Comparative genomics of the non-legume *Parasponia* reveals insights into evolution of nitrogen-fixing rhizobium symbioses

Robin van Velzen[†], Rens Holmer[†], Fengjiao Bu[‡], Luuk Rutten[‡], Arjan van Zeijl, Wei Liu, Luca Santuari, Qingqin Cao, Trupti Sharma, Defeng Shen, Yuda Roswanjaya, Titis A.K. Wardhani, Maryam Seifi Kalhor, Joelle Jansen, Johan van den Hoogen, Berivan Güngör, Marijke Hartog, Jan Hontelez, Jan Verver, Wei-Cai Yang, Elio Schijlen, Rimi Repin, Menno Schilthuizen, M. Eric Schranz, Renze Heidstra, Kana Miyata, Elena Fedorova, Wouter Kohlen, Ton Bisseling, Sandra Smit, & Rene Geurts

Supplementary figures, tables, and methods.

†,‡: These authors contributed equally

www.pnas.org/cgi/doi/10.1073/pnas.1721395115

Supplementary figures

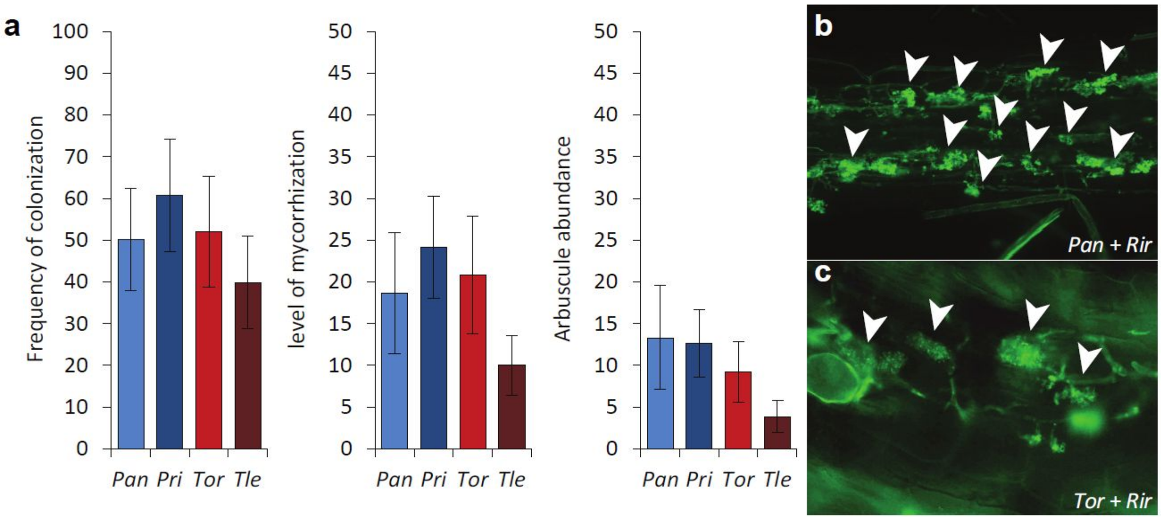


Fig. S1

Arbuscular mycorrhization of *Parasponia* and *Trema* species. **a** Mycorrhization efficiency of *Parasponia andersonii* WU01.14 (Pan), *P. rigida* WU20 (Pri), *Trema orientalis* RG33 (Tor) and *T. levigata* WU50 (Tle), 6 weeks post inoculation with *Rhizophagus irregularis* (*Rir*, n=10, error bars denote standard errors). **b, c** Confocal image of WGA-Alexafluor 488-stained arbuscules in root segment of either *P. andersonii* (**b**) or *T. orientalis* (**c**).

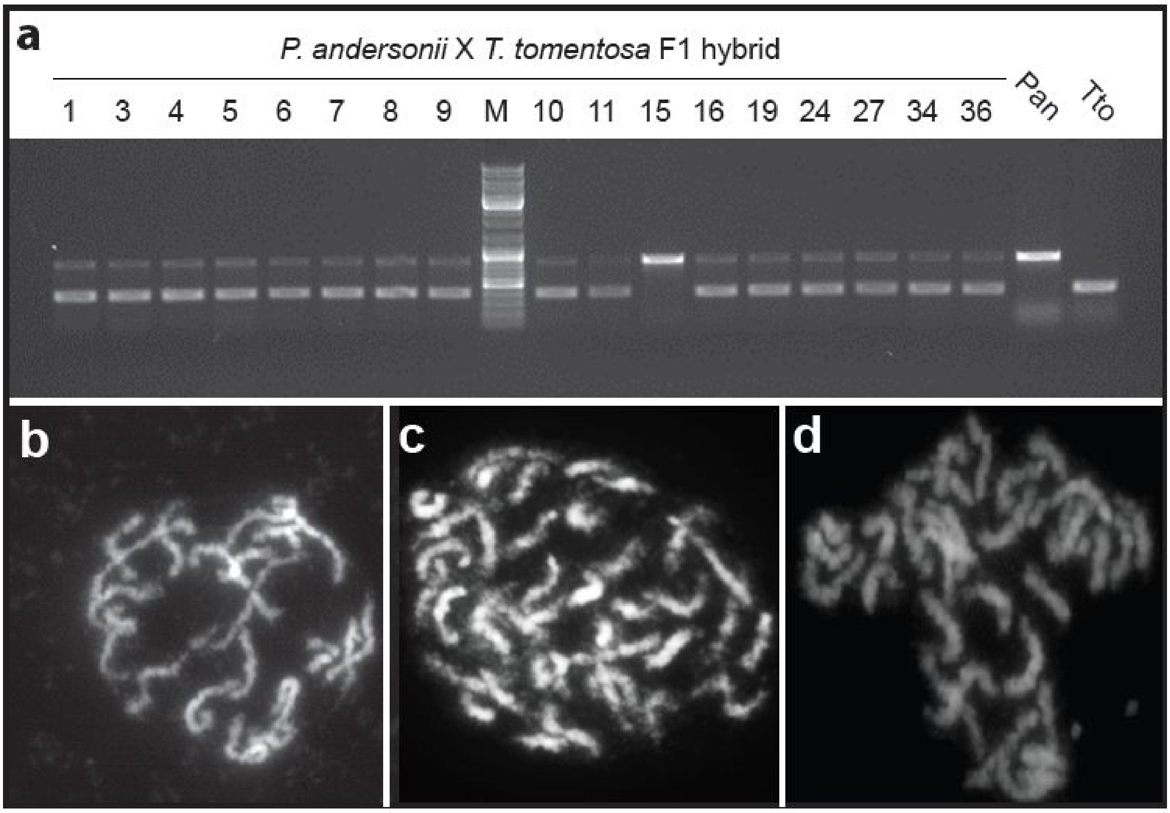


Fig. S2

Genotyping of *Parasponia andersonii* x *Trema tomentosa* F_1 hybrid plants. **a** Genotyping of 17 putative F_1 hybrid plants of the cross *P. andersonii* (Pan) x *T. tomentosa* (Tto) using amplified length polymorphism due to an indel in the *LAX1* promoter. M: generuler DNA ladder mix (ThermoFisher Scientific). Hybrid plants 4, 8, 9, 16, 19 and 36 were used for further experiments. **b-d** Mitotic metaphase chromosome complement of *P. andersonii* (2n=2x=20) (**b**), *T. tomentosa* (2n=4x=40) (**c**), and F1 hybrid (2n=3x=30) (**d**).

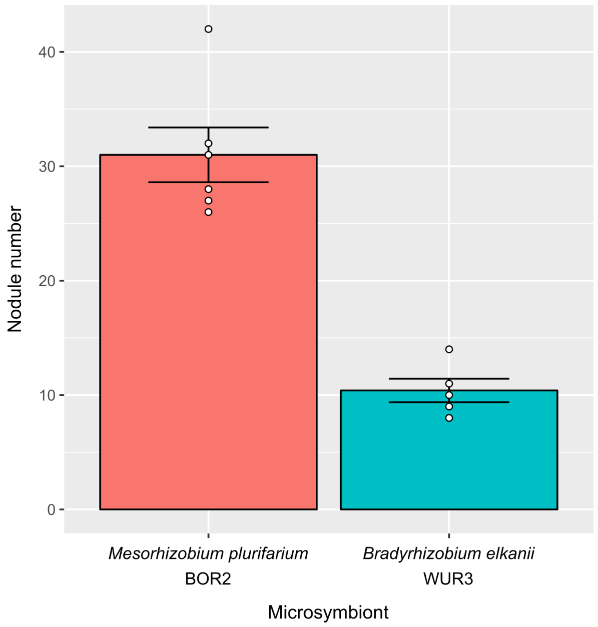


Fig. S3

Nodulation efficiency of *Parasponia andersonii*. Mean number of nodules on roots of *P. andersonii* inoculated with either *Mesorhizobium plurifarium* BOR2 (n=6) or *Bradyrhizobium elkanii* WUR3 microsymbionts (n=5) (6 weeks post inoculation). Dots represent individual measurements.

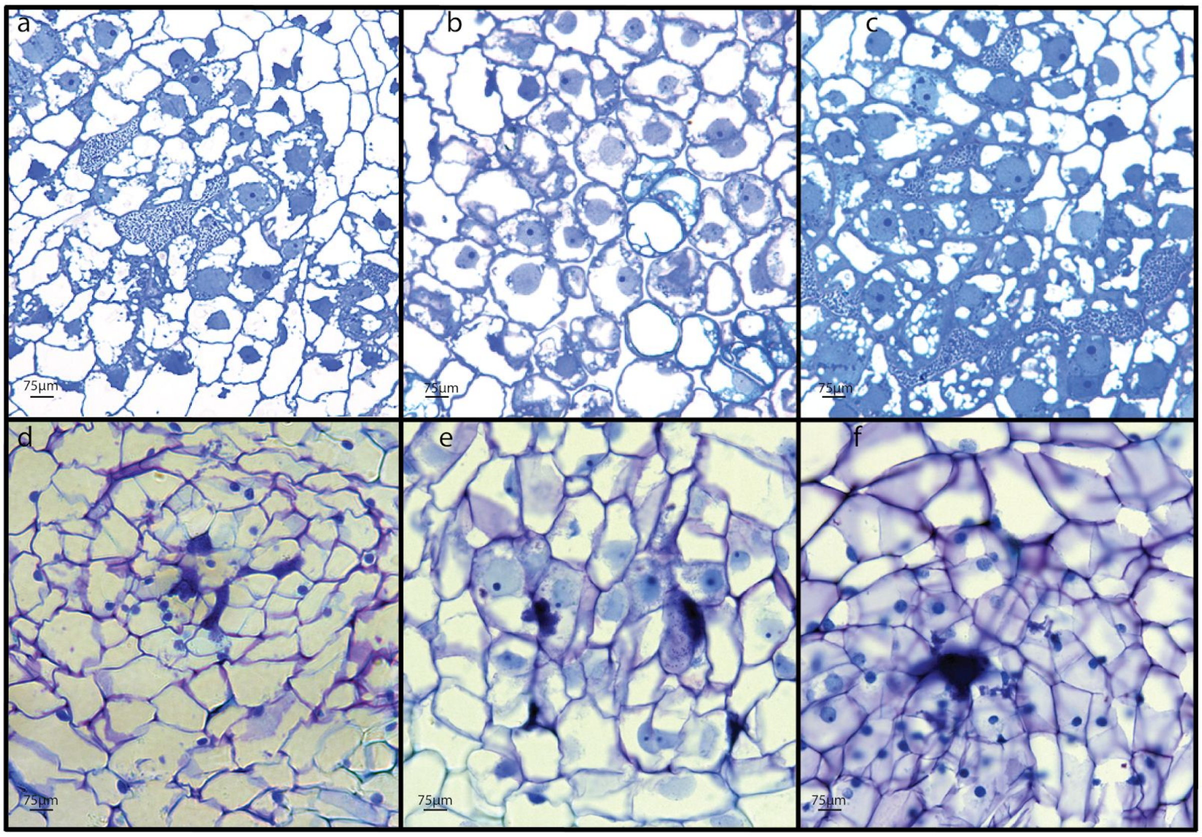


Fig. S4

Longitudinal sections of root nodules of *Parasponia andersonii* x *Trema tomentosa* F1 hybrid plants. Hybrid plants H4, H8, H9, H16, H19 and H36 were clonally propagated and inoculated and inoculated with either *Bradyrhizobium elkanii* WUR3 (**a-c**) or *Mesorhizobium plurifarium* BOR2 (**d-f**). **a** H4 nodule induced by *B. elkanii* WUR3. **b** H8 nodule induced by *B. elkanii* WUR3. **c** H9 nodule induced by *B. elkanii* WUR3. **d** H16 nodule induced by *M. plurifarium* BOR2. **e** H19 nodule induced by *M. plurifarium* BOR2. **f** H36 nodule induced by *M. plurifarium* BOR2. Note absence of intracellular infection in all sectioned nodules.

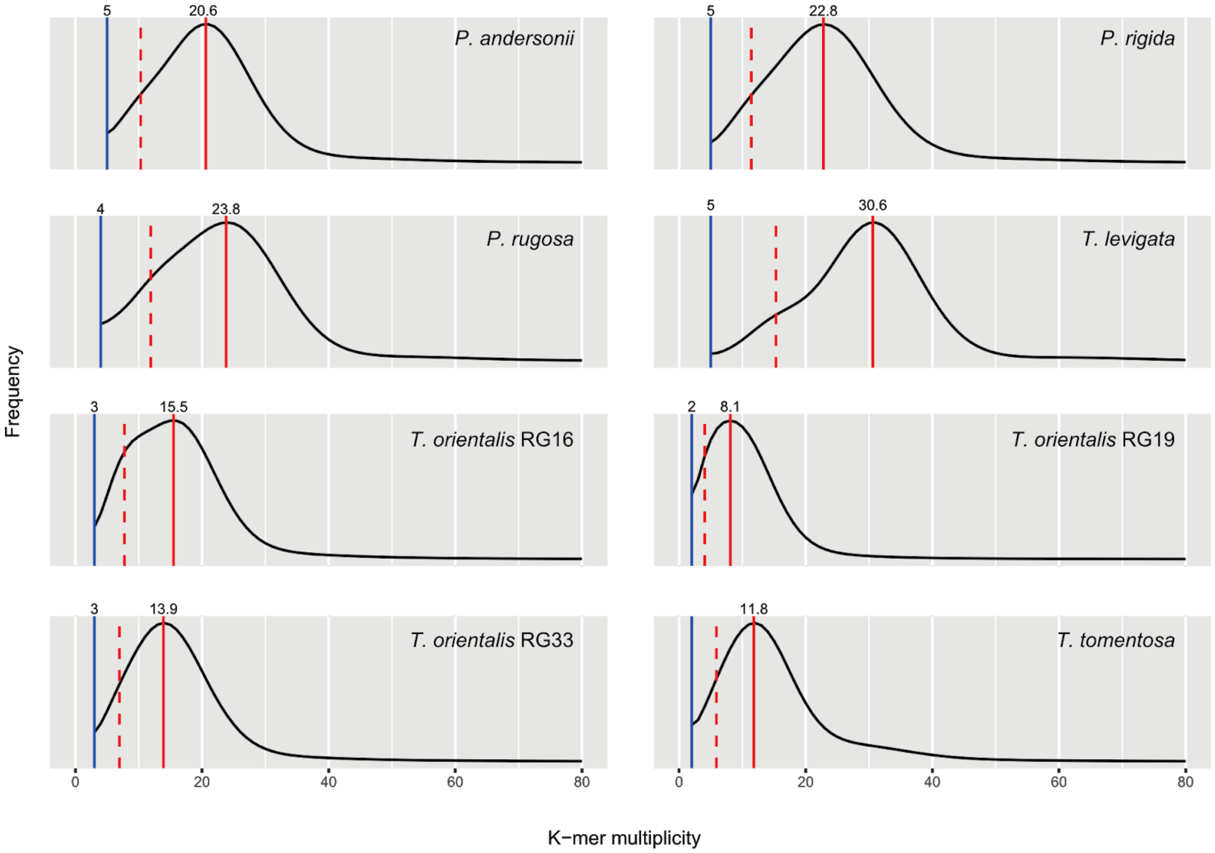


Fig. S5

Genome coverage and heterozygosity estimates based on k-mer analysis of *Parasponia* and *Trema* species. Plots of 21-mer multiplicity frequencies based on jellyfish output showing that *T. levigata* and *T. orientalis* RG16 are relatively heterozygous. Solid red lines indicate estimated genome coverage corresponding to homozygous sequence; dashed red lines indicate half the estimated genome coverage corresponding to heterozygous sequence; blue lines indicate estimated error multiplicity threshold.

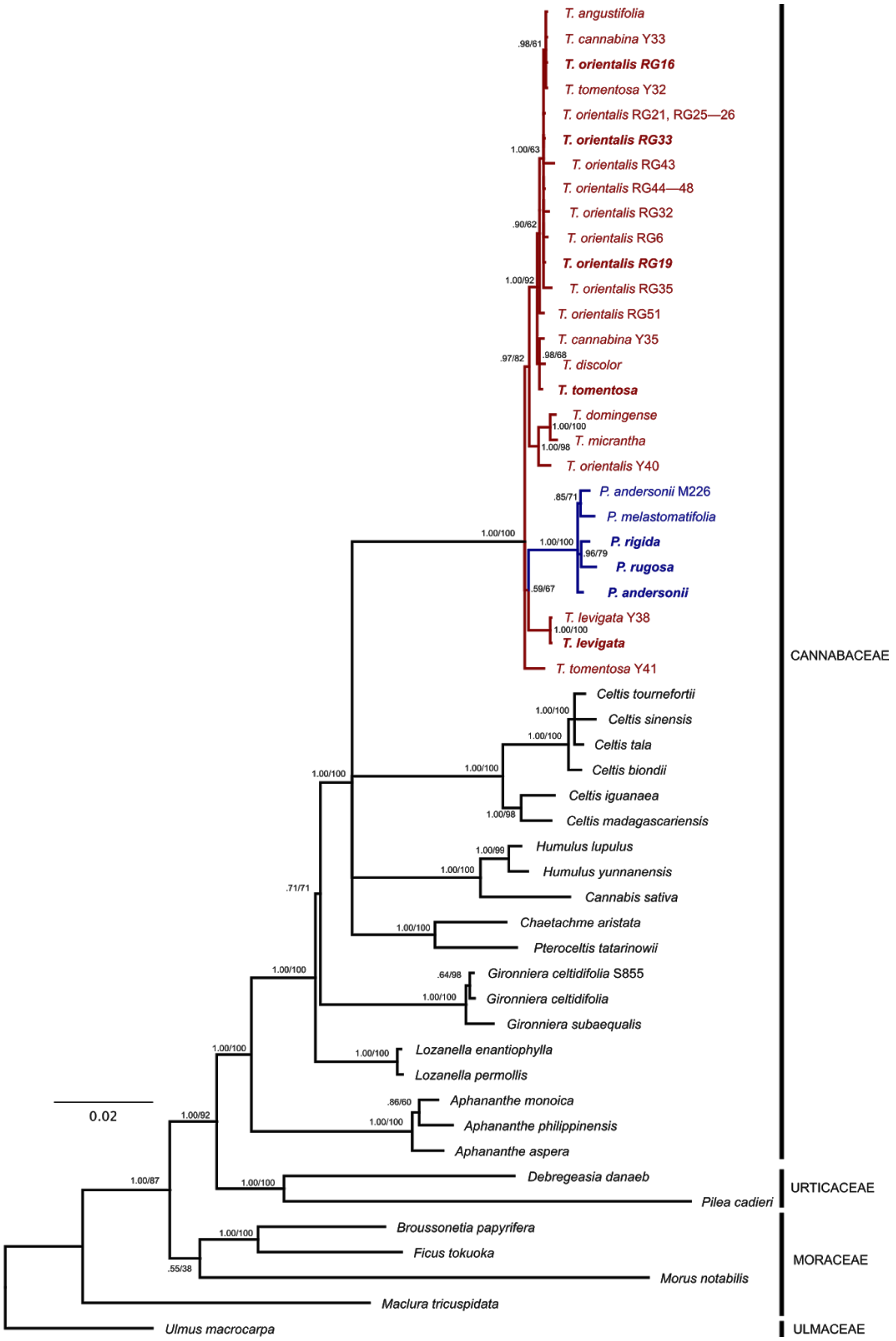


Fig. S6

Phylogenetic reconstruction of the Cannabaceae based on combined analysis of four plastid markers. Bayesian tree based on five optimal partitions and models of sequence evolution: atpB-rbcL combined with trnL-F (GTR+I+G); first codon position of rbcL (GTR+I+G); second position of rbcL (SYM+I+G); third position of rbcL (GTR+G); rps16 (GTR+G). Node values indicate posterior probability / RAxML bootstrap support; scale bar represents substitutions per site. *Parasponia* lineage is in blue, *Trema* lineages are in red. Note that sister relationship of *Parasponia* and *T. levigata* has low bootstrap support, but is independently supported by four shared sequence insertions (Fig. S8). Accessions selected for comparative genome analysis in bold. GenBank accession numbers are in Dataset S7.

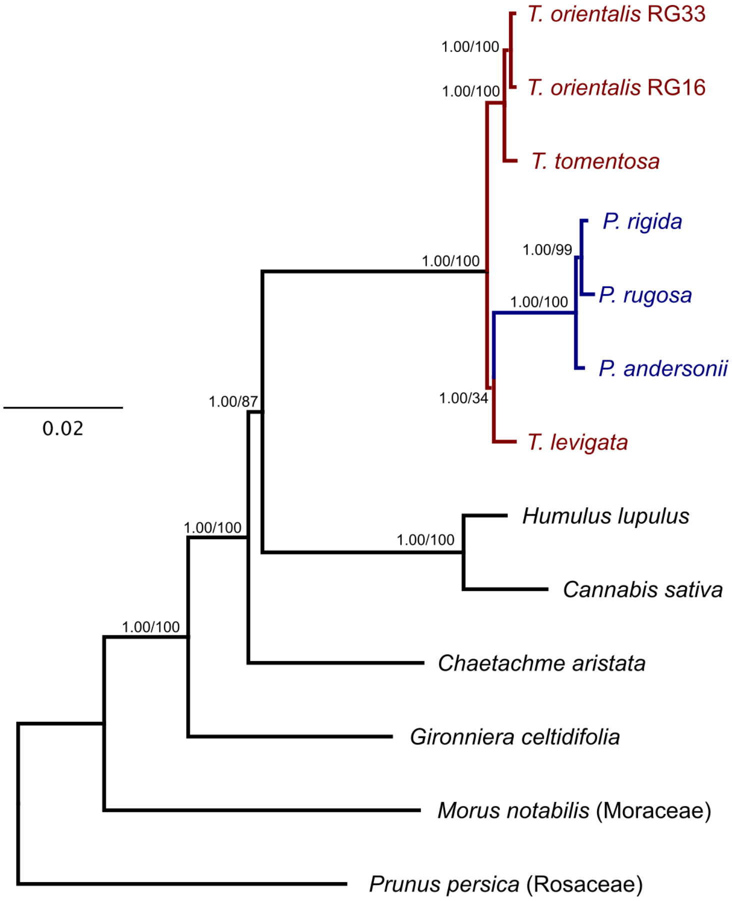


Fig. S7

Phylogenetic reconstruction of the Cannabaceae based on chloroplast genomes. Bayesian tree based on eight optimal partitions and models of sequence evolution: tRNA sequence (HKY+I), rRNA sequence (GTR+I), long single copy region (LSC) coding sequence (GTR+I+G), LSC non-coding sequence (GTR+G), short single copy region (SSC) coding sequence (GTR+G), SSC non-coding sequence (GTR+G), inverted repeat region (IR) coding sequence (GTR+G), and IR non-coding sequence (GTR+G). *Parasponia* lineage is in blue, *Trema* lineages are in red. Note that sister relationship of *Parasponia* and *T. levigata* has low bootstrap support but is independently supported by four shared sequence insertions (Fig. S8). Node values indicate posterior probability / RAxML bootstrap support; scale bar represents substitutions per site. GenBank accession numbers are in Dataset S7.

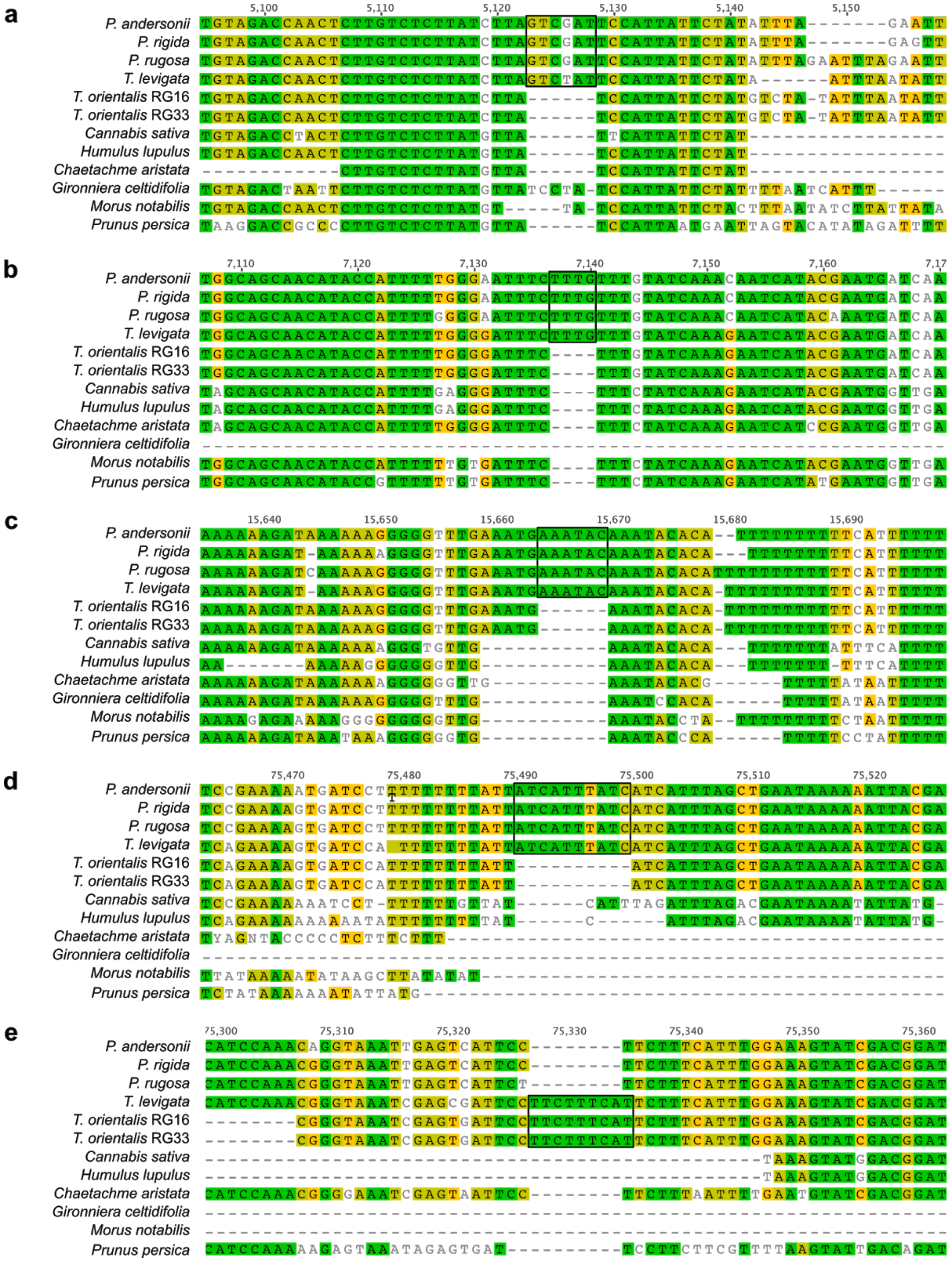

a
5,100
5,110
5,120
5,130
5,140
5,150
P. andersonii
P. rigida
P. rugosa
T. levigata
T. orientalis RG16
T. orientalis RG33
Cannabis sativa
Humulus lupulus
Chaetachme aristata
Gironniera celtidifolia
Morus notabilis
Prunus persica
b
7,110
7,120
7,130
7,140
7,150
7,160
7,17
c
15,640
15,650
15,660
15,670
15,680
15,690
d
75,470
75,480
75,490
75,500
75,510
75,520
e
75,300
75,310
75,320
75,330
75,340
75,350
75,360

Fig. S8

Chloroplast genome insertions in Cannabaceae. Shared sequence insertions in chloroplast genomes supporting (**a-d**) or refuting (**e**) sister relationship of *Parasponia* and *Trema levigata*. **a** *matK-rps16* intergenic spacer, **b** *rps16-psbK* intergenic spacer, **c** *atpF* intron, **d, e** *petA-psbJ* intergenic spacer. Numbers indicate alignment coordinates; colours indicate percent identity while ignoring gaps: green = 100%, olive = 80-100%, yellow = 60-80%; black rectangles mark shared sequence insertions concerned.

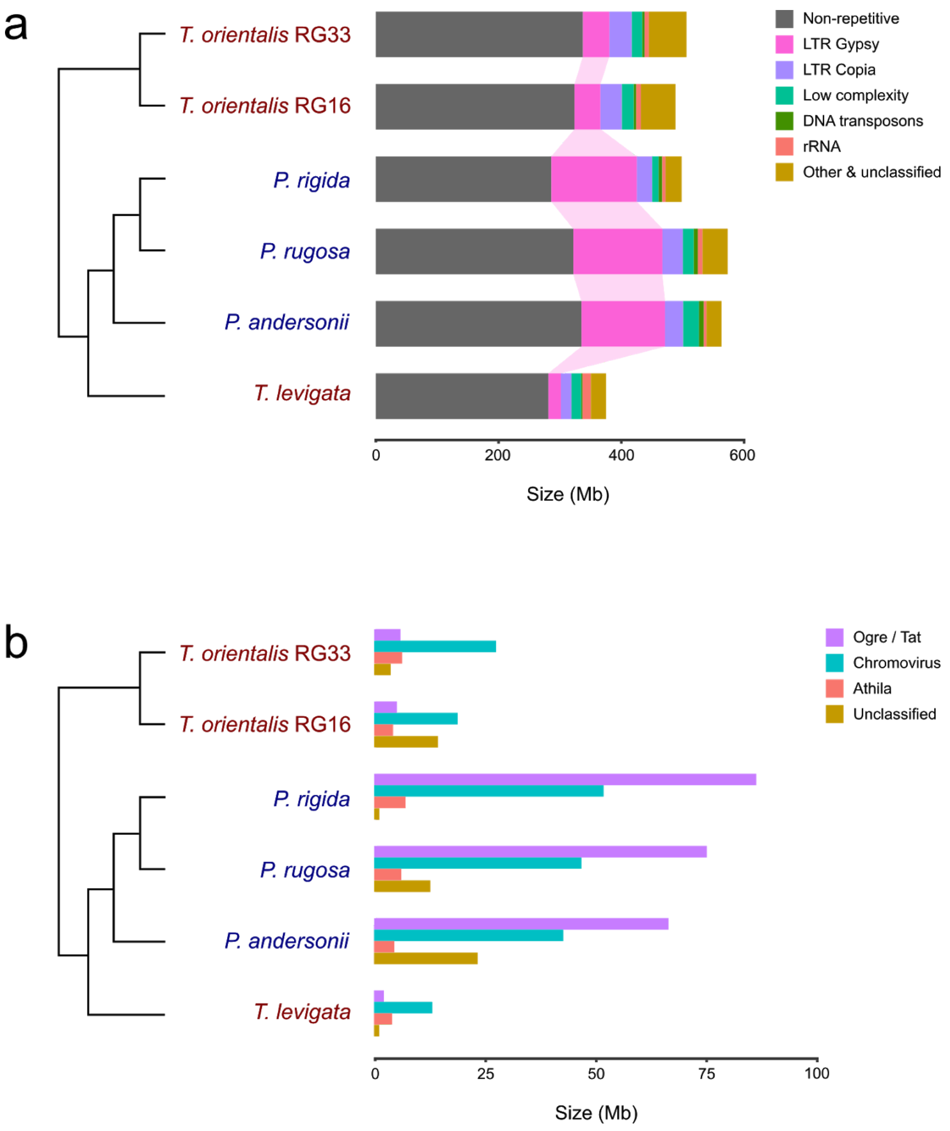


Fig. S9

Parasponia and *Trema* genome structure. Estimated genome sizes and fractions of different classes of repeats as detected by RepeatExplorer, calibrated using k-mer based genome size estimates. **a** Total genome sizes and fractions of major repeat classes showing 1) a conserved size of around 300 Mb of non-repetitive sequence, and 2) a large expansion of gypsy-type LTR retrotransposons in all *Parasponia* compared with all *Trema* species. **b** Estimated size of gypsy-type LTR subclasses in *Parasponia* and *Trema* showing that expansion of this class was mainly due to a tenfold increase of Ogre/Tat to around 75Mb in *Parasponia*.

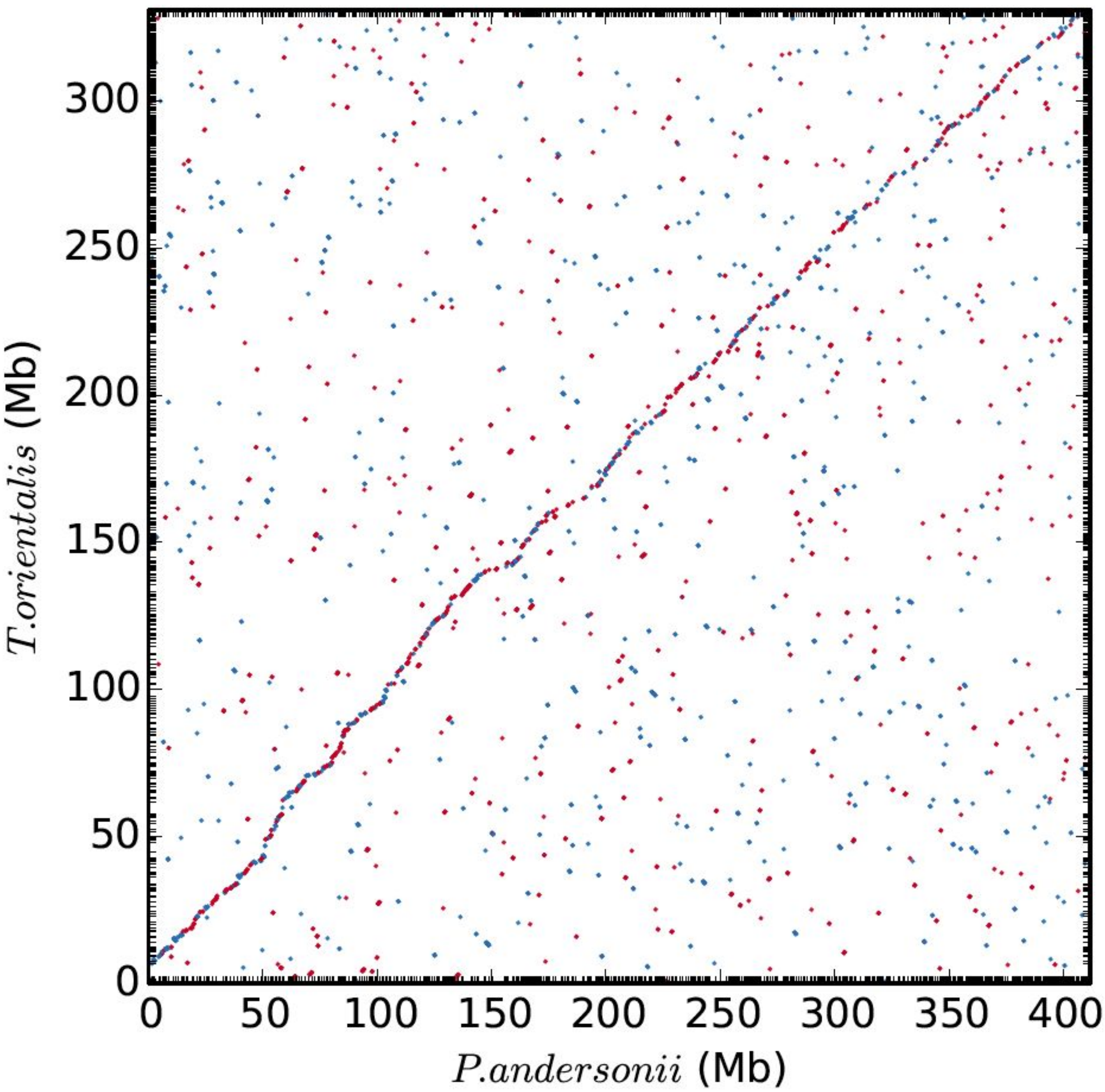


Fig. S10

Whole genome alignment dotplot for *Parasponia andersonii* and *Trema orientalis* RG33. Maximal unique matching (MUM) alignments were generated using nucmer 4.0.0beta with the following settings: breaklength 500, mincluster 200, maxgap 100, minmatch 80, minalign 7000. Forward alignments are red, reverse alignments are blue. Scaffolds are ordered by alignment size, which results in a clear diagonal line indicating the collinearity of the two genomes.

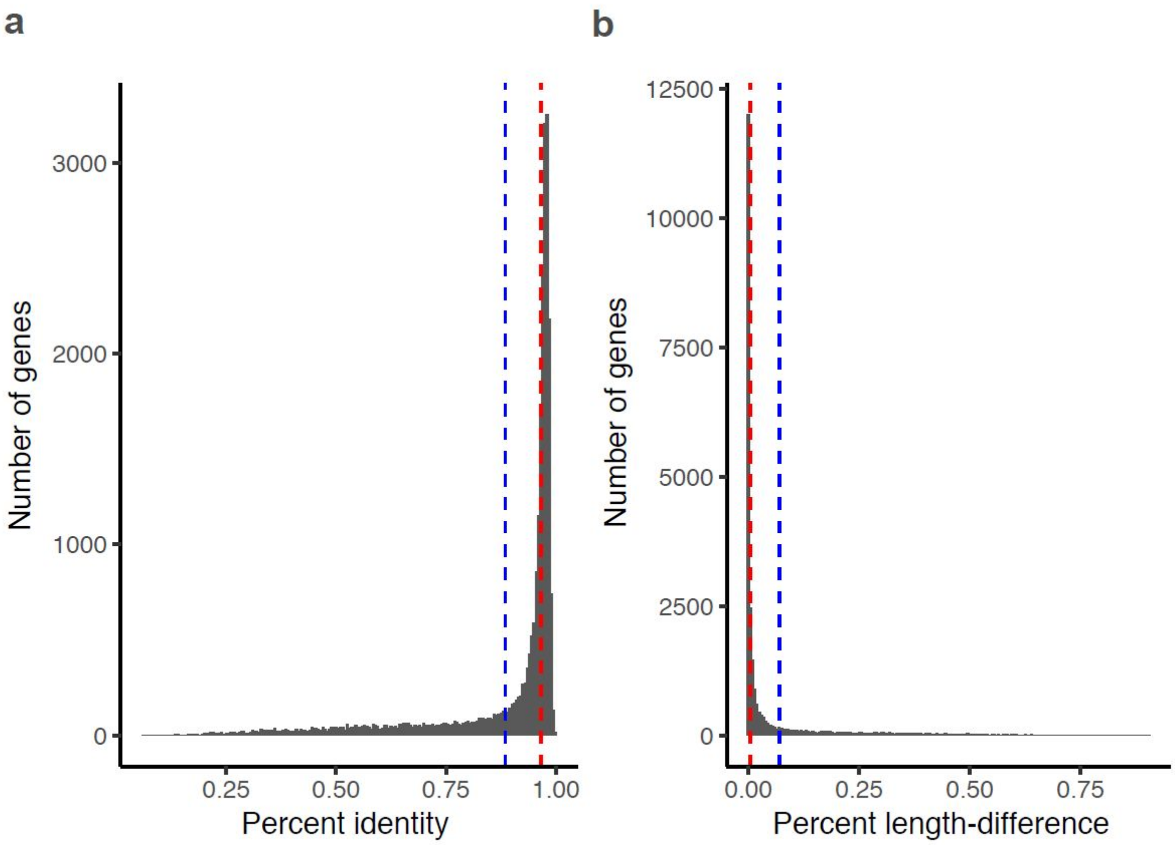


Fig. S11

Identity of *Parasponia andersonii* - *Trema orientalis* putative orthologous gene pairs. Histograms of (**a**) percent nucleotide identity (calculated by taking the fraction of identical nucleotides ignoring end gaps using global alignments produced by MAFFT) and (**b**) length difference of all 25,605 orthologous gene pairs from *P. andersonii* and *T. orientalis* as a percentage of the longest gene. Red line indicates median, blue line indicates mean**.**

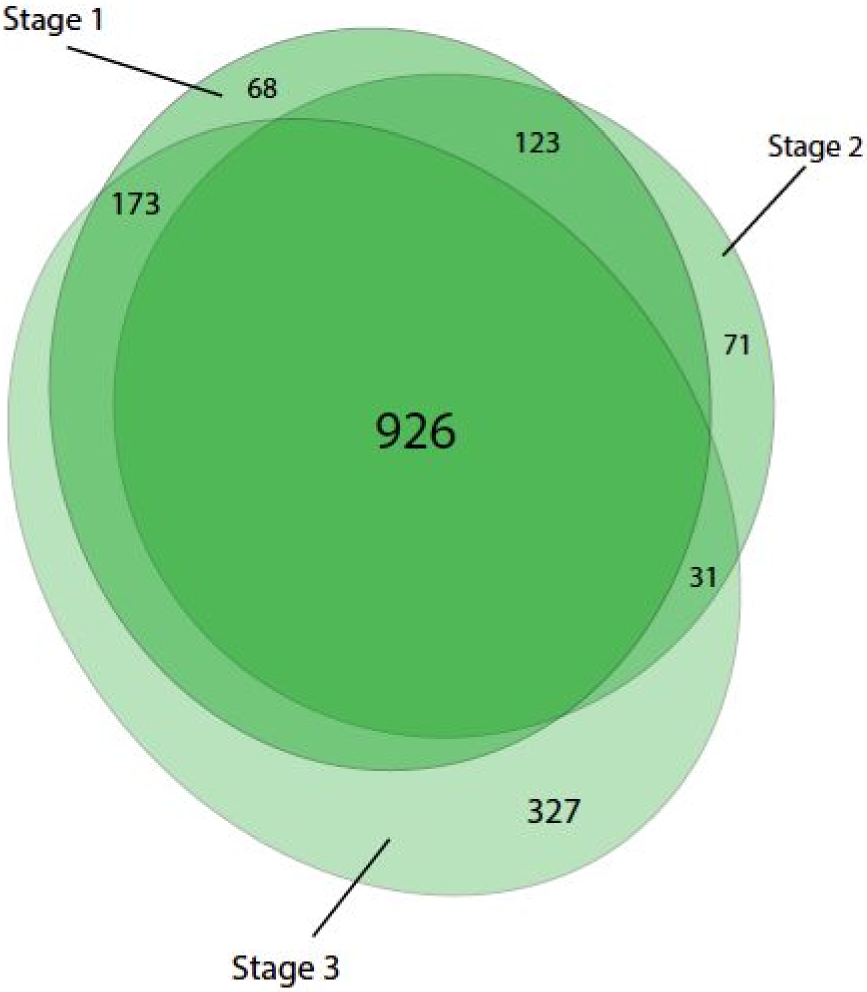


Fig. S12

Venn diagram of *Parasponia andersonii* nodule enhanced genes in 3 developmental stages. Nodule developmental stages according to Fig. 1H-J. List of genes is given in Dataset S4. *P. andersonii* genes are considered ‘nodule enhanced’ when expression is increased >2-fold in any of 3 nodule developmental stages when compared to non-inoculated root sample. Largest fraction concerns genes enhanced in all 3 stages.

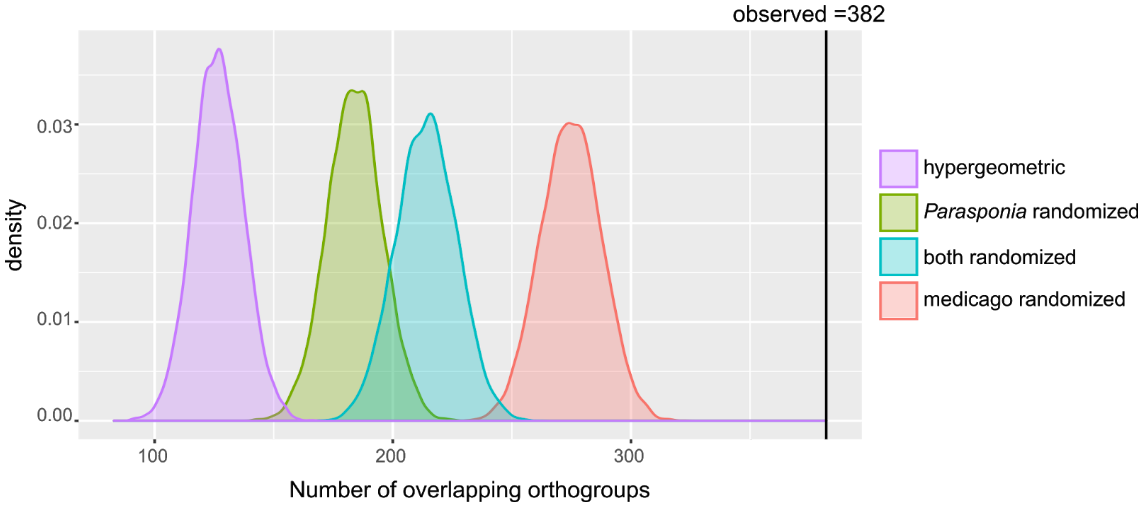
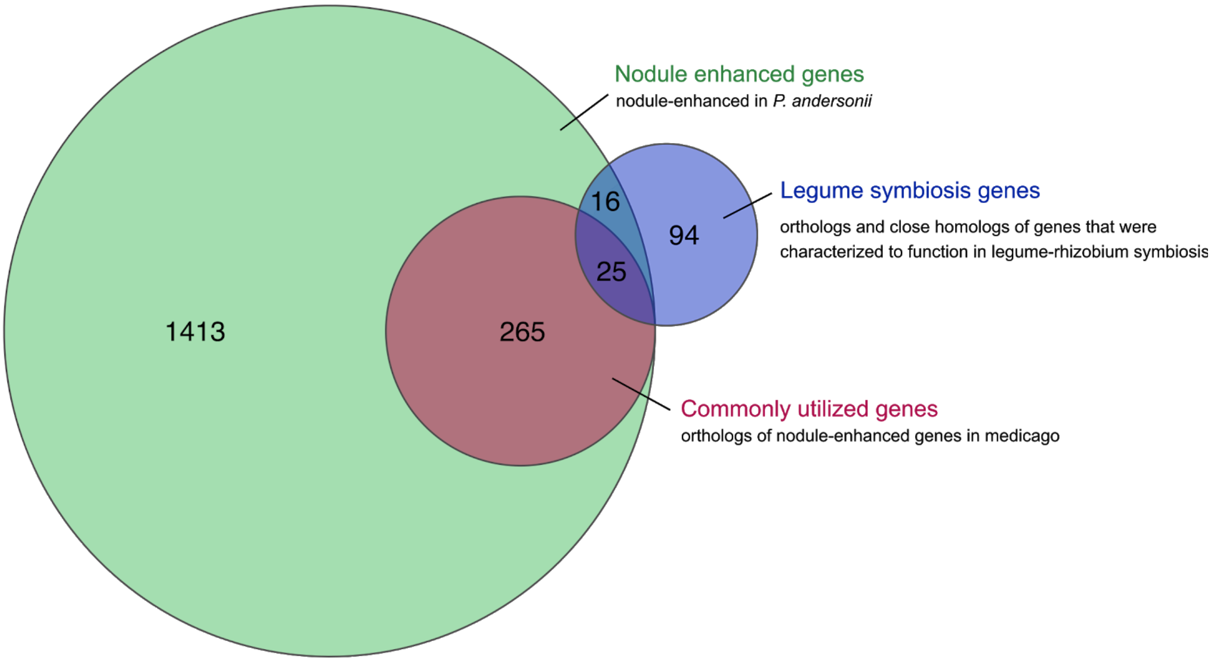


Fig. S13

Statistical testing of common utilization of genes in *Parasponia* and medicago. To assess common utilization of genes in *Parasponia* and medicago nodules we performed statistical testing of overlap between *P. andersonii* and medicago nodule-enhanced genes. Overlap was calculated based on orthogroup membership (i.e. when an orthogroup contains nodule-enhanced genes from *P. andersonii* and medicago it is scored as overlap). Significance of set overlaps is usually calculated based on the hypergeometric distribution. However, because larger orthogroups have higher chance of overlap, the hypergeometric is not suitable. We therefore assessed significance with a permutation test where the null distribution is based on overlap found when gene-orthogroup membership is randomized in the *Parasponia* gene set, the medicago gene set, or both sets (n=10,000). Figure shows density plots of both hypergeometric distribution and permutation random variates. Vertical line shows the observed number of 382 overlapping orthogroups (p<0.0001 based on permutation tests).



Nodule enhanced genes
nodule-enhanced in *P. andersonii*
Legume symbiosis genes
orthologs and close homologs of genes that were characterized to function in legume-rhizobium symbiosis
Commonly utilized genes
orthologs of nodule-enhanced genes in medicago
1413
265
16
25
94

Fig. S14

Venn diagram of *Parasponia andersonii* symbiosis gene sets. Nodule enhanced genes have a significantly enhanced expression level (fold change > 2, p < 0.05, DESeq2 Wald test) in any of three developmental stages (N = 1,719; Fig. S12, Dataset S4). Commonly utilized genes are nodule-enhanced in *P. andersonii* as well as in the legume medicago (N = 290; Dataset S5, S6). Legume symbiosis genes are putative orthologs or close homologs of genes that were characterized to function in legume-rhizobium symbiosis (N = 135; Dataset S1, S2).

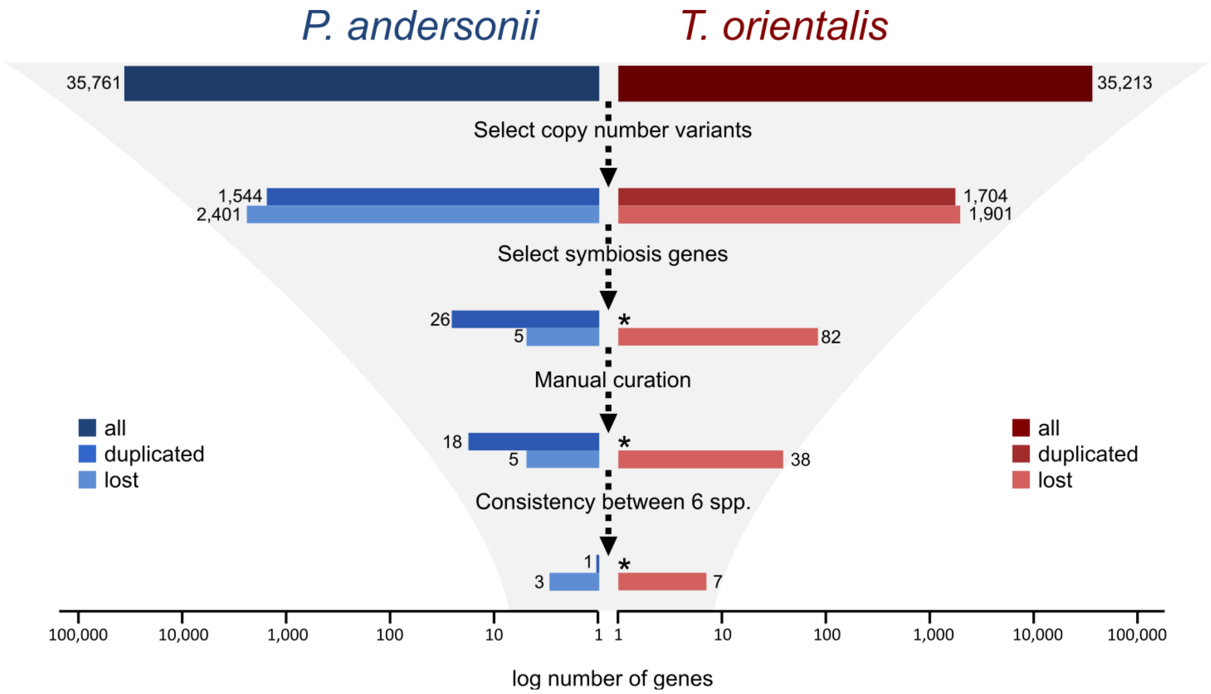


Fig. S15

Filtering steps to establish consistent copy number variants (CNVs) in symbiosis genes. Starting from all annotated gene models, we sequentially i) selected automatically predicted CNVs based on phylogenetic analysis of orthogroups, ii) selected CNVs in 1,817 symbiosis genes, iii) manually curated automatically predicted CNVs using whole-genome alignments between *Parasponia andersonii* and *Trema orientalis*, and iv) selected only those CNVs that were consistent between all sampled *Parasponia* and *Trema* species. *Note that *Trema*-specific duplications were ignored in the second step because we considered these irrelevant in a symbiotic context.

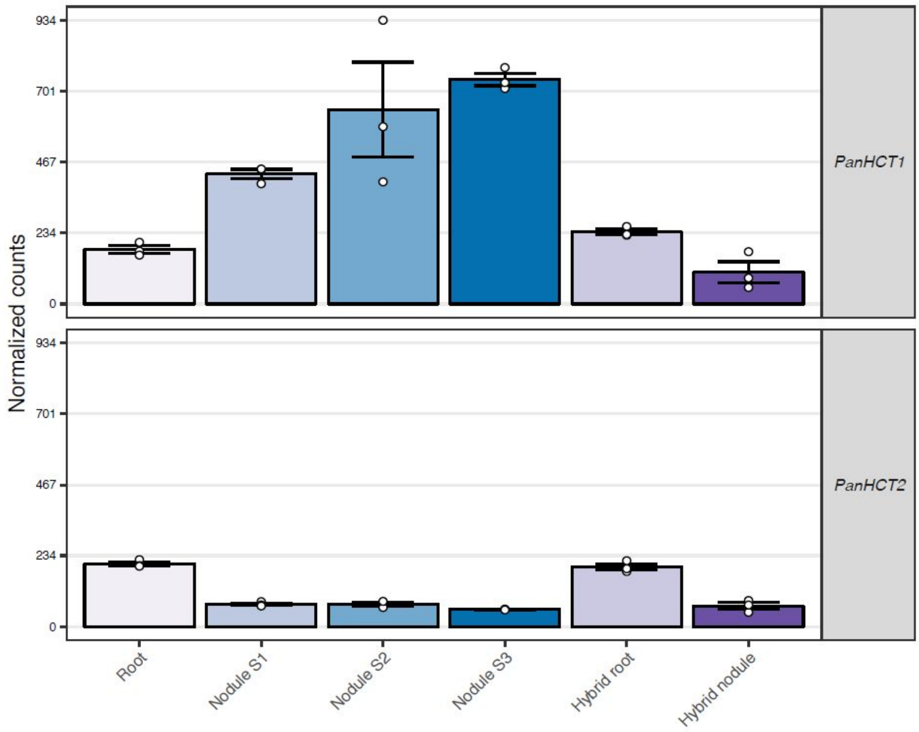


Fig. S16

Expression profile of *PanHCT1* and *PanHCT2* genes. Expression of *Parasponia andersonii HYDROXYCINNAMOYL-COA SHIKIMATE TRANSFERASE 1* (*PanHCT1*) and *PanHCT2* in *P. andersonii* roots, stage 1-3 nodules, and in *P. andersonii* x *Trema tomentosa* F_1 hybrid roots and nodules (line H9). *PanHCT1* and *PanHCT2* represent the only *Parasponia*-specific gene duplication in the defined symbiosis gene set, as *PanHCT1* is upregulated in nodules. Expression is given in DESeq2 normalized read counts, error bars represent standard error of three biological replicates, dots represent individual expression levels.

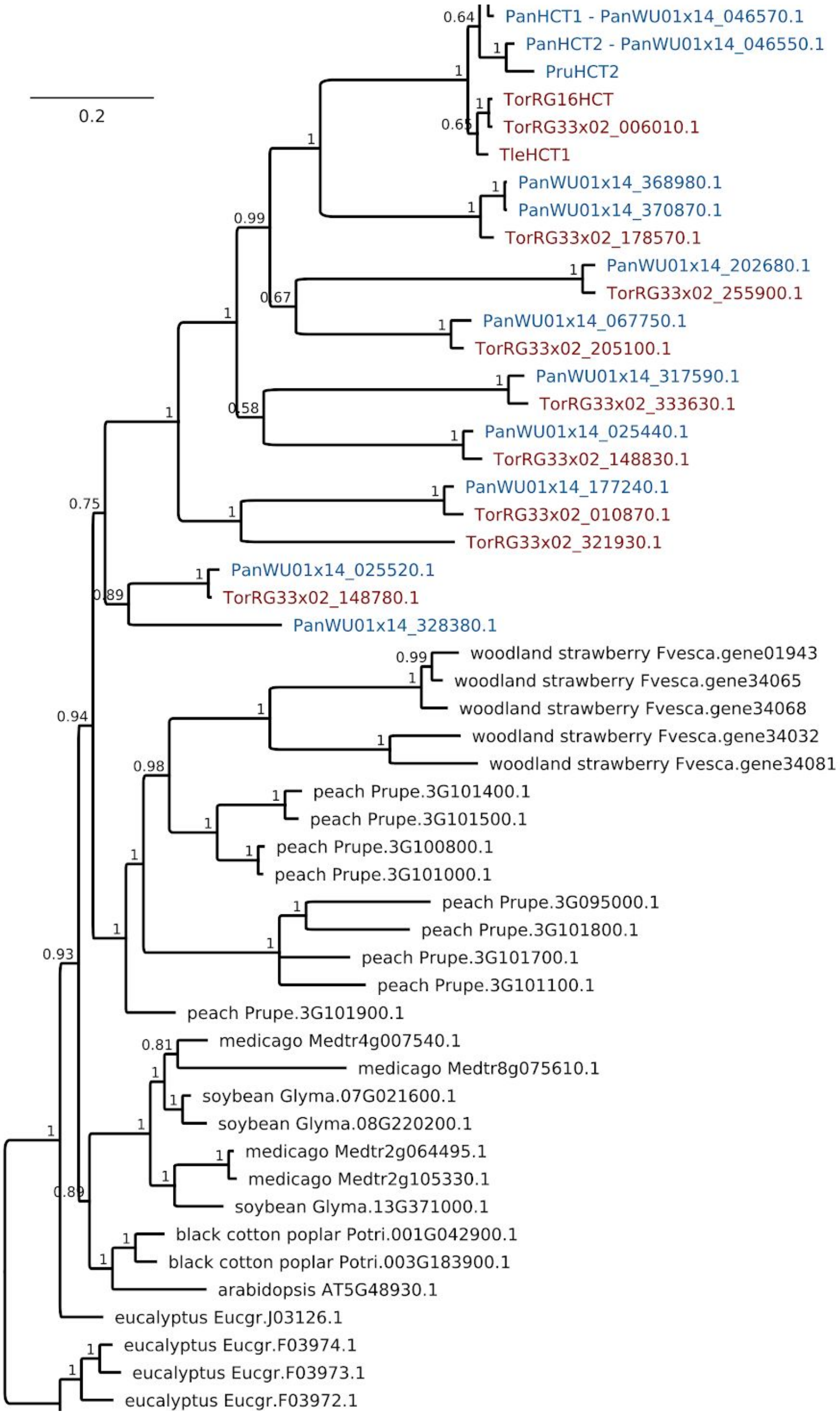


Fig. S17

Phylogenetic reconstruction of Hydroxycinnamoyl-CoA Shikimate Transferase (HCT) orthogroup. *HCT* orthogroup was created by merging OG0001291, OG0016758, OG0016791, OG0018560, OG0020327, OG0020921, OG0022256 & OG0023772, supplemented with HCT1 and HCT2 orthologs of *Parasponia rigida, P. rugosa*, *Trema* *orientalis* RG16 and *T. levigata. PriHCT2* is a putative pseudogene and was not included. HCT1 and HCT2 represent the only *Parasponia* specific gene duplication in the defined symbiosis gene set, as *PanHCT1* was found to be upregulated in nodules. Species included: *Parasponia andersonii* (Pan); *P. rigida* (Pri); *P. rugosa* (Pru) (all in blue); *Trema orientalis* (Tor); *T. orientalis* RG16 (TorRG16); *T. levigata* (Tle) (all in red); *Medicago truncatula* (Mt); *Glycine max* (Glyma),*Prunus persica* (Prupe), *Populus trichocarpa* (Potri); *Fragaria vesca* (Fvesca); *Eucalyptus grandis* (Eugr); *Arabidopsis thaliana* (AT). Node numbers indicate posterior probabilities below 1, scale bar represents substitutions per site.

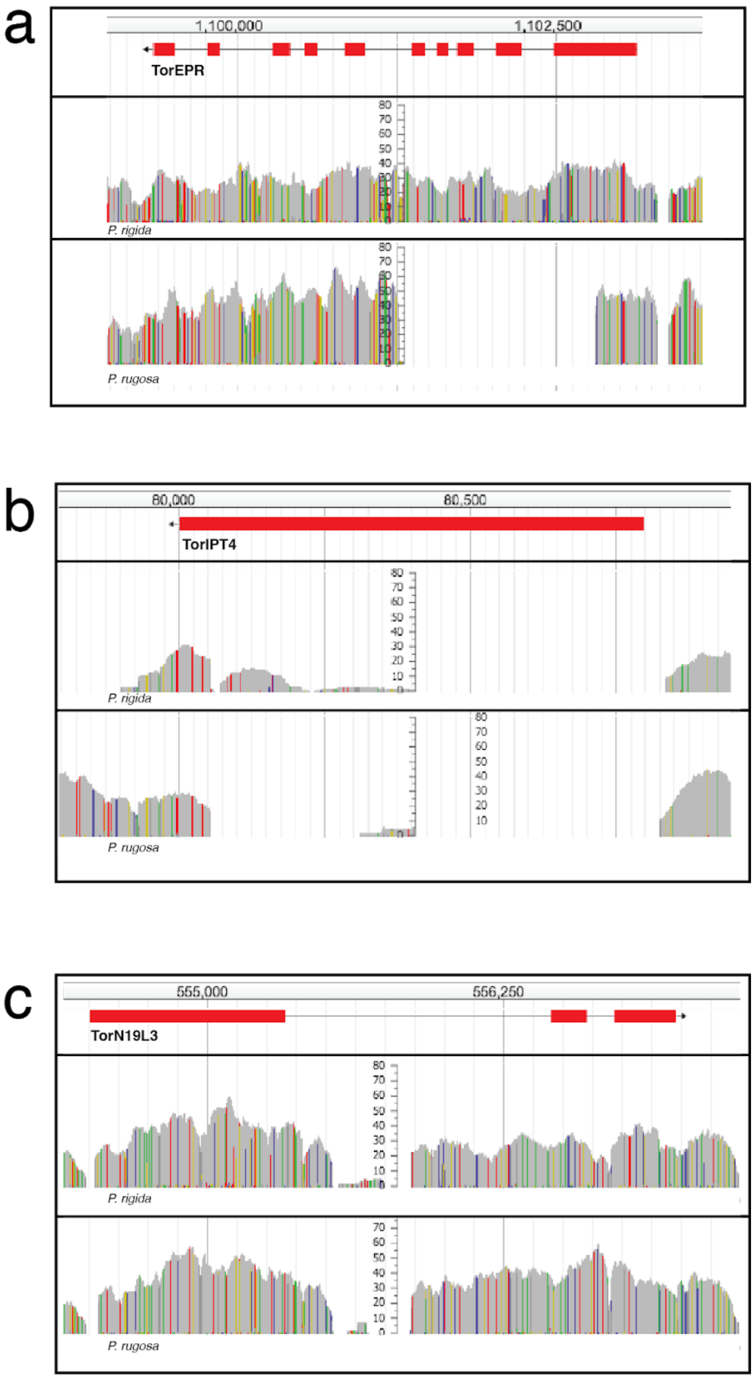


Fig. S18

Read mappings of *Parasponia rigida* and *P. rugosa* to the *Trema orientalis EPR, IPT4* and *N19L3* genes. Read mappings to genic regions of (**a**) *TorEPR*, illustrating absence of a large part of the gene in *P. rugosa*. In *P. rigida EPR* is a pseudogene due to a single bp insertion causing a frame-shift in the first exon (Fig. S20), (**b**) *TorIPT4*, illustrating absence of most of the gene in both *P. rigida* and *P. rugosa*, (**c**) *TorN19L3*, which is a pseudogene in both *P. rigida* and *P. rugosa* due to a large sequence insertion and a 10bp deletion causing frame shifts in the first exon (note that annotated pseudogene sequences are deposited on

GenBank, see Dataset S7 for accession numbers). Coordinates on the x-axis correspond to those of the *T. orientalis* scaffold; red bars depict *T. orientalis* gene models; histograms depict read coverage in grey; nucleotide differences from the *T. orientalis* reference scaffold are in color (green = adenine, blue = cytosine, yellow = guanine, red = thymine).

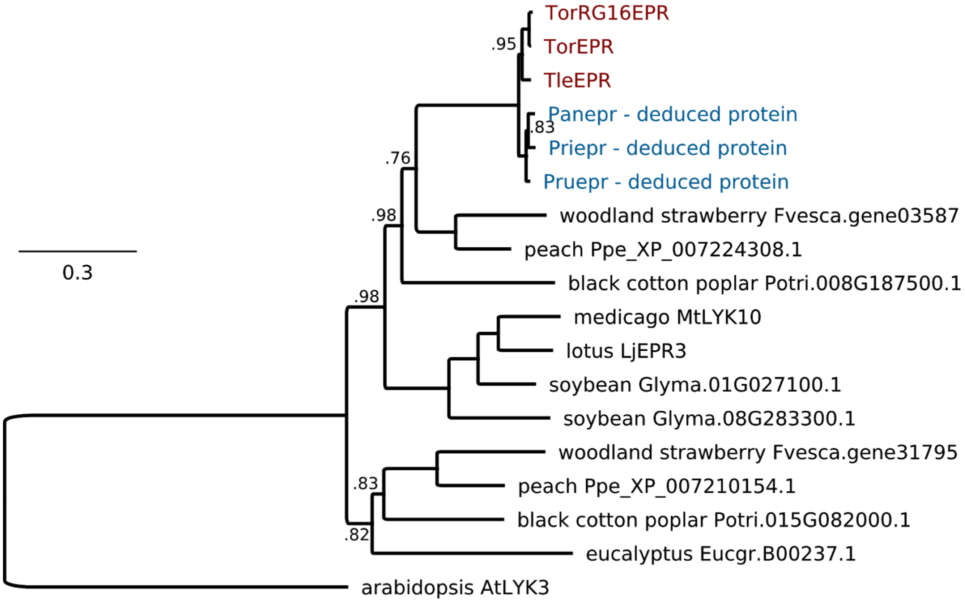

TorRG16EPR
.95
TorEPR
TleEPR
Panepr - deduced protein
.83
Priepr - deduced protein
Pruepr - deduced protein
.76
.98
woodland strawberry Fvesca.gene03587
peach Ppe_XP_007224308.1
0.3
black cotton poplar Potri.008G187500.1
.98
medicago MtLYK10
lotus LjEPR3
soybean Glyma.01G027100.1
soybean Glyma.08G283300.1
woodland strawberry Fvesca.gene31795
.83
peach Ppe_XP_007210154.1
black cotton poplar Potri.015G082000.1
.82
eucalyptus Eucgr.B00237.1
arabidopsis AtLYK3

Fig. S19

Phylogenetic reconstruction of the EPR3 orthogroup. Alignment of orthogroup OG0010070 containing exopolysaccharide receptor LjEPR3. Note that all *Parasponia* species lack a functional *EPR* (Fig. S18, S20). Species included: *Trema orientalis* RG33 (Tor); *Trema orientalis* RG16 (TorRG16); *Trema levigata* (Tle) (all in red); *Parasponia Andersonii* (Pan); *Parasponia Rigida* (Pri) *Parasponia Rugosa* (Pru) (all in blue). *Medicago truncatula* (Mt); *Glycine max* (Glyma), *Populus trichocarpa* (Potri); *Fragaria vesca* (Fvesca); *Eucalyptus grandis* (Eugr). Node numbers indicate posterior probabilities below 1, scale bar represents substitutions per site.

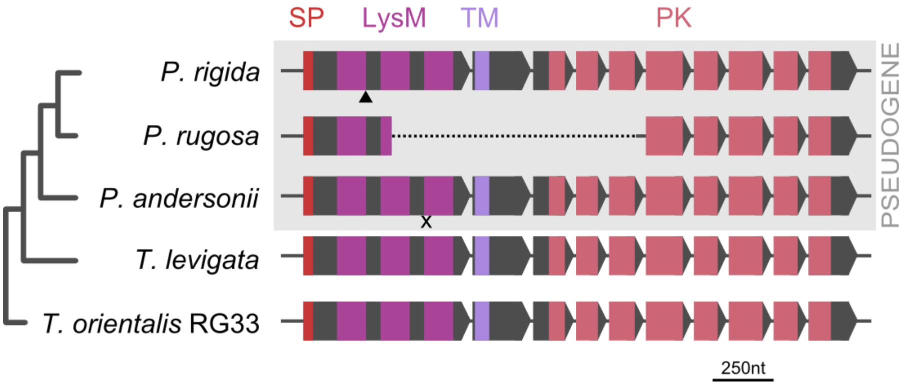


Fig. S20

Independent pseudogenization in *Parasponia* species of *EPR* that is orthologous to the *Lotus japonicus* exopolysaccharide receptor *LjEPR3*. Introns are indicated, but not scaled. X indicates premature stop codon in *P. andersonii epr*, triangle indicate frame-shift in *P. rigida epr*, whereas *P. rugosa epr* contains a large deletion. SP = signal peptide (red); LysM: 3 Lysin Motif domains (magenta); TM = transmembrane domain (lilac); PK = protein kinase (pink).

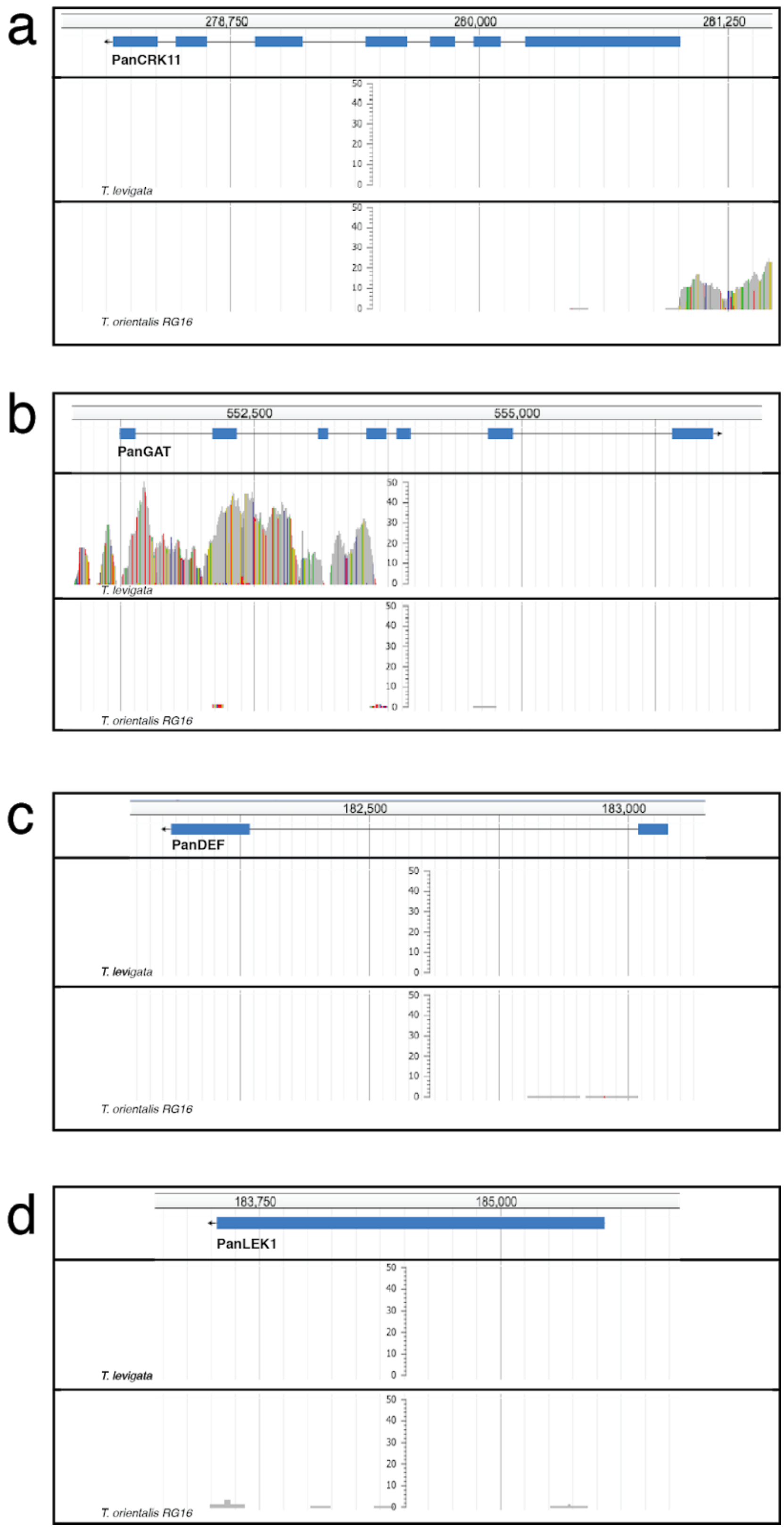


Fig. S21

Read mappings of *Trema orientalis* RG16 and *T. levigata* to the *Parasponia andersonii CRK11*, *GAT*, *DEF* and *LEK1* genes. Read mappings to genic regions of (**a**) *PanCRK11*, illustrating absence of the gene in both *T. levigata* and *T. orientalis* RG16, (**b**) *PanGAT*, illustrating absence of most of the gene in *T. levigata* and total absence in *T. orientalis* RG16, (**c**) *PanDEF*, illustrating absence of the gene in both *T. levigata* and *T. orientalis* RG16. (**d**) *PanLEK1*, illustrating absence of the gene in both *T. levigata* and *T. orientalis* RG16. Coordinates on the x-axis correspond to those of the *P. andersonii* scaffold; orange bars depict *P. andersonii* gene models; histograms depict read coverage in grey; nucleotide differences from the *P. andersonii* reference scaffold are in color (green = adenine, blue = cytosine, yellow = guanine, red = thymine).

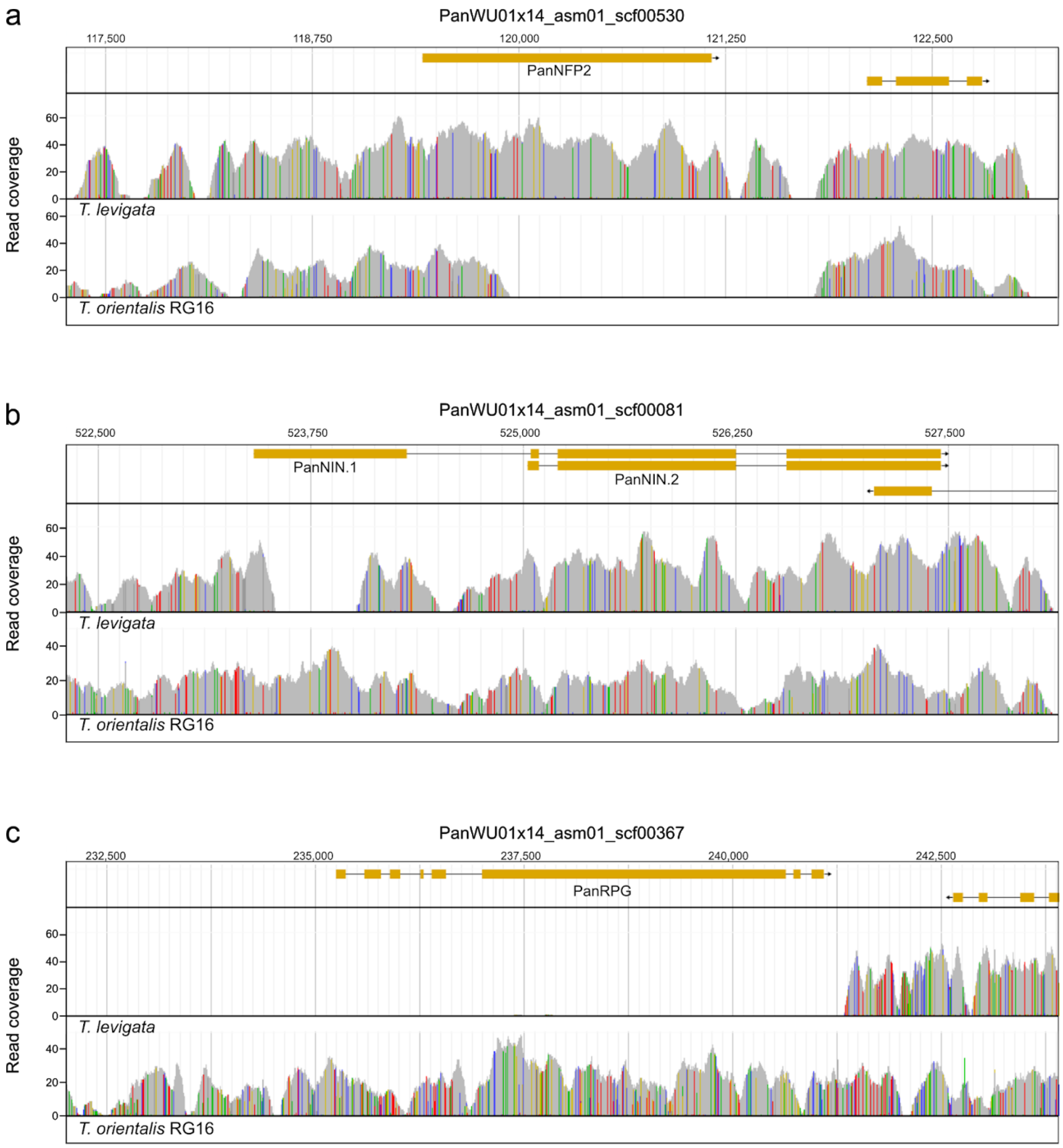
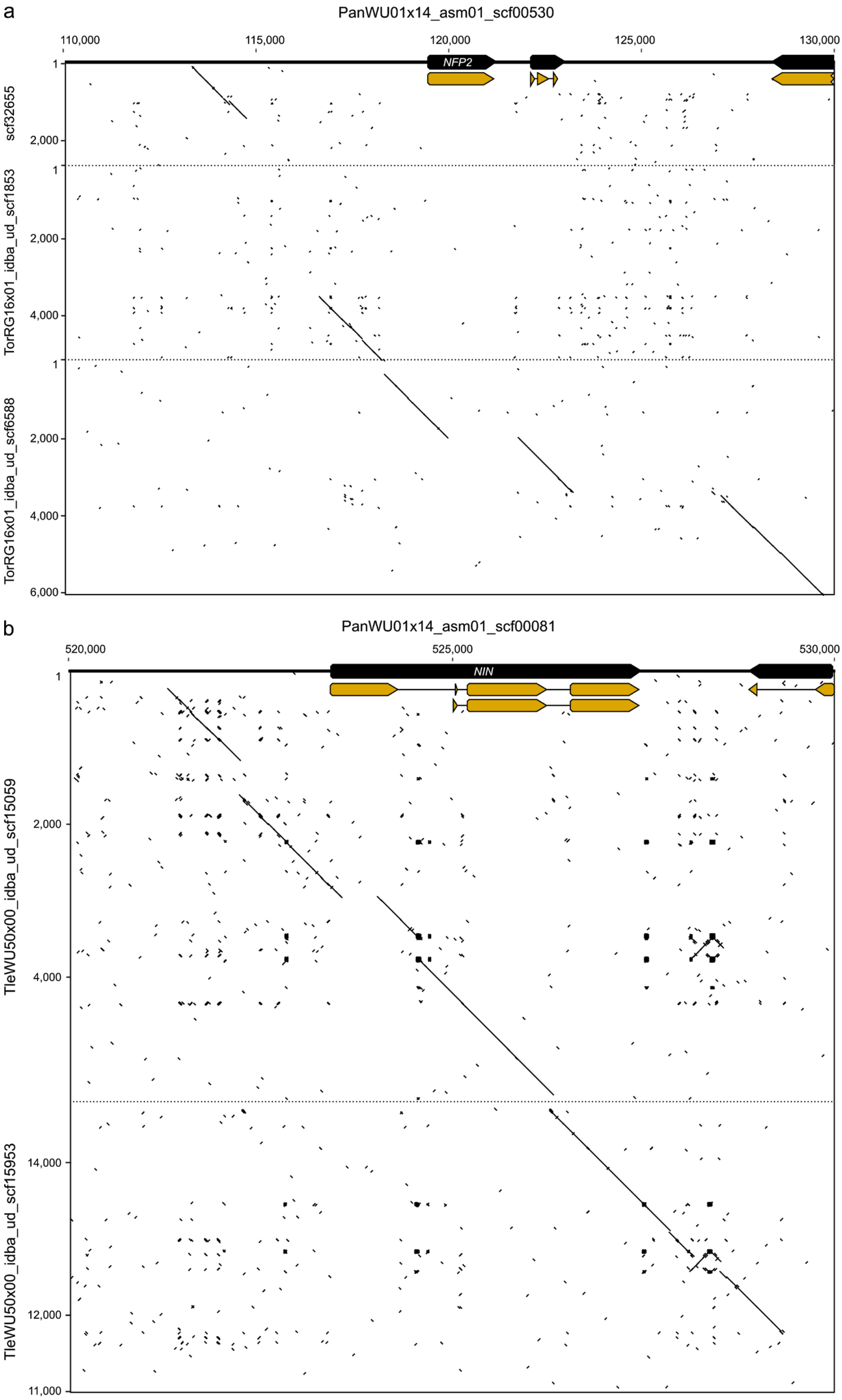


Fig. S22

Read mappings of *Trema orientalis* RG16 and *T. levigata* to the *Parasponia andersonii NFP2*, *NIN* and *RPG* genes. Read mappings to genic regions of (**a**) *PanNFP2*, illustrating absence of a large part of the gene in *T. orientalis* RG16, (**b**) *PanNIN*, illustrating absence of a large part of the canonical first exon in *T. levigata*, (**c**) *PanRPG*, illustrating absence of the gene in *T. levigata*. Coordinates on the x-axis correspond to those of the *P. andersonii* scaffold; orange bars depict *P. andersonii* gene models; histograms depict read coverage in grey; nucleotide differences from the *P. andersonii* reference scaffold are in color (green = adenine, blue = cytosine, yellow = guanine, red = thymine).


a
PanWU01x14_asm01_scf00530
110,000
115,000
120,000
125,000
130,000
NFP2
scf32655
1
2,000
TorRG16x01_idba_ud_scf1853
1
2,000
4,000
TorRG16x01_idba_ud_scf6588
1
2,000
4,000
6,000
b
PanWU01x14_asm01_scf00081
520,000
525,000
530,000
NIN
TleWU50x00_idba_ud_scf15059
1
2,000
4,000
TleWU50x00_idba_ud_scf15953
14,000
12,000
11,000

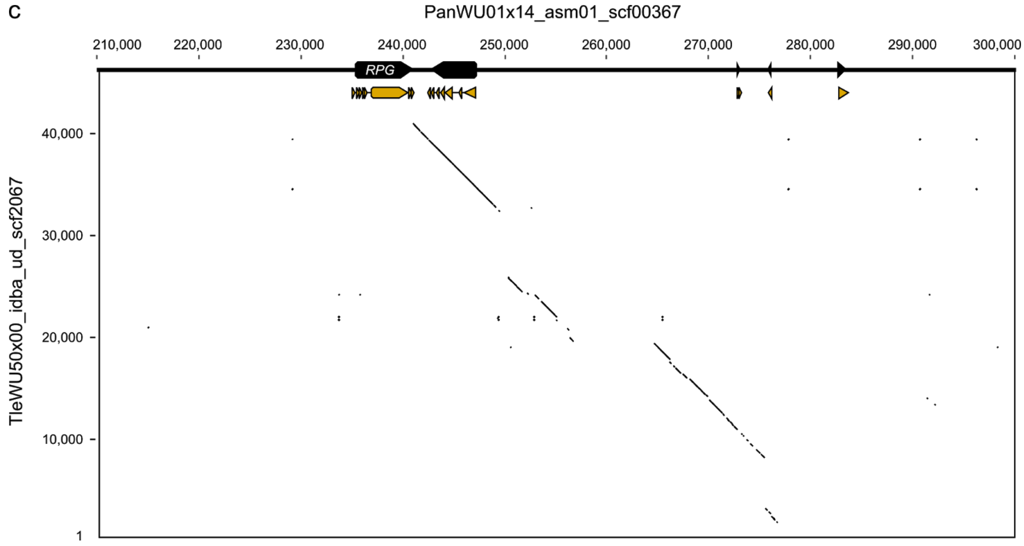


Fig. S23

Genomic alignments of *Trema orientalis* RG16 or *Trema levigata* to *Parasponia andersonii NFP2, NIN, and RPG* gene regions. Genome alignment(s) of (**a**) *T. orientalis* RG16 with *PanNFP2* gene region, (**b**) *T. levigata* with *PanNIN* gene region, (**c**) *T. levigata* with *PanRPG* gene region. Coordinates correspond to those on the draft genome scaffolds; *Parasponia andersonii* gene and CDS models are depicted in black and orange, respectively; different genomic scaffolds are separated by dashed lines. Genomic alignments were performed with the EMBOSS 6.5.7 tool dotmatcher as implemented in the Geneious function dotplot.

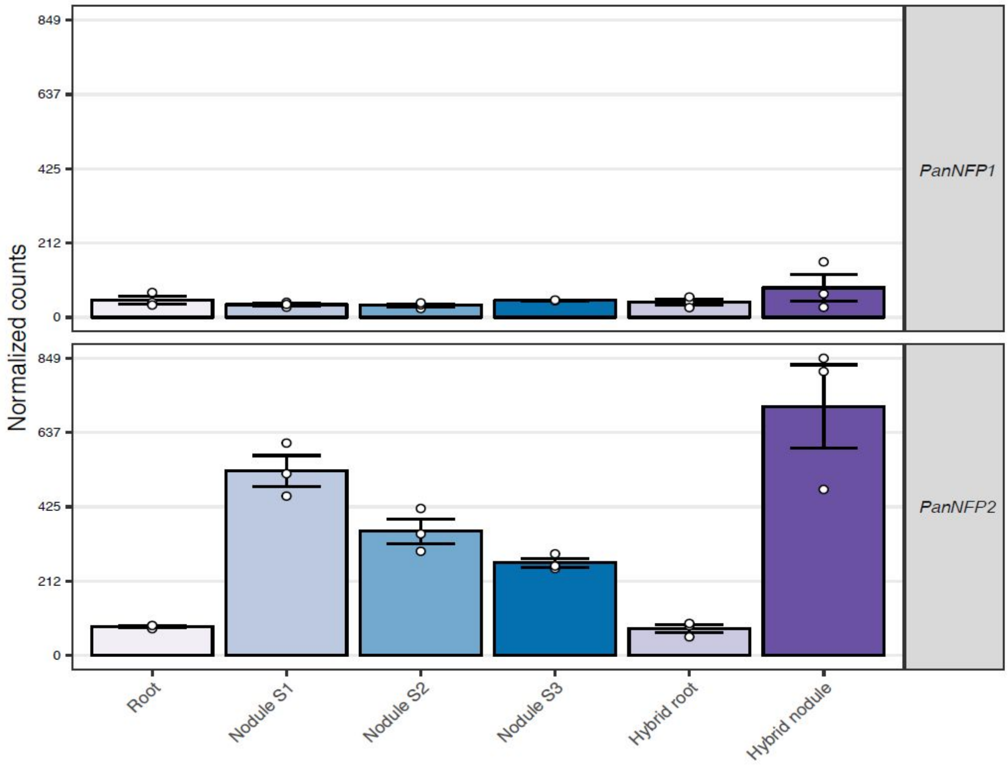


Fig. S24

Expression profile of *PanNFP1* and *PanNFP2* genes. Expression of *P. andersonii NOD FACTOR PERCEPTION 1* (*PanNFP1*) and *PanNFP2* in *P. andersonii* roots, stage 1-3 nodules, and in *P. andersonii x T. tomentosa* F1 hybrid roots and nodules. Expression is given in DESeq2 normalized read counts, error bars represent standard error of three biological replicates, dots represent individual expression levels.

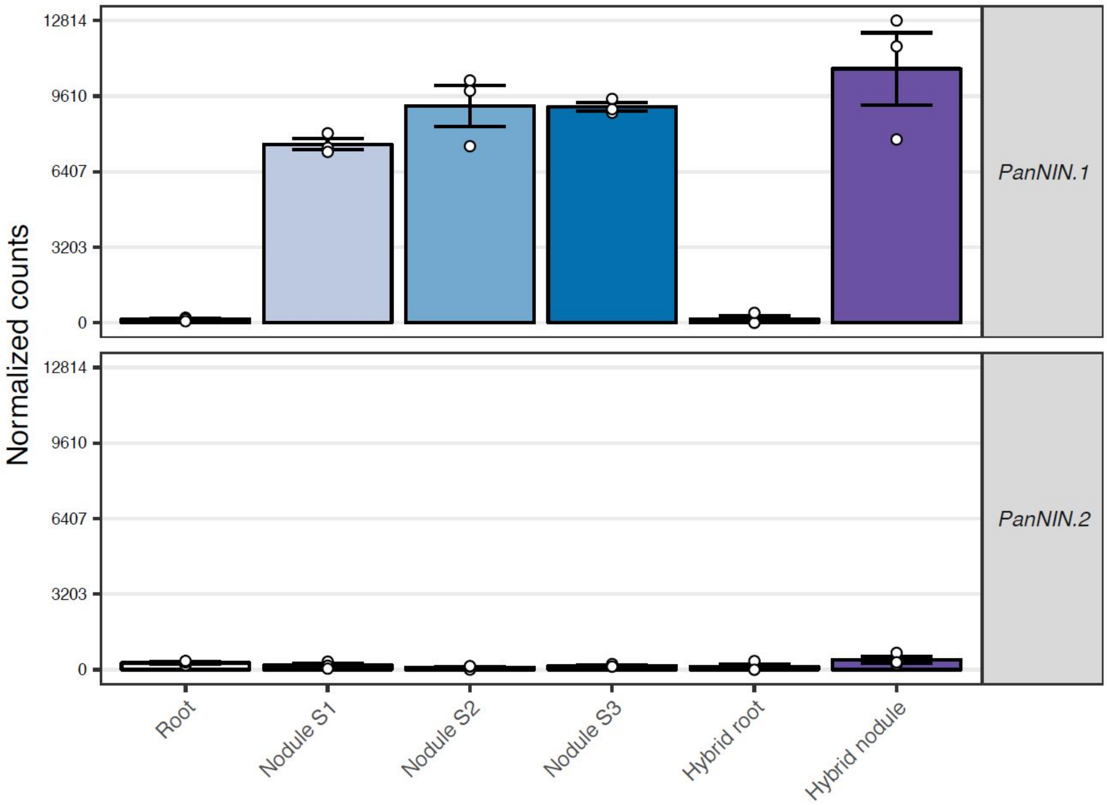


Fig. S25

Expression of *Parasponia andersonii NODULE INCEPTION* (*PanNIN*) gene splice variants. *PanNIN.1* encodes a canonical symbiotic protein, whereas *PanNIN.2* encodes a shorter protein variant that is the result of an alternative start site in an intron. Expression levels were determined by identifying unique DNA sequences for both variants; spanning the intron in case of *PanNIN.1* (CTGCCAAGCGCTTGAGGCTGTTGATCTT), or including the start site of *PanNIN.2* (GCCAATTACCTTGCAGGCTGTTGATCTT) and counting all occurrences in the RNA-seq reads. DESeq2 size factors were used to normalize these counts. The fraction of these normalized counts between *PanNIN.1* and *PanNIN.2* was used to scale the expression levels. Error bars represent standard error of three biological replicates, dots represent individual expression levels.

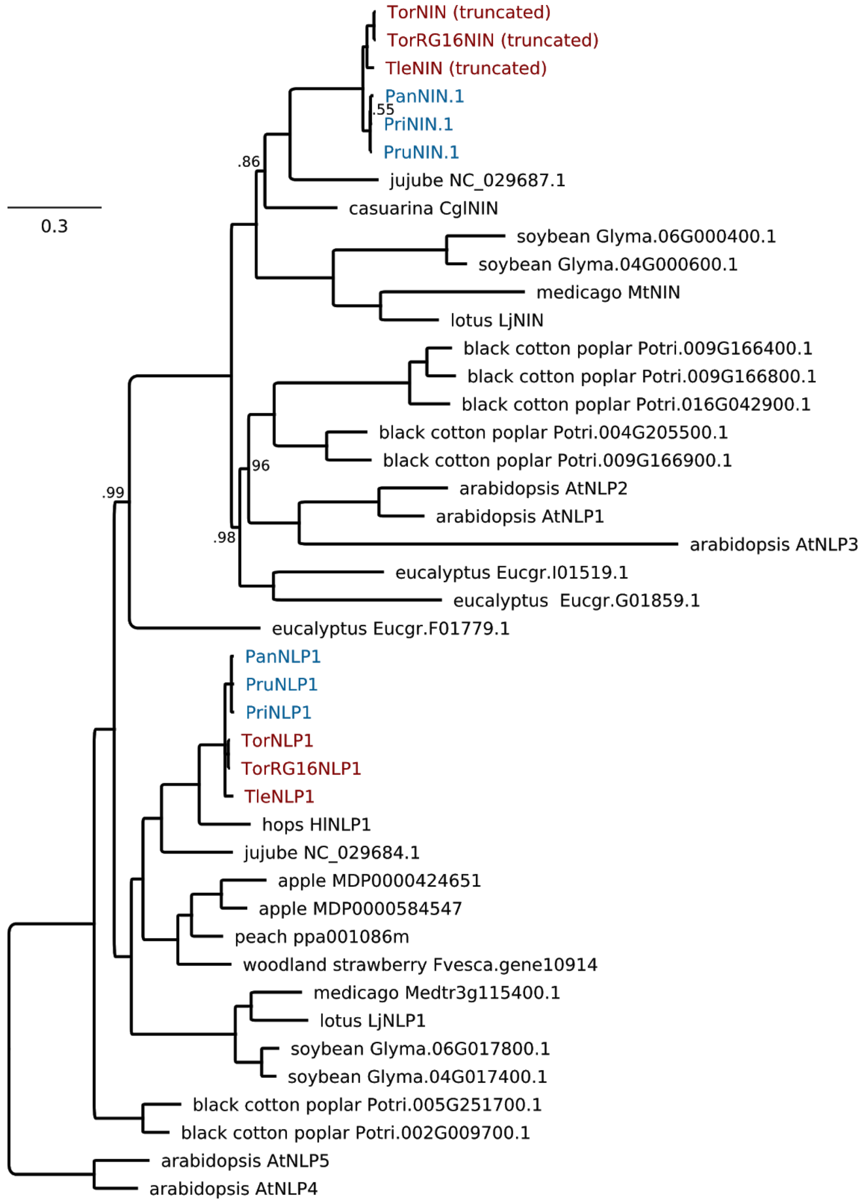

TorNIN (truncated)
TorRG16NIN (truncated)
TleNIN (truncated)
PanNIN.1
.55
PriNIN.1
PruNIN.1
.86
jujube NC_029687.1
casuarina CglNIN
0.3
soybean Glyma.06G000400.1
soybean Glyma.04G000600.1
medicago MtNIN
lotus LjNIN
black cotton poplar Potri.009G166400.1
black cotton poplar Potri.009G166800.1
black cotton poplar Potri.016G042900.1
black cotton poplar Potri.004G205500.1
black cotton poplar Potri.009G166900.1
96
arabidopsis AtNLP2
.99
arabidopsis AtNLP1
.98
arabidopsis AtNLP3
eucalyptus Eucgr.I01519.1
eucalyptus Eucgr.G01859.1
eucalyptus Eucgr.F01779.1
PanNLP1
PruNLP1
PriNLP1
TorNLP1
TorRG16NLP1
TleNLP1
hops HlNLP1
jujube NC_029684.1
apple MDP0000424651
apple MDP0000584547
peach ppa001086m
woodland strawberry Fvesca.gene10914
medicago Medtr3g115400.1
lotus LjNLP1
soybean Glyma.06G017800.1
soybean Glyma.04G017400.1
black cotton poplar Potri.005G251700.1
black cotton poplar Potri.002G009700.1
arabidopsis AtNLP5
arabidopsis AtNLP4

Fig. S26

Phylogenetic reconstruction of NIN orthogroup. Alignment of OG0001118, which includes NIN and NLP1 (NIN-LIKE PROTEIN 1)-like proteins, supplemented with additional species. AtNLP4 and AtNLP5 were included as outgroup. *Parasponia* spp. marked in blue, *Trema* spp. In red. Note that in *Trema* species NIN only occurs in truncated forms (Fig. 6). Included species: *Parasponia andersonii* (Pan); *Parasponia rigida* (Pri); *Parasponia rugosa* (Pru); *Trema orientalis* RG33 (Tor); *Trema orientalis* RG16 (TorRG16); *Trema levigata* (Tle); medicago (*Medicago truncatula,* Mt); lotus (*Lotus japonicus,* Lj); soybean (*Glycine max,* Glyma); peach (*Prunus persica,* ppe); woodland strawberry (*Fragaria vesca,* Fvesca); back cotton poplar (*Populus trichocarpa,* Potri); eucalyptus (*Eucalyptus grandis,* Eugr); arabidopsis (*Arabidopsis thaliana,* At), jujube (*Ziziphus Jujube*) apple (*Malus x domestica*), mulberry (*Morus Notabilis*), hop (*Humulus Lupulus* (*natsume.shinsuwase.v1.0*)), and casuarina (*Casuarina glauca*)*.* Node numbers indicate posterior probabilities below 1, scale bar represents substitutions per site.

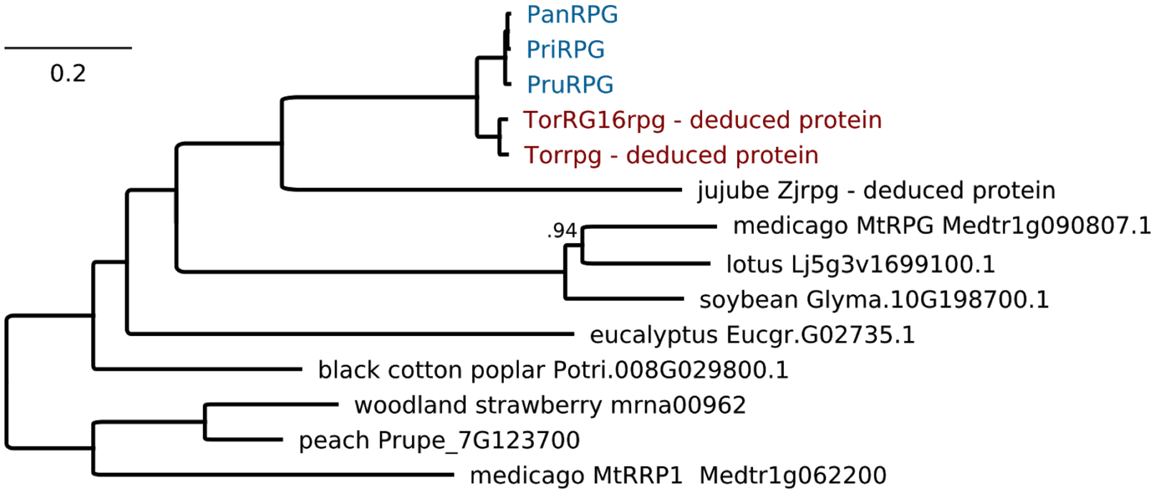

0.2
PanRPG
PriRPG
PruRPG
TorRG16rpg - deduced protein
Torrpg - deduced protein
jujube Zjrpg - deduced protein
.94
medicago MtRPG Medtr1g090807.1
lotus Lj5g3v1699100.1
soybean Glyma.10G198700.1
eucalyptus Eucgr.G02735.1
black cotton poplar Potri.008G029800.1
woodland strawberry mrna00962
peach Prupe_7G123700
medicago MtRRP1 Medtr1g062200

Fig. S27

Phylogenetic reconstruction of the RPG orthogroup. Alignment of OG0014072 was supplemented with RPG homologs of additional species. Included species: *Parasponia andersonii* (Pan) *Parasponia rigida* (Pri); *Parasponia rugosa* (Pru) *Medicago truncatula* (Mt); *Lotus japonicus* (Lj); *Glycine max* (Glyma), *Populus trichocarpa* (Potri); *Eucalyptus grandis* (Eugr). *Trema orientalis* RG33 (Tor); *Trema. orientalis* RG16 (TorRG16). *Ziziphus jujube* (Zj). No other functional RPG proteins could be detected in Rosales species, including *Fragaria vesca Ziziphus Jujube, Malus Domestica, Morus Notabilis,* and *Humulus Lupulus (natsume.shinsuwase.v1.0)*. Outgroup: *M. truncatula* MtRRP1 (RPG RELATED PROTEIN 1, Medtr1g062200.1). Node numbers indicate posterior probabilities below 1, scale bar represents substitutions per site.

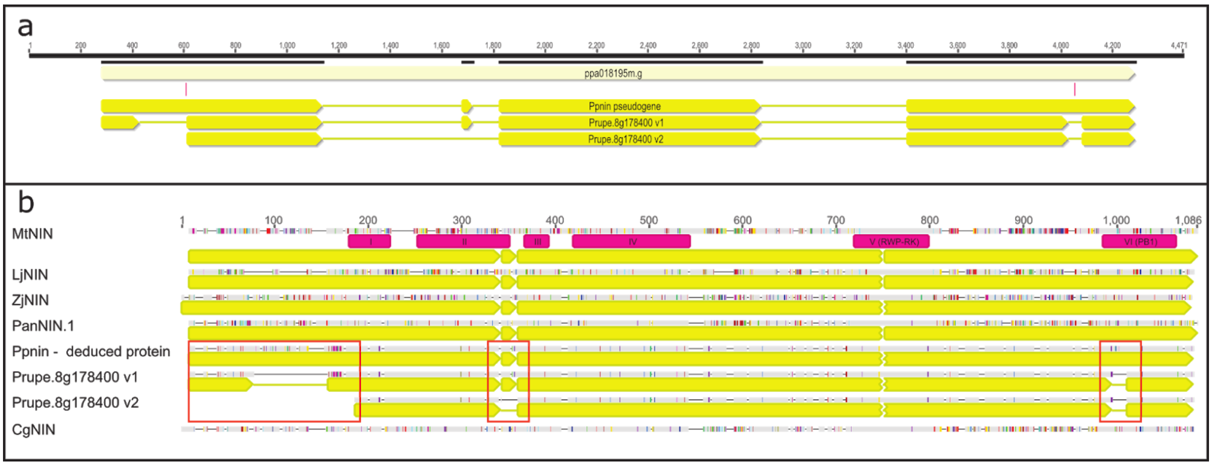


Fig. S28

Annotation of *Prunus persica* locus ppa018195m.g representing *PpNIN*. **a** Comparison of the exon-intron structure of two publicly released gene models (named Prupe.8g17800_v1 and Prupe.8g178400_v2) and the gene model used here (*Ppnin* pseudogene). Yellow arrows: exons. Red bars indicate 2 single-nucleotide insertions that affect the coding region of the *Ppnin* pseudogene. **b** Alignment of derived/deduced NIN proteins of 3 *Prunus persica* gene models Prupe.8g17800_v1, Prupe.8g178400_v2, and *Ppnin* pseudogene, with *Medicago truncatula* MtNIN, *Lotus japonicus* LjNIN, *Ziziphus jujube* ZjNIN, *Parasponia andersonii* PanNIN.1, and *Casuarina glauca* CgNIN. Six conserved domains are annotated in MtNIN (cyan). Exon structure for all *NIN* genes indicated in yellow (except CgNIN for which no gene sequence is available). Deviations in the three *Prunus persica* derived/deduced NIN proteins are marked in red boxes.

Supplementary tables

Table S1

Intergeneric crossings between *Parasponia* and *Trema* species.

Maternal parent	**Paternal parent**	**Result**
P. andersonii	*T. tomentosa*	positive
P. andersonii	*T. orientalis* RG33	negative
P. andersonii	*T. levigata*	negative
P. rigida	*T. tomentosa*	negative
P. rigida	*T. orientalis* RG33	negative
T. orientalis RG33	*P. andersonii*	negative

Result column indicates whether intergeneric crosses could be obtained (positive) or not (negative).

Table S2

Parasponia-Trema germplasm collection.

Species	Accession	Chromosome number	Estimated genome size		Origin	NCBI bioproject
			flow cytometry	k-mers		
P. andersonii	PanWU01x14	2n=2x=20	551	563 (536-591)	Papua New Guinea	PRJNA272473
P. rigida	PriWU20x00	n.d.	521	573 (549-600)	Papua New Guinea	PRJNA272486
P. rugosa	PruLW88x56	n.d.	n.d.	498 (478-520)	Philippines	PRJNA272880
T. orientalis	RG6	n.d.	1,931	n.d.	Malaysia: Borneo	n.a.
T. orientalis	RG16	n.d.	483	488 (458-521)	Malaysia: Borneo	PRJNA272878
T. orientalis	RG19	n.d.	488	629 (560-717)	Malaysia: Borneo	n.a.
T. orientalis	RG21	n.d.	677	n.d.	Malaysia: Borneo	n.a.
T. orientalis	RG25	n.d.	508	n.d.	Malaysia: Borneo	n.a.
T. orientalis	RG26	n.d.	296	n.d.	Malaysia: Borneo	n.a.
T. orientalis	RG32	n.d.	593	n.d.	Malaysia: Borneo	n.a.
T. orientalis	RG33	n.d.	501	506 (472-545)	Malaysia: Borneo	PRJNA272482
T. orientalis	RG35	n.d.	593	n.d.	Malaysia: Borneo	n.a.
T. orientalis	RG43	n.d.	2,286	n.d.	Malaysia: Borneo	n.a.
T. orientalis	RG44	n.d.	593	n.d.	Malaysia: Borneo	n.a.
T. orientalis	RG45	n.d.	508	n.d.	Malaysia: Borneo	n.a.
T. orientalis	RG46	n.d.	423	n.d.	Malaysia: Borneo	n.a.
T. orientalis	RG47	n.d.	1,016	n.d.	Malaysia: Borneo	n.a.
T. orientalis	RG48	n.d.	508	n.d.	Malaysia: Borneo	n.a.
T. orientalis	RG51	n.d.	1,524	n.d.	Malaysia: Borneo	n.a.
T. orientalis	WU30	n.d.	n.d.	n.d.	Australia	n.a.
T. levigata	TleWU50x00	n.d.	271	375 (363-388)	China	PRJNA38059
T. tomentosa	TtoWU10x00	2n=4x=40	910	997 (919-1,090)	Australia	PRJNA388567

Note that we consider estimated genome sizes based on flow cytometry as less reliable than those based on k-mer analysis. n.d. = not determined, n.a. = not available.

Table S3

Genome size estimations based on estimated genome coverage.

Genome	Total kmers	Error-free kmers	Error threshold	Peak multiplicity	Estimated coverage	Estimated genome size (Mb)
P. andersonii	11,991,734,213	11,588,276,008	5	21	20.6	563
P. rigida	13,365,390,696	13,070,993,890	5	23	22.8	573
P. rugosa	12,031,642,877	11,861,824,695	4	24	23.8	498
T. levigata	11,734,970,688	11,483,093,120	5	31	30.6	375
T. orientalis RG16	7,681,138,750	7,558,426,418	3	15	15.5	488
T. orientalis RG19	5,158,370,893	5,092,882,606	2	8	8.1	629
T. orientalis RG33	7,144,168,594	7,031,752,162	3	14	13.9	506
T. tomentosa	11,856,769,186	11,767,550,254	2	12	11.8	997

Table S4

Genome sequencing strategy.

Species	**Platform**	**Library type**	**Exp. insert size (bp)**	**Read length (bp)**	**SRA accession**	**Raw (Gb)**	**Clean (Gb)**	**Est. av. cov.**	**Mapped (%)**	**Properly paired (%)**	**Median insert size**
P. andersonii	MiSeq	pe	400	250	SRR5457665	7,287	6,470	11	97.52	82.34	456
P. andersonii	MiSeq	pe	400	250	SRR5457666	6,982	6,244	11	97.72	82.30	457
P. andersonii	MiSeq	pe	400	250	SRR5457667	3,603	3,208	6	97.56	82.34	458
P. andersonii	HiSeq	pe	500	100	SRR5457668	13,161	10,837	19	91.75	83.58	254
P. andersonii	HiSeq	pe	500	100	SRR5457669	4,430	3,324	6	90.54	83.61	246
P. andersonii	HiSeq	mp	3,000	100	SRR5457672	12,662	6,065	11	95.32	81.05	3,676
P. andersonii	HiSeq	mp	7,000	100	SRR5457671	17,207	7,886	14	95.45	79.08	6,186
P. andersonii	HiSeq	mp	10,000	100	SRR5457670	9,461	4,553	8	96.25	62.96	10,412
T. orientalis RG33	MiSeq	pe	400	250	SRR5464169	5,455	4,687	9	95.99	82.66	481
T. orientalis RG33	MiSeq	pe	400	250	SRR5464170	6,616	5,721	11	97.66	85.14	482
T. orientalis RG33	MiSeq	pe	400	250	SRR5464171	2,465	2,108	4	97.34	85.11	482
T. orientalis RG33	HiSeq	pe	500	100	SRR5464172	9,950	8,638	17	96.50	92.87	243
T. orientalis RG33	HiSeq	mp	3,000	100	SRR5464176	6,576	2,931	6	98.08	81.89	3,169
T. orientalis RG33	HiSeq	mp	3,000	100	SRR5464175	6,645	3,117	6	98.11	82.12	3,172
T. orientalis RG33	HiSeq	mp	7,000	100	SRR5464173	7,060	3,282	6	97.64	78.93	5,633
T. orientalis RG33	HiSeq	mp	7,000	100	SRR5464174	7,130	3,501	7	97.67	78.90	5,635
T. orientalis RG33	HiSeq	mp	10,000	100	SRR5464178	4,479	2,043	4	98.56	72.06	9,204
T. orientalis RG33	HiSeq	mp	10,000	100	SRR5464177	4,505	2,279	5	98.58	72.03	9,211
P. rigida	HiSeq	pe	500	100	SRR5626387	14,263	13,147	23	99.79	89.35	253
P. rigida	HiSeq	pe	500	100	SRR5626388	4,723	3,931	7	99.76	90.81	245
P. rugosa	HiSeq	pe	500	100	SRR5631337	16,141	15,726	32	99.95	83.60	480
T. orientalis RG16	HiSeq	pe	500	100	SRR5674478	11,009	9,862	20	99.89	88.52	243
T. levigata	HiSeq	pe	500	100	SRR5631617	16,119	15,483	41	99.91	86.85	467
T. tomentosa	HiSeq	pe	500	100	SRR5679053	8,130	8,018	8	99.51	68.59	481
T. tomentosa	HiSeq	pe	500	100	SRR5679052	7,678	7,574	8	99.50	67.89	481

pe = paired end, mp = mate pair, Exp. = expected, Est. av. cov. = estimated average coverage.

Table S5

Assembly results of *Parasponia* and *Trema* genome sequences.

		P. andersonii	***P. rigida***	***P. rugosa***	***T. orientalis* RG33**	***T. orientalis* RG16**	***T. levigata***
Assembly size (bp)		475,834,238	473,976,781	440,853,656	387,958,189	391,454,457	350,420,318
# N		24,563,530	14,507	340,003	18,061,089	12,092	501,912
# scaffolds		2,732	213,690	199,834	2,756	182,478	190,006
N50 length (bp)		712,846	12,913	12,790	656,203	8,006	14,367
N50 index		193	9,257	9,637	157	12,384	6,875
N95 length (bp)		101,596	464	397	76,336	461	320
N95 index		828	116,579	87,938	809	110,056	73,056
Min seq length (bp)		501	100	100	501	100	100
Max seq length (bp)		4,735,547	173,338	128,802	5,284,217	94,643	117,780
GC (%)		34.4	34.3	34.2	33.3	33.3	33.2
BUSCO	%complete	95.42	92.01	90.69	95.00	87.71	89.44
	%fragmented	0.83	3.19	3.96	1.39	6.46	5.21
	%missing	3.75	4.79	5.35	3.61	5.83	5.35
CEGMA	%complete	91.1	87.1	87.5	90.7	85.5	88.7
	%partial	97.6	97.2	97.2	97.2	96.4	97.6

For BUSCO a set of 1,440 plant-specific genes was used. #N = number of gap sequences, GC% = guanine-cytosine content.

Table S6

Gene models in *Parasponia andersonii* and *Trema orientalis* RG33 reference genomes.

	P. andersonii	***T. orientalis* RG33**
Total gene models	35,761	35,213
in orthogroups	29,393	29,958
orthologous pairs	25,605	25,605
inparalogs	1,544	1,704
singletons	1,901	2,401
multi-orthologs	343	248
total CNVs	3,445	4,105
BUSCO complete	93.30%	92.80%
BUSCO fragmented	3.80%	4.10%
BUSCO missing	2.90%	3.10%

Inparalogs = species specific duplications, singletons = loss of gene in other species, multi-orthologs = duplication in the other species, CNVs = copy number variants. We found no significant enrichment of total CNVs in the symbiosis genes (hypergeometric test, p = 0.99). For BUSCO a set of 1,440 plant specific genes was used.

Table S7.

Copy number variants in symbiosis genes that are consistent between *Parasponia* and *Trema* genera.

Name	Gene ID	CNV type	Class	Description
PanNFP2	PanWU01x14_320250	loss in *Trema*	LS,NE	LysM domain containing receptor kinase, putative rhizobium LCO receptor
PanCRK11	PanWU01x14_285030	loss in *Trema*	NE	Cysteine rich receptor like kinase
PanLEK1	PanWU01x14_069780	loss in *Trema*	NE	Concanavalin A-like lectin receptor kinase
PanNIN	PanWU01x14_111140	loss in *Trema*	LS,CR	Ortholog of transcription factor NODULE INCEPTION
PanRPG	PanWU01x14_272380	loss in *Trema*	LS,CR	Ortholog of long coiled-coil protein RHIZOBIUM-DIRECTED POLAR GROWTH
PanDEF1	PanWU01x14_187760	loss in *Trema*	NE	Defensin-like protein
PanGAT	PanWU01x14_150960	loss in *Trema*	NE	Gamma-aminobutyric acid (GABA) transporter
PanHCT1	PanWU01x14_046570	duplication in *Parasponia*	NE	Hydroxycinnamoyl-CoA shikimate / Quinate hydroxycinnamoyl transferase
TorEPR	TorRG33x02_052550	loss in *Parasponia*	LS	LysM domain containing receptor kinase, putative rhizobium exopolysaccharide receptor
TorN19L3	TorRG33x02_066920	loss in *Parasponia*	LS	NODULIN19-like protein
TorIPT4	TorRG33x02_307000	loss in *Parasponia*	LS	Isopentenyltransferase

Gene ID corresponds to that in *P. andersonii*, or *T. orientalis* in case of gene loss in *Parasponia* species. LS: putative ortholog of legume genes that function in symbiosis (Dataset S1), NE: nodule enhanced expression in *P. andersonii* (Dataset S4), CR: genes that are commonly utilized in *P. andersonii* and medicago (Dataset S5). Expression profiles of the *P. andersonii* genes are depicted in Fig. 4.

Table S8

Sequenced RNA samples.

Species	Description	Used in annotation	SRA accession	Mapped (%)	Properly paired (%)
P. andersonii	leaves	yes	SRR5631161	96	93.5
P. andersonii	stems	yes	SRR5631160	95.7	93.3
P. andersonii	female flowers	yes	SRR5631163	95.6	93
P. andersonii	male flowers	yes	SRR5631162	95.1	92.2
P. andersonii	young berries	yes	SRR5631157	95.5	92.7
P. andersonii	uninoculated root	no	SRR5631165	91.44	88.34
P. andersonii	uninoculated root	no	SRR5631166	91.72	88.36
P. andersonii	uninoculated root	no	SRR5631164	91.32	88.26
P. andersonii	nodule stage 1 M. plurifarium BOR2	no	SRR5631152	86.45	83.78
P. andersonii	nodule stage 1 M. plurifarium BOR2	no	SRR5631151	79.63	76.96
P. andersonii	nodule stage 1 M. plurifarium BOR2	no	SRR5631150	90.7	87.78
P. andersonii	nodule stage 2 M. plurifarium BOR2	no	SRR5631149	93.01	90
P. andersonii	nodule stage 2 M. plurifarium BOR2	no	SRR5631148	93.9	90.61
P. andersonii	nodule stage 2 M. plurifarium BOR2	no	SRR5631147	91.4	88.48
P. andersonii	nodule stage 3 M. plurifarium BOR2	no	SRR5631146	92.31	88.79
P. andersonii	nodule stage 3 M. plurifarium BOR2	no	SRR5631145	75.34	66.26
P. andersonii	nodule stage 3 M. plurifarium BOR2	no	SRR5631144	85.65	80.71
P. andersonii	lateral root primordia	yes	SRR5631155	100	87.83
P. andersonii	lateral root primordia	yes	SRR5631154	100	88.38
P. andersonii	lateral root primordia	yes	SRR5631153	100	77.92
P. andersonii	2 cm of root tip.	yes	SRR5631156	100	73.97
P. andersonii	2 cm of root tip.	yes	SRR5631159	100	81.04
P. andersonii	2 cm of root tip.	yes	SRR5631158	100	75.5
T. orientalis RG33	lateral root primordia	yes	SRR5681931	100	74.77
T. orientalis RG33	lateral root primordia	yes	SRR5681932	100	75.46
T. orientalis RG33	2 cm of root tip.	yes	SRR5681933	100	83.58
T. orientalis RG33	2 cm of root tip.	yes	SRR5681934	100	84.8
T. orientalis RG33	2 cm of root tip.	yes	SRR5681930	100	77.61
Hybrid	uninoculated hybrid root	no	SRR5641455	84.54	78.97
Hybrid	uninoculated hybrid root	no	SRR5641456	83.47	77.73
Hybrid	uninoculated hybrid root	no	SRR5641457	84.7	78.63
Hybrid	hybrid nodule M. plurifarium BOR2	no	SRR5641458	86.8	81.28
Hybrid	hybrid nodule M. plurifarium BOR2	no	SRR5641459	76.45	70.65
Hybrid	hybrid nodule M. plurifarium BOR2	no	SRR5641460	75.94	70.29

Supplementary methods

Seed germination

Parasponia and *Trema* seeds were surface sterilized in 4% sodium hypochlorite containing 0.02% (v:v) Tween20, and rinsed thoroughly with sterile water. Sterilized seeds were subjected to 6 temperature cycles (11h 28°C, 6h 4°C) to induce germination. Germinating seeds were transferred to sterile 1.0% agar plates and grown at 28°C with a photoperiod of 16 day and 8h night.

Arbuscular mycorrhization assay

Two week old seedlings were transferred to 800 ml Sand:Granule:*Rhizophagus irregularis* (*Rir,* INOQ TOP- INOQ GmbH, Schnega Germany) inoculum mixture (1:1:0.01), irrigated with 80 ml ½ strength modified Hoagland solution containing 20 μM K_2HPO_4 (1) and grown for an additional 6 weeks at 28°C, under a photoperiod of 16/8h (day/night). 50 ml additional nutrient solution was provided once a week. Mycorrhization efficiency was analysed as previously described (2) for three aspects: 1) frequency of fungal colonization in 1 cm root segments; 2) average level of mycorrhization in all root fragments, and 3) arbuscular abundance in all root fragments (Additional file 2: Fig. S1). Arbuscules were WGA-Alexafluor 488-stained and imaged according to Huisman *et al* 2015 (3).

Nodulation assay

All nodulation assays were conducted with *Mesorhizobium plurifarium* BOR2. This strain was isolated from *P. andersonii* root nodules grown in soil samples collected from the root rhizosphere of *Trema orientalis* plants in Malaysian Borneo, province of Sabah (4). *M. plurifarium* was grown on yeast extract mannitol medium at 28°C (5). Plants were grown in sterile plastic 1 liter pots containing perlite and EKM medium supplemented with 0.375 mM NH_4NO_3 and rhizobium (OD600:0.05) (6). Nodule number per plant was quantified 6 weeks post inoculation.

To isolate *P. andersonii* nodules at 3 developmental stages nodules were separated based on morphology and size. Stage 1: nodules are round and < 1mm in diameter in size. The outer cell layers of stage 1 nodules are transparent. Light microscopy confirmed that at this stage, rhizobia already reach the central part of the nodule, but are mainly present in the apoplast (Fig. 1h). Stage 2: nodules are brownish, and ~2 mm in size. Nodules have formed an apical meristem and 2-3 cell layers have been infected by rhizobia (Fig. 1i). Stage 3:

nodules are pinkish on the outside due to an accumulation of hemoglobin and > 2 mm in size. Light microscopy showed that stage 3 nodules contain zones of fully infected cells (Fig. 1j). For each of these stages, three biological replicates were used for RNA sequencing.

Acetylene reduction assays (ARA)

Acetylene reduction assays (7) were conducted on nodules harvested 6 weeks post inoculation with *Mesorhizobium plurifarium* strain BOR2. Nodules were sampled per plant and collected in 15 ml headspace vials with screw lids. 2.5 ml of acetylene was injected into the vial and incubated for about 10 minutes, after which 1 ml headspace was used to quantify ethylene nitrogenase activity using an ETD 300 detector (Sensor Sense, Nijmegen, The Netherlands; Isogen, Wageningen, The Netherlands) (8).

Microscopy

Tissue fixation and embedding were done as described by Fedorova et al. 1999 (9). Semi-thin (0.6 µm) sections were cut using a Leica Ultracut microtome and examined by Leica FL light microscope. Electron microscopy analysis was performed using a JEOL JEM 2100 transmission electron microscope equipped with a Gatan US4000 4K×4K camera.

Genomic DNA isolation for sequencing

For comparative genomic analyses, DNA samples were isolated from three *Parasponia* accessions and three *Trema* accessions. Two grams of young leaves were ground in liquid nitrogen, and the still frozen powder resuspend in 50 ml of cold Nuclei Purification Buffer (NPB: 20 mM MOPS pH7, 40 mM NaCl, 90 mM KCl, 0.5 mM EGTA, 2 mM EDTA, 0.5 mM Spermidine3HCl and 0.2 mM Spermin4HCl). The suspension was filtered through a layer of Miracloth, and a 70 µm cell strainer, and subsequently centrifuged at 1,500 g for 10 min, 4°C. The pellet was resuspended in 15 ml of cold NPB + 0.9% (w/w) Triton X100, and incubated on ice for 15 min to allow chloroplasts to denature. The suspension was centrifuged at 1,000 g for 10 min, 4°C to collect nuclei. The pellet was resuspended in 1ml of 65°C Nuclei Lysis Buffer (2% CTAB, 1.4M NaCl, 100mM TrisHCl pH8, 5mM EDTA) supplemented with 100 µg RNAse A and incubated for 30 min. at 65°C. Finally, the DNA was cleaned by phenol and chloroform extractions and ethanol precipitation.

For phylogenetic reconstruction DNA was isolated from additional specimens (Additional file 9: Table S12). In case of fresh material the protocol above was used; in case of dried material, 50 mg of dried leaf material was ground in liquid nitrogen, and the still frozen powder was resuspended in 900 ul CTAB buffer (100mM Tris pH 8.0, 1.4M NaCl, 20mM

EDTA pH8.0, 2% CTAB, 1.2% β-mercaptoethanol), and heated for 1 hr at 55°C, with every 15 minutes vortexing for a few seconds. This mixture was two times extracted with (1 volume and half a volume, respectively) chloroform/isoamylalcohol (24/1). The nucleic acids were precipitated by addition of 0.6 volume of cold isopropanol, followed by overnight storage at -20°C. After centrifugation, the pellet was washed with 70% ethanol, air-dried and overnight left in 0.5mM Tris pH8.5/50µM EDTA pH8.0 at 8°C to dissolve. DNA concentration was measured using a Qubit Fluorometer. 1µg of DNA was diluted to 100µl in 10mM Tris pH8.0/0.1mM EDTA pH8.0, 10µg RNase A was added, followed by 10 minutes incubation at 37°C. The DNA from this mixture was cleaned-up using a NucleoSpin gDNA Clean-up XS kit (Machery-Nagel, 52355 Duerren, Germany).

RNA isolation for sequencing

RNA samples from various tissues and nodulation stages were isolated from *P. andersonii* and *Trema orientalis* RG33 (Table S8). 10 to 50 mg tissue was ground in liquid nitrogen, transferred in 0.7 ml 65°C extraction buffer (2% CTAB, 2.5% PVP-40 in 2M NaCl, 25mM EDTA, 100mM TrisHCl pH 8) and incubated at 65°C for 10 min. 75 µl 3M NaAc pH5.6 was added to the suspension, which was subsequently extracted with phenol-chloroform and chloroform. The RNA was precipitated by addition of 2/3 volume of isopropanol and 10 µg glycogen as a carrier. The pellet was dissolved in 100µl MilliQ water and treated with DNase according to the manufacturer's protocol (Qiagen RNeasy handbook, Appendix E). To remove the DNAse phenol and chloroform extractions were performed followed by an ethanol precipitation. Library preparation and RNA sequencing was conducted by B.G.I. (Shenzhen, China).

DNA Library preparation and sequencing

Paired-end Illumina genomic DNA libraries (insert size 500bp, 100bp reads) were prepared for all accessions (Table S4). Mate-pair libraries (3Kb, 7Kb, and 10Kb) and overlapping fragment libraries (450bp insert size, 250bp reads) were prepared for the reference accessions (*P. andersonii* accession WU01 and *T. orientalis* accession RG33). Paired-end and mate-pair libraries were sequenced on an Illumina HiSeq2000, overlapping libraries were sequenced on an Illumina MiSeq. For the *P. andersonii* and *T. orientalis* reference genomes, a total of 75Gb (~132x genome coverage) and 61Gb (~121x coverage) of data was produced respectively. The other accessions were sequenced at an average coverage of ~30X.

Illumina libraries were prepared: 500bp insert size paired-end libraries for all accessions, mate-pair libraries (3Kb, 7Kb, and 10Kb) and overlapping fragment libraries (450bp insert size, 250bp reads) for the reference accessions (*P. andersonii* accession WU01 and *T. orientalis* accession RG33). Paired-end and mate-pair libraries were sequenced on an Illumina HiSeq2000 instrument using 101, 7, 101 flow cycles for forward, index and reverse reads, respectively. DNA Mate Pair libraries for *P. andersonii* and *T. orientalis* RG33 were made according to Nextera Mate Pair sample preparation Guide (Illumina) with few adaptations. Approximately 4 µg DNA was used for tagmentation in a 400 µL volume at 55°C for 30 minutes. Tagmented DNA was purified using a Zymoclean purification column and eluted in 30 µL elution buffer. Strand displacement of tagmented DNA was done for 30 minutes at room temperature. DNA was then purified using AmpureXP beads (Agencourt). Yield and fragment size were analyzed using Qubit fluorescence quantification (Thermo Fisher Scientific) and Bioanalyzer12000 DNA chip (Agilent technologies) respectively. Approximately 750 ng tagmented repaired DNA was loaded on a 0.6% Megabase agarose gel (Bio-rad) with SYBR safe (Thermo Fisher Scientific) staining. After electrophoresis for 3 hours at 100 Volt, a clear smear of 2 to 15 kb DNA fragments was visible.. Of both *Trema* and *Parasponia* tagmented DNA three fractions were isolated from gel, i.e. 3 to 6 kb, 7 to 9 kb and 10 to 15 kb. DNA was recovered from gel slices using Large fragment DNA recovery kit (Zymo) followed by DNA circularization for 18 hours at 30°C, exonuclease treatment at 37°C for 1 hour and inactivation at 70°C both for 30 minutes using a water bath. Remaining circularized DNA molecules were sheared using a Covaris E210 focused ultrasonicator to approximately 500 bp target size. Sheared fragments containing a biotinylated circularization adapter were enriched using M280 streptavidin Dynabeads (Thermo Fisher Scientific), followed by standard end repair, A-tailing and barcoded adapter ligation according to manufacturer's protocol (Illumina), all incubation steps using a 2720 thermocycler (Thermo Fisher Scientific). Adapter-ligated fragments were then amplified using 15 PCR cycles, purified twice with ampureXP beads and re-suspend in 20 µL elution buffer. Final libraries were quantified by Qubit and Bioanalyzer High Sensitivity DNA assay.

In addition, overlapping libraries were prepared. Approximately 750 ng was sheared in a 120µL volume using a Covaris E210 device for 450bp target fragment size. End repair, A-tailing, barcoded Adapter ligation and PCR library amplification were all performed according to Illumina TruSeq LT DNA gel free library prep guidelines. Adapter Ligated DNA fragments were purified using AmpureXP with 20 µL elution buffer and size selected using two slots on a 2% Agarose dye free gel (Blue Pippin, Sage Science). Size selection of *Parasponia* DNA was done using a tight selection protocol (575 bp target), whereas *Trema*

DNA size selection was done using a narrow range selection protocol (525 to 615 bp range). Size-selected fragments were Ampure XP purified and finally eluted in 20 µL Elution Buffer TE. Libraries were quantified by Qubit fluorescence and library fragment size was analyzed by Bioanalyzer High Sensitivity DNA assay.

The barcoded overlapping libraries of *Trema* and *Parasponia* were sequenced on a MiSeq instrument (2*251 cycles for Paired End sequencing plus 7 cycles for the indexing reads). For the *Parasponia* and *Trema* reference genomes, a total of 75Gb (~132x genome coverage) and 61Gb (~121x coverage) of data was produced respectively. The barcoded Mate Pair libraries were loaded for DNA clustering on two lanes of an Illumina Paired-End flow cell using a cBot. Sequencing was done on an HiSeq2000 instrument using 101, 7, 101 flow cycles for forward, index and reverse reads. The barcoded overlapping libraries of *Trema* and *Parasponia* were equimolar pooled for clustering and sequencing on an illumina MiSeq instrument using six MiSeq V3 flow-cells and 2*300 cycles for Paired End sequencing plus 7 cycles for the indexing reads. De-multiplexing of resulting data was carried out using Casava 1.8 software. A list of DNA samples is given in Additional file 3: Table S5.

Estimation of heterozygosity levels and genome size

To generate preliminary estimates of genome size of available germplasm we used flow cytometry generally as described previously (10, 11). In short; nuclei were isolated from leaf tissue ground in liquid nitrogen and suspended in 10 ml ice-cold Nuclear preparation buffer (NPB; 20 mM MOPS pH 7 40 mM NaCl, 90 mM KCl, 0.5 mM EGTA, 2 mM EDTA, 0.5 mM Spermidine, 0.2 mM Spermine and Complete Protease Inhibitor cocktail of Sigma-Aldrigh). The suspension was filtered through a 70 µM cell strainer and spun 10 min at 1,000 g at 4°C. The pellet was resuspended in 1 ml NPB complemented with 0.1% Triton X-100, transfer to 1.5 ml LoBind Eppendorf tube, and spun at 1,000 g for 10 min at 4 °C. Subsequently, the pellet containing nuclei were resuspended in NPB buffer containing 50 ug/ml propidium iodide (PI). These were analysed FACSAria II (Becton Dickenson) using a 488 nm laser with the following modifications: Biparametric contour plots of FL1-A (616/23 nm band-pass filter) versus FL2-A (530/30 nm band-pass filter) were generated followed by gating the nuclei derived PI signals and representing these in a uniparametric histogram of FL1-A fluorescence that was used to estimate genome size. To calibrate the flow cytometry results *Medicago truncatula* cv. Jemalong A17 (~500 Mb) was included as a standard genome.

To assess levels of heterozygosity and more accurately estimate genome size we performed k-mer analyses based on medium-coverage sequence data. Multiplicities of 21-mers were extracted from the reads using Jellyfish (version 2.2.0) (12) and processed using custom R scripts. First, a multiplicity threshold was determined below which most k-mers are considered to represent sequencing errors and which were excluded from further analysis. In principle, errors occur randomly and this generates a high-frequency peak at multiplicity 1 after which frequency decreases and subsequently increases due to a broad frequency peak around the mean genome coverage. The error multiplicity threshold was therefore set at the multiplicity with the lowest frequency between these two peaks. Next, we identified the peak multiplicity as the one with the highest frequency. Homozygous genome coverage was estimated by scaling the peak multiplicity proportional to the difference of its frequency with that of multiplicities one below and above. Heterozygous coverage was defined as half that of the homozygous coverage (Additional file 2: Fig. S5). Finally, genome size was calculated as the total number of error-free k-mers divided by the estimated homozygous genome coverage (Additional file 3: Table S4). These estimates are generally comparable to those based on our flow cytometry measurements (Additional file 3: Table S3) except for genomes that differed much from the standard genome used to calibrate measurements. This inconsistency is probably due to the use of a single standard to calibrate FACS measurements. We therefore considered the quantitative genome estimates based on k-mer analysis a more accurate estimation of genome size.

Characterization of repetitive sequences

Repetitive sequences are inherently difficult to assemble. We therefore characterized and quantified repetitive element using the ab initio graph-based clustering approach implemented in RepeatExplorer (13). Analyses were based on random subsamples of 20,000 paired-end reads and included a reclustering step where clusters with shared mate pairs are merged (threshold k=0.2). Repeat classification was based on the RepeatExplorer Viridiplantae dataset and on plant organellar sequences. Relative sizes of repetitive sequences in the genome were scaled by the genome size estimations based on k-mer analysis to generate absolute sizes in Mb (Fig. S9).

Assembly of reference genomes

Raw sequencing data were preprocessed as follows: first, adapters (standard and junction) were removed and reads were trimmed using fastq-mcf (version 1.04.676) (14). Minimum remaining sequence length was set to 50 for HiSeq data and 230 for MiSeq data. Duplicates were removed using FastUniq (version 1.1) (14, 15). Chloroplast and mitochondrial genomes were assembled first with IOGA (version 1) using reference sets of plant chloroplast and mitochondrial genomes (16). Chloroplast and mitochondrial reads were identified and separated from the nuclear reads by mapping to four organellar assemblies (*Parasponia andersonii*, *Trema orientalis*, *Morus indica*, *Malus x domestica*) using BWA-MEM (version 0.7.10) (17). (2.5% of reads were filtered in *P. andersonii* and 1.7% in *T. orientalis* RG33). Finally, a contamination database was produced by BLASTing preliminary in-house *Parasponia andersonii* and *Trema orientalis* draft genome assemblies against NCBI's nt database. Hits outside the plant kingdom were selected using a custom script and corresponding sequences were downloaded from GenBank. The contamination database was then supplemented with plant virus sequences (http://www.dpvweb.net/seqs/allplantfasta.zip). Genomics reads were filtered by mapping to this contamination database (0.1% reads were filtered in *P. andersonii* and 0.2% in *T. orientalis* RG33).

The preprocessed data were *de novo* assembled using ALLPATHS-LG (release 48961) (18). Relevant parameters were PLOIDY=2 and GENOME_SIZE=600000000. The assemblies were performed on the Breed4Food High-Performance Cluster from Wageningen UR (http://breed4food.com).

Remaining contamination in the ALLPATHS-LG assembly was identified by blasting the assembled contigs to their respective chloroplast and mitochondrial genomes, the NCBI nr and univec databases (Downloaded 29 October 2014) and by mapping back genomic reads of the HiSeq 500bp insert size library. Regions were removed if they matched all of the following criteria: (1) significant blast hits with more than 98% identity (for the nr database only blast results that were not plant-derived were selected); (2) read coverage lower than 2 or higher than 50 (average coverage for the HiSeq 500bp insert size library is ~30x); (3) number of properly paired reads lower than 2.

Resulting contigs were subsequently scaffolded with two rounds of SSPACE-standard (v3.0) (19) with the mate-pair libraries using default settings. In order to use reads mapped with BWA-MEM (v0.7.10) we used a python script written by Peter Cock to convert sam files to tabular format (available on Github). We used the output of the second run of SSPACE scaffolding as the final assembly.

Assembly of *Parasponia* and *Trema* draft genomes

To assess whether gene copy number variants of interest are also present in other, non-reference *Parasponia* and *Trema* genomes, we assembled genomic sequences of *P. rigida*, *P. rugosa*, *T. levigata*, and *T. orientalis* accession RG16 based on the medium-coverage sequence data that was also used for *k*-mer analysis (Table S4, S5). Assembly was performed with the iterative de Bruijn graph assembler IDBA-UD version 1.1.1 (20), iterating from 30-mers (assembling low-coverage regions) to 120-mers (accurately assembling regions of high coverage), with incremental steps of 20. Genes of interest were manually annotated and putatively lost genes or gene fragments were confirmed based on (I.) genomic alignments performed with the EMBOSS 6.5.7 tool (21) dotmatcher as implemented in the Geneious function dotplot (Fig. S23), and (II.) mapping the medium-coverage reads of *P. rigida* and *P. rugosa* to the *T. orientalis* RG33 reference genome and *T. orientalis* RG16 and *T. levigata* to the *P. andersonii* reference genome (Fig. S18,21-22). Mapping was done with BWA-MEM version 0.7.10 (17) with default parameters and unfiltered reads. Mappings covered on average ~95-96% of all exons comprising CDS with an average coverage of ~27x for *P. rigida*, ~34x for *P. rugosa*, ~34x for *T. levigata*, and ~20x for *T. orientalis* RG16 as determined using the samtools bedcov function (22).

Phylogenetic reconstructions

Multiple sequence alignments were generated using MAFFT version 7.017 (23). Phylogenetic analyses of Cannabaceae were based on nucleotide alignments and performed using MrBayes version 3.2.2 (24). The first phylogenetic reconstruction of Cannabaceae was based on four markers comprising data from Yang et al. 2013 (25) supplemented with new data generated with primers and protocols published therein (Dataset S7) with five optimal partitions and models of sequence evolution: atpB-rbcL combined with trnL-F (GTR+I+G); first codon position of rbcL (GTR+I+G); second position of rbcL (SYM+I+G); third position of rbcL (GTR+G); rps16 (GTR+G). The second phylogenetic reconstruction of Cannabaceae was based on whole chloroplast genomes (Dataset S7) with eight optimal partitions and models of sequence evolution: tRNA sequence (HKY+I), rRNA sequence (GTR+I), long single copy region (LSC) coding sequence (GTR+I+G), LSC non-coding sequence (GTR+G), short single copy region (SSC) coding sequence (GTR+G), SSC non-coding sequence (GTR+G), inverted repeat region (IR) coding sequence (GTR+G), and IR non-coding sequence (GTR+G).

Phylogenetic analyses of genes were based on amino acid sequences. Analyses of orthogroups comprising legume symbiosis genes (146 orthogroups) and nodule-enhanced genes (414 orthogroups) were performed using RAxML version 8.2.11 setting gamma-distributed rate variation, estimating optimal models of amino acid sequence evolution (PROTGAMMAAUTO), and running 100 fast bootstrap replicates to assess clade support. Analyses of HB, *NFP*, *HCT*, *EPR*, *NIN*, and *RPG* genes were performed using MrBayes version 3.2.6 running 2.2 million generations, setting gamma-distributed rate variation and integrating over different models of amino acid sequence evolution (aamodelpr=mixed). For NFP analyses was based on the kinase domain only, because based on the full-length sequences the relationships between the NFP1 and NFP2 paralogs remained unresolved.

References

1. Hoagland DR, Arnon DI (1950) The water-culture method for growing plants without soil. *Circular California Agricultural Experiment Station* 347:1–32.

2. Trouvelot A, Kough JL, Gianinazzi-Pearson V (1986) Mesure de taux de mycorhization VA d'un système radiculaire. Recherche de méthodes d'estimation ayant une signification fonctionnelle. *Physiological and Genetic Aspects of Mycorrhizae*, ed Gianinazzi-Pearson GS V (INRA Press, Paris), pp 217–221.

3. Huisman R, et al. (2015) Haustorium formation in *Medicago truncatula* roots infected by *Phytophthora palmivora* does not involve the common endosymbiotic program shared by arbuscular mycorrhizal fungi and rhizobia. *Mol Plant Microbe Interact* 28(12):1271–1280.

4. Merckx VSFT, et al. (2015) Evolution of endemism on a young tropical mountain. *Nature* 524:347–350.

5. Op den Camp RHM, et al. (2012) Nonlegume *Parasponia andersonii* deploys a broad rhizobium host range strategy resulting in largely variable symbiotic effectiveness. *Mol Plant Microbe Interact* 25(7):954–963.

6. Becking JH (1983) The *Parasponia parviflora—Rhizobium* symbiosis. Host specificity, growth and nitrogen fixation under various conditions. *Plant Soil* 75(3):309–342.

7. Bergersen FJ (1970) The quantitative relationship between nitrogen fixation and the acetylene-reduction assay. *Aust Jnl Of Bio Sci* 23(4):1015–1026.

8. Cristescu SM, Persijn ST, te Lintel Hekkert S, Harren FJM (2008) Laser-based systems for trace gas detection in life sciences. *Appl Phys B* 92(3):343–349.

9. Fedorova E, et al. (1999) Localization of H(+)-ATPases in soybean root nodules. *Planta* 209(1):25–32.

10. Hare EE, Johnston JS (2011) Genome size determination using flow cytometry of propidium iodide-stained nuclei. *Methods Mol Biol* 772:3–12.

11. Galbraith DW, Lambert GM (2012) High-throughput monitoring of plant nuclear DNA contents via flow cytometry. *Methods Mol Biol* 918:311–325.

12. Marçais G, Kingsford C (2011) A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* 27(6):764–770.

13. Novák P, Neumann P, Pech J, Steinhaisl J, Macas J (2013) RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. *Bioinformatics* 29(6):792–793.

14. Aronesty E (2013) Comparison of sequencing utility programs. *Open Bioinforma J* 7(1):1–8.

15. Xu H, et al. (2012) FastUniq: a fast de novo duplicates removal tool for paired short reads. *PLoS One* 7(12):e52249.

16. Bakker FT, et al. (2016) Herbarium genomics: plastome sequence assembly from a range of herbarium specimens using an Iterative Organelle Genome Assembly pipeline. *Biol J Linn Soc Lond* 117(1):33–43.

17. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* 25(14):1754–1760.

18. Gnerre S, et al. (2011) High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc Natl Acad Sci U S A* 108(4):1513–1518.

19. Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W (2011) Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* 27(4):578–579.

20. Peng Y, Leung HCM, Yiu SM, Chin FYL (2012) IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics*

28(11):1420–1428.

21. Rice P, Longden I, Bleasby A (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* 16(6):276–277.

22. Li H, et al. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25(16):2078–2079.

23. Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* 30(4):772–780.

24. Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19(12):1572–1574.

25. Yang M-Q, et al. (2013) Molecular phylogenetics and character evolution of Cannabaceae. *Taxon* 62(3):473–485.