at similar positions. Each treatment was performed with three replications.

Quantitative RT-PCR

The roots or nodules were immediately frozen in liquid nitrogen for RNA extraction. The qRT-PCR was performed to analyze the relative expression levels of genes in plants infected with different concentrations of rhizobia, using a housekeeping gene for the control. The reaction volume was 10 μL,

containing 0.2 μL of each gene-specific primer, 0.2 μL of cDNA, 5 μL of SYBR Green, and 4.4 μL of sterile distilled water. The qRT-PCR cycling conditions were as follows: 95°C for 30 sec, followed by 40 cycles at 95°C for 10 sec and 60°C for 30 sec; solubility curve of 95°C for 15 sec, 60°C for 60 sec, and 95°C for 15 sec. All reactions were performed in triplicate, and the relative expression level ($2^{-\Delta\Delta Ct}$ value) was calculated based on normalization to the *Actin* gene. The gene-specific primers are listed in **Table S2**. The primers were designed with Primer 5.0 software and checked for specificity by BLASTN. A significance test of differential expression was carried out with Student's *t*-test.

Abbreviations

NFC: nitrogen-fixation clade; NUE: N use efficiency (NUE); NIN: nodule inception; NLPs: NIN-like proteins; WGCNA: weighted correlation network analysis; TFs: transcriptional factors; TRs: transcriptional regulators; PKs: protein kinases

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and material

The plant samples that are used in the present study are available from the corresponding author on reasonable request. The genome information in our study can be found in Table S1 in the supplemental data.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by grant number 2019CFB214 of Hubei Provincial Natural Science Foundation of China, grant number 2014FY110100 of The National Special Fund on Basic Research of Science and Technology of China, grant number 2018ABA093 of The Major Technological Innovation of Hubei

Province of China, grant number 2018BFC360 of Fund for Key Laboratory Construction of Hubei Province, and grant number 2017BEC014 of The Construction Plan of Hubei Province Science and Technology Basic Conditions Platform.

Authors' contributions

RQ, ZHW and HL conceived and designed the experiments. ZHW, WH, LSY, EDQ, TQY and YXZ performed the bioinformatic analysis, CA, YQG, XRW and JW performed the experiments. ZHW, RQ and HL wrote the paper. All authors read and approved the final manuscript.

Acknowledgements

Not applicable.

References

1. Gojon A: **Nitrogen nutrition in plants: rapid progress and new challenges**. *Journal of experimental botany* 2017, **68**(10):2457-2462.
2. Oldroyd GE: **Speak, friend, and enter: signalling systems that promote beneficial symbiotic associations in plants**. *Nature reviews Microbiology* 2013, **11**(4):252-263.
3. Good A: **Toward nitrogen-fixing plants**. *Science (New York, NY)* 2018, **359**(6378):869-870.
4. Yan D, Easwaran V, Chau V, Okamoto M, Ierullo M, Kimura M, Endo A, Yano R, Pasha A, Gong Y *et al*: **NIN-like protein 8 is a master regulator of nitrate-promoted seed germination in Arabidopsis**. *Nature communications* 2016, **7**:13179.
5. Hu B, Jiang Z, Wang W, Qiu Y, Zhang Z, Liu Y, Li A, Gao X, Liu L, Qian Y *et al*: **Nitrate–NRT1.1B–SPX4 cascade integrates nitrogen and phosphorus signalling networks in plants**. *Nature Plants* 2019.
6. Lin JS, Li X, Luo Z, Mysore KS, Wen J, Xie F: **NIN interacts with NLPs to mediate nitrate inhibition of nodulation in Medicago truncatula**. *Nat Plants* 2018, **4**(11):942-952.
7. Griesmann M, Chang Y, Liu X, Song Y, Haberer G, Crook MB, Billault-Penneteau B, Lauressergues D, Keller J, Imanishi L *et al*: **Phylogenomics reveals multiple losses of nitrogen-fixing root nodule symbiosis**. *Science (New York, NY)* 2018, **361**(6398).
8. Chardin C, Girin T, Roudier F, Meyer C, Krapp A: **The plant RWP-RK transcription factors: key regulators of nitrogen responses and of gametophyte development**. *Journal of experimental botany* 2014, **65**(19):5577-5587.
9. Konishi M, Yanagisawa S: **Arabidopsis NIN-like transcription factors have a central role in nitrate signalling**. *Nature communications* 2013, **4**:1617.

10. van Velzen R, Holmer R, Bu F, Rutten L, van Zeijl A, Liu W, Santuari L, Cao Q, Sharma T, Shen D *et al*: **Comparative genomics of the nonlegume Parasponia reveals insights into evolution of nitrogen-fixing rhizobium symbioses**. *Proc Natl Acad Sci U S A* 2018, **115**(20):E4700-e4709.
11. Streeter JG: **Effect of nitrate in the rooting medium on carbohydrate composition of soybean nodules**. *Plant Physiol* 1981, **68**(4):840-844.
12. Carroll BJ, McNeil DL, Gresshoff PM: **Isolation and properties of soybean [Glycine max (L.) Merr.] mutants that nodulate in the presence of high nitrate concentrations**. *Proc Natl Acad Sci U S A* 1985, **82**(12):4162-4166.
13. Li H, Hu B, Chu C: **Nitrogen use efficiency in crops: lessons from Arabidopsis and rice**. *Journal of experimental botany* 2017, **68**(10):2477-2488.
14. Gough C, Vasse J, Galera C, Webster G, Cocking E, Denari ´ e J: **Interactions between bacterial diazotrophs and non-legume dicots: Arabidopsis thaliana as a model plant**. *Plant and Soil* 1997, **194**:123–130.
15. Wu Z, Wang M, Yang S, Chen S, Chen X, Liu C, Wang S, Wang H, Zhang B, Liu H *et al*: **A global coexpression network of soybean genes gives insight into the evolution of nodulation in non-legumes and legumes**. *The New phytologist* 2019.
16. Emms DM, Kelly S: **STRIDE: Species Tree Root Inference from Gene Duplication Events**. *Mol Biol Evol* 2017, **34**(12):3267-3278.
17. Emms DM, Kelly S: **OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy**. *Genome biology* 2015, **16**:157.
18. Jin J, Tian F, Yang DC, Meng YQ, Kong L, Luo J, Gao G: **PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants**. *Nucleic acids research* 2017, **45**(D1):D1040-D1045.
19. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL: **BLAST+: architecture and applications**. *BMC bioinformatics* 2009, **10**:421.
20. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, Qureshi M, Richardson LJ, Salazar GA, Smart A *et al*: **The Pfam protein families database in 2019**. *Nucleic acids research* 2018.
21. Finn RD, Clements J, Arndt W, Miller BL, Wheeler TJ, Schreiber F, Bateman A, Eddy SR: **HMMER web server: 2015 update**. *Nucleic acids research* 2015, **43**(W1):W30-38.
22. Marchler-Bauer A, Bo Y, Han L, He J, Lanczycki CJ, Lu S, Chitsaz F, Derbyshire MK, Geer RC, Gonzales NR *et al*: **CDD/SPARCLE: functional classification of proteins via subfamily domain architectures**. *Nucleic acids research* 2017, **45**(D1):D200-D203.
23. Letunic I, Doerks T, Bork P: **SMART: recent updates, new developments and status in 2015**. *Nucleic acids research* 2015, **43**(Database issue):D257-260.
24. Savojardo C, Martelli PL, Fariselli P, Profiti G, Casadio R: **BUSCA: an integrative web server to predict subcellular localization of proteins**. *Nucleic acids research* 2018, **46**(W1):W459-W466.

25. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS: **MEME SUITE: tools for motif discovery and searching**. *Nucleic acids research* 2009, **37**(Web Server issue):W202-208.
26. Kumar S, Stecher G, Li M, Knyaz C, Tamura K: **MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms**. *Mol Biol Evol* 2018, **35**(6):1547-1549.
27. Langfelder P, Horvath S: **WGCNA: an R package for weighted correlation network analysis**. *BMC bioinformatics* 2008, **9**:559.
28. Bolger AM, Lohse M, Usadel B: **Trimmomatic: a flexible trimmer for Illumina sequence data**. *Bioinformatics* 2014, **30**(15):2114-2120.
29. Kim D, Langmead B, Salzberg SL: **HISAT: a fast spliced aligner with low memory requirements**. *Nature methods* 2015, **12**(4):357-360.
30. Liao Y, Smyth GK, Shi W: **featureCounts: an efficient general purpose program for assigning sequence reads to genomic features**. *Bioinformatics* 2014, **30**(7):923-930.
31. Love MI, Huber W, Anders S: **Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2**. *Genome biology* 2014, **15**(12):550.
32. Robinson MD, McCarthy DJ, Smyth GK: **edgeR: a Bioconductor package for differential expression analysis of digital gene expression data**. *Bioinformatics* 2010, **26**(1):139-140.
33. Borisov AY, Madsen LH, Tsyganov VE, Umehara Y, Voroshilova VA, Batagov AO, Sandal N, Mortensen A, Schauser L, Ellis N *et al*: **The Sym35 gene required for root nodule development in pea is an ortholog of Nin from Lotus japonicus**. *Plant Physiol* 2003, **131**(3):1009-1017.
34. Clavijo F, Diedhiou I, Vaissayre V, Brottier L, Acolatse J, Moukouanga D, Crabos A, Auguy F, Franche C, Gherbi H *et al*: **The Casuarina NIN gene is transcriptionally activated throughout Frankia root infection as well as in response to bacterial diffusible signals**. *The New phytologist* 2015, **208**(3):887-903.
35. Bjellqvist B, Hughes GJ, Pasquali C, Paquet N, Ravier F, Sanchez JC, Frutiger S, Hochstrasser D: **The focusing positions of polypeptides in immobilized pH gradients can be predicted from their amino acid sequences**. *Electrophoresis* 1993, **14**(10):1023-1031.
36. Hu B, Jin J, Guo AY, Zhang H, Luo J, Gao G: **GSDS 2.0: an upgraded gene feature visualization server**. *Bioinformatics* 2015, **31**(8):1296-1297.
37. Blom N, Sicheritz-Ponten T, Gupta R, Gammeltoft S, Brunak S: **Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence**. *Proteomics* 2004, **4**(6):1633-1649.
38. Varala K, Marshall-Colon A, Cirrone J, Brooks MD, Pasquino AV, Leran S, Mittal S, Rock TM, Edwards MB, Kim GJ *et al*: **Temporal transcriptional logic of dynamic regulatory networks underlying nitrogen signaling and use in plants**. *Proc Natl Acad Sci U S A* 2018, **115**(25):6494-6499.
39. Libault M, Farmer A, Joshi T, Takahashi K, Langley RJ, Franklin LD, He J, Xu D, May G, Stacey G: **An integrated transcriptome atlas of the crop model Glycine max, and its use in comparative analyses in plants**. *The Plant journal : for cell and molecular biology* 2010, **63**(1):86-99.

40. O'Rourke JA, Iniguez LP, Fu F, Bucciarelli B, Miller SS, Jackson SA, McClean PE, Li J, Dai X, Zhao PX *et al*: **An RNA-Seq based gene expression atlas of the common bean**. *BMC genomics* 2014, **15**:866.
41. Oldham MC, Horvath S, Geschwind DH: **Conservation and evolution of gene coexpression networks in human and chimpanzee brains**. *Proc Natl Acad Sci U S A* 2006, **103**(47):17973-17978.
42. Marchive C, Roudier F, Castaings L, Brehaut V, Blondet E, Colot V, Meyer C, Krapp A: **Nuclear retention of the transcription factor NLP7 orchestrates the early response to nitrate in plants**. *Nature communications* 2013, **4**:1713.
43. Medici A, Marshall-Colon A, Ronzier E, Szponarski W, Wang R, Gojon A, Crawford NM, Ruffel S, Coruzzi GM, Krouk G: **AtNIGT1/HRS1 integrates nitrate and phosphate signals at the Arabidopsis root tip**. *Nature communications* 2015, **6**:6274.
44. Konishi M, Yanagisawa S: **The role of protein-protein interactions mediated by the PB1 domain of NLP transcription factors in nitrate-inducible gene expression**. *BMC Plant Biol* 2019, **19**(1):90.
45. Zheng Y, Jiao C, Sun H, Rosli HG, Pombo MA, Zhang P, Banf M, Dai X, Martin GB, Giovannoni JJ *et al*: **iTAK: A Program for Genome-wide Prediction and Classification of Plant Transcription Factors, Transcriptional Regulators, and Protein Kinases**. *Molecular plant* 2016, **9**(12):1667-1670.
46. Kelly S: **The amount of nitrogen used for photosynthesis modulates molecular evolution in plants**. *Mol Biol Evol* 2018.
47. Nishida H, Tanaka S, Handa Y, Ito M, Sakamoto Y, Matsunaga S, Betsuyaku S, Miura K, Soyano T, Kawaguchi M *et al*: **A NIN-LIKE PROTEIN mediates nitrate-induced control of root nodule symbiosis in Lotus japonicus**. *Nature communications* 2018, **9**(1):499.
48. Werner GD, Cornwell WK, Sprent JI, Kattge J, Kiers ET: **A single evolutionary innovation drives the deep evolution of symbiotic N2-fixation in angiosperms**. *Nature communications* 2014, **5**:4087.
49. Jiao Y, Wickett NJ, Ayyampalayam S, Chanderbali AS, Landherr L, Ralph PE, Tomsho LP, Hu Y, Liang H, Soltis PS *et al*: **Ancestral polyploidy in seed plants and angiosperms**. *Nature* 2011, **473**(7345):97-100.
50. Emms D, Kelly S: **STAG: Species Tree Inference from All Genes**. *bioRxiv* 2018:1-29.
51. Cannon SB, McKain MR, Harkess A, Nelson MN, Dash S, Deyholos MK, Peng Y, Joyce B, Stewart CN, Jr., Rolf M *et al*: **Multiple polyploidy events in the early radiation of nodulating and nonnodulating legumes**. *Mol Biol Evol* 2015, **32**(1):193-210.
52. Zhang H, Forde BG: **Regulation of Arabidopsis root development by nitrate availability**. *Journal of experimental botany* 2000, **51**(342):51-59.
53. Azani N, Babineau M, Bailey CD, Banks H, Barbosa A, Pinto RB, Boatwright J, Borges L, Brown G, Bruneau A *et al*: **A new subfamily classification of the Leguminosae based on a taxonomically comprehensive phylogeny – The Legume Phylogeny Working Group (LPWG)**. *Taxon* 2017, **66**(1):44-77.
54. Sumimoto H, Kamakura S, Ito T: **Structure and function of the PB1 domain, a protein interaction module conserved in animals, fungi, amoebas, and plants**. *Sci STKE* 2007, **2007**(401):re6.

55. Baralle FE, Giudice J: **Alternative splicing as a regulator of development and tissue identity**. *Nature reviews Molecular cell biology* 2017, **18**(7):437-451.
56. Long M, Rosenberg C, Gilbert W: **Intron phase correlations and the evolution of the intron/exon structure of genes**. *Proc Natl Acad Sci U S A* 1995, **92**(26):12495-12499.
57. Long M, de Souza SJ, Gilbert W: **Evolution of the intron-exon structure of eukaryotic genes**. *Current opinion in genetics & development* 1995, **5**(6):774-778.
58. Movahedi S, Van Bel M, Heyndrickx KS, Vandepoele K: **Comparative co-expression analysis in plant biology**. *Plant, cell & environment* 2012, **35**(10):1787-1798.
59. Tedeschi F, Rizzo P, Rutten T, Altschmied L, Baumlein H: **RWP-RK domain-containing transcription factors control cell differentiation during female gametophyte development in Arabidopsis**. *The New phytologist* 2017, **213**(4):1909-1924.
60. Koszegi D, Johnston AJ, Rutten T, Czihal A, Altschmied L, Kumlehn J, Wust SE, Kirioukhova O, Gheyselinck J, Grossniklaus U *et al*: **Members of the RKD transcription factor family induce an egg cell-like gene expression program**. *The Plant journal : for cell and molecular biology* 2011, **67**(2):280-291.
61. Breakspear A, Liu C, Roy S, Stacey N, Rogers C, Trick M, Morieri G, Mysore KS, Wen J, Oldroyd GE *et al*: **The root hair "infectome" of Medicago truncatula uncovers changes in cell cycle genes and reveals a requirement for Auxin signaling in rhizobial infection**. *Plant Cell* 2014, **26**(12):4680-4701.
62. Mortier V, Holsters M, Goormachtig S: **Never too many? How legumes control nodule numbers**. *Plant, cell & environment* 2012, **35**(2):245-258.
63. Xia X, Ma C, Dong S, Xu Y, Gong Z: **Effects of nitrogen concentrations on nodulation and nitrogenase activity in dual root systems of soybean plants**. *Soil Science and Plant Nutrition* 2017, **63**(5):470-482.

Figures

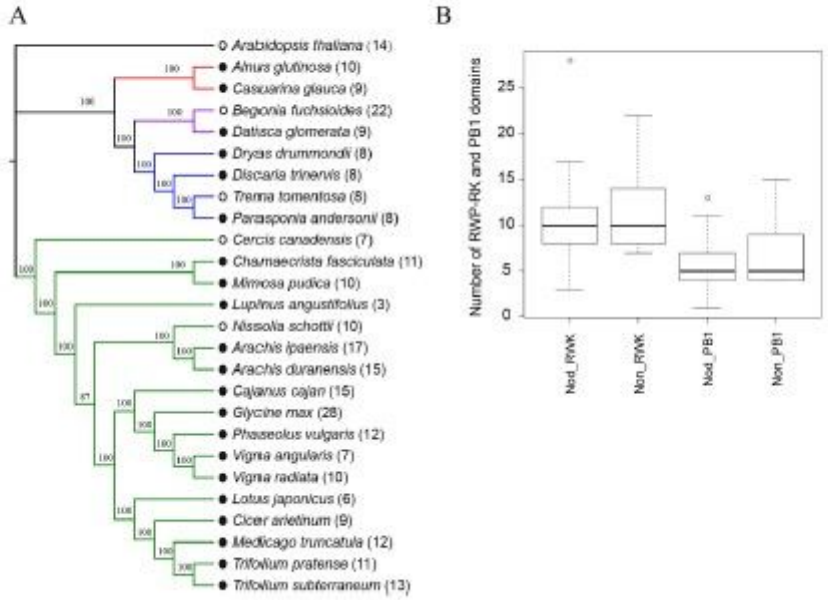

Figure 1

Species tree based on genome-wide duplication events constructed using the maximum-likelihood method. (A) Species tree. (B) Distribution of RWP-RK and PB domains between nodule-forming plants (Nod) and non-nodule-forming plants (Non). The numbers on branches represent bootstrap values, and the numbers in round brackets represent the number of RWP-RK family members. The nodule-forming plants are indicated by filled circles, and the non-nodule-forming plants are indicated by empty circles.

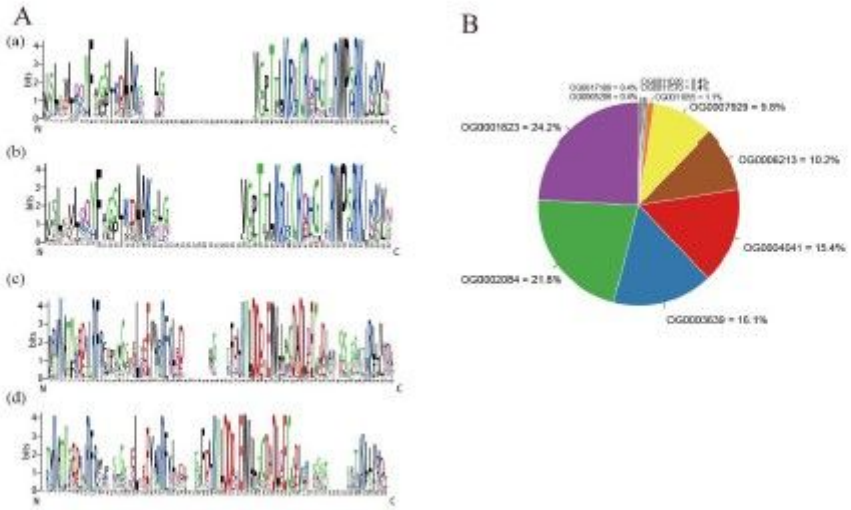

Figure 2

Consensus domains of RWP-RK and PB1 in nodule-forming plants and non-nodule-forming plants and origins of RWP-RKs. In (A), (a) consensus RWP-RK domains of nodule-forming plants; (b) consensus RWP-RK domains of non-nodule-forming plants; (c) consensus PB1 domains of nodule-forming plants; (d) consensus PB1 domains of non-nodule-forming plants. (B) Origins of RWP-RKs, as indicated by the distributions of different orthogroups.

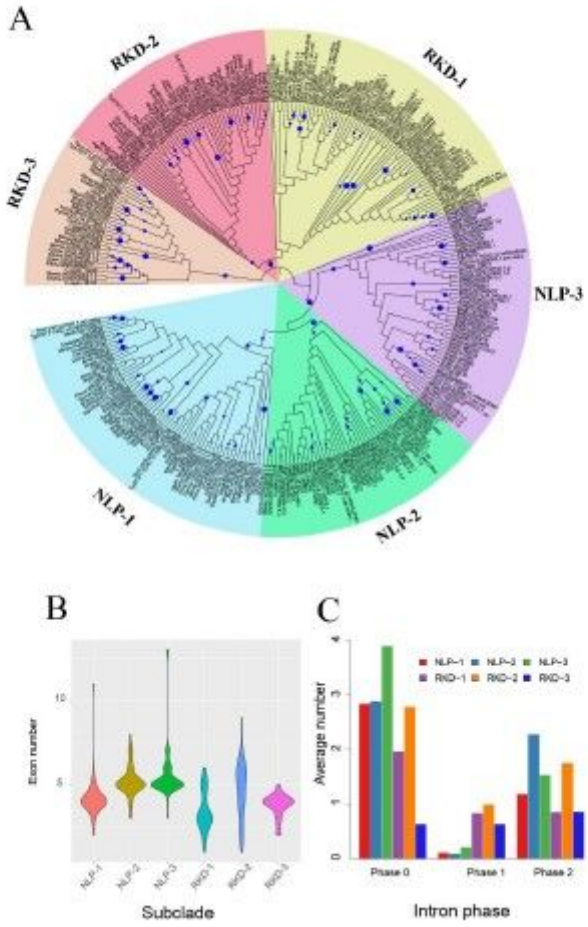

Figure 3

Phylogeny and exon features of RWP-RK proteins. (A) Gene tree constructed by the neighbor-joining method with 1000 bootstrap values. (B) Exon numbers indicated by violin plots. Circle sizes on branches represent bootstrap values ranging from 50 to 100. (C) Average number of exons for six subclades of RWP-RKs with phase 0, phase 1 and phase 2.

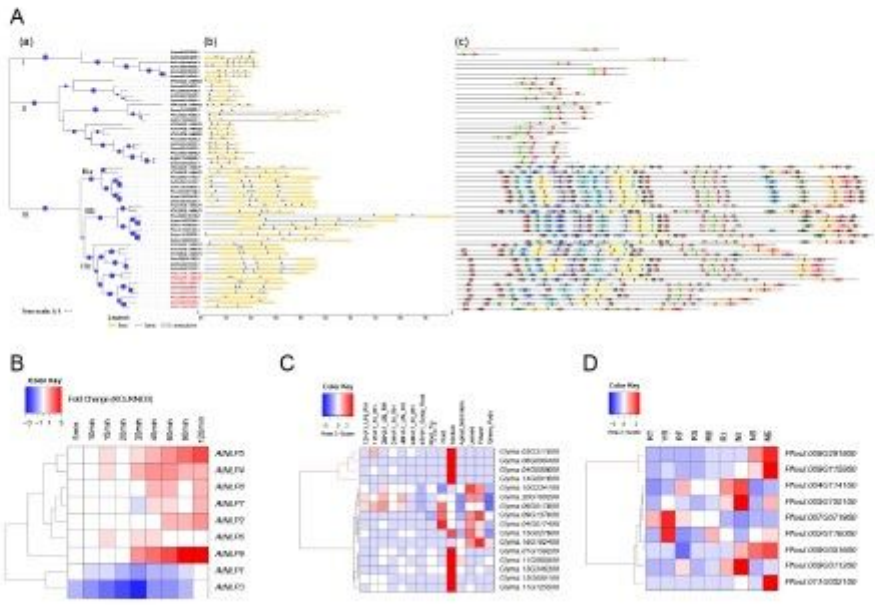

Figure 4

Comparison of gene expression patterns and homology among Arabidopsis thaliana, Phaseolus vulgaris, and Glycine max. Gene tree of RWP-RK proteins with additional gene structures and protein motifs (A) and expression patterns in A. thaliana (B), P. vulgaris (C), and G. max (D).

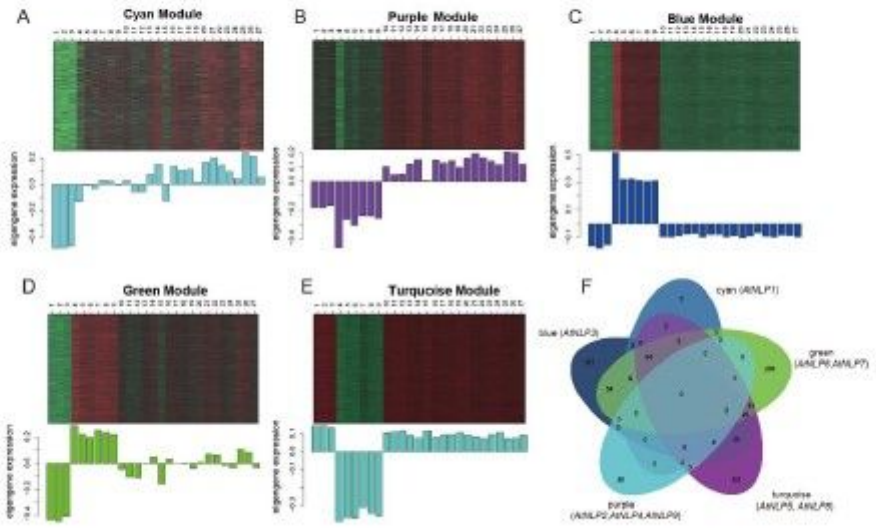

Figure 5

Modules containing AtNLPs correspond to functional subdivisions of the KCl-datasets. (A) Cyan module with AtNLP1. (B) Purple module with AtNLP2, AtNLP4, and AtNLP9. (C) Blue module with AtNLP3. (D) Green module with AtNLP6 and AtNLP7. (E) Turquoise module with AtNLP5 and AtNLP8. (F) Venn diagram of enriched gene ontology terms for each module at the biological process level. The sample accessions represent each time point with three replications, 5 min (1–3), 10 min (4–6), 15 min (7–9), 20 min (10–12), 30 min (13–15), 45 min (16–18), 60min (19–21), 90 min (22–24), and 120 min (25–27).

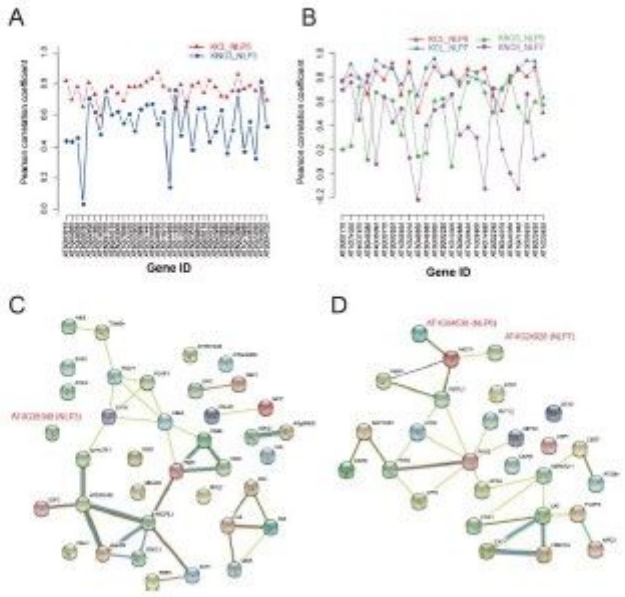

Figure 6

Relationships of AtNLP3, AtNLP6, and AtNLP7 to genes involved in symbiosis (GO:0044403) and starvation (GO:0009267) under N starvation. (A) Correlation of AtNLP3 with genes from the enriched symbiosis-related GO term in the blue module. (B) Correlation of AtNLP6 and AtNLP7 with genes from the symbiosis-related GO term in the green module. (C) Correlation of AtNLP6 and AtNLP7 with genes from the starvation-related GO term in the green module. (D) Protein association network of AtNLP3 with genes from the symbiosis-related GO term constructed with STRING 11.0. (E) Protein association network of AtNLP6 and AtNLP7 with genes from the symbiosis-related GO term constructed with STRING 11.0. (F) Protein association network of AtNLP6 and AtNLP7 with genes from the starvation-related GO term constructed with STRING 11.0.

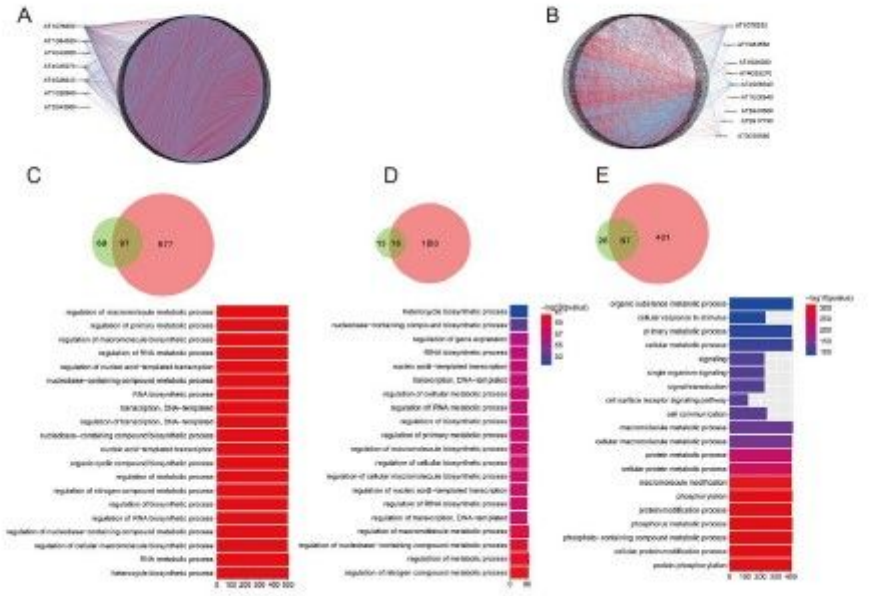

Figure 7

Pearson’s correlation coefficient (absolute value ≥ 0.8) of RWP-RKs to transcription factors, transcriptional regulators, and protein kinases under N starvation (A), N supplementation (B) from A. thaliana and enrichment analysis of genes highly correlated to TFs (C), TRs (D), and PKs (E) under N-starvation. The blue line represents positive correlation and the red line represents negative correlation. The node color represents the module of each gene, and the node size represents the connectivity within each module.

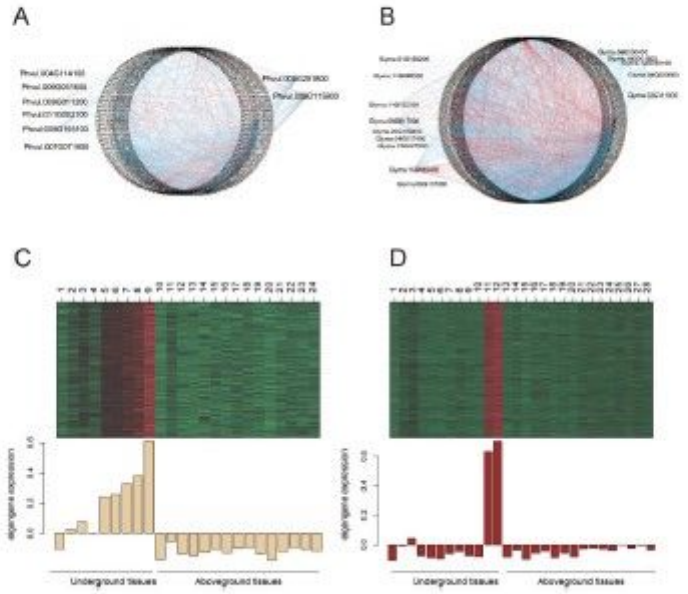

Figure 8

Pearson's correlation coefficient (absolute value ≥ 0.8) of RWP-RKs vs. transcription factors, transcriptional regulators, and protein kinases from the transcriptome atlas of G. max (A) and P. vulgaris (B). The RWP-RKs closely related to AtNLP1 and AtNLP2 within clade IIIc were highly expressed in the nodules of G. max (C) and P. vulgaris (D), respectively.

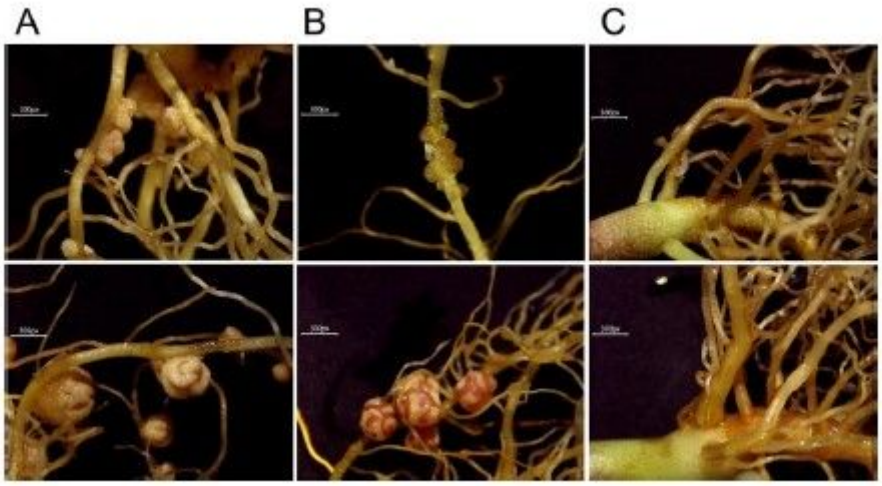

Figure 9

Phenotypes of early and mature nodules under different concentrations of nitrate. (A) 0 mM nitrate. (B) 5 mM nitrate. (C) 10 mM nitrate. For each concentration, the photograph at the top shows early nodules

and the photograph at the bottom shows mature nodules.

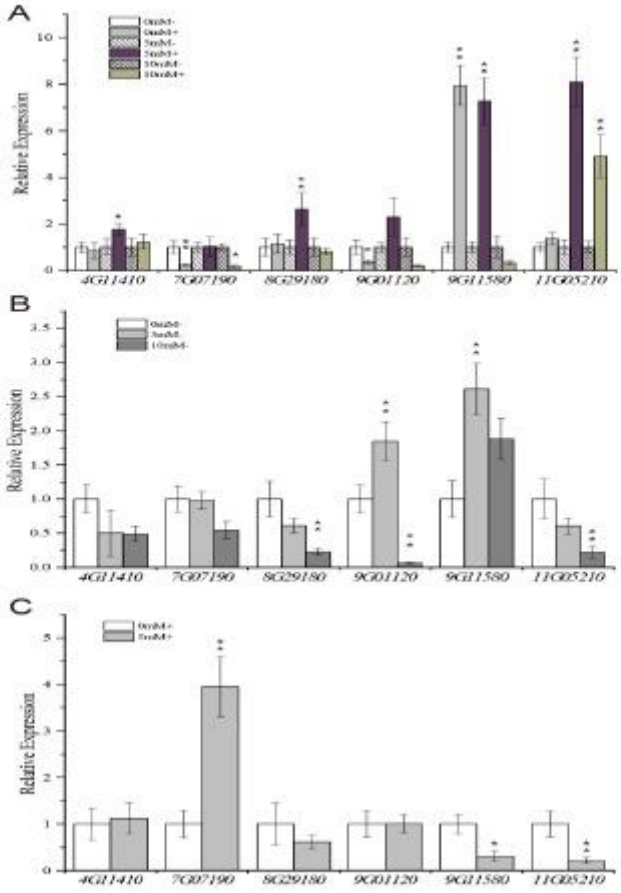

Figure 10

Expression analysis of selected RWP-RK genes in early roots, early nodules, mature roots, and mature nodules treated with different concentrations of nitrates. (A) Early roots and root-nodule mixtures treated with or without inoculation and 0 mM, 5 mM, or 10 mM KNO3. (B) Mature roots treated without inoculation and 0 mM, 5 mM, or 10 mM KNO3. (C) Mature nodules treated with inoculation and 0 mM or 5 mM. Student’s t-test, *P < 0.05, **P < 0.01, ***P < 0.001.