Estimating Divergence Times and Substitution Rates in Rhizobia

Rim Chriki-Adeeb and Ali Chriki
Département de Biologie, Laboratoire de Génétique, Faculté des Sciences de Bizerte, Jarzouna, Tunisie.

ABSTRACT: Accurate estimation of divergence times of soil bacteria that form nitrogen-fixing associations with most leguminous plants is challenging because of a limited fossil record and complexities associated with molecular clocks and phylogenetic diversity of root nodule bacteria, collectively called rhizobia. To overcome the lack of fossil record in bacteria, divergence times of host legumes were used to calibrate molecular clocks and perform phylogenetic analyses in rhizobia. The 16S rRNA gene and intergenic spacer region remain among the favored molecular markers to reconstruct the timescale of rhizobia. We evaluate the performance of the random local clock model and the classical uncorrelated lognormal relaxed clock model, in combination with four tree models (coalescent constant size, birth–death, birth–death incomplete sampling, and Yule processes) on rhizobial divergence time estimates. Bayes factor tests based on the marginal likelihoods estimated from the stepping-stone sampling analyses strongly favored the random local clock model in combination with Yule process. Our results on the divergence time estimation from 16S rRNA gene and intergenic spacer region sequences are compatible with age estimates based on the conserved core genes but significantly older than those obtained from symbiotic genes, such as nodD genes. This difference may be due to the accelerated evolutionary rates of symbiotic genes compared to those of other genomic regions not directly implicated in nodulation processes.

KEYWORDS: rhizobia, 16S rRNA gene, ITS, nodD genes, random local clock, uncorrelated lognormal clock, Bayes factors

Introduction
Nitrogen-fixing bacteria in legume nodules collectively named rhizobia are currently classified into several genera within the α-proteobacteria class, including Agrobacterium, Rhizobium, Allorhizobium, Aminobacter, Azorhizobium, Devosia, Mesorhizobium, Methylobacterium, Microvirga, Ochrobacterium, Phyllobacterium, Sinella, Sinorhizobium (syn. Ensifer), and Bradyrhizobium. Recently, Mousavi et al. proposed the genus Neorhizobium for the Rhizobium galegae complex (including R. galegae, Rhizobium vignae, Rhizobium bauletense, and Rhizobium alkalisolii) that formed a separate clade in the family Rhizobiaceae. Members of the genus Agrobacterium are predominantly soil-inhabiting and plant-associated bacteria. Some Agrobacterium strains may carry symbiotic plasmid and have nodulating activity on legume plants.3,4

Until the isolation of legume-nodulating bacteria species of the genera Burkholderia and Cupriavidus in the β-proteobacteria class in 2001,5,6 it has been assumed that rhizobia were limited to α-proteobacteria. Although most legumes are symbiotic with α-proteobacteria species (α-rhizobia), several legumes are nodulated by β-proteobacteria (β-rhizobia).7–9

Nodule formation and nitrogen fixation have been well studied in rhizobia, and different symbiotic genes, such as nod and nif genes, are known.1 Different host specificities are determined by the symbiotic gene content.1,10,11 Moreover, rhizobia symbiovars1 have been used to distinguish symbiotically distinct subgroups within a single rhizobial species, such as Rhizobium leguminosarum, Rhizobium etli, and also Rhizobium gallicum (a closely related species to Rhizobium sullae).12

After over two decades of polyphasic characterization,12,13 R. sullae (syn. Rhizobium hedysari) has been defined as the host-specific symbiont of sull (Hedysarum coronarium). Some strains of R. sullae are able to nodulate other Hedysarum legumes, such as Hedysarum spinosissimum and Hedysarum flexuosum.14,15

Although the evident role of the symbiotic gene content is to determine the host specificity,1 the taxonomic and phylogenetic studies of rhizobia are mainly based on the highly conserved 16S rRNA gene.16 However, multilocus sequence analysis of housekeeping genes is thought to be a more powerful approach for resolving some of the taxonomic issues.17,18 The 16S–23S rRNA intergenic spacer (ITS) region was also used in combination with the 16S rRNA gene to study the phylogenetic relationships between rhizobia.19–21 In this study, we determined the nucleotide sequences of both 16S rRNA and ITS genes from 150 species of herbaceous legumes with a wide natural distribution.
The species *H. coronarium*, synonym *Sulla coronaria*, is distributed within the Mediterranean basin from Northern Africa to Southern Spain and centrally to Southern Italy. *H. spinosi- sum* sp. capitatum, synonym *Sulla capitata*, is an indigenous arid and a semi-arid forage plant adapted to desert rangelands in Africa and the Middle East. Finally, *Hedyasperum carnosum*, synonym *Sulla carnosa*, is an endemic species distributed in the arid and semi-arid regions of Tunisia and Algeria. Our interest in the three *Hedyasperum* species arises from their proven forage value under arid, semi-arid, and subhumid conditions in Tunisia.

The genus *Hedyasperum* is a member of the tribe Hedysareae, which is included in the inverted repeat-lacking clade group. The tribe Hedysareae (including genera, such as *Taverniera, Hedyasperum, Albagi, Onosyrichis*, and *Caragana*) is a sister group to the Astragalalean clade, which includes genera, such as *Astragalus, Oxytropis*, and *Colutea*. The most recent common ancestor of Hedysareae and Astragalalean clade originated between 25.0 and 39.2 million years ago (Mya), and the divergence time between *Caragana* and *Hedyasperum* was estimated as 29.3 ± 3.0 Mya. The later dating estimate was used in this study to calibrate molecular clocks for rhizobia.

In order to calibrate molecular clocks for estimating the age of bacterial lineages, the codivergence of endosymbiotic bacteria with their host species is used. The concordance between the molecular phylogenies of the bacteria and their hosts permits the application of the hosts’ fossil record to their endosymbionts. This so-called primary calibration method for estimating the divergence times of bacteria has been applied in different works and also by Turner and Young to estimate the divergence times in rhizobia using core genes that code for two related forms, GSI and GSII, of the glutamine synthetase. These studies were based on a limited sampling within the α-proteobacteria class and, thus, do not provide reliable approximations for the crown node age and early diversification history. Secondary calibration schemes in which the primary fossils are not included in age estimates were also used in rhizobia. Recently, Aoki et al. used nod genes to estimate the divergence time between α- and β-rhizobia. However, most estimates of rhizobia ages have focused only on a limited number of genera without including *Agrobacterium* genus and other related taxa.

Therefore, the purpose of this study was to determine the divergence times between the main group of rhizobia, including *Agrobacterium* and *Neorhizobium* genera. Because the *Agrobacterium* strains were isolated from root nodules of *Sulla* legumes, the age of *Hedyasperum* clade (29.3 ± 3.0 Mya) was used as a calibration point to perform the molecular phylogenetic analyses in rhizobia. Both 16S rRNA gene and ITS region sequences were used in this study instead of core genes (proteins), which often have differed from the analyses of rRNA genes leading to an overall uncertainty in prokaroyote phylogeny.

Despite its use as a barcode for bacteria, the 16S rRNA sequence often fails to provide sufficient information for species-level identification. In contrast, the ITS fragment located between 16S and 23S genes in fast-growing rhizobia is the most hypervariable chromosomal region and has been recognized as providing the superior resolution of closely related bacterial taxa.

It is well established that the estimating divergence times in phylogenies using a molecular clock depends on the accurate modeling of nucleotide substitution rates in DNA sequences. Also, the assumption that nucleotide substitutions accumulate at a constant rate over time (strict molecular clock) is often rejected in favor of variable rate hypotheses (relaxed molecular clocks), among them the uncorrelated log-normal (UCLN) model and the random local clock (RLC) model are used.

In this study, we employ Bayesian phylogenetic analyses using the RLC and UCLN clock substitution models for analyzing the divergence times in rhizobia. A maximum likelihood (ML) method integrated in RelTime program was also used for comparisons. We principally focus on the following three questions: (i) how different are the divergence ages estimated with a random local clock model or an UCLN clock model; (ii) what is the best-fit molecular clock model for the dataset used; and (iii) can divergence times be inferred by using host legume ages as an alternative calibration instead of a fossil record.

**Materials and Methods**

**Phylogenetic inference.** Specimens of the legume hosts growing spontaneously in different Tunisian regions were sampled. *S. capitata* plants were harvested near Sousse, at the Kanatoui locality (35°53′N, 10°34′E; climate semiarid). *S. carnosa* was collected at the EL-Alam region (35°48′N, 10°8′E; climate arid). *S. coronaria* was harvested at the Bizerte locality (37°29′N, 9°45′E; climate sub humid).

In order to analyze the phylogenetic position of our native isolates by the molecular approaches currently used for bacterial species definition, the ITS region was sequenced in addition to the 16S rRNA gene sequences. Methods of amplification and sequencing of 16S rRNA gene and ITS region were described previously. The sequences obtained were compared with available 16S rRNA gene and ITS region sequences retrieved from the GenBank using the BLAST program (http://www.ncbi.nlm.nih.gov/blast/) to determine an approximate phylogenetic affiliation (Table 1). Percent identity between sequences was estimated using the FASTA programs (http://fasta.bioch.virginia.edu/fasta www2 fasta www.cgi).

Supplementary sequences of related species were retrieved from GenBank (Table 2) and included in the current phylogenetic analysis. A total of 31 sequences were aligned for 16S rRNA gene and ITS region using Muscle (with default settings) integrated in MEGA6.06. For each rhizobial species, the same strain was used for both
molecular markers. Final alignments included 1624 sites for the 16S rRNA gene and 1442 characters for the ITS region. However, these alignments were concatenated using BioEdit v7.2.5 because RelTime program did not perform analyses on partitioned data.

The best-fit nucleotide substitution model was selected according to the Akaike information criterion (AIC) using MEGA6. The model GTR + G + I was retained for the concatenated dataset. Therefore, it was applied for the ML phylogenetic analysis as well as for all further analyses conducted on concatenated data. The Bayesian inference (BI) was achieved using the program MrBayes v3.2.1.

Three β-proteobacteria species (Ralstonia solanacearum LMG 2299, Burkholderia cepacia strain 8111, and B. cepacia strain 8201) and a γ-proteobacteria strain (Hcar02) were used as outgroup taxa for the α-rhizobia species included in this study. Bayesian analyses used two sets of four simultaneous chains (three cold and one heated; the default setting in MrBayes) and one million generations.

Table 1. List of bacterial strains isolated from root nodules of three species of the genus *Sulla* (Fabaceae).

| STRAIN | GENBANK ID (16S/ITS) | HOST PLANT | CLOSEST SPECIES | GENBANK ID* |
|--------|----------------------|------------|-----------------|-------------|
| Hc01   | JN944178/JN944179    | S. coronaria | *R. nepotum*    | KP762553    |
| Hc04   | JN944189/JN944182    | S. coronaria | *R. huautlense* | AM237359    |
| Hsc01  | JN944190/JN944183    | S. capitata  | *R. sullae*     | Y10170      |
| Hsc02  | JN944191/JN944184    | S. capitata  | *A. rubi*       | NR_113608   |
| Hsc03  | JN966893/JN966897    | S. capitata  | *A. rubi*       | NR_113608   |
| Hcar01 | JN944192/JN944185    | S. carnosa   | *R. pusense*    | KF676889    |
| Hcar02 | JN944193/JQ081302    | S. carnosa   | *P. agglomerans*| HM038120    |

Note: *For 16S.

Table 2. List of reference strains used in this study.

| GENUS         | SPECIES        | STRAIN         | ACCESSION NUMBER (16S/ITS) |
|---------------|---------------|----------------|---------------------------|
| *Ralstonia*   | *solanaeearum*| LMG 2299       | EF016361/KC756967         |
| *Burkholderia*| *cepacia*     | 8201           | FJ870554/FJ870554         |
| *Burkholderia*| *cepacia*     | 8111           | FJ870551/FJ870551         |
| *Rhizobium*   | *leguminosarum*| LPB0205      | GQ863505/GQ863516         |
| *Rhizobium*   | *phaseoli*    | ATCC 14482     | EF141340/EF141341         |
| *Rhizobium*   | *fabae*       | CCBAU 3320     | DQ835306/FJ392873         |
| *Rhizobium*   | *rhizogenes*  | A4             | AB247607/AB247607         |
| *Rhizobium*   | *huautlense*  | OS-49.b        | AM237359/AF345270*        |
| *Rhizobium*   | *galegae*     | CCBAU 05104    | HM070174/EU418348         |
| *Rhizobium*   | *galegae*     | LMG 6214       | X67226/AF345265           |
| *Rhizobium*   | *giardini*    | H 152          | U63444/AF345268           |
| *Rhizobium*   | *giardini*    | CCBAU 85040    | EU256415/EU288740         |
| *Rhizobium*   | *hainanense*  | I66            | U71078/AF321872           |
| *Rhizobium*   | *tropici*     | CCBAU 25295    | EU399715/EU418365         |
| *Rhizobium*   | *mesosinicum* | CCBAU 25010    | NR_043548/EU120729        |
| *Rhizobium*   | *mesosinicum* | CCBAU 25217    | EU070130/EU120730         |
| *Rhizobium*   | *gallicum*    | CCBAU 85013    | EU256408/EU288729         |
| *Rhizobium*   | *mongolense*  | CCBAU 05122    | EU399697/EU418349         |
| *Sinorhizobium*| *fredii*      | USDA 194       | AB433352/EU152398         |
| *Sinorhizobium*| *melloti*     | CCBAU 05183    | EU399710/EU418357         |
| *Mesorhizobium*| *tarimense*   | CCBAU 83306    | EF035058/EF050771         |
| *Mesorhizobium*| *lot*         | LMG 6125       | X67229/AF345260           |
| *Bradyrhizobium*| *japonicum*  | USDA 110       | AF363150/AB100749         |
| *Bradyrhizobium*| *elkanii*     | C8-1780        | AB513452/AB513476         |

Note: *Strain LMG 18254 for ITS.
sampled trees every 1000 generations and assessed convergence by examining the standard deviation of split frequencies within the output. Conservatively, the first 25% of the sampled trees were discarded as burn in, and the remaining 75% of the sampled trees were used to calculate the Bayesian posterior probabilities (PPs). A majority-rule consensus tree was used to summarize trees sampled from the stationary posterior by using the sumt command. The Bayesian inferred (BI) tree was then used as a fixed phylogeny for further analyses.

**Divergence times.** We first tested for the violation of the molecular clock using a likelihood ratio test (LRT)\(^44\) with the ML-inferred tree. Likelihoods’ values were estimated using baseml in PAML v4.8\(^45\) under rate constant and rate variable models and used to compute the LRT statistic according to the following equation:

\[
LRT = -2(\log L_0 - \log L_1)
\]

In this equation, \(L_1\) is the unconstrained (nonclock) likelihood value, and \(L_0\) is the value obtained under the rate-constancy assumption. LRT is distributed approximately as a chi-square random variable with \((m - 2)\) degrees of freedom (df), \(m\) being the number of branches.

To assess the impact of molecular clocks and tree models on divergence time estimation, two clock models (UCLN and RLC) were analyzed in combination with three tree models (constant size, pure birth, and birth–death processes). BEAST program v1.8.2\(^46\) on CIPRES Science Gateway\(^47\) was used to estimate the divergence dates under each pair of clock/tree model combinations. All analyses were conducted on concatenated data under GTR + G + I model with five gamma categories. Monophyly was enforced for two taxon sets: (i) all taxa except for \(\gamma\)- and \(\beta\)-proteobacterial taxa (used for rooting the tree) and (ii) \(R.\) gallicum and \(R.\) mongolense (in order to calibrate the tree via the split between this clade and \(R.\) sullae).

For each BEAST analysis, we ran two to four independent Markov chains Monte Carlo runs of 10–20 million generations and sampled every 1000 generations. For each run, 2.5–5 million generations were discarded as burn-in. Numbers of runs and replicates are variable whether UCLN or RLC model is used. Convergence of chains was checked using TRACER v1.6\(^48\) throughout the effective sample size quantification, which ensures convergence with effective sample size values above 200 for each parameter of the different models tested. The remaining generations for each run were combined with LOGCOMBINER and were used to construct the maximum clade credibility tree and the associated combined nodal posterior probabilities (HDP) distributions around the estimated node ages using TREEANNOTATOR v1.8.2 in BEAST package,\(^46\) and visualized using FigTreev1.4.3pre.\(^49\)

Relative fit of tree priors and clock models was evaluated using the values of \((\log_\theta)\)Bayes factors’ (BFs),\(^50\) calculated from the \((\log)\)marginal likelihoods (mlls) estimated by path sampling (PS)\(^51,52\) and stepping-stone sampling (SS)\(^51,52\) methods implemented in BEAST. We followed the method of Raftery\(^50\) in interpreting BFs in terms of decision making that is a BF value between 0 and 1 is not worth more than a bare mention, whereas a BF value between 1 and 3 is considered to give positive evidence favoring the model with the higher \((\log)\)ml. Values higher than 3 and 5 are considered to give strong and very strong evidence, respectively, in favor of the model with the higher \((\log)\)ml.

For comparison with the Bayesian divergence time estimation, we recently used a maximum likelihood (ML) approach integrated in the RelTime\(^58\) program implemented in MEGA6.\(^41\) This ML method produces branching times in the phylogenetic tree without the need of a prior selection of statistical distributions to model the heterogeneity of rates among branches and does not require reliable knowledge of a prior divergence times.

For the RelTime analysis that needed the use of a start tree as input, we used the same sequence alignment and the same substitution model as that of BEAST analyses. All our ML and Bayesian dating analyses were based on the same calibration point. The divergence time of \(R.\) sullae, deduced from the split of the Hedysaroid clade (comprising Caragana and \(H.\) hedysarum genera) at 29.3 ± 3.0 Mya,\(^23\) was used to calibrate the molecular clocks and estimate the absolute divergence times for the main groups of rhizobia.

In order to evaluate the impact of molecular markers on divergence time estimates, we used 16S and \(nod\)I\(J\) gene sequences selected from the same \(\alpha\)- and \(\beta\)-rhizobia taxa included on the phylogenetic analyses conducted by Aoki et al.\(^29\) Bayesian phylogenetic analyses using BEAST were performed separately on 16S rRNA and \(nod\)I\(J\) gene sequences under the same clock model and tree model combination that has been selected for our 16S and ITS dataset (see “Results” section). The divergence time estimate of 26.5 Mya (±10%) for \(R.\) leguminosarum (obtained in this study) was used to calibrate the molecular clocks instead of \(R.\) sullae, which is not yet characterized for \(nod\)I\(J\) genes. The degree of similarity between 16S and \(nod\)I\(J\) phylogenies was measured using the combined Mantel\(^53\) and the distance-based congruence among distance matrices (CADM) tests.\(^53\) The Log-Det DNA distance\(^55\) giving symmetrical matrix was used.

**Results**

**Phylogenetic inference.** Because the main objective of this study consisted to estimate divergence times between rhizobia, only seven native strains (Hc01, Hsc02, Hsc03, Hcar01, Hcar02, Hc04, and Hsc01) identified in \(Sulla\) legumes (Table 1) were included to avoid the redundancy of sequences and correctly detect evolutionary events.

For 16S rRNA gene, pairwise sequence identity analyses showed that Hsc01 (JN944190) was most similar to \(R.\) sullae-type strain IS123T (Y10170) with a sequence identity of 99.4%. The strain Hc04 (JN944189) was affiliated to \(R.\) huautlense.
isolate OS-49.b (AM237359) with a sequence identity of 97.8%. The other native strains belonged to the *Agrobacterium tumefaciens* complex with relatively high sequence identity values (100% between Hc01 and *Rhizobium nepotum* strain IHB B 13640, 98.2% between Hsc02 and *Agrobacterium rubi* strain NBRC 13261, 98.9% between Hsc03 and *A. rubi* NBRC 13261, and 100% between Hc01 and *Rhizobium pusense* strain SM-T3). The strain Hc02 isolated from *S. carnosa* exhibited a sequence identity of 99.9% with the strain TW1 of *Pantoea agglomerans* (HM038120).

A similar pattern but with lower sequence identity values emerged in parallel comparisons among the ITS region sequences. The strains Hc01, Hsc02, Hsc03, and Hc01 were included in the *A. tumefaciens* complex, with sequence identity values ranging from 90.2% (between Hsc02 and *A. tumefaciens* strain MAFF 03-0122) to 97.9% (between Hc01 and *A. tumefaciens* strain MAFF 03-01222). The lowest sequence identity (72.0%) was obtained between Hc04 (JN944182) and *Rhizobium giardinii* strain CCBAU 85014 (EU288738). A sequence identity value of 75.9% was obtained between Hsc01 (JN944183) and *R. mongolense* strain CCBAU 05122 (EU418349), whereas Hc02 (JQ081302) belonged to *P. agglomerans* (EU306596) with a sequence identity of 93.3%. It should be noted that most native strains produced one band in polymerase chain reaction amplification of the ITS region. Only Hc02 produced two types of sequences (Hc02a and Hc02b) with 97.2% identity.

In order to confirm the molecular affiliation of the native endosymbionts isolated from the root nodules of *Sulla* legume plants, some related α- and β-proteobacteria sequences (Table 2) were added. A combined approach that included Bayesian and ML inference of phylogeny was applied using the concatenated dataset that includes 31 taxa, with an aligned length of 3066 bp (including gaps). The ML construction (Fig. 1) and BI tree (Supplementary Fig. 1) differed only in the placement of the native strain Hc04, which belonged to *R. giardinii* sequences and *Agrobacterium* strains in the ML tree (Fig. 1). Otherwise, these phylogenetic trees were well congruent topologically and showed strong supports with ML bootstrap scores (BS) >90 and BI PP >95 for most nodes. Because independent Bayesian and bootstrap analyses of the dataset yielded congruent topologies, we used the ML-inferred tree (Fig. 1) to define the following well-supported clades:

- **β-Proteobacteria clade** (BS 100), containing *R. solanacearum* LMG 2299, *B. cepacia* strain 8111, and *B. cepacia* strain 8201;
- **Bradyrhizobium clade** (BS 100), including *Bradyrhizobium japonicum* USDA 110 and *Bradyrhizobium elkanii* C8-1780;
- **Mesorhizobium clade** (BS 100), including *Mesorhizobium tarinense* CCBAU 83306 and *Mesorhizobium loti* LMG 6125;

![Figure 1](https://example.com/figure1.png)

**Figure 1.** ML-inferred tree for rhizobia based on the concatenated dataset of 16S rRNA gene and ITS region sequences. Bootstrap probabilities based on 500 replicates are indicated at each node. The scale bar represents the number of nucleotide substitutions per site.

---

**Table 2.** Related α- and β-proteobacteria sequences included in the ML and Bayesian phylogenetic analysis.
– Sinorhizobium clade (BS 100), containing Sinorhizobium fredii USDA 194 and Sinorhizobium melliloti CCBAU 05183;
– Agrobacterium clade (BS 100), including our native strains Hc01, Hcar01, Hsc02, and Hsc03;
– Neorhizobium clade (BS 100), containing R. huautlense, R. galegae CCBAU 05104, and R. galegae LMG 6214;
– Rhizobium clade (BS 76), containing the majority of Rhizobium species except for Hc04 and R. giardinii, which were associated with the Agrobacterium clade (Fig. 1).

**Divergence times.** The strict clock hypothesis was rejected on the basis of the log likelihood test (LRT) (log $L_0 = -6409.773$; log $L_1 = -6352.013$; degrees of freedom = 25; $\chi^2 = 115.52; P < 0.001$). This finding validates the use of relaxed molecular clock approach to estimate the node ages throughout Bayesian analyses by BEAST. Different clock/tree model combinations were tested on BEAST. We particularly focused on RLC and UCLN clock models, in combination with the following four tree priors: coalescent constant size model, Yule process, birth–death process, and birth–death incomplete sampling model.

The divergence times estimated for different rhizobial groups under the height combinations of clock/tree models are reported in Table 3. Here, we explore the impact of these various relaxed-clock models on estimated divergence times by focusing on the inferred ages of six key nodes: (1) the root node, corresponding to the split of α-rhizobia; (2) the split of Bradyrhizobium; (3) the split of Mesorhizobium; (4) the split between Sinorhizobium and Rhizobium/Agrobacterium groups; (5) the split between Rhizobium and Agrobacterium; and (6) the divergence between Neorhizobium and Rhizobium.

Except for the split between Rhizobium and Neorhizobium, the UCLN clock model yielded younger ages than the RLC one (Table 3). Fluctuations in divergence time estimation across tree models were also noted within each clock model group. For both RLC and UCLN models, the constant size tree model invariably yielded the oldest ages for all the key nodes considered (Table 3).

These patterns of fluctuation in divergence time estimation were confirmed by a substantial variation of the inferred substitution rates across clock and tree models (Table 4). The RLC model produced low substitution rates in the range of 0.0012 substitutions/site/million years (My) for constant size model and 0.0018 substitutions/site/My for both Yule and birth–death processes. In contrast, the UCLN clock model yielded high level of substitution rates varying between 0.0018 substitutions/site/My for constant size model and 0.0025 substitutions/site/My for Yule process.

The (log)mls of the UCLN and RLC models, in combination with the four tree models, are shown in Table 5. PS and SS gave similar results. So, only (log)mls estimated by the SS method were used to perform BF tests. A total of 28 BF tests were performed to compare all pairs of clock/tree model combinations. The log(BF) values are reported in Table 6. For models involving UCLN clock, the birth–death incomplete sampling process is significantly better than the three other tree models (BF > 5), whereas Yule process in combination with RLC model is favored (BF > 5) in all comparisons involving RLC as well as UCLN clock models (Table 6). Based on these BF's, we therefore retained RLC and Yule tree models for all our divergence time analyses conducted on BEAST. A ML method of divergence time estimates implemented in using RelTime program was also used for comparisons.

Divergence times between α-, β-, and γ-proteobacteria strains included in this study are shown in Figure 2 and Supplementary Figure 2. Specific age estimates for the root node and for the divergence between the main groups of rhizobia that were inferred from both BEAST (under RLC model and Yule tree prior) and RelTime analyses are given in Table 3. Inclusion of earlier diverging taxa from γ- and β-proteobacteria allowed us to date the root of rhizobial genera at about 603 Mya (95% HPD = 281–998 Mya) and 875 Mya (confidence interval

---

**Table 3.** Ages in million years ago for rhizobial group divergences estimated with BEAST under different combinations of clock/tree models (Brady: Bradyrhizobium; Meso: Mesorhizobium; Sino: Sinorhizobium; Agro: Agrobacterium; Neorhizo: Neorhizobium).

| CLOCK MODEL | TREE MODEL | ROOT   | BRADY | MESO   | SINO   | AGRO   | NEORHIZO |
|-------------|------------|--------|-------|--------|--------|--------|----------|
| RLC         | Birth-Death| 639.72 | 424.73| 372.00 | 195.17 | 142.42 | 53.28    |
|             | Yule       | 603.46 | 384.62| 343.96 | 200.87 | 149.07 | 54.50    |
|             | Birth-Death| 892.97 | 497.03| 429.90 | 198.83 | 144.95 | 52.64    |
|             | Constant Size| 1166.42| 618.06| 505.17 | 221.86 | 160.87 | 56.54    |
|             | Average    | 825.64 | 481.11| 412.76 | 204.18 | 149.33 | 54.24    |
| UCLN        | Birth-Death| 522.11 | 264.81| 214.58 | 140.98 | 104.19 | 76.94    |
|             | Birth-Death| 529.85 | 269.87| 217.60 | 141.95 | 104.68 | 76.47    |
|             | Constant Size| 774.76| 332.17| 258.43 | 160.58 | 117.29 | 86.02    |
|             | Yule       | 462.85 | 255.70| 211.20 | 143.06 | 105.46 | 78.21    |
|             | Average    | 572.39 | 280.64| 225.45 | 146.64 | 107.91 | 79.41    |

**Note:** *Birth–death (incomplete sampling).*

**Abbreviations:** RLC, random local clock; UCLN, uncorrelated lognormal.
Divergence times and diversification rates in rhizobia

Evolutionary Bioinformatics 2016:12

Table 4. Mean substitution rates estimated by BEAST under RLC (random local clock) and UCLN (uncorrelated lognormal) clock combined with four tree models.

| CLOCK MODEL¹ | TREE MODEL | MEAN RATE* | CONFIDENCE INTERVAL |
|--------------|------------|------------|---------------------|
| RLC          | Birth-Death | 0.0018     | (0.0008, 0.0028)    |
|               | Yule       | 0.0018     | (0.0009, 0.0028)    |
|               | Birth-Death** | 0.0014     | (0.0007, 0.0021)    |
|               | Constant Size | 0.0012     | (0.0005, 0.0019)    |
|               | Average    | 0.0015     |                     |
| UCLN         | Birth-Death** | 0.0023     | (0.0009, 0.0041)    |
|               | Birth-Death | 0.0023     | (0.0010, 0.0038)    |
|               | Constant Size | 0.0018     | (0.0008, 0.0029)    |
|               | Yule       | 0.0025     | (0.0009, 0.0044)    |
|               | Average    | 0.0022     |                     |

Notes: *(Mean) substitution rate unit: substitutions/site/million years (My). **Birth–death (incomplete sampling).

Table 5. (Log)marginal likelihood (ml) estimations performed with BEAST using the path sampling (PS) and stepping-stone sampling (SS) methods under different combinations of clock/tree models.

| CLOCK MODEL¹ | TREE MODEL | LOG ML (PS) | LN ML (SS) |
|--------------|------------|-------------|------------|
| UCLN         | Constant size | −22988.18   | −22988.35  |
|               | Birth-death | −22989.07   | −22989.28  |
|               | Birth-death* | −22349.77   | −22349.70  |
|               | Yule       | −22992.10   | −22992.18  |
| RLC          | Constant size | −21089.77   | −21089.87  |
|               | Birth-death | −20265.45   | −20265.47  |
|               | Birth-death* | −19934.40   | −19934.44  |
|               | Yule       | −19574.18   | −19574.14  |

Notes: *Birth–death (incomplete sampling).
Abbreviations: UCLN, uncorrelated lognormal; RLC, random local clock.

[CI] = 154–1596 Mya) from BEAST and RelTime analyses, respectively (Table 7). Both estimates are younger than the minimum time for the divergence of α- and β-proteobacteria of 1640 Mya suggested by Battistuzzi and Hedges.³⁰

Overall, date estimates determined using BEAST were similar to those estimated using RelTime program given the overlapping of the CIs for the two algorithms (Table 7). However, the CIs determined using RelTime were more conservative (wider) than the 95% high posterior density (HPD) credibility intervals associated with the Bayesian analysis using BEAST (Table 7). This difference may lie in the fact that BEAST uses relaxed clock approach to derive the posterior of rates and times and allows the specification of different types of calibration distributions to model calibration uncertainty (Rutschmann and Strijk et al.). In contrast, RelTime uses user-supplied minimum and maximum age constraints as starting priors to determine the node age estimates.

Table 6. Model comparisons using Bayes factors calculated from marginal likelihoods (mls) in BEAST.

| MODEL* | UCLN (CS) | UCLN (BD) | UCLN (Y) | RLC (CS) | RLC (BD) | RLC (Y) |
|--------|-----------|-----------|----------|----------|----------|---------|
| UCLN (CS) | 1.00*     | 1.00      | 1.00     | 1.00     | 1.00     | 1.00    |
| UCLN (BD) | 0.00      | −0.00     | −0.00    | 0.00     | 0.00     | 0.00    |
| UCLN (Y)  | 0.00      | −0.00     | −0.00    | 0.00     | 0.00     | 0.00    |
| RLC (CS)  | 0.00      | −0.00     | −0.00    | 0.00     | 0.00     | 0.00    |
| RLC (BD)  | 0.00      | −0.00     | −0.00    | 0.00     | 0.00     | 0.00    |
| RLC (Y)   | 0.00      | −0.00     | −0.00    | 0.00     | 0.00     | 0.00    |

Notes: *Birth–death (incomplete sampling).
Abbreviations: UCLN, uncorrelated lognormal; RLC, random local clock.
As inferred by both ML and Bayesian methods, \(\alpha\)-rhizobia began diversifying around 385–348 Mya with the split of the genus *Bradyrhizobium* from the remaining taxa (Fig. 2 and Table 7), succeeded by the split of the genera: *Mesorhizobium* (344–285 Mya) and *Sinorhizobium* (201–140 Mya). It seems that the divergence of the *Sinorhizobium* genus coincided with the debut of the Jurassic period, which corresponds to the middle segment of the Mesozoic Era (199.6–145.5 Mya), whereas the divergence of the *Neorhizobium* genus (71–55 Mya; Table 7) occurred at Paleocene (66–55.8 Mya). The *Rhizobium* genus began diversifying in Cretaceous (145.5–65.5 Mya) with the split of *Agrobacterium* complex around 149–100 Mya (Table 7). The analysis of the chronograms (Fig. 2 and Supplementary Fig. 2) showed that most of the *Rhizobium* species diverged later (between 47 and 29 Mya for *R. huautlense* and 26 and 19 Mya for *R. leguminosarum*), during the Eocene (56.0–33.9 Mya), Oligocene (33.9–23.03 Mya), and Miocene (5.33–23.03 Mya) periods that corresponded to the expansion of the Fabaceae family that began diversifying around 60 Mya.  

To evaluate the impact of the molecular markers on divergence time estimates, the 16S rRNA and *nod*IJ gene sequences were selected from the same \(\alpha\) - and \(\beta\)-rhizobia taxa included in the phylogenetic analyses conducted by Aoki et al. Bayesian phylogenetic analyses using BEAST were performed separately on 16S rRNA and *nod*IJ gene sequences under the RLC model and Yule tree prior combination that has been selected for our 16S and ITS dataset on the basis of BF

### Table 7. Dates (Mya) and 95% confidence intervals (CIs) for rhizobial groups estimated using Bayesian (BEAST) and ML (RelTime) methods.

| NODE            | BEAST* | 95% CI         | RELTIME* | 95% CI         |
|-----------------|--------|----------------|----------|----------------|
| Root            | 603.46 | (280.92, 998.22)| 874.78   | (153.59, 1596.0) |
| Bradyrhizobium  | 384.62 | (174.00, 661.31)| 347.71   | (61.02, 634.40)  |
| Mesorhizobium   | 343.96 | (159.78, 598.13)| 285.19   | (49.98, 520.40)  |
| Sinorhizobium   | 200.87 | (109.04, 319.75)| 140.10   | (24.41, 255.78)  |
| Agrobacterium   | 149.07 | (78.11, 236.36)| 100.23   | (17.12, 183.34)  |
| Neorhizobium    | 54.50  | (35.25, 79.53)| 71.38    | (12.26, 130.50)  |

Note: *The RelTime and BEAST estimates from the combined 16S rRNA gene and ITS region sequence analyses were converted into absolute times by using one calibration: *R. sullae* (29.3 ± 3.0 Mya).
test. The divergence time estimate of 26.5 Mya (±10%) for *R. leguminosarum* (obtained in this study) was used to calibrate the molecular clocks instead of *R. sullae* because the later species is not characterized for its *nodIJ* genes. Chronograms for 16S and *nodI* genes are, respectively, shown in Supplementary Figures 3 and 4. Results of this comparative study are also summarized in Table 8.

We noted that plasmid genes (e.g., *nodIJ* genes) and chromosomal markers (such as the 16S rRNA gene) yielded different ages (Table 8) despite the use of the same taxa (from Aoki et al.29), the same clock and tree models, and the same calibration point. The youngest age estimates were obtained for *nodIJ* genes that are directly implicated in the symbiotic relationship between rhizobia and host legume plants (Table 8). When concatenated with the ITS alignment, the 16S rRNA gene sequences gave moderate age estimates (Table 8). As shown in Table 8, these differences in divergence time estimation may be due to the variation of the substitution rates among molecular markers. The highest (mean) substitution rate of 3.2 × 10⁻³ substitutions/site/My was obtained for *nodIJ* marker against 2 × 10⁻⁴ substitutions/site/My for 16S rRNA gene and 1.8 × 10⁻³ substitutions/site/My for concatenated 16S rRNA gene and ITS region.

Variation of the substitution rates may also affect the inferred topologies from different molecular markers. To test for this hypothesis, the degree of similarity between 16S and *nodIJ* phylogenies was measured using the combined Mantel53 and the distance-based CADM tests.54 The Log-Det DNA distance55 giving symmetrical matrix was used.

The results of these tests showed the absence of congruence between the rRNA and *nodIJ* gene matrices (CADM Kendall’s coefficient *W* = 0.674, *P*-value = 0.001; CADM Mantel correlation *r* = 0.347, *P*-value = 0.001), and confirmed the hypothesis that incongruence between divergence time estimates is principally due to the type of the molecular markers used.

### Table 8. Divergence times estimated in rhizobia under random local clock (RLC) model in combination with Yule process from different molecular markers and calibration points.

| NODE          | MOLECULAR MARKER | 16S-ITS ² | 16S ² | *nodIJ* ² |
|---------------|------------------|-----------|-------|----------|
| *Bradyrhizobium*/*Mesorhizobium* | 384.62          | 441.92    | 75.32 |
| *Mesorhizobium*/*Sinorhizobium* | 343.96          | 283.69    | 63.23 |
| *Sinorhizobium*/Rhzobium* | 200.87          | 198.36    | 42.67 |
| Mean rate ³   | 0.0018          | 0.0002    | 0.0032 |

Notes: Calibration points: split of *R. sullae* at 29.3 ± 3.0 Mya for concatenated 16S–ITS marker and divergence of *R. leguminosarum* at 26.5 ± 2.6 Mya for 16S rRNA and *nodIJ* genes. * Concatenated 16S rRNA gene and ITS region sequences (this study). ² Concatenated 16S rRNA and *nodI* and *nodIJ* gene sequences from taxa used by Aoki et al. ³ Substitutions per site per million years.

Discussion

**Divergence times.** Several studies showed that *R. sullae* is the specific bacterial partner of *S. coronaria*13 and *S. spinosissima*.14 Host preferences appeared to be a general tendency in different plant taxa.38 Recently, Lemaire et al.59 noted that *Mesorhizobium* symbionts exhibit a general host preference for the tribe Psoraleeae. On the other hand, the same study showed that *Burkholderia* prevailed in the Podalyrieae family.

Host genotype may be the main factor determining rhizobial recruitment via specific chemical signaling between the symbiotic partners.30 The standing hypothesis is that the recognition of Nod factors by legume host plants is a “driving force in coevolution of both symbiotic partners and will result in host specificity”.60

Given the highly specific association between rhizobia and their host plants, we can assume that the record of divergence of the host legumes could be used to calibrate molecular clocks for rhizobia. Thus, it becomes possible to work around the lack of fossil data in bacteria. In this case, one solution is to calibrate a node with an age estimate from a previous molecular dating study that applied a fossil calibration. Compared to primary calibration strategies that rely on fossil records, age estimates obtained in molecular dating analyses relying on a secondary calibration point are younger.51 Although secondary calibration methods are subject to criticisms,62,63 they are used for divergence time estimation in different organisms including bacteria. A recent study carried out by Hipksley and Müller64 showed that 15% of over 600 analyses based on the molecular divergence dating methods used secondary calibration schemes.

In this study, we have used the time split of the genus *Sulla* to calibrate the molecular trees inferred from the Bayesian and ML analyses. The calibrated molecular dating of our rhizobia tree provides for the first time minimum age estimations for all major groups of α-rhizobia, including *Neorhizobium* and *Agrobacterium* genera (Table 7). The resulting BEAST chronogram (Fig. 2) shows that the diversification of extant lineages of α-rhizobia started with the split of the slow-growing *Bradyrhizobium* at about 385 Mya (95% CI [174, 661] Mya), whereas the divergence time between the fast-growing *Sinorhizobium* and *Rhzobium* has occurred at 201 Mya (95% CI [109, 320] Mya) (Table 7).

These divergence time estimates were slightly younger, but not significantly different, from those obtained by Turner and Young27 using two core genes and multiple calibration points. They estimated the divergence times for *Bradyrhizobium* to be between 507 and 553 Mya and for *Sinorhizobium* and *Rhzobium* genera in the range of 203–324 Mya.

In contrast, the age estimates obtained in this study were significantly older than those reported in Aoki et al.29 using the *nodIJ* gene sequences that have been selected from different α-proteobacterial taxa representing the main groups of rhizobia. These authors also included some β-proteobacterial
taxa in order to investigate the origin and evolution of the common nodulation genes nodIJ. Because the host legumes of these α- and β-proteobacteria ranged from Mimosoideae to Faboideae, Aoki et al.29 decided to set the basal divergence time of rhizobia at 60 Mya. This time prior constraint on the root would be the cause of the significant difference between the age estimates of rhizobia obtained in this study (Table 7 and Fig. 2 and Supplementary Fig. 2) and those of Aoki et al. (see Fig. 4 in this article).29

In addition to the eventual calibration effect on the divergence time estimation, the choice of the molecular markers themselves may be responsible for the incongruence between divergence time estimates among different phylogenetic studies. The nodIJ genes used as molecular markers by Aoki et al.29 are directly implicated in nodulation process in all α-rhizobia and also in some β-proteobacteria (called β-rhizobia) able to nodulate different legumes particularly of the Burkholderia, Cupriavidus, and Mimosas genera.6-9 For this reason, these molecular markers were often used in different phylogenetic analyses for addressing the origin of rhizobia instead of housekeeping and ribosomal RNA genes that do not interact with legumes.

The discordance between divergence time estimates across molecular markers may be due to the variation of nucleotide substitution rates among genomic regions. According to our phylogenetic analyses on BEAST, this hypothesis could not be rejected for the reason that the substitution rate estimate of 1.8 × 10^{-3} substitutions/site/My (for 16S rRNA gene and ITS region sequences; Table 8) was approximately half of the estimated rate for nodIJ genes (3.2 × 10^{-3} substitutions/site/My; Table 8).

Conclusion

The current phylogenetic analyses of rhizobia are based on the assumption that historical information (divergence times) on host legumes can be used to calibrate the molecular clocks. Our results from ribosomal markers are consistent with rhizobial divergence times inferred from core gene analyses. In contrast, our divergence time estimates were slightly older than those inferred from symbiotic genes. Additional ribosomal gene sequences together with housekeeping and symbiotic genes are therefore needed for timescale reconstruction in rhizobia.

Acknowledgment

We are grateful to Pr. Ahmed Landoulsi for help.

Author Contributions

Performed the experimental studies: RC-A. Performed the bioinformatics analyses: RC-A, AC. Wrote the first draft of the article: RC-A. Contributed to the writing of the manuscript: AC. Agreed the article results and conclusions: RC-A, AC. Both authors reviewed and approved the final article.

Supplementary Material

**Supplementary Figure 1.** Bayesian inference (BI) tree for rhizobia based on the concatenated dataset of 16S rRNA gene and ITS region sequences. PP values are indicated at each node. The scale bar represents the number of nucleotide substitutions per site.

**Supplementary Figure 2.** ML, time tree of rhizobia inferred from concatenated 16S rRNA gene and ITS region data under the GTR + G + I evolutionary substitution model. The RelTime estimates were converted into absolute times by using the divergence time of R. sullae (29.3 ± 3.0 Mya) as a calibration point. Node heights represent age estimates in Mya. The scale axis represents age estimates in Mya.

**Supplementary Figure 3.** BEAST divergence time estimates from 16S rRNA sequence data under RLC model and Yule tree process. Node heights represent age estimates in Mya. Divergence time of R. leguminosarum was used to calibrate the clock (26.5 ± 2.6 Mya). Methylcoccus capsulatus Bath is used as a outgroup. The scale axis represents age estimates in Mya.

**Supplementary Figure 4.** BEAST divergence time estimates from nodIJ sequence data under RLC model and Yule tree process. Node heights represent age estimates in Mya. Divergence time of R. leguminosarum (26.5 ± 2.6 Mya) was used to calibrate the clock. M. capsulatus Bath is used as an outgroup. The scale axis represents age estimates in Mya.

REFERENCES

1. Rogel MA, Ormeño-Orillo E, Martínez-Romero E. Symbiobars in rhizobia reflect bacterial adaptation to legumes. Syst Appl Microbiol. 2011;34:96–104.
2. Mousavi SA, Österman J, Wahlberg N, et al. Phylogeny of the Rhizobium–Allorhizobium–Agyrobacterium clade supports the delineation of Neorhizobium gen. nov. Syst Appl Microbiol. 2014;37:208–15.
3. Cummings SP, Gyneshwar P, Vinuesa P, et al. Nodulation of Sesbania species by Rhizobium (agrobacterium) strain IRBG74 and other rhizobia. Environ Microbiol. 2009;11:2510–25.
4. Zhao L, Fan M, Zhang D, et al. Distribution and diversity of rhizobia associated with wild soybean (Glycine soja Sieb. & Zucc.) in Northwest China. Syst Appl Microbiol. 2014;37:449–56.
5. Chen WM, Laeven S, Lee TM, et al.Ralstonia taiwanensis sp. nov., isolated from root nodules of Mimosas species and spurtum of a cystic fibrosis patient. Int J Syst Evol Microbiol. 2001;51:1279–35.
6. Moulou L, Munique A, Dreyfyus B, Boivin-Mosson C. Nodulation of legumes by members of the beta-subclass of Proteobacteria. Nature. 2003;411:948–50.
7. Chen WM, de Faria SM, Stralottio R, et al. Proof that Burkholderia forms effective symbioses with legumes: a study of novel Mimosas-nodulating strains from South America. Appl Environ Microbiol. 2007;71:7461–71.
8. Andam CP, Mondo SJ, Parker MA. Monophyly of nodA and nifH genes across Texan and Costa Rican populations of Cupriavidus nodulating symbionts. Appl Environ Microbiol. 2003;69:4686–90.
9. Mishra RPN, Tisseyre P, Melkonian R, et al. Genetic diversity of Mimosas pudica rhizobial symbionts in soils of French Guiana: investigating the origin and diversity of Burkholderia phytochrom and other beta-rhizobia. FEMS Microbiol Ecol. 2012;79:487–503.
10. Cooper JE. Early interaction between legumes and rhizobia: disclosing complexity in a molecular dialogue. J Appl Microbiol. 2007;103:1355–65.
11. Masson-Boivin C, Giraud E, Perret X, Batut J. Establishing nitrogen-fixing symbiosis with legumes: a study of novel legume/nodulating strains. South America. Appl Environ Microbiol. 2009;75:7461–71.
12. Andam CP, Mondo SJ, Parker MA. Monophyly of nodA and nifH genes across Texan and Costa Rican populations of Cupriavidus nodulating symbionts. Appl Environ Microbiol. 2003;69:4686–90.
13. Mishra RPN, Tisseyre P, Melkonian R, et al. Genetic diversity of Mimosas pudica rhizobial symbionts in soils of French Guiana: investigating the origin and diversity of Burkholderia phytochrom and other beta-rhizobia. FEMS Microbiol Ecol. 2012;79:487–503.
14. Cooper JE. Early interaction between legumes and rhizobia: disclosing complexity in a molecular dialogue. J Appl Microbiol. 2007;103:1355–65.
15. Masson-Boivin C, Giraud E, Perret X, Batut J. Establishing nitrogen-fixing symbiosis with legumes: a study of novel legume/nodulating strains. South America. Appl Environ Microbiol. 2009;75:7461–71.
16. Andam CP, Mondo SJ, Parker MA. Monophyly of nodA and nifH genes across Texan and Costa Rican populations of Cupriavidus nodulating symbionts. Appl Environ Microbiol. 2003;69:4686–90.
