

1 **Parallel loss of symbiosis genes in relatives of nitrogen-fixing non-**
2 **legume *Parasponia***

3

4 Robin van Velzen^{1†}, Rens Holmer^{1,2†}, Fengjiao Bu^{1‡}, Luuk Rutten^{1‡}, Arjan van Zeijl¹, Wei Liu³,
5 Luca Santuari^{1,4}, Qingqin Cao^{1,5}, Trupti Sharma¹, Defeng Shen¹, Yuda P. Roswanjaya¹, Titis
6 A.K. Wardhani¹, Maryam Seifi Kalhor¹ Joëlle Jansen¹, D. Johan van den Hoogen¹, Berivan
7 Güngör¹, Marijke Hartog¹, Jan Hontelez¹, Jan Verver¹, Wei-Cai Yang³, Elio Schijlen⁶, Rimi
8 Repin⁷, Menno Schilthuizen^{8,9}, M. Eric Schranz¹⁰, Renze Heidstra⁴, Kana Miyata¹, Elena
9 Fedorova¹, Wouter Kohlen¹, Ton Bisseling¹, Sandra Smit² & Rene Geurts^{1*}

10 **Affiliations:**

11 ¹Wageningen University, Department of Plant Science, Laboratory of Molecular Biology, Wageningen,
12 The Netherlands.

13 ²Wageningen University, Department of Plant Science, Bioinformatics Group, Wageningen, The
14 Netherlands.

15 ³Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing, China.

16 ⁴Department of Plant Science, Laboratory of Plant Developmental Biology, Wageningen, The
17 Netherlands.

18 ⁵Beijing University of Agriculture, College of Biological Science and Engineering & Beijing
19 Collaborative Innovation Center for Eco-Environmental Improvement with Forestry and Fruit trees,
20 Beijing, China.

21 ⁶Wageningen UR, Plant Research, BU Bioscience, Wageningen, The Netherlands.

22 ⁷Sabah Parks, Kota Kinabalu, Sabah, Malaysia.

23 ⁸Naturalis Biodiversity Center, Leiden, The Netherlands.

24 ⁹Institute for Tropical Biology and Conservation, Universiti Malaysia Sabah, Malaysia.

25 ¹⁰Wageningen University, Department of Plant Science, Laboratory of Biosystematics, Wageningen,
26 The Netherlands.

27 †,‡: These authors contributed equally

28 *Corresponding author: Rene Geurts, rene.geurts@wur.nl

29 **Abstract**

30

31 Rhizobium nitrogen-fixing nodules are a well-known trait of legumes, but nodules also occur
32 in other plant lineages either with rhizobium or the actinomycete *Frankia* as microsymbiont.
33 The widely accepted hypothesis is that nodulation evolved independently multiple times, with
34 only a few losses. However, insight in the evolutionary trajectory of nodulation is lacking. We
35 conducted comparative studies using *Parasponia* (Cannabaceae), the only non-legume able
36 to establish nitrogen fixing nodules with rhizobium. This revealed that *Parasponia* and
37 legumes utilize a large set of orthologous symbiosis genes. Comparing genomes of
38 *Parasponia* and its non-nodulating relative *Trema* did not reveal specific gene duplications
39 that could explain a recent gain of nodulation in *Parasponia*. Rather, *Trema* and other non-
40 nodulating species in the Order Rosales show evidence of pseudogenization or loss of key
41 symbiosis genes. This demonstrates that these species have lost the potential to nodulate.
42 This finding challenges a long-standing hypothesis on evolution of nitrogen-fixing symbioses,
43 and has profound implications for translational approaches aimed at engineering nitrogen-
44 fixing nodules in crop plants.

45 Introduction

46

47 Nitrogen sources such as nitrate or ammonia are key nutrients for plant growth, but their
48 availability is frequently limited. Some plant species in the related orders Fabales, Fagales,
49 Rosales and Cucurbitales -collectively known as the nitrogen fixation clade- can overcome
50 this limitation by establishing a nitrogen-fixing endosymbiosis with either *Frankia* or rhizobium
51 bacteria¹. These symbioses require specialized root organs, known as nodules, that provide
52 optimal physiological conditions for nitrogen fixation². For example, nodules of legumes
53 (Fabaceae, order Fabales) contain a high concentration of hemoglobin that is essential to
54 control oxygen homeostasis and protect the rhizobial nitrogenase enzyme complex from
55 oxidation^{2,3}. Legumes, such as soybean (*Glycine max*) and common bean (*Phaseolus*
56 *vulgaris*), represent the only crops that possess nitrogen-fixing nodules, and engineering this
57 trait in other crop plants is a long-term vision in sustainable agriculture^{4,5}.

58

59 Nodulating plants represent ~10 clades that diverged >100 million years ago and which are
60 nested in many non-nodulating lineages^{6,7}. Consequently, the widely accepted hypothesis is
61 that nodulation originated independently multiple times, preceded by a shared hypothetical
62 predisposition event in a common ancestor of the nitrogen fixation clade^{1,6-9}. Genetic
63 dissection of rhizobium symbiosis in two legume models -*Medicago truncatula* (medicago)
64 and *Lotus japonicus* (lotus)- has uncovered symbiosis genes that are essential for nodule
65 organogenesis, bacterial infection, and nitrogen fixation (Supplementary Table 1). These
66 include genes encoding LysM-type receptors that perceive rhizobial lipo-
67 chitoligosaccharides (LCOs) and transcriptionally activate the *NODULE INCEPTION (NIN)*
68 transcription factor¹⁰⁻¹⁵. Expression of *NIN* is essential and sufficient to set in motion nodule
69 organogenesis^{14,16-18}. Some symbiosis genes have been co-opted from the more ancient and
70 widespread arbuscular mycorrhizae symbiosis^{19,20}. However, causal genetic differences
71 between nodulating and non-nodulating species have not been identified²¹.

72

73 To obtain insight in the evolution of rhizobium symbiosis we conducted comparative studies
74 using *Parasponia* (Cannabaceae, order Rosales). The genus *Parasponia* is the only lineage
75 outside the legume family establishing a nodule symbiosis with rhizobium²²⁻²⁵. Similar as
76 shown for legumes, nodule formation in *Parasponia* is initiated by rhizobium-secreted LCOs
77 and this involves a close homolog of the legume LysM-type receptor NOD FACTOR
78 PERCEPTION / NOD FACTOR RECEPTOR 5 (NFP/NFR5)²⁶⁻²⁸. This suggests that
79 *Parasponia* and legumes utilize a similar set of genes to control nodulation. The genus
80 *Parasponia* represents a clade of five species that is phylogenetically embedded in the
81 closely related *Trema* genus²⁹. Like *Parasponia* and most other land plants, *Trema* species
82 can establish an arbuscular mycorrhizae symbiosis (Supplementary Fig. 1)., However, they
83 are non-responsive to rhizobium LCOs and do not form nodules^{25,28}. Taken together,
84 *Parasponia* is an excellent system for comparative studies with legumes and non-nodulating
85 *Trema* species to provide insights into the evolutionary trajectory of nitrogen-fixing root
86 nodules.

87

88 RESULTS

89

90 Nodule organogenesis is a dominant genetic trait

91

92 First, we took a genetics approach for understanding the rhizobium symbiosis trait of
93 *Parasponia* by making intergeneric crosses (Supplementary Table 2). Viable F₁ hybrid plants
94 were obtained only from the cross *Parasponia andersonii* (2n=20) x *Trema tomentosa*
95 (2n=4x=40) (Fig. 1a, Supplementary Fig. 2). These triploid hybrids (2n=3x=30) were infertile,
96 but could be propagated clonally. We noted that F₁ hybrid plants formed root nodules when
97 grown in potting soil, similar as earlier observations for *P. andersonii*³⁰. To further investigate
98 the nodulation phenotype of these hybrid plants, clonally propagated plants were inoculated

99 with two different strains; *Bradyrhizobium elkanii* strain WUR3³⁰ or *Mesorhizobium*
100 *plurifarium* strain BOR2. The latter strain was isolated from the rhizosphere of *Trema*
101 *orientalis* in Malaysian Borneo and showed to be an effective nodulator of *P. andersonii*
102 (Supplementary Fig. 3). Both strains induced nodules on F₁ hybrid plants (Fig. 1b,d,e;
103 Supplementary Fig. 4) but, as expected, not on *T. tomentosa*, nor on any other *Trema*
104 species investigated. Using an acetylene reduction assay we noted that, in contrast to *P.*
105 *andersonii* nodules, in F₁ hybrid nodules of plant H9 infected with *M. plurifarium* BOR2 there
106 is no nitrogenase activity (Fig. 1c). To further examine this discrepancy, we studied the
107 cytoarchitecture of these nodules. In *P. andersonii* nodules, apoplastic *M. plurifarium* BOR2
108 colonies infect cells to form so-called fixation threads (Fig. 1f,h-j), whereas in F₁ hybrid
109 nodules these colonies remain apoplastic, and fail to establish intracellular infections (Fig.
110 1g,k). To exclude the possibility that the lack of intracellular infection is caused by
111 heterozygosity of *P. andersonii* where only a nonfunctional allele was transmitted to the F₁
112 hybrid genotype, or by the particular rhizobium strain used for this experiment, we examined
113 five independent F₁ hybrid plants either inoculated with *M. plurifarium* BOR2 or *B. elkanii*
114 WUR3. This revealed a lack of intracellular infection structures in nodules of all F₁ hybrid
115 plants tested, irrespective which of both rhizobium strains was used (Fig. 1g,k,
116 Supplementary Fig. 4), confirming that heterozygosity of *P. andersonii* does not play a role in
117 the F₁ hybrid infection phenotype. These results suggest, at least partly, independent genetic
118 control of nodule organogenesis and rhizobium infection. Since F₁ hybrids are nodulated with
119 similar efficiency as *P. andersonii* (Fig. 1b), we conclude that the network controlling nodule
120 organogenesis is genetically dominant.

121

122 ***Parasponia* and *Trema* genomes are highly similar**

123

124 Based on preliminary genome size estimates using FACS measurements, three *Parasponia*
125 and five *Trema* species were selected for comparative genome analysis (Supplementary

126 Table 3). K-mer analysis of medium-coverage genome sequence data (~30x) revealed that
127 all genomes had low levels of heterozygosity, except those of *Trema levigata* and *T.*
128 *orientalis* (accession RG16) (Supplementary Fig. 5). Based on these k-mer data we also
129 generated more accurate estimates of genome sizes. Additionally, we used these data to
130 assemble chloroplast genomes based on which we obtained additional phylogenetic
131 evidence that *T. levigata* is sister to *Parasponia* (Fig. 1a, Supplementary Fig. 6-8). Graph-
132 based clustering of repetitive elements in the genomes (calibrated with the genome size
133 estimates based on k-mers) revealed that all selected species contain roughly 300 Mb of
134 non-repetitive sequence, and a variable repeat content that correlates with the estimated
135 genome size that ranges from 375 to 625 Mb (Fig. 2a, Supplementary table 4). Notably, we
136 found a *Parasponia*-specific expansion of *ogre/tat* LTR retrotransposons comprising 65 to 85
137 Mb (Fig. 2b). We then generated annotated reference genomes using high-coverage (~125X)
138 sequencing of *P. andersonii* (accession WU1)²⁷ and *T. orientalis* (accession RG33). These
139 species were selected based on their low heterozygosity levels in combination with relatively
140 small genomes. *T. tomentosa* was not used for a high-quality genome assembly because it
141 is an allotetraploid (Supplementary Fig. 5, Supplementary Table 5-6).

142

143 We generated orthogroups for *P. andersonii* and *T. orientalis* genes and six other Eurosid
144 species, including *Arabidopsis thaliana* and the legumes medicago and
145 soybean. From both *P. andersonii* and *T. orientalis* ~35,000 genes could be clustered
146 into >29,000 orthogroups (Supplementary Table 7-8). Within these orthogroups we identified
147 25,605 *P. andersonii* - *T. orientalis* orthologous gene pairs based on phylogenetic analysis as
148 well as whole genome alignments (Supplementary Table 8, note that there can be multiple
149 orthologous gene pairs per orthogroup). These orthologous gene pairs had a median
150 percentage nucleotide identity of 97% for coding regions (Supplementary Fig. 9-10). This
151 further supports the recent divergence of the two species and facilitates their genomic
152 comparison.

153 **Common utilization of symbiosis genes in *Parasponia* and medicago**

154

155 To assess commonalities in the utilization of symbiosis genes in *Parasponia* species and
156 legumes we employed two strategies. First, we identified close homologs of genes that were
157 characterized to function in legume-rhizobium symbiosis. This revealed that *P. andersonii*
158 contains orthologs of the vast majority of these legume symbiosis genes (117 out of 124)
159 (Supplementary Table 1, Supplementary Data File 1). Second, we compared the sets of
160 genes with enhanced expression in nodules of *Parasponia* and medicago. RNA sequencing
161 of *P. andersonii* nodules revealed 1,725 genes that have a significantly enhanced expression
162 level (fold change > 2, $p < 0.05$, DESeq2 Wald test) in any of three nodule developmental
163 stages compared with uninoculated roots (Supplementary Fig. 11; Supplementary Table 9).
164 For medicago, we used a comparable set of nodule-enhanced genes (1,463 genes)³¹. We
165 then determined the overlap of these two gene sets based on orthogroup membership and
166 found that 102 orthogroups comprise both *P. andersonii* and medicago nodule-enhanced
167 genes. This number is significantly larger than is to be expected by chance (permutation test,
168 $p < 0.02$)(Supplementary Fig. 12). Based on phylogenetic analysis of these orthogroups we
169 found that in 85 cases (out of 1,725) putative orthologs have been utilized in both *P.*
170 *andersonii* and medicago root nodules (Supplementary Table 10, Supplementary Data File
171 2). Among these 85 commonly utilized genes are 15 that we have identified in the first
172 strategy; e.g. the LCO-responsive transcription factor *NIN* and its downstream target
173 *NUCLEAR TRANSCRIPTION FACTOR-YA1 (NFYA1)* that are essential for nodule
174 organogenesis^{16,17,32,33}, and *RHIZOBIUM DIRECTED POLAR GROWTH (RPG)* involved in
175 intracellular infection³⁴. A notable exception to this pattern of common utilization are the
176 oxygen-binding hemoglobins. Earlier studies showed that *Parasponia* and legumes have
177 recruited hemoglobin genes by divergent evolution³⁵. Whereas legumes use class II
178 LEGHEMOGLOBIN to control oxygen homeostasis, *Parasponia* recruited the paralogous
179 class I HEMOGLOBIN 1 (HB1) for this function (Fig. 3a,b).

180 By exploiting the insight that nodule organogenesis and rhizobial infection can be genetically
181 dissected using hybrid plants we classified these commonly utilized genes into two
182 categories based on their expression profiles in both *P. andersonii* and F1 hybrid roots and
183 nodules (Fig. 4). The first category comprises genes that are upregulated in both *P.*
184 *andersonii* and hybrid nodules and that we associate with nodule organogenesis. The
185 second category comprises genes that are only upregulated in the *P. andersonii* nodule that
186 we therefore associate with infection and/or fixation. These variations in expression show
187 that the commonly utilized genes commit functions in various developmental stages of the *P.*
188 *andersonii* root nodule.

189

190 **Lineage-specific adaptation in *Parasponia* HEMOGLOBIN 1**

191

192 We further examined *HB1* as it was recruited independently from legumes (Fig. 3a,b)³⁵.
193 Biochemical studies have revealed that *P. andersonii* PanHB1 has oxygen affinities and
194 kinetics that are adapted to their symbiotic function, whereas this is not the case for *T.*
195 *tomentosa* *TtoHB1*^{35,36}. We therefore examined HB1 from *Parasponia* species, *Trema*
196 species, and other non-symbiotic Rosales species to see if these differences are due to a
197 gain of function in *Parasponia* or a loss of function in the non-symbiotic species. Based on
198 protein alignment we identified *Parasponia*-specific adaptations in 7 amino acids (Fig. 3c,d).
199 Among these is Ile(101) for which it is speculated to be causal for a functional change in *P.*
200 *andersonii* HB1³⁶. HEMOGLOBIN-controlled oxygen homeostasis in rhizobium-infected
201 nodule cells is crucial to protect the nitrogen-fixing enzyme complex Nitrogenase^{2,3}.
202 Therefore, *Parasponia*-specific gain of function adaptations in *HB1* most likely were an
203 essential evolutionary step towards functional rhizobium nitrogen fixing root nodules.

204

205

206

207 **Parallel loss of symbiosis genes in *Trema* and other relatives of *Parasponia***

208

209 Evolution of complex genetic traits is often associated with gene copy number variations
210 (CNVs)³⁷. To test if CNVs were associated with a potential independent evolution of
211 nodulation in *Parasponia*, we focussed on two gene sets: (1) the 117 symbiosis genes that
212 have been characterized in legumes, and (2) the 1,725 genes with a nodule-enhanced
213 expression in *P. andersonii* (these sets partially overlap and add up to 1,791 genes; see
214 Supplementary Fig. 13). To ensure that our findings are consistent between the *Parasponia*
215 and *Trema* genera and not due to species-specific events, we analyzed the additional draft
216 genome assemblies of two *Parasponia* and two *Trema* species (Supplementary Table 6).
217 Finally, we discarded *Trema*-specific duplications as we considered them irrelevant for the
218 nodulation phenotype. This resulted in only 11 consistent CNVs in the 1,791 symbiosis
219 genes examined, further supporting the recent divergence between *Parasponia* and *Trema*.
220 Due to the dominant inheritance of nodule organogenesis in F₁ hybrid plants, we anticipated
221 finding *Parasponia*-specific gene duplications that could be uniquely associated with
222 nodulation. Surprisingly, we found only one consistent *Parasponia*-specific duplication in
223 symbiosis genes; namely for a *HYDROXYCINNAMOYL-COA SHIKIMATE TRANSFERASE*
224 (*HCT*) (Supplementary Fig. 14-15). This gene has been investigated in the legume forage
225 crop alfalfa (*Medicago sativa*), where it was shown that HCT expression correlates negatively
226 with nodule organogenesis^{38,39}. Therefore, we do not consider this duplication relevant for the
227 nodulation capacity of *Parasponia*. Additionally, we identified three consistent gene losses in
228 *Parasponia* among which is the ortholog of LysM-type *EXOPOLYSACCHARIDE*
229 *RECEPTOR 3* that in lotus inhibits infection of rhizobia with incompatible
230 exopolysaccharides^{40,41} (Table 1, Supplementary Fig. 16-17). Such gene losses may have
231 contributed to effective rhizobium infection in *Parasponia* and their presence in *T. tomentosa*
232 could explain the lack of intracellular infection in the F1 hybrid nodules.

233

234 Contrary to our initial expectations, we discovered consistent loss or pseudogenization of
235 seven symbiosis genes in *Trema*. These genes have a nodule-specific expression profile in
236 *Parasponia*, suggesting that they function exclusively in symbiosis (Fig. 5). Three of these
237 are orthologs of genes that are essential for establishment of nitrogen-fixing nodules in
238 legumes: *NIN*, *RPG*, and the LysM-type LCO receptor *NFP/NFR5* (Supplementary Fig. 18-
239 19). In the case of *NFP/NFR5*, we found two close homologs of this gene, *NFP1* and *NFP2*,
240 of which the latter is consistently pseudogenized in *Trema* species (Fig. 6). In an earlier
241 study we used RNA interference (RNAi) to target *PanNFP1*, which led to reduced nodule
242 numbers and a block of intracellular infection by rhizobia as well as arbuscular mycorrhiza²⁷.
243 Most probably, however, the RNAi construct unintentionally also targeted *PanNFP2*, as both
244 genes are 69% identical in the 422 bp RNAi target region. Phylogenetic reconstruction
245 revealed that the *NFP1-NFP2* duplication predates the divergence of legumes and
246 *Parasponia*, and that *Parasponia NFP2* is most closely related to legume *MtNFP/LjNFR5*
247 rhizobium LCO receptors (Fig. 6). Additionally, in *P. andersonii* nodules *PanNFP2* is
248 significantly higher expressed than *PanNFP1* (Supplementary Fig. 20). Taken-together, this
249 suggests that *PanNFP2* represents a rhizobium LCO receptor that functions in nodule
250 formation and intracellular infection in *Parasponia*.

251
252 Based on expression profiles and phylogenetic relationships we postulate that also
253 *Parasponia NIN* and *RPG* commit essential symbiotic functions similar as in other nodulating
254 species (Fig. 5; Supplementary Fig. 21-23). Expression of *PanRPG* increases >300 fold in *P.*
255 *andersonii* nodules that become intracellularly infected (nodule stage 2), whereas in F₁ hybrid
256 nodules -which are devoid of intracellular rhizobium infection- *PanRPG* upregulation is less
257 than 20-fold (Fig. 5). This suggests that *PanRPG* commits a function in rhizobium infection,
258 similar as found in medicago³⁴. The transcription factor *NIN* has been studied in several
259 legume species as well as in the actinorhizal plant casuarina (*Casuarina glauca*) and in all
260 cases shown to be essential for nodule organogenesis^{14,16,42,43}. Loss of *NIN* and/or *NFP2* in

261 *Trema* species can explain the genetic dominance of nodule organogenesis in the
262 *Parasponia* x *Trema* F1 hybrid plants.

263

264 Next, we questioned whether loss of these symbiosis genes also occurred in more distant
265 relatives of *Parasponia*. We analysed non-nodulating species representing 6 additional
266 lineages of the Rosales clade; namely hops (*Humulus lupulus*, Cannabaceae)⁴⁴, mulberry
267 (*Morus notabilis*, Moraceae)⁴⁵, jujube (*Ziziphus jujuba*, Rhamnaceae)⁴⁶, peach (*Prunus*
268 *persica*, Rosaceae)⁴⁷, woodland strawberry (*Fragaria vesca*, Rosaceae)⁴⁸, and apple (*Malus*
269 *x domestica*, Rosaceae)⁴⁹. This revealed a consistent pattern of pseudogenization or loss of
270 *NFP2*, *NIN* and *RPG* orthologs, the intact jujube *ZjNIN* being the only exception (Fig. 7). We
271 note that for peach *NIN* was previously annotated as protein-coding gene⁴⁷. However, based
272 on comparative analysis of conserved exon structures we found two out-of-frame mutations
273 (see supplementary Fig. 24). Because the pseudogenized symbiosis genes are largely intact
274 in most of these species and differ in their deleterious mutations, the loss of function of these
275 essential symbiosis genes should have occurred recently and in parallel in at least seven
276 Rosales lineages. As we hypothesize that *NFP2*, *NIN* and *RPG* are essential for nodulation,
277 we argue that *Trema* species, hops, mulberry, jujube, woodland strawberry, apple, and
278 peach irreversibly, recently, and independently lost the potential to nodulate.

279

280 **DISCUSSION**

281

282 Here we present the nodulating Cannabaceae species *Parasponia* as a comparative system
283 to obtain insights in the evolutionary trajectory of nitrogen-fixing symbioses. Instead of finding
284 gene duplications that can explain a gain of symbiosis in *Parasponia*, we found parallel loss
285 or pseudogenization of symbiosis genes in non-nodulating Rosales species. This indicates
286 that in non-nodulating Rosales lineages these symbiosis genes experienced a recent period
287 of reduced functional constraints. This challenges current hypotheses on the evolution of

288 nitrogen fixing plant-microbe symbiosis.

289

290 Evolution of nodules is widely considered to be a two-step process: first an unspecified
291 predisposition event in the ancestor of all nodulating species, bringing species in the
292 nitrogen fixation clade to a precursor state for nodulation^{1,6}. Subsequently, nodulation
293 originated in parallel; eight times with *Frankia* and twice with rhizobium^{1,6-9}. This hypothesis
294 is most parsimonious and suggests a minimum number of independent losses of symbiosis.
295 *NFP/NFR5*, *NIN* and *RPG* are essential for nodulation in legumes and -in case of *NIN*- the
296 non-legume casuarina⁴³. Consequently, the non-nodulating species that have lost these
297 genes irreversibly lost the potential to nodulate. This opposes the current view that non-host
298 relatives of nodulating species are generally in a precursor state for nodulation^{1,6}.

299

300 The loss of symbiosis genes in non-nodulating plants is difficult to explain under the current
301 hypothesis of parallel origins of nodulation. These genes commit functions that currently
302 cannot be linked to any non-symbiotic processes. As a consequence, the interpretation of
303 why such a diverse set of genes repeatedly experienced reduced functional constraints in
304 non-nodulating plant lineages requires these genes to be linked to some other, yet unknown,
305 common process. Additionally, the current hypothesis of parallel origins would imply
306 convergent recruitment of at least 85 genes to commit symbiotic functions in *Parasponia* and
307 legumes. This implies parallel evolution of a highly complex trait.

308

309 Alternatively, the parallel loss of symbiosis genes in non-nodulating plants can be interpreted
310 as a single gain and massive loss of nodulation. In this new hypothesis nodulation is much
311 older than generally anticipated, and possibly represents the hypothetical predisposition
312 event in the nitrogen fixation clade. Subsequently, nodulation was lost in most descendent
313 lineages (hence massive). This new hypothesis fits our data better in four ways. (I.) It more
314 convincingly explains the parallel loss of symbiosis genes in non-nodulating plants, because

315 then gene loss correlates directly with loss of nodulation. (II.) A single gain of nodulation
316 explains the origin of the conserved set of more than 80 symbiosis genes utilized by
317 *Parasponia* and medicago. (III.) It is in line with the lack of *Parasponia*-specific gene
318 duplications that associate with nodulation. (IV.) The duplication in the NFP/NFR5 clade
319 encoding putative LCO LysM-type receptors predates the Rosales and Fabales split, thereby
320 coinciding with the origin of the nitrogen fixation clade. A single gain of nodulation would
321 require only a single (sub)neofunctionalization event of LCO-receptors to function in root
322 nodule formation. Additionally, the single gain-massive loss hypothesis eliminates the
323 predisposition event, a theoretical concept that currently cannot be addressed
324 mechanistically. Therefore, we consider a single gain-massive loss hypothesis as plausible.

325

326 Loss of nodulation is not controversial, as it is generally considered to have occurred >20
327 times in the legume family^{6,7}. Nevertheless, the single gain-massive loss hypothesis implies
328 many more events than the current hypothesis of parallel gains. Based on phylogenetic
329 evidence, the minimum number of losses required to explain the pattern of nodulation in the
330 nitrogen fixation clade implies 20 events in Rosales (8 of which in the Cannabaceae), 5 in
331 Fagales, 3 in Cucurbitales and 2 in Fabales (not taken into account Fabaceae)⁷. However,
332 as the identified pseudogenes in *Trema* species, mulberry, jujube, apple, and peach are
333 relatively intact we hypothesize that loss of nodulation has occurred relatively recent,
334 which would imply significantly more events. In either case, this hypothesis is not the
335 most parsimonious. On the other hand, it is conceptually easier to lose a complex trait, such
336 as nodulation, rather than to gain it⁹. Genetic studies in legumes indeed demonstrated that
337 nitrogen-fixing symbiosis can be abolished by a single knockout mutation in tens of different
338 genes, among which are *NFP/NFR5*, *NIN* and *RPG* (Supplementary Table 1). This suggests
339 that simple parsimony may not be the best way to model the evolution of nodulation.

340

341 Massive, recent and parallel loss of nodulation may have been triggered by changes at a

342 geological scale, e.g. a glacial maximum. During periods of glacial maxima, which
343 occurred between 18,000 and 800,000 years ago, atmospheric CO₂ levels dropped
344 below 200 ppm^{50,51}. Experiments show that such CO₂ concentrations have profound
345 effects on photosynthesis and plant growth in general⁵². Under such conditions
346 photosynthates may have been the growth limiting factor, rather than fixed nitrogen⁵²⁻⁵⁴.
347 In line with this nitrogen-fixation rates in the legume *Prosopis glandulosa* (honey
348 mesquite) can drop to zero when grown at 200 ppm CO₂⁵³. Therefore it is likely that the
349 nitrogen fixation trait has experienced relaxed constraints during periods of low
350 atmospheric CO₂ concentration, leading to genetic defects.

351
352 Based on the single gain-massive loss hypothesis we can make the following predictions.
353 First, the hypothesis implies that many (if not most) ancestral species in the nitrogen-fixing
354 clade were nodulators. This should be substantiated by fossil evidence. Currently, fossil data
355 on nodules are basically absent with only a single report on a fossilized nodule that is
356 estimated to be 11,5 thousand years old^{55,56}. An alternative strategy is to infer the presence
357 of nitrogen-fixing symbiosis from N isotope variation in fossil tree rings. This method was
358 successfully applied to discriminate tree species that predominantly utilize biologically fixed
359 nitrogen from tree species that use nitrogen resources retrieved from soil⁵⁷. Secondly, we
360 predict that actinorhizal plant species maintained *NIN*, *RPG*, and possibly *NFP2* (in case
361 LCOs are used as symbiotic signal⁵⁸), and that these genes are essential for nodulation. This
362 can be shown experimentally, as was done for *NIN* in casuarina⁴³.

363
364 The loss of symbiosis genes in non-nodulating plant species is not absolute, as we observed
365 a functional copy of *NIN* in jujube. This pattern is similar to the pattern of gene loss in species
366 that lost endomycorrhizal symbiosis^{59,60}. Also in that case, occasionally such genes have
367 been maintained in non-mycorrhizal plants. Conservation of *NIN* in jujube suggests that this
368 gene has a non-symbiotic function. Contrary to *NFP2*, which is the result of a gene

369 duplication near the origin of the nitrogen-fixing clade, functional copies of *NIN* are also
370 present in species outside the nitrogen-fixing clade (Supplementary Fig. 22). This suggests
371 that these genes may have retained -at least in part- an unknown ancestral non-symbiotic
372 function in some lineages within the nitrogen-fixing clade. Alternatively, *NIN* may have
373 acquired a new non-symbiotic function within some lineages in the nitrogen-fixing clade.

374

375 As hemoglobin is crucial for rhizobium symbiosis³, it is striking that *Parasponia* and legumes
376 do not use orthologous copies of hemoglobin genes in their nodules. At first sight this seems
377 inconsistent with a single gain of nodulation. However, a scenario that incorporates a switch
378 in microsymbionts can reconcile the use of paralogous hemoglobin genes with the
379 occurrence of two types of microsymbiont in the nitrogen fixing clade. This scenario would
380 dictate a single gain of actinorhizal symbiosis in the nitrogen fixing clade, and a switch from
381 *Frankia* to rhizobium in the ancestors of both *Parasponia* and legumes. As *Frankia* species
382 possess intrinsic physical characteristics to protect the Nitrogenase enzyme for oxidation,
383 expression of plant encoded hemoglobin in nodules is not a prerequisite for nitrogen fixation
384 in actinorhizal plants⁶¹⁻⁶⁴. In line with this, there is no evidence that *Ceanothus* spp.
385 (Rhamnaceae, Rosales) - which represent the closest nodulating relatives of *Parasponia* -
386 express a hemoglobin gene in *Frankia*-infected nodules⁶²⁻⁶⁴. A microsymbiont switch from
387 *Frankia* to rhizobium would therefore require adaptations in hemoglobin. Based on the fact
388 that *Parasponia* acquired lineage-specific adaptations in HB1 that are considered to be
389 essential to control oxygen homeostasis in rhizobium root nodules^{35,36}, such a symbiont
390 switch may have occurred early in the *Parasponia* lineage.

391

392 The uncovered evolutionary trajectory of a rhizobium nitrogen-fixing symbiosis provides
393 novel leads in attempts to engineer nitrogen-fixing root nodules in agricultural crop plants.
394 Such a translational approach is anticipated to be challenging⁶⁵, and the only published study
395 so far, describing transfer of 8 LCO signaling genes, was unsuccessful⁶⁶. If we interpret the

396 parallel loss of symbiosis genes in non-nodulating plants as evidence that these genes have
397 been neofunctionalized to commit symbiotic functions, then this gene set is essential in any
398 engineering approach. However, transfer of symbiosis genes may not be sufficient to obtain
399 functional nodules. The lack of infection in nodules on hybrid plants that contain a full
400 genome complement of *Parasponia* indicates the presence of an inhibitory mechanism in *T.*
401 *tomentosa*. Such a mechanism may also be present in other non-host species.
402 Consequently, engineering nitrogen-fixing nodules requires gene knockouts in non-
403 nodulating plants to overcome inhibition of intracellular infection. *Trema* may be the best
404 candidate species for such a (re)engineering approach, due to its high genetic similarity with
405 *Parasponia* and the availability of transformation protocols⁶⁷. Therefore, the presented
406 *Parasponia-Trema* comparative system may not only be suited for evolutionary studies, but
407 also can form an excellent experimental platform to obtain essential insights to engineer
408 nitrogen fixing root nodules.

409 MATERIAL AND METHODS

410

411 ***Parasponia* - *Trema* intergeneric crossing and hybrid genotyping**

412 *Parasponia* and *Trema* are wind-pollinated species. A female-flowering *P. andersonii*
413 individual WU1.14 was placed in a plastic shed together with a flowering *T. tomentosa* WU10
414 plant. Putative F₁ hybrid seeds were germinated (see Supplementary Methods) and
415 transferred to potting soil. To confirm the hybrid genotype a PCR marker was used that
416 visualizes a length difference in the promoter region of *LIKE-AUXIN 1 (LAX1)* (primers:
417 LAX1-f: ACATGATAATTTGGGCATGCAACA, LAX1-r: TCCCGAATTTTCTACGAATTGAAA,
418 amplicon size *P. andersonii*: 974 bp; *T. tomentosa*: 483 bp). Hybrid plant H9 was propagated
419 *in vitro*^{27,68}. The karyotype of the selected plants was determined according to Geurts and De
420 Jong 2013⁶⁹.

421

422 **Nodulation and nitrogenase activity assays**

423 All nodulation assays were conducted with *Mesorhizobium plurifarium* BOR2. This strain was
424 isolated from *P. andersonii* root nodules grown in soil samples collected from the root
425 rhizosphere of *Trema orientalis* plants in Malaysian Borneo, province of Sabah⁷⁰. *M.*
426 *plurifarium* was grown on yeast extract mannitol medium at 28°C³⁰. Plants were grown in
427 sterile plastic 1 liter pots containing perlite and EKM medium supplemented with 0.375 mM
428 NH₄NO₃ and rhizobium (OD600:0.05)⁷¹. Nodule number per plant was quantified 6 weeks
429 post inoculation.

430

431 Acetylene reduction assays⁷² were conducted on nodules harvested 6 weeks post
432 inoculation with *Mesorhizobium plurifarium* strain BOR2. Nodules were sampled per plant
433 and collected in 15 ml headspace vials with screw lids. 2.5 ml of acetylene was injected into
434 the vial and incubated for about 10 minutes, after which 1 ml headspace was used to
435 quantify ethylene nitrogenase activity using an ETD 300 detector (Sensor Sense, Nijmegen,

436 The Netherlands; Isogen, Wageningen, The Netherlands)⁷³.

437

438 To isolate *P. andersonii* nodules at 3 developmental stages nodules were separated based
439 on morphology and size. Stage 1: nodules are round and < 1mm in diameter in size. The
440 outer cell layers of stage 1 nodules are transparent. Light microscopy confirmed that at this
441 stage, rhizobia already reach the central part of the nodule, but are mainly present in the
442 apoplast (Fig. 1h). Stage 2: nodules are brownish, and ~2 mm in size. Nodules have formed
443 an apical meristem and 2-3 cell layers have been infected by rhizobia (Fig. 1i). Stage 3:
444 nodules are pinkish on the outside due to accumulation of haemoglobin and > 2 mm in size.
445 Light microscopy showed that stage 3 nodules contain zones of fully infected cells (Fig. 1j).
446 For each of these stages three biological replicates were used for RNA sequencing.

447

448 **Arbuscular mycorrhization assay**

449 Two week old seedlings were transferred to 800 ml Sand:Granule:*Rhizophagus irregularis*
450 (*Rir*, INOQ TOP- INOQ GmbH, Schnega Germany) inoculum mixture (1:1:0.01), irrigated
451 with 80 ml ½ strength modified Hoagland solution containing 20 µM K₂HPO₄⁷⁴ and grown for
452 an additional 6 weeks at 28°C, under a photoperiod of 16/8h (day/night). 50 ml additional
453 nutrient solution was provided once a week. Mycorrhization efficiency was analysed as
454 previously described⁷⁵ for three aspects: 1) frequency of fungal colonization in 1 cm root
455 segments; 2) average level of mycorrhization in all root fragments, and 3) arbuscular
456 abundance in all root fragments (Supplementary Fig. 1). Arbuscules were WGA-Alexafluor
457 488-stained and imaged according to Huisman *et al* 2015⁷⁶.

458

459 **DNA/RNA sequencing**

460 Paired-end Illumina libraries (insert size 500bp, 100bp reads) were prepared for all
461 accessions (Supplementary Table 5), mate-pair libraries (3Kb, 7Kb, and 10Kb) and
462 overlapping fragment libraries (450bp insert size, 250bp reads) were prepared for the

463 reference accessions (*P. andersonii* accession WU01 and *T. orientalis* accession RG33).
464 Paired-end and mate-pair libraries were sequenced on an Illumina HiSeq2000, overlapping
465 libraries were sequenced on an Illumina MiSeq. For the *P. andersonii* and *T. orientalis*
466 reference genomes a total of 75Gb (~132x genome coverage) and 61Gb (~121x coverage)
467 of data was produced respectively. The other accessions were sequenced at an average
468 coverage of ~30X. See Supplementary Methods for further details on library preparation and
469 sequencing. RNA samples from various tissues and nodulation stages were isolated from *P.*
470 *andersonii* and *Trema orientalis* RG33 (Supplementary Methods, Supplementary Table 11).
471 Library preparation and RNA sequencing was conducted by B.G.I. (Shenzhen, China).

472

473 **Estimation of heterozygosity levels and genome size**

474 To assess levels of heterozygosity and genome size we performed *k*-mer analyses.
475 Multiplicities of 21-mers were extracted from the reads using Jellyfish (version 2.2.0)⁷⁷ and
476 processed using custom R scripts. First, a multiplicity threshold was determined below which
477 most *k*-mers are considered to represent sequencing errors and which were excluded from
478 further analysis. In principle, errors occur randomly and this generates a high frequency peak
479 at multiplicity 1 after which frequency decreases and subsequently increases due to a broad
480 frequency peak around the mean genome coverage. The error multiplicity threshold was
481 therefore set at the multiplicity with the lowest frequency between these two peaks. Next, we
482 identified the peak multiplicity as the one with the highest frequency. Homozygous genome
483 coverage was estimated by scaling the peak multiplicity proportional to the difference of its
484 frequency with that of multiplicities one below and above. Heterozygous coverage was
485 defined as half that of the homozygous coverage (Supplementary Fig. 5). Finally, genome
486 size was calculated as the total number of error-free *k*-mers divided by the estimated
487 homozygous genome coverage (Supplementary Table 4). These estimates are generally
488 comparable to those based on FACS measurements⁷⁸ (Supplementary Table 3) except for
489 genomes that differ much from the reference used to calibrate the FACS results (*Medicago*

490 *truncatula*, ~500Mb). This inconsistency is probably due to the non-linearity of the FACS
491 measurements. We therefore consider the quantitative genome estimates based on k-mer
492 analysis a more accurate estimation of genome size.

493

494 **Characterization of repetitive sequences**

495 Repetitive sequences are inherently difficult to assemble. We therefore characterized and
496 quantified repetitive element using the ab initio graph-based clustering approach
497 implemented in RepeatExplorer⁷⁹. Analyses were based on random subsamples of 20,000
498 paired-end reads and included a reclustering step where clusters with shared mate pairs are
499 merged (threshold $k=0.2$). Repeat classification was based on the RepeatExplorer
500 Viridiplantae dataset and on plant organellar sequences. Relative sizes of repetitive
501 sequences in the genome were scaled by the genome size estimations based on *k*-mer
502 analysis to generate absolute sizes in Mb (Fig. 2).

503

504 **Assembly of reference genomes**

505 The raw sequencing data were preprocessed. First, adapters (standard and junction) were
506 removed and reads were trimmed using fastq-mcf (version 1.04.676)⁸⁰. Minimum remaining
507 sequence length was set to 50 for HiSeq data and 230 for MiSeq data. Duplicates were
508 removed using FastUniq (version 1.1)^{80,81}. Chloroplast and mitochondrial genomes were
509 assembled first with IOGA (version 1) using reference sets of plant chloroplast and
510 mitochondrial genomes⁸². Chloroplast and mitochondrial reads were identified and separated
511 from the nuclear reads by mapping to four organellar assemblies (*Parasponia andersonii*,
512 *Trema orientalis*, *Morus indica*, *Malus x domestica*) using BWA (version 0.7.10)⁸³. Finally, a
513 contamination database was produced by BLASTing contigs from earlier in-house draft
514 genome assemblies from *Parasponia andersonii* and *Trema orientalis* against NCBI's nt
515 database. Hits outside the plant kingdom were extracted using a custom script and
516 corresponding sequences were downloaded from GenBank and a database of plant viruses

517 was added (<http://www.dpvweb.net/seqs/allplantfasta.zip>). Genomics reads were cleaned by
518 mapping to this contamination database.

519

520 The preprocessed data were *de novo* assembled using ALLPATHS-LG (release 48961)⁸⁴.
521 Relevant parameters were PLOIDY=2 and GENOME_SIZE=600000000. The assemblies
522 were performed on the Breed4Food High Performance Cluster from Wageningen UR
523 (<http://breed4food.com>).

524

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

533

534 Resulting contigs were subsequently scaffolded with two rounds of SSPACE (v3.0)⁸⁵,
535 standard with the mate pair libraries. In order to use reads mapped with BWA (v0.7.10) the
536 SSPACE utility sam_bam2tab.pl was used. We used the output of the second run of
537 SSPACE scaffolding as the final assembly.

538

539 Validation of the final assemblies showed that 90-100% of the genomic reads mapped back
540 to the assemblies (Supplementary Table 5), and 94-98% of CEGMA⁸⁶ and BUSCO⁸⁷ genes
541 were detected (Supplementary Table 6).

542

543

544 **Annotation of reference genomes**

545 Repetitive elements were identified following the standard Maker-P recipe
546 ([http://weatherby.genetics.utah.edu/MAKER/wiki/index.php/Repeat_Library_Construction-
547 Advanced](http://weatherby.genetics.utah.edu/MAKER/wiki/index.php/Repeat_Library_Construction-Advanced) accessed october 2015) as described on the GMOD site: (1) RepeatModeler with
548 Repeatscout v1.0.5, Recon v1.08, RepeatMasker version open4.0.5, using RepBase version
549 20140131⁸⁸ and TandemRepeatFinder; (2) GenomeTools: LTRharvest and LTRdigest⁸⁹; (3)
550 MITEhunter with default parameters⁹⁰. We created species-specific repeat libraries for both
551 *P. andersonii* and *T. orientalis* separately and combined these into a single repeat library,
552 filtering out sequences that are >98% similar. We masked both genomes using
553 RepeatMasker with this shared repeat library.

554

555 To aid the structural annotation we used 11 *P. andersonii* and 6 *T. orientalis* RNA
556 sequencing datasets (Supplementary Table 11). All RNA-seq samples were assembled *de*
557 *novo* using genome-guided Trinity⁹¹, resulting in one combined transcriptome assembly per
558 species. In addition all samples were mapped to their respective reference genomes using
559 BWA and processed into putative transcripts using cufflinks⁹² and transdecoder⁹³. This
560 resulted in one annotation file (gff) per transcriptome sample per species. As protein
561 homolog evidence, only Swiss-Prot⁹⁴ entries filtered for plant proteins were used. This way
562 we only included manually verified protein sequences and prevented the incorporation of
563 erroneous predictions. Finally, four gene-predictor tracks were used: 1) SNAP⁹⁵, trained on
564 *P. andersonii* transdecoder transcript annotations; 2) SNAP, trained on *T. orientalis*
565 transdecoder transcript annotations; 3) Augustus⁹⁶ as used in the BRAKER pipeline, trained
566 on RNA-seq alignments⁹⁷; 4) GeneMark-ET as used in the BRAKER pipeline, trained on
567 RNA-seq alignments⁹⁸.

568

569 First, all evidence tracks were processed by Maker-P^{87,99}. The results were refined with
570 EVIDENCEModeler (EVM)¹⁰⁰, which was used with all the same tracks as Maker-P, except for

571 the Maker-P blast tracks and with the addition of the Maker-P consensus track as additional
572 evidence. Ultimately, EVM gene models were preferred over Maker-P gene models, except
573 when there was no overlapping EVM gene model. Where possible, evidence of both species
574 was used to annotate each genome (i.e. *de novo* RNA-seq assemblies of both species were
575 aligned to both genomes).

576

577 To take maximum advantage of annotating two highly similar genomes simultaneously we
578 developed a custom reconciliation procedure involving whole genome alignments. The
579 consensus annotations from merging the EVM and Maker-P annotations were transferred to
580 their respective partner genome using nucmer¹⁰¹ and RATT revision 18¹⁰² (i.e. the *P.*
581 *andersonii* annotation was transferred to *T. orientalis* and *vice versa*), based on nucmer
582 whole genome alignments (Supplementary Fig. 9). Through this reciprocal transfer, both
583 genomes had two candidate annotation tracks, the original (called P and T) and the
584 transferred (called P' on *T. orientalis* and T' on *P. andersonii*). This allowed for validation of
585 both annotations simultaneously, assuming that two orthologous regions containing a single
586 gene that has not changed since its common ancestor, should be annotated identically. If
587 annotations between orthologous regions differ, we used RNA-seq evidence and protein
588 alignments for curation, by picking one of four annotation combinations: P and T, P and T', P'
589 and T, or P' & T'. Picking one of these options is based on transcriptome coverage: the
590 combination with the highest percentage of covered introns per annotation is the most likely.
591 If there is insufficient coverage in any of the genomes, the combination with the highest
592 pairwise identity based on protein alignments of the translated annotation is selected.

593

594 For the reconciliation procedure we developed a custom Python script. To deal with
595 orthologous regions containing different numbers of annotations, we identified 'annotation
596 clusters'. This was done iteratively by selecting overlapping gene models and transferred
597 gene models with the same gene ID. Two annotations were considered to be overlapping if

598 they were on the same strand and at least one of each of their exons overlapped. This
599 allowed for separate processing of genes on opposing strands, and 'genes within genes', i.e.
600 a gene within the intron of another gene. The validation of annotation differences between *P.*
601 *andersonii* and *T. orientalis* greatly reduces technical variation and improves all downstream
602 analyses.

603

604 After automatic annotation and reconciliation 1,693 *P. andersonii* genes and 1,788 *T.*
605 *orientalis* genes were manually curated. These were mainly homologs of legume symbiosis
606 genes and genes that were selected based on initial data exploration.

607

608 To assign putative product names to the predicted genes we combined BLAST results
609 against Swiss-Prot, TrEMBL and nr with InterProScan results (custom script). To annotate
610 GO terms and KEGG enzyme codes Blast2GO was used with the nr BLAST results and
611 interproscan results. Finally, we filtered all gene models with hits to InterPro domains that are
612 specific to repetitive elements.

613

614 **Phylogenetic reconstruction of Cannabaceae**

615 Multiple sequence alignments were generated using MAFFT (version 7.017)¹⁰³ and
616 phylogenetic analyses were performed using MrBayes (version 3.2.2)¹⁰⁴. The first
617 phylogenetic reconstruction of the Cannabaceae was based on four markers comprising data
618 from Yang et al. 2013²⁹ supplemented with new data generated with primers and protocols
619 published in this manuscript (Supplementary Table 12). Analysis was based on five optimal
620 partitions and models of sequence evolution as estimated by PartitionFinder (version
621 2.0.0)¹⁰⁵: atpB-rbcL combined with trnL-F (GTR+I+G); first codon position of rbcL (GTR+I+G);
622 second position of rbcL (SYM+I+G); third position of rbcL (GTR+G); rps16 (GTR+G). An
623 additional phylogenetic reconstruction of the Cannabaceae was based on whole chloroplast
624 genomes (Supplementary Table 12). Analysis was based on eight optimal partitions and

625 models of sequence evolution as estimated by PartitionFinder: tRNA sequence (HKY+I),
626 rRNA sequence (GTR+I), long single copy region (LSC) coding sequence (GTR+I+G), LSC
627 non-coding sequence (GTR+G), short single copy region (SSC) coding sequence (GTR+G),
628 SSC non-coding sequence (GTR+G), inverted repeat region (IR) coding sequence (GTR+G),
629 and IR non-coding sequence (GTR+G). For both Cannabaceae reconstructions additional
630 bootstrap support values were calculated using RAxML (version 8.2.9)^{105,106} using the same
631 partitions applying the GTR+G model. All gene tree reconstructions were based on
632 unpartitioned analysis of protein sequence with the POISSON+G model.

633

634 **Orthogroup inference**

635 To determine the relationships between *P. andersonii* and *T. orientalis* genes, as well as with
636 other plant species we inferred orthogroups with OrthoFinder (version 0.4.0)¹⁰⁷. Since
637 orthogroups are defined as the set of genes that are descended from a single gene in the
638 last common ancestor of all the species being considered, they can comprise orthologous as
639 well as paralogous genes. Our analysis included proteomes of selected species from the
640 Eurosid clade: *Arabidopsis thaliana* TAIR10 (Brassicaceae, Brassicales)¹⁰⁸ and *Eucalyptus*
641 *grandis* v2.0 (Myrtaceae, Myrtales) from the Malvid clade¹⁰⁹; *Populus trichocarpa* v3.0
642 (Salicaceae, Malpighiales)¹¹⁰, legumes *Medicago truncatula* Mt4.0v1¹¹¹ and *Glycine max*
643 *Wm82.a2.v1* (Fabaceae, Fabales)¹¹², *Fragaria vesca* v1.1 (Rosaceae, Rosales)⁴⁸, *P.*
644 *andersonii* and *T. orientalis* (Cannabaceae, Rosales) from the Fabid clade (Supplementary
645 Table 7). Sequences were retrieved from phytozome (www.phytozome.net).

646

647 **Gene copy number variant detection**

648 To assess orthologous and paralogous relationships between *Parasponia* and *Trema* genes,
649 we inferred phylogenetic gene trees for each orthogroup comprising *Parasponia* and/or
650 *Trema* genes using the neighbour-joining algorithm¹¹³. Based on these gene trees, for each
651 *Parasponia* gene, its relationship to other *Parasponia* and *Trema* genes was defined as

652 follows. 1) orthologous pair: the sister lineage is a single gene from the *Trema* genome
653 suggesting that they are the result of a speciation event; 2) inparalog: the sister lineage is a
654 gene from the *Parasponia* genome, suggesting that they are the result of a gene duplication
655 event; 3) singleton: the sister lineage is a gene from a species other than *Trema*, suggesting
656 that the *Trema* gene was lost; 4) multi-ortholog: the sister lineage comprises multiple genes
657 from the *Trema* genome, suggesting that the latter are inparalogs. For each *Trema* gene,
658 relationship was defined in the same way but with respect to the *Parasponia* genome
659 (Supplementary Table 8). Because phylogenetic analysis relies on homology we assessed
660 the level of conservation in the multiple-sequence alignments by calculating the trident score
661 using MstatX (<https://github.com/gcollet/MstatX>)¹¹⁴. Orthogroups with a score below 0.1 were
662 excluded from the analysis. Examination of orthogroups comprising >20 inparalogs revealed
663 that some represented repetitive elements; these were also excluded. Finally, orthologous
664 pairs were validated based on the whole-genome alignments used in the annotation
665 reconciliation.

666

667 **Assembly of Parasponia and Trema draft genomes**

668 To assess whether gene copy number variants of interest are also present in other, non-
669 reference *Parasponia* and *Trema* genomes, we assembled genomic sequences of *P. rigida*,
670 *P. rugosa*, *T. levigata*, and *T. orientalis* accession RG16 based on the medium-coverage
671 sequence data that was also used for *k*-mer analysis (Supplementary Table 4-5). Assembly
672 was performed with the iterative de Bruijn graph assembler IDBA-UD (version 1.1.1)¹¹⁵,
673 iterating from 30-mers (assembling low-coverage regions) to 120-mers (accurately
674 assembling regions of high coverage), with incremental steps of 20. Genes of interest were
675 manually annotated and putatively lost genes or gene fragments were confirmed based on
676 (I.) mapping the medium-coverage reads to the respective *P. andersonii* or *T. orientalis*
677 RG33 reference genome and (II.) genomic alignments (Supplementary Fig. 18-19).

678

679 **Nodule-enhanced genes**

680 To assess gene expression in *Parasponia* nodules, RNA was sequenced from the three
681 nodule stages described above as well as uninoculated roots (Supplementary Table 11).
682 RNA-seq reads were mapped to the *Parasponia* reference genome with HISAT2 (version
683 2.02)¹¹⁶ using an index that includes exon and splice site information in the RNA-seq
684 alignments. Mapped reads were assigned to transcripts with featureCounts (version 1.5.0)¹¹⁷.
685 Normalization and differential gene expression were performed with DESeq2. Nodule
686 enhanced genes were selected based on >2.0 fold-change and $p \leq 0.05$ in any nodule stage
687 compared with uninoculated root controls. Genes without functional annotation or orthogroup
688 membership were excluded. To assess expression of *Parasponia* genes in the hybrid
689 nodules, RNA was sequenced from nodules and uninoculated roots. Here, RNA-seq reads
690 were mapped to a combined reference comprising two parent genomes from *P. andersonii*
691 and *T. tomentosa*.

692 **References**

- 693 1. Soltis, D. E. *et al.* Chloroplast gene sequence data suggest a single origin of the
694 predisposition for symbiotic nitrogen fixation in angiosperms. *Proc. Natl. Acad. Sci. U. S.*
695 *A.* **92**, 2647–2651 (1995).
- 696 2. Udvardi, M. & Poole, P. S. Transport and metabolism in legume-rhizobia symbioses.
697 *Annu. Rev. Plant Biol.* **64**, 781–805 (2013).
- 698 3. Ott, T. *et al.* Symbiotic leghemoglobins are crucial for nitrogen fixation in legume root
699 nodules but not for general plant growth and development. *Curr. Biol.* **15**, 531–535
700 (2005).
- 701 4. Burrill, T. J. & Hansen, R. Is symbiosis possible between legume bacteria and non-
702 legume plants? *Bulletin (University of Illinois (Urbana-Champaign campus). Agricultural*
703 *Experiment Station); no. 202* **202**, 115–181 (1917).
- 704 5. Stokstad, E. The nitrogen fix. *Science* **353**, 1225–1227 (2016).
- 705 6. Werner, G. D. A., Cornwell, W. K., Sprent, J. I., Kattge, J. & Kiers, E. T. A single
706 evolutionary innovation drives the deep evolution of symbiotic N₂-fixation in
707 angiosperms. *Nat. Commun.* **5**, 4087 (2014).
- 708 7. Li, H.-L. *et al.* Large-scale phylogenetic analyses reveal multiple gains of actinorhizal
709 nitrogen-fixing symbioses in angiosperms associated with climate change. *Sci. Rep.* **5**,
710 14023 (2015).
- 711 8. Doyle, J. J. Phylogenetic perspectives on the origins of nodulation. *Mol. Plant. Microbe.*
712 *Interact.* **24**, 1289–1295 (2011).
- 713 9. Doyle, J. J. Chasing unicorns: Nodulation origins and the paradox of novelty. *Am. J. Bot.*
714 **103**, 1865–1868 (2016).
- 715 10. Limpens, E. *et al.* LysM domain receptor kinases regulating rhizobial Nod factor-induced
716 infection. *Science* **302**, 630–633 (2003).
- 717 11. Madsen, E. B. *et al.* A receptor kinase gene of the LysM type is involved in legume
718 perception of rhizobial signals. *Nature* **425**, 637–640 (2003).

- 719 12. Radutoiu, S. *et al.* Plant recognition of symbiotic bacteria requires two LysM receptor-
720 like kinases. *Nature* **425**, 585–592 (2003).
- 721 13. Arrighi, J. F. *et al.* The *Medicago truncatula* lysine motif-receptor-like kinase gene family
722 includes *NFP* and new nodule-expressed genes. *Plant Physiol.* **142**, 265–279 (2006).
- 723 14. Marsh, J. F. *et al.* *Medicago truncatula* *NIN* is essential for rhizobial-independent nodule
724 organogenesis induced by autoactive Calcium/Calmodulin-Dependent Protein Kinase.
725 *Plant Physiol.* **144**, 324–335 (2007).
- 726 15. Broghammer, A. *et al.* Legume receptors perceive the rhizobial lipochitin oligosaccharide
727 signal molecules by direct binding. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 13859–13864
728 (2012).
- 729 16. Schauser, L., Roussis, A., Stiller, J. & Stougaard, J. A plant regulator controlling
730 development of symbiotic root nodules. *Nature* **402**, 191–195 (1999).
- 731 17. Soyano, T., Kouchi, H., Hirota, A. & Hayashi, M. NODULE INCEPTION directly targets
732 *NF-Y* subunit genes to regulate essential processes of root nodule development in *Lotus*
733 *japonicus*. *PLoS Genet.* **9**, e1003352 (2013).
- 734 18. Vernié, T. *et al.* The *NIN* transcription factor coordinates diverse nodulation programs in
735 different tissues of the *Medicago truncatula* root. *Plant Cell* tpc.15.00461 (2015).
- 736 19. Parniske, M. Arbuscular mycorrhiza: the mother of plant root endosymbioses. *Nat. Rev.*
737 *Microbiol.* **6**, 763–775 (2008).
- 738 20. Oldroyd, G. E. D. Speak, friend, and enter: signalling systems that promote beneficial
739 symbiotic associations in plants. *Nat. Rev. Microbiol.* **11**, 252–263 (2013).
- 740 21. Geurts, R., Xiao, T. T. & Reinhold-Hurek, B. What does it take to evolve a nitrogen-fixing
741 endosymbiosis? *Trends Plant Sci.* **21**, 199–208 (2016).
- 742 22. Clason, E. W. The vegetation of the upper-Badak region of mount Kelut (east java).
743 *Bulletin Jard. Bot. Buitenzorg Serie III*, 509–518 (1936).
- 744 23. Trinick, M. J. Symbiosis between *Rhizobium* and the non-legume, *Trema aspera*. *Nature*
745 **244**, 459–460 (1973).

- 746 24. Akkermans, A. D. L., Abdulkadir, S. & Trinick, M. J. Nitrogen-fixing root nodules in
747 Ulmaceae. *Nature* **274**, 190–190 (1978).
- 748 25. Becking, J. H. In: *Biological nitrogen fixation* (eds. Stacey, G., Burris, R. H. & Evans, H.
749 J.) 497–559 (Routledge, Chapman and Hall, 1992).
- 750 26. Marvel, D. J., Torrey, J. G. & Ausubel, F. M. Rhizobium symbiotic genes required for
751 nodulation of legume and nonlegume hosts. *Proc. Natl. Acad. Sci. U. S. A.* **84**, 1319–
752 1323 (1987).
- 753 27. Op den Camp, R. *et al.* LysM-type mycorrhizal receptor recruited for rhizobium
754 symbiosis in nonlegume *Parasponia*. *Science* **331**, 909–912 (2011).
- 755 28. Granqvist, E. *et al.* Bacterial-induced calcium oscillations are common to nitrogen-fixing
756 associations of nodulating legumes and non-legumes. *New Phytol.* **207**, 551–558
757 (2015).
- 758 29. Yang, M.-Q. *et al.* Molecular phylogenetics and character evolution of Cannabaceae.
759 *Taxon* **62**, 473–485 (2013).
- 760 30. Op den Camp, R. H. M. *et al.* Nonlegume *Parasponia andersonii* deploys a broad
761 rhizobium host range strategy resulting in largely variable symbiotic effectiveness. *Mol.*
762 *Plant. Microbe. Interact.* **25**, 954–963 (2012).
- 763 31. Roux, B. *et al.* An integrated analysis of plant and bacterial gene expression in symbiotic
764 root nodules using laser-capture microdissection coupled to RNA sequencing. *Plant J.*
765 **77**, 817–837 (2014).
- 766 32. Combier, J.-P. P. *et al.* MthAP2-1 is a key transcriptional regulator of symbiotic nodule
767 development regulated by microRNA169 in *Medicago truncatula*. *Genes Dev.* **20**, 3084–
768 3088 (2006).
- 769 33. Baudin, M. *et al.* A Phylogenetically conserved group of Nuclear Factor-Y transcription
770 factors interact to control nodulation in legumes. *Plant Physiol.* **169**, 2761–2773 (2015).
- 771 34. Arrighi, J.-F. *et al.* The *RPG* gene of *Medicago truncatula* controls *Rhizobium*-directed
772 polar growth during infection. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 9817–9822 (2008).

- 773 35. Sturms, R., Kakar, S., Trent, J. & Hargrove, M. S. *Trema* and *Parasponia* hemoglobins
774 reveal convergent evolution of oxygen transport in plants. *Biochemistry* **49**, 4085–4093
775 (2010).
- 776 36. Kakar, S. *et al.* Crystal structures of *Parasponia* and *Trema* hemoglobins: Differential
777 heme coordination is linked to quaternary structure. *Biochemistry* **50**, 4273–4280 (2011).
- 778 37. Żmieńko, A., Samelak, A., Kozłowski, P. & Figlerowicz, M. Copy number polymorphism
779 in plant genomes. *Theor. Appl. Genet.* **127**, 1–18 (2014).
- 780 38. Shadle, G. *et al.* Down-regulation of hydroxycinnamoyl CoA: shikimate
781 hydroxycinnamoyl transferase in transgenic alfalfa affects lignification, development and
782 forage quality. *Phytochemistry* **68**, 1521–1529 (2007).
- 783 39. Gallego-Giraldo, L. *et al.* Lignin modification leads to increased nodule numbers in
784 alfalfa. *Plant Physiol.* **164**, 1139–1150 (2014).
- 785 40. Kawaharada, Y. *et al.* Receptor-mediated exopolysaccharide perception controls
786 bacterial infection. *Nature* **523**, 308–312 (2015).
- 787 41. Kawaharada, Y. *et al.* Differential regulation of the Epr3 receptor coordinates
788 membrane-restricted rhizobial colonization of root nodule primordia. *Nat. Commun.* **8**,
789 14534 (2017).
- 790 42. Borisov, A. Y. *et al.* The Sym35 gene required for root nodule development in pea is an
791 ortholog of Nin from *Lotus japonicus*. *Plant Physiol.* **131**, 1009–1017 (2003).
- 792 43. Clavijo, F. *et al.* The *Casuarina NIN* gene is transcriptionally activated throughout
793 *Frankia* root infection as well as in response to bacterial diffusible signals. *New Phytol.*
794 **208**, 887–903 (2015).
- 795 44. Natsume, S. *et al.* The draft genome of hop (*Humulus lupulus*), an essence for brewing.
796 *Plant Cell Physiol.* **56**, 428–441 (2015).
- 797 45. He, N. *et al.* Draft genome sequence of the mulberry tree *Morus notabilis*. *Nat. Commun.*
798 **4**, ncomms3445 (2013).
- 799 46. Huang, J. *et al.* The jujube genome provides insights into genome evolution and the

- 800 domestication of sweetness/acidity taste in fruit trees. *PLoS Genet.* **12**, e1006433
801 (2016).
- 802 47. Verde, I. *et al.* The high-quality draft genome of peach (*Prunus persica*) identifies unique
803 patterns of genetic diversity, domestication and genome evolution. *Nat. Genet.* **45**, 487–
804 494 (2013).
- 805 48. Shulaev, V. *et al.* The genome of woodland strawberry (*Fragaria vesca*). *Nat. Genet.* **43**,
806 109–116 (2011).
- 807 49. Velasco, R. *et al.* The genome of the domesticated apple (*Malus × domestica* Borkh.).
808 *Nat. Genet.* **42**, 833–839 (2010).
- 809 50. Sigman, D. M., Hain, M. P. & Haug, G. H. The polar ocean and glacial cycles in
810 atmospheric CO₂ concentration. *Nature* **466**, 47–55 (2010).
- 811 51. Lüthi, D. *et al.* High-resolution carbon dioxide concentration record 650,000–800,000
812 years before present. *Nature* **453**, 379–382 (2008).
- 813 52. Gerhart, L. M. & Ward, J. K. Plant responses to low [CO₂] of the past. *New Phytol.* **188**,
814 674–695 (2010).
- 815 53. Polley, H. W., Johnson, H. B. & Mayeux, H. S. Increasing CO₂: Comparative responses
816 of the C₄ grass *Schizachyrium* and grassland invader *Prosopis*. *Ecology* **75**, 976–988
817 (1994).
- 818 54. Sage, R. F. Was low atmospheric CO₂ during the Pleistocene a limiting factor for the
819 origin of agriculture? *Glob. Chang. Biol.* **1**, 93–106 (1995).
- 820 55. Baker, D. & Miller, N. G. Ultrastructural evidence for the existence of actinorhizal
821 symbioses in the late Pleistocene. *Can. J. Bot.* **58**, 1612–1620 (1980).
- 822 56. Sprent, J. I. Evolving ideas of legume evolution and diversity: a taxonomic perspective
823 on the occurrence of nodulation. *New Phytol.* **174**, 11–25 (2007).
- 824 57. Gulbranson, E. L. *et al.* Nitrogen-fixing symbiosis inferred from stable isotope analysis of
825 fossil tree rings from the Oligocene of Ethiopia. *Geology* (2017). doi:10.1130/G39213.1
- 826 58. Nguyen, T. V. *et al.* An assemblage of *Frankia* Cluster II strains from California contains

- 827 the canonical nod genes and also the sulfotransferase gene nodH. *BMC Genomics* **17**,
828 796 (2016).
- 829 59. Delaux, P.-M. *et al.* Algal ancestor of land plants was preadapted for symbiosis. *Proc.*
830 *Natl. Acad. Sci. U. S. A.* **112**, 13390–13395 (2015).
- 831 60. Kamel, L., Keller-Pearson, M., Roux, C. & Ané, J.-M. Biology and evolution of arbuscular
832 mycorrhizal symbiosis in the light of genomics. *New Phytol.* **213**, 531–536 (2017).
- 833 61. Winship, L. J., Martin, K. J. & Sellstedt, A. The acetylene reduction assay inactivates
834 root nodule uptake hydrogenase in some actinorhizal plants. *Physiol. Plant.* **70**, 361–366
835 (1987).
- 836 62. Silvester, W. B. & Winship, L. J. Transient responses of nitrogenase to acetylene and
837 oxygen in actinorhizal nodules and cultured frankia. *Plant Physiol.* **92**, 480–486 (1990).
- 838 63. Silvester, W. B., Berg, R. H., Schwintzer, C. R. & Tjepkema, J. D. in *Nitrogen-fixing*
839 *Actinorhizal Symbioses* (eds. Pawlowski, K. & Newton, W. E.) 105–146 (Springer
840 Netherlands, 2007).
- 841 64. Silvester, Warwick B., Harris Sharon L., Tjepkema, John D. in *The Biology of Frankia*
842 *and Actinorhizal Plants* (ed. Schwintzer, Christa R. , , Tjepkema, John D) 157–176
843 (Academic Press, 2012).
- 844 65. Rogers, C. & Oldroyd, G. E. D. Synthetic biology approaches to engineering the nitrogen
845 symbiosis in cereals. *J. Exp. Bot.* **65**, 1939–1946 (2014).
- 846 66. Untergasser, A. *et al.* One-step *Agrobacterium* mediated transformation of eight genes
847 essential for rhizobium symbiotic signaling using the novel binary vector system pHUGE.
848 *PLoS One* **7**, e47885 (2012).
- 849 67. Cao, Q., den Camp, R. O., Kalhor, M. S., Bisseling, T. & Geurts, R. Efficiency of
850 *Agrobacterium rhizogenes*–mediated root transformation of *Parasponia* and *Trema* is
851 temperature dependent. *Plant Growth Regul.* **68**, 459–465 (2012).
- 852 68. Davey, M. R. *et al.* Effective nodulation of micro-propagated shoots of the non-legume
853 *Parasponia andersonii* by *Bradyrhizobium*. *J. Exp. Bot.* **44**, 863–867 (1993).

- 854 69. Geurts, R. & de Jong, H. Fluorescent In Situ Hybridization (FISH) on pachytene
855 chromosomes as a tool for genome characterization. *Methods Mol. Biol.* **1069**, 15–24
856 (2013).
- 857 70. Merckx, V. S. F. T. *et al.* Evolution of endemism on a young tropical mountain. *Nature*
858 **524**, 347–350 (2015).
- 859 71. Becking, J. H. The *Parasponia parviflora*—*Rhizobium* symbiosis. Host specificity, growth
860 and nitrogen fixation under various conditions. *Plant Soil* **75**, 309–342 (1983).
- 861 72. Bergersen, F. J. The quantitative relationship between nitrogen fixation and the
862 acetylene-reduction assay. *Aust. Jnl. Of Bio. Sci.* **23**, 1015–1026 (1970).
- 863 73. Cristescu, S. M., Persijn, S. T., te Lintel Hekkert, S. & Harren, F. J. M. Laser-based
864 systems for trace gas detection in life sciences. *Appl. Phys. B* **92**, 343–349 (2008).
- 865 74. Hoagland, D. R., Arnon, D. I. & Others. The water-culture method for growing plants
866 without soil. *Circular. California Agricultural Experiment Station* **347**, (1950).
- 867 75. Trouvelot, A, Kough J L, Gianinazzi-Pearson V. in *Physiological and Genetic Aspects of*
868 *Mycorrhizae* (ed. Gianinazzi-Pearson, G. S., V.) 217–221 (INRA Press, 1986).
- 869 76. Huisman, R. *et al.* Haustorium formation in *Medicago truncatula* roots infected by
870 *Phytophthora palmivora* does not involve the common endosymbiotic program shared by
871 arbuscular mycorrhizal fungi and rhizobia. *Mol. Plant. Microbe. Interact.* **28**, 1271–1280
872 (2015).
- 873 77. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of
874 occurrences of k-mers. *Bioinformatics* **27**, 764–770 (2011).
- 875 78. Hare, E. E. & Johnston, J. S. Genome size determination using flow cytometry of
876 propidium iodide-stained nuclei. *Methods Mol. Biol.* **772**, 3–12 (2011).
- 877 79. Novák, P., Neumann, P., Pech, J., Steinhaisl, J. & Macas, J. RepeatExplorer: a Galaxy-
878 based web server for genome-wide characterization of eukaryotic repetitive elements
879 from next-generation sequence reads. *Bioinformatics* **29**, 792–793 (2013).
- 880 80. Aronesty, E. Comparison of sequencing utility programs. *Open Bioinforma. J.* **7**, 1–8

- 881 (2013).
- 882 81. Xu, H. *et al.* FastUniq: a fast de novo duplicates removal tool for paired short reads.
883 *PLoS One* **7**, e52249 (2012).
- 884 82. Bakker, F. T. *et al.* Herbarium genomics: plastome sequence assembly from a range of
885 herbarium specimens using an Iterative Organelle Genome Assembly pipeline. *Biol. J.*
886 *Linn. Soc. Lond.* **117**, 33–43 (2016).
- 887 83. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler
888 transform. *Bioinformatics* **25**, 1754–1760 (2009).
- 889 84. Gnerre, S. *et al.* High-quality draft assemblies of mammalian genomes from massively
890 parallel sequence data. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 1513–1518 (2011).
- 891 85. Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D. & Pirovano, W. Scaffolding pre-
892 assembled contigs using SSPACE. *Bioinformatics* **27**, 578–579 (2011).
- 893 86. Parra, G., Bradnam, K. & Korf, I. CEGMA: a pipeline to accurately annotate core genes
894 in eukaryotic genomes. *Bioinformatics* **23**, 1061–1067 (2007).
- 895 87. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M.
896 BUSCO: assessing genome assembly and annotation completeness with single-copy
897 orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
- 898 88. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements
899 in eukaryotic genomes. *Mob. DNA* **6**, 11 (2015).
- 900 89. Gremme, G., Steinbiss, S. & Kurtz, S. GenomeTools: a comprehensive software library
901 for efficient processing of structured genome annotations. *IEEE/ACM Trans. Comput.*
902 *Biol. Bioinform.* **10**, 645–656 (2013).
- 903 90. Han, Y. & Wessler, S. R. MITE-Hunter: a program for discovering miniature inverted-
904 repeat transposable elements from genomic sequences. *Nucleic Acids Res.* **38**, e199
905 (2010).
- 906 91. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a
907 reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).

- 908 92. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals
909 unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.*
910 **28**, 511–515 (2010).
- 911 93. Haas, B. J. *et al.* De novo transcript sequence reconstruction from RNA-seq using the
912 Trinity platform for reference generation and analysis. *Nat. Protoc.* **8**, 1494–1512 (2013).
- 913 94. UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Res.* **43**,
914 D204–12 (2015).
- 915 95. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59 (2004).
- 916 96. Stanke, M., Diekhans, M., Baertsch, R. & Haussler, D. Using native and syntenically
917 mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **24**, 637–644
918 (2008).
- 919 97. Hoff, K. J., Lange, S., Lomsadze, A., Borodovsky, M. & Stanke, M. BRAKER1:
920 Unsupervised RNA-seq-based genome annotation with GeneMark-ET and AUGUSTUS.
921 *Bioinformatics* **32**, 767–769 (2016).
- 922 98. Lomsadze, A., Burns, P. D. & Borodovsky, M. Integration of mapped RNA-Seq reads
923 into automatic training of eukaryotic gene finding algorithm. *Nucleic Acids Res.* **42**, e119
924 (2014).
- 925 99. Campbell, M. S. *et al.* MAKER-P: a tool kit for the rapid creation, management, and
926 quality control of plant genome annotations. *Plant Physiol.* **164**, 513–524 (2014).
- 927 100. Haas, B. J. *et al.* Automated eukaryotic gene structure annotation using
928 EVIDENCEModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**,
929 R7 (2008).
- 930 101. Kurtz, S. *et al.* Versatile and open software for comparing large genomes. *Genome Biol.*
931 **5**, R12 (2004).
- 932 102. Otto, T. D., Dillon, G. P., Degraeve, W. S. & Berriman, M. RATT: Rapid Annotation
933 Transfer Tool. *Nucleic Acids Res.* **39**, e57 (2011).
- 934 103. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7:

- 935 improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
- 936 104. Ronquist, F. & Huelsenbeck, J. P. MrBayes 3: Bayesian phylogenetic inference under
937 mixed models. *Bioinformatics* **19**, 1572–1574 (2003).
- 938 105. Lanfear, R., Frandsen, P. B., Wright, A. M., Senfeld, T. & Calcott, B. PartitionFinder 2:
939 New methods for selecting partitioned models of evolution for molecular and
940 morphological phylogenetic analyses. *Mol. Biol. Evol.* **34**, 772–773 (2016).
- 941 106. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of
942 large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
- 943 107. Emms, D. M. & Kelly, S. OrthoFinder: solving fundamental biases in whole genome
944 comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* **16**,
945 157 (2015).
- 946 108. Swarbreck, D. *et al.* The Arabidopsis Information Resource (TAIR): gene structure and
947 function annotation. *Nucleic Acids Res.* **36**, D1009–14 (2008).
- 948 109. Myburg, A. A. *et al.* The genome of *Eucalyptus grandis*. *Nature* **510**, 356–362 (2014).
- 949 110. Tuskan, G. A. *et al.* The Genome of Black Cottonwood, *Populus trichocarpa* (Torr. &
950 Gray). *Science* **313**, 1596–1604 (2006).
- 951 111. Young, N. D. *et al.* The *Medicago* genome provides insight into the evolution of rhizobial
952 symbioses. *Nature* **480**, 520–524 (2011).
- 953 112. Schmutz, J. *et al.* Genome sequence of the palaeopolyploid soybean. *Nature* **463**, 178–
954 183 (2010).
- 955 113. Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing
956 phylogenetic trees. *Mol. Biol. Evol.* **4**, 406–425 (1987).
- 957 114. Valdar, W. S. J. Scoring residue conservation. *Proteins* **48**, 227–241 (2002).
- 958 115. Peng, Y., Leung, H. C. M., Yiu, S. M. & Chin, F. Y. L. IDBA-UD: a de novo assembler for
959 single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics*
960 **28**, 1420–1428 (2012).
- 961 116. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory

962 requirements. *Nat. Methods* **12**, 357–360 (2015).

963 117. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for

964 assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).

965 **Acknowledgments**

966 This work was supported by NWO-NSFC Joint Research project (846.11.005) to WCY, TB
967 and RG, NWO-VICI (865.13.001) to RG, NWO-VENI (863.15.010) to WK, the European
968 Research Council (ERC-2011-AdG294790) to TB and China Scholarship Councils
969 (201303250067) to FB and (201306040120) to DS. We thank Shelley James and Giles
970 Oldroyd for providing germplasm.

971
972 All custom scripts and code are available on https://github.com/holmrenser/parasponia_code.
973 The data reported in this paper are tabulated in the Supplementary Materials and archived at
974 NCBI under BioProject numbers PRJNA272473 and PRJNA272482. All analyzed data can
975 be browsed or downloaded through a WebPortal on www.parasponia.org. [For reviewing
976 purposes, an account is available with username *reviewer* and password *D7yGNEkNv25e*].

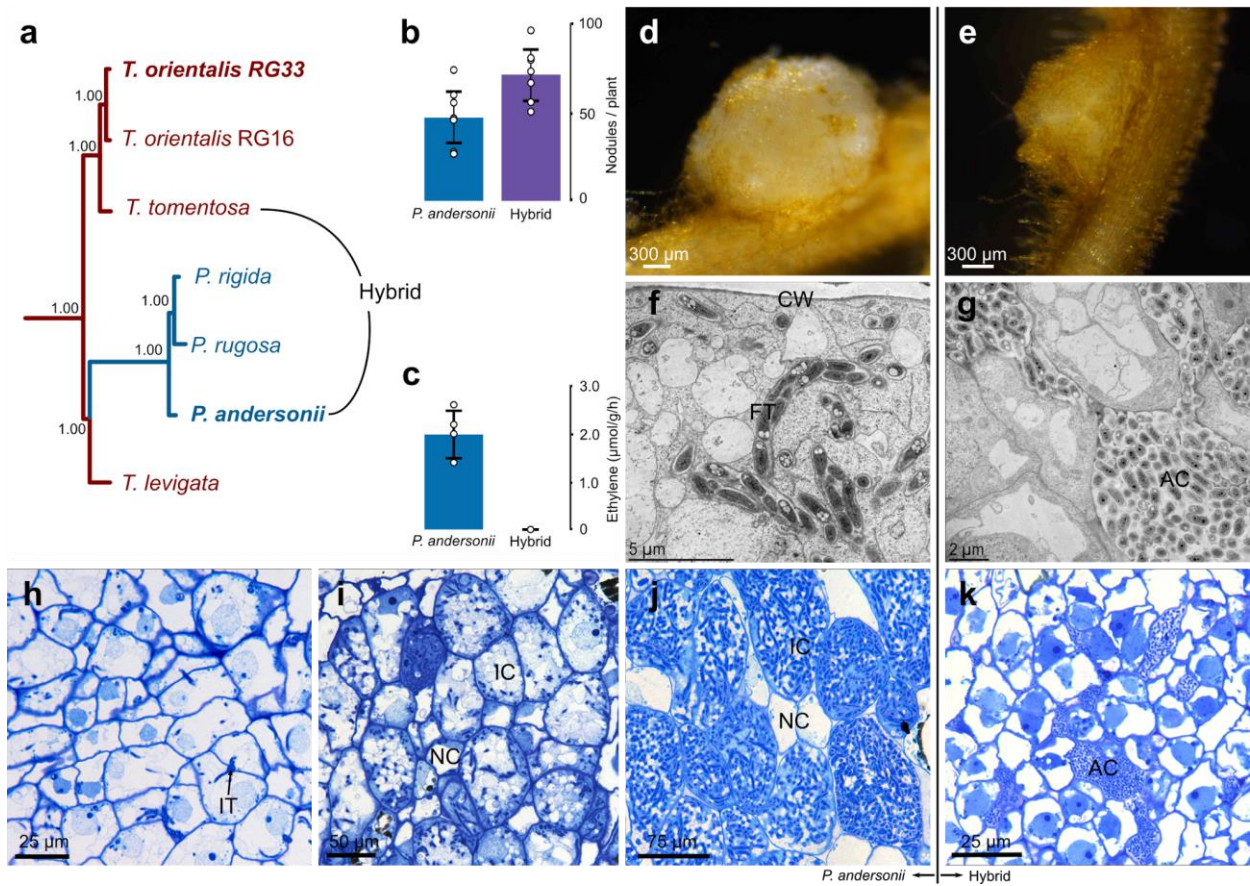
977

978 **Contributions**

979 This research was led by RG, who together with TB conceived the project. *Trema orientalis*
980 accessions, including rhizosphere samples, were collected in Sabah Parks (Malaysian
981 Borneo) by RG in an expedition organised by MS and RR. FACS studies to estimate genome
982 sizes were done by FB and RHe. Plant propagations and tissue isolations were done by FB,
983 WL, QC, TS, DS, YR, MH, WY and RG. Arbuscular mycorrhiza assays were done by TS, YR
984 and WK. Studies on hybrid plants were done by FB, QC, DJvdH and EF, and ARA assays by
985 FB and EF. Light and electron microscopy studies were conducted by FB and EF. DNA and
986 RNA was isolated by JV, JH and WL, and sequencing was done by ES. Chloroplast analysis
987 was conducted by RvV, RHo and BG, Bioinformatic analyses were done by RvV, RHo, LS,
988 JJ and SS, and manual curations by RvV, RHo, LR, AvZ, TAKW, JJ, KM, WK, and RG. RvV,
989 RHo, LR, MES, TB, SS, and RG wrote the manuscript.

990

991



992

993

994 **Figure 1: Nodulation phenotype of *Parasponia* and interspecific *Parasponia* x *Trema* F₁**

995 **hybrid plants. (a)** Phylogenetic reconstruction based on whole chloroplast of *Parasponia*

996 and *Trema*. The *Parasponia* lineage (marked blue) is embedded in the *Trema* genus

997 (marked red). Species selected for interspecific crosses are indicated, species used for

998 reference genome assembly are in bold. Node labels indicate posterior probabilities. (b)

999 Mean number of nodules on roots of *P. andersonii* and *P. andersonii* x *T. tomentosa* F₁

1000 hybrid plants (n=7). (c) Mean nitrogenase activity in acetylene reductase assay of *P.*

1001 *andersonii* and *P. andersonii* x *T. tomentosa* F₁ hybrid nodules (n=4). Barplot error bars

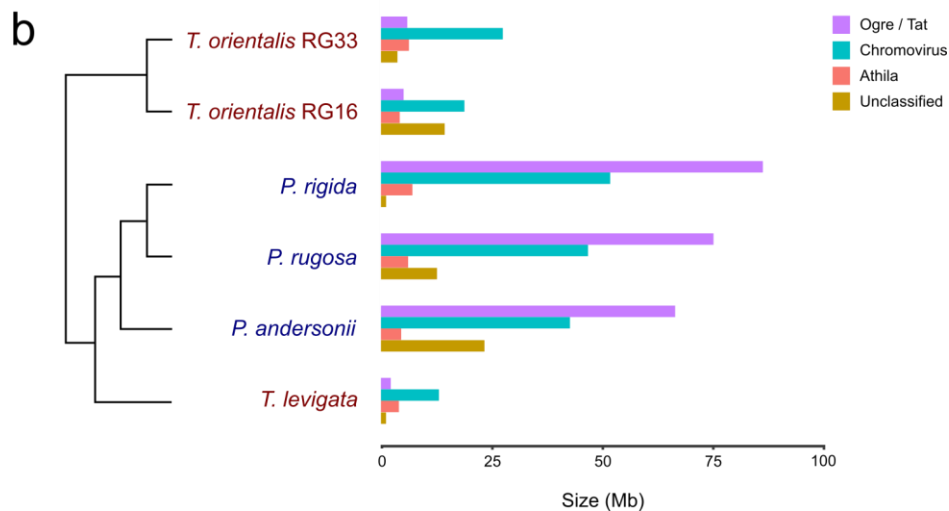
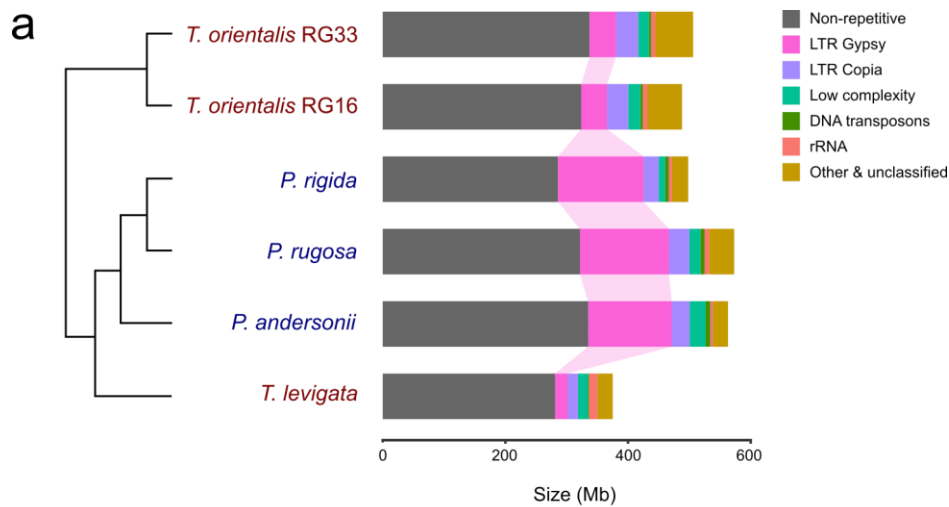
1002 indicate standard deviations; dots represent individual measurements (d) *P. andersonii*

1003 nodule. (e) *P. andersonii* x *T. tomentosa* F₁ hybrid nodule. (f,g) Ultrastructure of nodule

1004 tissue of *P. andersonii* (f) and F₁ hybrid (g). Note the intracellular fixation thread (FT) in the

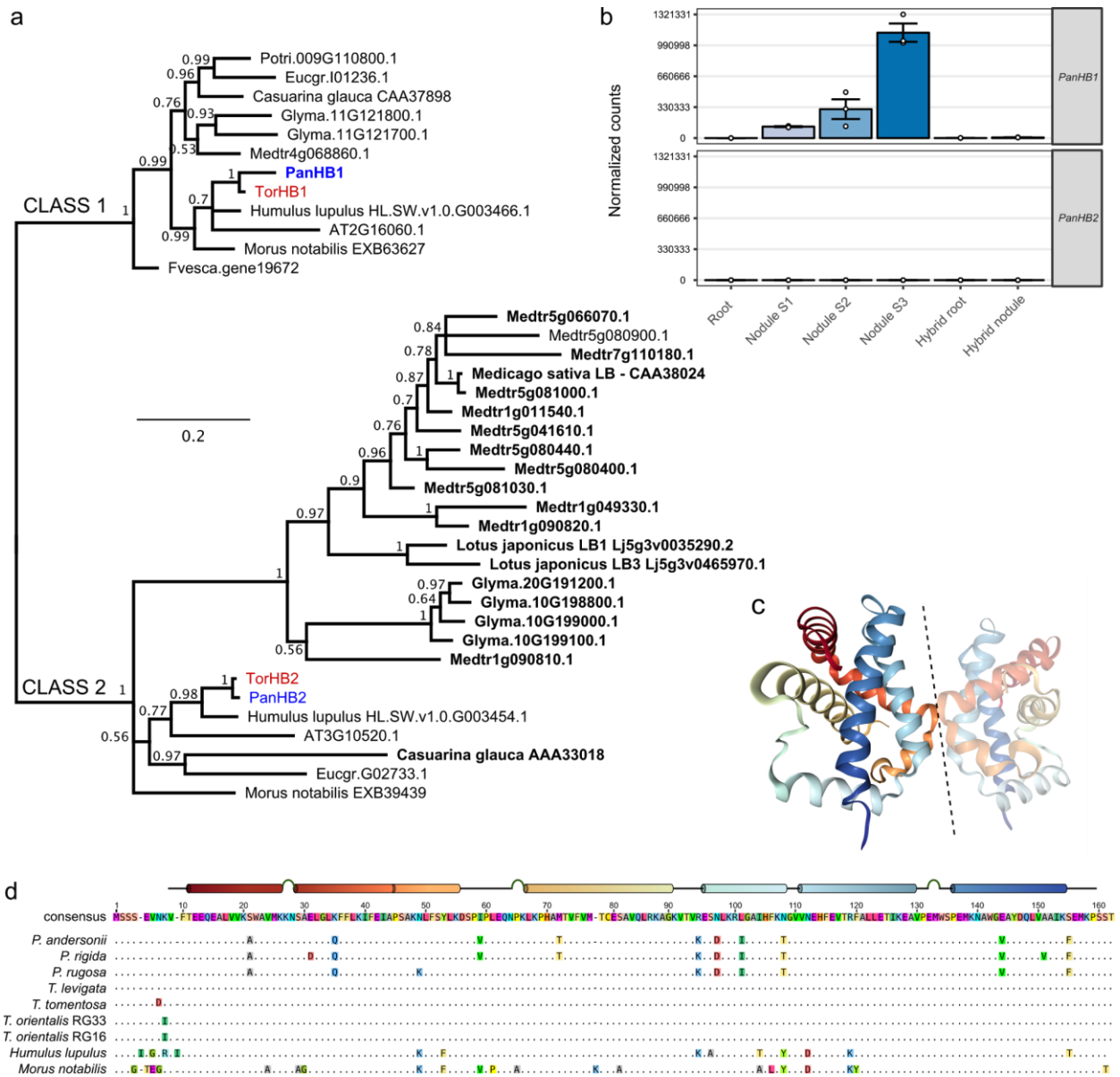
1005 cell of *P. andersonii* in comparison with the extracellular, apoplastic colonies of rhizobia (AC)
1006 in the hybrid nodule. **(h-i)** Light microscopy images of *P. andersonii* nodules in three
1007 subsequent developmental stages. **(h)** Stage 1: initial stages of colonization when infection
1008 threads (IT) enter the host cells. **(i)** Stage 2: progression of rhizobium infection in nodule host
1009 cell, **(j)** Stage 3: nodule cells completely filled with fixation threads. Note difference in size
1010 between the infected (IC) and non-infected cells (NC). **(k)** Light microscopy image of F1
1011 hybrid nodule cells. Note rhizobium colonies in apoplast, surrounding the host cells (AC).
1012 Nodules have been analysed 6 weeks post inoculation with *Mesorhizobium plurifarum*
1013 BOR2. Abbreviations: FT: fixation thread, CW: cell wall, AC: apoplastic colony of rhizobia, IT:
1014 infection threads, IC: infected cell, NC: non-infected cell.

1015



1016

1017 **Figure 2: *Parasponia* and *Trema* genome structure.** Estimated genome sizes and
 1018 fractions of different classes of repeats as detected by RepeatExplorer, calibrated using k-
 1019 mer based genome size estimates. (a) Total genome sizes and fractions of major repeat
 1020 classes showing 1) a conserved size of around 300 Mb of non-repetitive sequence, and 2) a
 1021 large expansion of gypsy-type LTR retrotransposons in all *Parasponia* compared with all
 1022 *Trema* species. (b) Estimated size of gypsy-type LTR subclasses in *Parasponia* and *Trema*
 1023 showing that expansion of this class was mainly due to a tenfold increase of Ogre/Tat to
 1024 around 75Mb in *Parasponia*.



1025

1026

1027 **Figure 3: *Parasponia*-specific adaptations in class 1 hemoglobin protein HB1. (a)**

1028 Phylogenetic reconstruction of class 1 (OG0010523) and class 2 hemoglobins (OG0002188).

1029 Symbiotic hemoglobins are marked in bold; legumes and the actinorhizal plant *casuarina*

1030 have recruited class 2 hemoglobins for balancing oxygen levels in their nodules. Conversely,

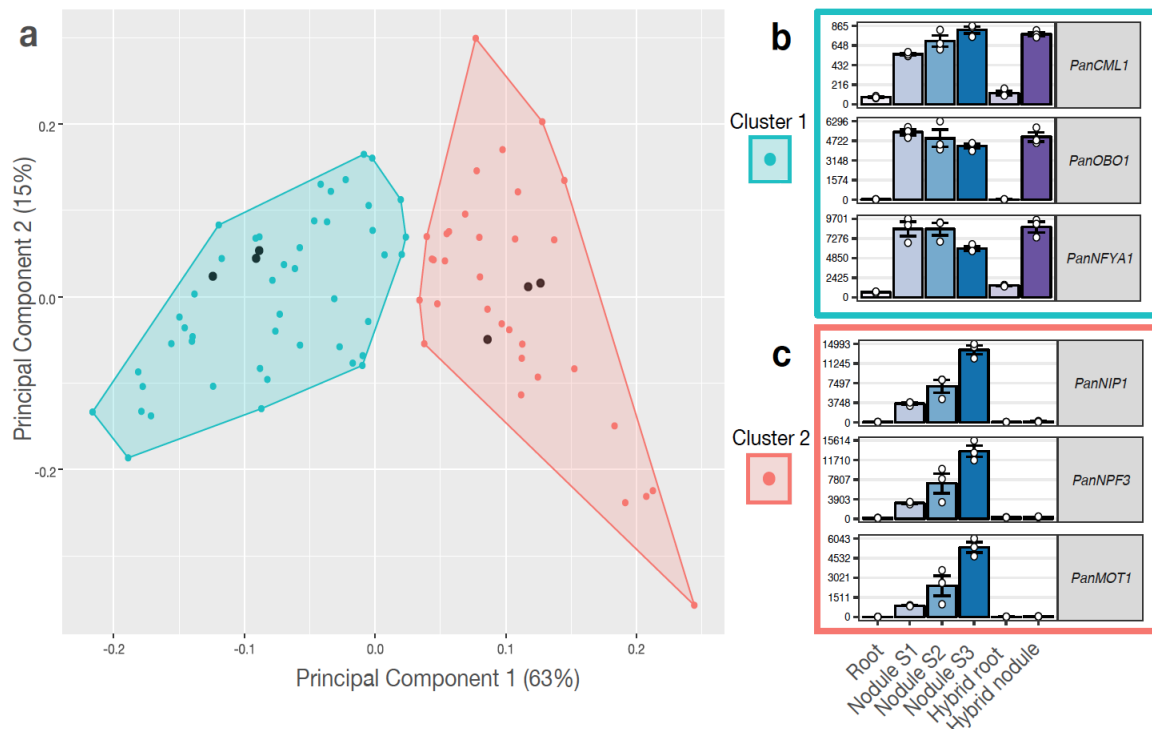
1031 *Parasponia* has recruited a class 1 hemoglobin *HB1* confirming parallel evolution of

1032 symbiotic oxygen transport in this lineage. *Medicago truncatula* (Medtr); *Glycine max*

1033 (*Glyma*), *Populus trichocarpa* (*Potri*); *Fragaria vesca* (*Fvesca*); *Eucalyptus grandis* (*Eugr*);

1034 *Arabidopsis thaliana* (*AT*). Node values indicate posterior probabilities; Scale bar represents

1035 substitutions per site. *Parasponia* marked in blue, *Trema* in red. **(b)** Expression profile of
1036 *PanHB1* and *PanHB2* in *P. andersonii* roots, stage 1-3 nodules, and in *P. andersonii* x *T.*
1037 *tomentosa* F₁ hybrid roots and nodules (line H9). Expression is given in DESeq2 normalized
1038 read counts, error bars represent standard error of three biological replicates, dots represent
1039 individual expression levels. **(c)** Crystal structure of the asymmetric dimer of PanHB1 as
1040 deduced by Kakar *et al.* 2011³⁶. Dashed line separates the two units. **(d)** Protein sequence
1041 alignment of class 1 hemoglobins from *Parasponia* spp., *Trema* spp., *Humulus lupulus*, and
1042 *Morus notabilis*. Only amino acids that differ from the consensus are drawn. A linear model of
1043 the crystal structure showing alpha helices and turns is depicted above the consensus
1044 sequence. There are seven amino acids that consistently differ between all *Parasponia* and
1045 all *Trema* species we sampled: Ala(21), Gln(35), Asp(97), Ile(101), Thr(108), Val(144), and
1046 Phe(155). These differences therefore correlate with the functional divergence between *P.*
1047 *andersonii* PanHB1 and *T. tomentosa* *TtoHB1*^{35,36}. All seven consistently different sites are
1048 identical for all sampled *Trema* species, and five are identical for *Trema*, *Humulus*, and
1049 *Morus*; at both the amino acid and nucleotide level. This shows that these sites are
1050 conserved in all species except *Parasponia* and therefore supports adaptation of HB1 in the
1051 common ancestor of *Parasponia*. This suggests that the ancestral form of HB1 had oxygen
1052 affinities and kinetics that were not adapted to rhizobium symbiosis.



1053

1054 **Figure 4: Clustering of commonly utilized symbiosis genes based on expression**

1055 **profile. (a)** Principal component analysis plot of the expression profile of 85 commonly

1056 utilized symbiosis genes in 18 transcriptome samples: *P. andersonii* roots and nodules

1057 (stage 1-3), hybrid roots and nodules (line H9). All samples have three biological replicates.

1058 First two components are shown, representing 78% of the variation in all samples. Colors

1059 indicate clusters (K-means clustering using pearson correlation as distance measure, k=2) of

1060 genes with similar expression patterns. The three genes with the highest pearson correlation

1061 to the cluster centroids are indicated as black dots. **(b-c)** Expression profiles of

1062 representative genes for each cluster. **(b)** Cluster 1 represents genes related to

1063 organogenesis: these genes are upregulated in both *P. andersonii* and hybrid nodules. **(c)**

1064 Cluster 2 represents genes related to infection and fixation: these genes are highly

1065 upregulated in *P. andersonii* nodules, but do not respond in the hybrid nodule. *PanCML1*: *P.*

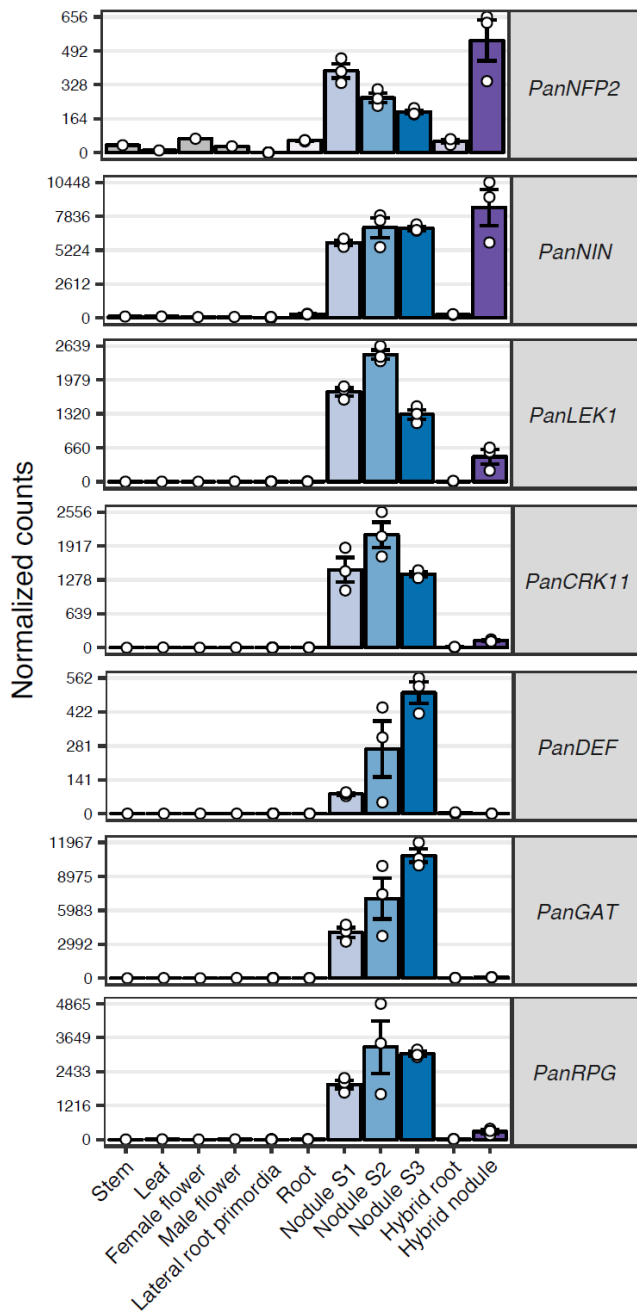
1066 *andersonii* CALMODULIN 1; *PanOBO1*: *P. andersonii* ORGAN BOUNDARY-LIKE 1;

1067 *PanNFYA1*: *P. andersonii* NUCLEAR TRANSCRIPTION FACTOR-YA 1; *PanNIP*: *P.*

1068 *andersonii* AQUAPORIN NIP NODULIN26-LIKE; *PanNPF3*: *P. andersonii*

1069 NITRATE/PEPTIDE TRANSPORTER FAMILY 3; *PanMOT1*: *P. andersonii* MOLYBDATE

1070 *TRANSPORTER 1.*



1071

1072

1073 **Figure 5: Expression profile of *Parasponia* symbiosis genes that are lost in *Trema***

1074 **species.** Expression of symbiosis genes in *P. andersonii* stem, leaf, female and male

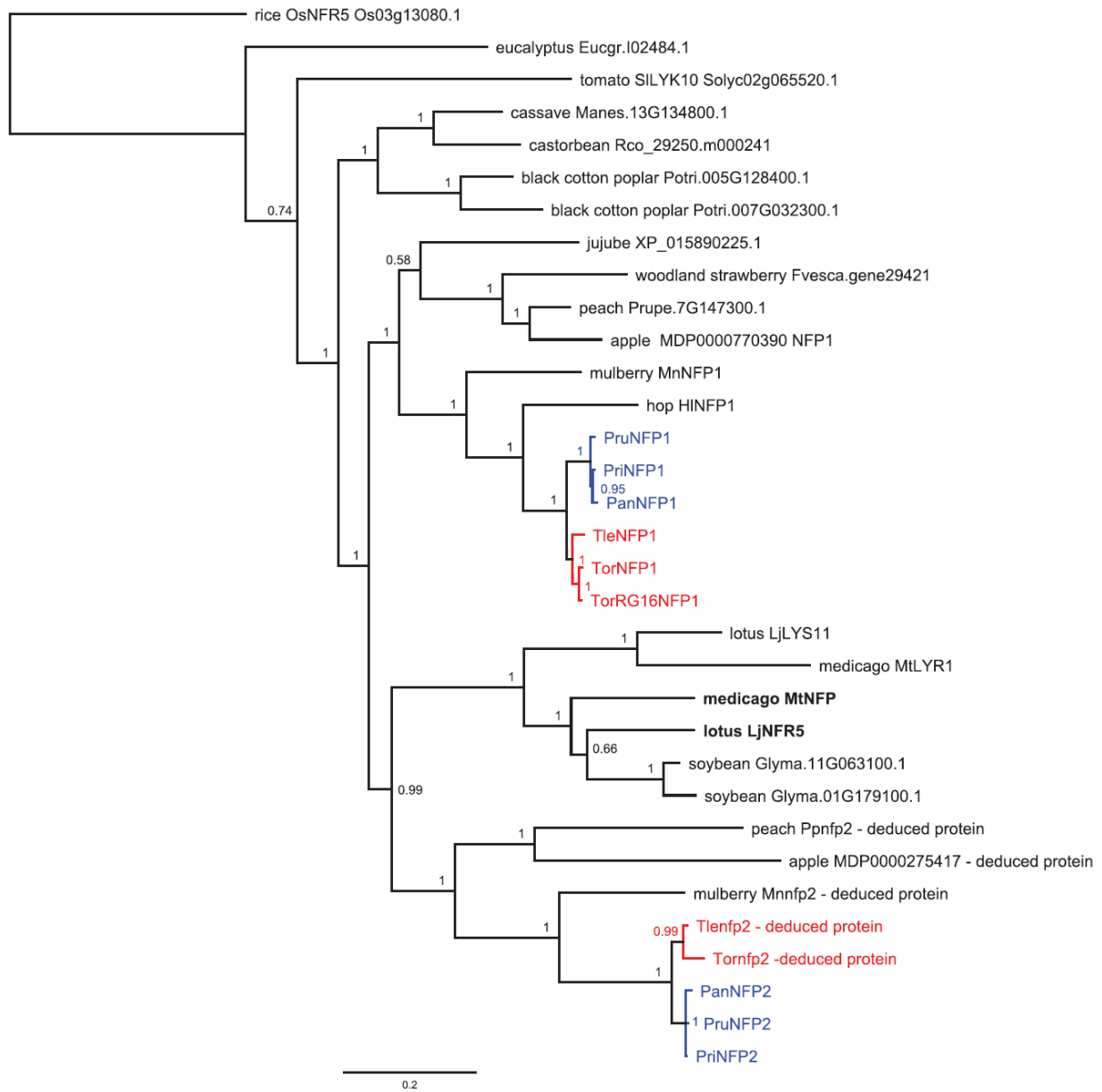
1075 flowers, lateral root primordia, roots and 3 nodule stages (S1-3), and in *P. andersonii* x *T.*

1076 *tomentosa* F₁ hybrid roots and nodules (line H9). Expression is given in DESeq2 normalized

1077 read counts, error bars represent standard error of three biological replicates for lateral root

1078 primordia, root, and nodule samples. Dots represent individual expression levels. *PanNFP2*:

1079 *P. andersonii* NOD FACTOR PERCEPTION 2, PanNIN: *P. andersonii* NODULE
1080 INCEPTION, PanLEK1: *P. andersonii* LECTIN RECEPTOR KINASE 1, PanCRK11: *P.*
1081 *andersonii* CYSTEINE-RICH RECEPTOR KINASE 11, PanDEF1: *P. andersonii* DEFENSIN
1082 1; PanRPG: *P. andersonii* RHIZOBIUM DIRECTED POLAR GROWTH.



1083

1084

1085 **Figure 6: *Parasponia* NFP1 and NFP2 are homologous to legume LCO receptors**

1086 **MtNFP/LjNFR5.** Phylogenetic reconstruction of the NFP/NFR5 orthogroup. NFP2 protein

1087 sequences of *T. levigata*, *T. orientalis*, *Morus notabilis*, *Malus x domestica* and *Prunus*

1088 *persica* have been deduced from pseudogenes. Included species: *Parasponia andersonii*

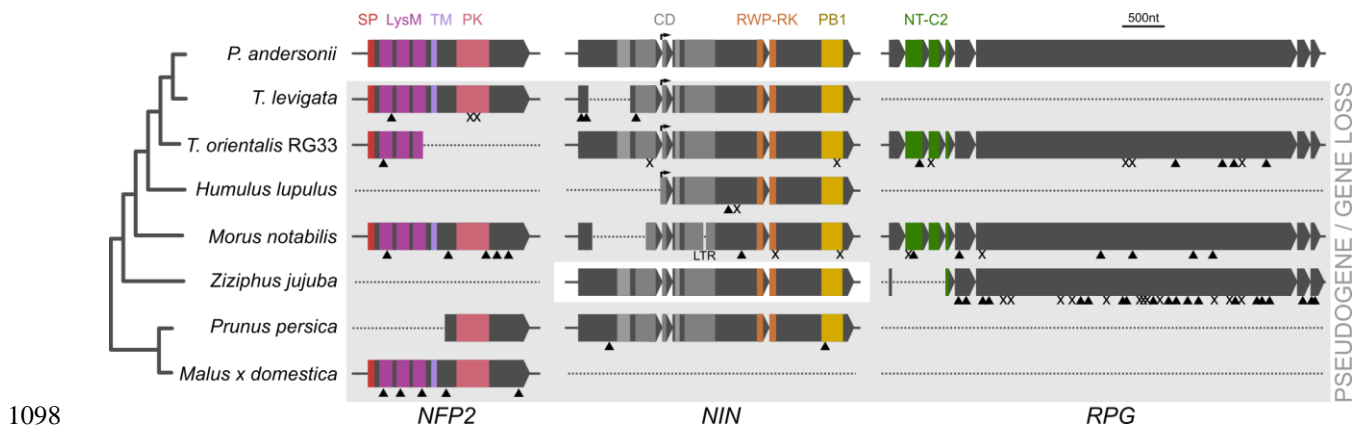
1089 (*Pan*); *Parasponia rigida* (*Pri*); *Parasponia rugosa* (*Pru*); *Trema orientalis* RG33 (*Tor*); *Trema*

1090 *orientalis* RG16 (*TorRG16*); *Trema levigata* (*Tle*); medicago (*Medicago truncatula*, *Mt*); lotus

1091 (*Lotus japonicus*, *Lj*); soybean (*Glycine max*, *Glyma*); peach (*Prunus persica*, *Ppe*);

1092 woodland strawberry (*Fragaria vesca*, *Fvesca*); black cotton poplar (*Populus trichocarpa*,

1093 Potri); eucalyptus (*Eucalyptus grandis*, Eogr); jujube (*Ziziphus jujube*), apple (*Malus x*
1094 *domestica*), mulberry (*Morus notabilis*), hops (*Humulus lupulus* (natsume.shinsuwase.v1.0)),
1095 cassave (*Manihot esculenta*), rice (*Oryza sativa*), tomato (*Solanum lycopersicum*), castor
1096 bean (*Ricinus communis*). Node numbers indicate posterior probabilities, scale bar
1097 represents substitutions per site. *Parasponia* proteins are marked in blue, *Trema* in red.



1098

1099

1100 **Figure 7: Parallel loss of symbiosis genes in non-nodulating Rosales species.**

1101 Pseudogenization or loss of *NOD FACTOR PERCEPTION 2* (*NFP2*), *NODULE INCEPTION*

1102 (*NIN*) and *RHIZOBIUM-DIRECTED POLAR GROWTH* (*RPG*) in two phylogenetically

1103 independent *Trema* lineages, *Humulus lupulus*, *Morus notabilis*, *Prunus persica*, and *Malus x*

1104 *domestica*. In *Ziziphus jujuba* *NFP2* is lost and *RPG* is pseudogenized, but *NIN* is intact. In

1105 *Fragaria vesca* all three genes are lost (not shown). Introns are indicated but not scaled.

1106 Triangles indicate frame-shifts; X indicate premature stop codons; LTR indicates long

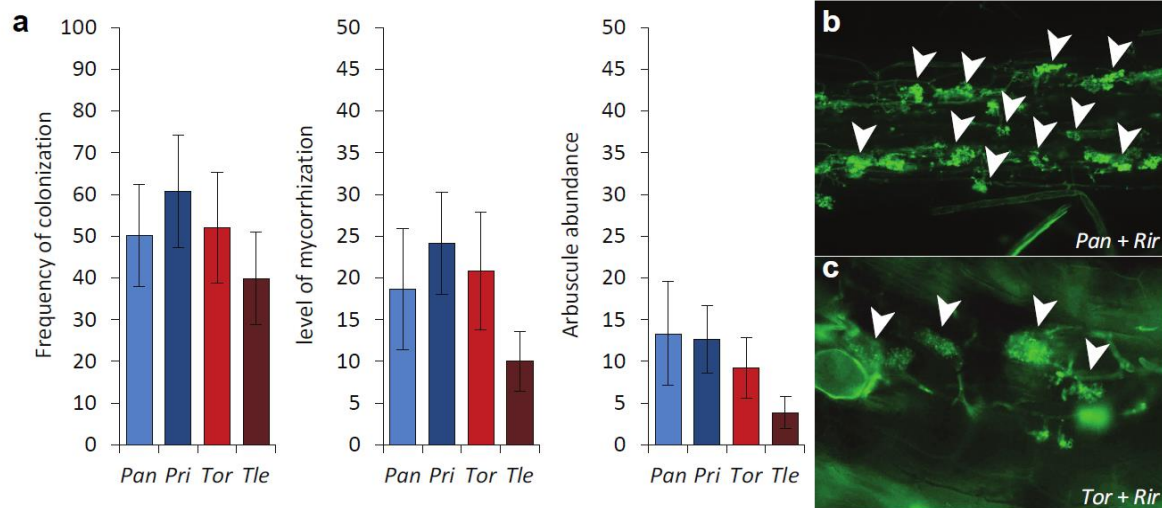
1107 terminal repeat retrotransposon insertion (not scaled); arrows indicate alternative

1108 transcriptional start site in *NIN*. SP = signal peptide (red); LysM: 3 Lysin Motif domains

1109 (magenta); TM = transmembrane domain (lilac); PK = protein kinase (pink); CD = 4

1110 conserved domains (grey); RWP-RK: conserved amino acid domain (orange); PB1 = Phox

1111 and Bem1 domain (yellow); NT-C2 = N-terminal C2 domain (green).

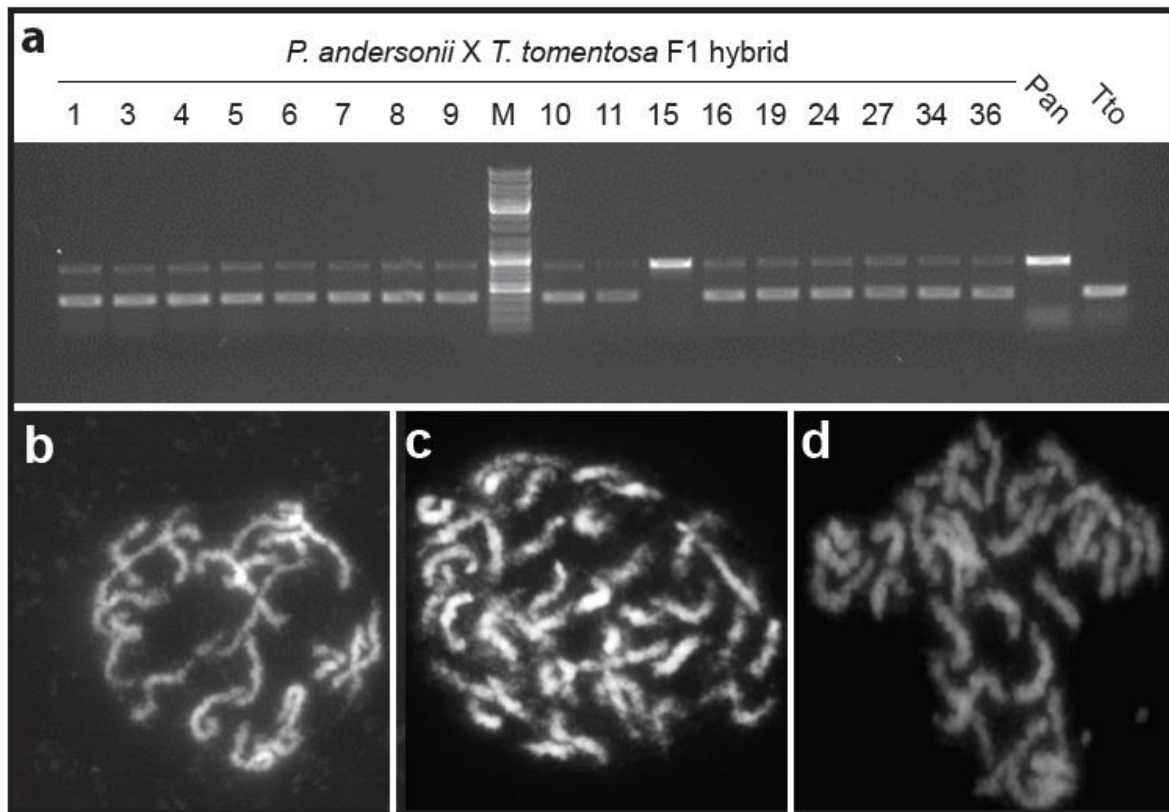


1112

1113 **Supplementary Figure 1: Arbuscular mycorrhization of *Parasponia* and *Trema* species.**

1114 **(a)** Mycorrhization efficiency of *Parasponia andersonii* WU01.14 (Pan), *Parasponia rigida*
1115 WU20 (Pri), *Trema orientalis* RG33 (Tor) and *Trema levigata* WU50 (Tle), 6 weeks post
1116 inoculation with *Rhizophagus irregularis* (*Rir*, n=10, error bars denote standard errors). **(b, c)**
1117 Confocal image of WGA-Alexafluor 488-stained arbuscules in root segment of either *P.*
1118 *andersonii* (Pan) **(b)** or *T. orientalis* (Tor) **(c)**.

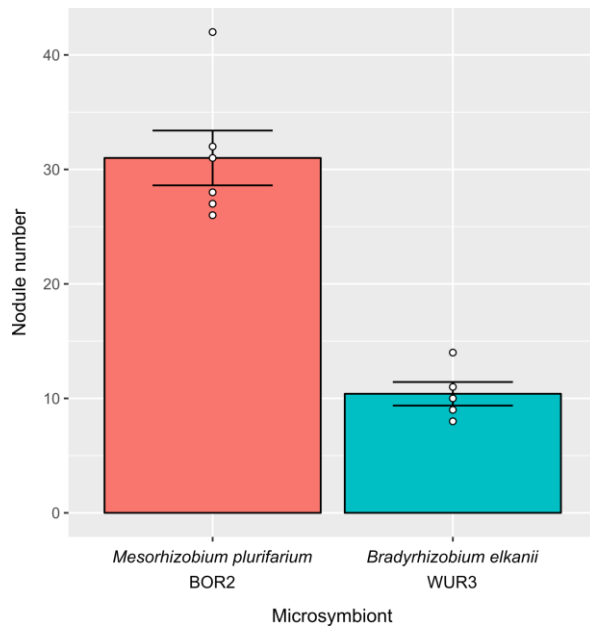
1119



1120

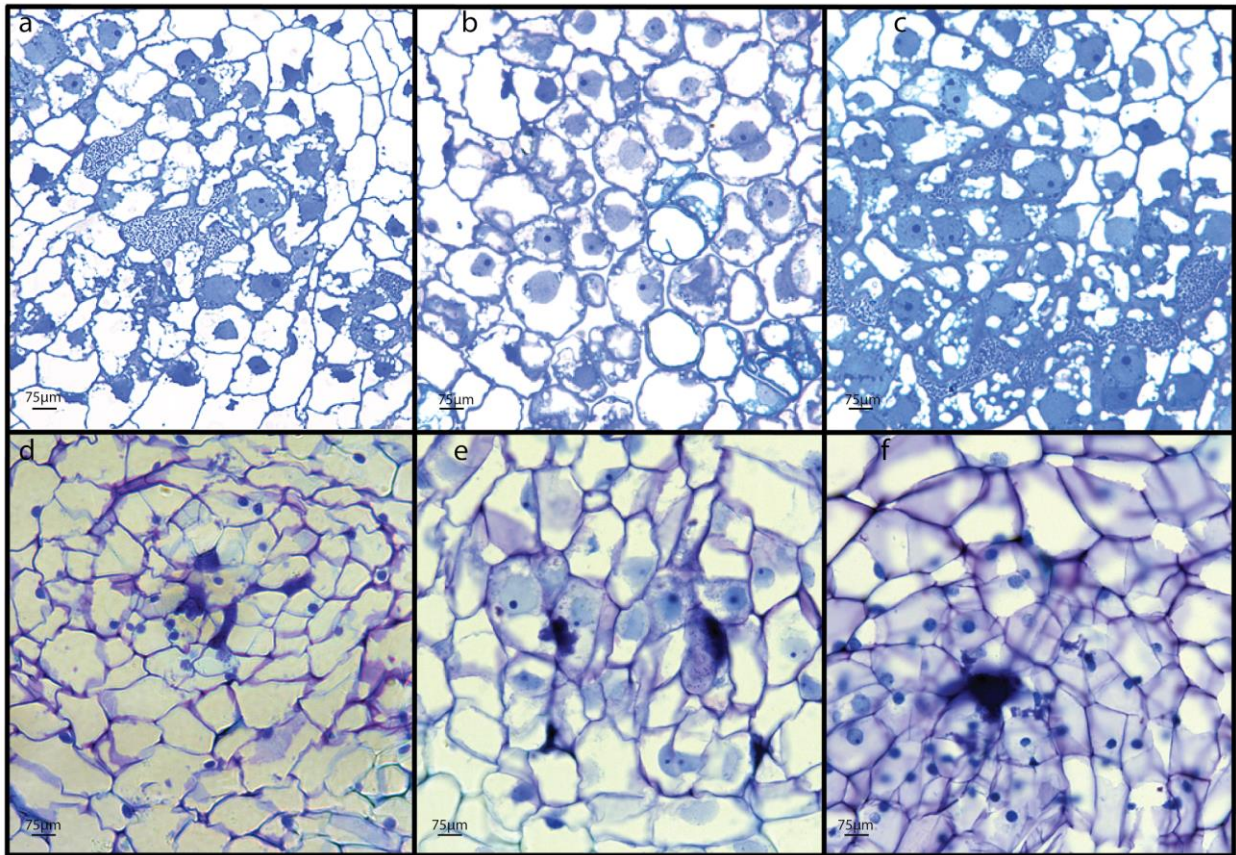
1121

1122 **Supplementary Figure 2: Genotyping of *Parasponia andersonii* x *Trema tomentosa* F₁**
1123 **hybrid plants.** (a) Genotyping of 17 putative F₁ hybrid plants of the cross *P. andersonii* (Pan)
1124 x *T. tomentosa* (Tto) using amplified length polymorphism due to an indel in the *LAX1*
1125 promoter. M: generuler DNA ladder mix (Fermentas). Hybrid plants 4, 8, 9, 16, 19 and 36
1126 were used for further experiments. (b-d) Mitotic metaphase chromosome complement of *P.*
1127 *andersonii* (2n=2x=20) (b), *T. tomentosa* (2n=4x=40) (c), and *P. andersonii* x *T. tomentosa*
1128 F1 hybrid (2n=3x=30) (d).



1129

1130 **Supplementary Figure 3: Nodulation efficiency of *Parasponia andersonii*.** Mean number
1131 of nodules on roots of *P. andersonii* inoculated with either *Mesorhizobium plurifarium* BOR2
1132 (n=6) or *Bradyrhizobium elkanii* WUR3 microsymbionts (n=5) (6 weeks post inoculation).
1133 Dots represent individual measurements.

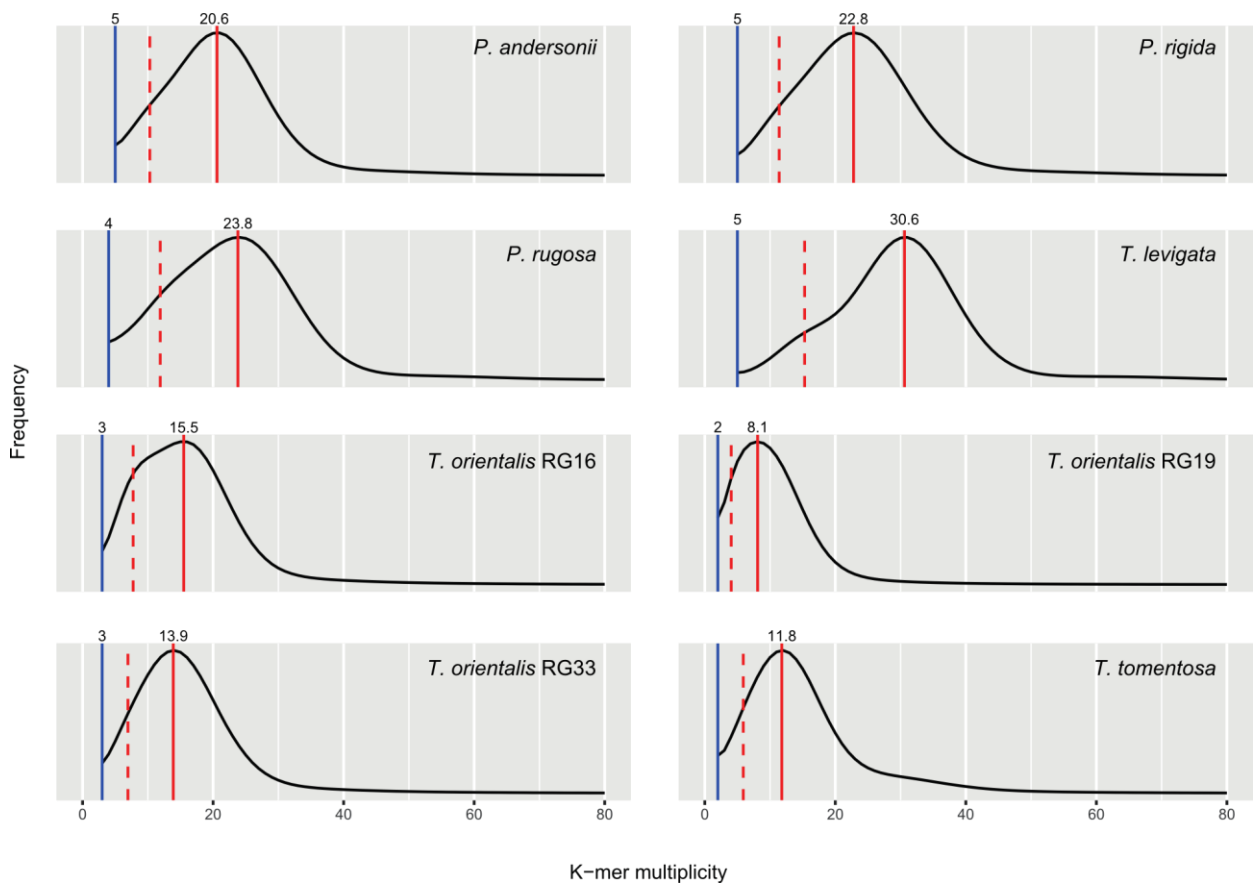


1134

1135 **Supplementary Figure 4: Longitudinal sections of root nodules of *Parasponia***
1136 ***andersonii* x *Trema tomentosa* F1 hybrid plants.** Hybrid plants H4, H8, H9, H16, H19 and
1137 H36 were clonally propagated and inoculated and inoculated with either *Bradyrhizobium*
1138 *elkanii* WUR3 (a-c) or *Mesorhizobium plurifarium* BOR2 (d-f). (a) H4 nodule induced by *B.*
1139 *elkanii* WUR3. (b) H8 nodule induced by *B. elkanii* WUR3. (c) H9 nodule induced by *B.*
1140 *elkanii* WUR3. (d) H16 nodule induced by *M. plurifarium* BOR2. (e) H19 nodule induced by
1141 *M. plurifarium* BOR2. (f) H36 nodule induced by *M. plurifarium* BOR2. Note absence of
1142 intracellular infection in all sectioned nodules.

1143

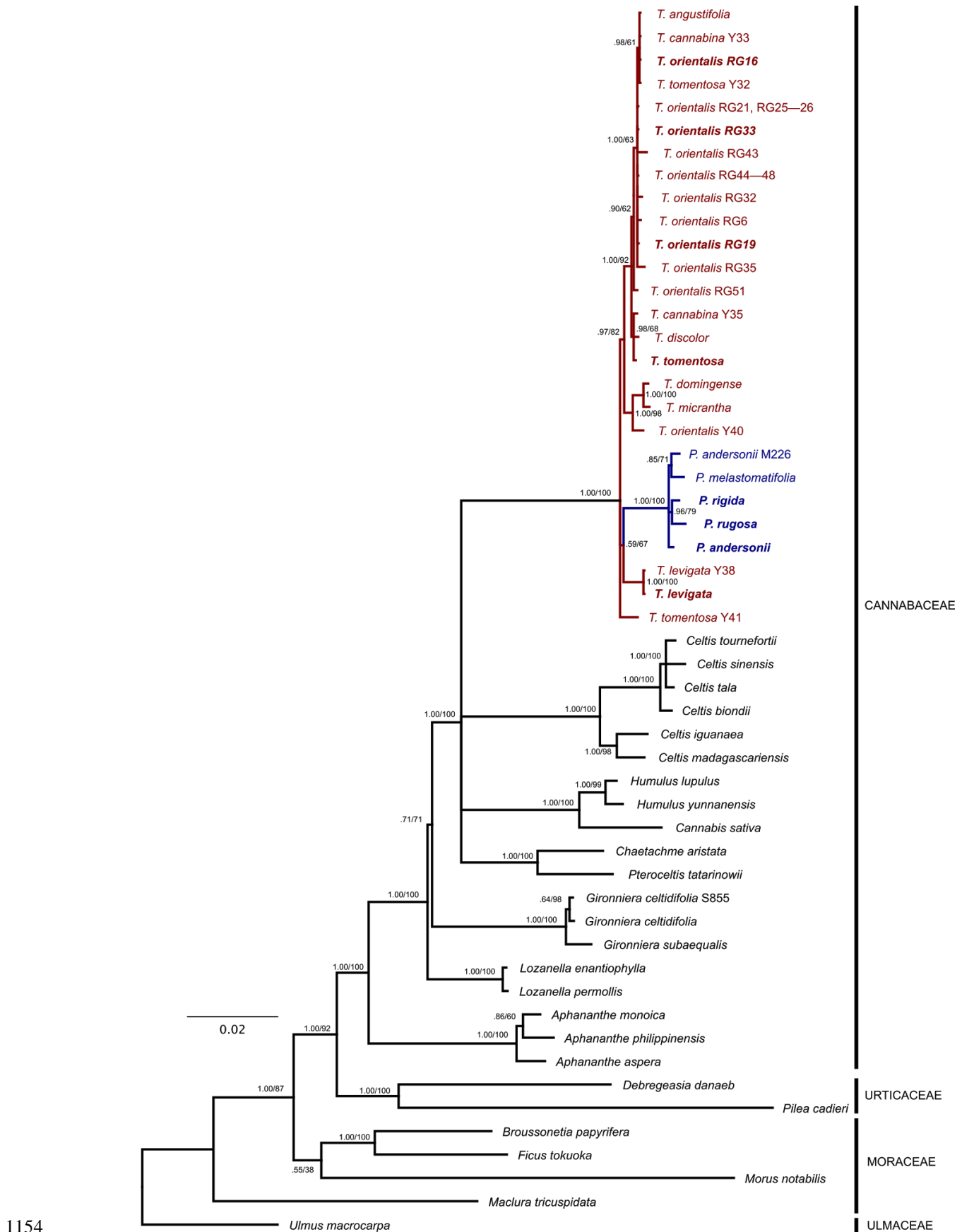
1144



1145

1146

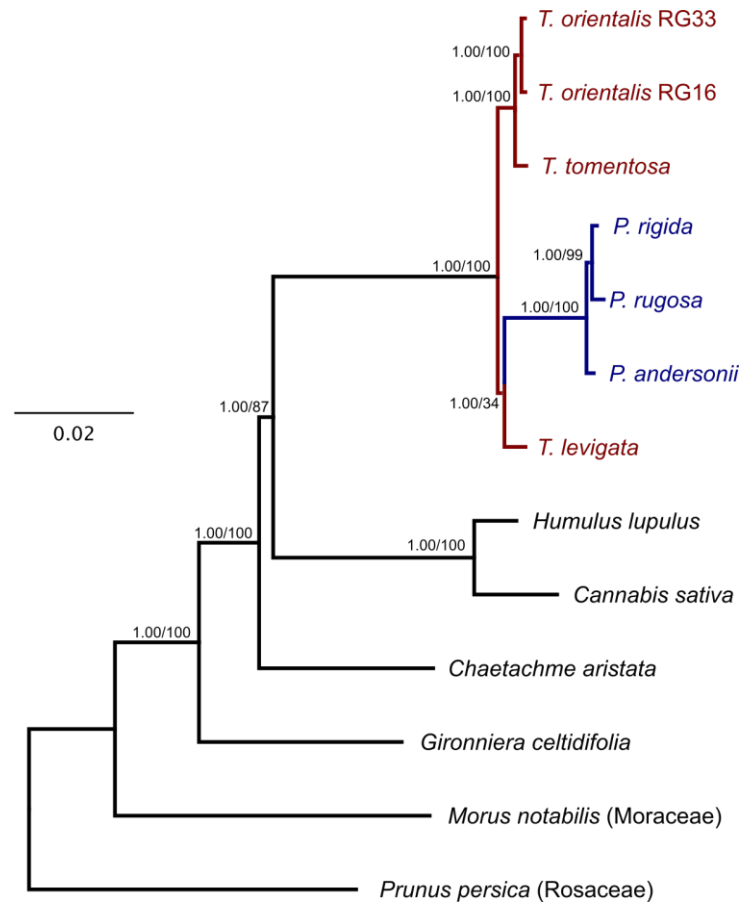
1147 **Supplementary Figure 5: Genome coverage and heterozygosity estimates based on k-**
1148 **mer analysis of *Trema* and *Parasponia* species.** Plots of 21-mer multiplicity frequencies
1149 based on jellyfish output showing that *T. levigata* and *T. orientalis* RG16 are relatively
1150 heterozygous. Solid red lines indicate estimated genome coverage corresponding to
1151 homozygous sequence; dashed red lines indicate half the estimated genome coverage
1152 corresponding to heterozygous sequence; blue lines indicate estimated error multiplicity
1153 threshold.



1154

1155 **Supplementary Figure 6: Phylogenetic reconstruction of the Cannabaceae based on**

1156 **combined analysis of four plastid markers.** Node values indicate posterior probability /
1157 RAxML bootstrap support; scale bar represents substitutions per site. *Parasponia* lineage is
1158 in blue, *Trema* lineages are in red. Note that sister relationship of *Parasponia* and *T. levigata*
1159 has low bootstrap support, but is independently supported by four shared sequence
1160 insertions (Supplementary Fig. 8). Accessions selected for comparative genome analysis in
1161 bold. GenBank accession numbers are in Supplementary Table 12.



1162

1163

1164 **Supplementary Figure 7: Phylogenetic reconstruction of the Cannabaceae based on**

1165 **chloroplast genomes.** Bayesian tree based on a combined analysis of eight data partitions

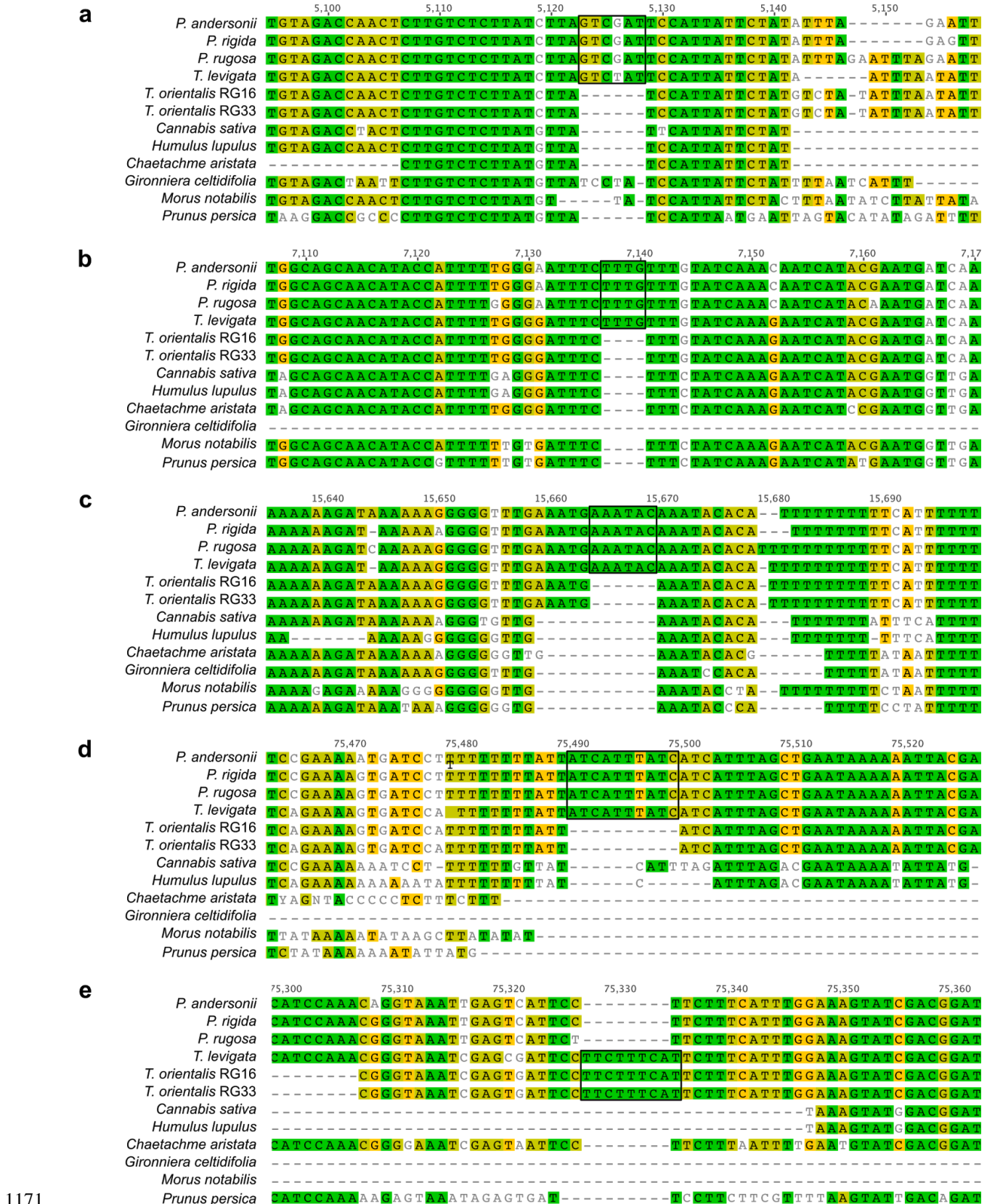
1166 (see Methods). *Parasponia* lineage is in blue, *Trema* lineages are in red. Note that sister

1167 relationship of *Parasponia* and *Trema* has low bootstrap support but is independently

1168 supported by four shared sequence insertions (Supplementary Fig. 8). Node values indicate

1169 posterior probability / RAXML bootstrap support; scale bar represents substitutions per site.

1170 GenBank accession numbers are in Supplementary Table 12.



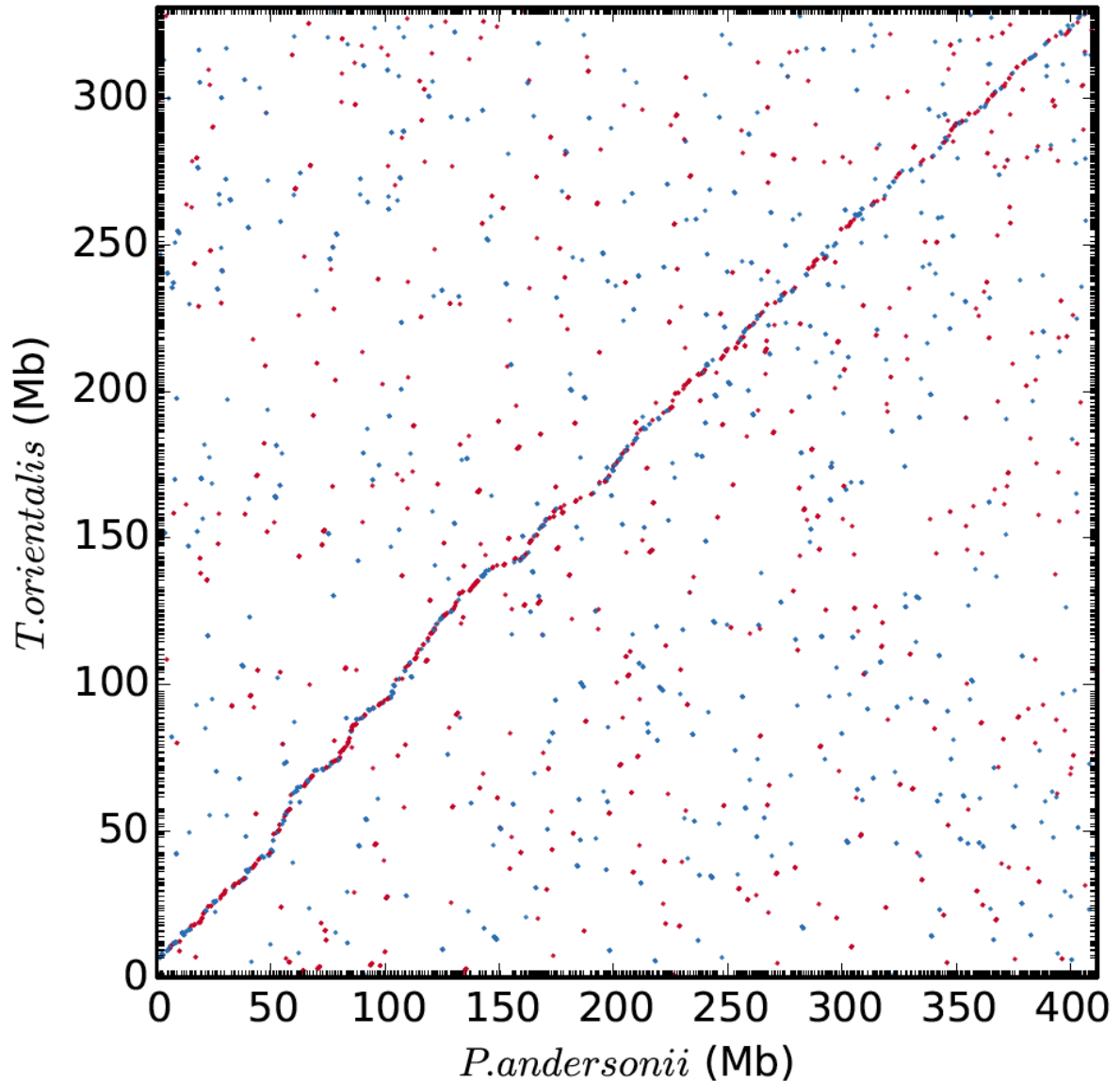
1171

1172

1173

Supplementary Figure 8: Chloroplast genome insertions in Cannabaceae. Shared

1174 sequence insertions in chloroplast genomes supporting **(a-d)** or refuting **(e)** sister relationship
1175 of *Parasponia* and *Trema levigata*. **(a)** *matK-rps16* intergenic spacer, **(b)** *rps16-psbK*
1176 intergenic spacer, **(c)** *atpF* intron, **(d, e)** *petA-psbJ* intergenic spacer. Numbers indicate
1177 alignment coordinates; colours indicate percent identity while ignoring gaps: green = 100%,
1178 olive = 80-100%, yellow = 60-80%; black rectangles mark shared sequence insertions
1179 concerned.



1180

1181

1182 **Supplementary Figure 9: Whole genome alignment dotplot for *P. andersonii* and *T.***

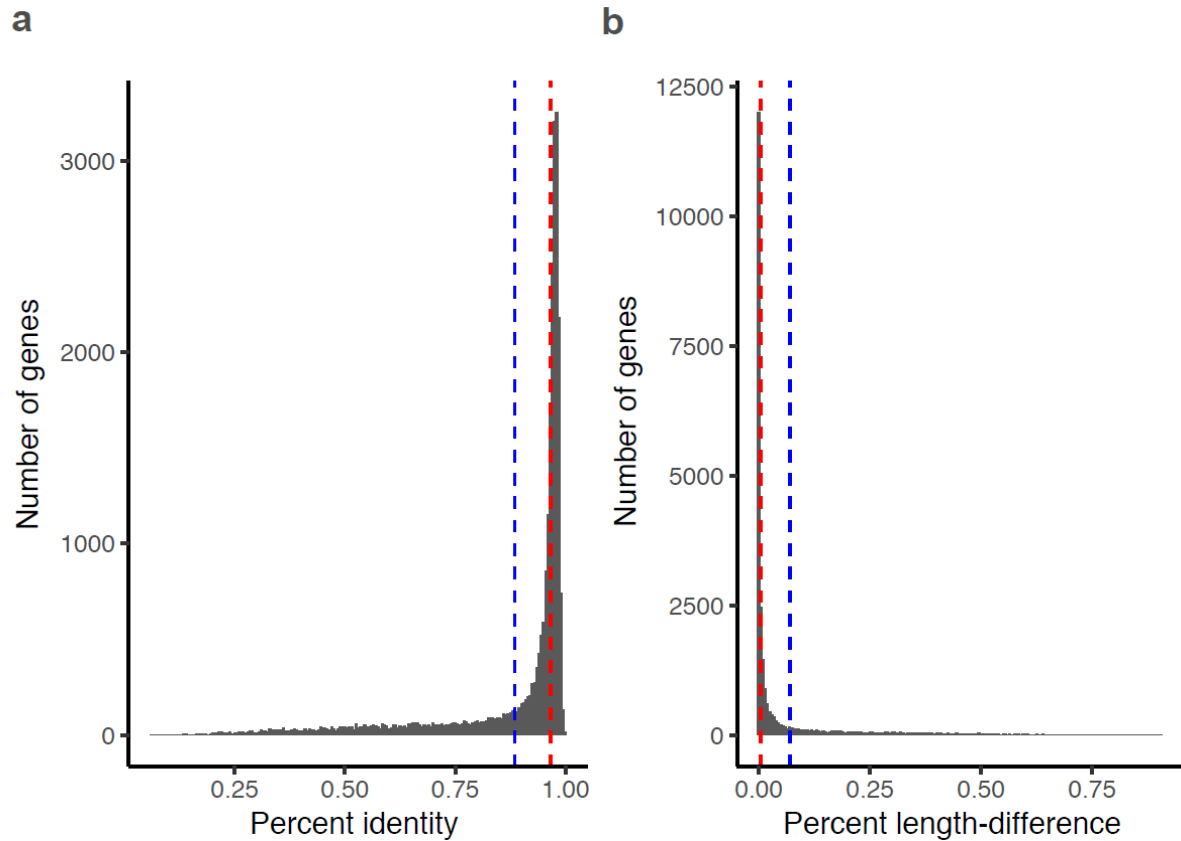
1183 ***orientalis* RG33.** Maximal unique matching (MUM) alignments were generated using nucmer

1184 4.0.0beta with the following settings: breaklength 500, mincluster 200, maxgap 100,

1185 minmatch 80, minalign 7000. Forward alignments are red, reverse alignments are blue.

1186 Scaffolds are ordered by alignment size, which results in a clear diagonal line indicating the

1187 collinearity of the two genomes.



1188

1189

1190 **Supplementary Figure 10: Identity of *P. andersonii* - *T. orientalis* putative orthologous**

1191 **gene pairs.** Histograms of (a) percent nucleotide identity (calculated by taking the fraction of

1192 identical nucleotides ignoring end gaps using global alignments produced by MAFFT version

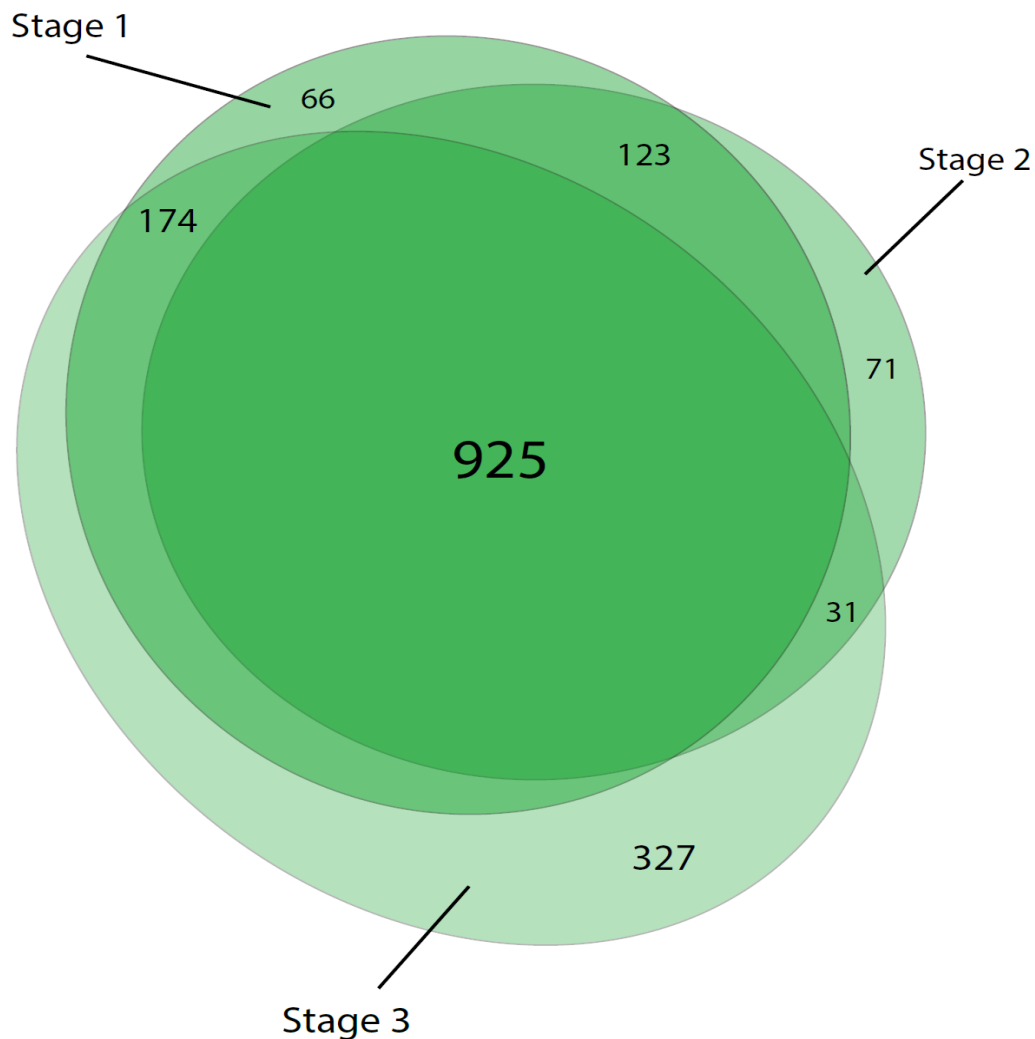
1193 7.017¹⁰³) and (b) length difference of all 25,605 orthologous gene pairs from *P. andersonii*

1194 and *T. orientalis* as a percentage of the longest gene. Red line indicates median, blue line

1195 indicates mean.

1196

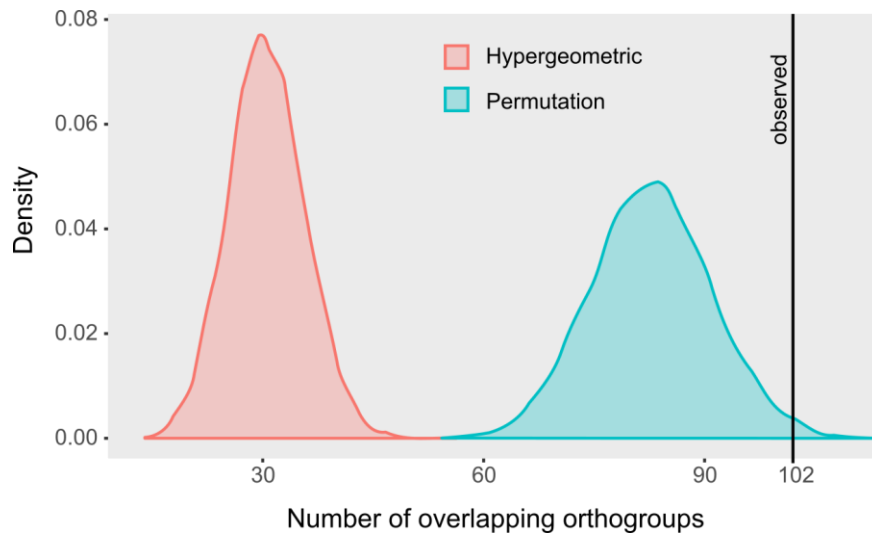
1197



1198

1199

1200 **Supplementary Figure 11: Venn diagram of *P. andersonii* nodule enhanced genes in 3**
1201 **developmental stages.** Nodule developmental stages according to Fig. 1h-j. List of genes is
1202 given in Supplementary Table 9. *Parasponia andersonii* genes are considered 'nodule
1203 enhanced' when expression is increased >2-fold in any of 3 nodule developmental stages
1204 when compared to non-inoculated root sample. Largest fraction concerns genes enhanced in
1205 all 3 stages.



1206

1207

1208 **Supplementary Figure 12: Statistical testing of common utilization of genes in**

1209 ***Parasponia* and medicago.** To assess common utilization of genes in *Parasponia* and

1210 medicago nodules we performed statistical testing of overlap between *Parasponia andersonii*

1211 and medicago nodule-enhanced genes. Overlap was calculated based on orthogroup

1212 membership (i.e. when an orthogroup contains nodule-enhanced genes from *P. andersonii*

1213 and medicago it is scored as overlap). Significance of set overlaps is usually calculated

1214 based on the hypergeometric distribution. However, because larger orthogroups have higher

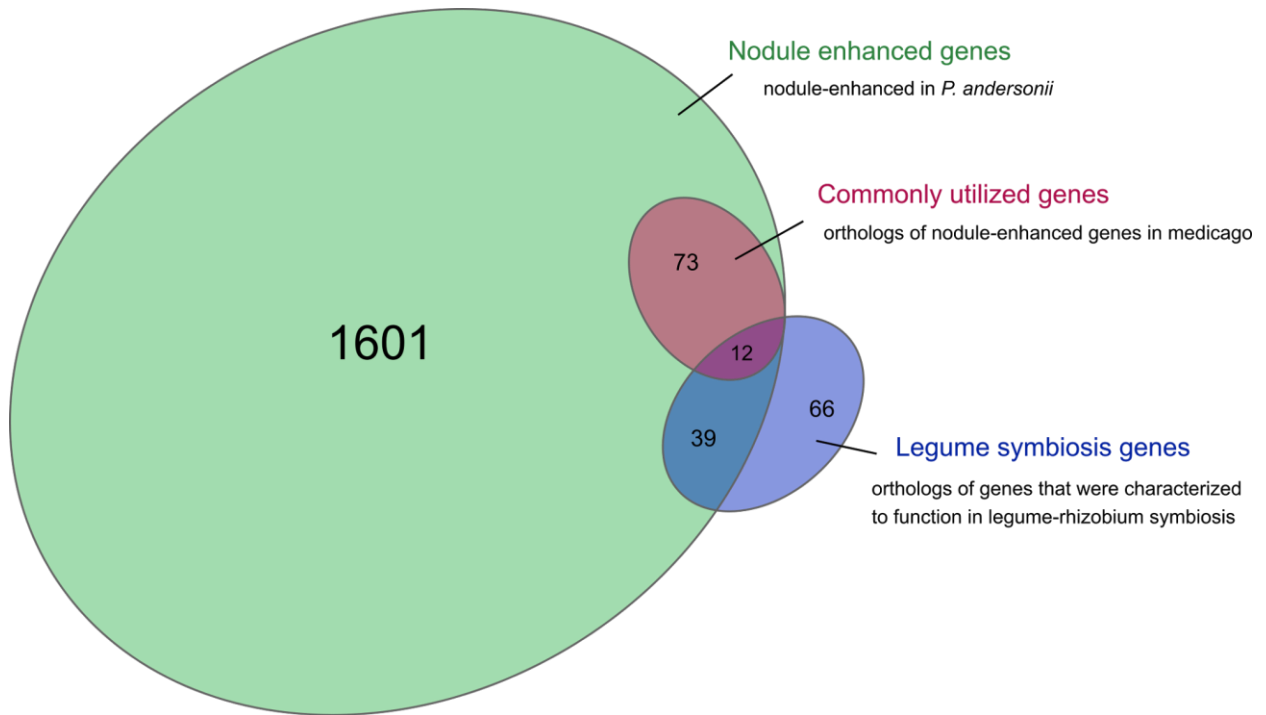
1215 chance of overlap, the hypergeometric is not suitable. We therefore assessed significance

1216 with a permutation test where the null distribution is based on overlap found when gene-

1217 orthogroup membership is randomized (n=10,000).Figure shows density plots of both

1218 hypergeometric distribution and permutation random variates. Vertical line shows the

1219 observed number of 102 overlapping orthogroups ($p < 0.02$ based on permutation test).



1220

1221

1222 **Supplementary Figure 13: Venn diagram of *P. andersonii* symbiosis gene sets.** Nodule

1223 enhanced genes have a significantly enhanced expression level (fold change > 2, $p < 0.05$,

1224 DESeq2 Wald test) in any of three developmental stages (N = 1725; Supplementary Fig. 11;

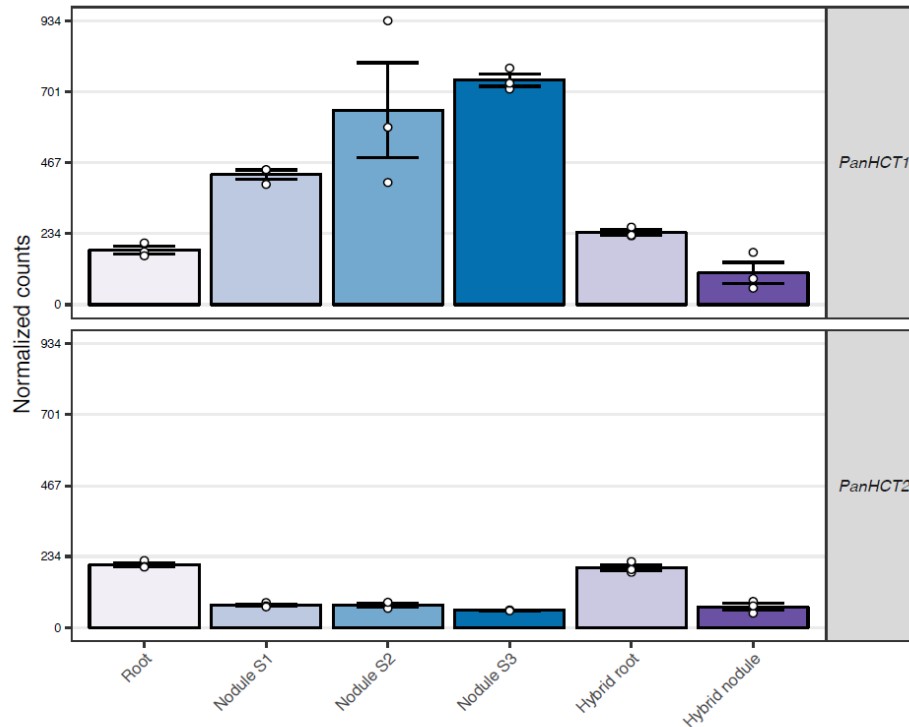
1225 Supplementary Table 9). Commonly utilized genes are nodule-enhanced in *P. andersonii* as

1226 well as in the legume medicago³¹ (N = 85; Supplementary Table 10, Supplementary Data

1227 File 2). Legume symbiosis genes are orthologs of genes that were characterized to function

1228 in legume-rhizobium symbiosis (N = 117; Supplementary Table 1, Supplementary Data File

1229 1).



1230

1231 **Supplementary Figure 14: Expression profile of *PanHCT1* and *PanHCT2* genes.**

1232 Expression of *P. andersonii* HYDROXYCINNAMOYL-COA SHIKIMATE TRANSFERASE 1

1233 (*PanHCT1*) and *PanHCT2* in *P. andersonii* roots, stage 1-3 nodules, and in *P. andersonii* x *T.*

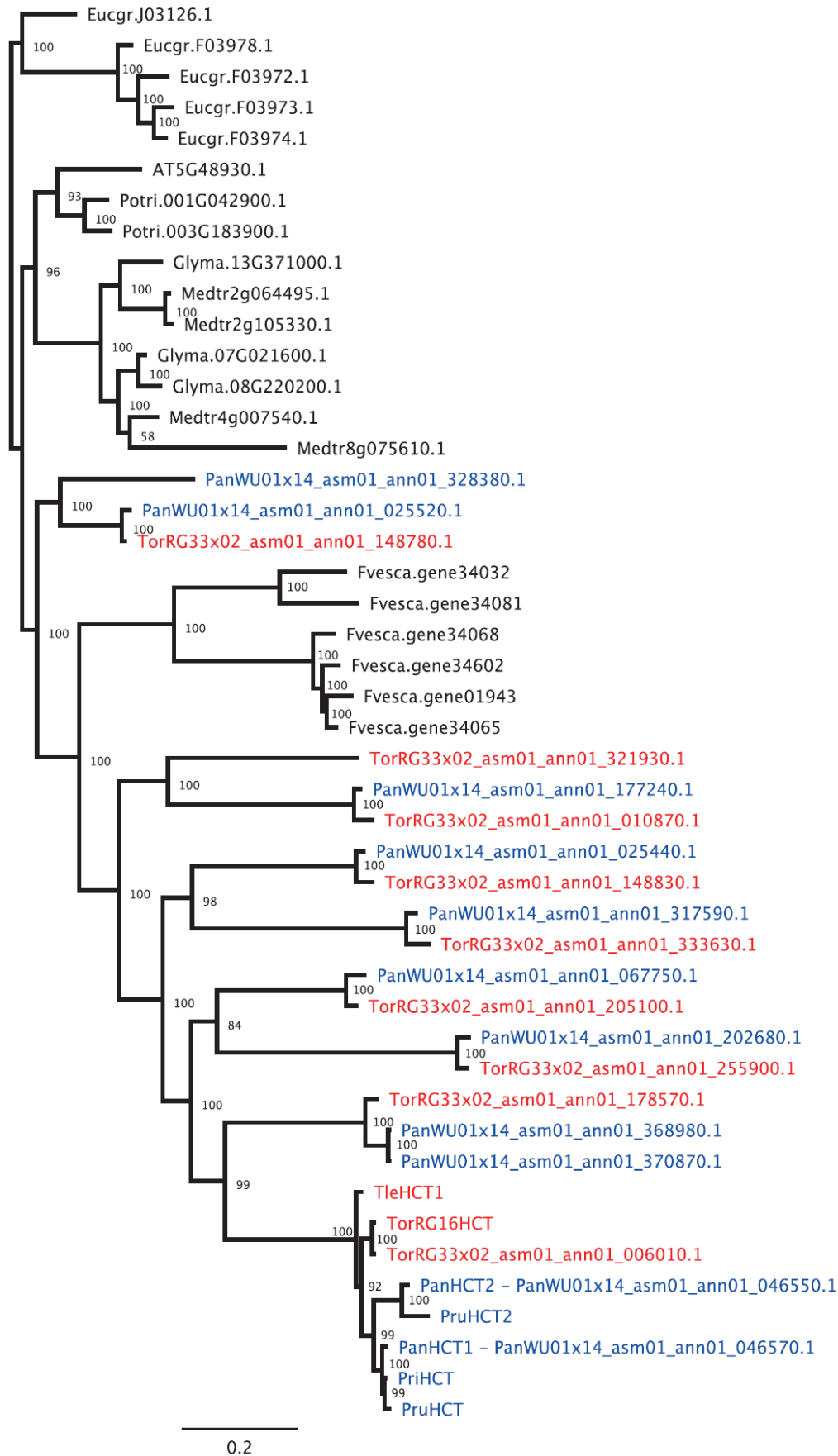
1234 *tomentosa* F₁ hybrid roots and nodules (line H9). *PanHCT1* and *PanHCT2* represent the only

1235 *Parasponia*-specific gene duplication in the defined symbiosis gene set, as *PanHCT1* is

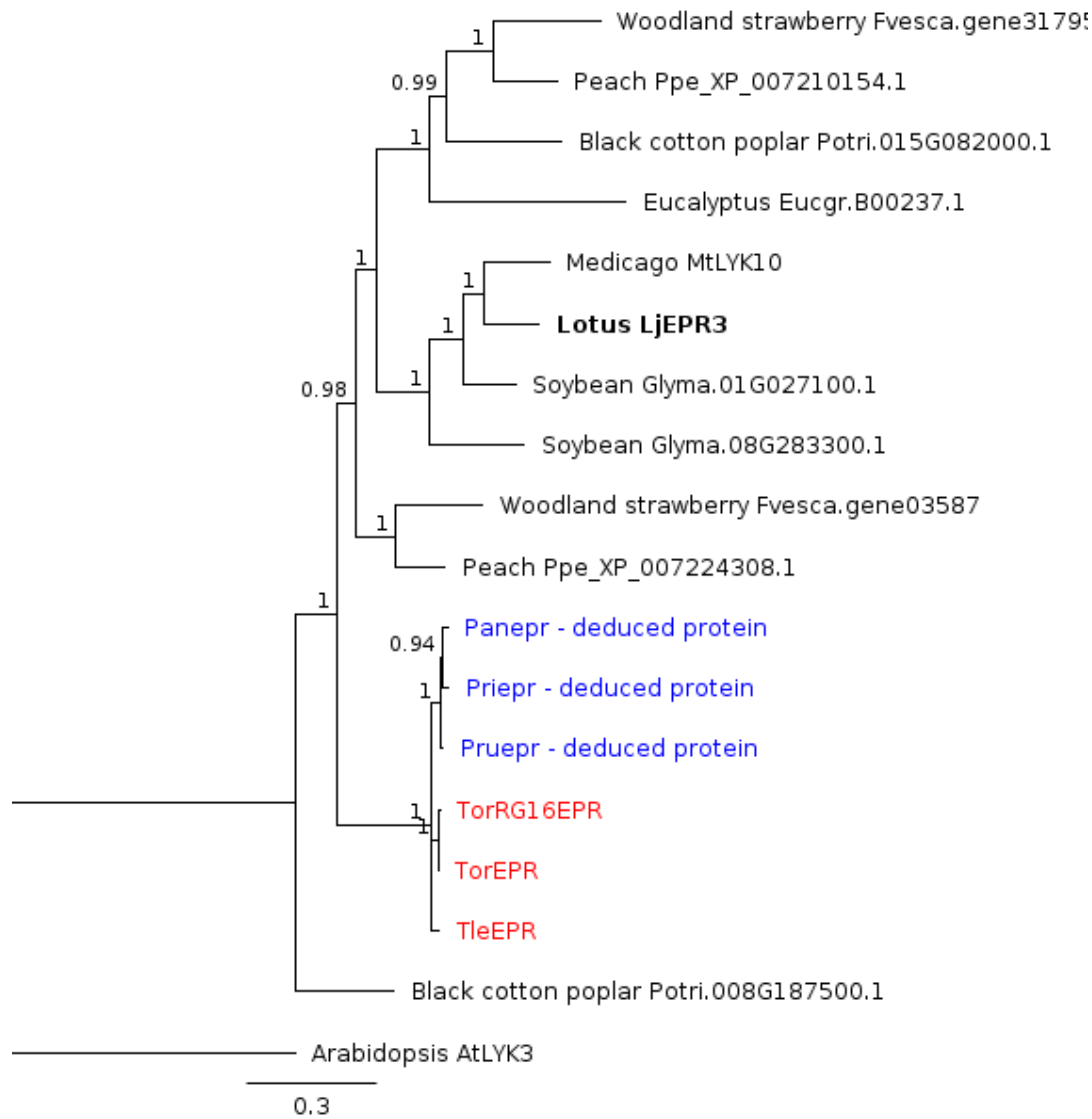
1236 upregulated in nodules. Expression is given in DESeq2 normalized read counts, error bars

1237 represent standard error of three biological replicates, dots represent individual expression

1238 levels.



1240 **Supplementary Figure 15: Phylogenetic reconstruction of Hydroxycinnamoyl-CoA**
1241 **Shikimate Transferase (HCT) orthogroup.** *HCT* orthogroup was created by merging
1242 OG0001291, OG0016758, OG0016791, OG0018560, OG0020327, OG0020921,
1243 OG0022256 & OG0023772, supplemented with HCT1 and HCT2 orthologs of *P. rigida*, *P.*
1244 *rugosa*, *T. orientalis* RG16 and *T. levigata*. *PriHCT2* is a putative pseudogene and was not
1245 included. HCT1 and HCT2 represent the only *Parasponia* specific gene duplication in the
1246 defined symbiosis gene set, as *PanHCT1* was found to be upregulated in nodules. Species
1247 included: *Parasponia andersonii* (Pan); *P. rigida* (Pri); *P. rugosa* (Pru) (all in blue); *Trema*
1248 *orientalis* (Tor); *T. orientalis* RG16 (TorRG16); *T. levigata* (Tle) (all in red); *Medicago*
1249 *truncatula* (Mt); *Glycine max* (Glyma), *Populus trichocarpa* (Potri); *Fragaria vesca* (Fvesca);
1250 *Eucalyptus grandis* (Eugr); *Arabidopsis thaliana* (AT). Phylogenetic inference was calculated
1251 using MrBayes 3.2.2. Scale bar represents substitutions per site.



1252

1253 **Supplementary Figure 16: Phylogenetic reconstruction of the EPR3 orthogroup.**

1254 Alignment of orthogroup OG0010070 containing exopolysaccharide receptor LjEPR3. Note

1255 that all *Parasponia* species lack a functional *EPR* (Supplementary Fig. 15). Species included:

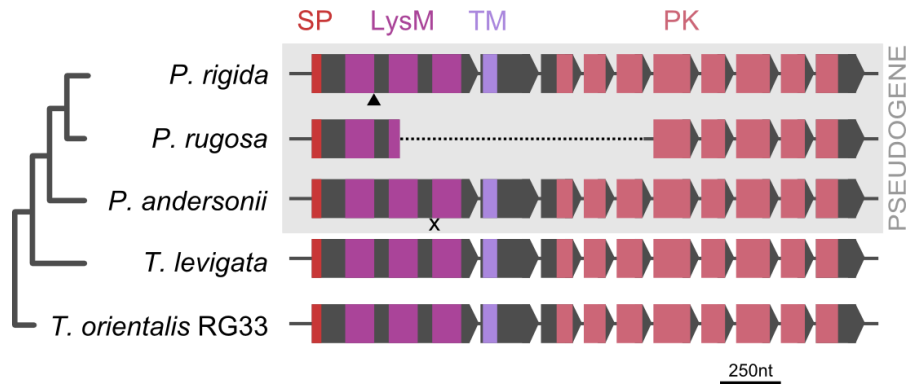
1256 *Trema orientalis* RG33 (Tor); *Trema orientalis* RG16 (TorRG16); *Trema levigata* (Tle) (all in

1257 red); *Parasponia Andersonii* (Pan); *Parasponia Rigida* (Pri) *Parasponia Rugosa* (Pru) (all in

1258 blue). *Medicago truncatula* (Mt); *Glycine max* (Glyma), *Populus trichocarpa* (Potri); *Fragaria*

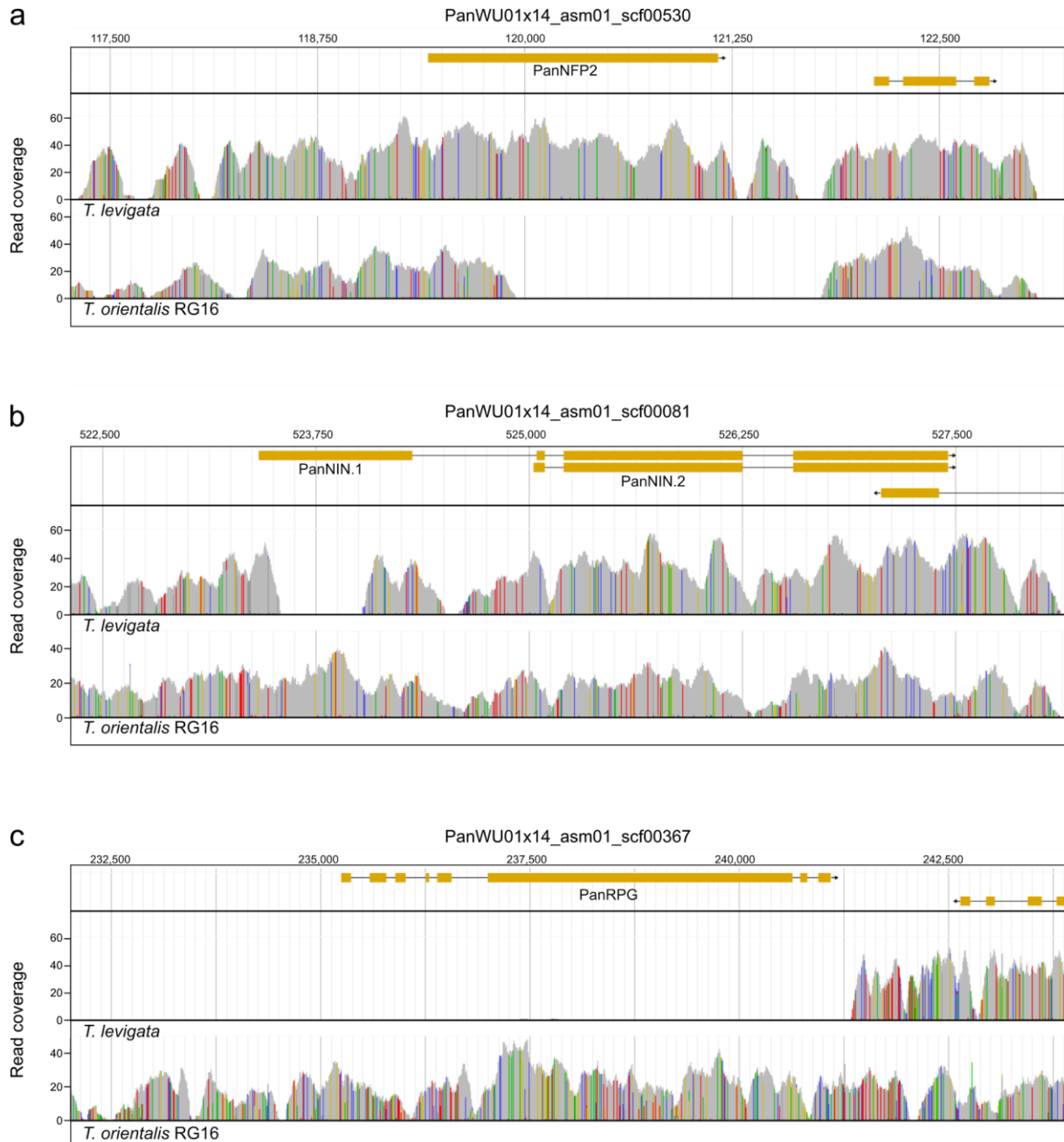
1259 *vesca* (Fvesca); *Eucalyptus grandis* (Eugr). Phylogenetic inference was calculated using

1260 MrBayes 3.2.2. Scale bar represents substitutions per site.



Supplementary Figure 17: 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).



1270

1271 **Supplementary Figure 18: Read mappings of *Trema orientalis* RG16 and *T. levigata* to**

1272 **the *Parasponia andersonii* genome.** Read mappings to gene region of (a) *PanNFP2*,

1273 illustrating absence of a large part of the gene in *T. orientalis* RG16, (b) *PanNIN*, illustrating

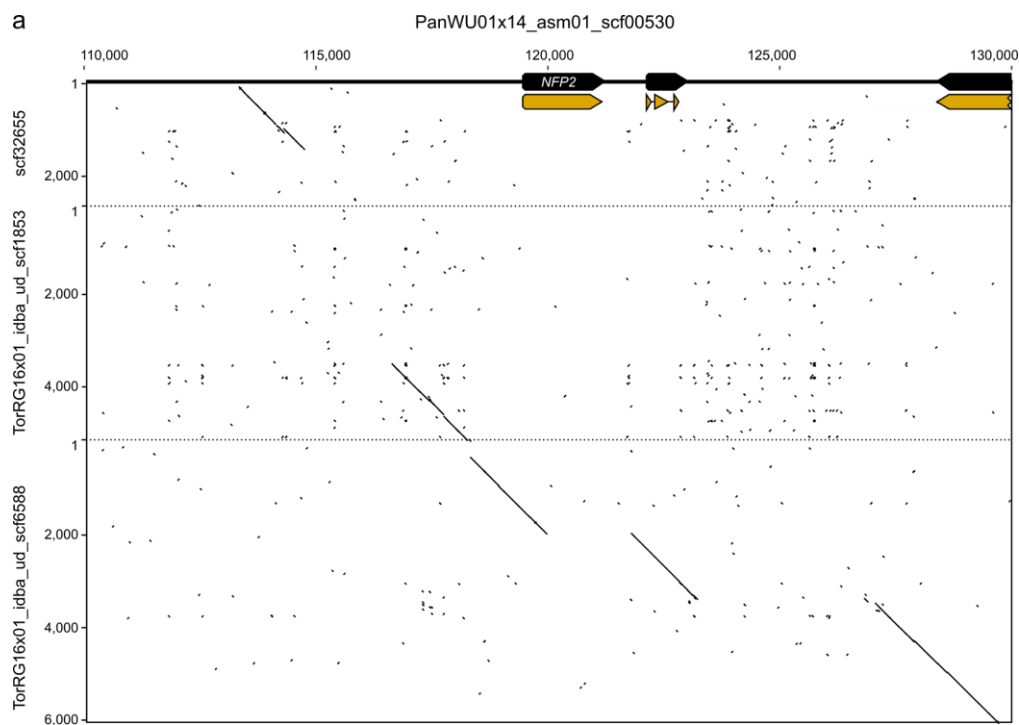
1274 absence of a large part of the canonical first exon in *T. levigata*, (c) *PanRPG*, illustrating

1275 absence of the gene in *T. levigata*. Coordinates on the x-axis correspond to those of the *P.*

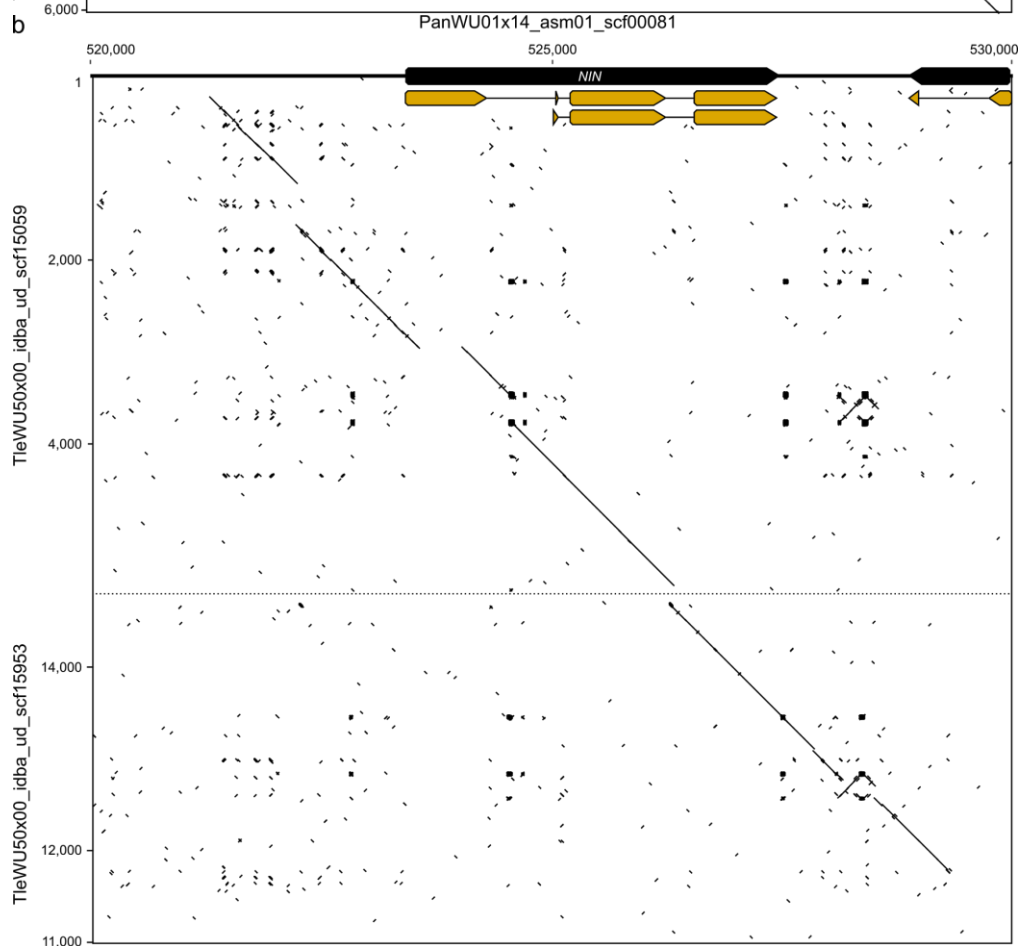
1276 *andersonii* scaffold; orange bars depict *P. andersonii* gene models; histograms depict read

1277 coverage in grey; nucleotide differences from the *P. andersonii* reference scaffold are in color

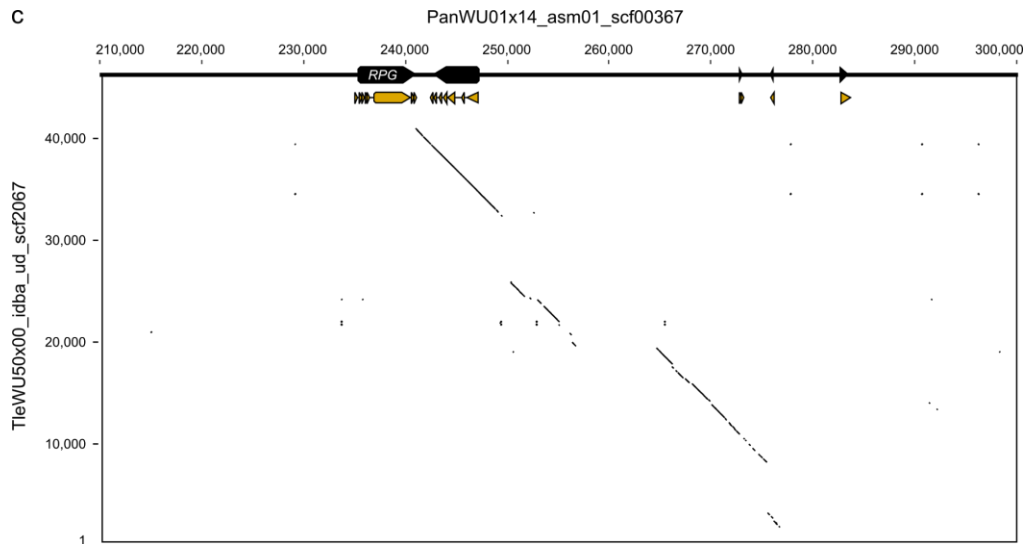
1278 (green = adenine, blue = cytosine, yellow = guanine, red = thymine).



1279



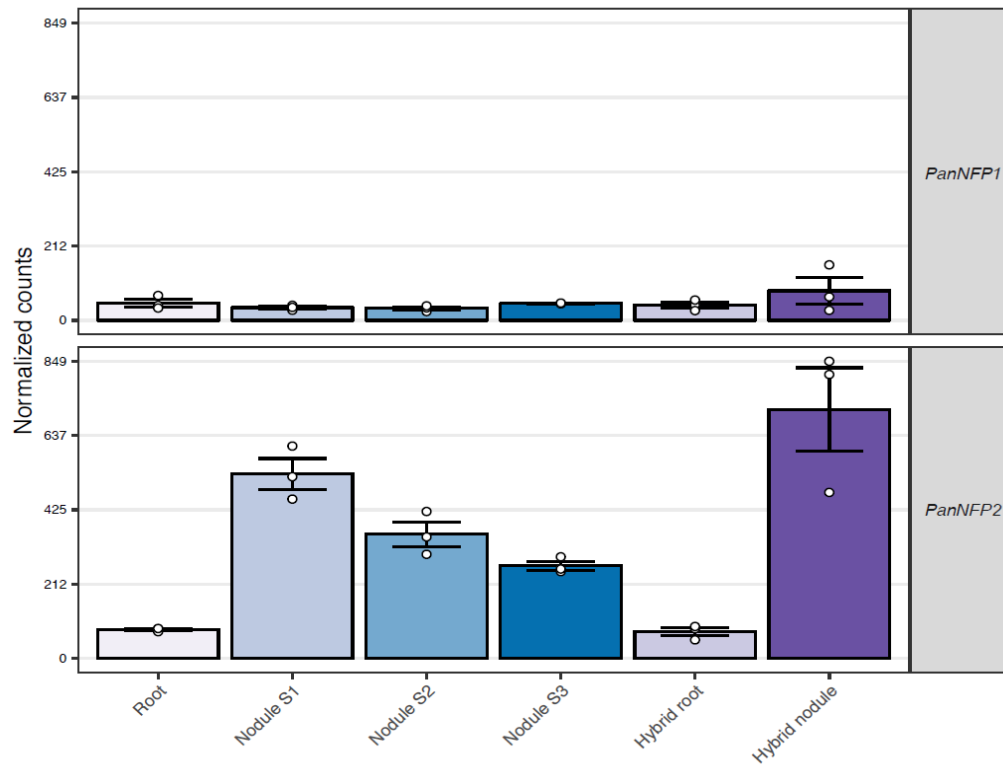
1280



1281

1282

1283 **Supplementary Figure 19: Genomic alignments of *Trema orientalis* RG16 or *Trema***
1284 ***levigata* to *Parasponia andersonii* NFP2, NIN, and RPG gene regions.** Genome
1285 alignment(s) of (a) *T. orientalis* RG16 with *PanNFP2* gene region, (b) *T. levigata* with
1286 *PanNIN* gene region, (c) *T. levigata* with *PanRPG* gene region. Coordinates correspond to
1287 those on the draft genome scaffolds; *Parasponia andersonii* gene and CDS models are
1288 depicted in black and orange, respectively; different genomic scaffolds are separated by
1289 dashed lines.



1290

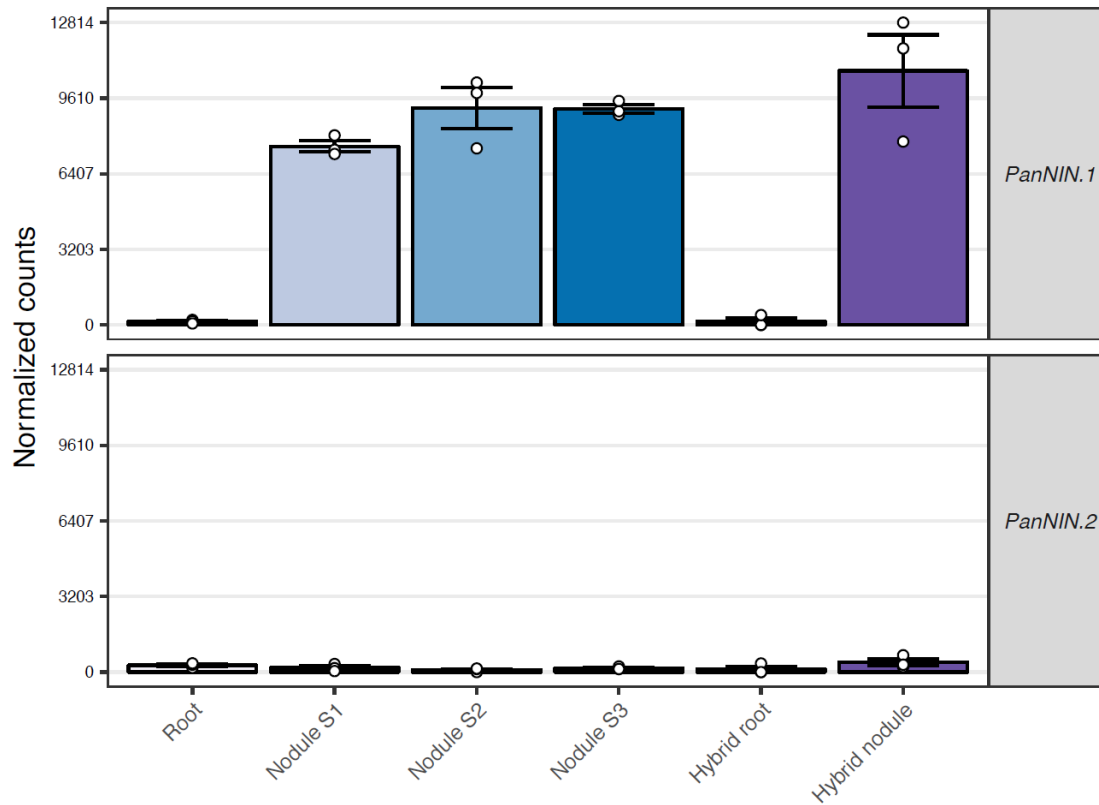
1291 **Supplementary Figure 20: Expression profile of *PanNFP1* and *PanNFP2* genes.**

1292 Expression of *P. andersonii* NOD FACTOR PERCEPTION 1 (*PanNFP1*) and *PanNFP2* in *P.*

1293 *andersonii* roots, stage 1-3 nodules, and in *P. andersonii* x *T. tomentosa* F1 hybrid roots and

1294 nodules. Expression is given in DESeq2 normalized read counts, error bars represent

1295 standard error of three biological replicates, dots represent individual expression levels.



1296

1297 **Supplementary Figure 21: Expression of *P. andersonii* NODULE INCEPTION (*PanNIN*)**

1298 **gene splice variants.** *PanNIN.1* encodes a canonical symbiotic protein, whereas *PanNIN.2*

1299 encodes a shorter protein variant that is the result of an alternative start site in an intron.

1300 Expression levels were determined by identifying unique DNA sequences for both variants;

1301 spanning the intron in case of *PanNIN.1* (CTGCCAAGCGCTTGAGGCTGTTGATCTT), or

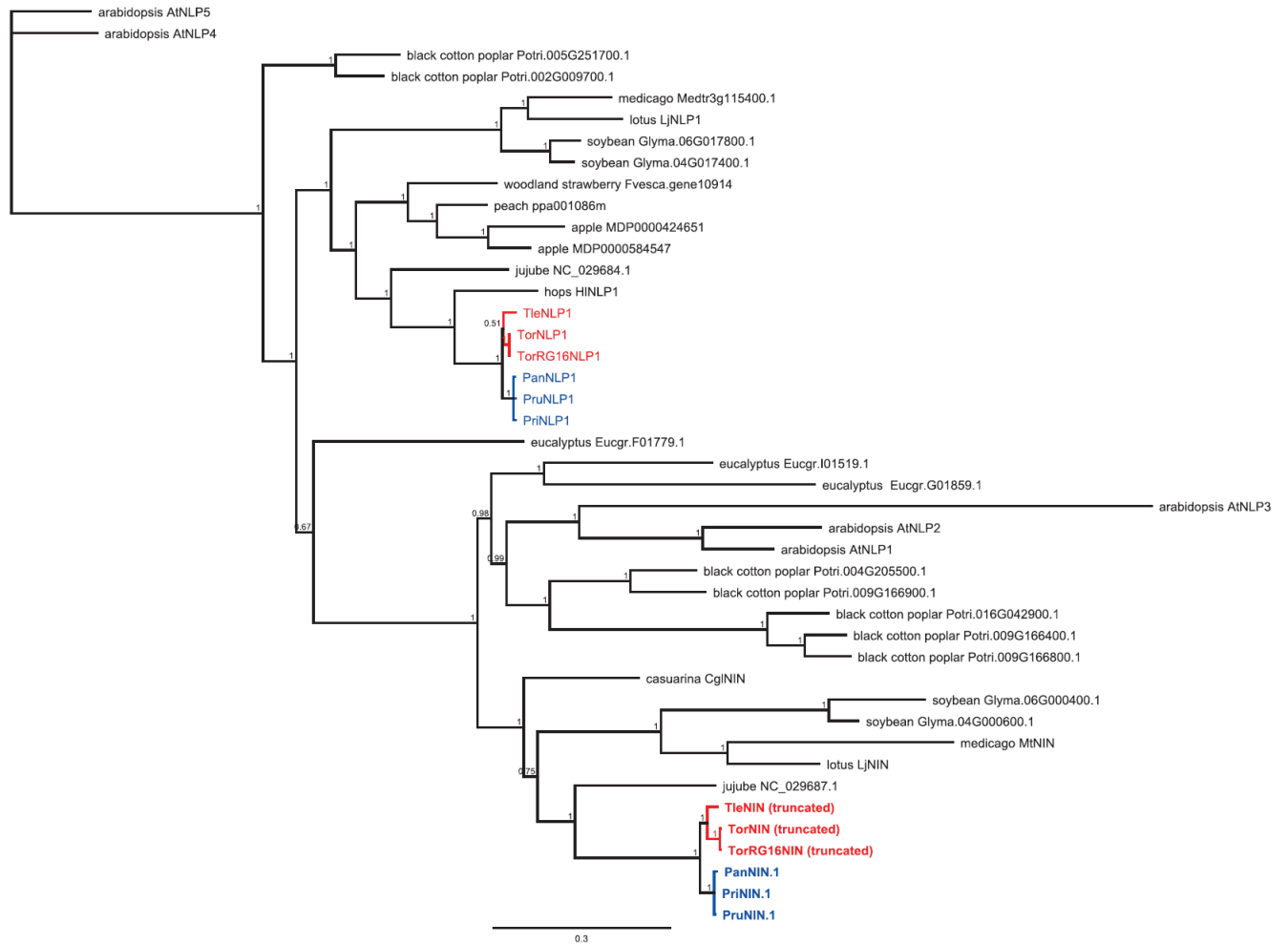
1302 including the start site of *PanNIN.2* (GCCAATTACCTTGCAGGCTGTTGATCTT) and

1303 counting all occurrences in the RNA-seq reads. DESeq2 size factors were used to normalize

1304 these counts. The fraction of these normalized counts between *PanNIN.1* and *PanNIN.2* was

1305 used to scale the expression levels. Error bars represent standard error of three biological

1306 replicates, dots represent individual expression levels.



1307

1308

1309 **Supplementary Figure 22: Phylogenetic reconstruction of NIN orthogroup.** Alignment of

1310 OG0001118, which includes NIN and NLP1 (NIN-LIKE PROTEIN 1)-like proteins,

1311 supplemented with additional species. AtNLP4 and AtNLP5 were included as outgroup.

1312 *Parasponia* spp. marked in blue, *Trema* spp. In red. Note that in *Trema* species NIN only

1313 occurs in truncated forms (Fig. 7). Included species: *Parasponia andersonii* (Pan);

1314 *Parasponia rigida* (Pri); *Parasponia rugosa* (Pru); *Trema orientalis* RG33 (Tor); *Trema*

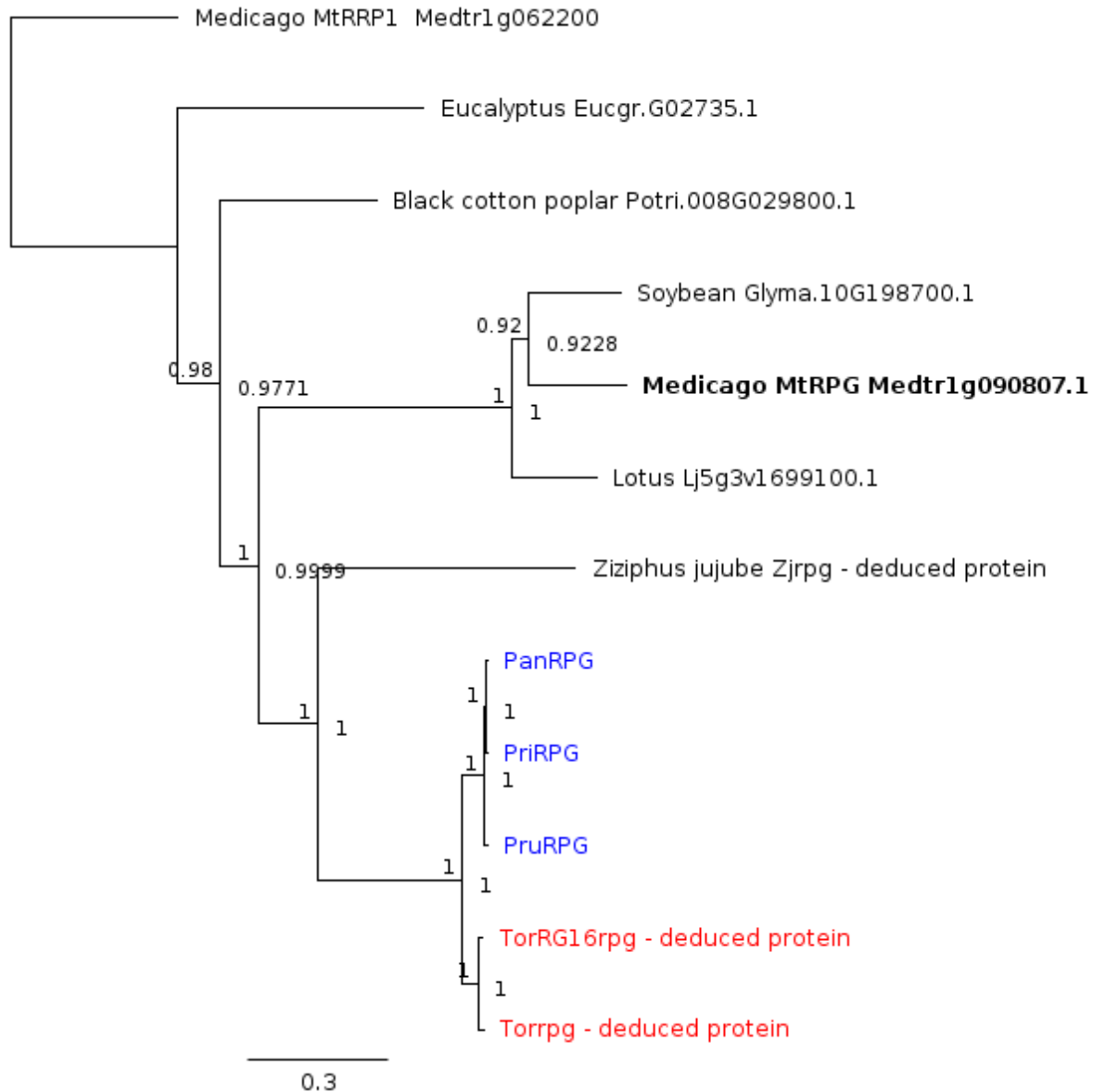
1315 *orientalis* RG16 (TorRG16); *Trema levigata* (Tle); medicago (*Medicago truncatula*, Mt); lotus

1316 (*Lotus japonicus*, Lj); soybean (*Glycine max*, Glyma); peach (*Prunus persica*, ppe); woodland

1317 strawberry (*Fragaria vesca*, Fvesca); back cotton poplar (*Populus trichocarpa*, Potri);

1318 eucalyptus (*Eucalyptus grandis*, Eugr); arabidopsis (*Arabidopsis thaliana*, At), jujube

1319 (*Ziziphus Jujube*) apple (*Malus x domestica*), mulberry (*Morus Notabilis*), hop (*Humulus*
1320 *Lupulus (natsume.shinsuwase.v1.0)*), and casuarina (*Casuarina glauca*). Node numbers
1321 indicate posterior probabilities, scale bar represents substitutions per site.



1322

1323

1324 **Supplementary Figure 23: Phylogenetic reconstruction of the RPG orthogroup.**

1325 Alignment of OG0014072 was supplemented with RPG homologs of additional species.

1326 *Parasponia* spp. marked in blue, Nitrogen fixation clade in bold. Included species:

1327 *Parasponia andersonii* (Pan) *Parasponia rigida* (Pri); *Parasponia rugosa* (Pru) *Medicago*

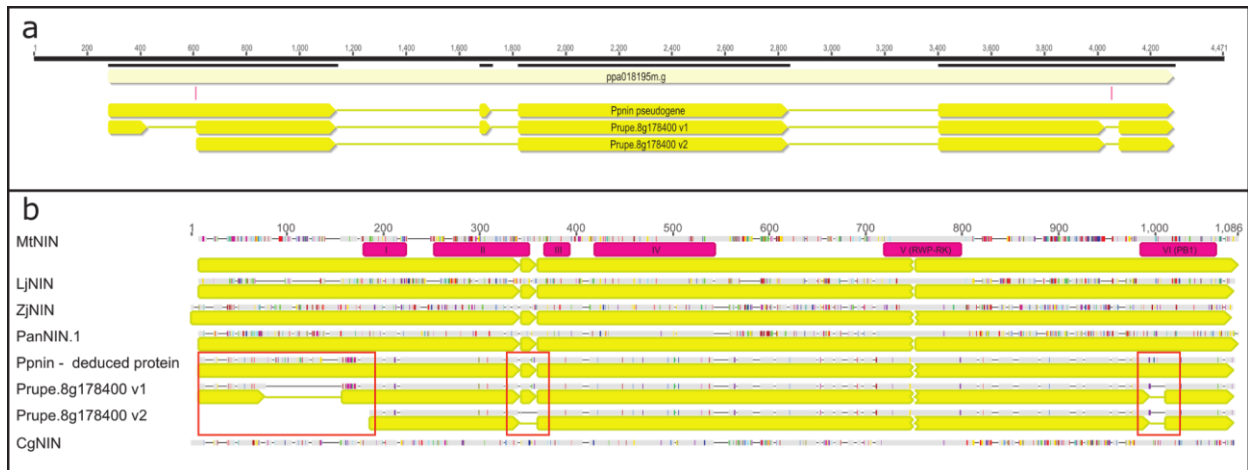
1328 *truncatula* (Mt); *Lotus japonicus* (Lj); *Glycine max* (Glyma), *Populus trichocarpa* (Potri);

1329 *Eucalyptus grandis* (Eugr). *Trema orientalis* RG33 (Tor); *Trema. orientalis* RG16 (TorRG16).

1330 *Ziziphus jujube* (Zj). No other functional RPG proteins could be detected in Rosales species,

1331 including *Fragaria vesca* *Ziziphus Jujube*, *Malus Domestica*, *Morus Notabilis*, and *Humulus*

1332 *Lupulus* (*natsume.shinsuwase.v1.0*). Outgroup: *M. truncatula* MtRRP1 (RPG RELATED
1333 PROTEIN 1, Medtr1g062200.1). Node numbers indicate posterior probabilities, scale bar
1334 represents substitutions per site.



1335

1336 **Supplementary Figure 24: Annotation of *Prunus persica* locus ppa018195m.g**

1337 **representing *PpNIN*.** (a) Comparison of the exon-intron structure of two publicly released

1338 gene models (named Prupe.8g17800_v1 and Prupe.8g178400_v2) and the gene model

1339 used here (*Ppnin* pseudogene). Yellow arrows: exons. Red bars indicate 2 single-nucleotide

1340 insertions that affect the coding region of the *Ppnin* pseudogene. (b) Alignment of

1341 derived/deduced NIN proteins of 3 *Prunus persica* gene models Prupe.8g17800_v1,

1342 Prupe.8g178400_v2, and *Ppnin* pseudogene, with *Medicago truncatula* MtNIN, *Lotus*

1343 *japonicus* LjNIN, *Ziziphus jujube* ZjNIN, *Parasponia andersonii* PanNIN.1, and *Casuarina*

1344 *glauca* CgNIN. Six conserved domains are annotated in MtNIN (cyan). Exon structure for all

1345 *NIN* genes indicated in yellow (except CgNIN for which no gene sequence is available).

1346 Deviations in the three *Prunus persica* derived/deduced NIN proteins are marked in red

1347 boxes.

1348 **Table 1.** Copy number variants in nodulation genes that are consistent between *Parasponia*
 1349 and *Trema* genera.
 1350

Name	ID	CNV type	Class	Description
PanNFP2	PanWU01x14_asm01_ann01_320250	loss in <i>Trema</i>	LS,NE	LysM domain containing receptor kinase, putative rhizobium LCO receptor
PanCRK11	PanWU01x14_asm01_ann01_285030	loss in <i>Trema</i>	NE	Cysteine rich receptor like kinase
PanLEK1	PanWU01x14_asm01_ann01_069780	loss in <i>Trema</i>	NE	Concanavalin A-like lectin receptor kinase
PanNIN	PanWU01x14_asm01_ann01_111140	loss in <i>Trema</i>	LS,CR	Ortholog of transcription factor NODULE INCEPTION
PanRPG	PanWU01x14_asm01_ann01_272380	loss in <i>Trema</i>	LS,CR	Ortholog of long coiled-coil protein RHIZOBIUM-DIRECTED POLAR GROWTH
PanDEF1	PanWU01x14_asm01_ann01_187760	loss in <i>Trema</i>	NE	Defensin-like protein
PanGAT	PanWU01x14_asm01_ann01_150960	loss in <i>Trema</i>	NE	Gamma-aminobutyric acid (GABA) transporter
PanHCT1	PanWU01x14_asm01_ann01_046570	duplication in <i>Parasponia</i>	NE	Hydroxycinnamoyl-CoA shikimate / Quinate hydroxycinnamoyl transferase
TorEPR	TorRG33x02_asm01_ann01_052550	loss in <i>Parasponia</i>	LS	LysM domain containing receptor kinase, putative rhizobium exopolysaccharide receptor
TorN19L3	TorRG33x02_asm01_ann01_066920	loss in <i>Parasponia</i>	LS	NODULIN19-like protein
TorIPT4	TorRG33x02_asm01_ann01_307000	loss in <i>Parasponia</i>	LS	Isopentenyltransferase

1351
 1352 Gene ID corresponds to that in *P. andersonii*, or *T. orientalis* in case of gene loss in
 1353 *Parasponia* species. LS: putative ortholog of legume genes that function in symbiosis (see:
 1354 Supplementary Table 1), NE: nodule enhanced expression in *P. andersonii* (see:
 1355 Supplementary Table 9), CR: genes that are commonly utilized in *P. andersonii* and
 1356 medicago (see: Supplementary Table 10). Expression profiles of the *P. andersonii* genes are
 1357 depicted in Fig. 4.

1358 **Supplementary Table 1:** *Parasponia andersonii* and *Trema orientalis* RG33 putative
1359 orthologs of legume genes that function in rhizobium symbiosis.

1360 Genes have been classified according to function of encoded proteins. CNV between *P.*
1361 *andersonii* and *T. orientalis* are marked in red. Genes for which no putative ortholog could be
1362 identified are indicated (not identified). The *P. andersonii* and *T. orientalis* genes are
1363 classified as either 'putative ortholog', 'closest homolog' or 'inparalog' depending on the
1364 phylogenetic relation with the legume symbiosis gene. Orthogroup number corresponds to
1365 orthogroups in Supplementary Table 7. It is indicated in case gene has been found to
1366 function in other symbiosis. AM: arbuscular mycorrhiza, and ANS: actinorhizal nodule
1367 symbiosis. Gm: *Glycine max*; Lj: *Lotus japonicus*; Ms: *Medicago sativa*; Mt: *Medicago*
1368 *truncatula*; Ps *Pisum sativum*; Pv: *Phaseolus vulgaris*.

1369

1370 **Supplementary Table 2:** Intergeneric crossings between *Parasponia* and *Trema* species.

1371 Results column indicates whether intergeneric crosses could be obtained (positive) or not
1372 (negative).

1373

1374 **Supplementary Table 3:** *Parasponia-Trema* germplasm collection.

1375 **Supplementary Table 4:** Genome size estimations based on estimated genome coverage.

1376 **Supplementary Table 5:** Genome sequencing strategy.

1377 **Supplementary Table 6:** Assembly results of *Parasponia - Trema* genome sequences.

1378 #N is number of gap sequences; GC% is guanine-cytosine content; BUSCO⁸⁷ and
1379 CEGMA^{86,87} are tools that assess completeness of genome assemblies by checking sets of
1380 conserved genes. For BUSCO a set of 1,440 plant specific genes was used.

1381

1382 **Supplementary Table 7:** Inferred orthogroups.

1383 Eurosids orthogroups generated by OrthoFinder based on gene models from *Parasponia*
1384 *andersonii*, *Trema orientalis*, *Medicago truncatula*, *Glycine max*, *Fragaria vesca*, *Populus*
1385 *trichocarpa*, *Arabidopsis thaliana* and *Eucalyptus grandis*.

1386

1387 **Supplementary Table 8:** Gene models in *Parasponia andersonii* and *Trema orientalis* RG33
1388 reference genomes.

1389 Inparalogs: species specific duplications; singletons: loss of gene in other species; multi-
1390 orthologs: duplication in the other species; CNVs: copy number variants. We found no
1391 significant enrichment of these CNVs in the symbiosis genes in Supplementary Table 1 and
1392 nodule enhanced genes in Supplementary table 9 (hypergeometric test, $p = 0.99$). For
1393 BUSCO a set of 1,440 plant specific genes was used.

1394

1395 **Supplementary Table 9:** *Parasponia* nodule-enhanced gene set.

1396 *P. andersonii* genes with enhanced expression in three nodule developmental stages
1397 compared to non-inoculated roots (>2-fold increase). Stage 1: initial stages of colonization
1398 when infection threads entering the host cells. Stage 2: progression of rhizobium infection in
1399 nodule host cells, Stage 3: nodule cells filled with fixation threads. Plants were inoculated
1400 with *M. plurifarium* BOR2. OrthoGroup number corresponds to Supplementary Table 7,

1401 STAT: trident alignment conservation score, CLASS: gene homology classification (nv =
1402 orthologous pair not validated by whole-genome alignments) FC: gene expression log fold-
1403 change in nodule versus root, P: P-value adjusted for multiple testing based on false
1404 discovery rate estimation. Genes that are putatively orthologous to legume genes with
1405 symbiotic function are classified as 'LEGUME SYMBIOSIS GENE'. Conserved CNVs
1406 between *Parasponia* and *Trema* species are shaded pink.

1407

1408 **Supplementary Table 10:** Commonly utilized symbiosis gene set.

1409 *P. andersonii* genes with enhanced expression in three nodule developmental stages
1410 compared to non-inoculated roots (>2-fold increase) (Supplementary Table 9), of which a
1411 putative ortholog in *Medicago truncatula* also has been identified as a gene with a nodule
1412 enhanced expression³¹. Cluster numbers indicate expression profile clusters of commonly
1413 utilized genes as shown in Fig. 5. DESeq2 normalized read counts are included for *P.*
1414 *andersonii* and hybrid roots, hybrid nodules and three stages of *P. andersonii* nodules. Stage
1415 1: initial stages of colonization when infection threads enter the host cells. Stage 2:
1416 progression of rhizobium infection in nodule host cells, Stage 3: nodule cells filled with
1417 fixation threads. Plants were inoculated with *M. plurifarium* BOR2. OrthoGroup number
1418 corresponds to Supplementary Table 7, STAT: trident alignment conservation score, CLASS:
1419 *Parasponia-Trema* gene homology classification (nv = orthologous pair not validated by
1420 whole-genome alignments), FC: gene expression log fold-change gene expression in nodule
1421 versus roots, P: P-value adjusted for multiple testing based on false discovery rate
1422 estimation. Genes that are putatively orthologous to legume and actinorhizal genes with
1423 symbiotic function are classified as 'LEGUME SYMBIOSIS GENE'. Conserved CNVs
1424 between *Parasponia* and *Trema* spp are in bold.

1425

1426 **Supplementary Table 11:** Sequenced RNA samples.

1427 **Supplementary Table 12:** GenBank Accession numbers of sequences.

1428 Tab 1: sequences used in phylogenetic reconstructions of Cannabaceae. Tab 2: sequences
1429 of genes with copy number variants. Sequences generated for this study are in bold;
1430 pseudogenes are marked with grey background.

1431

1432 **Supplementary Data File 1: Phylogenetic analysis of *Parasponia andersonii* and *Trema***
1433 ***orientalis* RG33 putative orthologs of legume genes that function in symbioses.**

1434 Phylogenetic trees (based on Neighbour Joining) of orthogroups containing published genes
1435 that function in symbiosis (see also Supplementary Table 1). OrthoGroup number
1436 corresponds to Supplementary Table 7; Node labels indicate bootstrap support values.
1437 Legume symbiosis genes and symbiotic homologs from actinorrhizal species are marked in
1438 bold; proteins from *Arabidopsis thaliana* (AT) are marked in green; *Eucalyptus grandis*
1439 (Eucgr) in olive; *Populus trichocarpa* (Potri) in light blue; *Medicago truncatula* (Medtr) in
1440 purple; *Glycine max* (Glyma) in mint; *Fragaria vesca* (Fvesca) in pink; *P. andersonii* (Pan) in
1441 dark blue; and *T. orientalis* (Tor) in dark red. Legume or actinorrhizal symbiosis genes from
1442 species not included in the orthogroup inferences are in black. Agl: *Alnus glutinosa*; Cgl:
1443 *Casuarina glauca*; Dgl *Datisca glomerata*; Lja: *Lotus japonicus*; Msa: *Medicago sativa*; Mtr:
1444 *Medicago truncatula*; Phy: *Petunia hybrida*; Psa: *Pisum sativum*; Pvu: *Phaseolus vulgaris*.

1445

1446 **Supplementary Data File 2: Phylogenetic analysis of genes utilized in *Parasponia* and**
1447 **medicago root nodules.** Phylogenetic trees (based on Neighbour Joining) of orthogroups
1448 containing genes with significantly enhanced expression level in any of three *Parasponia*
1449 nodule developmental stages (Supplementary Fig. 11; Supplementary Table 9) as well as
1450 genes with significantly enhanced expression in nodules of medicago³¹ (Supplementary

1451 Table 10). OrthoGroup number corresponds to Supplementary Table 7; Node labels indicate
1452 bootstrap support values. Nodule-enhanced genes are marked in bold; proteins from
1453 *Arabidopsis thaliana* (AT) are marked in green; *Eucalyptus grandis* (Eucgr) in olive; *Populus*
1454 *trichocarpa* (Potri) in light blue; *Medicago truncatula* (Medtr) in purple; *Glycine max* (Glyma)
1455 in mint; *Fragaria vesca* (Fvesca) in pink; *P. andersonii* (Pan) in dark blue; and *T. orientalis*
1456 (Tor) in dark red. Agl: *Alnus glutinosa*; Cgl: *Casuarina glauca*; Dgl *Datisca glomerata*; Lja:
1457 *Lotus japonicus*; Msa: *Medicago sativa*; Mtr: *Medicago truncatula*; Phy: *Petunia hybrida*; Psa:
1458 *Pisum sativum*; Pvu: *Phaseolus vulgaris*.