

Functional and comparative genomics reveals conserved noncoding sequences in the nitrogen-fixing clade

Wendell J. Pereira¹ , Sara Knaack² , Sanhita Chakraborty³ , Daniel Conde¹ , Ryan A. Folk⁴ , Paolo M. Triozzi¹ , Kelly M. Balmant¹, Christopher Dervinis¹ , Henry W. Schmidt¹ , Jean-Michel Ane^{3,5} , Sushmita Roy^{2,6}  and Matias Kirst^{1,7} 

¹School of Forest, Fisheries and Geomatics Sciences, University of Florida, Gainesville, FL 32611, USA; ²Wisconsin Institute for Discovery, University of Wisconsin-Madison, Madison, WI 53715, USA; ³Department of Bacteriology, University of Wisconsin-Madison, Madison, WI 53706, USA; ⁴Department of Biological Sciences, Mississippi State University, Starkville, MS 39762, USA; ⁵Department of Agronomy, University of Wisconsin-Madison, Madison, WI 53706, USA; ⁶Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, WI 53715, USA; ⁷Genetics Institute, University of Florida, Gainesville, FL 32611, USA

Summary

Author for correspondence:
Matias Kirst
Email: mkirst@ufl.edu

Received: 26 August 2021
Accepted: 16 January 2022

New Phytologist (2022) 234: 634–649
doi: 10.1111/nph.18006

Key words: comparative genomics, conserved noncoding sequences (CNS), *Medicago truncatula*, *MtCRE1*, nitrogen fixation, nodulation.

- Nitrogen is one of the most inaccessible plant nutrients, but certain species have overcome this limitation by establishing symbiotic interactions with nitrogen-fixing bacteria in the root nodule. This root–nodule symbiosis (RNS) is restricted to species within a single clade of angiosperms, suggesting a critical, but undetermined, evolutionary event at the base of this clade.
- To identify putative regulatory sequences implicated in the evolution of RNS, we evaluated the genomes of 25 species capable of nodulation and identified 3091 conserved noncoding sequences (CNS) in the nitrogen-fixing clade (NFC).
- We show that the chromatin accessibility of 452 CNS correlates significantly with the regulation of genes responding to lipochitooligosaccharides in *Medicago truncatula*. These included 38 CNS in proximity to 19 known genes involved in RNS. Five such regions are upstream of *MtCRE1*, *Cytokinin Response Element 1*, required to activate a suite of downstream transcription factors necessary for nodulation in *M. truncatula*. Genetic complementation of an *Mtcre1* mutant showed a significant decrease of nodulation in the absence of the five CNS, when they are driving the expression of a functional copy of *MtCRE1*.
- CNS identified in the NFC may harbor elements required for the regulation of genes controlling RNS in *M. truncatula*.

Introduction

Nitrogen is one of the most limiting plant nutrients, despite making up 78% of the Earth's atmosphere. Plants cannot access nitrogen directly and must obtain it from the soil. Some species have evolved to overcome this limitation by establishing symbiotic associations with nitrogen-fixing bacteria. In some of these mutually beneficial relationships, the host plant develops root nodules, new root organs that provide an environment for the bacteria to fix nitrogen. Root–nodule symbiosis (RNS) evolved from the recruitment of molecular and cellular mechanisms from arbuscular mycorrhizal symbiosis and the lateral root developmental program (Gualtieri & Bisseling, 2000; Streng *et al.*, 2011; Schiessl *et al.*, 2019). However, the specific molecular origins of the trait remain unknown.

Plants able to develop RNS are restricted to species in a single monophyletic clade of angiosperms, consisting of the orders Rosales, Fagales, Cucurbitales and Fabales (Soltis *et al.*, 1995). While this nitrogen-fixing clade (NFC) comprises 28 families, only 10 include species capable of developing nodules. Within the

majority of these families, most of the member species lack this capability. The scattered distribution of RNS across the NFC suggests it either evolved independently multiple times (convergent evolution) or was lost after one or more gain events. Recent work identified essential genes specific to RNS such as *NODULE INCEPTION (NIN)*, *RHIZOBIUM-DIRECTED POLAR GROWTH (RPG)* and *NOD FACTOR PERCEPTION (NFP)*. The loss of any of these genes is associated with the absence of the trait in the NFC (Griesmann *et al.*, 2018; van Velzen *et al.*, 2018). While this may partially explain the evolution of symbiosis in this clade, their presence in species outside the NFC indicates that other genetic factors are likely to underlie the origin of this trait. Moreover, the observation that nodulating species are confined to a specific clade provides evidence for a single origin or evolutionary event behind the genetic underpinning of nodulation at the NFC base (Soltis *et al.*, 1995; Werner *et al.*, 2014). The nature of this event has not been determined.

While genes implicated in nodulation are present in most species (nodulating or not) within the NFC and even outside the clade, gene sequence conservation alone does not imply

maintenance of similar function. Closely related species can exhibit extensive developmental or phenotypic divergence, despite having high gene content similarity (Kirst *et al.*, 2003). Such differences are often due to variation in transcription regulation (Pollard *et al.*, 2006; Prabhakar *et al.*, 2006). Gene expression regulation is partially determined by the interaction of transcription activators and repressors with specific regulatory sequences. Epigenetic factors, such as accessibility of the genomic regions containing these regulatory sequences, further modulate their contribution to gene expression.

Species capable of establishing symbiosis with nitrogen-fixing bacteria are expected to share conserved sequences due to negative (purifying) selection, including conserved noncoding sequences (CNS). These CNS often carry functionally critical phylogenetic footprints and harbor transcription factor binding motifs (TFBMs) or other *cis*-acting binding sites (Freeling & Subramaniam, 2009). CNS have been extensively documented in plants (Haudry *et al.*, 2013; Burgess & Freeling, 2014; Van de Velde *et al.*, 2014; Liang *et al.*, 2018) and shown to perform essential roles in regulatory networks and plant development. In maize, more than half of putative *cis*-regulatory sequences overlap with CNS. Similarly, TFBMs, gene expression quantitative trait loci and open chromatin regions are enriched in these noncoding sequences (Song *et al.*, 2021). In Arabidopsis, CNS overlaps with a large fraction of open chromatin sites and functional TFBMs (Van de Velde *et al.*, 2014). Therefore, regulatory sequences critical to RNS may be identified by searching for CNS in the genomes of nodulating species in the NFC. Furthermore, chromatin accessibility of these regions may also be associated with the developmental process of nodulation, contributing to the regulation of gene expression.

Identifying CNS among nodulating species in the NFC but diverging in species outside of the clade could point to regulatory sequences that underlie the origin of nodulation. Here we aimed to identify CNS that may harbor putative regulatory sequences associated with the origin of RNS based on the hypothesis that a genetic event (predisposition or gain) occurred at the NFC base. We identified several CNS potentially involved in the regulation of genes that are required for nitrogen fixation. Furthermore, we demonstrate that the deletion of such regions upstream of *MtCRE1* significantly reduces the number of nodules produced by *Medicago truncatula* when symbiosis is established with *Sinorhizobium meliloti*. Such regulatory sequences may have contributed to the differential regulation of genes critical for nodule development. Consequently, they may be necessary for engineering nodulation in crops outside of the NFC, a long-term goal for sustainable agriculture.

Materials and Methods

Species selection and genome sequence quality control

We searched the National Center for Biotechnology Information (NCBI) RefSeq and GenBank databases for all publicly available genomes belonging to orders containing species capable of nitrogen fixation (Fabales, Fagales, Cucurbitales and Rosales). For

some species, an improved version of the genome assembly available in PHYTOZOME (Goodstein *et al.*, 2012) was used. Next, we selected the genomes of species capable of associating with nitrogen-fixing bacteria based on existing information in the literature. Moreover, we selected a group of species not capable of engaging in RNS, belonging to orders outside the NFC, but phylogenetically near this clade (outgroup).

To verify the quality of these genomes, the assembly-stats v.1.0.1 algorithm (<https://github.com/sanger-pathogens/assembly-stats>) was implemented to investigate the contiguity of the assemblies, and BUSCO v.3.02 (Seppey *et al.*, 2019) was used to evaluate the completeness of the corresponding gene annotations based on the embryophyta_odb10 dataset (https://busco-archive.ezlab.org/v3/frame_wget.html). We constrained all analyses to species for which the assembled genome contained at least 80% of the BUSCO genes in this dataset.

Whole-genome alignments

Multiple sequence alignments of whole genomes were generated separately for the two sets of selected genomes (NFC and outgroup), using a previously developed workflow (Liang *et al.*, 2018). First, each genome was submitted to a pairwise whole-genome alignment to the genome of a reference species using LAST v.916 (Kielbasa *et al.*, 2011). The *M. truncatula* Gaertn. genome (v.5; Pecrix *et al.*, 2018) was used as the reference. Before the alignment, simple repetitive elements were masked from all genome sequences using TANTAN (Frith, 2011a,b). The alignments were generated using the lastdb argument -uMAM8, and lastal arguments -p HOXD70 -e 4000 -C 2 -m 100. Next, the split-last algorithm (Frith & Kawaguchi, 2015) was used to improve the selection of orthologous sequences. Finally, the pairwise whole-genome alignments were assembled using AXTCHAIN and CHAINNET (Kent *et al.*, 2003).

All pairwise alignments were combined in a multiple-alignments file using ROAST (reference-dependent multiple alignment tool), which merges the alignments using a phylogenetic tree-guided approach (<http://www.bx.psu.edu/~cathy/toast-roast.tmp/README.toast-roast.html>, last accessed June 2021). The phylogenetic tree used to guide ROAST is shown in Fig. 1 and was based on the backbone and taxon placements in a recent comprehensive land plant phylogeny (Gitzendanner *et al.*, 2018).

Identification of conserved noncoding regions in species capable of engaging in RNS

PHASTCONS (Siepel *et al.*, 2005) was used to estimate the conservation score of the *M. truncatula* genome regions contained within multiple alignments. Briefly, for each multiple genome alignment group (NFC and outgroup), we first applied phyloFit from the PHAST toolkit, v.1.5, to estimate a background model of sequence evolution. Next, we used the PHASTCONS estimate-tree function to learn models of conserved and nonconserved sequence evolution, respectively. Finally, the PHASTCONS most-conserved function was applied on each chromosome to identify conserved and diverged sequences.

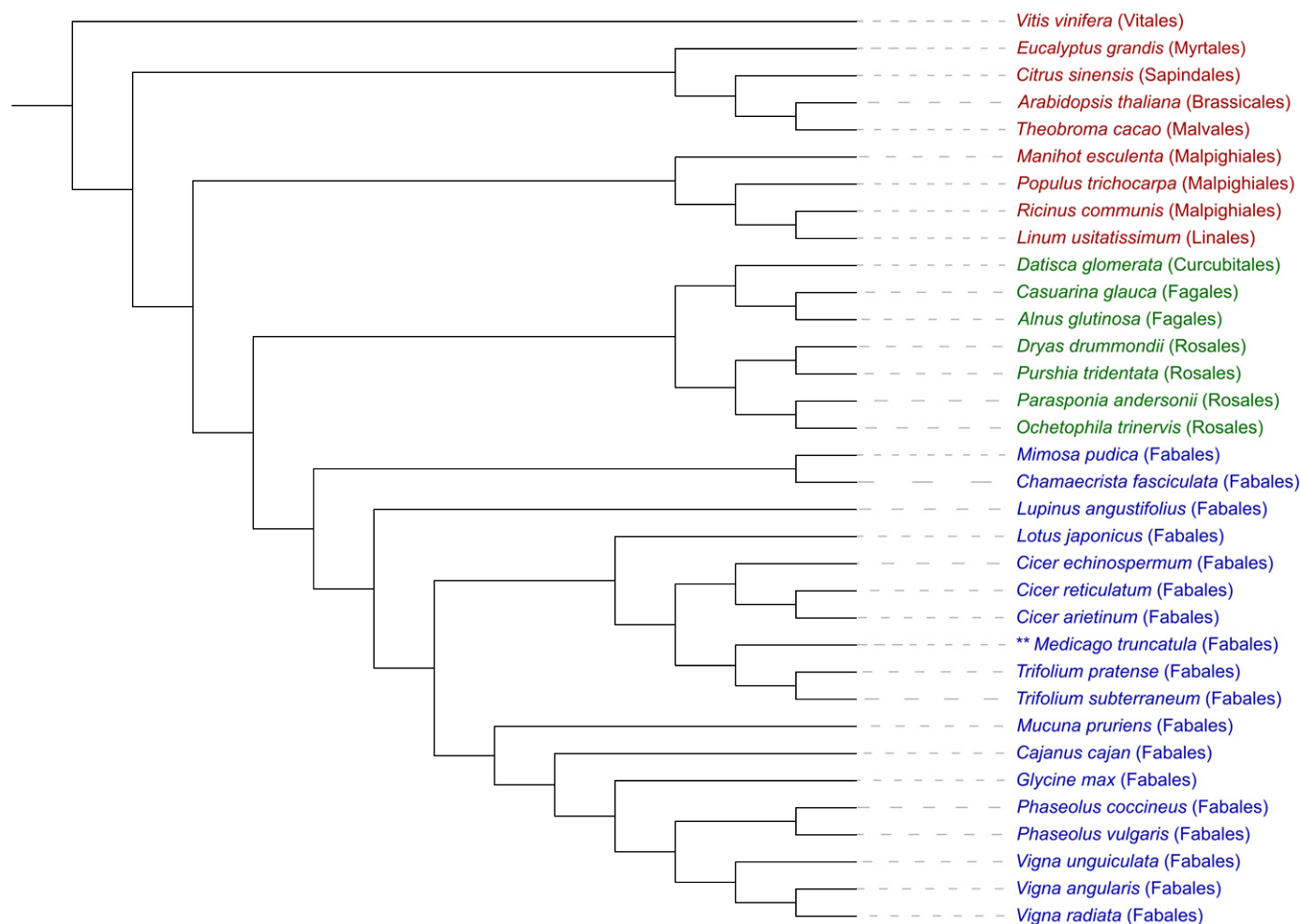


Fig. 1 Phylogenetic tree used to guide the creation of the multiple genome alignment by ROAST. Presented are 25 species included in the evolutionary analysis, belonging to the NFC clade (capable of nitrogen fixation) orders Fabales (blue), Fagales, Cucurbitales and Rosales (green). Additionally, nine outgroup species (red) were included. Asterisks indicate the species used as a reference to produce the whole-genome alignments. To generate this tree, the relationships among the four orders of the NFC clade were defined following a topology recovered in several phylogenetic studies (Werner *et al.*, 2014; The Angiosperm Phylogeny Group, 2016; Gitzendanner *et al.*, 2018; Griesmann *et al.*, 2018). Note that alternative topologies for the orders within the NFC have been proposed (One Thousand Plant Transcriptomes Initiative, 2019).

The regions identified were further filtered to exclude those: smaller than five nucleotides, overlapping (≥ 1 bp) transposable elements and noncoding RNAs, and contained within the mitochondrial or chloroplast genomes. We also excluded from further investigation the regions overlapping in one or more bases with regions annotated as coding sequences in the *M. truncatula* genome. The remaining conserved regions were compared with a subset of the nonredundant protein database (nr), containing all Viridiplantae species, using BLASTX v.2.10.1+. All regions with an *e*-value ≤ 0.01 were excluded. This step aims to remove coding regions that may exist but were not annotated in the *M. truncatula* genome v.5. The analysis workflow for the definition of the CNS regions is shown in Supporting Information Fig. S1.

Genomic context of CNS and ortholog comparison with *Glycine max* (L.) Merr

To perform filtering of the CNS based on synteny to other species, they were classified in six categories according to their

distance to the closest gene in the *M. truncatula* genome. The categories were: intronic (locate inside an intron of a gene), downstream (up to 2 kb downstream of the translation stop site), upstream (up to 2 kb upstream of the translation start site (TSS)), distal upstream (2–10 kb upstream of the TSS), distal downstream (2–10 kb downstream of the translation stop site) and intergenic (all CNS not classified in the previous five categories). We considered the TSS and translation stop site as the reference coordinates to permit the comparison with the soybean (*G. max*) genome, for which untranslated region (UTR) coordinates are not described in the NCBI's annotation. We located the CNS in the *G. max* genome and classified them following the same approach.

Next, we examined if the CNS identified in the *M. truncatula* genome were in orthologous regions of *G. max*, v.2.1 (Schmutz *et al.*, 2010). The soybean genome was selected for this analysis because gene orthology to *M. truncatula* is well established and readily available in the PLAZA database (Van Bel *et al.*, 2018). Moreover, for the *G. max* genome, it was possible to

recover the relationship between the NCBI's (Entrez) locus tags and the identification adopted in the PLAZA database, allowing the comparison. Because PLAZA uses the gene identification of v.4 of the *M. truncatula* genome assembly and annotation, we used the information from the genome portal of *M. truncatula* assembly v.5 (<https://medicago.toulouse.inra.fr/MtrunA17r5.0-ANR/>, last accessed June 2021) to recover the gene identification equivalent in the newer version (v.5). In some cases, two or more genes identified in v.4 corresponded to a single gene in v.5. We considered a gene an ortholog if at least one of the v.4 gene identifications was deemed orthologous of the corresponding gene in *G. max*. Only CNS in which the closest gene was considered an ortholog were kept for further evaluation.

Chromatin accessibility of CNS and gene expression of associated genes

To identify CNS that are associated with regulatory changes triggered by rhizobial infection and nodule formation, we examined chromatin accessibility (ATAC-seq) and gene expression (RNA-seq) data captured after *M. truncatula* root treatment lipochitooligosaccharides (LCOs) from *Sinorhizobium meliloti* (Dangeard 1926) De Lajudie *et al.* 1994 (Knaack *et al.*, 2021). Rhizobial LCOs act as symbiotic signals during nodule formation. Briefly, for generating these data, roots from wild-type (WT) seedlings (reference accession Jemalong A17) were immersed in a solution of purified LCOs derived from *S. meliloti* or 0.005% ethanol solution (control) for 1 h and collected posteriorly at specific intervals (0 h, 15 min, 30 min, 1 h, 2 h, 4 h, 8 h and 24 h). We obtained the quantile-normalized and log-transformed TPM (transcripts per kilobase million) expression values for all *M. truncatula* genes (Knaack *et al.*, 2021). Additionally, ATAC-seq read counts were aggregated and normalized for each CNS. Specifically, for each CNS the mean per-base pair coverage was obtained and log-ratio-transformed relative to genome-wide mean coverage for each time point, producing a normalized accessibility profile across time. Quantile normalization was subsequently applied across the time course to the log-ratio-normalized values.

To evaluate the relationship between gene expression and CNS accessibility, we carried out a correlation analysis, as described previously (Knaack *et al.*, 2021). Briefly, we first performed a zero-mean transformation of each gene's expression profile and each CNS's accessibility profile. Pearson's correlation (ρ) between the accessibility profiles of each site and the expression profile of the closest gene across all eight time points was calculated. In this step, only the CNS classified as downstream, upstream, distal upstream or distal downstream were included. To assess the significance of the relationship between gene expression and chromatin accessibility of the CNS, a null distribution of correlations from 1000 random permutations of the chromatin accessibility score and the gene expression in the eight time points was generated. We computed a P -value that estimates the probability of observing a correlation in the permuted data higher than the observed correlation, treating positive and negative

correlations separately. In practice, a P -value threshold of ≤ 0.05 for significant correlation was used. Also, we focus on the CNS for which chromatin accessibility and gene expression produced strong correlations ($\rho > 0.50$ or < -0.50).

Selection of CNS potentially involved in nitrogen fixation

To exclude CNS present in species outside the NFC, we removed those that overlap (≥ 1 bp) with CNS regions called for the out-group species. The remaining filtered CNS consist of those with a potential role in RNS because of their conservation within this clade.

Several genes have been previously implicated in RNS, collectively known as RNS genes. These genes were identified based on the current literature on *M. truncatula* symbiosis and orthology with other species within the NFC. To determine CNS that may contribute to regulating these genes, we selected those regions where a significant correlation between chromatin accessibility and gene expression was detected. For the selected CNS, a search for TFBMs was conducted using the PlantRegMap database (Jin *et al.*, 2017; Tian *et al.*, 2020).

Experimental validation of CNS associated with the R gene *MtCRE1*

To evaluate the functional role of CNS identified upstream of the *MtCRE1* gene, we used the Golden Gate modular cloning system to transform the *Mtcre1-1* mutant (Plet *et al.*, 2011) and insert a functional copy of *MtCRE1* fused with three different versions of the 2.5 kb region upstream of the translation starting point, containing the promoter region and the 5' UTR, synthesized by Synbio Technologies (Monmouth Junction, NJ, USA). In each version, the sequences corresponding to two or more of the five CNS identified in the 5' UTR of *MtCRE1* were deleted. In the first version, two CNS were deleted ($\Delta 2$ CNS), in the second version three were deleted ($\Delta 3$ CNS) and in the third version all five regions were deleted ($\Delta 5$ CNS) (see Fig. S2 and Table S1 for details). The two CNS deleted in $\Delta 2$ CNS are in the 5' UTR, while the three CNS deleted in $\Delta 3$ CNS are in an intron that divides the 5' UTR (based on Jemalong A17 genome v.5.0; Fig. S2). Golden Gate cloning was utilized as previously described (Engler *et al.*, 2014).

To assemble the transcriptional units, Level 0 parts of each version of the promoter were created with the *MtCRE1* coding region and 35S terminator (MoClo Plant Parts Kit-pICH41414) into the MoClo toolkit Level 1 acceptor-position2-reverse (pICH47811) vector using the Golden Gate BsaI/T4 restriction ligation reaction (Weber *et al.*, 2011; Engler *et al.*, 2014). Finally, each Level 1 transcriptional unit containing a different version of the promoter was assembled into the MoClo Level 2 acceptor (pAGM4673) together with the Level 1 transcriptional units containing the fluorescent marker of transformation, p35S::tdTOMATO-ER::tNOS, using the Level 2 BpI/T4 restriction ligation reaction (Weber *et al.*, 2011; Engler *et al.*, 2014).

Generation of composite plants and nodulation assay

Mtcre1 mutants or their wild-type sibling radicles were inoculated with *Agrobacterium rhizogenes* (Riker *et al.*, 1930) Conn, 1942 MSU440 as previously described (Boisson-Dernier *et al.*, 2001). Three weeks after transformation with MSU440, roots were screened for red fluorescence of *tdTomato* on the binary vector. The composite plants with red fluorescent roots were transferred to growth pouches (<https://mega-international.com/tech-info/>) containing Modified Nodulation Medium (Chakraborty *et al.*, 2021). The plants were acclimated for 1 week and then inoculated with *S. meliloti* 1021, harboring pXLGD4 and expressing *lacZ* under the *hemA* promoter (Leong *et al.*, 1985). The nutrient medium was replenished as required. Two weeks after inoculation, live seedlings were stained for *lacZ* (5 mM potassium ferrocyanide, 5 mM potassium ferricyanide, and 0.08% X-gal in 0.1 M Pipes, pH 7) overnight at 37°C. Roots were rinsed in distilled water and observed under a Leica fluorescence stereomicroscope. Nodule presence was scored at 14 d postinoculation.

Quantification of *MtCRE1* expression with qRT-PCR

To investigate if *MtCRE1* expression is affected by the deletions described above, total RNA was extracted from the transformed roots using a Qiagen RNeasy® Plant Mini kit 5 d postinoculation. Genomic DNA was removed using a Turbo DNA-free™ Kit (Ambion). First-strand cDNA was synthesized using RevertAid RT Reverse Transcription Kit (Thermo Scientific, Waltham, MA, USA). Quantitative reverse transcriptase PCR (qRT-PCR) was performed using Bio-Rad SsoAdvanced Universal SYBR Green Supermix on the Bio-Rad CFX96™ Real-time system: C1000 Touch™ Thermal cycler. *RNA Helicase 1 (Hel)* and *Ubiquitin-conjugating enzyme 9 (UBC9)* were used as endogenous controls. Three biological and technical replicates were used. The primers used in the qRT-PCR are shown in Table S2.

Results

An atlas of CNS in the nitrogen-fixing clade

Selection of genomes of species in the NFC and whole-genome alignments We identified genome sequences from 84 species in the NFC orders Fabales, Fagales, Cucurbitales and Rosales in NCBI. After removing those species unable to associate with nitrogen-fixing bacteria, a total of 33 genomes remained. Assessment of assembly contiguity and gene completeness resulted in the exclusion of eight additional genome sequences. Thus, the genomes of 25 species capable of RNS, including the reference *M. truncatula*, were used to identify CNS (Fig. 1; Notes S1). Moreover, we selected a group of nine species belonging to the orders Brassicales, Linales, Malpighiales (three species), Malvales, Myrtales, Sapindales and Vitales (all outside the NFC; Fig. 1) to be used as an outgroup (see the Materials and Methods section).

The pairwise whole-genome alignment of each species to *M. truncatula* covered on average 16.2% (SD = 0.06) of

nucleotides of the reference genome and 15.0% (SD = 0.06) of nucleotides in the genomes of the remaining species in the NFC. In the pairwise whole-genome alignments of the outgroup species, 8.6% (SD = 0.01) of nucleotides in the *M. truncatula* and 10.7% (SD = 0.05) of the nucleotides of the other species were covered, on average (Notes S1). Moreover, a large share of the regions covered by the alignment of the NFC group corresponds to coding sequences in *M. truncatula*. On average, 40.9% (SD = 11.6) of the nucleotides in alignments between the reference and target species were located within these regions. Considering the coding sequences in the *M. truncatula* genome, an average of 56.4% (SD = 0.08) of the nucleotides were covered in the alignment of the NFC group, and 45% (SD = 0.03) in the outgroup (Notes S1).

Identification of conserved sequences We used PHASTCONS to estimate the conservation score across the genome of species in the NFC. PHASTCONS identifies conserved sequences in multiple genome alignments while considering the phylogenetic relationship and nucleotide substitution in each site under a neutral evolutionary model. PHASTCONS detected 114 162 conserved regions from the alignments of 25 NFC species (Fig. 2a), with a conservation score > 0.5 and a size of ≥ 5 bp (Table 1). The conserved regions represented 2.5% of the *M. truncatula* genome. Next, we applied PHASTCONS to identify conserved regions among species in the outgroup and detected 97 302 such regions (mean length = 104.63 bp, SD = 104.41).

The conserved regions identified by PHASTCONS were mainly located within genic regions or their vicinity. Among the 114 162 conserved regions, 107 433 were within (overlap of one or more nucleotides) a coding sequence. For the remaining conserved regions, 1526 were located downstream (0–2 kb), 485 long-range (distal) downstream (2–10 kb), 2924 upstream (0–2 kb) and 537 long-range (distal) upstream (2–10 kb) of a gene. Moreover, 1093 of these regions were within introns and 164 in intergenic regions, further than 10 kb from the transcription start or transcription stop site of any gene.

Selection of CNS within the NFC and their genomic context To focus on the CNS, we excluded from further analysis the majority (>94%) of the 114 162 regions detected by PHASTCONS that overlapped with coding sequences (Fig. 2b). After removing the conserved regions within coding sequences, 6729 (5.9%) remained. The CNS located outside of coding sequences were further investigated to determine if they represent noncoding sequences such as noncoding RNAs, transposable elements (TEs) or coding genes not described in the *M. truncatula* genome annotation. The abundance of these noncoding regions in plant genomes may result in the detection of high PHASTCONS conservation scores while being unrelated to the evolutionary importance of RNS.

Among the 6729 conserved regions located outside coding sequences, 515 and 778 were excluded from further analysis due to overlap with noncoding RNAs and TEs, respectively. Moreover, 225 conserved regions had significant similarity (BLASTX, *e*-value < 0.01) to sequences in the nr protein database and were

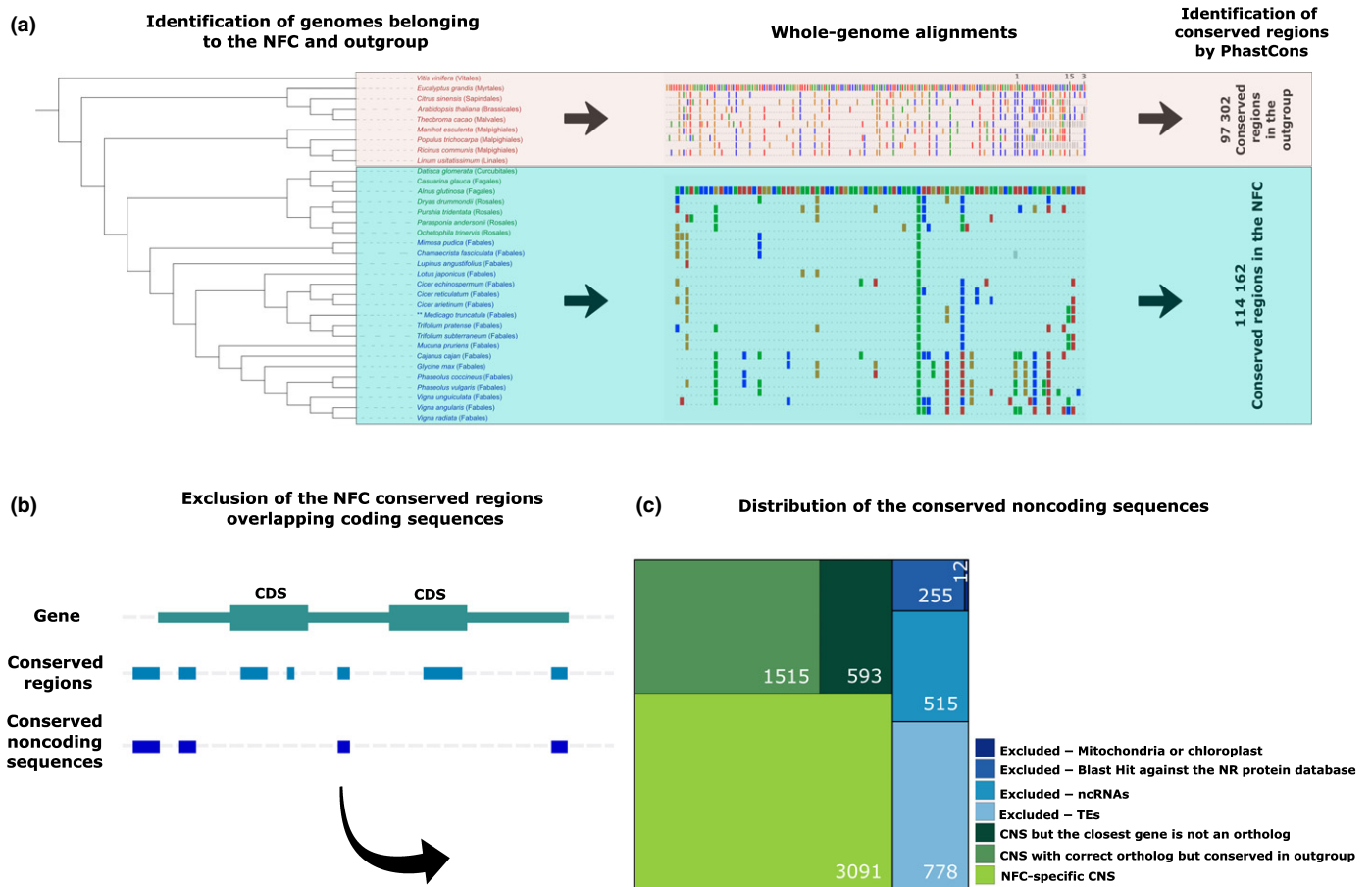


Fig. 2 Workflow representing the datasets and analytical steps used in this study to generate the whole-genome multiple alignments and define conserved regions in the nitrogen-fixing clade and outgroup (a); to identify the conserved noncoding sequences (CNS) in the nitrogen-fixing clade (NFC, green) and outgroup (red) (b); and to select the conserved noncoding sequences that are specific to the NFC (c). CDS, coding sequence; ncRNAs, noncoding RNAs; nr, protein database, NCBI's nonredundant protein database; TEs, transposable elements.

Table 1 Descriptive statistics of the length of all conserved regions identified by PHASTCONS and of the set of regions considered as true conserved noncoding sequences.

Group	Number of regions	Minimum (bp)	Maximum (bp)	Mean (bp)	Median (bp)	SD (mean)	Confidence interval (95%)
Conserved regions (PHASTCONS)	114 162	5	3111	93.65	70	107.7	93.02–94.27
Conserved noncoding sequences	5199	5	431	43.04	27	43.57	41.86–44.22

removed (Fig. 2c). Twelve regions were removed because they were from mitochondrial or chloroplast genomes. Further information about these regions is given in Notes S2. After filtering for the criteria described above, 5199 conserved regions were considered *bona fide* CNS. These represent 4.6% of all conserved regions identified by PHASTCONS. The length of these CNS was significantly smaller than that observed for the complete set of conserved regions identified by PHASTCONS (Table 1; Fig. S3). This observation possibly reflects shorter regulatory motifs detected in noncoding sequences compared to coding sequences.

A large fraction of the 5199 CNS (2441, or 46.95%) was detected upstream of the TSS of the closest gene (0–2 kb), based on the *M. truncatula* reference genome. Among the remaining

CNS, 1245 (23.95%) were downstream of the translation end site (0–2 kb) and 871 (16.75%) were in introns. An additional 294 (5.65%) CNS were located between 2 and 10 kb upstream of the transcription start site and were classified as distal upstream. Also, 275 CNS were located between 2 and 10 kb downstream the transcription stop site and were classified as distal downstream. All other CNS (73) were found in intergenic regions (Fig. S4).

Orthologous CNS in *M. truncatula* and *G. max*

For a CNS to be evolutionarily and biologically relevant across taxa, it is expected to occur in a similar genomic context (e.g.

within the promoter region of orthologous genes). Thus, we next examined if the CNS identified in the *M. truncatula* genome were in orthologous regions of the soybean genome, *G. max*, v.2.1 (Schmutz *et al.*, 2010). From the 5199 CNS, it was possible to recover the coordinates of 5165 in the *G. max* genome. The distances of these CNS relative to the closest gene were calculated and classified as described above for *M. truncatula*. Most of the CNS were classified similarly in both genomes (4098 CNS, 78.82%), especially in the upstream, downstream and intronic categories (Fig. 3).

The most prominent disagreement was observed for the CNS classified as intergenic, where only 34.25% of those in this category in *M. truncatula* were in the same category in *G. max*. The distal upstream and distal downstream categories also presented a significant disagreement in their classification. Approximately 50% of the CNS were classified equally in both genomes for these two categories (Fig. 3). This lack of agreement is expected because CNS are, in general, closely located to their target genes, and CNS found further than 2 kb from genes may not be functionally relevant. Nonetheless, long-range *cis*-regulatory elements are also known in plants and the distal CNS represent a potential approach for their detection.

Next, we assessed if the closest gene to each CNS classified in the same category (except for intergenic CNS) in both *M. truncatula* and *G. max* genomes corresponds to orthologous genes in these species. For several of the CNS it was not possible to determine if the closest genes in both species were orthologous. This occurred because there was no correspondence between *M. truncatula* v.4 and v.5 gene IDs, the Entrez ID and the *G. max* IDs used by PLAZA, or yet no orthology mapping was established from PLAZA for a given *M. truncatula* gene. Consequently, the number of CNS that had their closest genes evaluated varied from 2600 to 3805, depending on the method used to define the

orthology (Table 2). The fraction of CNS for which genes were considered orthologous ranged from 84.45 to 96.82% using PLAZA'S Best-Hits-and-Inparalogs (BHI) family or 'Orthologous gene family' methods, respectively (Table 2).

To characterize the largest number of CNS possible but also to focus on those most likely to have a biological effect, we opted to use the more inclusive 'Orthologous gene family' method to filter the CNS–gene pairings. We ultimately identified 3684 CNS mapped to orthologous genes for subsequent investigation (Notes S3).

CNS specific to the NFC clade

Considering the hypothesis that a predisposition or gain of the nodulation trait arose from a single event at the base of the NFC, we excluded all CNS detected in the NFC that were also observed within or overlapped in one or more nucleotides with a conserved region in the outgroup (593 CNS). Removal of these sequences resulted in 3091 (83.90%) remaining for further analysis. Hereinafter, those CNS are referred to as NFC-specific CNS. The genomic coordinates of these CNS are available in Notes S4. The NFC-specific CNS were generated using only species capable of RNS.

There are species within the NFC that do not engage in RNS. The exclusion of NFC-specific CNS conserved in these non-nodulator species may further contribute to identifying NFC-specific CNS that are not related to RNS. Therefore, the whole-genome alignments of 21 species within the NFC but not capable of RNS (Fig. S5) were generated. A total of 112 061 conserved regions, including coding regions, were identified in this group (Notes S5). Next, crossing the coordinates of 3091 NFC-specific CNS and these conserved regions showed that 1084 NFC-specific CNS overlap with one or more of the conserved regions in the NFC species not capable of RNS. The 1084 NFC-specific

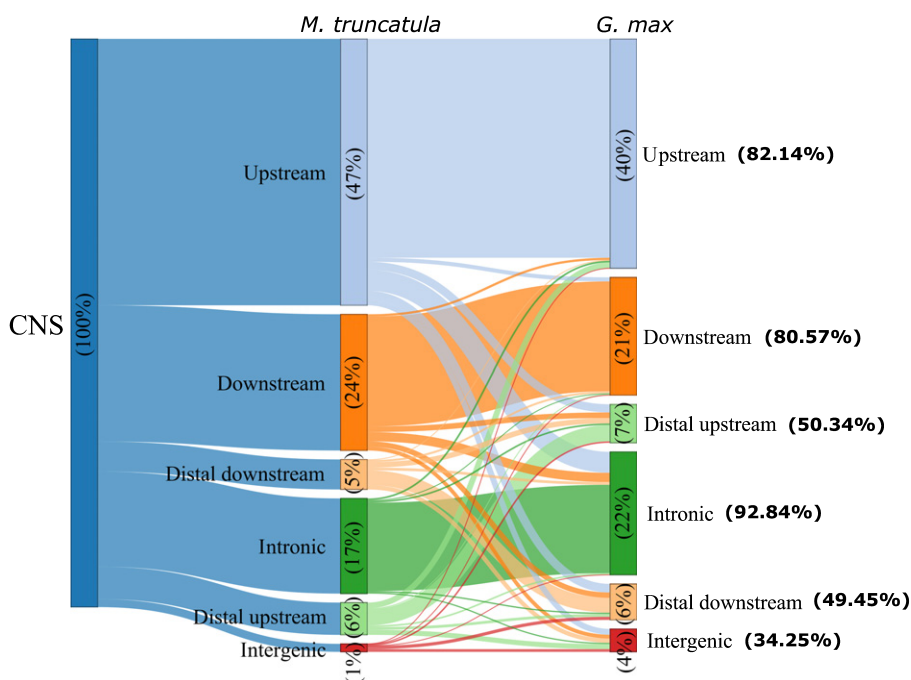


Fig. 3 Classification of conserved noncoding sequences (CNS) according to their distance to the closest genes in the genomes of *Medicago truncatula* and *Glycine max*. The percentages within the bars represent the distribution of the CNS in each species among the six categories. The percentage of CNS with the same classification in *G. max* as in *M. truncatula* is shown on the right.

Table 2 Evaluation of orthology of the closest gene of each conserved noncoding sequence specific to the nitrogen-fixing clade in *Medicago truncatula* and *Glycine max*, according to the orthology methods available in the PLAZA dicot v.4.0 database.

Method (PLAZA DB)	Orthologous?			%
	No	Yes	Total	
Tree-based ortholog	293	2307	2600	88.74
Orthologous gene family	121	3684	3805	96.82
Anchor point	260	3158	3418	92.39
Best-Hits-and-Inparalogs (BHI) family	590	3203	3793	84.45

CNS, and their genomic coordinates, are provided in Notes S6. Nonetheless, there are important caveats in excluding NFC-specific CNS based solely on the overlap with conserved regions in species not capable of RNS (see Discussion below). All 3091 NFC-specific CNS were kept in the subsequent analysis.

Chromatin accessibility of NFC-specific CNS correlates with gene expression during nodule development

Chromatin accessibility of regulatory regions often contributes to modulate the expression of nearby genes. We previously generated global transcriptome and genome-wide chromatin accessibility data for *M. truncatula* (genotype Jemalong A17) roots, 0 min, 15 min, 30 min, 1 h, 2 h, 4 h, 8 h and 24 h after treatment with rhizobium LCOs (Knaack *et al.*, 2021). Based on the ATAC-seq data, we observed that most of the 3901 NFC-specific CNS were in regions that display variation in chromatin accessibility following LCO treatment (Fig. S6).

The detection of variation in the chromatin accessibility of NFC-specific CNS in response to LCO treatment suggests a possible role of these regions in the transcriptional control of *Medicago* genes necessary for symbiotic signaling, rhizobial colonization and nodule development. We therefore sought to identify NFC-specific CNS that could be important in these biological processes. The RNA-seq data collected from root samples were used to quantify transcription levels of the closest gene to each NFC-specific CNS classified as upstream, downstream, distal upstream or distal downstream. In total, we tested 2459 NFC-specific CNS and 1155 genes for significant correlations between the chromatin accessibility profile of the respective NFC-specific CNS and a corresponding expression profile of a mapped gene, noting here that a single gene can be associated with multiple NFC-specific CNS. Based on Pearson's correlation analysis, we detected 452 instances where the chromatin accessibility of an NFC-specific CNS is significantly correlated with the expression of the closest gene ($P \leq 0.05$) responding to the LCO treatment. A total of 376 NFC-specific CNS correlated positively ($\rho > 0.5$; Fig. 4) and 76 negatively ($\rho < -0.5$; Fig. 5) with the expression profile of the closest gene.

NFC-specific CNS associated with expression of nodulation genes

Root–nodule symbiosis is a complex developmental phenomenon in which a large number of genes are differentially

regulated (Breakspear *et al.*, 2014; Larrainzar *et al.*, 2015; Jardinaud *et al.*, 2016; Schiessl *et al.*, 2019). These genes are collectively named as RNS genes and more than 200 are known to date (Roy *et al.*, 2020). From the 3091 NFC-specific CNS, we selected those in which the closest gene is involved in RNS, and evaluated their chromatin accessibility and gene expression (Fig. 6). A total of 38 NFC-specific CNS were located in proximity to 19 RNS genes (Notes S7). Ten NFC-specific CNS have a significant and positive correlation ($P \leq 0.05$ and $\rho > 0.5$), and two have a significant and negative correlation ($P \leq 0.05$ and $\rho < -0.5$).

The NFC-specific CNS correlated significantly with expression were associated with six RNS genes (Fig. 6). These include three genes with well-established roles in nodulation (Fig. 6): *Cytokinin Response Element 1* (*MtCRE1*), encoding a histidine kinase cytokinin receptor required for proper nodule organogenesis (Gonzalez-Rizzo *et al.*, 2006; Plet *et al.*, 2011); *Interacting Protein of NSP2* (*IPN2*), encoding a member of the MYB family of transcription factors, activates the key nodulation gene *Nodule Inception* (*NIN*) in *Lotus japonicus* (Xiao *et al.*, 2020); and *MtPUB2*, encoding a U-box (PUB)-type E3 ligase, is involved in nodule homeostasis (Liu *et al.*, 2018). We further observed association with three genes with potential roles in nodulation: *MtCLE1* is a member of the Clavata3/ESR (CLE) gene family and is induced in nodules (Hastwell *et al.*, 2017) – several CLE peptides are involved in the autoregulation of nodulation or regulation by nitrate or rhizobia (Nowak *et al.*, 2019; Mens *et al.*, 2021); *MtERF1*, encoding an APETALA2/ethylene response factor (AP2/ERF) transcription factor (TF) – *ERF Required for Nodulation 1 and 2* (*ERN1*, *ERN2*) play important roles during rhizobial infection (Cerrri *et al.*, 2012); and *MtKLV*, encoding a receptor-like kinase (*Klavier*). In *L. japonicus*, *klavier* mediates the systemic negative regulation of nodulation and *klavier* mutants develop hypernodulation (Oka-Kira *et al.*, 2005; Miyazawa *et al.*, 2010). The two CNS with significant negative correlation were located upstream of the gene *MtCLE1* and downstream of the gene *MtKLV*, respectively.

Three of the 10 positively correlated NFC-specific CNS were located upstream (one) or distal upstream (two) of the gene *MtERF1* (*ETHYLENE RESPONSE FACTOR 1*, encoding an ethylene-responsive AP2 transcription factor). The NFC-specific CNS classified as upstream was located at the 5' UTR while the two distal upstream NFC-specific CNS are c. 5 kb away from the gene. Another distal-upstream NFC-specific CNS, located c. 8 kb of the gene *MtIPN2*, was identified (Fig. 6). The remaining four significant and positively correlated NFC-specific CNS are located upstream of the gene *MtCRE1* (*MtrunA17Chr8g0392301*) (Fig. 6).

Since CNS are known to harbor TFBMs, we investigated if the 12 significantly correlated NFC-specific CNS contain TFBMs that could be implicated in regulating their corresponding RNS gene. PlantRegMap (Jin *et al.*, 2017; Tian *et al.*, 2020) was used to search for the presence of TFBMs, and 41 TFBMs were identified in six of the NFC-specific CNS. In some cases, all TFBMs identified in an NFC-specific CNS belong to the same

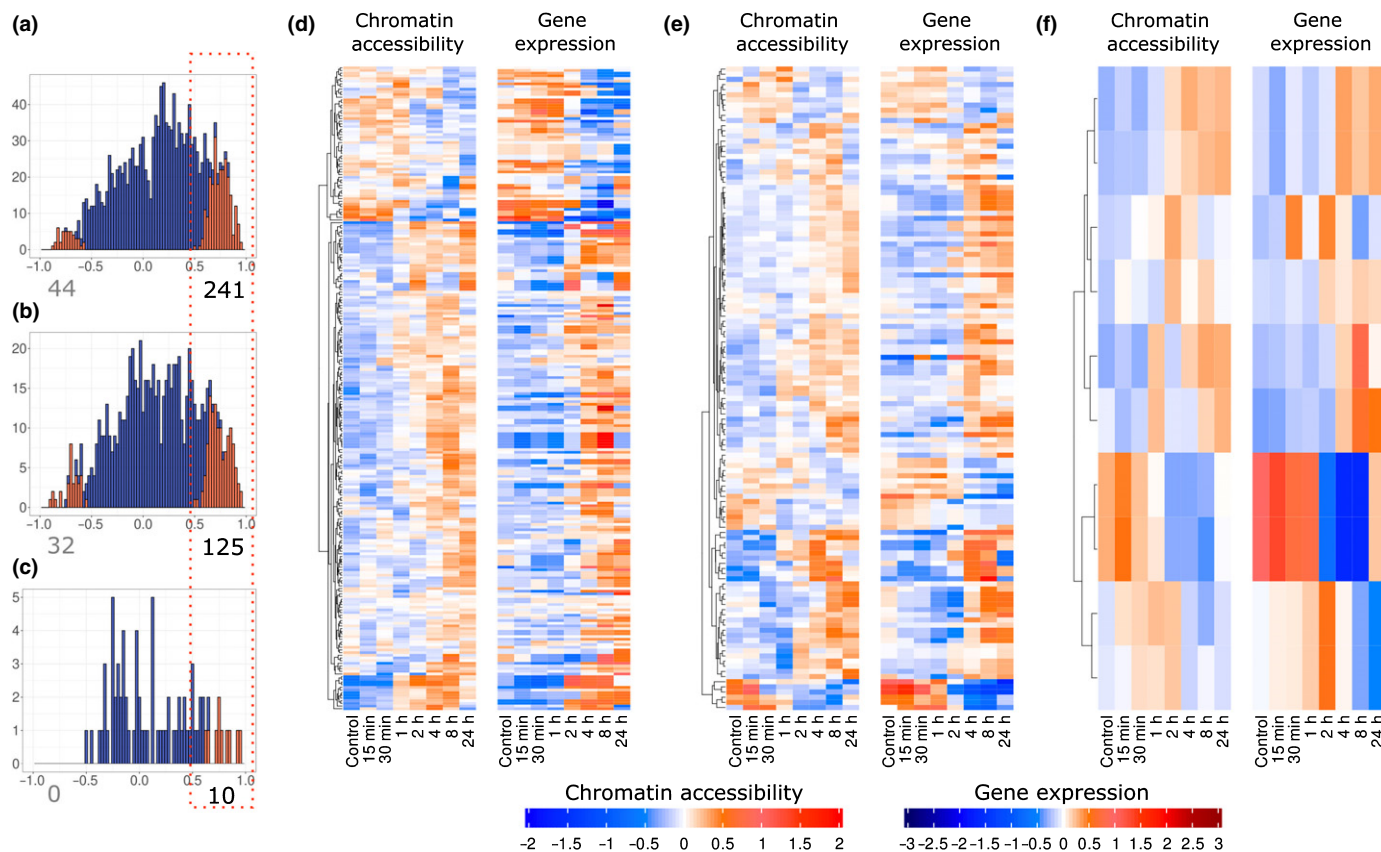


Fig. 4 Chromatin accessibility of nitrogen-fixing clade-specific conserved noncoding sequences (NFC-specific CNS) and (mapped) gene expression profiles positively correlated, shown in the red dashed boxes in (a), (b) and (c), as investigated in *Medicago truncatula*. The profiles shown are based on NFC-specific CNS location relative to the closest mapped gene: (a, d) NFC-specific CNS located upstream (0–2 kb); (b, e) downstream (0–2 kb); and (c, f) distal upstream (2–10 kb). The histograms (a–c) show the distribution of Pearson's correlation of NFC-specific CNS accessibility and gene expression profiles. The number of significant correlations ($P < 0.05$; highlighted in bold) is shown at the bottom of each histogram. For each pair of heatmaps (d–f), chromatin accessibility of the NFC-specific CNS (left) and the expression of the closest (mapped) gene (right) is shown. The rows of the heatmaps (both NFC-specific CNS chromatin accessibility and gene expression) are ordered by hierarchical clustering of the corresponding NFC-specific CNS accessibility profiles (note dendrograms, left of each heatmap). More than one NFC-specific CNS may be associated with a gene, and the heatmaps consequently include repeated gene expression entries. No significant correlation was identified for the NFC-specific CNS classified as distal and downstream of the nearest gene.

family of TFs. In contrast, in other NFC-specific CNS, TFBSs of multiple families of TFs were identified (Notes S8).

Validation of NFC-specific CNS potentially associated with the function of *MtCRE1* in nodulation

Due to its role as a central regulator of nodule organogenesis and the availability of *Mtcre1-1* mutants in the genotype background (Jemalong A17) used in this study (Plet *et al.*, 2011), *MtCRE1* represents a compelling candidate for the experimental investigation of the role of CNS in the regulation of genes related to nitrogen fixation. To test this hypothesis, we evaluated if deletion of the four NFC-specific CNS associated with *MtCRE1* would hinder the occurrence of RNS. We noted that an additional CNS located in the 5' UTR was removed during the filtering steps because it was not possible to recover its corresponding coordinates in the *G. max* genome. Therefore, it was not possible to confirm their orthology. Considering that all other NFC-specific CNS located in *MtCRE1* aligned to an orthologous gene in *G. max* (LOC100789894; also known as Glyma.05G241600), we

considered the absence of alignment to this region of the *G. max* genome to be a possible error and included it in the experimental validation (Fig. S2). To investigate if the five NFC-specific CNS identified in *MtCRE1* are required for nodulation, three versions of *MtCRE1*'s promoter containing deletions of the distal two NFC-specific CNS ($\Delta 2$ CNS), proximal three NFC-specific CNS ($\Delta 3$ CNS) or all NFC-specific CNS ($\Delta 5$ CNS) were engineered.

Composite *M. truncatula* plants in the *Mtcre1-1* background were generated by transforming roots with a construct containing *MtCRE1* either under the WT promoter (positive control) or one of the three engineered promoters. In addition, a construct lacking the *MtCRE1* cassette was used as an empty vector (negative control). Two weeks postinoculation, a gradual decrease in the number of nodules in plants containing the NFC-specific CNS deletions was observed (Fig. 7). Statistical analysis showed a significant reduction in the number of nodules in the comparison between the WT and $\Delta 5$ CNS roots, providing evidence that the NFC-specific CNS are required for nodule organogenesis. Moreover, while the differences were not statistically significant, $\Delta 2$ CNS and $\Delta 3$ CNS showed a reduced number of nodules,

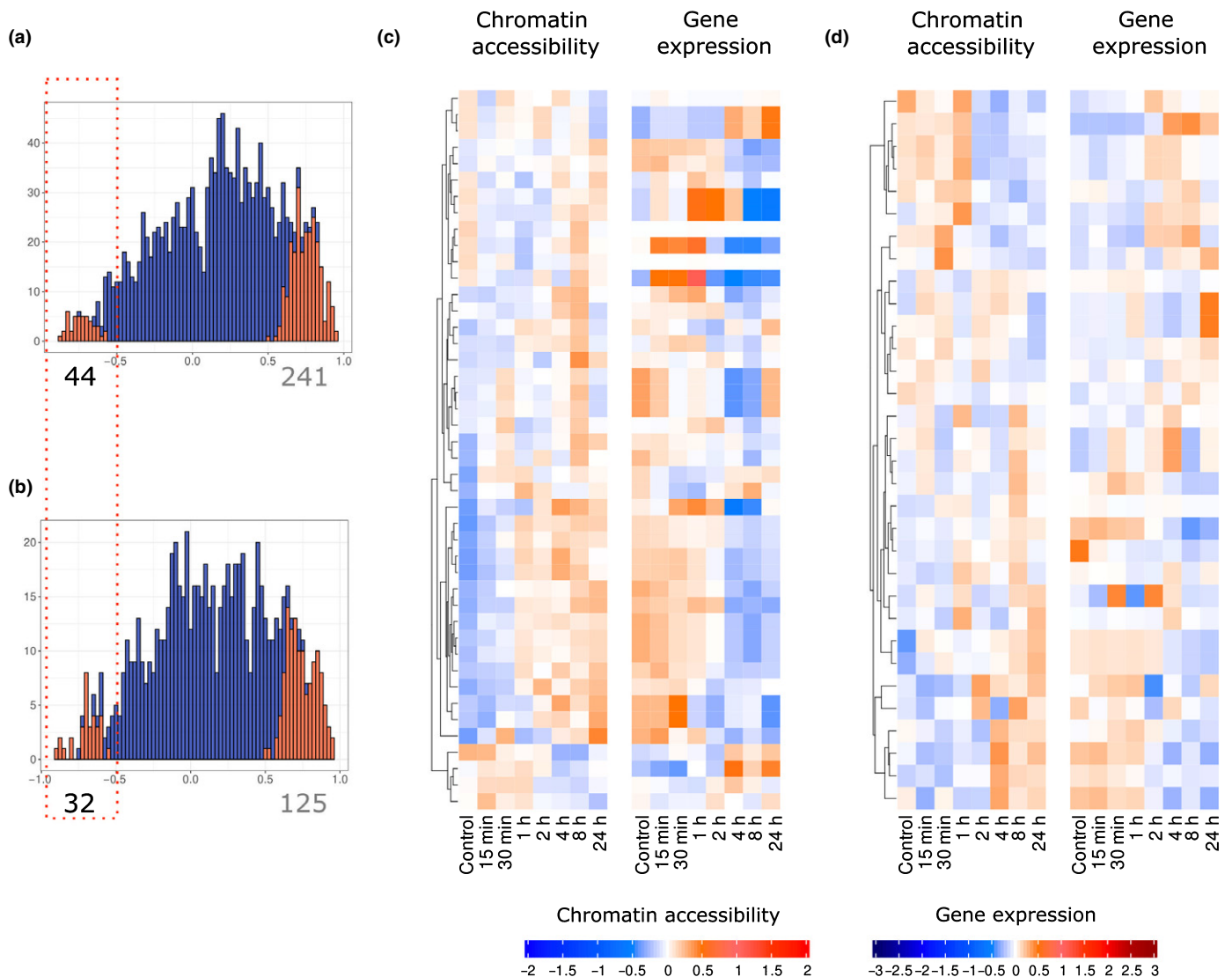


Fig. 5 Chromatin accessibility and gene expression profiles of all nitrogen-fixing clade-specific conserved noncoding sequences (NFC-specific CNS) negatively correlated, shown in the red dashed boxes in (a) and (b), as investigated in *Medicago truncatula*. The profiles are shown according to the NFC-specific CNS location from the closest gene. (a, c) NFC-specific CNS located upstream (0–2 kb); (b, d) downstream (0–2 kb). The histograms in (a) and (b) show the distribution of the correlation values. Significant correlations ($P < 0.05$, highlighted in bold) and the number of significant correlations is given beneath each histogram. For each pair of heatmaps (c, d), the profiles of the NFC-specific CNS chromatin accessibility (left) and expression of the closest gene are shown, where the rows of both are ordered by hierarchical clustering of the accessibility profiles (see dendrograms at the left of heatmaps). Note that more than one NFC-specific CNS can be associated with a gene and, consequently, gene expression can have repeated entries. No significant correlation was identified for the NFC-specific CNS classified as distal (upstream or downstream).

suggesting that deletions of these NFC-specific CNS may partially impair the capacity of the plants to engage in RNS. In addition, investigation of the expression of *MtCRE1* in these plants also revealed a reduction of expression in the $\Delta 5\text{CNS}$ construction (Fig. S7). Similar to the reduction in the number of nodules, the effect on expression was significant only in the $\Delta 5\text{CNS}$ but was also observed in $\Delta 2\text{CNS}$ and $\Delta 3\text{CNS}$. In this assay, expression of the WT (*Mtcre1-1* background transformed with the WT *MtCRE1*) was elevated in comparison to the empty vector. This was probably due to the presence of two copies of the gene, *Mtcre1* and *MtCRE1*, and that both WT and mutant versions of the gene are targeted by the primers used in the assay.

Discussion

Progress has been achieved in uncovering the genomic elements underpinning the capacity of plants to engage in RNS, with more than 200 genes identified to date (Roy *et al.*, 2020). Nonetheless, this has not enabled the long-standing goal of transferring the genetic toolkit required for RNS to plant species lacking this capability. Moreover, comparative genomics indicates that species lacking RNS (including those outside the NFC) contain the majority or all the required genes (Griesmann *et al.*, 2018). However, the gain of a trait can also be driven by the differential regulation of a similar gene ensemble. Thus, the evolution of

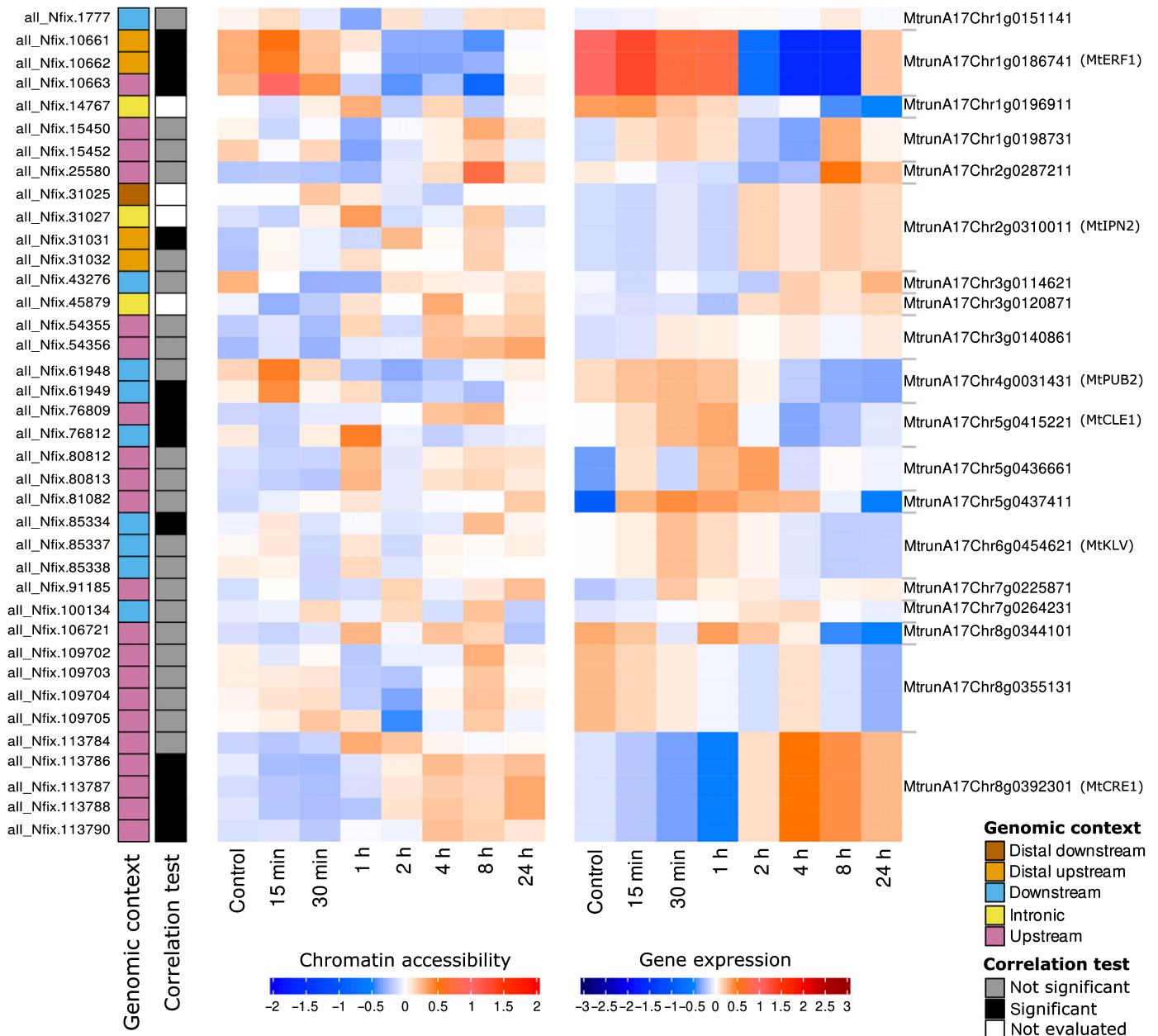


Fig. 6 Chromatin accessibility profile of each nitrogen-fixing clade-specific conserved noncoding sequence (NFC-specific CNS) whose closest gene is a root nodule symbiosis (RNS) gene (left) and their respective gene expression (right) in *Medicago truncatula*. The colors in the bar on the left indicate the genomic location of the NFC-specific CNS from the RNS gene (see key). The second bar from the left indicates if the correlation is significant (black) or not significant (gray). For the intronic NFC-specific CNS regions, the correlation was not calculated (white). Common names are included for genes presenting a significant correlation between their expression and the accessibility profile of at least one associated NFC-specific CNS. Only in the heatmaps of chromatin accessibility (left), the rows were hierarchically clustered. In the gene expression heatmaps, genes are organized in the same row as their associated NFC-specific CNS. Note that more than one NFC-specific CNS can be associated with a gene. Therefore, the heatmap representing gene expression has repeated entries. The functional annotation of these RNS genes is available in Supporting Information Notes S7.

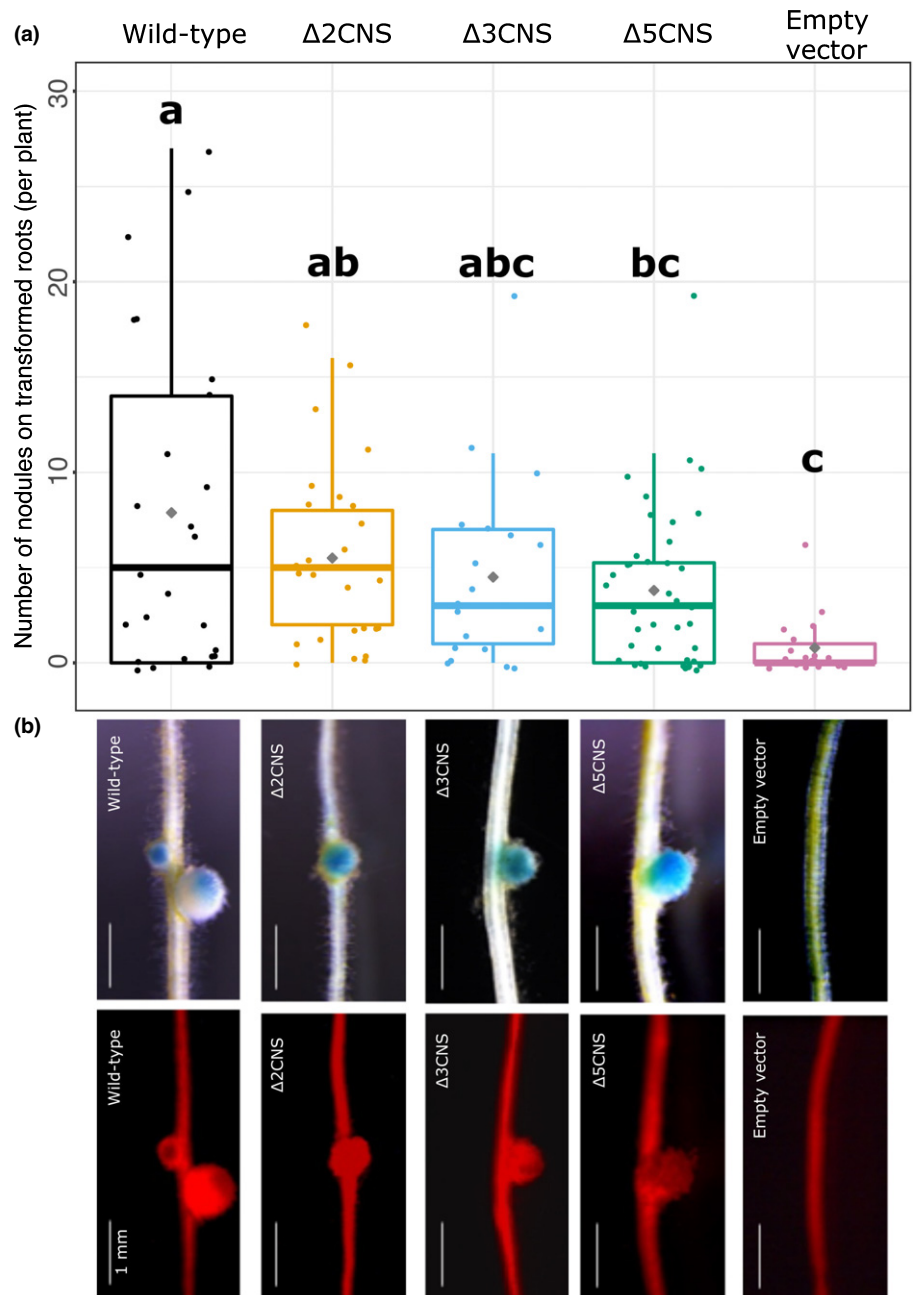
regulatory elements may have been critical for the appearance of RNS in the NFC.

With few exceptions (Liu *et al.*, 2019), the potential role of regulatory elements in RNS has remained largely unexplored. Here, we applied comparative genomics to identify thousands of CNS in nodulating species of the NFC clade. Those NFC-specific CNS have the potential to harbor regulatory elements that are critical for RNS. However, sequence conservation in

these noncoding regions could also be a feature of the NFC, and not directly related to RNS. To identify the NFC-specific CNS that are most prone to affect the function of genes essential to RNS, we evaluated the relationship between their chromatin accessibility and gene expression in *M. truncatula* roots responding to LCOs.

While the correlation between the NFC-specific CNS's chromatin accessibility and gene expression during the treatment with

Fig. 7 Role of the *CRE1* nitrogen-fixing clade-specific conserved noncoding sequences (NFC-specific CNS) in nodulation, in *Medicago truncatula*. (a) The distribution of the total number of nodules on transformed roots per plant. Columns not connected by the same letter(s) are significantly different at $\alpha = 0.05$ (ANOVA followed by Tukey's *post-hoc* test for multiple comparisons). Data from three independent rounds of transformation with $n = 7-19$ composite plants per genotype. In the boxplots, upper and lower hinges correspond to the first and third quartiles (the 25th and 75th percentiles) and the upper (or lower) whisker extends from the hinge to the highest (or lowest) value that is within 1.5 times the interquartile range, or distance between the first and third quartiles. The horizontal line represents the median and the gray diamonds represent the average. (b) Representative X-gal-stained nodules of the transformed roots visualized in the brightfield (upper panel) and tdTomato fluorescence (lower panel). Wild-type denotes the *Mtcre1-1* background transformed with the wild-type *MtCRE1*; $\Delta 2\text{CNS}$, $\Delta 3\text{CNS}$ and $\Delta 5\text{CNS}$ denote the deletion of two, three or five NFC-specific CNS from the wild-type *MtCRE1* gene, respectively.



LCO is a strong indication of functionality, before further validation, the role of these NFC-specific CNS in RNS cannot be established definitively. Also, the lack of correlation between chromatin accessibility and expression profile of the closest gene in NFC-specific CNS does not necessarily disqualify it, or a potential regulatory element within it, from being involved in RNS. These regions could be related to the regulation of genes in which the response is triggered at different stages of the response to the stimulus. Alternatively, they may regulate genes required for rhizobium infection that are not responsive to LCO treatment. Therefore, studies including more time points and following the plant response to the rhizobium infection are still needed to evaluate these NFC-specific CNS thoroughly.

In this study, a conservative approach was used to identify the set of CNS more likely to affect RNS. Regions with an overlap of one or more bases with the coding sequences, TEs, noncoding RNAs and sequences with matches in the nr protein database (Viridiplantae) were removed. Information on these regions is described in a publicly available repository (10.6084/m9.figshare.17124287). The repository contains all whole-genome alignments and the coordinates of all conserved regions identified in this study, facilitating their further exploration.

We validated the function of five NFC-specific CNS located upstream of the gene *MtCRE1*, four of which showed a significant correlation between their chromatin accessibility and the gene expression. The deletion of these regions produced fewer

nodules in *M. truncatula* roots after the infection by *S. meliloti* (Fig. 7), demonstrating that they are required for the correct functioning of *MtCRE1* during RNS. The observed effect may be associated in part with the disruption of the 5' UTR region of *MtCRE1*.

We anticipate that the growing accumulation of plant genomes in public databases, including species within the NFC, will allow the continuing identification of CNS in the NFC. Furthermore, the development of new software capable of generating better whole-genome alignments will provide more power and resolution to the investigation of their role in RNS. In this work, only part of the *M. truncatula* genome was represented in the final multiple-genome alignments. The absence of genomic regions is expected due to species-specific sequence diversification but can also be a consequence of the incapacity of the applied pipeline to produce more complete alignments.

All whole-genome alignments generated in this study used a reference-based approach, which can bias the results. Specifically, using a reference may produce better alignments for species that are phylogenetically close to *M. truncatula* relative to species in more distant branches of the NFC. *Medicago truncatula* was chosen as a reference because it is the species within the NFC with the best genomic resources available, and well-established protocols for nodulation assays are accessible. Nonetheless, investigation of the location of NFC-specific CNS in a second nodulating species, *G. max*, showed that the generated alignment captured the correct orthologous regions between these genomes. While it is not possible to guarantee that the correct orthologous regions were also captured for the other species used in this study, these results indicate strongly that the alignments were not generated spuriously.

Future work will benefit from reference-free multiple whole-genome alignment approaches (Armstrong *et al.*, 2020) as they are not biased by using a reference genome and may capture a larger fraction of the genomes in alignments. However, further bioinformatics developments are necessary for these methods to be computationally efficient in the analysis of large, complex and highly repetitive plant genomes.

In this study we searched for NFC-specific CNS among nodulating species within the clade. The NFC contains species unable to nodulate, and these were excluded from our search for CNS. This choice was made because genome sequence evolution within species not capable of RNS in the NFC imposes challenges to define their possible functional role in nodulation. The loss of any gene or regulatory region critical for RNS and nonessential for other plant growth and development functions is expected to result in the absence of nodulation. In that scenario, all genomic regions involved in the phenotype, and not required for other essential biological processes, are released from purifying selection. The accumulation of mutations in these regions will cause them to diverge and escape detection as conserved sequences by the approach deployed in this study. However, differences in the rate of evolutionary divergence and time of loss of nodulation within the NFC can result in regions related to RNS being detected as conserved in the nonnodulators. As a consequence, excluding NFC-specific CNS because of their detection in

nonnodulators may lead to the removal of CNS that are involved in RNS, but did not diverge sufficiently in the nonnodulating species of the NFC.

A complete understanding of the genomic elements required for a plant to engage in RNS will enable engineering this phenotype in plants outside the NFC. Achieving this goal will require the evaluation of genomes beyond the coding regions, to capture the elements involved in regulating essential genes in the noncoding segments of the genome. Here, we identified hundreds of NFC-specific CNS in *M. truncatula* that potentially affect RNS. Moreover, we show provide evidence that NFC-specific CNS are required for the correct functioning of *MtCRE1* and the establishment of RNS. To the best of our knowledge, this is the first genome-wide study attempting to connect CNS to RNS, providing a foundation for uncovering essential CNS sites in this process.












Acknowledgements

We thank the United States Department of Energy, Office of Science, Biological and Environmental Research program, for funding this study (DE-SC0018247 to MK, SR and J-MA).

Author contributions

WJP, SK, DC, RAF, SR, J-MA and MK planned and designed the research. WJP and SK conducted data analysis. DC, SC and PMT performed experiments. WJP and MK wrote the manuscript. WJP, SK, DC, SC, RAF, PMT, KMB, CD, HWS, J-MA, SR and MK performed data interpretation and edited the manuscript. J-MA, SR and MK supervised the research.

ORCID

Jean-Michel Ané  <https://orcid.org/0000-0002-3128-9439>
 Sanhita Chakraborty  <https://orcid.org/0000-0001-7991-3977>
 Daniel Conde  <https://orcid.org/0000-0001-8362-4190>
 Christopher Dervinis  <https://orcid.org/0000-0002-2176-4232>
 Ryan A. Folk  <https://orcid.org/0000-0002-5333-9273>
 Matias Kirst  <https://orcid.org/0000-0002-8186-3945>
 Sara Knaack  <https://orcid.org/0000-0002-1178-2517>
 Wendell J. Pereira  <https://orcid.org/0000-0003-1019-6281>
 Sushmita Roy  <https://orcid.org/0000-0002-3694-1705>
 Henry W. Schmidt  <https://orcid.org/0000-0003-3917-3219>
 Paolo M. Triozzi  <https://orcid.org/0000-0003-0192-3336>

Data availability

All genomes utilized in this study are available via NCBI and or PHYTOZOME (see Notes S1 for instructions of how to download them). The code to reproduce the analysis is also freely available and can be found on GitHub (https://github.com/KirstLab/CNS_Nitrogen_Fixing_Clade). Both RNA-seq and ATAC-seq data used in this publication were generated and published by Knaack *et al.* (2021). The data have also been deposited in

NCBI's Gene Expression Omnibus (<https://www.ncbi.nlm.nih.gov/geo/>) and are accessible through accession no. GSE154845 (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE154845>). Additional datafiles generated in this study are accessible in a publicly available repository on FigShare (10.6084/m9.figshare.17124287).

References

- Armstrong J, Hickey G, Diekhans M, Fiddes IT, Novak AM, Deran A, Fang QI, Xie D, Feng S, Stiller J *et al.* 2020. Progressive cactus is a multiple-genome aligner for the thousand-genome era. *Nature* 587: 246–251.
- Boisson-Dernier A, Chabaud M, Garcia F, Bécard G, Rosenberg C, Barker DG. 2001. *Agrobacterium rhizogenes*-transformed roots of *Medicago truncatula* for the study of nitrogen-fixing and endomycorrhizal symbiotic associations. *Molecular Plant–Microbe Interactions* 14: 695–700.
- Breakspear A, Liu C, Roy S, Stacey N, Rogers C, Trick M, Morieri G, Mysore KS, Wen J, Oldroyd GED *et al.* 2014. 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* 26: 4680–4701.
- Burgess D, Freeling M. 2014. The most deeply conserved noncoding sequences in plants serve similar functions to those in vertebrates despite large differences in evolutionary rates. *Plant Cell* 26: 946–961.
- Cerri MR, Frances L, Laloum T, Auriac M-C, Niebel A, Oldroyd GED, Barker DG, Fournier J, de Carvalho-Niebel F. 2012. *Medicago truncatula* ERN transcription factors: regulatory interplay with NSP1/NSP2 GRAS factors and expression dynamics throughout rhizobial infection. *Plant Physiology* 160: 2155–2172.
- Chakraborty S, Driscoll H, Abrahante Lloréns J, Zhang F, Fisher R, Harris JM. 2021. Salt stress enhances early symbiotic gene expression in *Medicago truncatula* and induces a stress-specific set of rhizobium-responsive genes. *Molecular Plant–Microbe Interactions* 34: 904–921.
- Engler C, Youles M, Gruetzner R, Ehnert T-M, Werner S, Jones JDG, Patron NJ, Marillonnet S. 2014. A golden gate modular cloning toolbox for plants. *ACS Synthetic Biology* 3: 839–843.
- Freeling M, Subramaniam S. 2009. Conserved noncoding sequences (CNSs) in higher plants. *Current Opinion in Plant Biology* 12: 126–132.
- Frith MC. 2011a. A new repeat-masking method enables specific detection of homologous sequences. *Nucleic Acids Research* 39: e23.
- Frith MC. 2011b. Gentle masking of low-complexity sequences improves homology search. *PLoS ONE* 6: e28819.
- Frith MC, Kawaguchi R. 2015. Split-alignment of genomes finds orthologies more accurately. *Genome Biology* 16: 106.
- Gitzendanner MA, Soltis PS, Wong GK-S, Ruhfel BR, Soltis DE. 2018. Plastid phylogenomic analysis of green plants: a billion years of evolutionary history. *American Journal of Botany* 105: 291–301.
- Gonzalez-Rizzo S, Crespi M, Frugier F. 2006. The *Medicago truncatula* CRE1 cytokinin receptor regulates lateral root development and early symbiotic interaction with *Sinorhizobium meliloti*. *Plant Cell* 18: 2680–2693.
- Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N *et al.* 2012. PHYTOZOME: a comparative platform for green plant genomics. *Nucleic Acids Research* 40: D1178–D1186.
- Griesmann M, Chang Y, Liu X, Song Y, Haberer G, Crook MB, Billault-Penneteau B, Laressergues D, Keller J, Imanishi L *et al.* 2018. Phylogenomics reveals multiple losses of nitrogen-fixing root nodule symbiosis. *Science* 361: eaat1743.
- Gualtieri G, Bisseling T. 2000. The evolution of nodulation. *Plant Molecular Biology* 42: 181–194.
- Hastwell AH, de Bang TC, Gresshoff PM, Ferguson BJ. 2017. CLE peptide-encoding gene families in *Medicago truncatula* and *Lotus japonicus*, compared with those of soybean, common bean and *Arabidopsis*. *Scientific Reports* 7: 9384.
- Haudry A, Platts AE, Vello E, Hoen DR, Leclercq M, Williamson RJ, Forczek E, Joly-Lopez Z, Steffen JG, Hazzouri KM *et al.* 2013. An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. *Nature Genetics* 45: 891–898.
- Jardinaud M-F, Boivin S, Rodde N, Catrice O, Kisiala A, Lepage A, Moreau S, Roux B, Cottret L, Sallet E *et al.* 2016. A laser dissection-RNAseq analysis highlights the activation of cytokinin pathways by Nod factors in the *Medicago truncatula* root epidermis. *Plant Physiology* 171: 2256–2276.
- Jin J, Tian F, Yang D-C, Meng Y-Q, Kong L, Luo J, Gao G. 2017. PLANTTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. *Nucleic Acids Research* 45: D1040–D1045.
- Kent WJ, Baertsch R, Hinrichs A, Miller W, Haussler D. 2003. Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proceedings of the National Academy of Sciences, USA* 100: 11484–11489.
- Kielbasa SM, Wan R, Sato K, Horton P, Frith MC. 2011. Adaptive seeds tame genomic sequence comparison. *Genome Research* 21: 487–493.
- Kirst M, Johnson AF, Baucom C, Ulrich E, Hubbard K, Staggs R, Paule C, Retzel E, Whetten R, Sederoff R. 2003. Apparent homology of expressed genes from wood-forming tissues of loblolly pine (*Pinus taeda* L.) with *Arabidopsis thaliana*. *Proceedings of the National Academy of Sciences, USA* 100: 7383–7388.
- Knaack SA, Conde D, Balmant KM, Irving TB, Maia LG, Triozzi PM, Dervinis C, Pereira WJ, Maeda J, Chakraborty S *et al.* 2021. Temporal change in chromatin accessibility predicts regulators of nodulation in *Medicago truncatula*. *bioRxiv*. doi: 10.1101/2021.08.07.455463.
- Larrainzar E, Riely BK, Kim SC, Carrasquilla-Garcia N, Yu H-J, Hwang H-J, Oh M, Kim GB, Surendrarao AK, Chasman D *et al.* 2015. Deep sequencing of the *Medicago truncatula* root transcriptome reveals a massive and early interaction between nodulation factor and ethylene signals. *Plant Physiology* 169: 233–265.
- Leong SA, Williams PH, Ditta GS. 1985. Analysis of the 5' regulatory region of the gene for delta-aminolevulinic acid synthetase of *Rhizobium meliloti*. *Nucleic Acids Research* 13: 5965–5976.
- Liang P, Saqib HSA, Zhang X, Zhang L, Tang H. 2018. Single-base resolution map of evolutionary constraints and annotation of conserved elements across major grass genomes. *Genome Biology and Evolution* 10: 473–488.
- Liu J, Deng J, Zhu F, Li Y, Lu Z, Qin P, Wang T, Dong J. 2018. The MtDMI2-MtPUB2 negative feedback loop plays a role in nodulation homeostasis. *Plant Physiology* 176: 3003–3026.
- Liu J, Rutten L, Limpens E, van der Molen T, van Velzen R, Chen R, Chen Y, Geurts R, Kohlen W, Kulikova O *et al.* 2019. A remote *cis*-regulatory region is required for NIN expression in the pericycle to initiate nodule primordium formation in *Medicago truncatula*. *Plant Cell* 31: 68–83.
- Mens C, Hastwell AH, Su H, Gresshoff PM, Mathesius U, Ferguson BJ. 2021. Characterisation of *Medicago truncatula* CLE34 and CLE35 in nitrate and rhizobia regulation of nodulation. *New Phytologist* 229: 2525–2534.
- Miyazawa H, Oka-Kira E, Sato N, Takahashi H, Wu G-J, Sato S, Hayashi M, Betsuyaku S, Nakazono M, Tabata S *et al.* 2010. The receptor-like kinase KLAVER mediates systemic regulation of nodulation and non-symbiotic shoot development in *Lotus japonicus*. *Development* 137: 4317–4325.
- Nowak S, Schnabel E, Frugoli J. 2019. The *Medicago truncatula* CLAVATA3-LIKE CLE12/13 signaling peptides regulate nodule number depending on the CORYNE but not the COMPACT ROOT ARCHITECTURE2 receptor. *Plant Signaling & Behavior* 14: 1598730.
- Oka-Kira E, Tateno K, Miura K, Haga T, Hayashi M, Harada K, Sato S, Tabata S, Shikazono N, Tanaka A *et al.* 2005. klavier (klv), a novel hypernodulation mutant of *Lotus japonicus* affected in vascular tissue organization and floral induction. *The Plant Journal* 44: 505–515.
- One Thousand Plant Transcriptomes Initiative. 2019. One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* 574: 679–685.
- Pecry Y, Staton SE, Sallet E, Lelandais-Brière C, Moreau S, Carrère S, Blein T, Jardinaud M-F, Latrasse D, Zouine M *et al.* 2018. Whole-genome landscape of *Medicago truncatula* symbiotic genes. *Nature Plants* 4: 1017–1025.
- Plet J, Wasson A, Ariel F, Le Signor C, Baker D, Mathesius U, Crespi M, Frugier F. 2011. MtCRE1-dependent cytokinin signaling integrates bacterial and plant cues to coordinate symbiotic nodule organogenesis in *Medicago truncatula*. *The Plant Journal* 65: 622–633.

- Pollard KS, Salama SR, King B, Kern AD, Dreszer T, Katzman S, Siepel A, Pedersen JS, Bejerano G, Baertsch R *et al.* 2006. Forces shaping the fastest evolving regions in the human genome. *PLoS Genetics* 2: e168.
- Prabhakar S, Noonan JP, Pääbo S, Rubin EM. 2006. Accelerated evolution of conserved noncoding sequences in humans. *Science* 314: 786.
- Roy S, Liu W, Nandety RS, Crook A, Mysore KS, Pislariu CI, Frugoli J, Dickstein R, Udvardi MK. 2020. Celebrating 20 years of genetic discoveries in legume nodulation and symbiotic nitrogen fixation. *Plant Cell* 32: 15–41.
- Schiessl K, Lilley JLS, Lee T, Tamvakis I, Kohlen W, Bailey PC, Thomas A, Luptak J, Ramakrishnan K, Carpenter MD *et al.* 2019. NODULE INCEPTION recruits the lateral root developmental program for symbiotic nodule organogenesis in *Medicago truncatula*. *Current Biology* 29: 3657–3668.e5.
- Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J *et al.* 2010. Genome sequence of the palaeopolyploid soybean. *Nature* 463: 178–183.
- Seppy M, Manni M, Zdobnov EM. 2019. BUSCO: assessing genome assembly and annotation completeness. *Methods in Molecular Biology* 1962: 227–245.
- Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S *et al.* 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research* 15: 1034–1050.
- Soltis DE, Soltis PS, Morgan DR, Swensen SM, Mullin BC, Dowd JM, Martin PG. 1995. Chloroplast gene sequence data suggest a single origin of the predisposition for symbiotic nitrogen fixation in angiosperms. *Proceedings of the National Academy of Sciences, USA* 92: 2647–2651.
- Song B, Buckler ES, Wang H, Wu Y, Rees E, Kellogg EA, Gates DJ, Khaiphoburch M, Bradbury PJ, Ross-Ibarra J *et al.* 2021. Conserved noncoding sequences provide insights into regulatory sequence and loss of gene expression in maize. *Genome Research* 31: 1245–1257.
- Streng A, op den Camp R, Bisseling T, Geurts R. 2011. Evolutionary origin of rhizobium Nod factor signaling. *Plant Signaling & Behavior* 6: 1510–1514.
- The Angiosperm Phylogeny Group. 2016. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. *Botanical Journal of the Linnean Society* 181: 1–20.
- Tian F, Yang D-C, Meng Y-Q, Jin J, Gao G. 2020. PlantRegMap: charting functional regulatory maps in plants. *Nucleic Acids Research* 48: D1104–D1113.
- Van Bel M, Diels T, Vancaester E, Kreft L, Botzki A, Van de Peer Y, Coppens F, Vandepoele K. 2018. PLAZA 4.0: an integrative resource for functional, evolutionary and comparative plant genomics. *Nucleic Acids Research* 46: D1190–D1196.
- Van de Velde J, Heyndrickx KS, Vandepoele K. 2014. Inference of transcriptional networks in Arabidopsis through conserved noncoding sequence analysis. *Plant Cell* 26: 2729–2745.
- van Velzen R, Holmer R, Bu F, Rutten L, van Zeijl A, Liu W, Santuari L, Cao Q, Sharma T, Shen D *et al.* 2018. Comparative genomics of the nonlegume *Parasponia* reveals insights into evolution of nitrogen-fixing rhizobium symbioses. *Proceedings of the National Academy of Sciences, USA* 115: E4700–E4709.
- Weber E, Engler C, Gruetzner R, Werner S, Marillonnet S. 2011. A modular cloning system for standardized assembly of multigene constructs. *PLoS ONE* 6: e16765.
- Werner GDA, Cornwell WK, Sprent JI, Kattge J, Kiers ET. 2014. A single evolutionary innovation drives the deep evolution of symbiotic N₂-fixation in angiosperms. *Nature Communications* 5: 4087.
- Xiao A, Yu H, Fan Y, Kang H, Ren Y, Huang X, Gao X, Wang C, Zhang Z, Zhu H *et al.* 2020. Transcriptional regulation of NIN expression by IPN2 is required for root nodule symbiosis in *Lotus japonicus*. *New Phytologist* 227: 513–528.

Supporting Information

Additional Supporting Information may be found online in the Supporting Information section at the end of the article.

Fig. S1 Workflow representing the main analytical steps to generate the alignments, define the conserved noncoding sequences (CNS) and select the CNS that are potentially related to root nodule symbiosis.

Fig. S2 Conserved noncoding regions associated with the gene *MtCRE1*.

Fig. S3 Conserved noncoding sequences specific to the nitrogen-fixing clade are in general smaller than the conserved regions in the genome.

Fig. S4 Distribution of the conserved noncoding sequences identified in the nitrogen-fixing clade according to their genome context.

Fig. S5 Phylogenetic tree used to guide the creation of the multiple genome alignment of the nonnodulators within the nitrogen-fixing clade by ROAST.

Fig. S6 Chromatin accessibility profile of the conserved noncoding sequences in response to LCO treatment, as measured by ATAC-seq in a time-series experiment.

Fig. S7 Deletion of conserved noncoding sequences upstream of the gene *MtCRE1* causes reduction of its expression in response to rhizobium infection.

Notes S1 List of genomes used to construct the whole-genome alignments, and descriptive statistics of their assemblies and alignments.

Notes S2 Analysis of the conserved regions removed from the nitrogen-fixing clade-specific conserved noncoding sequences dataset due to overlaps with transposons, noncoding RNAs or presence of hits when using BLAST against the NCBI's nonredundant protein database.

Notes S3 Orthology evaluation of genes from the *Medicago truncatula* and *Glycine max* genomes.

Notes S4 Genomic coordinates of the identified conserved noncoding sequences.

Notes S5 Genomic coordinates of all conserved regions identified in the nonnodulator species within the nitrogen-fixing clade.

Notes S6 Genomic coordinates of the conserved noncoding sequences that overlap with conserved regions of the nonnodulator species within the nitrogen-fixing clade.

Notes S7 Functional annotation of the root nodule symbiosis genes that contain nitrogen-fixing clade-specific conserved noncoding sequences in their vicinity.

Notes S8 Transcription factor motif binding prediction within conserved noncoding sequences that the chromatin accessibility correlates with the expression of a gene involved in root nodule symbiosis.

Table S1 Genome coordinates of the deletions used to remove the five conserved noncoding sequences upstream of the gene *MtCRE1*.

Table S2 Primers used for the quantitation of *MtCRE1* expression via qRT-PCR.

Please note: Wiley Blackwell are not responsible for the content or functionality of any Supporting Information supplied by the authors. Any queries (other than missing material) should be directed to the *New Phytologist* Central Office.