Heredity (2014) 113, 306–315
© 2014 Macmillan Publishers Limited All rights reserved 0018-067X/14
www.nature.com/hdy

OPEN

ORIGINAL ARTICLE

Insights into the history of a bacterial group II intron remnant from the genomes of the nitrogen-fixing symbionts *Sinorhizobium meliloti* and *Sinorhizobium medicae*

N Toro, L Martínez-Rodríguez and F Martínez-Abarca

Group II introns are self-splicing catalytic RNAs that act as mobile retroelements. In bacteria, they are thought to be tolerated to some extent because they self-splice and home preferentially to sites outside of functional genes, generally within intergenic regions or in other mobile genetic elements, by mechanisms including the divergence of DNA target specificity to prevent target site saturation. RmInt1 is a mobile group II intron that is widespread in natural populations of *Sinorhizobium meliloti* and was first described in the GR4 strain. Like other bacterial group II introns, RmInt1 tends to evolve toward an inactive form by fragmentation, with loss of the 3′ terminus. We identified genomic evidence of a fragmented intron closely related to RmInt1 buried in the genome of the extant *S. meliloti*/*S. medicae* species. By studying this intron, we obtained evidence for the occurrence of intron insertion before the divergence of ancient rhizobial species. This fragmented group II intron has thus existed for a long time and has provided sequence variation, on which selection can act, contributing to diverse genetic rearrangements, and to generate pan-genome divergence after strain differentiation. The data presented here suggest that fragmented group II introns within intergenic regions closed to functionally important neighboring genes may have been microevolutionary forces driving adaptive evolution of these rhizobial species.

Heredity (2014) **113,** 306–315; doi:10.1038/hdy.2014.32; published online 16 April 2014

INTRODUCTION

Rhizobia are a group of Gram-negative bacteria that form symbiotic associations with leguminous plants. They convert atmospheric nitrogen (N_2), which is unavailable to plants, into ammonia, which is used for the synthesis of amino acids. This fundamental process is essential for life on the Earth. Rhizobia include α-proteobacteria members of the genera *Rhizobium*, *Sinorhizobium* (*Ensifer*), *Bradyrhizobium*, *Azorhizobium*, *Mesorhizobium*, *Devosia*, *Methylobacterium*, *Microvirga*, *Ochrobactrum*, *Phyllobacterium* and *Shinella,* and β-proteobacteria members of the genera *Burkholderia* and *Cupriavidus* (Weir, 2012). *S. meliloti* and *S. medicae* are closely related species forming symbioses with the same host legume species (for example, *Medicago sativa*, *Medicago truncatula*, *Melilotus* and *Trigonella*). The genomes of both *S. meliloti* and *S. medicae* consist of a single circular chromosome (∼3.65 Mb) plus two large symbiotic (sym) plasmids of ∼1.3 (megaplasmids) and ∼1.6 Mb (chromids) in size (Barloy-Hubler *et al.*, 2000; Barnett *et al.*, 2001; Capela *et al.*, 2001; Finan *et al.*, 2001; Galiber *et al.*, 2001; Reeve *et al.*, 2010), and additional smaller plasmids optimizing adaptation to environmental changes.

RmInt1 is a mobile bacterial group II intron that is widespread in natural populations of *S. meliloti* (Muñoz *et al.*, 2001) and was first described in the GR4 strain (Martínez-Abarca *et al.*, 1998). The complete genome sequence of this strain has recently been reported (Martinez-Abarca *et al.*, 2013). Group II introns are self-splicing catalytic RNAs that act as mobile retroelements. They consist of a structured RNA that folds into a conserved three-dimensional structure organized into six double-helical domains, DI to DVI (Michel *et al.*, 2009). Most bacterial group II introns have an open reading frame (ORF) encoding an intron-encoded protein (IEP) in DIV. This IEP consists of a reverse transcriptase followed by a putative RNA-binding domain with RNA splicing or maturase activity (the X domain), and, in some intron lineages, a C-terminal DNA-binding and endonuclease domain (Mohr *et al.*, 1993; San Filippo and Lambowitz, 2002; Toro and Martínez-Abarca, 2013). Group II intron mobility is mediated by a ribonucleoprotein complex consisting of the IEP encoded by the ORF and the spliced intron lariat RNA, which remains associated with the IEP. The ribonucleoprotein complex recognizes the intron target via both the IEP and the intron lariat RNA. The central part of the target containing the intron insertion site is recognized by base pairing between the exon-binding sites (EBSs) in the lariat RNA and the complementary region in the DNA target, the intron-binding site (IBS; Michel and Ferat, 1995). In *S. meliloti*, RmInt1 is mostly found located within the (IS)*Rm2011-2* insertion sequence (Martínez-Abarca *et al.*, 1998; Biondi *et al.*, 2011). It propagates at high frequency in the *S. meliloti* genome, principally via the homing of an RNA intermediate to cognate-homing sites (retrohoming), with a strand bias related to the replication of the chromosome and the plasmids harbored (Martínez-Abarca *et al.*, 2004; Nisa-Martínez *et al.*, 2007). RmInt1-like elements have also

Grupo de Ecología Genética de la Rizosfera, Estación Experimental del Zaidín, Consejo Superior de Investigaciones Científicas (CSIC), Profesor Albareda 1, Granada, Spain
Correspondence: Professor N Toro, Grupo de Ecología Genética de la Rizosfera, Estación Experimental del Zaidín, Consejo Superior de Investigaciones Científicas (CSIC), Profesor Albareda 1, Granada, 18008 Spain.
E-mail: nicolas.toro@eez.csic.es

Received 1 October 2013; revised 12 February 2014; accepted 25 February 2014; published online 16 April 2014

been identified in other *Sinorhizobium* and *Rhizobium* species (Fernandez-Lopez *et al.*, 2005). The intron-homing sites in these species are IS elements of the IS*Rm2011-2* group, as in *S. meliloti*. It has been suggested that these related bacteria have acquired RmInt1-like elements by vertical inheritance from a common ancestor and by independent horizontal transfer events. Interestingly, ectopic transposition has also been observed in natural populations (insertion into target sites other than the usual homing site, occurring at a lower frequency), providing a possible means of transfer to new genomic locations (Muñoz *et al.*, 2001; Fernandez-Lopez *et al.*, 2005).

We report here that a fragmented group II intron from a closely related RmInt1-like element provides a genomic record of ancient intron insertion, probably occurring before the divergence of the *S. meliloti*/*S. medicae* species. This ancient intron record provided us with an opportunity to investigate the long-term evolutionary dynamics of group II introns and the associated microevolutionary processes. Our results suggest that the gradual eradication of group II introns by the host during evolution would not result in some cases in the complete elimination of intron sequences, with some intron fragments remaining and continuing to evolve in the genome making a contribution to the symbiotic capacity and environmental adaptations of these rhizobial species.

MATERIALS AND METHODS

Search for homologous sequences in databases

The nr database (GenBank + EMBL + DDBJ + PDB + RefSeq or GenBank + PDB + Swissprot + PIR + PRF (AA or DNA)) maintained by National Center for Biotechnology Information (entries with absolutely identical sequences have been merged) was used as a target for BLAST searches of sequences homologous to RmInt1 (Y11597.2), the fragmented RmInt1-like element ($FRE_{652,}$ see the results section) and the associated digualinate cyclase (DGC) sequence, using BlastN (nucleotides) and tBlastN (amino acids). Homologous sequences were identified on the basis of *e*-value, size and % pairwise identity, and by careful examination of the corresponding retrieved sequences. Sequences were downloaded, analyzed and processed with Geneious Pro software (Biomatters Ltd, Auckland, New Zealand). FRE_{652} encompasses the locus tag C770_GR4pB086, annotated as group II catalytic intron D1-D4-ncRNA. The entry for DGC associated with GR4 FRE_{652} is annotated as C770_GR4pB085 (Gene-ID: 14254915). A search for the *S. meliloti* strain GR4 FRE_{652}-associated DGC amino-acid sequence yielded 154 hits, 4 of which were identified as additional homologous sequences from strain GR4, as indicated in the results section. Searches of the NCBI and KEGG complete organism databases for the strain GR4 sequence identified another 15 annotated DGCs GGDEF genes.

Comparison of complete genome sequences

Complete genome sequences were aligned with the progressive Mauve algorithm (Darling *et al.*, 2004) and locally collinear blocks were calculated and extracted with Geneious Pro software (Biomatters Ltd).

Phylogenetic analyses

Multiple sequence alignments were generated with MAFFT v7.017, using the FFT-NS-ix2 algorithm and the BLOSUM62 scoring matrix for protein alignment and the automatic algorithm and the 200PAM/k = 2 scoring matrix for nucleotide alignment (Geneious Pro software). For nucleotide phylogeny, a consensus unrooted tree and 100 bootstraps were generated with PhyML or Bayesian methods using the model of nucleotides substitution HKY85 and gamma model with four categories. We applied the GUIDANCE method (Penn *et al.*, 2010) to the original alignment of DGC GGDEF domain sequences (169) with a confidence score of 0.948650, to remove aligned positions considered unreliable. The final alignment used for the phylogenetic analysis retained 90.3% of the columns (cutoff: 0.582) and 196 informative positions. GUIDANCE is freely available from http://guidance.tau.ac.il. Unrooted trees and 100 bootstraps were generated with PhyML (Guindon and Gascuel, 2003, Guindon *et al.*, 2010), with the WAG amino-acid substitution model (similar tree topologies was inferred with the LG model) and a discrete gamma model with four categories. In some analyses, we carried out Bayesian analysis with the parallel version of MrBayes 3.1 (Huelsenbeck and Ronquist, 2001). Two independent runs of four chains were completed for 1 100 000 Metropolis-coupled Markov chain Monte Carlo generations, using the default priors for model parameters, the WAG (amino acids) and HKY85 (nucleotides) model as the rate matrix (fixed) and the gamma model for between-site rate variation. Trees were sampled every 200 generations, and 110 000 samples were discarded as the 'burn-in', to produce a 50% majority-rule consensus tree. The phylogenetic tree generated with PhyML for rhizobial species was based on the concatenated alignment of DnaK and RpoB (Tian *et al.*, 2012) sequences retrieved from the GenBank database.

RESULTS

Presence of full-length and fragmented RmInt1-like elements in rhizobia

BlastN search of nr database carried out here revealed that the closest (⩾85% identity in pairwise comparisons) full-length known relatives of the *S. meliloti* RmInt1 intron were present in the closest relatives *S. medicae*, *E. adhaerens* and *S. terangae* (85–99% identity), whereas the most closely related fragmented introns (90–99% identity) were present in the same bacterial species and in other *Sinorhizobium* (*S. fredii*) and *Rhizobium* (*R. etli*, *R. tropici* and *R. leguminosarum* bv. *phaseoli*) species.

In addition to the chromosome and the expected symbiotic plasmids pRmeGR4c (pSymA) and pRmeGR4d (pSymB), *S. meliloti* strain GR4 harbors two accessory plasmids, pRmeGR4a and pRmeGR4b (Martinez-Abarca *et al.*, 2013). GR4 has 10 copies of RmInt1 that are 99.9% identical (Table 1). In *S. meliloti*, the currently existing RmInt1 copies display considerable sequence conservation (>99% identity). The numbers of these copies differ between the *S. meliloti* strains for which complete genome sequences are available (Table 1), such as 1021 (Galibert *et al.*, 2001), Rm41 (Weidner *et al.*, 2013), BL225C (Galardini *et al.*, 2011) and SM11 (Schneiker-Bekel *et al.*, 2011), and strain AK83 (Galardini *et al.*, 2011) harbors a single copy in plasmid pSINME01 (pSymA) including an internal deletion from nucleotides 1520–1844 (98% identity at the 5′ end of the fragment). In addition to its full-length copies of RmInt1, strain GR4 harbors a cryptic plasmid, pRmeGR4b, bearing a fragmented RmInt1-like element (Table 1) of 652 nt (hereafter referred to as FRE_{652}) that is 88.9% identical to RmInt1 and 89.7% identical to its closest relative, *S. medicae* intron Sr.md.I1. BlastN search and sequence alignments revealed the presence of copies of a similar fragmented intron to FRE_{652} in other *S. meliloti* strains (Figure 1). Strain AK83 harbors two copies on the so-called 'chromosome 3' (pSymA), whereas strain SM11 has one copy on pSmeSM11c (pSymA) and strain C017 has one copy on the sequenced cryptic plasmid pHRC017 (Crook *et al.*, 2012). In strain BL225C, a 145-nt copy of this intron fragment, trimmed at its 5′ end, was identified on pSINMEB01 (pSymA). FRE_{652} was found to be absent from the genomes of strains 1021 and Rm41, but one copy of 94% identical to FRE_{652} was present in the closely related species *S. medicae*, WSM419, for which complete genome sequence is available located on plasmid pSMED02 (orthologous to pSymA).

FRE_{652} provides a genomic record of the history of an intron closely related to RmInt1

An analysis of the sequences of FRE_{652} copies in *S. meliloti* and *S. medicae* strains (Figure 1) revealed that the sequence of this intron fragment spanned ribozyme domains I to III and included part of domain IV, which was truncated at a position corresponding to

Table 1 Full-length and fragmented RmInt1 phylogenetically related elements in Rhizobia.

Name[a]	*Species*	*Strain*	*Host replicon*[b]	*Size*	*Accesion (intron boundaries)*	*Reference*
S. me GR4 RmInt1	*S. meliloti*	GR4	Chr(4); pRmeGR4b (1); pRmeGR4c(5)	1884	Y11597 (1–1884)	Martínez-Abarca *et al.* (1998, 2013)
S. me 1021 RmInt1	*S. meliloti*	1021	pSymA (1); pSymB(2)	1884	AL591985 (675027–676910)	Galibert *et al.* (2001)
S. me SM11 RmInt1	*S. meliloti*	SM11	Chr(2); pSmeSM11c (3); pSmeSM11d(1)	1884	CP001830 (3677924–3679807)	Schneiker-Bekel *et al.* (2011)
S. me BL225C RmInt1	*S. meliloti*	BL225C	Chr(2); pSINMEB01 (1); pSINMEB02(2)	1884	CP002740 (667127–669010)	Galardini *et al.* (2011)
S. me Rm41 RmInt1	*S. meliloti*	Rm41	pSymB(1)	1884	NC_018701 (1129686–1127803)	Weidner *et al.* (2013)
E. ad 5D19 RmInt1	*Ensifer adhaerens*	5D19	ND	1884	AY248839 (1-1884)	Fernandez-Lopez *et al.* (2005)
S. fred NGR234 RmInt1 F	*S. fredii*	NGR234	pNGR234b (1)	487[c]	NC_012586 (796157–796643)	Schmeisser *et al.* (2009)
S. me AK83 RmInt1 F	*S. meliloti*	AK83	pSINME01(1)	1519[d]	CP002784 (8422-9931)	Galardini *et al.* (2011)
S. te ORS22 RmInt1-like	*S. teranga*	ORS22	ND	1884	AY608908	Fernandez-Lopez *et al.* (2005)
S. md WSM419 Sr.md.I1	*S. medicae*	WSM419	Chr (1); pSMED01(1); pSMED02(2)	1885	CP000738 (1305475–1307359)	Reeve *et al.* (2010)
S. me GR4 RmInt2	*S. meliloti*	GR4	pRmeGR4c(4); pRmeGR4d(3)	1887	NC_019849 (1678444–1680330)	Martinez-Abarca *et al.* (2013)
R. et 8C-3 F	*Rhizobium etli*	8C-3	ND	1262	DQ058416 (91480–92624)	Flores *et al.* (2005)
R. et CIAT652 F	*R. etli*	CIAT 652	pB	1219	CP001076 (126968–125954)	Gonzalez *et al.* (2010)
R. et CFN42 F	*R. etli*	CFN 42	p42d	1208	REU80928 (105330–104316)	Girard *et al.* (1991)
R. et CFN42 (2) F	*R. etli*	CFN 42	p42a	856	CP000134 (63346–62493)	Gonzalez *et al.* (2006)
R. et CE3 F	*R. etli*	MB043	pa	856	AF176227 (10069–9216)	Bittinger *et al.* (2000)
E. ad LMG20582	*E. adhaerens*	R-6387	ND	764	AY608901 (162–925)	Fernandez-Lopez *et al.* (2005)
S. me GR4 pRmeGR4b FRE652	*S. meliloti*	GR4	pRmeGR4b	652	NC_019847 (73323–73974)	Martinez-Abarca *et al.* (2013)
S. me AK83 FRE652 (2)	*S. meliloti*	AK83	Chr3 (pSymA-like)	653	NC_015591 (330796–331448)	Galardini *et al.* (2011)
S. me C017 pHRC017 FRE652	*S. meliloti*	C017	pHRC017	651	JQ665880 (63063–63713)	Crook *et al.* (2012)
S. me SM11 FRE652	*S. meliloti*	SM11	pSmeSM11c (pSymA-like)	651	CP001831 (95351–96001)	Schneiker-Bekel *et al.* (2011)
S. me AK83 FRE652 (1)	*S. meliloti*	AK83	Chr3 (pSymA-like)	644	NC_015591 (1249601–1248958)	Galardini *et al.* (2011)
S. me BL225C FRE652	*S. meliloti*	BL225C	pSINMEB01 (pSymA-like)	145	CP002741 (821391–821535)	Galardini *et al.* (2011)
S. md WSM419 FRE652	*S. medicae*	WSM419	pSMED02 (pSymA-like)	639	CP000740 (1062090–1061452)	Reeve *et al.* (2010)
S. md RMO02 FRE652	*S. medicae*	RMO02	ND	650	AY608903 (148–797)	Fernandez-Lopez *et al.* (2005)

Abbreviations: F, fragmented; ND, not determined.
[a] Name corresponding to Figure 2.
[b] Replicon and no of copies (in parenthesis).
[c] Group II intron interrupted at nt 487 by a DNA fragment (3803 nt) containing several transposases.
[d] Group II intron with an internal deletion from nucleotides 1520–1844.

position 653 of RmInt1. Likewise, the 5′ end of the ORF had been subject to an earlier frameshift mutation, deleting the T residue of the ATG start codon. Interestingly, all copies of FRE_{652} have lost the first G residue of the intron, but the exon-binding sequences (EBS1, EBS2 and EBS3) are identical to those of RmInt1, suggesting that the original full-length intron had the same potential targets. The phylogenetic tree (Figure 2) generated by maximum-likelihood methods and Bayesian analyses from alignments (Supplementary Figure 1) of 24 sequences covering 664 informative nucleotide positions in FRE_{652} copies from various hosts, RmInt1 and other fragmented RmInt1-like elements identified in *Sinorhizobium/Ensifer* and *Rhizobium* species revealed that all FRE_{652} sequences branched from a common node with strong bootstrap support (100% bootstrap support and a posterior probability of 99.98%), suggesting that they arose from a single ancestral intron. Furthermore, the group of FRE_{652} elements had a statistically supported node in common with the RmInt1 group (71% bootstrap support and a posterior probability of 95.37%). The 487 nt intron fragment (also truncated at its 3′ end) of *S. fredii* strain NGR234 also clustered within the RmInt1 group, and the fragmented intron copies of *R. etli*/*Ensifer* species formed a differentiated group with strong statistical support (93% bootstrap support and a posterior probability of 98.55%). The mutation of existing RmInt1 elements is therefore unlikely to account for FRE_{652}, and the most plausible explanation seems to be that FRE_{652} represents a genomic record of the history of an intron closely related to RmInt1.

FRE_{652} represents an ancient intron insertion event

We investigated whether FRE_{652} resulted from an ancient intron insertion event, by analyzing the sequences flanking FRE_{652} in *S. meliloti* and *S. medicae*, by using Mauve to align whole-genome sequences carrying copies of the ancient fragmented intron. We found that the sequences flanking the copies of FRE_{652} on pSymA in *S. meliloti* strains AK83 and SM11, the accessory plasmids pRmeGR4b

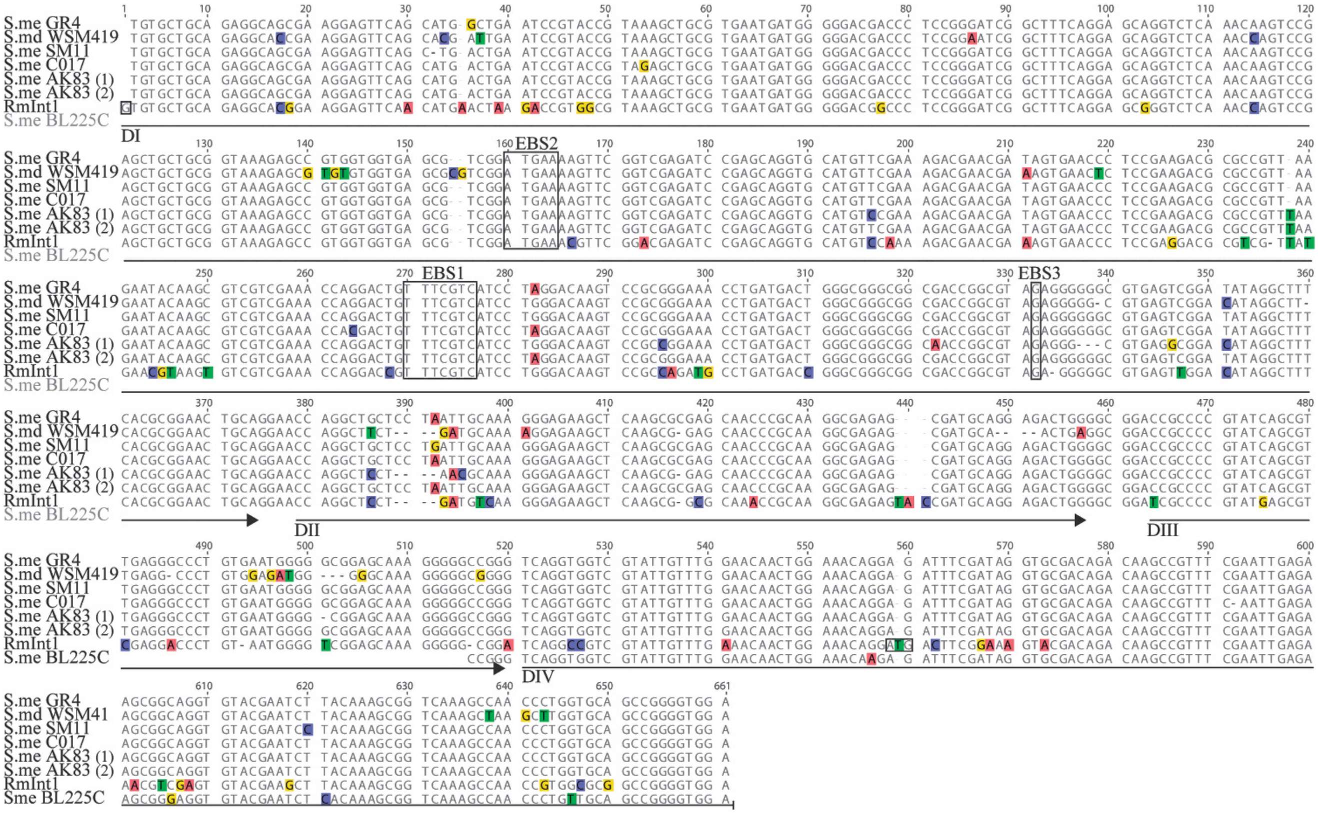

Figure 1 Alignment of FRE$_{652}$ and RmInt1 sequences. The ribozyme domains DI to DIV are underlined. Exon-binding sequences (EBS1, EBS2 and EBS3) are boxed, and relevant nucleotides corresponding to the first G residue of the RmInt1 intron and the ATG start codon of the IEP absent from FRE$_{652}$ sequences are also boxed. Residues differing between FRE$_{652}$ and the consensus are highlighted. Genomic positions, as shown from left to right, are: S. me GR4 pRmeGR4b (bases 73 323–73 974), S. md WSM419 pSMED02 (bases 1 062 090–1 061 452), S. me SM11 pSymA (bases 95 351–96 001), S. me C017 pHRC017 (bases 63 063–63 713), S. me AK83 pSymA copy 1 (bases 1 249 601–1 248 958), S. me AK83 pSymA copy 2 (bases 330 796–331 448) and S. me BL225C (bases 821 534–821 391).

and pHRC017 from strains GR4 and C017, respectively, and pSMED02 from *S. medicae* strain WSM419 displayed synteny over a region of about 7 kb (Figure 3a). Synteny extended over a longer distance in pHRC017 and pRmeGR4b (not shown), suggesting that these plasmids have a common origin. In all cases, a predicted helix–turn–helix transcriptional regulator (referred to hereafter as TR) transcribed in the opposite orientation to the putative intron insertion was found downstream from the FRE$_{652}$ element. However, this putative regulator sequence was interrupted by a copy of the IS*Bm1* insertion sequence in *S. medicae*. Upstream from the putative intron insertion, in the opposite orientation, there is a diguanylate cyclase/phosphodiesterase (DGC/PDEA) gene, followed by genes encoding a pectate lyase and a carbonate dehydratase. In strain BL225C, only the putative TR remains, the upstream flanking sequences of the syntenic block being absent (not shown). The complete block was found to be absent from the genomes of strains 1021 and Rm41.

As the 3′ end of the original intron from which FRE$_{652}$ was derived is currently missing, the boundary and presence of the 3′ exon remain uncertain. We therefore investigated the sequence of the putative 5′ exon. The 5′ exon of RmInt1 contains the intron-binding sequences IBS1 (7 nt) and IBS2 (5 nt), which are separated by a single nucleotide (C residue), and extends to the 5′ distal exon region via seven additional nucleotides, including the critical T residue in position −15, a DNA-target region probably recognized by the IEP (Jiménez-Zurdo *et al.*, 2003). Presently, only limited pairing was observed at the putative IBS1 (4 of 7 residues) and IBS2 (3 of 5 residues) for FRE$_{652}$; likewise, instead of a T in position −15 a G residue was found, and possibly the A in position −1 of the target site was also lost, like the first nucleotide of the intron (Figure 3b). Overall, these results suggest that intron insertion at this genomic location could result from infrequent retrotransposition or retrohoming event with later alterations in the EBS/IBS pairing.

The DGC sequence associated with FRE$_{652}$ derived from a common ancestor in the Rhizobiaceae/Phyllobacteriaceae proteobacteria families

We tested our proposed evolutionary hypothesis further, by analyzing the phylogenetic information available for the DGC ORF adjacent to FRE$_{652}$. The DGC and phophodiesterase (PDE) enzymes control intracellular c-di-GMP concentration by the synthesis and degradation, respectively, of this molecule. This ubiquitous second messenger is known to have a key role in several cellular functions, including exopolysaccharide production, attachment and motility, and in adhesion and biofilm formation in bacteria (for a review see Jenal and Malone, 2006; Hengge, 2009; Schirmer and Jenal, 2009; Römling *et al.*, 2013). These functions are relevant to rhizosphere colonization and host plant nodulation by *S. meliloti* and *S. medicae* (Gage, 2004; Fujishige *et al.*, 2006). The active site of DGCs contains a conserved GGDEF domain, characterized by the GG(D/E)EF motif (A site), whereas PDE activity is associated with C-terminal EAL (PDEA) or HD-GYP domains. The DGCs linked to FRE$_{652}$ are composite

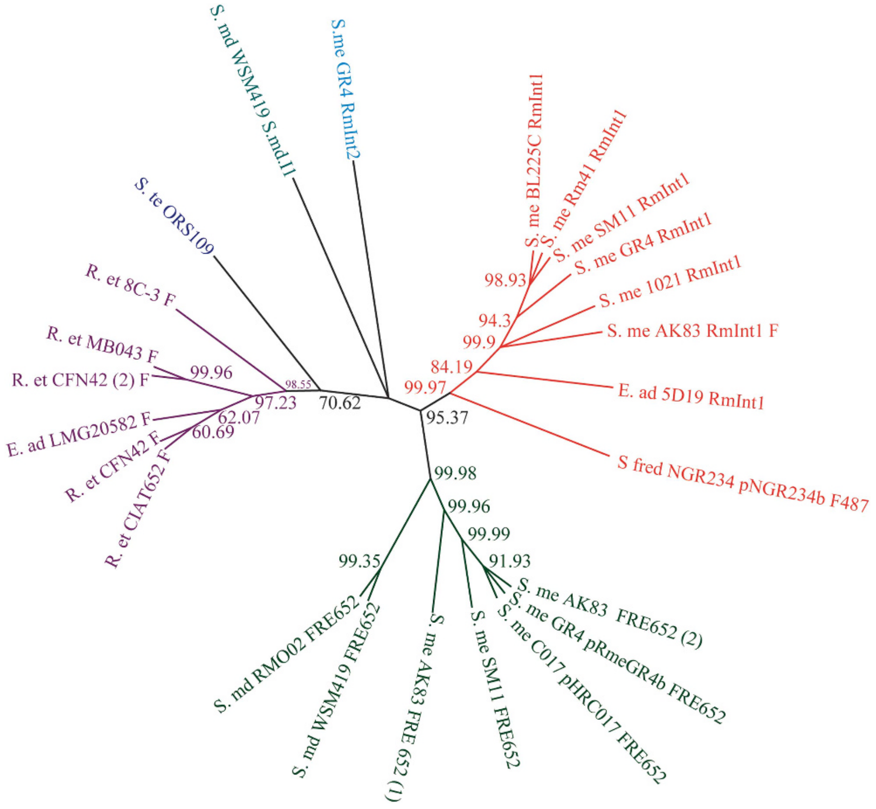

Figure 2 Phylogeny of FRE_{652}, RmInt1 and other full-length and fragmented RmInt1-like elements. A consensus unrooted tree is shown and the corresponding sequences are labeled. The clusters are highlighted in color and posterior probabilities for Bayesian analyses are indicated at the nodes. The multiple sequence alignment used for the analysis is shown in Supplementary Figure 1. F, fragmented intron. Accession numbers for sequences are provided in Table 1. S. me, *S. meliloti*; S. md, *S. medicae*; E. ad, *E. adhaerens*; R. et, *R. etli*; S. te, *S. terangae*. S. me GR4 RmInt2 is a distant relative of RmInt1.

DGC-PDEA structural proteins with the A site motif RxAGDEF but without the I site (characterized by an RXXD motif) subject to allosteric product inhibition (Supplementary Figure 2).

TBlastN searches using the amino-acid sequence of the DGC (543 aa) encoded by a gene adjacent to FRE_{652} in strain GR4 as a query identified a large number of homologous sequences in databases. The DGC sequences identified in the searches (154 sequences) were mostly from bacterial species of the order Rhizobiales (Figure 4) and all were DGC/PDEA domain proteins. These hits included four additional homologous sequences from strain GR4 displaying distinct similarities to the query sequence. Two of these sequences were located on pSymA, about ~682 kb apart, and encoded proteins displaying 81.4% (551 aa) and 44.8% (1071 aa) identity. One was located on pSymB and was 62.5% (564 aa) identical and the remaining sequence (742 aa) was located on pRmeGR4b and was 46.6% identical in pairwise comparisons. This last sequence was separated from that associated with FRE_{652} by ~28 kb. The larger ORFs found on pSymA- and pRmeGR4b-encoded proteins that also contained a PAS domain sensory box at the N-terminus. Strain GR4 actually harbors 20 proteins annotated as DGCs and three annotated as EALs, highlighting the potential importance of c-di-GMP in the lifestyle of these symbiotic bacteria.

Phylogenetic analyses were performed on an alignment of 154 amino-acid sequences (identity $\geqslant$40%) and the other 15 DGC sequences harbored by strain GR4. Similar tree topologies were obtained from independent alignments (not shown) for the GGDEF and EAL domains. Some of the DGCs in strain GR4 have no EAL domain. We therefore present here only the phylogenetic analysis based on the GGDEF domain alignment (196 informative positions, Supplementary Figure 2) of 169 sequences after the removal of unreliable positions from the alignment by GUIDANCE. The estimated phylogenetic tree (Figure 4) indicates that the DGCs associated with FRE_{652} in *S. meliloti* and *S. medicae* cluster together in a well-supported (88% bootstrap value) common node, consistent with a monophyletic group and the occurrence of an intron insertion event in this genomic region before speciation. The homologous DGC found on the strain GR4 pSymA (551 aa), in particular, branched from a common node (96% bootstrap value) shared with the cluster described above (referred to hereafter as pSymA-DGC1). The other sequenced strains of *S. meliloti* and *S. medicae* have no counterpart of former DGC gene, suggesting that this gene and the DGC associated to FRE_{652} in GR4 strain may be paralogs.

Similarly, the DGCs/PDEAs found on *S. meliloti* pSymB and the orthologous plasmid of *S. medicae,* pSMED01, and on the symbiotic plasmid of *S. fredii* (pNGR234b/pSfHH103e) are monophyletic (97% bootstrap support). This cluster (referred to hereafter as pSymB-DGC1) and pSymA-DGC1 have an internal node, with a bootstrap value of 99%, in common with other DGCs from other members of the Rhizobiaceae (*R. leguminosarum* bv. trifolii and viciae, *R. etli, R. tropici*, and *A. radiobacter* species) and Phyllobacteriaceae (*M. loti* species) families. The tree topology of this node resembles that for the species phylogeny (Figure 5): *Mesorhizobium*, *Rhizobium* and *Sinorhizobium* cluster together with a high level of bootstrap support (100%); these results suggest that these DGCs of these species

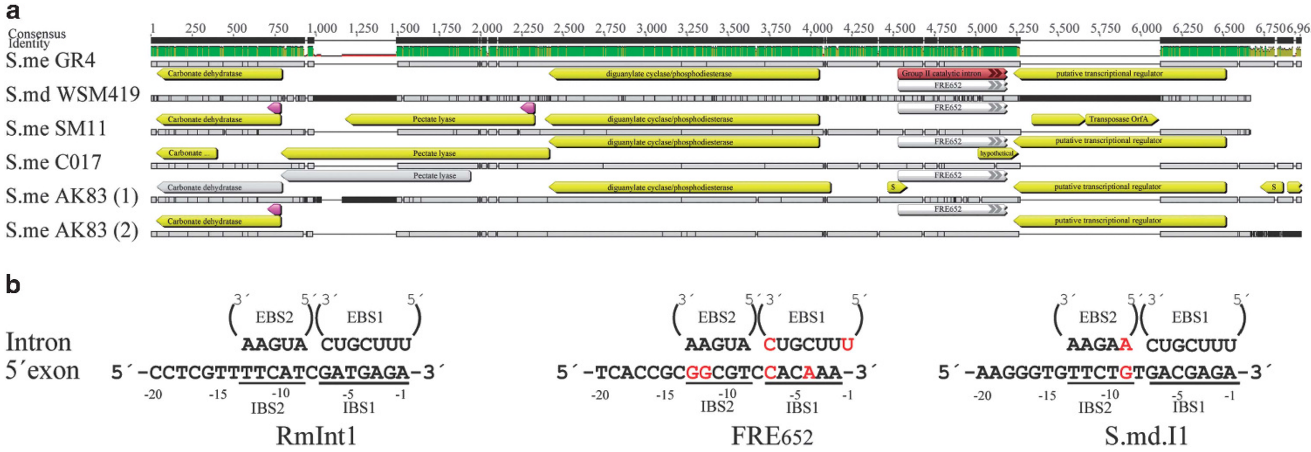

Figure 3 Analysis of the neighborhood of FRE_{652} in the *S. meliloti* and *S. medicae* genomes. (**a**) Locally collinear block corresponding to a conserved segment containing FRE_{652} sequences identified by Mauve. Complete sequences of megaplasmid pSymA from *S. meliloti* strains AK83 and SM11, the cryptic plasmids pHRC017 and pRmeGR4b from *S. meliloti* strains CO17 and GR4, respectively, and that of megaplasmid pSMED02 from *S. medicae* strain WSM419 were aligned with the progressive Mauve algorithm, and locally collinear blocks were extracted with Geneious Pro software (Biomatters Ltd). The relevant ORFs (coding sequence (CDS)) are depicted (yellow). The authors of the original studies in the database annotated carbonate dehydratase as carbonic anhydrase (GR4 and pHRC017); diguanylate cyclase/phosphodiesterase as GGDEF domain/EAL domain protein (GR4), conserved hypothetical protein (SM11) and PAS/PAC sensor-containing diguanylate cyclase (GGDEF)/phosphodiesterase (EAL) (pHRC017); the putative transcriptional regulator as hypothetical protein (SM11), XRE family helix-turn-helix transcriptional regulator (pHRC017) and helix-turn-helix domain protein (AK83). Pink open arrows above some of the CDS indicate putative signal peptides and an S within some of the CDS correspond to annotated short hypothetical proteins. FRE_{652} annotations and annotations for a group II intron derivative have been introduced manually here. The consensus identity is shown above the block, and the green color reflects the regions of higher identity. Genomic positions, as shown from left to right, are: S. me GR4 pRmeGR4b (bases 69 345–74 866), S. md WSM419 pSMED02 (bases 1 066 557–1 060 009), S. me SM11 pSymA (bases 91 375–96 621), S. me C017 pHRC017 (bases 59 086–64 605), S. me AK83 pSymA copy 1 (bases 1 253 952 to 1 248 066) and S. me AK83 pSymA copy 2 (bases 326 826–332 352). (**b**) DNA target sites recognized by group II introns RmInt1 and Sr.md.I1, and the putative sequences recognized by the intron from which FRE_{652} was derived. 5′ Exon positions are indicated relative to the intron insertion site, and the EBS and IBS sequences are shown. The unpaired residues are highlighted in red. Note that for FRE_{652}, the −1 residue of the 5′ exon is absent; as it was probably deleted as the first G residue of the intron (see Figure 1).

are probably derived from a single ancestral gene that subsequently underwent rearrangements, gene duplications and losses after species and strain differentiation.

The additional DGCs in GR4, which are distributed in various replicons, including the chromosome, did not cluster with these nodes in the tree and their relationships with the other sequences included in the alignment remain uncertain, but some of them may have been acquired by horizontal gene transfer (see the pSymA-DGC2 clade in Figure 4).

DISCUSSION

We identified a genomic record of a closely RmInt1-related intron buried in the genome of extant *S. meliloti*/*S. medicae* species, and obtained evidence to suggest that the insertion of this intron probably occurred before the divergence of these rhizobial species. The intron subsequently underwent deletion events and accumulated diverse mutations. In *S. meliloti* strains, this genomic region carrying genes that might be important to rhizosphere colonization and host plant nodulation undertook further deletions and diverse genetic rearrangements, including losses and duplications, and in some strains, a block of ∼7 kb containing the intron fragment and neighboring genes was hijacked by other smaller accessory replicons. This fragmented form of the intron has been maintained over extensive evolutionary time, which suggests it may confer a selective advantage on the host.

Rhizobial genes involved in symbiosis are often clustered on large plasmids (pSym), a feature differentiating these nitrogen-fixing plant endosymbionts from other nonsymbiotic saprophytes. Rhizobial genomes appear to be highly dynamic, probably due to the presence of repeated DNA sequences, IS elements and transposons, together with multiple replicons (MacLean *et al.*, 2007). *S. meliloti* has a large, typically multipartite genome with a chromosome, a chromid (pSymB), a large replicon containing not only plasmid-type replication systems but also genes essential for growth and survival (Harrison *et al.*, 2010), a megaplasmid (pSymA), and several additional smaller plasmids. The smaller plasmids and the megaplasmids are considered to be essentially strain-specific and of recent origin, whereas the chromid and the chromosome are thought to be less variable and more genus-specific and of ancient origin. It has been suggested that the pSymA megaplasmid is involved principally in structural fluidity and the emergence of new functions (Galardini *et al.*, 2013). Consistent with this assumption, we found that the insertion of the ancient RmInt1-like intron into the ancestor of pSymA has, to some extent, contributed to generate pan-genome divergence after strain differentiation.

Like other bacterial group II introns (Dai and Zimmerly, 2002), RmInt1 is tending to evolve toward an inactive form by fragmentation, with the loss of the 3′ terminus, including the IEP (Fernandez-Lopez *et al.*, 2005). The significance of fragmented introns within a particular genome remains unclear. They generally have no counterpart from the same intron in the same genome (Leclercq and Cordaux, 2012) and are thus considered to have been inactivated before proliferation, or they are overlooked as dying copies of a particular intron that is currently active. Only 25% of the bacterial genomes sequenced to date (Lambowitz and Zimmerly, 2011) harbor recognizable group II introns, arguing against a role as a broad and important force promoting evolutionary change, but caution is required in the interpretation of these observations. The overall group

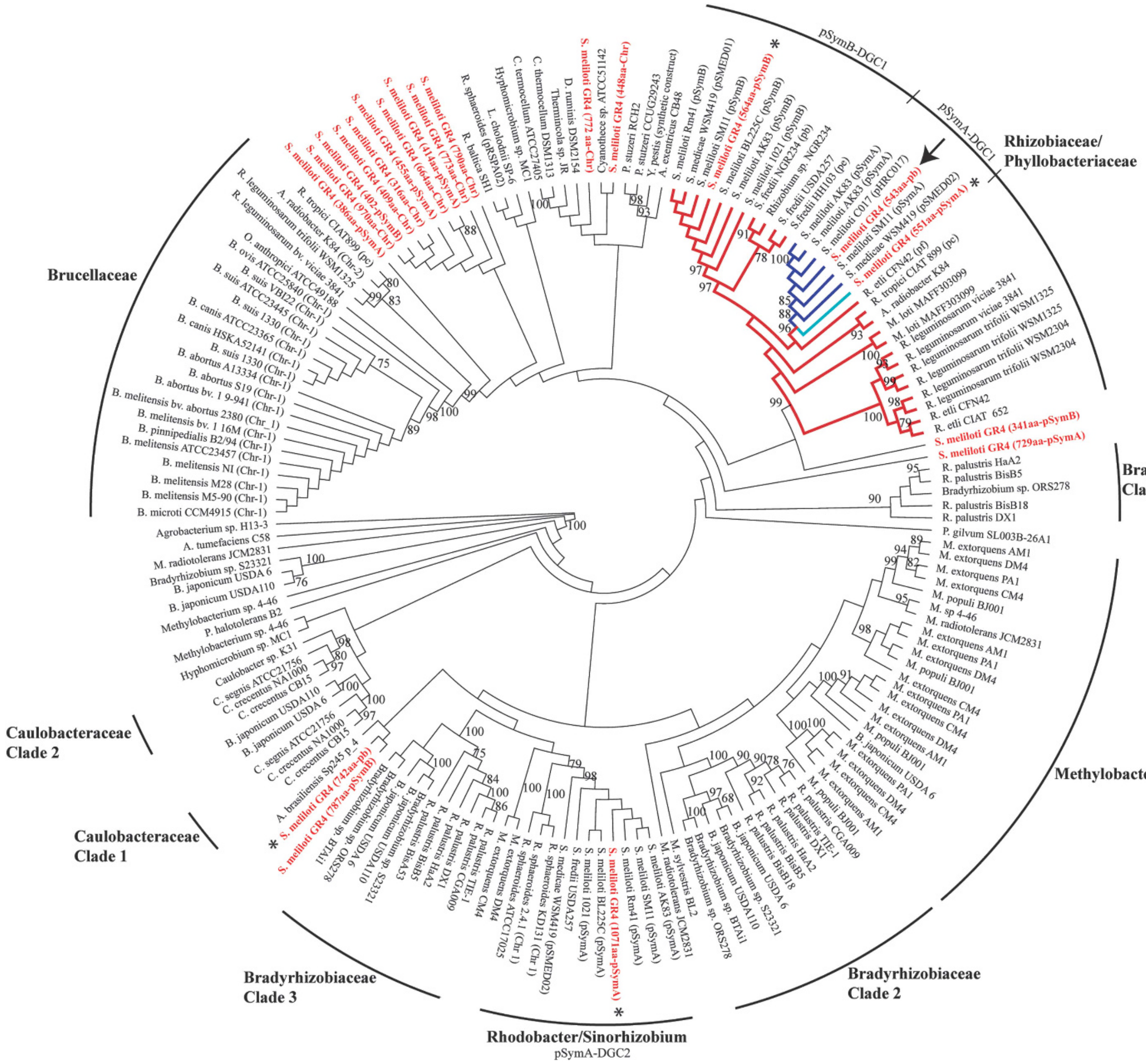

Figure 4 Phylogeny of DGCs homologous to that associated with FRE_{652}. The consensus unrooted tree estimated by ML (maximum likelihood) methods is presented as a radial cladogram, and bootstrap values of >75% are shown at the nodes. The phylogenetic tree is based on the alignment of the GGDEFF domains (196 informative positions) of 169 sequences (see Supplementary Figure 2), including all the DGCs (20) annotated in the genome of *S. meliloti* strain GR4, which are highlighted in red, with an indication of their size and genomic location. The bacterial species harboring the DGCs used in the alignment are indicated at each external branch. Major clades and relevant clusters are also indicated and the corresponding bacterial families are shown. The branches corresponding to the Rhizobiaceae/Phyllobacteriaceae clade containing the most closely related orthologs and paralogs of FRE_{652}-associated DGC are highlighted in red. The DGCs associated to FRE_{652} are indicated in dark blue and the possible paralog in strain GR4 pSymA in light blue. The additional four DGC homologs identified in strain GR4 in the Blast search within the 154 hits are indicated by an asterisk, whereas the DGC associated to FRE_{652} is indicated by an arrow.

II intron primary sequence is not well conserved, other than for RNA domain (DV), so the 5′ end of intron sequences lacking the encoded ORF is unlikely to have been detected in sequenced bacterial genomes.

Nuclear pre-messenger RNA introns (Michel and Ferat, 1995) and non-long terminal repeat retrotransposons are both thought to be descended from mobile group II introns (Eickbush, 1994), but the role of group II introns in generating genetic novelty and bacterial evolution remains unclear. It has recently been suggested that, as for transposable elements, the dispersal and dynamics of group II intron spread within a bacterial genome would follow a selection-driven extinction model, predicting the removal of highly colonized genomes from the population by purifying selection (Leclercq and Cordaux, 2012). It has been reported (Muñoz *et al.*, 2001) that 10% of *S. meliloti* strains and isolates seem to lack RmInt1 but do not appear to have any active mechanism for controlling intron invasion or proliferation (Martínez-Abarca *et al.*, 2004; Nisa-Martinez *et al.*, 2007). It is generally accepted that the ‘selfish’ features of mobile elements underlie their acquisition and maintenance in bacterial genomes, but these elements may also be beneficial to their hosts. In bacteria, group II introns are thought to be tolerated to some extent because they self-splice and preferential home to sites outside of function genes, generally within intergenic regions or in other mobile genetic elements (Simon *et al.*, 2008), through mechanisms including

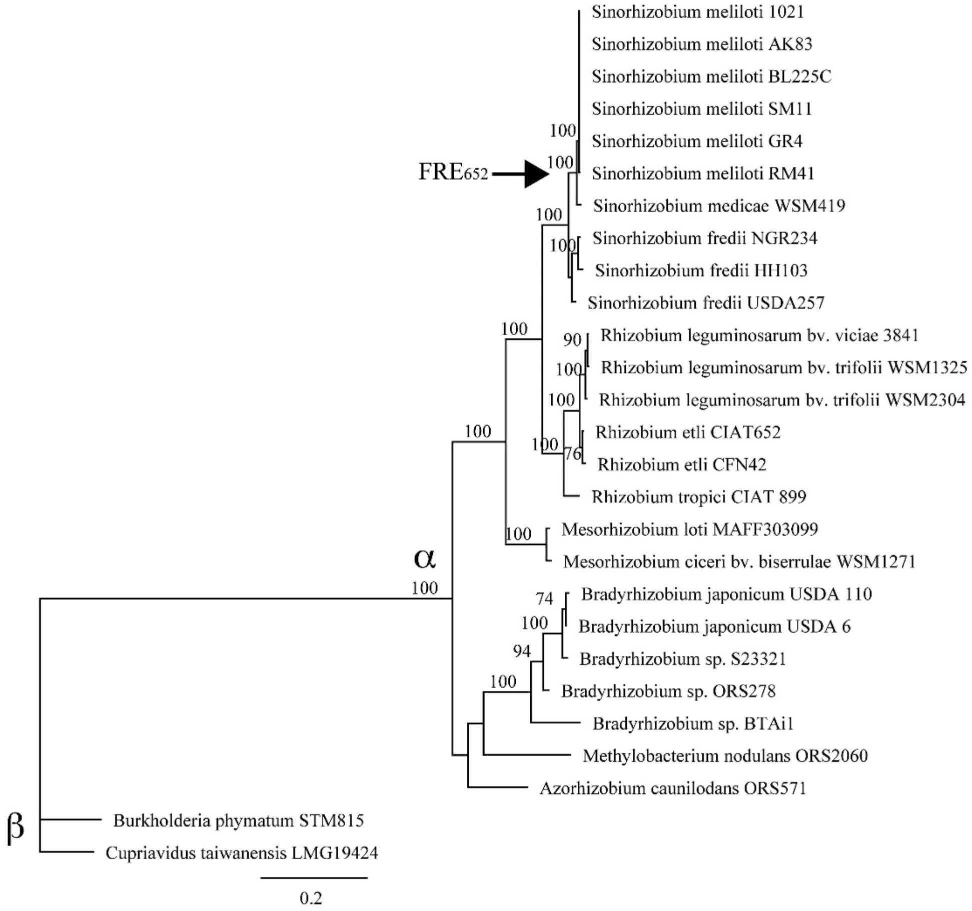
Sinorhizobium meliloti 1021
Sinorhizobium meliloti AK83
Sinorhizobium meliloti BL225C
Sinorhizobium meliloti SM11
Sinorhizobium meliloti GR4
Sinorhizobium meliloti RM41
Sinorhizobium medicae WSM419
Sinorhizobium fredii NGR234
Sinorhizobium fredii HH103
Sinorhizobium fredii USDA257
Rhizobium leguminosarum bv. viciae 3841
Rhizobium leguminosarum bv. trifolii WSM1325
Rhizobium leguminosarum bv. trifolii WSM2304
Rhizobium etli CIAT652
Rhizobium etli CFN42
Rhizobium tropici CIAT 899
Mesorhizobium loti MAFF303099
Mesorhizobium ciceri bv. biserrulae WSM1271
Bradyrhizobium japonicum USDA 110
Bradyrhizobium japonicum USDA 6
Bradyrhizobium sp. S23321
Bradyrhizobium sp. ORS278
Bradyrhizobium sp. BTAi1
Methylobacterium nodulans ORS2060
Azorhizobium caunilodans ORS571
Burkholderia phymatum STM815
Cupriavidus taiwanensis LMG19424
FRE652
100
100
100
100
100
90
100
100
100
76
100
100
100
74
100
94
100
α
β
0.2

Figure 5 Species tree for rhizobia. Consensus unrooted phylogenetic tree estimated by ML methods from the concatenated sequences of *dnaK* and *rpoB* housekeeping genes for the rhizobial species indicated aligned using the Geneious Pro Software (Biomatters Ltd). Bootstrap values of >75% are indicated at each node. α, the branch corresponding to α-proteobacteria; β, the branch corresponding to β-proteobacteria. The arrow indicates the predicted intron insertion into the ancestor of *S. meliloti*/*S. medicae*.

the divergence of DNA target specificity to prevent target site saturation (Mohr *et al.*, 2010). Other studies have suggested that group II introns are beneficial to their hosts because they control other potentially harmful mobile genetic elements (Chillón *et al.*, 2010), and contribute to the generation of diversity and remodel genomes in time of stress (Coros *et al.*, 2009). These features may decrease negative effects on the host organism, resulting in the maintenance of these retroelements for longer periods in bacterial populations. Our results suggest that the gradual eradication of group II introns by the host during evolution would not result in the complete elimination of intron sequences, with some intron fragments remaining and continuing to evolve in the genome.

The divergence of the rapidly growing rhizobial genera (*Sinorhizobium–Rhizobium–Mesorhizobium*) has been dated to 203–324 million years ago (MYA), before the emergence of legumes estimated to 60 MYA (Turner and Young, 2000; Sprent and James, 2007). *S. meliloti* and *S. medicae* are taxonomically and symbiotically related species, but DNA–DNA hybridization was found to exhibit only 42–60% DNA homology (Rome *et al.*, 1996), and have important differences in gene content (Sugawara *et al.*, 2013). There are no reports dating the divergence of *S. meliloti* and *S. medicae* species, but phylogenies based on a sample of housekeeping genes suggest that *S. meliloti* and *S. medicae* are not sister species (Martens *et al.*, 2007), and suggest that *S. medicae* might be the first emerging taxon within a clade including *S. medicae*, *S. meliloti* and *S. arboris*, suggesting a rather ancient speciation event leading to the first two species (Bailly *et al.*, 2007). The identification in this study of a genomic record of a group II intron in *S. meliloti*/*S. medicae* genomes reveals that fragmented introns from ancient insertions within intergenic regions can persist for long periods, probably because their removal increases the likelihood of harmful effects on adjacent genes, as suggested for other fragmented transposable elements in eukaryotes (Werren, 2011). A search for conserved fragmented introns in *S. meliloti* species (N. Toro, unpublished) revealed that FRE_{652} is not a unique case, and similar ribozyme 5′ end fragments of other group II introns are conserved buried in the genomes of this bacterial species closed to conserved actively transcribed regions. These group II intron remnants could just represent a stochastic persistence of some intron fragments under a very slow process of degradation, but the possibility remains that they could provide sequence variation on which selection can act remaining and continuing to evolve in the genome in some bacterial lineages. We hypothesize that, as for other fragmented transposable elements in eukaryotes (Werren 2011), these fragmented intron sequences in bacteria may have evolved into functional *cis*-regulatory elements making a direct contribution to bacterial speciation. The data presented here raise novel issues concerning the significance of group II introns in bacterial evolution, which need to be further investigated.

DATA ARCHIVING

There were no data to deposit.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

This work was supported by research grants CSD 2009-0006 of the Consolider-Ingenio programme including ERDF (European Regional Development Funds), and BIO2011-24401 from the Spanish *Ministerio de Ciencia e Innovación*, currently the *Ministerio de Economía y Competitividad*.

Bailly X, Olivieri I, Brunel B, Cleyet-Marel J-C, Bena G (2007). Horizontal gene transfer and homologous recombination drive the evolution of the nitrogen-fixing symbionts of *Medicago* species. *J Bacteriol* **189**: 5223–5236.

Barloy-Hubler F, Capela D, Batut J, Galibert F (2000). High-resolution physical map of the pSymb megaplasmid and comparison of the three replicons of *Sinorhizobium meliloti* strain 1021. *Curr Microbiol* **41**: 109–113.

Barnett MJ, Fisher RF, Jones T, Komp C, Abola AP, Barloy-Hubler F *et al.* (2001). Nucleotide sequence and predicted functions of the entire *Sinorhizobium meliloti* pSymA megaplasmid. *Proc Natl Acad Sci USA* **98**: 9883–9888.

Biondi E, Toro N, Bazzicalupo M, Martínez-Abarca F (2011). Spread of the group II intron RmInt1 and its insertion sequence target sites in the plant endosymbiont *Sinorhizobium meliloti*. *Mob Genet Elements* **1**: 1–7.

Bittinger MA, Gross JA, Widom J, Clardy J, Handelsman J (2000). *Rhizobium etli* CE3 carries virgene homologs on a self-transmissible plasmid. *Mol Plant Microbe Interact* **13**: 1019–1021.

Capela D, Barloy-Hubler F, Gouzy J, Bothe G, Ampe F, Batut J *et al.* (2001). Analysis of the chromosome sequence of the legume symbiont *Sinorhizobium meliloti* strain 1021. *Proc Natl Acad Sci USA* **98**: 9877–9882.

Chillón I, Martínez-Abarca F, Toro N (2010). Splicing of the *Sinorhizobium meliloti* RmInt1group II intron provides evidence of retroelement behavior. *Nucleic Acids Res* **39**: 1095–1104.

Coros CJ, Piazza CJ, Chalamcharla VR, Smith D, Belfort M (2009). Global regulators orchestrate group II intron retromobility. *Mol Cell* **34**: 250–256.

Crook MB, Lindsay DP, Biggs MB, Bentley JS, Price JC, Clement SC *et al.* (2012). Rhizobial plasmids that cause impaired symbiotic nitrogen fixation and enhanced host invasion. *Mol Plant Microbe Interact* **25**: 1026–1033.

Dai LX, Zimmerly S (2002). The dispersal of five group II introns among natural populations of *Escherichia coli*. *RNA* **8**: 1294–1307.

Darling AC, Mau B, Blattner FR, Perna NT (2004). Mauve: multiple alignmentof conserved genomic sequence with rearrangements. *Genome Res* **14**: 1394–1403.

Eickbush TH (1994). Origins and evolutionary relationships of retroelements. In: Morse SS (ed.) *The Evolutionary Biology of Viruses*. Raven Press: New York, pp 121–157.

Fernandez-Lopez M, Munoz-Adelantado E, Gillis M, Willems A, Toro N (2005). Dispersal and evolution of the *Sinorhizobium meliloti* group II RmInt1 intron in bacteria that interact with plants. *Mol Biol Evol* **22**: 1518–1528.

Finan TM, Weidner S, Wong K, Buhrmester J, Chain P, Vorhölter FJ *et al.* (2001). The complete sequence of the 1,683-kb pSymB megaplasmid from the N-2-fixing endosymbiont *Sinorhizobium meliloti*. *Proc Natl Acad Sci USA* **98**: 9889–9894.

Flores M, Morales L, Avila A, González V, Bustos P, García D *et al.* (2005). Diversification of DNA sequences in the symbiotic genome of *Rhizobium etli*. *J Bacteriol* **187**: 7185–7192.

Fujishige NA, Kapadia NN, De Hoff PL, Hirsch AM (2006). Investigations of *Rhizobium* biofilm formation. *FEMS Microbiol Ecol* **56**: 195–206.

Gage DJ (2004). Infection and invasion of roots by symbiotic, nitrogen-fixing rhizobia during nodulation of temperate legumesMicrobiol. *Mol Biol Rev* **68**: 280–300.

Galardini M, Mengoni A, Brilli M, Pini F, Fioravanti A, Lucas S *et al.* (2011). Exploring the symbiotic pangenome of the nitrogen-fixing bacterium *Sinorhizobium meliloti*. *BMC Genomic* **12**: 235.

Galardini M, Pini F, Bazzicalupo M, Biondi EG, Mengoni A (2013). Replicon-dependent bacterial genome evolution: the case of *Sinorhizobium meliloti*. *Genome Biol Evol* **5**: 542–558.

Galibert F, Finan TM, Long SR, Puhler A, Abola P, Ampe F *et al.* (2001). The composite genome of the legume symbiont *Sinorhizobium* meliloti. *Science* **293**: 668–672.

Girard ML, Flores M, Brom S, Romero D, Palacios R, Dávila G (1991). Structuralcomplexity of the symbiotic plasmid of *Rhizobium leguminosarum* bv. phaseoli. *J Bacteriol* **173**: 2411–2419.

González V, Santamaría RI, Bustos P, Hernández-González I, Medrano-Soto A, Moreno-Hagelsieb G *et al.* (2006). The partitioned *Rhizobium etli* genome: genetic and metabolic redundancyin seven interacting replicons. *Proc Natl Acad Sci USA* **103**: 3834–3839.

González V, Acosta JL, Santamaría RI, Bustos P, Fernández JL, Hernández González IL *et al.* (2010). Conserved symbiotic plasmid DNA sequences in the multireplicon pangenomic structure of *Rhizobium etli*. *Appl Environ Microbiol* **76**: 1604–1614.

Guindon S, Gascuel O (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* **52**: 696–704.

Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* **59**: 307–321.

Harrison PW, Lower RPJ, Kim NKD, Young JPW (2010). Introducing the bacterial 'chromid': not a chromosome, not a plasmid. *Trends Microbiol* **18**: 141–148.

Hengge R (2009). Principles of c-di-GMP signalling in bacteria. *Nat Rev Microbiol* **7**: 263–273.

Huelsenbeck JP, Ronquist F (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**: 754–755.

Jenal U, Malone J (2006). Mechanisms of cyclic-di-GMP signaling in bacteria. *Ann Rev Genet* **40**: 385–407.

Jiménez-Zurdo JI, García-Rodríguez FM, Barrientos-Durán A, Toro N (2003). DNA target site requirements for homing *in vivo* of a bacterial group II intron encoding a protein lacking the DNA endonuclease domain. *J Mol Biol* **326**: 413–423.

Leclercq S, Cordaux R (2012). Selection-driven extinction dynamics for group II introns in Enterobacteriales. *PLoS One* **7**: e52268.

Lambowitz AM, Zimmerly S (2011). Group II introns: mobile ribozymes that invade DNA. *Cold Spring Harb Perspect Biol* **3**: a003616.

MacLean AM, Finan TM, Sadowsky MJ (2007). Genomes of the symbiotic nitrogen-fixing bacteria of legumes. *Plant Physiol* **144**: 615–622.

Martínez-Abarca F, Zekri S, Toro N (1998). Characterization and splicing *in vivo* of a *Sinorhizobium meliloti* group II intron associated with particular insertion sequences of the IS630-Tc1/IS3 retroposon superfamily. *Mol Microbiol* **28**: 1295–1306.

Martínez-Abarca F, Barrientos-Durán A, Fernández-López M, Toro N (2004). The RmInt1 group II intron has two different retrohoming pathways for mobility using predominantly the nascent lagging strand at DNA replication forks for priming. *Nucleic Acids Res* **32**: 2880–2888.

Martinez-Abarca F, Martinez-Rodriguez L, Lopez-Contreras JA, Jiménez-Zurdo JI, Toro N (2013). Complete genome sequence of the alfalfa symbiont *Sinorhizobium/Ensifer melilotistrain* GR4. *Genome Announc* **1**: e00174–12.

Martens M, Delaere M, Coopman R, De Vos P, Gillis M, Willens A (2007). Multilocus sequence analysis of *Ensifer* and related taxa. *Int J Syst Evol Microbiol* **57**: 489–503.

Michel F, Ferat JL (1995). Structure and activities of group II introns. *Annu Rev Biochem* **64**: 435–461.

Michel F, Costa M, Westhof E (2009). The ribozyme core of group II introns: a structure in want of partners. *Trends Biochem Sci* **34**: 189–199.

Mohr G, Perlman PS, Lambowitz AM (1993). Evolutionary relationships among group II intron-encoded proteins and identification of a conserved domain that may be related to maturase function. *Nucleic Acids Res* **21**: 4991–4997.

Mohr G, Ghanem E, Lambowitz AM (2010). Mechanisms used for genomic proliferation by thermophilic group II Introns. *PLoS Biol* **8**: e1000391.

Muñoz E, Villadas PJ, Toro N (2001). Ectopic transposition of a group II intron in natural bacterial populations. *Mol Microbiol* **41**: 645–652.

Nisa-Martínez R, Jiménez-Zurdo JI, Martínez-Abarca F, Muñoz-Adelantado E Toro N (2007). Dispersion of the RmInt1 group II intron in the *Sinorhizobium meliloti* genome upon acquisition by conjugative transfer. *Nucleic Acids Res* **35**: 214–222.

Penn O, Privman E, Landan G, Graur D, Pupko T (2010). An alignment confidence score capturing robustness to guide tree uncertainty. *Mol Biol Evol* **27**: 1759–1767.

Reeve W, Chain P, O'Hara G, Ardley J, Nandesena K, Bräu L *et al.* (2010). Complete genome sequence of the Medicago microsymbiont Ensifer (Sinorhizobium) medicae strain WSM419. *Stand Genomic Sci* **2**: 77–86.

Rome S, Fernández M-P, Brunel B, Normand Ph, Cleyet-Marel J-C (1996). *Sinorhizobium medicae* sp. Nov., isolated from Annual Medicago spp. *Int J Syst Bacteriol* **46**: 972–980.

Römling U, Galperin MY, Gomelsky M (2013). Cyclic di-GMP: the first 25 years of a universal bacterial second messenger. *Microbiol Mol Biol Rev* **77**: 1–52.

San Filippo J, Lambowitz AM (2002). Characterization of the C-terminal DNA-binding/DNA endonuclease region of a group II intron-encoded protein. *J Mol Biol* **324**: 933–951.

Schmeisser C, Liesegang H, Krysciak D, Bakkou N, Le Quéré A, Wollherr A *et al.* (2009). *Rhizobium* sp. strain NGR234 possesses a remarkable number of secretion systems. *Appl Environ Microbiol* **75**: 4035–4045.

Schirmer T, Jenal U (2009). Structural and mechanistic determinants of c-di-GMP signalling. *Nature Rev Microbiol* **7**: 724–735.

Schneiker-Bekel S, Wibberg D, Bekel T, Blom J, Linke B, Neuweger H *et al.* (2011). The complete genome sequence of the dominant *Sinorhizobiummeliloti* field isolate SM11 extends the *S. meliloti* pan-genome. *J Biotech* **155**: 20–33.

Simon DM, Clarke NA, McNeil BA, Johnson I, Pantuso D, Dai L *et al.* (2008). Group II introns in eubacteria and archaea: ORF-less introns and new varieties. *RNA* **14**: 1704–1713.

Sprent JI, James EK (2007). Legume evolution: Where do nodules and mycorrhizas fit in? *Plant Physiol* **144**: 575–581.

Sugawara M, Epstein B, Badgley BD, Unno T, Xu L, Reese J *et al.* (2013). Comparative genomics of the core and accessory genomes of 48 *Sinorhizobium* strains comprising five genospecies. *Genome Biol* **14**: R17.

Tian CF, Zhou YJ, Zhang YM, Li QQ, Zhang YZ, Li DF *et al.* (2012). Comparative genomics of rhizobia nodulating soybean suggests extensive recruitment of lineage-specific genes in adaptations. *Proc Natl Acad Sci USA* **109**: 8629–8634.

Toro N, Martínez-Abarca F (2013). Comprehensive phylogenetic analysis of bacterial group II intron-encoded ORFs lacking the DNA endonuclease domain reveals new varieties. *PLoS One* **8**: e55102.

Turner SL, Young JPW (2000). The glutamine synthetases of rhizobia: phylogenetics and evolutionary implications. *Mol Biol Evol* **17**: 309–319.

Weidner S, Baumgarth B, Göttfert M, Jaenicke S, Pühler A, Schneiker-Bekel S *et al.* (2013). Genome sequence of *Sinorhizobium meliloti* Rm41. *Genome Announc* **1**: e00013–12.

Weir BS (2012). The current taxonomy of rhizobia. NZ Rhizobia website. http://www.rhizobia.co.nz/taxonomy/rhizobia. Last updated: 10 April 2012

Werren JH (2011). Selfish genetic elements, genetic conflict, and evolutionary innovation. *Proc Natl Acad Sci USA* **108**: 10863–10870.

Supplementary Information accompanies this paper on Heredity website (http://www.nature.com/hdy)