nature plants

LETTERS

https://doi.org/10.1038/s41477-018-0286-7

# Whole-genome landscape of *Medicago truncatula* symbiotic genes

Yann Pecrix[1,13], S. Evan Staton[2,13], Erika Sallet[1,13], Christine Lelandais-Brière[3,4,13], Sandra Moreau[1], Sébastien Carrère[1], Thomas Blein[3,4], Marie-Françoise Jardinaud[5], David Latrasse[3,4], Mohamed Zouine[6], Margot Zahm[6], Jonathan Kreplak[7], Baptiste Mayjonade[1], Carine Satgé[1,12], Magali Perez[3,4], Stéphane Cauet[8], William Marande[8], Céline Chantry-Darmon[8], Céline Lopez-Roques[9], Olivier Bouchez[9], Aurélie Bérard[10], Frédéric Debellé[1], Stéphane Muños[1], Abdelhafid Bendahmane[3,4], Hélène Bergès[8], Andreas Niebel[1], Julia Buitink[11], Florian Frugier[3,4], Moussa Benhamed[3,4], Martin Crespi[3,4,14], Jérôme Gouzy[1,14]* and Pascal Gamas[1,14]*

**Advances in deciphering the functional architecture of eukaryotic genomes have been facilitated by recent breakthroughs in sequencing technologies, enabling a more comprehensive representation of genes and repeat elements in genome sequence assemblies, as well as more sensitive and tissue-specific analyses of gene expression. Here we show that PacBio sequencing has led to a substantially improved genome assembly of *Medicago truncatula* A17, a legume model species notable for endosymbiosis studies[1], and has enabled the identification of genome rearrangements between genotypes at a near-base-pair resolution. Annotation of the new *M. truncatula* genome sequence has allowed for a thorough analysis of transposable elements and their dynamics, as well as the identification of new players involved in symbiotic nodule development, in particular 1,037 upregulated long non-coding RNAs (lncRNAs). We have also discovered that a substantial proportion (~35% and 38%, respectively) of the genes upregulated in nodules or expressed in the nodule differentiation zone colocalize in genomic clusters (270 and 211, respectively), here termed symbiotic islands. These islands contain numerous expressed lncRNA genes and display differentially both DNA methylation and histone marks. Epigenetic regulations and lncRNAs are therefore attractive candidate elements for the orchestration of symbiotic gene expression in the *M. truncatula* genome.**

Because of the wide range of genetic and genomic resources[2], *Medicago truncatula* has been used to study various aspects of legume biology, in addition to bacterial and mycorrhizal fungal endosymbioses, such as organ development (root, leaf, flower, fruit and seed), responses and adaptation to biotic and abiotic stresses, or secondary metabolism. In this study, following the production of a new *M. truncatula* A17 genome sequence, we investigated the functional architecture of the genome, through a study of the *M. truncatula* genes regulated during the development of root nodules following a symbiotic interaction with a nitrogen-fixing bacterium, *Sinorhizobium meliloti* 2011.

The first draft genome assembly (Mt3.5) of *M. truncatula* A17, which was mainly based on a BAC-tiling path with Sanger sequencing and high-quality optical mapping, spanned 418 Mb in 60,143 sequence contigs[3]. A second version (Mt4.0) took advantage of whole-genome shotgun sequencing with high-depth short reads (Illumina), which, combined with the previous data[3], allowed a reduced number of contigs (10,160) spanning 412 Mb to be obtained[4]. To further improve the genome assembly, we used high-depth (more than 100×) long-read (PacBio) sequencing, as well as previous[3] and new (BioNano technology) optical maps. Following a meta-assembly protocol based on a combination of several assemblers (Supplementary Note I, Supplementary Fig. 1 and Supplementary Table 1), a highly contiguous reference genome of 430 Mb (termed Mt5.0) was generated in only 64 sequence contigs (including 3.59 Mb in 32 unanchored contigs). This assembly spans chromosome arms from telomeres to centromeres, with half of the eight *M. truncatula* pseudo-chromosomes containing a single sequence gap at the centromere position (Fig. 1). The genome structure was compared with the previous *M. truncatula* A17 Mt4.0[4] and R108[5] genome assemblies and differences were assessed in light of the synteny with closely related legume species (Supplementary Note I.5 and I.6). The chromosomal rearrangement that was previously identified in A17 (translocation between chr4 and chr8[6]) could thus be localized with a 14-nt resolution on the R108 genome (Supplementary Note I.6.1.1), illustrating the advantage of having several high-resolution PacBio genomes of the same species. At the gene level, the improved quality of this new assembly is illustrated by the presence of two sets of duplicated genes missing in the previous Mt4.0 version (namely 7 *CEP*[7] and 12 *CRP*[8] genes; Supplementary Fig. 2 and Supplementary Note I.6.4).

Next, a deep genome annotation was performed using strand-oriented transcriptomic data[9,10] and strand-specific gene modelling, leading to the identification of 44,623 inferred protein-coding genes and 4,081 lncRNAs (no coding sequence (CDS) > 39 amino acids) (Supplementary Note II). In addition, structure- and homology-based

[1] LIPM, Université de Toulouse, INRA, CNRS, Castanet-Tolosan, France. [2] University of British Columbia, Vancouver, Canada. [3] IPS2, CNRS, INRA, Universities of Paris Diderot and Sorbonne Paris Cité, Gif sur Yvette, France. [4] IPS2, CNRS, INRA, Universities of Paris Diderot, Paris Sud, Evry and Paris-Saclay, Gif sur Yvette, France. [5] LIPM, Université de Toulouse, INPT, ENSAT, Castanet-Tolosan, France. [6] GBF, Université de Toulouse, INPT, ENSAT, Castanet-Tolosan, France. [7] AGROECOLOGIE, INRA, Dijon, France. [8] CNRGV, INRA, Castanet-Tolosan, France. [9] INRA, US1426, GeT-PlaGe, Genotoul, Castanet-Tolosan, France. [10] INRA, US 1279 EPGV, Université Paris-Saclay, Evry, France. [11] IRHS, Agrocampus-Ouest, INRA, Université d'Angers, Beaucouzé, France. [12] Present address: CNRGV, INRA, Castanet-Tolosan, France. [13] These authors contributed equally: Y. Pecrix, S. Evan Staton, E. Sallet, C. Lelandais-Brière. [14] These authors jointly supervised this work: J. Gouzy, M. Crespi, P. Gamas. *e-mail: Jerome.Gouzy@inra.fr; Pascal.Gamas@inra.fr

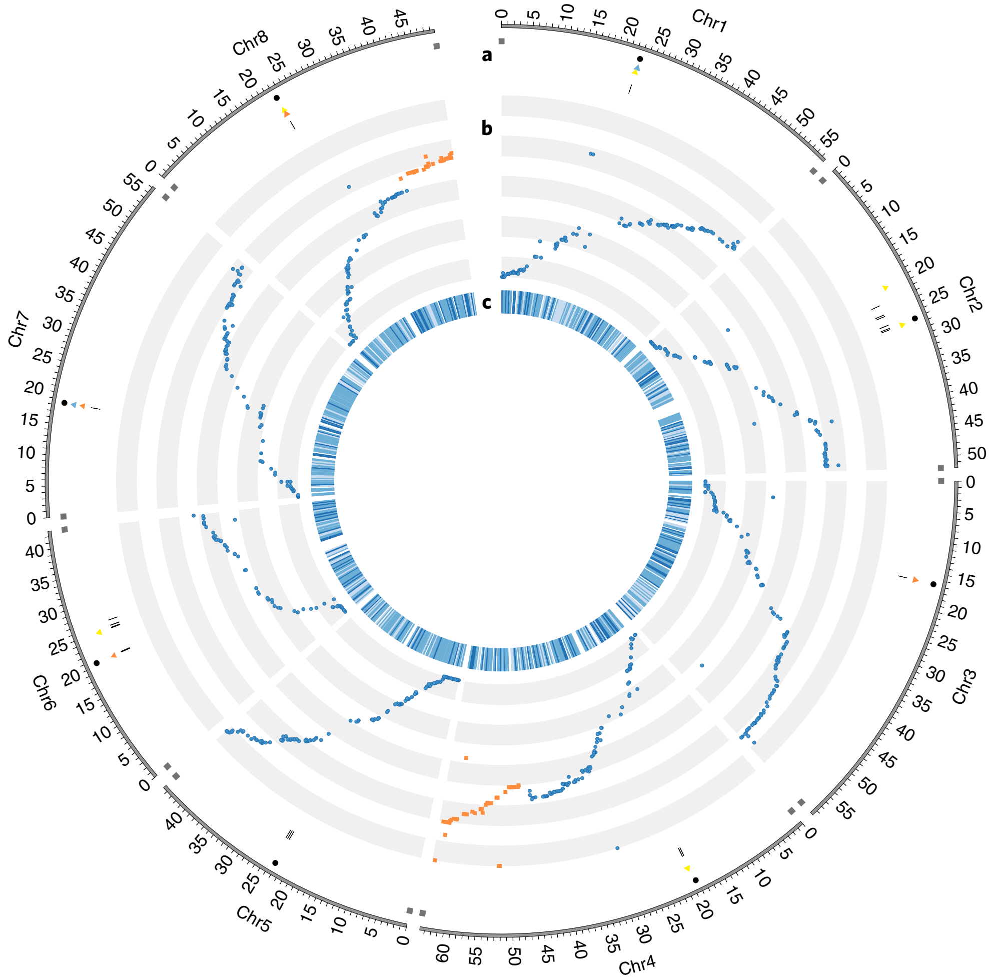
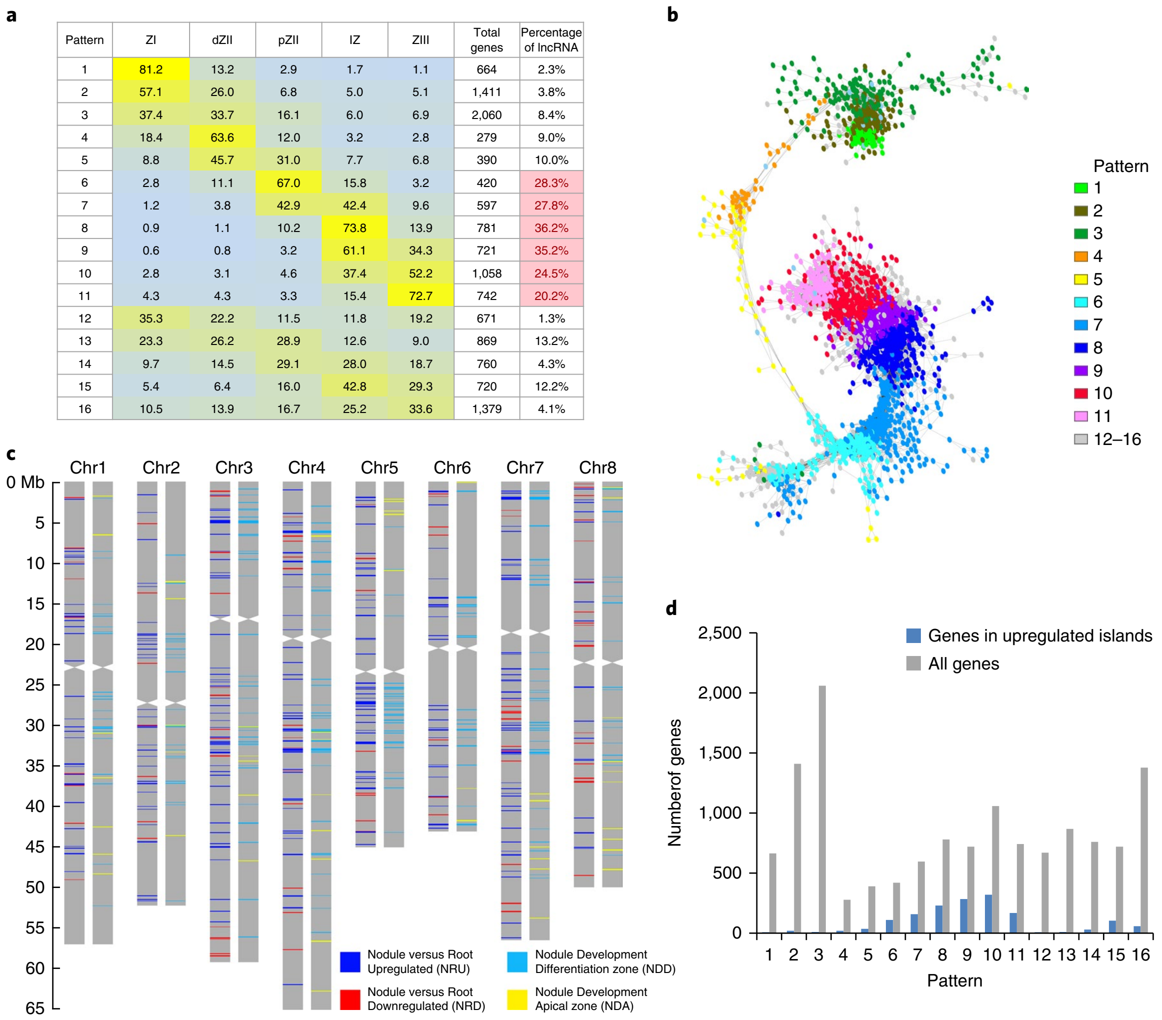


**Fig. 1 | Overview of the *M. truncatula* A17 genome. a**, Centromere position (black circles), telomere sequences (grey rectangles), pericentromeric repeats MtR1 (blue triangles), MtR2 (yellow triangles), MtR3 (orange triangles), sequence gaps (black bars). **b**, Blue circles represent markers of the genetic map[3] (0–90 cM) mapped with e-PCR software (maximum mismatches, 2; maximum amplicon size, 2,000), orange rectangles at the end of Chr4 and Chr8 represent markers associated with the translocation in the A17 genotype (that is associated with LG8 and LG4, respectively; see Supplementary Note I.6.1.1 for more information on the translocation boundaries). **c**, Heat map representing the congruence between the *Bam*HI optical map[3] and the sequence: a sliding window of 500 kb, on a four-colour scale: white, light blue, blue, dark blue (from the lowest to the higher density score of the alignment). White regions represent gaps in the *Bam*H1 map.

tools identified 24,645 intact transposable elements (~24% of the assembled genome; Supplementary Table 2), belonging to more than 2,600 families and with divergent patterns of historical activity, genomic distribution and host gene incorporation (Supplementary Fig. 3, Supplementary Note III and Supplementary Table 3). Small RNA (sRNA) populations were also characterized, using a large set of sRNA libraries, including root, developing and mature nodule libraries (Supplementary Note II.3). We identified 1,402 micro RNA (miRNA) genes, including 376 belonging to novel families, and 167,853 short interfering RNA (siRNA) clusters (Supplementary Table 4 and Supplementary Table 5, respectively). An integrative web portal https://medicago.toulouse.inra.fr/MtrunA17r5.0-ANR/ enables an easy transfer of knowledge from all *M. truncatula* gene nomenclatures used over the last 20 years (NickNames), with the visualization of gene models from previous genome sequence releases[3,4], in addition to transcriptomic, epigenetic and natural variation (HapMap project[11]; 262 *Medicago* accessions, 38 million single nucleotide polymorphisms; Supplementary Note II.4) datasets.

To identify genes involved in symbiotic nodule development, RNAseq data were analysed from roots and nodules[9], encompassing five laser-capture-microdissected (LCM) nodule zones. A total of 5,513 genes (~18.2% of all expressed genes) were identified that are strongly upregulated in nodules as compared to roots (log$_2$-fold change (LFC) $> 2$, false discovery rate (FDR) $< 0.01$; Supplementary Table 6). The upregulated transcripts include 1,037 lncRNAs, 996 of which were previously undescribed (average and median size: 1,525 nt and 921 nt, respectively). In contrast, from the 3,997 downregulated genes (LFC $< -2$, FDR $< 0.01$) only 132 transcripts represented lncRNAs. Hierarchical clustering of the LCM RNAseq data distinguished 16 expression patterns (FDR $< 0.001$; Fig. 2a and Supplementary Table 6), related to their spatial regulation in the five microdissected nodule zones (meristematic region (zone I); (pre) infection region (distal zone II); early and late differentiation zone (proximal zone II and interzone II–III, respectively); nitrogen-fixation zone (zone III)). A coexpressed gene network illustrates the successive waves of genes involved, from the nodule apical genes to the nitrogen-fixation genes (Fig. 2b). Interestingly, a majority (67%) of the 2,783 lncRNAs detected in the LCM RNAseq data can be found in expression patterns 6 to 11, corresponding to transcripts upregulated in the nodule differentiation region (Fig. 2a). In comparison,



| Pattern | ZI | dZII | pZII | IZ | ZIII | Total genes | Percentage of lncRNA |
|---|---|---|---|---|---|---|---|
| 1 | 81.2 | 13.2 | 2.9 | 1.7 | 1.1 | 664 | 2.3% |
| 2 | 57.1 | 26.0 | 6.8 | 5.0 | 5.1 | 1,411 | 3.8% |
| 3 | 37.4 | 33.7 | 16.1 | 6.0 | 6.9 | 2,060 | 8.4% |
| 4 | 18.4 | 63.6 | 12.0 | 3.2 | 2.8 | 279 | 9.0% |
| 5 | 8.8 | 45.7 | 31.0 | 7.7 | 6.8 | 390 | 10.0% |
| 6 | 2.8 | 11.1 | 67.0 | 15.8 | 3.2 | 420 | 28.3% |
| 7 | 1.2 | 3.8 | 42.9 | 42.4 | 9.6 | 597 | 27.8% |
| 8 | 0.9 | 1.1 | 10.2 | 73.8 | 13.9 | 781 | 36.2% |
| 9 | 0.6 | 0.8 | 3.2 | 61.1 | 34.3 | 721 | 35.2% |
| 10 | 2.8 | 3.1 | 4.6 | 37.4 | 52.2 | 1,058 | 24.5% |
| 11 | 4.3 | 4.3 | 3.3 | 15.4 | 72.7 | 742 | 20.2% |
| 12 | 35.3 | 22.2 | 11.5 | 11.8 | 19.2 | 671 | 1.3% |
| 13 | 23.3 | 26.2 | 28.9 | 12.6 | 9.0 | 869 | 13.2% |
| 14 | 9.7 | 14.5 | 29.1 | 28.0 | 18.7 | 760 | 4.3% |
| 15 | 5.4 | 6.4 | 16.0 | 42.8 | 29.3 | 720 | 12.2% |
| 16 | 10.5 | 13.9 | 16.7 | 25.2 | 33.6 | 1,379 | 4.1% |

**Fig. 2 | Symbiotic gene expression patterns and organization in *M. truncatula*. a**, Hierarchical clustering analysis of genes exhibiting differential expression amongst five laser-dissected nodule zones (ZI: zone I, meristematic region; dZII: distal zone II, (pre)infection region; pZII: proximal zone II, early differentiation region; IZ: interzone II–III, late differentiation zone; ZIII: zone III, nitrogen-fixation zone). Expression patterns 1 to 11 exhibit very strong differences between nodule zones, while patterns 12 to 16 show gradual differences (with one preferential zone). **b**, Gene network analysis of genes exhibiting differential expression amongst five laser-dissected nodule zones, using Pearson correlation; different colours correspond to the different patterns defined in **a**. **c**, Symbiosis-related genomic islands represent physical clusters of genes strongly up- or downregulated when comparing whole nodules versus root systems (blue and red lines, respectively) or differentially regulated between laser-dissected nodule zones (specifically expressed in the nodule apex or differentiation zone: yellow and light blue, respectively). **d**, Most of the NRU genes belong to the differentiation patterns defined in **a** (preferential expression in pZII or IZ).

only 26% of the differentially expressed messenger RNAs exhibit these patterns.

Bearing in mind that lncRNAs and antisense transcripts can regulate neighbouring mRNAs in plants[12–16], we then asked whether the expression of nodule lncRNAs and their mRNA neighbour(s) was correlated in roots and/or nodules[9]. We found that the expression profiles of 1,693 lncRNAs were positively correlated with the closest mRNA in the five nodule zones (Spearman correlation, FDR < 0.05; Supplementary Table 7 and Supplementary Note IV.1), whereas only 334 lncRNAs were negatively correlated with the closest mRNA. The positively correlated lncRNA–mRNA pairs were physically closer than the negatively correlated pairs (average median distance of ~3.46 kb versus ~14.3 kb, when considering the five nodule zones; Supplementary Table 7). We also identified hundreds of overlapping antisense lncRNA–mRNA or mRNA–mRNA pairs, being generally positively correlated (Supplementary Table 7 and Supplementary Note IV.1). An interesting example is the Mt*EFD* gene, a key regulator of nodule development[17], and its antisense transcript, which are both induced in nodules but in different zones (Supplementary Table 7). A positive correlation at the whole organ level can therefore hide significant differences at the tissue level. A whole set of other symbiotic effectors and regulator genes (for example, Mt*DME*[18], Mt*NSP2*[19], Mt*RSD*[20], Mt*NIN*[21], Mt*IPD3*[22], Mt*SYMREM1*[23], Mt*SYMCRK*[24]; Supplementary Table 7) exhibit correlated lncRNAs or antisense transcripts, the biological importance of which will be interesting to explore in future studies.

Taking advantage of the highly continuous genome sequence now available, we then investigated the physical distribution on the genome of the genes differentially regulated during nodule development. Following our observation that coregulated genes often colocalize, we set up an automatic identification of what we termed symbiosis-related islands (SRIs). A genomic region was called SRI when ≥4 coregulated and colocalized genes represented >60% of the expressed genes (see Methods and Supplementary Note V.1).

**Table 1 | Features of SRIs**

| Island types | Total number of genes[a] with relevant pattern[b] | Number of islands | Mean island size[c] (nt) | Number of genes with relevant pattern[b] in islands | Mean fold change[c] of island genes (median) | Number of expressed island lncRNAs (number and percentage of islands with lncRNAs) | Mean number[c] of genes with relevant pattern[b] per island [d] (max) | Mean percentage[c] of genes with relevant pattern[b] per island | Mean number[c] of different coregulated gene families per island |
|---|---|---|---|---|---|---|---|---|---|
| NRU | 5,499 | 270 | 45,964 ± 2,629 | 1,960 | 758.3 ± 24.8 (295.4) | 431 (188; 69.6%) | 7.26 ± 0.31 (39) | 86.9 ± 0.77 | 5.58 ± 0.22 |
| NRD | 3,981 | 89 | 48,178 ± 2,457 | 468 | 35.6 ± 2.9 (12.2) | 14 (12; 13.5%) | 5.26 ± 0.23 (18) | 82.0 ± 1.4 | 3.04 ± 0.15 |
| NRN[e] | 5,102 | 84 | 51,203 ± 2,076 | 396 | 1.0 ± 0.0 (1.0) | 3 (3; 3.6%) | 4.71 ± 0.12 (9) | 81.4 ± 1.5 | 4.30 ± 0.15 |
| NDA | 4,068 | 49 | 52,167 ± 3,501 | 242 | | 20 (10; 20.4%) | 4.94 ± 0.20 (10) | 80.6 ± 1.8 | 4.31 ± 0.19 |
| NDD | 4,309 | 211 | 37,897 ± 2,692 | 1,558 | | 552 (192; 91.0%) | 7.38 ± 0.37 (53) | 87.1 ± 0.82 | 6.14 ± 0.26 |
| NDN[f] | 4,142 | 57 | 44,939 ± 2,681 | 275 | | 21 (15; 26.3%) | 4.82 ± 0.15 (10) | 80.6 ± 1.6 | 4.16 ± 0.14 |

[a] Nuclear genes, whether in islands or not. [b] That is coregulated (NRU, NRD, NDA, NDD) or non-regulated (NRN, NDN). [c] ±s.e.m. [d] Considering only expressed genes. [e] Nodule versus root LFC < 0.25 and > −0.25. [f] FDR zone effect = 1 and non-adjusted $P > 0.08$ (global comparison of the five laser-dissected nodule zones). Non-regulated gene islands were used as a control to evaluate the impact of *M. truncatula* gene duplications and the frequency of serendipitous clustering of genes with similar patterns.

At the organ level, two types of SRI were defined, referred to as nodule versus root upregulated (NRU) and downregulated (NRD). At the tissue level, two SRIs were identified and referred to as nodule development, differentiation (NDD) and apical zone (NDA) (LCM data). In addition, control islands were defined from non-regulated genes for organ (NRN) and tissue (NDN) data. A total of 1,960 genes, representing 35.6% of total nodule upregulated genes, colocalized in 270 NRU islands (12.4 Mb in total; Fig. 2c, Table 1 and Supplementary Table 8), a number significantly higher than that obtained by random sampling using the same pipeline (Supplementary Note V.1). These NRU islands have a mean size of ~46 kb with ~7.3 coregulated genes on average, representing ~87% of the expressed genes in the islands (Table 1). In contrast, only 11.7% (468 genes) of the downregulated genes are located in 89 NRD islands while 7.8% (396 genes) of the non-regulated genes ($0.25 < \mathrm{LFC} < 0.25$) are present in 84 NRN islands (Table 1). The NRU islands include numerous lncRNA genes (431 versus 14 and 3 in downregulated and non-regulated islands, respectively; Table 1) and mostly genes expressed in the nodule differentiation zone (Fig. 2d). Analysis on the tissue level identified 211 NDD islands with 1,558 genes mostly expressed in the differentiation zone (expression patterns 6 to 11; Fig. 2a,c and Supplementary Table 8), with 143 NDD islands overlapping with NRU islands. By contrast, only 49 and 57 islands were found with genes expressed in the nodule apex (NDA; expression patterns 1 to 3) (Fig. 2c) or non-spatially regulated (NDN), respectively (Table 1). A comparative analysis with the R108 genotype showed that, inside the different SRIs, the percentage of conserved genes between A17 and R108 varies from 81.4% to 95.8% (NDD and NDA expressed genes, respectively; Supplementary Table 9 and Supplementary Note V.7).

Gene duplications are known to be frequent in *M. truncatula*[3,8] and genomic clusters probably resulting from local gene amplifications have been reported for nodule-associated genes, notably encoding nodule-specific cysteine-rich peptides (NCRs[3,25–27]), glycine-rich proteins (GRPs[25]) or calmodulin-like proteins (CaML[28]). It was therefore important to assess whether SRIs are not simply due to gene duplications. Indeed we found that many of these genomic clusters are present within SRIs. However, we discovered that they group together with other coregulated gene types, in particular encoding lncRNAs and short hypothetical proteins, which together represent about half of the NRU and NDD expressed genes (Supplementary Notes V.2.1 and V.3; see Supplementary Fig. 4 for an example). Thus, while NCR genes are the most abundant class in SRIs (18.6% and 22.8% of NRU and NDD genes, respectively), they are found at > 1 copy in only 82 NRU (~30%) and 91 NDD (~43%) islands. More generally, nodule-associated genes known to be locally amplified in clusters (NCRs[3,25–27], GRPs[25], CaMLs[28], LEED…PEED (LP) antimicrobial proteins and defensin-like proteins[7]) represent no more than 21% and 25% of NRU and NDD expressed genes, respectively, while on average > 5 distinct gene families (including lncRNA genes) are found per NRU and NDD island (Table 1 and Supplementary Note V.3). This indicates that gene duplications cannot be the sole explanation of gene organization in islands.

As for the biological function of NRU and NDD protein-coding genes, an amplification of NCR and defensin-like genes (with conservation of the expression pattern) probably enables protein functional diversification, potentially useful for the plant to cope with rhizobium diversity and evolution[27], as well as possible coinfecting opportunist microbes[29]. Several other important symbiotic genes are also expressed from NRU and NDD islands (Supplementary Note V.2) (for example, Mt*ENOD11*, Mt*ENOD12*, Mt*ENOD40*, Mt*RPG*, Mt*SYMREM1*, Mt*RSD*, Mt*SYMCRK*, Mt*IRE*), as well as genes encoding ERF transcription factors, C2 domain ($Ca^{++}$-dependent membrane-targeting module) proteins, and a large number of peptides (383 and 270 hypothetical proteins < 100 amino acids (aa) from NRU and NDD islands, respectively), potentially representing new symbiotic players.

To better understand the regulation mechanisms, one hypothesis to be tested was that the local chromatin structure could contribute to the coregulation of neighbour gene sets. We therefore examined factors potentially impacting chromatin structure and gene expression, namely the DNA methylation status, small non-coding RNAs (ncRNAs) and histone marks. We focused on the nodule differentiation (NDD) islands, precisely defined by LCM data. Figure 3a recalls the large number of lncRNA genes in NDD islands, as compared to apical or non-spatially regulated islands.

We first analysed the distribution of differentially methylated regions (DMRs) previously found to be associated with nodule development, using a genomic capture approach for specific regions[18] (~12.4 Mb in total). The differentiation islands (but not the apical or non-regulated islands) are well represented on the

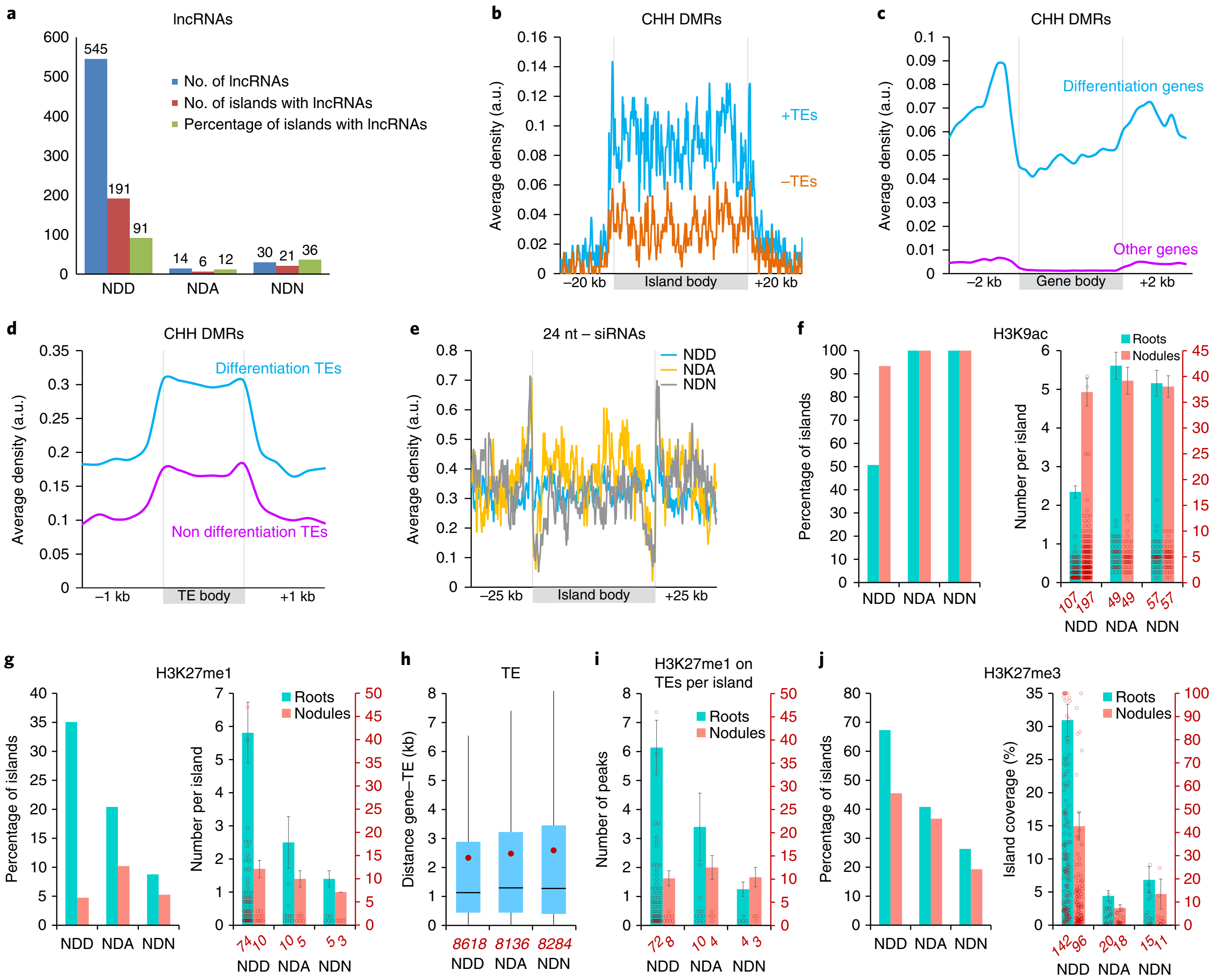


**Fig. 3 | Analysis of potential regulatory elements in nodule development islands.** **a**, lncRNA abundance. **b**, Profile of DMRs, CHH context, in NDD islands (length normalized to 50 kb) and 20 kb flanking regions, with (blue) or without (brown) DMRs intersecting with transposable elements and repeats. **c**, Average CHH DMR profile at the gene level (gene bodies normalized to 3 kb; 1,324 differentiation and 1,189 'other' genes, neither upregulated nor expressed in the nodule differentiation zone. **d**, CHH DMR profile on transposable elements and repeats (and 1 kb flanking regions) within or outside NDD islands. **e**, Profile of 24-nt siRNA clusters on islands (length normalized to 50 kb) and 25 kb flanking regions; the siRNA peaks at island ends reflect the presence of a gene at each island end. **f**,**g**, Percentage of islands with H3K9ac and H3K27me1 marks in roots and nodules (left) and average histone peak number (right). **h**, Distance between island genes and closest upstream and downstream transposable elements (TEs). The boxes show the second and third quartiles together with the median and average values (horizontal lines and red dots, respectively). **i**, Average number of H3K27me1 peaks on transposable elements and repeats per island showing those peaks. **j**, Percentage of islands with H3K27me3 marks in roots and nodules and average coverage per island showing these marks. NDD, nodule development differentiation islands; NDA, nodule development apical islands; NDN, nodule development non-spatially regulated islands; the number of islands (**f**, **g**, **i**, **j**) or genes (**h**) considered are given below the *x* axis. In (**f**, **g**, **i**, **j**), the data are plotted as dots in red using a second *y* axis to display the whole data distribution. Error bars represent s.e.m.

captured DNA (170 NDD islands, for 3.7 Mb on the captured DNA). We found that 129 NDD islands carry 13.1 CHH (three-nucleotide cytosine context where H can be A, C or T) DMRs on average (covering 8.8% of the island), with a strong increase as compared to the flanking regions (Fig. 3b). The CHH DMR density is maximal at gene promoter regions (Fig. 3c) and transposable element bodies (Fig. 3d). Subtracting the transposable element-associated DMRs did not qualitatively change the DMR profile (Fig. 3b). Together with CHH DMRs, we also observed CG-CHG DMRs in 95 islands, but with a lower frequency (on average 3.0 DMRs per island and 2.2% island length coverage).

Next, the distribution of small ncRNAs was analysed, revealing 24-nt siRNAs that matched the CHH DMR profile at gene and transposable element level (Supplementary Note IV.2). No obvious difference in siRNA accumulation could be observed between the islands and their flanking regions or between differentiation and non-regulated (NDN) islands (Fig. 3e), even though hundreds of gene bodies and promoter regions exhibit differential siRNA signal during nodule development (Supplementary Fig. 5 and Supplementary Table 5). In addition, we identified 1,239 miRNA targets in the islands, from which 156 corresponded to differentially accumulated miRNAs (Supplementary Table 4). There was, however, no specific enrichment of miRNA targets in the NDD islands (Supplementary Note V.4). This suggests that the miRNA or the RNA-dependent DNA methylation pathways[30] are not major global regulatory mechanisms of the symbiotic islands, even though they are certainly important for the regulation of individual genes (including epigenetic regulators; Supplementary Note IV.2).

Finally, we examined three repressive histone marks (two heterochromatic: H3K9me2 and H3K27me1; one euchromatic: H3K27me3) and one active mark (H3K9ac) by whole-genome ChIP-seq (Supplementary Fig. 6 and Supplementary Note V.6). The level of H3K9ac marks in the differentiation islands was higher in nodules than roots, as expected, while it was similar in both organs for apical and non-regulated islands (Fig. 3f). By contrast, the two heterochromatic marks were much more abundant in roots than in nodules, both in terms of number of islands and peaks per island, particularly in differentiation islands (Fig. 3g; example in Supplementary Fig. 4). Genome-wide analyses indicated that a major fraction (>85%) of the H3K27me1 and H3K9me2 marks are found on transposable elements and repeats. Interestingly, while the global transposable element composition and the distance between transposable elements and island genes do not differ much between the three island types (Fig. 3h and Supplementary Note V.5), the differentiation islands (and to a lower extent the apical islands) appear more targeted by heterochromatic marks in roots than the non-regulated islands (Fig. 3i). Transposable element/repeat-related epigenetic regulation might thus play a role in the repression of symbiotic island genes in roots (~35% and ~42% of differentiation islands carrying H3K27me1 and H3K9me2 marks, respectively). Moreover, the euchromatic repressive mark H3K27me3 is also more frequent in differentiation islands (~67% on average in roots, with a coverage of ~31% of the island length, versus ~26% of non-regulated islands with a coverage of ~7% of the island), again with a strong decrease in nodules (Fig. 3j).

In conclusion, a complete high-quality genome assembly together with sensitive and tissue-specific RNA sequencing approaches has proved to be extremely valuable to refine the structure and the functional organization of the *M. truncatula* genome and to improve our understanding of the gene regulation during nodulation. The physical clustering of nodule-expressed genes is potentially valuable for the coinheritance of favourable gene sets[31] and for the cell economy, through the coordinated gene regulation over entire genomic regions. It is somewhat reminiscent of gene clustering for specific biosynthetic pathways in various plants[32] or for stamen development in *Arabidopsis thaliana*[33] (Supplementary Note V.2.5). One peculiar feature of the symbiotic NDD and NRU islands, however, is the abundance of lowly expressed but strongly regulated transcripts (lncRNAs and peptide-encoding mRNAs). A conservative hypothesis would be that these transcripts are produced from cryptic promoters as a result of the strong transcriptional activation of genomic regions. Alternatively, some of these peptides may encode new signalling molecules, as described in various developmental processes[7,34], while lncRNAs might actually be directly involved in the transcriptional activation of the symbiotic islands, along with epigenetic marks. Indeed, recent works have highlighted the possible importance for the genome structuring of physically clustered genes with a shared transcriptional status[35–37], in addition to the known impact of lncRNAs on chromatin conformation, documented both in animal and plant cells[12,38,39]. An attractive model was recently proposed where the transcription of low-abundance tissue-specific lncRNAs could serve as guide-posts for shaping the three-dimensional genome organization[37]. Future studies will precisely decipher the role of these new actors of symbiosis.

## Methods

**Pacbio sequencing.** DNA was extracted as described[40] from *M. truncatula* A17 leaves. The following steps were performed using the GeT-PlaGe core facility (INRA, Toulouse, France; http://get.genotoul.fr). Seven libraries were produced, using the SMRTbell Template Prep Kit 1.0, each from 7.5 μg DNA fragmented with a Megaruptor system (Diagenode, RRID:SCR_014807). DNA was repaired (SMRTbellTM Template Prep Kit 1.0; PacBio) then ligated to hairpin adapters. Following digestion of incompletely formed SMRTbell templates with Exonuclease III and VII, DNA molecules between 17 and 90 kb were selected by BluePippin electrophoresis (Sage Science, RRID:SCR_014808). Sequencing was performed using a PacBio RS II with 49 SMRT Cells (P6/C4 chemistry, 6 h run), yielding a total estimated genome coverage of 109× (see Supplementary Note I.1).

**Genome assembly.** The 4,367,592 PacBio raw subreads (N50: 18,347 bp; 53.93 Gb in total) were first corrected with a modified version of PBcR[41,42] aiming at reducing computational time and storage[43]. Then, the 2,448,716 corrected sequences (N50: 16,449 bp; 29.2 Gb) were assembled with CANU 1.3[44], PBcR wgs8.3rc1[41,42] and FALCON 0.7.3[45] (overlaps dataset filtered with til-r[46], http://lipm-bioinfo.toulouse.inra.fr/download/til-r/) using different sets of parameters generating 12 primary assemblies with different metrics (Supplementary Table 1). By mapping the contigs of these assemblies on the reference *Bam*HI optical map[3] (Supplementary Note I.2), we observed that, for certain regions, assembly problems with one assembler could be solved using another assembler or an alternative set of parameters (Supplementary Fig. 1). As, to date, any assembly process based either on sequence or restriction data can generate chimeric contigs, the strategy we used aimed at building an assembly giving a high priority to solutions where sequence and optical data (Supplementary Note I.2) were consistent at least in one sequence assembly.

Fourteen sequence-based assemblies (Supplementary Note I.4.1 and Supplementary Table 1) were mapped on the 25 anchored contigs of the *Bam*HI maps[3] and filtered for reciprocal best hits (Supplementary Note I.2.3). The 14 mapping results were merged. Then, a second reciprocal best hit filter was applied on the merged file to build a minimal tiling path of contigs (originating from different assemblies). The overlaps between adjacent contigs were detected and a new consensus sequence was built by using the downstream sequence as consensus sequence of the overlapping region. Dangling parts of the selected contigs at the boundaries of the optical map (telomeres, centromeres, gaps) were used to extend the anchored contigs. This protocol enabled the identification of the 16 telomere sequences.

Next, to fill the gaps, we collected PacBio subreads with a minimum of 20 kb in length that could not be fully mapped on the anchored contigs. Then, the reads were assembled by CANU 1.3 with stringent parameters (ovlErrorRate = 0.015, utgErrorRate = 0.015, utgGraphErrorRate = 0.015, utgMergeErrorRate = 0.025, ovlMinLen = 5000). A sequence assembly spanning 17.5 Mb was obtained (214 contigs, N50 = 204 kb). The gap contigs were mapped on anchored contigs and on mitochondrial and chloroplast genomes. Gap contigs overlapping with the ends of anchored contigs were used to extend the anchored contigs. Gap contigs fully included (80% of their length) in anchored contigs, in mitochondrial or chloroplast genomes or in longer gap contigs, were removed.

The scaffolding of the remaining contigs was performed by using the Sanger BAC/Fosmids-ends generated by the *Medicago* genome consortium[3] (five libraries: *Hin*dIII x 2, *Eco*RI, Random sharing, Fosmid ends). The coverage of the EcoR1 library was increased via Illumina sequencing (Supplementary Note I.3). A first round of scaffolding was performed by taking into account inserts for which one end was uniquely mapped (a unique best scoring hit). Then, a second step of scaffolding was performed without this filter. A region close to the centromere of chr6 was significantly extended and improved. The sequence is fully collinear with the genetic map developed in this region[47]. In addition, the centromeric repeated markers MtR1, MtR2 and MtR3[48] were found in the centromeric regions of the pseudo-chromosomes (Fig. 1) as well as in the remaining unanchored contigs, as expected.

The genome was first polished by Quiver[49]. A second round of polishing was performed with one Illumina paired-ends library and two mate-pair libraries (3 and 5 kb insert sizes). The reads were mapped with glint software. The bam files were analysed by Pilon[50] (version 1.20) (see Supplementary Note I.4.5 for more details).

Similarity searches of long sequences (Pacbio subreads and sequence contigs) were performed with blastn. Short reads mapping used glint (BES mapping, polishing), http://lipm-bioinfo.toulouse.inra.fr/download/glint/. The management of the tiling path was performed by dedicated in-house Perl scripts. The mapping of optical maps on sequence contigs, filtering of links, contig scaffolding and HSPs chaining were performed by homemade Perl software (http://lipm-bioinfo.toulouse.inra.fr/download/lynx). Program parameters used for the different contig assemblies as well as assembly statistics at the various stages of the process are reported in Supplementary Table 1. The optical data are described in Supplementary Note I.2. More details of the different steps of the genome assembly are reported in Supplementary Note I.4.

**Genome annotation.** Gene models were predicted by the plant genome annotation pipeline egn-ep (http://eugene.toulouse.inra.fr/Downloads/egnep-Linux-x86_64.1.4.tar.gz release 1.4). The pipeline automatically manages probabilistic sequence model training, genome masking, transcript and protein alignment computation, alternative splice site detection and integrative gene modelling by the EuGene[51] software release 4.2a (http://eugene.toulouse.inra.fr/Downloads/eugene-4.2a.tar.gz).

Four protein databases were aligned with blastx to contribute to the detection of translated regions: (1) TAIR10; (2) Swiss-Prot – October 2016; (3) a plant subset of Uniprot proteins – October 2016 and (4) the proteome of *Brachypodium distachyon* release 192. Proteins similar to REPBASE[52] were removed from the four datasets. Chained alignments spanning less than 50% of the length of the database protein were removed.

Illumina-based RNAseq data were collected to cover a large spectrum of organs. The RNAseq datasets were assembled, library per library, with an iterative *k*-mer strategy based on the Velvet assembler[53]. A set of 3,284,874 transcript fragments was used in the annotation process (downloads section of https://medicago.toulouse.inra.fr/MtrunA17r5.0-ANR/). The integration of RNAseq data via contig assemblies allowed a homogenous integration with the sequences collected in Genbank on the 27 November 2016. Transcript fragments were aligned on the genome using GMAP[54] and only the best scoring hit was kept.

The standard EuGene 4.2a configuration file was modified to: (1) define a minimum CDS length of 120 nt (40 aa); (2) accept non-canonical GC/donor sites; (3) permit transcribed regions longer than 200 nt without any predicted CDS to be reported as ncRNAs; (4) permit independent forward and reverse annotations, thereby enabling the prediction of overlapping gene models on opposite strands; (5) increase the probability that regions identified by the peak detections of H3K9ac marks in roots and nodules were not intergenic regions and (6) increase the probability that the regions mapped by the peptides identified in the proteomic atlas encode for a protein.

In addition to the detection of ncRNAs based on transcriptomic data performed by EuGene, ncRNA genes were also predicted by tRNAScan-SE, RNAMMER and infernal 1.1.2.

After removal of redundant ncRNA predictions, 44,623 protein-coding genes, 974 tRNAs, 62 rRNAs and 3,762 non-protein-coding genes were retained. The set of predicted peptides was evaluated with BUSCO (release 3, embryophyta_odb9 dataset). 1,373 complete plus 19 fragmented gene models out of a total of 1,440 (95.4% and 1.3%, respectively) were detected.

An analysis on the genome browser of the mapped oriented paired-end reads used for RNAseq analysis of roots and nodules showed that several expressed regions were not annotated. The mapping results were analysed strand by strand and 39,568 regions with a read coverage greater than 20 and spanning at least 200 nt were identified. 1,895 of these regions (mean length = 1,369 nt) did not overlap with repeats and were well expressed (>0.6 normalized counts per million read pairs [cpm]) in at least one root or nodule condition and not overlapping by more than 100 nt with a mRNA on the same strand. These 1,895 regions were annotated as ncRNA genes and tagged as ope_rescue in the gff3 annotation file.

Finally, 44,623 protein-coding genes, 974 tRNAs, 62 rRNAs and 5,657 lncRNAs were annotated in the annotation release 1.6 (available at https://medicago.toulouse.inra.fr/MtrunA17r5.0-ANR/ and used in this study). The Supplementary Note II section contains more details on the protocol.

Protein-coding genes were functionally annotated by integrating six sources of information. Results were successively integrated depending on the expected accuracy of the source of information. Priority was successively given to: (1) 930 gene names manually annotated; (2) a blastp search of reciprocal best hits with the 1,938 Fabaceae proteins tagged as 'reviewed' in the Uniprot database (90% span, 80% identity) as of June 2017; (3) the Enzyme Commission (EC) number of 11,638 enzymes annotated using the described protocol with blastp e-value cutoff lowered to $1 \times 10^{-5}$ and pathway-prediction-score set to 0.3 in pathway-tools to increase stringency; (4) the transcription factors and kinases identified (2,716 and 1,615, respectively) by ITAK release 1.7; (5) the 4,174 transcription factors identified by PlantTFCat and (6) the Interpro (release 61.0) search matching 32,831 proteins. The EC numbers were tested against the ENZYME database, downloaded on 7 June 2017, updated when deprecated and then used to get the description of the enzymes. At each step, the description provided by the source of information was edited (when necessary) to ensure compliance with a submission to Genbank. Finally, the protein annotations were validated by the tbl2asn software (https://www.ncbi.nlm.nih.gov/genbank/tbl2asn2, 30 June 2017, r25.3). A putative function was assigned to 31,362 proteins and 13,261 proteins were tagged as 'hypothetical proteins'.

**Transposable element annotation.** Full-length transposable elements were first identified based on structural criteria, using Tephra[55] (version 0.09.3) as described[43] with the following modifications. First we omitted LTR retrotransposons (LTR-RTs) lacking coding domains from our classification procedures (Tephra configuration option 'domains_required' set to 'YES'). Second, the family-level classification method for LTR-RTs involved first combining the filtered set of LTR-RTs with all terminal-repeat retrotransposons in miniature, then running the 'tephra classifyltrs' command as described[43]; a family-level classification method was added for *Helitrons*, non-LTR retrotransposons, and terminal inverted repeat transposons based on global pairwise similarity. This method used blastn with an e-value threshold of $1 \times 10^{-10}$, and a modification of the 80–80–80 rule where pairwise matches must be over 80 bp, must cover at least 50% of the shorter element and must be over 80% identity to be considered members of the same family. Fourth, the final FASTA file of classified transposons used the standard header format: '>code_familyN_element_chromosome_start_end' where 'code' is a three-letter code designation for each transposon superfamily, 'familyN' is the family name/number determined by Tephra, 'element' is the individual transposon identifier, 'chromosome' is the source of the element, and 'start' and 'end' indicated the physical location of 'element' on 'chromosome.' Lastly, the GFF3 file of classified transposons used the following conventions in the attributes field: 'ID = element;family = familyN' where 'element' is the element identifier as in the FASTA header and 'familyN' is the family name/number. All other aspects of the GFF3 follow the Sequence Ontology specification[56]. Lastly, we refined the method for identifying fragmented transposable elements produced by the 'findfragments' subcommand of Tephra by changing the length threshold to 100 bp and by implementing an efficient range-based method for reporting only non-overlapping fragments.

**Repeat composition estimation.** Using the above methods, we obtained a set of 24,645 full-length transposable elements called MtTEdb hereafter (Supplementary Table 2). We used RepeatMasker (version open-4.0.7) with the command "RepeatMasker -qq -no_is -nolow -norna -species 'medicago truncatula'" and the Tephra (v0.09.3[55]) command 'tephra maskref' with default settings to mask the genome with MtTEdb for obtaining an estimate of the total repeat abundance. Based on the RepeatMasker and Tephra analyses, the *M. truncatula* genome was 24.02% and 24.11% transposons, respectively. To obtain an unbiased estimate of genome composition, we used Transposome[57] (v0.11.3) with a set of paired-end whole-genome sequencing (WGS) reads randomly sampled at varying levels of genome coverage (based on a genome size of 465 Mb for *M. trunculata;* Plant C-values database (http://data.kew.org/cvalues/). For this analysis, we took three random WGS samples at 0.01×, 0.03×, 0.05×, 0.07×, 0.09× and 0.10× genome coverages and performed an analysis with Transposome on each sequence set. We estimated the transposon content in *M. truncatula* to be 42.16 ± 4.69%, which was taken as the mean estimate of the annotated repeat content from all of the simulated reads sets (Supplementary Table 2). Notably, Young et al.[3] report the transposon content for the *M. truncatula* genome to be ~30%, which is in the range we determined for the A17 reference used in this study and an unbiased estimate from an analysis of WGS reads. To obtain a mathematical estimate of repeat abundance, we constructed an index of 20 million *k*-mers that were 20 bp in length with the Tallymer program 'mkindex' and searched these against the reference genome with the Tallymer 'search' command[58]. Our final estimates were filtered to remove simple repeats (di- and tri-nucleotide repeats) that were above 80% of the *k*-mer to reduce the number of spurious matches.

**Genome scale visualizations.** Fig. 1 and Supplementary Figs. 1 and 3 were generated with the Circos software[59] (http://circos.ca). Figure 2c was generated with the DensityMap software[60].

**Small RNA analyses.** Samples were collected from plants aeroponically grown for one week in the presence of 5 mM $NH_4NO_3$ and then nitrogen-starved for three days. Root samples (without root tips) were collected eight plants per replicate. Isolated nodules were harvested at 4, 6 and 10 days post inoculation with *Sinorhizobium meliloti* 2011. Small RNAs (<200 nt) were extracted in triplicate using the miRVana miRNA isolation kit (Thermo Fisher Scientific). Multiplexed libraries were then constructed using the Ion Total RNA-Seq Kit v2 for Small RNA Libraries (Thermo Fisher Scientific) and sequenced using the Ion PI Sequencing 200 Kit v3 on a Ion Proton Sequencer (Thermo Fisher Scientific). Between 20.3 and 42.2 million reads were obtained for each nodulation time.

Small RNA clusters in the genome were predicted using the ShortStack (v3.8.2) pipeline. For each replicate, reads were separated in different fastq files according to their size (from 20 nt to 25 nt). They were mapped through ShortStack on the genome without tolerating any mismatch (–mismatches 0), reporting all the possible locations returned by bowtie (–bowtie_m all) and placing the multi-mapping reads a unique position guided by uniquely mapping reads (–mmap u). The analysis part of ShortStack was then used to predict the clusters and describe their characteristics (Supplementary Table 5) using default parameters. The neighbourhood of the clusters was extracted with bedtools (v2.26.0). The clusters were considered phased if their ShortStack phasing score was > 10.

Additional information can be found in Supplementary Note II.3.

**Transcriptome analyses.** Previously generated[9] RNAseq data were mapped on the new genome sequence. The EdgeR Bioconductor package version 3.16.5 for R was used to detect differentially expressed genes. Genes with no counts across all libraries were not retained for further analysis. Normalization was performed using a trimmed mean of *M*-values method. Quality control plots of normalized datasets were generated by principal component analysis using Ade4 version 1.7-5 package and heatmaps were obtained on sample-to-sample Euclidean distances with the package pheatmap version 1.0.8.

Multiple factor (biological repetition and factor of interest) analyses were carried out using fitting generalized linear models (GLMs) with a design matrix. Dispersion was estimated by the Cox-Reid profile-adjusted likelihood method.

Differentially expressed genes (DEGs) or siRNA clusters were called using the GLM likelihood ratio test, with a FDR adjusted $q$-value $< 0.01$.

Hierarchical clustering on filtered DEGs (normalized cpm $> 0.6$ in at least one biological condition and $q$-value $< 0.01$) was generated with the heatmap.2 function as available in the gplots Bioconductor package version 3.0.1, using Ward's minimum variance clustering method on Euclidean distances (setting $k = 16$).

The coexpressed gene network was constructed from pairwise calculations of Pearson correlations with a threshold set at 0.95 and built using Cytoscape.

For analyses of lncRNAs and mRNAs correlations, Spearman's rank correlations and associated FDR were calculated using R with 'cor' and 'cortest' functions, respectively. The relative position between lncRNAs and mRNA were obtained using the bedtools toolkit. The overlapping genes were identified using the 'intersect' command, the closest genes were obtained with the 'closest' command and for both approaches the –s and –S arguments were used to take the DNA strandedness into account.

**Detection and analysis of SRIs.** Symbiotic islands were detected based on normalized RNAseq expression data (previous section) using the following workflow embedded in an in-house R script (see a graphical workflow in Supplementary Note V.1; script Pecrix_et_al-Suppl-Notes-V.Symbiotic-island-analysis.R available in the download section of https://medicago.toulouse.inra.fr/MtrunA17r5.0-ANR/). Windows of 50 kb centred on each expressed gene were defined. Filtering was applied so as to only consider the windows containing at least three coregulated genes representing at least 60% of the expressed gene content (bedtools command 'coverage'). Non-expressed genes (potentially including pseudogenes) were not taken into consideration (<0.6 normalized cpm (sum of three biological replicates) for the root and/or nodule polyA libraries (NRU, NRD and NRN islands) or for at least one LCM zone (NDD, NDA and NDN islands)). Overlapping filtered windows were then stacked to generate islands (bedtools command 'merge'). The borders of each island were trimmed to the limits of the first and last coregulated island gene (bedtools command 'intersect'). Finally, when the closest gene positioned up to 25 kb upstream or downstream from an island was also coregulated, it was incorporated into the island (bedtools command 'closest' with the '-iu' or '-id' arguments to search respectively upstream or downstream genes). Following these steps, only islands with at least four coregulated genes were retained.

To estimate the number of islands expected by random sampling, this workflow was iterated 1,000 times using $n$ randomly chosen genes with the R function 'sample', where $n$ is the total number of genes considered for each island type (see Supplementary Note V.1).

The DMR, siRNA cluster and histone mark composition of symbiotic islands were assessed using the bedtools toolkit, while metaplots were calculated using the 'profile' function of DANPOS2 software with a bin size set to 200 bp.

**Histone mark analyses.** ChIP assays were performed using 1 µg of anti-H3K9ac (Millipore, ref. 07–352), anti-H3K27me3 (Millipore, ref. 07–449), anti-H3K27me1 (Millipore, ref. 07–448) and anti-H3K9me2 (Abcam, ref. ab1220) antibodies, using a procedure adapted from Veluchamy et al.[61]. Briefly, after plant material fixation in 1% (v/v) formaldehyde, tissues were homogenized, nuclei isolated and lysed. Cross-linked chromatin was sonicated using a Covaris S220 (peak incident power: 175 W; duty factor: 20%; cycles per burst: 200; time: 10 min). Protein/DNA complexes were immunoprecipitated with antibodies, overnight at 4 °C with gentle shaking, and incubated for 1 h at 4 °C with 50 µl of Dynabeads Protein A (Invitrogen, Ref. 100-02D). The beads were washed for 2 × 5 min in ChIP Wash Buffer 1 (0.1% SDS, 1% Triton X-100, 20 mM Tris-HCl pH 8, 2 mM EDTA pH 8, 150 mM NaCl), for 2 × 5 min in ChIP Wash Buffer 2 (0.1% SDS, 1% Triton X-100, 20 mM Tris-HCl pH 8, 2 mM EDTA pH 8, 500 mM NaCl), for 2 × 5 min in ChIP Wash Buffer 3 (0.25 M LiCl, 1% NP-40, 1% sodium deoxycholate, 10 mM Tris-HCl pH 8,1 mM EDTA pH 8) and twice in transposable element (10 mM Tris-HCl pH 8, 1 mM EDTA pH 8). ChIPed DNA was eluted by two 15-min incubations at 65 °C with 250 µl of Elution Buffer (1% SDS, 0.1 M $NaHCO_3$). Chromatin was reverse-cross-linked by adding 20 µl of NaCl 5 M overnight at 65 °C. Reverse-cross-linked DNA was submitted to RNase and proteinase K digestion, and extracted with phenol-chloroform. DNA was ethanol precipitated in the presence of 20 µg of glycogen and resuspended in 20 µl of nuclease-free water (Ambion) in a DNA low-bind tube. 10 ng of immunoprecipitation or input DNA was used for ChIP-Seq library construction using the NEB-Next Ultra II DNA Library Prep Kit for Illumina (New England Biolabs) according to the manufacturer's recommendations. For all libraries, 11 cycles of PCR were used. The quality of the libraries was assessed with the Agilent 2100 Bioanalyzer (Agilent), and the libraries were subjected to high-throughput sequencing by NextSeq 500 (Illumina).

Single-end sequencing reads of 76 nt were quality controlled using FASTQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc). Raw data from histone marks and input libraries were mapped on the new *M. truncatula* genome release using glint software (Faraut & Courcelle; http://lipm-bioinfo.toulouse.inra.fr/download/glint/, unpublished) with the following parameters: map –best-score –mmis 3 –lrmin 80. Each biological replicate was analysed in parallel, with histone marks and input data from each replicate analysed together.

Detection of H3K9Ac, H3K27me1 and H3K9me2 narrow peaks was performed using MACS2 software (version: 2.1.1.20160309, method: callpeak, custom parameters: –shift 100 –extsize 200). Identification of H3K27me3 broad domains was done with SICER software (version: 1.1, parameters: redundancy threshold = 1; window size = 200; fragment size = 150; effective genome fraction = 0.860794380127; gap size = 600; FDR = 0.01). Two biological replicates were obtained for all samples, except for H3K9me2 where one replicate was retained.

**URLs.** Integrative web portal, including a *M. truncatula* genome browser: https://medicago.toulouse.inra.fr/MtrunA17r5.0-ANR/

**Reporting Summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

**Code availability.** The custom Perl scripts developed to manage the genome assembly process are available at https://medicago.toulouse.inra.fr/MtrunA17r5.0-ANR/downloads/1.6/Pecrix-et-al.Suppl-Notes-I.lynx-toolkit-20180223.tar.gz and http://lipm-bioinfo.toulouse.inra.fr/download/lynx. The R script developed for the definition of symbiosis-related islands is available at https://medicago.toulouse.inra.fr/MtrunA17r5.0-ANR/downloads/1.6/Pecrix_et_al-Suppl-Notes-V.Symbiotic-island-analysis.R. Others custom scripts mentioned in the manuscripts are available at https://medicago.toulouse.inra.fr/MtrunA17r5.0-ANR/downloads/1.6/Pecrix_et_al-misc_custom_scripts.zip.

## Data availability

This Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the accession PSQE00000000. The version described in this paper is version PSQE01000000. Raw reads from PacBio, ChIP-seq and small RNAseq experiments have been deposited at the Sequence Read Archive (SRA) (project accession number: SRP131849). Data related to gene annotation, transposable element annotation and ChIP-seq analyses, as well as Supplementary Table 6, are available at the web portal: https://medicago.toulouse.inra.fr/MtrunA17r5.0-ANR/; downloads section.

Received: 7 March 2018; Accepted: 21 September 2018;
Published online: 05 November 2018

## References

1. Martin, F.M., Uroz, S. & Barker, D.G. Ancestral alliances: plant mutualistic symbioses with fungi and bacteria. *Science* **356**, eaad4501 (2017).
2. Young, N. D. & Udvardi, M. Translating *Medicago truncatula* genomics to crop legumes. *Curr. Opin. Plant Biol.* **12**, 193–201 (2009).
3. Young, N. D. et al. The *Medicago* genome provides insight into the evolution of rhizobial symbioses. *Nature* **480**, 520–524 (2011).
4. Tang, H. et al. An improved genome release (version Mt4.0) for the model legume *Medicago truncatula*. *BMC Genomics* **15**, 312 (2014).
5. Moll, K. M. et al. Strategies for optimizing BioNano and Dovetail explored through a second reference quality assembly for the legume model, Medicago truncatula. *BMC Genomics* **18**, 578 (2017).
6. Kamphuis, L. G. et al. The *Medicago truncatula* reference accession A17 has an aberrant chromosomal configuration. *New Phytol.* **174**, 299–303 (2007).
7. de Bang, T. et al. Genome-wide identification of *Medicago* peptides involved in macronutrient responses and nodulation. *Plant Physiol.* **175**, 1669–1689 (2017).
8. Miller, J. R. et al. Hybrid assembly with long and short reads improves discovery of gene family expansions. *BMC Genomics* **18**, 541 (2017).
9. Roux, B. et al. An integrated analysis of plant and bacterial gene expression in symbiotic root nodules using laser-capture microdissection coupled to RNA sequencing. *Plant J.* **77**, 817–837 (2014).
10. Jardinaud, M. F. et al. A laser dissection-RNAseq analysis highlights the activation of cytokinin pathways by nod factors in the *Medicago truncatula* root epidermis. *Plant Physiol.* **171**, 2256–2276 (2016).
11. Stanton-Geddes, J. et al. Candidate genes and genetic architecture of symbiotic and agronomic traits revealed by whole-genome, sequence-based association genetics in *Medicago truncatula*. *PLoS ONE* **8**, e65688 (2013).
12. Ariel, F. et al. Noncoding transcription by alternative RNA polymerases dynamically regulates an auxin-driven chromatin loop. *Mol. Cell* **55**, 383–396 (2014).
13. Krzyczmonik, K., Wroblewska-Swiniarska, A. & Swiezewski, S. Developmental transitions in *Arabidopsis* are regulated by antisense RNAs resulting from bidirectionally transcribed genes. *RNA Biol.* **14**, 838–842 (2017).
14. Swiezewski, S., Liu, F., Magusin, A. & Dean, C. Cold-induced silencing by long antisense transcripts of an *Arabidopsis* polycomb target. *Nature* **462**, 799–802 (2009).
15. Fedak, H. et al. Control of seed dormancy in *Arabidopsis* by a cis-acting noncoding antisense transcript. *Proc. Natl Acad. Sci. USA* **113**, E7846–E7855 (2016).
16. Henriques, R. et al. The antiphasic regulatory module comprising CDF5 and its antisense RNA FLORE links the circadian clock to photoperiodic flowering. *New Phytol.* **216**, 854–867 (2017).
17. Vernié, T. et al. EFD is an ERF transcription factor involved in the control of nodule number and differentiation in *Medicago truncatula*. *Plant Cell* **20**, 2696–2713 (2008).

18. Satgé, C. et al. Reprogramming of DNA methylation is critical for nodule development in *Medicago truncatula*. *Nat. Plants* **2**, 16166 (2016).
19. Kalo, P. et al. Nodulation signaling in legumes requires NSP2, a member of the GRAS family of transcriptional regulators. *Science* **308**, 1786–1789 (2005).
20. Sinharoy, S. et al. The C2H2 transcription factor regulator of symbiosome differentiation represses transcription of the secretory pathway gene VAMP721a and promotes symbiosome development in *Medicago truncatula*. *Plant Cell* **25**, 3584–3601 (2013).
21. Marsh, J. F. et al. *Medicago truncatula* NIN is essential for rhizobial-independent nodule organogenesis induced by autoactive calcium/calmodulin-dependent protein kinase. *Plant Physiol.* **144**, 324–335 (2007).
22. Ovchinnikova, E. et al. IPD3 controls the formation of nitrogen-fixing symbiosomes in pea and *Medicago* Spp. *Mol. Plant Microbe Interact.* **24**, 1333–1344 (2011).
23. Lefebvre, B. et al. A remorin protein interacts with symbiotic receptors and regulates bacterial infection. *Proc. Natl Acad. Sci. USA* **107**, 2343–2348 (2010).
24. Berrabah, F. et al. A nonRD receptor-like kinase prevents nodule early senescence and defense-like reactions during symbiosis. *New Phytol.* **203**, 1305–1314 (2014).
25. Alunni, B. et al. Genomic organization and evolutionary insights on GRP and NCR genes, two large nodule-specific gene families in Medicago truncatula. *Mol. Plant Microbe Interact.* **20**, 1138–1148 (2007).
26. Graham, M. A., Silverstein, K. A., Cannon, S. B. & VandenBosch, K. A. Computational identification and characterization of novel genes from legumes. *Plant Physiol.* **135**, 1179–1197 (2004).
27. Pan, H. & Wang, D. Nodule cysteine-rich peptides maintain a working balance during nitrogen-fixing symbiosis. *Nat. Plants* **3**, 17048 (2017).
28. Liu, J. et al. Recruitment of novel calcium-binding proteins for root nodule symbiosis in *Medicago truncatula*. *Plant Physiol.* **141**, 167–177 (2006).
29. Alunni, B. & Gourion, B. Terminal bacteroid differentiation in the legume-rhizobium symbiosis: nodule-specific cysteine-rich peptides and beyond. *New Phytol.* **211**, 411–417 (2016).
30. Matzke, M. A. & Mosher, R. A. RNA-directed DNA methylation: an epigenetic pathway of increasing complexity. *Nat. Rev. Genet.* **15**, 394–408 (2014).
31. Hurst, L. D., Pal, C. & Lercher, M. J. The evolutionary dynamics of eukaryotic gene order. *Nat. Rev. Genet.* **5**, 299–310 (2004).
32. Nutzmann, H. W., Huang, A. & Osbourn, A. Plant metabolic clusters – from genetics to genomics. *New Phytol.* **211**, 771–789 (2016).
33. Reimegard, J. et al. Genome-wide identification of physically clustered genes suggests chromatin-level co-regulation in male reproductive development in *Arabidopsis thaliana*. *Nucleic Acids Res.* **45**, 3253–3265 (2017).
34. Plaza, S., Menschaert, G. & Payre, F. In search of lost small peptides. *Annu. Rev. Cell Dev. Biol.* **33**, 391–416 (2017).
35. Hnisz, D. & Young, R. A. New insights into genome structure: genes of a feather stick together. *Mol. Cell* **67**, 730–731 (2017).
36. Rowley, M. J. et al. Evolutionarily conserved principles predict 3D chromatin organization. *Mol. Cell* **67**, 837–852 e7 (2017).
37. Mele, M. & Rinn, J. L. 'Cat's cradling' the 3D genome by the act of LncRNA transcription. *Mol. Cell* **62**, 657–664 (2016).
38. Mercer, T. R. & Mattick, J. S. Structure and function of long noncoding RNAs in epigenetic regulation. *Nat. Struct. Mol. Biol.* **20**, 300–307 (2013).
39. Heo, J. B. & Sung, S. Vernalization-mediated epigenetic silencing by a long intronic noncoding RNA. *Science* **331**, 76–79 (2011).
40. Mayjonade, B. et al. Extraction of high-molecular-weight genomic DNA for long-read sequencing of single molecules. *Biotechniques* **61**, 203–205 (2016).
41. Berlin, K. et al. Corrigendum: assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nat. Biotechnol.* **33**, 1109 (2015).
42. Berlin, K. et al. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nat. Biotechnol.* **33**, 623–630 (2015).
43. Badouin, H. et al. The sunflower genome provides insights into oil metabolism, flowering and Asterid evolution. *Nature* **546**, 148–152 (2017).
44. Koren, S. et al. Canu: scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017).
45. Chin, C. S. et al. Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods* **13**, 1050–1054 (2016).
46. Raymond, O. et al. The Rosa genome provides new insights into the domestication of modern roses. *Nat. Genet.* **50**, 772–777 (2018).
47. Tayeh, N. et al. A tandem array of CBF/DREB1 genes is located in a major freezing tolerance QTL region on *Medicago truncatula* chromosome 6. *BMC Genomics* **14**, 814 (2013).
48. Kulikova, O. et al. Satellite repeats in the functional centromere and pericentromeric heterochromatin of *Medicago truncatula*. *Chromosoma* **113**, 276–283 (2004).
49. Chin, C. S. et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* **10**, 563–569 (2013).
50. Walker, B. J. et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE* **9**, e112963 (2014).
51. Foissac, S. et al. Genome annotation in plants and fungi: EuGene as a model platform. *Current Bioinformatics* **3**, 87–97 (2008).
52. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA* **6**, 11 (2015).
53. Zerbino, D. R. Using the Velvet de novo assembler for short-read sequencing technologies. *Curr. Protoc. Bioinformatics* **11**, Unit11 5 (2010).
54. Wu, T. D. & Watanabe, C. K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859–1875 (2005).
55. Tephra: A Tool for Discovering Transposable Elements and Describing Patterns of Genome Evolution v.0.12.2 (Staton, S., 2017); https://github.com/sestaton/tephra
56. Generic Feature Format Version 3 (GFF3) v.1.23 (Stein, L., 2013); https://github.com/The-Sequence-Ontology/Specifications/blob/master/gff3.md
57. Staton, S. E. & Burke, J. M. Transposome: a toolkit for annotation of transposable element families from unassembled sequence reads. *Bioinformatics* **31**, 1827–1829 (2015).
58. Kurtz, S., Narechania, A., Stein, J. C. & Ware, D. A new method to compute *K*-mer frequencies and its application to annotate large repetitive plant genomes. *BMC Genomics* **9**, 517 (2008).
59. Krzywinski, M. et al. Circos: an information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645 (2009).
60. Guizard, S., Piegu, B. & Bigot, Y. DensityMap: a genome viewer for illustrating the densities of features. *BMC Bioinformatics* **17**, 204 (2016).
61. Veluchamy, A. et al. LHP1 regulates H3K27me3 spreading and shapes the three-dimensional conformation of the *Arabidopsis* genome. *PLoS ONE* **11**, e0158936 (2016).

## Acknowledgements

We thank C. Ben and L. Gentzbittel (EcoLab, Université de Toulouse, CNRS, Toulouse INP, UPS, France), G. Aubert, R. Thompson and K. Gallardo (INRA, UMR 1347, Agroécologie, Dijon, France) and B. Gronenborn (I2BC, CNRS, Paris Sud, CEA, University of Paris Saclay, Gif sur Yvette, France) for providing small RNA data on disease responses, seeds and viroid-infected plants, respectively, as well as N. Peeters (LIPM, Toulouse) for mRNA data used for genome annotation. We thank M.C. Le Paslier for her help in Illumina sequencing. This work was supported by the ANR grants EPISYM (grant no. ANR-15-CE20-0002), NODCCAAT (no. ANR-15-CE20-0012), REGULEG (no. ANR-15-CE20-0001), the 'Laboratoire d'Excellence (LABEX)' TULIP (no. ANR-10-LABX-41), the LABEX Saclay Plant Sciences (SPS; no. ANR-10-LABX-40) and the European Research Council (no. ERC-SEXYPARTH), and we made use of data previously generated in the ANR SYMbiMICS (ANR-08-GENO-106) and the INRA SPE EPINOD projects. The sequencing platform was supported by France Génomique National infrastructure (grant no. ANR-10-INBS-09) and by the GET-PACBIO programme (Programme opérationnel FEDER-FSE MIDI-PYRENEES ET GARONNE 2014-2020). We are grateful to the Genotoul bioinformatics platform Toulouse Midi-Pyrenees (Bioinfo Genotoul) for providing computing and storage resources. C. Satgé was supported by a doctoral grant from the French Ministry of Education and Research.

## Author contributions

S.Mo., B.M., C.L-R. and O.B. prepared DNA samples and performed PacBio sequencing. S.Cau., C.C-D., W.M. and H.B. built the Bionano optical maps. B.M., J.G., W.M., S.Mu. and A.Ber. designed and performed Illumina seq of BAC end sequencing (EcoR1 library). F.D. prepared DNA samples and managed Illumina sequencing. J.G. assembled the genome. E.S., S.Car. and J.G. annotated protein-coding genes and miRNAs. S.E.S. annotated and analysed repeats and transposable elements. S.Car. developed the Medicago bioinformatics portal. J.K. positioned HapMap data on the new reference genome. C.S. and C.L.-B. prepared samples for the sRNA analyses. C.L.-B. conducted the miRNA analyses. T.B., C.L.-B. and Y.P. conducted the siRNA analyses. S.Mo. and M.P. prepared the histone mark samples. D.L. and M.P. performed the ChIP experiments. M.B., D.L. and A.Ben. performed ChIP-seq. M.Za., M.Zo., M.B., S.Car., Y.P. and P.G. performed the analysis of the ChIP-seq data. Y.P. and P.G. conducted the lncRNA analyses. Y.P., S.Car. and P.G. performed the gene family analyses. J.G. and T.B. performed the sRNA and mRNA expression analyses. M.-F.J. performed the gene and siRNA differential expression analyses. Y.P. and P.G. performed the integrated analyses of the symbiotic islands. P.G., J.G., M.C., A.N. and J.B. contributed to the project set-up. P.G., J.G., S.E.S. and C.L.-B. wrote the manuscript, with contributions from M.C., F.F., J.B., B.M., Y.P., F.D., A.N., M.Zo., E.S. and S.Mu. P.G., J.G. and M.C. coordinated the project.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** is available for this paper at https://doi.org/10.1038/s41477-018-0286-7.

**Reprints and permissions information** is available at www.nature.com/reprints.

**Correspondence and requests for materials** should be addressed to J.G. or P.G.

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

nature research

Corresponding author(s): Pascal Gamas, Jérôme Gouzy

# Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

## Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

| n/a | Confirmed | |
|---|---|---|
| □ | ☒ | The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement |
| □ | ☒ | An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| □ | ☒ | The statistical test(s) used AND whether they are one- or two-sided<br>*Only common tests should be described solely by name; describe more complex techniques in the Methods section.* |
| □ | ☒ | A description of all covariates tested |
| □ | ☒ | A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| □ | ☒ | A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| □ | ☒ | For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted<br>*Give P values as exact values whenever suitable.* |
| ☒ | □ | For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| □ | ☒ | For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| ☒ | □ | Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated |
| □ | ☒ | Clearly defined error bars<br>*State explicitly what error bars represent (e.g. SD, SE, CI)* |

*Our web collection on statistics for biologists may be useful.*

## Software and code

Policy information about availability of computer code

| | |
|---|---|
| Data collection | Pacbio data were collected on a RS II system, the Instrument control software version (InstCtrlVer) was 2.3.0.3.154799, the signal processing software versions (SigProcVer) were SwVer=2303.154799 and HwVer=1.0. The basecaller was V1 with contiguration file 2-3-0_P6-C4.xml. Label Density CalculatorTM 1.3.0, AutoDetectTM 2.1.4.9159 and BioNano Solve (from BioNano Genomics) |
| Data analysis | CANU 1.3; PBcR wgs8.3rc1 and FALCON 0.7.3; til-r 20160717; Quiver; ncbi-blast-2.2.31+; ncbi-blast-2.2.31+ and 2.6; Pilon (version 1.20); glint 1.0.rc12 ; egnep 1.4; Eugene 4.2a; Velvet 1.2.10; GMAP-2017-02-15 ; tRNAScan-SE 1.3.1; RNAMMER 1.2; infernal 1.1.1; BUSCO 3; ITAK release 1.7; PlantTFCat; InterproScan/Interpro (release 61.0); tbl2asn, r25.3; Tephra55 (version 0.09.3); Transposome (v0.11.3); RepeatMasker (version open-4.0.7); Tallymer; Circos v0.69-5; DensityMap; ShortStack v3.8.2; bedtools v 2.24.0 and v2.26.0; EdgeR Bioconductor package version 3.16.5 for R; Ade4 version1.7-5; pheatmap version 1.0.8; gplots Bioconductor package version 3.0.1; Cytoscape; DANPOS2; R package (cor, cortest, sample...); FASTQC; MACS2 2.1.1.20160309; SICER 1.1; D-GENIES; megablast; ALLMAPS; Minimap2; e-PCR; EMBOSS release 6.6.0.0; Red 05/22/2015; genometools 1.5.6 (LTRHarvest); smallA (mirfold 0.2b); miRanda release 3.3; CleaveLand release 4.4; ViennaRNA; Blat; LiftOver; Picard Tools; topGO; MCscan; OrthoFinder-2.0.0 (with diamond v0.8.24.86); OrthoMCL 1.4 (with blastall 2.2.23); JBrowse 1.12.5; Blast2GO 5 (beta release); usearch v8.0.1623_i86linux64; GMAP version 2017-09-05; E2P2 version 3.1; Microsoft Excel 2010.<br>Custom codes specifically set-up for this study are described in the method and supplementary notes sections of the manuscript. Source codes are available in the download section of the project web https://medicago.toulouse.inra.fr/MtrunA17r5.0-ANR/, see Pecrix-et-al.Suppl-Notes-I.lynx-toolkit-20180223.tar.gz (assembly) and Pecrix_et_al-Suppl-Notes-V.Symbiotic-island-analysis.R (island definition) |

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

## Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Code availability
The custom Perl scripts developed to manage the genome assembly process are available at https://medicago.toulouse.inra.fr/MtrunA17r5.0-ANR/downloads/1.6/Pecrix-et-al.Suppl-Notes-I.lynx-toolkit-20180223.tar.gz and http://lipm-bioinfo.toulouse.inra.fr/download/lynx. The R script developed for the definition of Symbiosis-related islands is available at https://medicago.toulouse.inra.fr/MtrunA17r5.0-ANR/downloads/1.6/Pecrix_et_al-Suppl-Notes-V.Symbiotic-island-analysis.R. Others custom scripts mentioned in the manuscripts are available at https://medicago.toulouse.inra.fr/MtrunA17r5.0-ANR/downloads/1.6/Pecrix_et_al-misc_custom_scripts.zip.

Data availability:
This Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the accession PSQE00000000. The version described in this paper is version PSQE01000000. Raw reads from PacBio, ChIPseq and small RNAseq experiments have been deposited at the Sequence Read Archive (SRA) (project accession number: SRP131849). Data related to gene annotation, TE annotation, and ChIP seq analyses are available at the web portal: https://medicago.toulouse.inra.fr/MtrunA17r5.0-ANR/; downloads section.

Correspondence and requests for material should be addressed to Pascal Gamas and / or Jérôme Gouzy.

# Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

# Life sciences

## Study design

All studies must disclose on these points even when the disclosure is negative.

| | |
|---|---|
| Sample size | No statistical methods were used to predetermine sample sizes. The number of replicates was dictated by financial constraints (genome-wide analyses) and followed what is commonly done in the field (namely three replicates for small RNA transcriptomes and two replicates for genome-wide ChIP seq analyses). Each replicate consisted of a pool of at least 10 individual plant samples. For RNAseq analyses, an a posteriori evaluation of the detection power of differentially expressed genes in our conditions (54 000 genes, 8037 differentially expressed genes with a minimum fold change of 4) indicated a probability (power) of 0.82 with three replicates [R Package: RnaSeqSampleSize. version 1.12.0. (Zhao S, Li C, Guo Y, Sheng Q, Shyr Y, 2017); FDR associated with this test=0.01]. |
| Data exclusions | One set of small RNA samples was not retained because it strongly differed from the two other replicates based on PCA, correlation matrix, euclidian distance matrix and triplot analyses. One ChIPseq replicate (H3K9me2) was not retained because too few peaks with FDR<0.05% were obtained. |
| Replication | Experimental findings were reliably reproduced, except in one set of small RNA samples as indicated above. |
| Randomization | Randomization was used to evaluate the robustness of symbiosis-related islands, with 1000 iterations as described in supplementary notes V.1 (Supplementary Notes Fig.21). |
| Blinding | No phenotypic analyses, where blinding is essential for reliability of results, were carried out in this study. All genomic analyses, including transcriptome and ChiPseq data analyses as well as symbiosis-related island identification, were conducted using the same automatic pipelines, regardless of the samples considered. |

## Materials & experimental systems

Policy information about availability of materials

| n/a | Involved in the study |
|---|---|
| ☒ | □ Unique materials |
| □ | ☒ Antibodies |
| ☒ | □ Eukaryotic cell lines |
| ☒ | □ Research animals |
| ☒ | □ Human research participants |

### Antibodies

Antibodies used

anti-H3K9ac (Millipore, ref. 07-352), anti-H3K27me3 (Millipore, ref. 07-449), anti-H3K27me1 (Millipore, ref. 07-448), anti-H3K9me2 (Abcam, ref. ab1220), one microgram per ChipSeq experiment.

Validation

all antibodies used in this study were commercially available and validated by the supplier (see http://www.merckmillipore.com and http://www.abcam.com/). Their use was already published.

# Method-specific reporting

| n/a | Involved in the study |
|---|---|
| □ | ☒ ChIP-seq |
| ☒ | □ Flow cytometry |
| ☒ | □ Magnetic resonance imaging |

## ChIP-seq

### Data deposition

☒ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

☒ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links
*May remain private before publication.*

https://medicago.toulouse.inra.fr/MtrunA17r5.0-ANR/

Files in database submission

SRX3663671; SRX3663672; SRX3663673; SRX3663674; SRX3663675; SRX3663676; SRX3663677; SRX3663678; SRX3663679; SRX3663680; SRX3663685; SRX3663686; SRX3663687; SRX3663688; SRX3663689; SRX3663690; SRX3663691; SRX3663692; SRX3663694

Genome browser session
(e.g. UCSC)

https://medicago.toulouse.inra.fr/MtrunA17r5.0-ANR

### Methodology

Replicates

two biological replicates per sample (except for H3K9me2: one replicate retained), with consistent results (shown in Fig.4 and supplementary notes Fig.13).

Sequencing depth

Single-end sequencing reads of 76 nt. H3K27me1_Nodules-R1: 167 167 340 reads; H3K27me1_Nodules-R2: 60 743 865 reads; H3K27me1_Roots-R1: 154 291 923 reads; H3K27me1_Roots-R2: 56 295 732 reads; H3K27me3_Nodules-R1: 78 342 883 reads; H3K27me3_Nodules-R2: 70 564 518 reads; H3K27me3_Roots-R1: 83 045 890 reads; H3K27me3_Roots-R2: 62 520 192 reads; H3K9ac_Nodules-R1: 173 458 638 reads; H3K9ac_Nodules-R2: 106 667 995 reads; H3K9ac_Roots-R1: 151 141 193 reads; H3K9ac_Roots-R2: 90 895 343 reads; H3K9me2_Nodules-R1: 213 509 464 reads; H3K9me2_Roots-R1: 241 696 034 reads; Input_Nodules-R1: 139 243 932 reads; Input_Nodules-R2: 129 650 431 reads; Input_Roots-R1: 153 250 967 reads; Input_Roots-R2: 108 610 399 reads.

Antibodies

anti-H3K9ac (Millipore, ref. 07-352), anti-H3K27me3 (Millipore, ref. 07-449), anti-H3K27me1 (Millipore, ref. 07-448) and anti-H3K9me2 (Abcam, ref. ab1220), one microgram per ChipSeq experiment.

Peak calling parameters

Detection of H3K9Ac, H3K27me1 and H3K9me2 narrow peaks was performed using MACS2 software (version: 2.1.1.20160309, method: callpeak, custom parameters : --shift 100 --extsize 200). Identification of H3K27me3 broad domains was done with SICER software (version: 1.1, parameters: redundancy threshold=1; window size=200; fragment size=150; effective genome fraction=0.860794380127; gap size=600; FDR=0.01).

Data quality

FDR <= 0.05, FOLD ENR >=2; H3K27me1_Nodules-R1: 317 peaks; H3K27me1_Nodules-R2: 243 peaks; H3K27me1_Roots-R1: 1877 peaks; H3K27me1_Roots-R2: 28708 peaks; H3K9ac_Nodules-R1: 19221 peaks; H3K9ac_Nodules-R2: 20240 peaks; H3K9ac_Roots-R1: 18224 peaks; H3K9ac_Roots-R2: 18718 peaks; H3K9me2_Nodules-R1: 373 peaks; H3K9me2_Roots-R1: 2554 peaks; H3K27me3_Nodules-R1: 4890 peaks; H3K27me3_Nodules-R2: 3414 peaks; H3K27me3_Roots-R1: 5216 peaks; H3K27me3_Roots-R2: 7433 peaks;

Software

MACS2 (version: 2.1.1.20160309); SICER  (version: 1.1).