MICROBIAL GENOMICS

RESEARCH ARTICLE

Nelson *et al*., *Microbial Genomics* 2018;4

DOI 10.1099/mgen.0.000174

The complete replicons of 16 *Ensifer meliloti* strains offer insights into intra- and inter-replicon gene transfer, transposon-associated loci, and repeat elements

Matthew Nelson,[1] Joseph Guhlin,[2] Brendan Epstein,[2] Peter Tiffin[2] and Michael J. Sadowsky[1,*]

Abstract

Ensifer meliloti (formerly *Rhizobium meliloti* and *Sinorhizobium meliloti*) is a model bacterium for understanding legume–rhizobial symbioses. The tripartite genome of *E. meliloti* consists of a chromosome, pSymA and pSymB, and in some instances strain-specific accessory plasmids. The majority of previous sequencing studies have relied on the use of assemblies generated from short read sequencing, which leads to gaps and assembly errors. Here we used PacBio-based, long-read assemblies and were able to assemble, *de novo*, complete circular replicons. In this study, we sequenced, *de novo*-assembled and analysed 10 *E. meliloti* strains. Sequence comparisons were also done with data from six previously published genomes. We identified genome differences between the replicons, including mol% G+C and gene content, nucleotide repeats, and transposon-associated loci. Additionally, genomic rearrangements both within and between replicons were identified, providing insight into evolutionary processes at the structural level. There were few cases of inter-replicon gene transfer of core genes between the main replicons. Accessory plasmids were more similar to pSymA than to either pSymB or the chromosome, with respect to gene content, transposon content and G+C content. In our population, the accessory plasmids appeared to share an open genome with pSymA, which contains many nodulation- and nitrogen fixation-related genes. This may explain previous observations that horizontal gene transfer has a greater effect on the content of pSymA than pSymB, or the chromosome, and why some rhizobia show unstable nodulation phenotypes on legume hosts.

DATA SUMMARY

All genomes were deposited in GenBank under Bioproject PRJNA388336 with the following accession numbers: HM006_ SRR6032557, SRR6032556, SRR6032555, SRR6032554; KH35C_ SRR6032553; KH46_ SRR6032552; M162_ SRR6032551; M270_ SRR6032550; Rm41_ SRR6032549; T073_ SRR6032548; USDA1021_ SRR6032559; USDA1106_ SRR6032558; USDA1157_ SRR6032560.

INTRODUCTION

Rhizobia are an important group of bacteria because of the symbioses they form with legume plants. These bacteria provide the plant with fixed nitrogen by converting atmospheric N_2 into a plant-usable form. In exchange, the plant provides carbon to the rhizobia located within root or stem nodules, thereby supporting greater bacterial growth and reproduction. The *Medicago truncatula*–*Ensifer meliloti* (*Sinorhizobium meliloti*) symbiosis is a model system to better understand the genetic basis and evolution of rhizobial–legume symbioses and the N_2-fixation process [1]. *E. meliloti* strains contain more than one large replicon, which is similar to roughly 10 % of assayed bacterial species [2]. In *E. meliloti* large non-chromosomal replicons are referred to as megaplasmids or symbiotic (Sym) plasmids [2].

The *E. meliloti* reference genome, Rm1021, has a chromosome and two megaplasmids, pSymA and pSymB, but not other small accessory plasmids [1]. Previous work has shown that genes involved in similar functions tend to be concentrated on particular replicons: pSymA contains genes playing essential roles in symbiosis, including nodule formation and symbiotic nitrogen fixation [1], whereas pSymB contains a large proportion of genes involved in import/

Received 21 January 2018; Accepted 26 March 2018

Author affiliations: [1]Biotechnology Institute and Department of Soil, Water, and Climate, University of Minnesota, St. Paul, MN 55108, USA; [2]Department of Plant and Microbial Biology, University of Minnesota, St. Paul, MN 55108, USA.

***Correspondence:** Michael J. Sadowsky, sadowsky@umn.edu

Keywords: genome evolution; structural variation; repeat and IS elements; accessory plasmids; comparative genomics; rhizobia.

Abbreviations: HMM, hidden Markov model; IS, insertional sequence; MCL, Markov clustering; PCA, principal component analysis; RE, repeat elements; TAL, transposon-associated loci.

Data statement: All supporting data, code and protocols have been provided within the article or through supplementary data files. Seven supplementary tables and eight supplementary figures are available with the online version of this article.

export functions [3], and the chromosome contains most of the housekeeping genes [4–6]. These replicon-specific gene functions have been hypothesized to be the result of the initial acquisition of plasmids followed by horizontal gene transfer events [6]. Some strains also contain smaller accessory plasmids [7], some of which have been shown to affect nodulation and metabolic potential [8, 9].

Previous studies have shown that the three main replicons of *E. meliloti* have distinct evolutionary histories [10], and inter-replicon differences can be seen in the: (1) levels of standing nucleotide variation [10], (2) effects of purifying and positive selection [5, 11], (3) proportion of duplicated and horizontally transferred genes [11], and (4) structural rearrangements and core gene content [5]. These studies have relied primarily on the use of short-read sequencing (mainly Illumina-based Hiseq or Miseq platforms) and a limited number of complete genomes. However, it is well known that mapping short-read sequences back to a single reference genome biases assembly and downstream analyses to what was found in the reference genome, and thus provides limited insight into large structural variation, genes missing from the reference genome, intra-replicon gene movement and quantification of repeated sequences [12, 13].

Fully assembled, reference-quality genomes generated from long-read technologies, such as Pacific Biosciences (PacBio) sequencing [14], allow for better assignment of genome rearrangements, repetitive sequences, gene content and evolutionary histories of populations. This enables characterization of transposon-associated loci (TALs), genes that encode proteins that may mediate the transposition and duplications of DNA within the genome [15], and repeat elements (REs), sequences of DNA repeated one or more times within a single genome. These features are difficult to assemble and differentiate using short-read data. The TALs and REs have been inferred to facilitate roles in gene movement among bacterial lineages [15], and presumably facilitate movement of genic regions between replicons [16].

Accessory plasmids are small replicons present in some, but not all, *E. meliloti* strains [7, 9, 13] and some have been identified as facilitating important biological roles such as metabolic potential, host incompatibility and nodulation competitiveness [9]. Relatively little is known about the origin or evolution of accessory plasmids in *Ensifer*, although they are generally thought to be transient components (easily gained or lost) from the species pan-genome. The limited data on *E. meliloti* accessory plasmids suggest that genes found on these plasmids are similar to those found on pSymA or pSymB. However, no extensive genomic analyses of multiple plasmids have been done [but see 7, 13, 17], and reference-quality genomes can provide data for such analyses.

Here we describe the complete genome sequence of 16 *E. meliloti* reference-quality genomes, 10 of which were newly sequenced in this study using high-coverage PacBio data. The fully assembled genome sequences were used to characterize: (1) the diversity of gene content, TALs and REs found in *E. meliloti*, (2) gene transfer events between replicons and (3) genomic composition and relationships among *E. meliloti* accessory plasmids. Our research shows that there is benefit in analysing *E. meliloti* in a replicon-independent manner, and describes how *E. meliloti* can gain and lose genes in a manner consistent with maintenance of overall genome stability and functionality.

IMPACT STATEMENT

This article provides evidence that the three main replicons of *Ensifer meliloti* have distinct gene content and gene patterns that are maintained even as strains genetically diverge. The symbiotically important replicon, pSymA, appears to preferentially recombine or transfer genes with the accessory plasmids. Some accessory plasmids were previously shown to be involved in host specificity. This suggests a mechanism for gaining or losing genes involved in symbiosis and host selection. The replication protein for the accessory plasmids, RepA, is found in numerous other bacterial species and may have been acquired by plasmid transfer from other soil microbiota. Our results reveal that some structural rearrangements of replicons in *E. meliloti* are common, but that gene translocation between replicons is relatively rare, or selected against. Intra-replicon gene transfer is associated with repeat elements, but not transposon-associated loci. The gene transfer events that occurred between the accessory plasmids and pSymA in *E. meliloti* demonstrate one mechanism by which the *Ensifer*–legume symbiosis is constantly evolving. Overall, and despite what has been found in other rhizobia, *E. meliloti* has a fairly stable genome structure on an evolutionary timescale.

METHODS

Genome sequencing, assembly and annotation

Genomic DNA from the 10 *E. meliloti* strains was isolated by using UltraClean Microbial DNA Isolation Kits from MoBio Laboratories. Strains were previously obtained [18] or newly acquired from the USDA culture collection. Cultures were grown at 30 °C in TY medium [19]. Genomic sequence data were generated using a Pacific Biosciences (PacBio) RS II sequencer at the Mayo Clinic with one PacBio single molecule real time (SMRT) cell per strain. Genomes were assembled using HGAP version 3.0 [20]. Each genome was assembled multiple times, adjusting the predicted size of the genome during each subsequent assembly run. The genomes were circularized and replicons were confirmed using gepard version 1.3.1 [21]: coverage range, 33.3–153.6; read range, 20 333–100 711; N50 range, 17 362–20 321; and average read length range, 10 826–20 341. The assembled genomes were individually polished with Pilon 1.16 [22], using Illumina reads from previous studies that were mapped to the PacBio assemblies using BWA version 0.7.12-r1039, and using the 'mem' algorithm [23]. Files were converted from SAM to BAM files using samtools view 1.3 [24]. Pilon was run with the required arguments and additionally '—changes —fix "bases" '. The complete commands for running BWA, samtools and Pilon on the M2 genome are available at: https://github.com/jguhlin/pacbio-paper-code/blob/master/pilon-M2-genome/run-pilon.sh. Base pair changes during the polishing stage for each assembly ranged from 4 to 624 (<0.01 % of the genome) [22]. The protein-coding genes of each assembled genome were predicted using prodigal, with no specialized parameters [25]. Replicon names were assigned based on sequence similarity to the reference strain *E. meliloti* Rm2011 [26]. The six previously sequenced strains were imported from NCBI, and all genomes were also independently assembled and annotated using the MaGE and the MicroScope platform (http://www.genoscope.cns.fr/agc/microscope/home/index.php). Annotations are also available at: https://github.com/jguhlin/pacbio-paper-code/tree/master/gene-prediction.

Identification of syntenic regions, core and pan genomes

Synteny analyses was performed using NUCMER from the MUMMER package (version 3.1) [27]. Plots and downstream analyses were done using custom code. Core and pan genomes were generated by performing an all-vs-all BLAST+ comparison on the predicted protein sequences, clustering based on BLAST+ bitscores using mcl with an inflation value of 10.0, due to strains belonging to a single species. Orthology-based approaches are not appropriate for this analysis [28, 29]. Identification of single-copy core genes was performed using an ODG database and a CYPHER query, available at the Github repository referenced below, with additional analysis with custom code [30]. Custom code, commands and scripts are available at: http://github.com/jguhlin/pacbio-paper-code. The 'Pan/Core-Genome' and the 'Gene Phyloprofile' tool of the MicroScope platform was also used to identify similar genes between bacterial strains and individual replicons, with thresholds set at the recommended MicroScope specifications of 80 % amino acid identity and 80 % alignment coverage [31, 32]. The MicroScope protein-coding gene annotations were used to generate core-genome and core-replicon gene content. Inter-replicon gene movements were defined at genes present in the core-genome, but absent in the core-replicon for the chromosome, pSymA and pSymB.

Identification of transposon-associated elements and repetitive elements

To predict TALs, *de novo* gene prediction was performed on each genome using Prodigal v2.6.3, and exporting coding sequences, peptide sequences and a GFFv3 file detailing predicted genes. The specific commands used are available at: https://gist.github.com/jguhlin/67811311c36e35b0c1ac2ef772c129cb [25]. Functional predictions were generated using orthologous functional prediction via eggNOG v4.5.1 and matched using HMMER v3.1b2 [33, 34]. The TALs were identified based on matching one or more eggNOG-based annotations associated with transposable elements. TALs were subsequently clustered using MCL V14–137 based on sequence similarity bit scores determined by BLAST+
[28, 29, 35].

Repeats and repeated elements were identified using NUCMER, from the MUMMER version 3.23 package [27], as follows. Total genomic content of each strain was compared against itself (using the command '*nucmer –p ID-vs-ID –maxmatch –nosimplify*' *ID.fasta ID.fasta,* where ID is the strain ID), and the sequences of matches were extracted and combined with the matches from all strains. NUCMER was subsequently used to compare all repeats to each other, and MCL was used to cluster repeats based on sequence matching coverage pairwise for each repeat [35]. Repeats were aligned using MAFFT (with arguments —*maxiterate 1000 —localpair –adjustdirection*), and poor alignments were trimmed or removed using TrimAL (-*resoverlap 0.6 -seqoverlap 60*). Repeats were re-aligned with MAFFT, using the same arguments as previously [36, 37]. HMM profiles were built from these alignments using HMMER [34], and the genome of each strain was analysed for repeat content using nhmmscan, with an e-value cutoff of 0.0001 [34]. Regions matching multiple HMM profiles were assigned to the best match, by either length, identity percentage or score, in that order. Principal component analyses (PCAs) were performed on both the repetitive element and the transposable-associated element contents of each replicon in each strain [38].

Mantel test

Correlations among pairwise genetic distance matrices were tested for each replicon. A genetic distance matrix for each replicon was constructed using the concatenated alignments of single copy core genes and the dist.dna function from the R [39] package ape [39]. This was done using the TN93 model of evolution [40]. A Mantel test [41] was implemented in the ade4 package [39], with 10 000 permutations, to calculate the correlation between distance matrices from different replicons and to test for significance.

Relationships among RepA sequences

The relationships among the RepA protein-coding genes from several bacterial species were characterized by constructing a maximum-likelihood phylogeny using the PhyML implementation on thewww.phylogeny.fr web-server, a part of Méthodes et algorithmes pour la bio informatique [42–48].

RESULTS

Replicon overview

All replicons in the 10 new strains sequenced in this study were *de novo* and completely assembled, and fully circularized using both PacBio and Illumina sequencing data. Our analyses were supplemented with six previously published complete genomes [1, 13, 26, 49, 50]. The total sizes of genomes from these 16 strains ranged from 6.68 to 7.27 Mb (Tables 1 and S1, available in the online version of this article). The strains were obtained from geographically diverse regions of the USA, Europe, Australia and the Middle East (Table S1). Each genome contained three main replicons – a chromosome and the two megaplasmids, pSymA and pSymB – and in many cases from one to three accessory plasmids. The chromosomes had an average G+C content of 62.72 mol% – similar to pSymB, which had an average of 62.40 mol% (pairwise *t*-test for difference in G+C content, P=0.34 after Bonferroni correction). In contrast, replicon pSymA had substantially lower G+C content, 60.31 mol% (*t*-test P<0.001 compared to the chromosome and pSymB), similar to the 59.18 mol% average of the accessory plasmids. The accessory plasmids showed great variance in G+C content, with $\sigma^2_{\text{accessory}} = 0.888$, $\sigma^2_{\text{psyma}} = 0.020$, $\sigma^2_{\text{psymb}} = 0.119$ and $\sigma^2_{\text{main}} = 0.0052$, perhaps reflecting their diverse origins.

General core genome and replicon-specific core genomes

The core genome of *E. meliloti* was defined by us as consisting of the intersection of all gene families (clusters) found in each of our assayed strains. We clustered genes based on predicted protein similarity from strains: AK83, SM11, BL225c, GR4, RM2011, RM1021, HM006, KH35c, KH46c, M270, RM41, T073, USDA1106, USDA1157, M162 and USDA1021. The core genome of our sample was composed of 4315 gene clusters (Fig. 1). We also performed this analysis using individual replicons, identifying 2472 core gene clusters for the chromosome, 308 for pSymA and 1242 for pSymB (Figs S1–S3). The core genome of the chromosome segments into three distinct patterns due to the inclusion of strains M270 and USDA1021. The pattern resulting from M270 is probably due, for the most part, to a deletion in the chromosome; the chromosome of M270 is the second smallest in our population at 3.5 Mbp. USDA1021 contains a translocation from the chromosome to pSymB, as seen in Fig. 2(b). Fig. S2 segments into two because of a likely split of the pSymA plasmid in strain M162. Because of the large translocations present in strains M162 and USDA1021 (Fig. 2), separate analysis on core genome structure was done excluding these two strains. For this slightly smaller population, 4389 core gene clusters were identified. This accounted for 69–75 % of all genes in the 16 strains (Table S2).

Core genes specific to the chromosome (2734 genes) accounted for 79–87 % of the total genes on each chromosome. Core genes on pSymA (536 genes) accounted for 40–53 % and pSymB (1255 genes) accounted for 85–90 % of the total genes on each replicon (Table S2). The percentage of genes that were part of the core were significantly different between replicons: the chromosome and pSymA (pairwise *t*-tests, Bonferroni-corrected: $P<2\times10^{-16}$), the chromosome and pSymB (P=0.023), and pSymA and pSymB ($P<2\times10^{-16}$).

Inter-replicon gene movement

The complete genome assemblies allowed for the identification of inter-replicon gene movement that would not have been identifiable by using short-read sequencing. This is due, in large part, to gaps, potentially unplaced contigs and misplaced contigs due to insertional sequence (IS) elements. To identify inter-replicon gene movement, the gene content for each replicon was compared to those in all strains via BLAST. Based on the analysis of single-copy core genes, 102 gene translocation events were identified. These were primarily from pSymB to pSymA (62 genes) and the chromosome to pSymB (33 genes). In all cases, the genes were found in 15 strains on an equivalent replicon, and in a single strain on a different replicon. In addition, there was one case where a gene moved from pSymA to pSymB in one strain, but that same gene moved to an accessory plasmid in another strain.

Two large sequence translocation events were identified in *E. meliloti* strains M162 and USDA1021 (Fig. 2). In strain M162, we identified a translocation of 300 kbp from pSymA into a large accessory plasmid. The MaGe annotation for the genes on the accessory plasmid in M162 which match pSymA genes is M162 (pA0019–pA0464). None of these genes was annotated as being *nod*, *nif* or *fix* genes. Moreover, the pSymA in strain M162 had a replication protein (RepA) identical to that found on an accessory plasmid in

Table 1. Genomic properties of the chromosomes, megaplasmids and accessory plasmids from 16 characterized *E. melioti* strains

For each entry, the range of values is presented in parentheses below the mean. Detailed data are available in Table S1. While no accessory plasmid exceeded 1 Mb in size, the number of TALs and REs per megabase is presented to allow for a comparison.

Characteristic	Replicon			
	Chromosome	pSymA	pSymB	Accessory
Size (Mb)	3.69 (3.43–3.91)	1.41 (0.89–1.63)	1.69 (1.62–2.01)	0.23 (0.07–0.42)
Number of genes	3807 (3545–4203)	1699 (1439–2100)	1700 (1620–1747)	298 (107–497)
% genes that are core	0.72 (0.69–0.75)	0.45 (0.40–0.46)	0.87 (0.85–0.87)	0
G+C content (mol%)	62.72 (62.57–62.83)	60.31 (59.97–60.54)	62.40 (62.18–62.6)	59.18 (57.45–60.66)
RE number/Mb	431.4 (417.0–446.5)	437.1 (380.0–493.8)	399.2 (390.0–408.3)	571.1 (268.0–853.3)
RE families	464	360	453	289
TAL number/Mb	86.9 (80.4–92.9)	134.9 (109.8–160.3)	115.8 (106.5–141.7)	251.0 (112.5–440.0)
TAL families	281	184	140	205

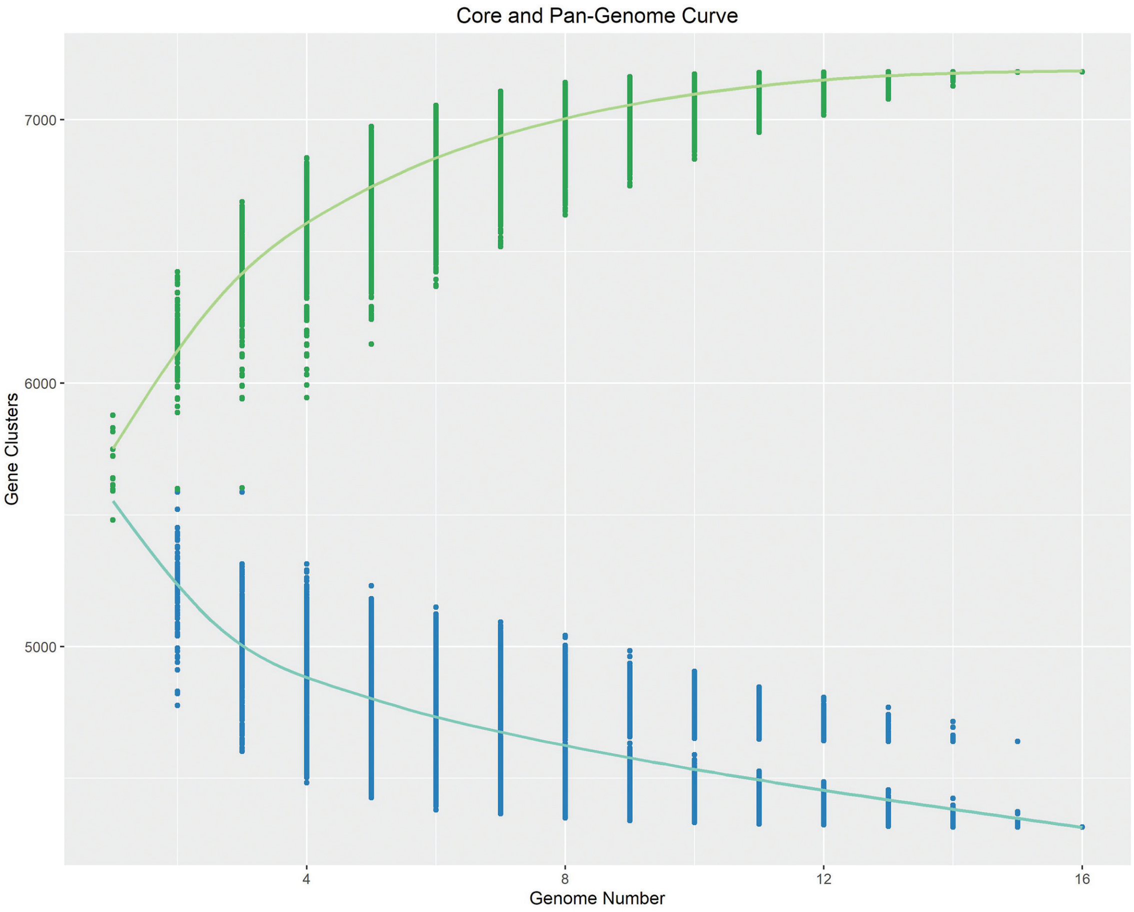

Fig. 1. Core- and pan-genome curves for our 16 *E. meliloti* strains. Green represents the pan-genome curve, and each point represents a combination of different genomes. Blue represents the core-genome curve, with a combination of our 16 genomes specified by genome number. Genomes are represented by distinct gene clusters, rather than individual genes.

Ensifer medicae strains M2 and WSM419, suggesting a likely horizontal gene transfer event also occurred. RepA is a main replication protein found on non-chromosomal plasmids [51]. Furthermore, the RepA from accessory plasmid M162 was identical to those found on the pSymA replicons in the other *E. meliloti* strains. This indicated that a genomic translocation event involving pSymA and an accessory plasmid occurred in strain M162.

A second large gene transfer event was detected in strain USDA1021, where a 325 kbp region, containing 337 genes, moved from the chromosome to pSymB. The MaGe gene annotations (mpb0525–mpb0862) show these genes are involved in flagella biosynthesis and ATP production. This resulted in a 325 kb increase in the size of pSymB in strain USDA1021 and similar sized reduction in the chromosome as in the other strains (Table S1).

Transposon-associated loci and repeat elements by replicon

TALs were identified by running a motif search on predicted protein sequences for our strains, and identifying related eggNOG categories. These 1768 eggNOG classifications included transposase, integrase, recombinase, phage integrase family and IS family members (Table S3). About 12 139 TALs were identified in our bacterial population, with 683–981 TALs per genome. The TAL sequences comprised 655–856 kb, accounting for 9.3–11.7 % of the total genome length (Table 1). Of the three main replicons, pSymA had the highest average density of TALs, with a mean 134.9 TALs/Mb. In contrast, pSymB had 115.8 TALs/Mb and the chromosome 86.92 TALs/Mb. The accessory plasmids had an average of 251.0 TALs/Mb, although the accessory plasmids varied widely with a range of 112.5–440 TALs/Mb. Based on sequence identity, the 12 139 TALs were clustered into 474 TAL families, 173 of which were found in every strain, and 42 of which were found in only a single strain. Of the 173 core TALs, 68 were exclusive to a single replicon and 29 were found on all three of the main replicons.

Despite being much smaller than the other replicons, the accessory plasmids contained almost half of all the TAL families (Table 1). The TALs on accessory plasmids were

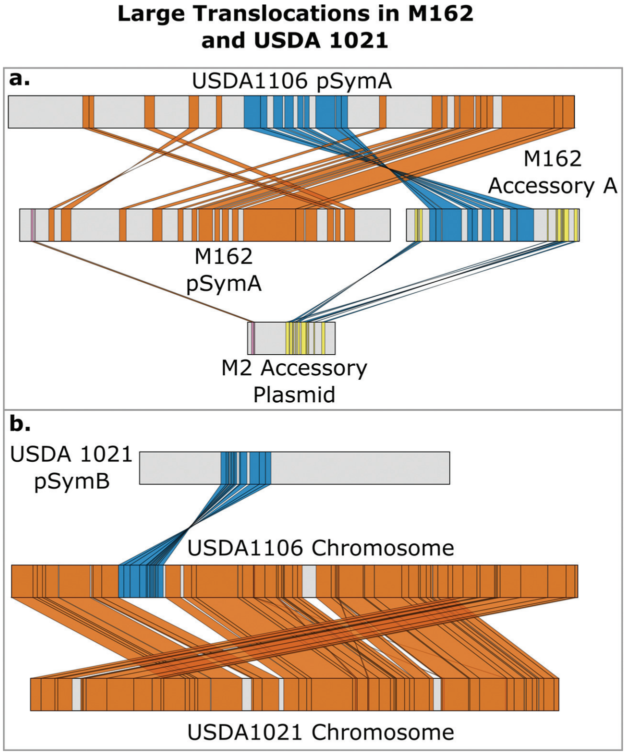

Fig. 2. Large translocation events identified in strains M162 and USDA1021, with USDA1106 serving as a typical example of the replicons. (a) *E. meliloti* pSymA appears to have split in strain M162, as identified by a typical pSymA. Furthermore, M162 pSymA shares a replication protein with an accessory plasmid of *E. medicae* strain M2. (b) USDA1021 shows a large translocation from the chromosome to pSymB. For clarity, links shown are as follows: USDA1106–M162 required ≥ 10 kbp and ID ≥ 92 %, while M162–M2 required ≥ 1.5 kbp and ID ≥ 90 %. USDA1106–USDA1021 ≥ 5 kbp and ID ≥ 96.5 %

also much more likely to be found in only a single strain; more than half of the accessory plasmid TALs were found in only a single strain whereas <25 % of the chromosomal TALs were found in a single strain. In contrast, >50 % of the chromosomal TALs and >40 % of pSymB TALs were found in nearly all of the strains (Fig. 3).

A total of 327, 83 and 49 of the TAL families were found on only one, two or three of the main replicons, respectively. If the accessory plasmids were also counted as a single replicon class, 232, 156 and 39 TAL families were found on one, two or three main replicons, respectively. Moreover, 47 TALs were found on all three main replicons and one or more of the accessory plasmids. The presence or absence of specific TALs was more related to a replicon than to a specific strain (Fig. 4).

We also identified repeat sequence elements (REs) by first searching for regions of high similarity within each genome [27], and then clustering these sequences, based on sequence identity, across all genomes to form hidden Markov models (HMMs). This process identified 48 133 repeated elements, ranging in size from 66 to 12 593 bp, that clustered into 688 repeat families by similarity Markov clustering (MCL) [52].

Each strain contained from 2702 to 3132 REs, accounting for 1.46–2.04 Mb of DNA, and comprising ~22–28 % of the total sequence content of each strain. The REs were more evenly distributed amongst the three main replicons than were the TALs: the chromosome contained 417–446 repeats/Mb, while pSymA contained 380–561 repeats/Mb, and pSymB 390–408 repeats/Mb (Table 1). In contrast to TALs (Fig. 3), approximately 50 % of the repeat families were found in all strains, and less than 5 % were found in only a single strain (Fig. 3). Although the majority of REs were not replicon-specific, the distribution of REs among all replicons and strains was more related to a replicon than to a strain (Fig. 4). Like TALs, the REs could be clustered into families with 78, 22 and nine found only on pSymA, pSymB or the accessory plasmids, respectively. The accessory plasmids and pSymA shared 68 RE families that were not found on the chromosome or on pSymB, whereas there were 14 RE families shared by the accessory plasmids and the chromosome but not with pSymA, and five RE families were shared by the accessory plasmids and pSymB but not with pSymA.

Accessory plasmids as part of the pSymA pan-replicon

The 14 accessory plasmids we identified in our sample contained 3215 predicted genes. These accessory plasmids had a similar ratio of predicted genes, including TALs (1.01 genes/kb), to the other replicons, the chromosome (1.03 genes/kb), pSymB (1.01 genes/kb) and pSymA (1.17 genes/kb). As expected, there were no genes in common among all the accessory plasmids, and several of these plasmids shared little sequence identity with any of the other accessory plasmids. Adding the accessory genome to the pSymA core and pan genome increased the number of core genes from 536 to 553 (excluding M162 and USDA1021, Fig. S4). The RepA protein was not identical across all accessory plasmids based on amino acid comparison (Fig. S5), indicating a potentially diverse source of genetic material [53]. RepA proteins that were highly similar to those found on our accessory plasmids were also found in other rhizobial species, including *E. medicae*, *Ensifer fredii* and *Agrobacterium tumefaciens*.

The gene content of the accessory plasmids was compared to that of the pan-replicon of each of the three main replicons for all 16 strains, based on gene annotation from MaGe (Table 2). The results of this analysis indicated that more genes found on the accessory genome were exclusively found on pSymA, rather than pSymB or the chromosome. This suggests that the accessory plasmids and pSymA have a greater ability to exchange genetic elements than the other replicons. TALs accounted for about 12 % of the total gene content of the accessory plasmids. Less than 2 % of TALs were exclusive to the chromosome or pSymB, whereas 32 % of the accessory plasmid genes were also found exclusively on pSymA.

Further evidence of a shared pSymA/accessory plasmid genome was found by performing a BLAST analysis against

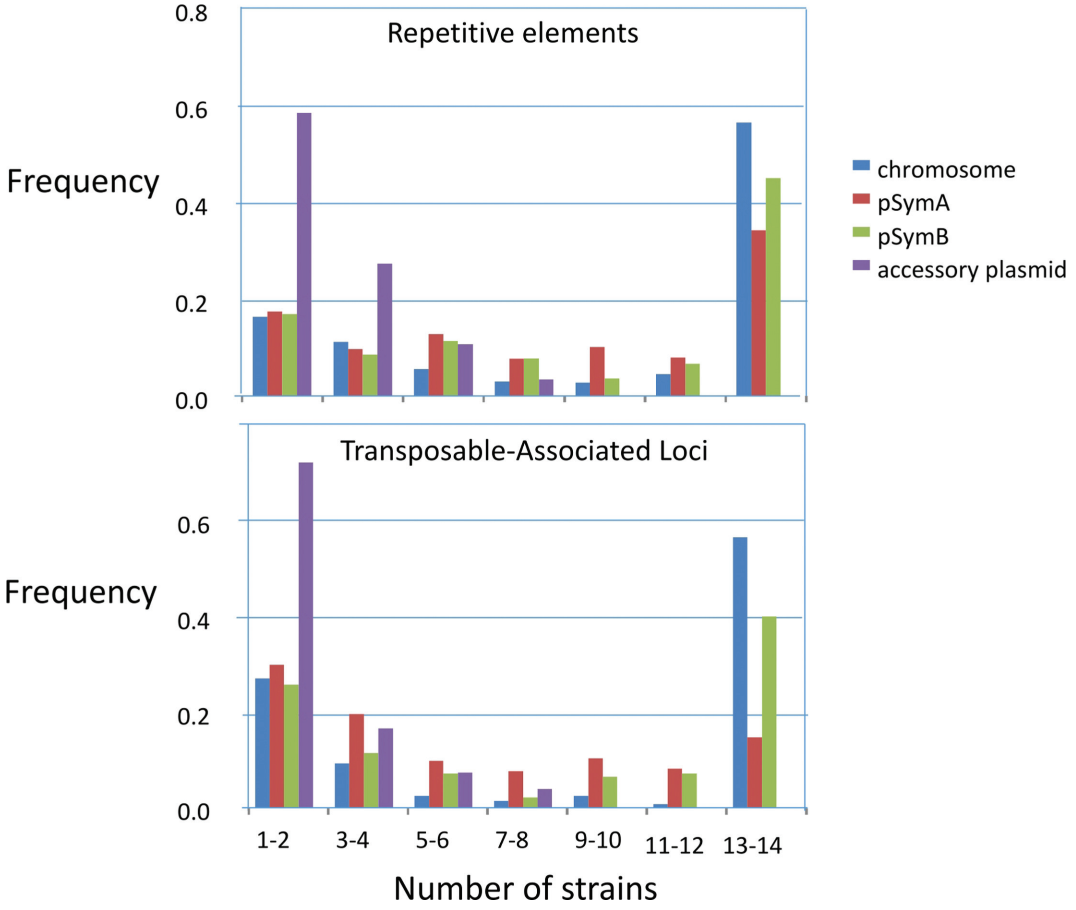

Fig. 3. The frequency of occurrence of repetitive and transposable elements found in different numbers of strains. The majority of REs and TALs that are present on accessory plasmids are found on only a single strain, whereas a majority or plurality of REs and TALs found on the chromosome or pSymB are found in all strains. Because of large translocations between plasmids, two strains (USDA1021 and M162) were excluded from analyses leaving a total of 14 primary replicons and 11 accessory plasmids.

the NCBI NT database using each of the accessory plasmids as a query sequence. Many of the accessory plasmids had BLAST alignments to plasmids found in other rhizobial species, and many contained genes similar to those found on pSymA (Table S4). For example, the accessory plasmid from strain T073 contained a 13 kb gene region matching *E. fredii*, an accessory plasmid from *Ensifer sojae* and plasmids A and C in *Ensifer americanum* at ~95 % identity.

Perhaps the most interesting match found was between accessory plasmid B in strain M270, which contained 64 genes (>20 % of the total) that were also found on the Ti-plasmid in *A. tumefaciens* strain C58 (Fig. S6, Table S5). Genes involved in agrocinopine synthases, transport and catabolism were present, but those required for T-DNA transfer [54] were missing. Agrocinopines are a sugar-phosphodiester subclass of opines (amino acid derivatives) that are typically found in tumours induced by *A. tumefaciens* [55]. While agrocinopines were originally thought to only be synthesized in crown gall tumours, a wide variety of bacteria are capable of utilizing them [56], and *E. meliloti* strain M270 probably gained the ability to synthesize and catabolize these opines via horizontal gene transfer. Additionally, the genes involved in conjugation were also present. These genes were also found on other accessory plasmids, which may indicate a transmission advantage for accessory plasmids with a conjugation gene cluster.

Replicons have distinct evolutionary lineages

A Mantel test was used to determine if the nucleotide content of the replicons (chromosome, pSymA and pSymB) in a specific strain diverged together or separately. The Mantel test revealed no statistically significant correlations in pairwise divergence among strains for any pairs of genes on replicons (all P>0.3, Table S6, Fig. S7). This shows that two strains might have a similar chromosome but distinct pSymA or pSymB replicons. The results in Fig. S8 show that the phylogenetic trees of single-copy core genes for each replicon were distinct. Additionally, the Microscope Gene Phyloprofile tool was used to examine the proportion of genes found in each strain as compared to the reference strain Rm2011 (Table S7). The lack of correlation in the rate of divergences can be seen by comparing the order of

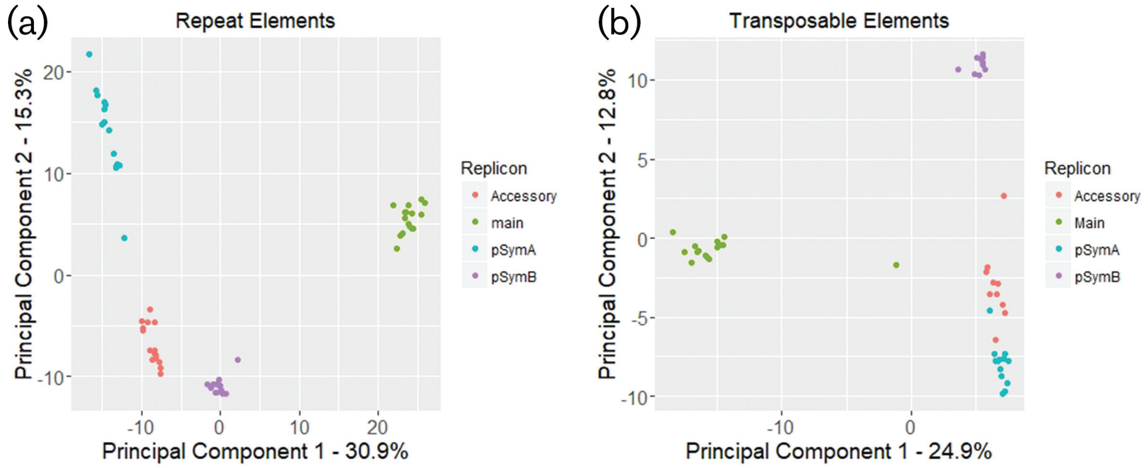

Fig. 4. PCA plot of the number of RE (a) and TAL (b) family members on each replicon for each strain. Replicons, rather than strains, primarily cluster together.

strains. For example, strain T073 had the fourth most similar chromosome, but was only the 14th closest to pSymA when compared to Rm 2011. In contrast, strain Rm41 had the 12th most similar chromosome and was the fourth closest to pSymA.

DISCUSSION

Characterizing genomic diversity is an important step in identifying genes responsible for naturally occurring variation, as well as gaining insight into past adaptation and evolutionary processes. The *E. meliloti* genome consists of three large replicons found in every strain, as well as smaller accessory plasmids that were present in only some strains. Previous analyses found that the replicons differ in their evolutionary histories, due to the strength of purifying selection to which they are subjected, the extent of horizontal gene transfer and the proportion of core versus accessory genes [5, 10]. Here, we show that the three primary replicons also do not often exchange genes with each other, leading to replicons with distinctive G+C, gene, RE and TAL contents.

Perhaps most strikingly, we saw evidence for only 40 non-transposable-element core genes having moved between replicons – fewer than 1 % of the 4600 core genes, with the exception of a single large translocation event involving the movement of a ~300 kb region from the chromosome to pSymB. The nearly complete lack of core gene movement between replicons is puzzling given that there is experimental evidence for frequent genome rearrangements in *Ensifer* under laboratory conditions [57], and some closely related rhizobia have shown genome instability and plasticity. Additionally, essential genes from an *Ensifer* ancestor are thought to have moved from the chromosome to pSymB [58].

Table 2. Accessory plasmid genes shared with main *Ensifer* replicons

Strain and accessory plasmid names	Total genes	Chromosome-exclusive	pSymA-exclusive	pSymB-exclusive	Multiple replicons
AK83 – plasmid NC_015592.1	107	4	16	1	2
AK83 – plasmid NC_015597.1	391	10	128	4	62
GR4 – plasmid NC_019846.1	213	2	90	1	17
GR4 – plasmid NC_019847.1	374	0	132	2	82
HM006 – accessory plasmid A	248	5	63	2	28
Kh35c – accessory plasmid A	215	3	88	1	24
M162 – accessory plasmid A	497	1	359	1	37
M270 – accessory plasmid A	431	9	83	3	88
M270 – accessory plasmid B	304	11	17	1	46
M270 – accessory plasmid C	181	4	46	2	22
Rm41 – accessory plasmid A	277	3	10	2	12
T073 – accessory plasmid A	197	6	50	0	10
USDA1157 – accessory plasmid A	353	1	216	2	38
USDA1021 – accessory plasmid A	379	13	152	1	87

TALs and REs, both of which are able to contribute to gene movement through the translocation of genes between cells and mediate horizontal gene transfer, are abundant on each replicon and often found on multiple replicons [15, 59–62]. We identified a relatively large numbers of REs and TALs compared to many bacterial genomes. The main chromosome had ~18 % repeated sequence in our population. In another study prokaryotic genomes with 20 %+ repeated sequence had the highest repeat coverage from a sample of 720 genomes [63]. Our study was conducted using with different metrics than those used by Treangen *et al.* [63], and probably identified more repeated sequences to identify more diverged repeated elements.

The presence of TALs and REs on multiple replicons indicates that movement between replicons is possible, as do the two large translocation events we detected – a ~300 kb region that moved from pSymA to an accessory plasmid, and a 325 kb region that moved from the chromosome to pSymB. Although the G+C content consistently differed among replicons, a phenomenon also found for the multiple chromosomes in *Burkholderia cenocepacia* [64], the magnitude of the difference (<3 %) is not expected to act as an appreciable barrier to gene exchange. The TALs, unlike the REs, did cluster more closely between pSymA and the accessory plasmid (Fig. 1, Table 1), suggesting that TALs may be involved in inter-replicon gene transfer between pSymA and accessory plasmids.

Understanding inter-replicon gene movement is also important to understand the evolution of *Ensifer* strains as foreign DNA inside the cell will either be eliminated, persist autonomously as a plasmid, or become co-integrated into an existing plasmid through inter-replicon gene transfer [65]. Although TALs can clearly have important roles in gene movement, only seven of 40 possible inter-replicon translocation events we detected had TALs in the regions flanking the translocated regions. In contrast, a high frequency of rearrangements has been found in *Rhizobium etli* [66, 67], and all of the genes showing evidence of inter-replicon movement were found in multiple copies, in at least some strains. This suggests that gene duplication may be involved in inter-replicon gene transfer between replicons in *Ensifer*. Another possibility is that gaining a functionally redundant gene on a different replicon allows for the loss of the original gene without loss of critical function.

Despite finding that <1 % of core genes are found on different replicons in different strains, we detected evidence for extensive gene movement within replicons, and within the chromosome and two megaplasmids. We also found that movement of genes between pSymA and the accessory plasmids was great, with >40 % of accessory plasmid genes also found on pSymA. Given that small plasmids can play central roles in inter-strain gene transfer through conjugation [62], the high rate of gene sharing between pSymA and accessory plasmids suggests that these accessory plasmids might be an important mechanism by which genes are moved between strains, and potentially other bacterial species. This is particularly important from a symbiotic and host-range perspective, given that many of the genes that are essential for establishing a functional symbiosis are found on pSymA [68].

Accessory plasmids in *E. meliloti* have been shown to cause host incompatibility or increase nodulation competitiveness [9]. Indeed, the type IV secretion system, which has been shown to have a variety of effects on nodulation [69, 70], is found on an accessory plasmid or pSymA [69]. Because accessory plasmids can exchange gene content with pSymA, this may allow for the rapid gain or loss of symbiosis-related genes. This phenomenon may lead to some of the symbiotic instability noted for these and other fast-growing rhizobia, where symbiosis genes are plasmid-borne.

While genes on the accessory plasmid were also found as part of the pSymA pan-genome, this was not the case for all genes and some were found in bacterial species other than rhizobia, presumably the result of horizontal gene transfer. Most striking was that one of the *E. meliloti* accessory plasmids had 64 *A. tumefaciens*-like genes. This indicates that rhizobia can probably obtain genes from *Agrobacterium* and other soil microbiota in their free-living, saprophytic, soil phase of existence. *Agrobacterium* and *Ensifer* are closely related and represent different genera within the family *Rhizobiaceae* [71]. *Agrobacterium* Ti-plasmids can be maintained and expressed by *E. meliloti*, although this rhizobial transconjugant is still unable to form tumours on plants [72] although *Rhizobium trifolii* can induce tumours with the addition of a Ti-plasmid [73]. Although the *Ensifer* M270 accessory plasmid B did not contain the tumour-inducing genes, it did contain opine metabolism genes that are found in *A. tumefaciens* Ti-plasmids [74]. This may give this bacterium a selective advantage for growth in some soils and in association with plant roots.

Conclusion

The complete sequences and analyses of 16 *E. meliloti* genomes offer important insight into the evolution of symbiosis-related loci in this bacterium. Our analyses, done using *de novo* assembled long read sequence data, revealed that the three main replicons have different characteristics with respect to gene content, REs, TALs and G+C content. Ten of the strains harboured accessory plasmids, often with distinct replication proteins, and their gene content was more similar to that of pSymA than to the other replicons. Further studies should investigate this phenomenon, which may give insight into how accessory plasmids form and interact in populations of rhizobia in soils. Intra-replicon gene transfer is associated with REs, but not TALs. The gene transfer events that occurred between the accessory plasmids and pSymA demonstrate one mechanism by which the *Ensifer* symbiosis is constantly evolving.

Funding information

These studies were funded, in part, by grants DBI-1237993 from The National Science Foundation and USDA-HATCH award number MIN-71-030.

Acknowledgements

We thank Claudine Médigue, David Vallenet and Zoe Rouy of Genoscope for help with annotations and Jon Badalamenti, University of Minnesota, for help with genome assembly. Sequence analysis was performed using the resources of the University of Minnesota Supercomputing Institute.

Conflicts of interest

The authors declare that there are no conflicts of interest.

Ethical statement

All experiments were performed in accordance with relevant guidelines and regulations.

Data bibliography

DNA sequences have been deposited to NCBI under bioproject number PRJNA388336 (2017). Sequences used in this study were also obtained from:

Galibert F, Finan TM, Long SR, Puhler A, Abola P *et al*. The composite genome of the legume symbiont *Sinorhizobium meliloti*. DOI: 10.1126/science.1060966 (2001).

Galardini M, Pini F, Bazzicalupo M, Biondi EG, Mengoni A. NCBI Bioproject PRJNA41993, PRJNA42477 (2013).

Sallet E, Roux B, Sauviac L, Jardinaud MF, Carrère S *et al*. NCBI Bioproject PRJNA187276 (2013).

Martínez-Abarca F, Martínez-Rodríguez L, López-Contreras JA, Jiménez-Zurdo JI, Toro N. NCBI Bioproject PRJNA175860 (2012).

Schneiker-Bekel S, Wibberg D, Bekel T, Blom J, Linke B *et al*. NCBI Bioproject PRJNA41117 (2011).

References

1. **Galibert F, Finan TM, Long SR, Puhler A, Abola P *et al.*** The composite genome of the legume symbiont *Sinorhizobium meliloti*. *Science* 2001;293:668–672 http://science.sciencemag.org/content/293/5530/668. (accessed 15 June 2017).
2. **Harrison PW, Lower RP, Kim NK, Young JP.** Introducing the bacterial 'chromid': not a chromosome, not a plasmid. *Trends Microbiol* 2010;18:141–148.
3. **Finan TM, Weidner S, Wong K, Buhrmester J, Chain P *et al.*** The complete sequence of the 1,683-kb pSymB megaplasmid from the N2-fixing endosymbiont *Sinorhizobium meliloti*. *Proc Natl Acad Sci USA* 2001;98:9889–9894.
4. **Dicenzo GC, MacLean AM, Milunovic B, Golding GB, Finan TM.** Examination of prokaryotic multipartite genome evolution through experimental genome reduction. *PLoS Genet* 2014;10:e1004742.
5. **Galardini M, Pini F, Bazzicalupo M, Biondi EG, Mengoni A.** Replicon-dependent bacterial genome evolution: the case of *Sinorhizobium meliloti*. *Genome Biol Evol* 2013;5:542–558.
6. **Dicenzo GC, Checcucci A, Bazzicalupo M, Mengoni A, Viti C *et al.*** Metabolic modelling reveals the specialization of secondary replicons for niche adaptation in *Sinorhizobium meliloti*. *Nat Commun* 2016;7:12219.
7. **Kuhn S, Stiens M, Pühler A, Schlüter A.** Prevalence of pSmeSM11a-like plasmids in indigenous *Sinorhizobium meliloti* strains isolated in the course of a field release experiment with genetically modified *S. meliloti* strains. *FEMS Microbiol Ecol* 2008; 63:118–131.
8. **Mazur A, Koper P.** Rhizobial plasmids — replication, structure and biological role. *Open Life Sci* 2012;7:571–586.
9. **Crook MB, Lindsay DP, Biggs MB, Bentley JS, Price JC *et al.*** Rhizobial plasmids that cause impaired symbiotic nitrogen fixation and enhanced host invasion. *Mol Plant Microbe Interact* 2012;25: 1026–1033.
10. **Epstein B, Branca A, Mudge J, Bharti AK, Briskine R *et al.*** Population genomics of the facultatively mutualistic bacteria *Sinorhizobium meliloti* and *S. medicae*. *PLoS Genet* 2012;8:e1002868.
11. **Epstein B, Sadowsky MJ, Tiffin P.** Selection on horizontally transferred and duplicated genes in *Sinorhizobium* (*Ensifer*), the root-nodule symbionts of medicago. *Genome Biol Evol* 2014;6: 1199–1209.
12. **Herring CD, Palsson BØ.** An evaluation of comparative genome sequencing (CGS) by comparing two previously-sequenced bacterial genomes. *BMC Genomics* 2007;8:274.
13. **Galardini M, Mengoni A, Brilli M, Pini F, Fioravanti A *et al.*** Exploring the symbiotic pangenome of the nitrogen-fixing bacterium *Sinorhizobium meliloti*. *BMC Genomics* 2011;12:235.
14. **Rhoads A, Au KF.** PacBio sequencing and its applications. *Genomics Proteomics Bioinformatics* 2015;13:278–289.
15. **Frost LS, Leplae R, Summers AO, Toussaint A.** Mobile genetic elements: the agents of open source evolution. *Nat Rev Microbiol* 2005;3:722–732.
16. **Guo X, Flores M, Mavingui P, Fuentes SI, Hernández G *et al.*** Natural genomic design in *Sinorhizobium meliloti*: novel genomic architectures. *Genome Res* 2003;13:1810–1817.
17. **Stiens M, Schneiker S, Pühler A, Schlüter A.** Sequence analysis of the 181-kb accessory plasmid pSmeSM11b, isolated from a dominant *Sinorhizobium meliloti* strain identified during a long-term field release experiment. *FEMS Microbiol Lett* 2007;271:297–309.
18. **Sugawara M, Epstein B, Badgley BD, Unno T, Xu L *et al.*** Comparative genomics of the core and accessory genomes of 48 *Sinorhizobium* strains comprising five genospecies. *Genome Biol* 2013;14: R17.
19. **Beringer JE.** R factor transfer in *Rhizobium leguminosarum*. *J Gen Microbiol* 1974;84:188–198.
20. **Chin CS, Alexander DH, Marks P, Klammer AA, Drake J *et al.*** Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat Methods* 2013;10:563–569.
21. **Krumsiek J, Arnold R, Rattei T.** Gepard: a rapid and sensitive tool for creating dotplots on genome scale. *Bioinformatics* 2007;23: 1026–1028.
22. **Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A *et al.*** Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* 2014;9:e112963.
23. **Li H, Durbin R.** Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* 2010;26:589–595.
24. **Li H, Handsaker B, Wysoker A, Fennell T, Ruan J *et al.*** The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009;25:2078–2079.
25. **Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW *et al.*** Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 2010;11:119.
26. **Sallet E, Roux B, Sauviac L, Jardinaud MF, Carrère S *et al.*** Next-generation annotation of prokaryotic genomes with EuGene-P: application to *Sinorhizobium meliloti* 2011. *DNA Res* 2013;20:339–354.
27. **Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M *et al.*** Versatile and open software for comparing large genomes. *Genome Biol* 2004;5:R12.
28. **Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J *et al.*** BLAST+: architecture and applications. *BMC Bioinformatics* 2009; 10:421.
29. **Gibbons TR, Mount SM, Cooper ED, Delwiche CF.** Evaluation of BLAST-based edge-weighting metrics used for homology inference with the Markov Clustering algorithm. *BMC Bioinformatics* 2015;16:218.
30. **Guhlin J, Silverstein KAT, Zhou P, Tiffin P, Young ND.** ODG: Omics database generator – a tool for generating, querying, and analyzing multi-omics comparative databases to facilitate biological understanding. *BMC Bioinformatics* 2017;18:367.
31. **Vallenet D, Calteau A, Cruveiller S, Gachet M, Lajus A *et al.*** MicroScope in 2017: an expanding and evolving integrated resource for community expertise of microbial genomes. *Nucleic Acids Res* 2017;45:D517–D528.

32. **Miele V, Penel S, Duret L.** Ultra-fast sequence clustering from similarity networks with SiLiX. *BMC Bioinformatics* 2011;12:116.

33. **Huerta-Cepas J, Szklarczyk D, Forslund K, Cook H, Heller D *et al.*** eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res* 2016;44:D286–D293.

34. **Finn RD, Clements J, Eddy SR.** HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res* 2011;39:W29–W37.

35. **Enright AJ, van Dongen S, Ouzounis CA.** An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* 2002; 30:1575–1584.

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

37. **Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T.** trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 2009;25:1972–1973.

38. Principal component analysis. *Chemom Intell Lab Syst* 1987;2:37–52.

39. **Paradis E, Claude J, Strimmer K.** APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* 2004;20:289–290.

40. **Tamura K, Nei M.** Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol Biol Evol* 1993;10:512–526.

41. **Mantel N.** The detection of disease clustering and a generalized regression approach. *Cancer Res* 1967;27:209–220.

42. **Dereeper A, Audic S, Claverie JM, Blanc G.** BLAST-EXPLORER helps you building datasets for phylogenetic analysis. *BMC Evol Biol* 2010;10:8.

43. **Dereeper A, Guignon V, Blanc G, Audic S, Buffet S *et al.*** Phylogeny.fr: robust phylogenetic analysis for the non-specialist. *Nucleic Acids Res* 2008;36:W465–W469.

44. **Edgar RC.** MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 2004;32:1792–1797.

45. **Castresana J.** Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* 2000; 17:540–552.

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

47. **Anisimova M, Gascuel O, Sullivan J.** Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. *Syst Biol* 2006;55:539–552.

48. **Chevenet F, Brun C, Bañuls AL, Jacq B, Christen R.** TreeDyn: towards dynamic graphics and annotations for analyses of trees. *BMC Bioinformatics* 2006;7:439.

49. **Martínez-Abarca F, Martínez-Rodríguez L, López-Contreras JA, Jiménez-Zurdo JI, Toro N.** Complete genome sequence of the alfalfa symbiont *Sinorhizobium/Ensifer meliloti* Strain GR4. *Genome Announc* 2013;1:e00174-12-e00174-12:e00174-12.

50. **Schneiker-Bekel S, Wibberg D, Bekel T, Blom J, Linke B *et al.*** The complete genome sequence of the dominant *Sinorhizobium meliloti* field isolate SM11 extends the *S. meliloti* pan-genome. *J Biotechnol* 2011;155:20–33.

51. **Cevallos MA, Cervantes-Rivera R, Gutiérrez-Ríos RM.** The repABC plasmid family. *Plasmid* 2008;60:19–37.

52. **van DSM.** Graph clustering by flow simulation. 2000 https://dspace.library.uu.nl/handle/1874/848. (accessed 13 November 2017).

53. **Manen D, Upegui-Gonzalez LC, Caro L.** Monomers and dimers of the RepA protein in plasmid pSC101 replication: domains in RepA. *Proc Natl Acad Sci USA* 1992;89:8923–8927.

54. **Hiei Y, Ohta S, Komari T, Kumashiro T.** Efficient transformation of rice (*Oryza sativa* L.) mediated by *Agrobacterium* and sequence analysis of the boundaries of the T-DNA. *Plant J* 1994;6:271–282.

55. **Ellis JG, Murphy PJ.** Four new opines from crown gall tumours – Their detection and properties. *MGG Mol Gen Genet* 1981;181: 36–43.

56. **Moore LW, Chilton WS, Canfield ML.** Diversity of opines and opine-catabolizing bacteria isolated from naturally occurring crown gall tumors. *Appl Environ Microbiol* 1997;63:201–207.

57. **Mavingui P, Flores M, Guo X, Dávila G, Perret X *et al.*** Dynamics of genome architecture in *Rhizobium* sp. strain NGR234. *J Bacteriol* 2002;184:171–176.

58. **Dicenzo G, Milunovic B, Cheng J, Finan TM.** The tRNAarg gene and engA are essential genes on the 1.7-Mb pSymB megaplasmid of *Sinorhizobium meliloti* and were translocated together from the chromosome in an ancestral strain. *J Bacteriol* 2013;195:202–212.

59. **Wei M, Zhang JJ, Liu H, Wang SJ, Fu H *et al.*** A transposable class I composite transposon carrying mph (methyl parathion hydrolase) from *Pseudomona*s sp. strain WBC-3. *FEMS Microbiol Lett* 2009;292:85–91.

60. **Mijnendonckx K, Provoost A, Monsieurs P, Leys N, Mergeay M *et al.*** Insertion sequence elements in *Cupriavidus metallidurans* CH34: distribution and role in adaptation. *Plasmid* 2011;65:193–203.

61. **Thomas CM, Nielsen KM.** Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nat Rev Microbiol* 2005;3:711–721.

62. **Darmon E, Leach DRF.** Bacterial genome instability. *Microbiol Mol Biol Rev* 2014;78:1–39.

63. **Treangen TJ, Abraham AL, Touchon M, Rocha EP.** Genesis, effects and fates of repeats in prokaryotic genomes. *FEMS Microbiol Rev* 2009;33:539–571.

64. **Dillon MM, Sung W, Lynch M, Cooper VS.** The rate and molecular spectrum of spontaneous mutations in the GC-Rich multichromosome genome of *Burkholderia cenocepacia*. *Genetics* 2015;200: 935–946.

65. **Daubin V, Szöllősi GJ.** Horizontal gene transfer and the history of life. *Cold Spring Harb Perspect Biol* 2016;8:a018036.

66. **Romero D, Martínez-Salazar J, Girard L, Brom S, Dávilla G *et al.*** Discrete amplifiable regions (amplicons) in the symbiotic plasmid of *Rhizobium etli* CFN42. *J Bacteriol* 1995;177:973–980.

67. **Romero D, Brom S, Martínez-Salazar J, Girard ML, Palacios R *et al.*** Amplification and deletion of a nod-nif region in the symbiotic plasmid of *Rhizobium phaseoli*. *J Bacteriol* 1991;173:2435–2441.

68. **Barnett MJ, Kahn ML.** Psyma of *Sinorhizobium meliloti*: nitrogen fixation and more. *Genomes and Genomics of Nitrogen-fixing Organisms*. Berlin/Heidelberg: Springer-Verlag. pp. 113–132.

69. **Nelson MS, Chun CL, Sadowsky MJ.** Type IV effector proteins involved in the *Medicago – Sinorhizobium* symbiosis. *Mol Plant-Microbe Interact* 2017;30:28–34.

70. **Hubber A, Vergunst AC, Sullivan JT, Hooykaas PJ, Ronson CW.** Symbiotic phenotypes and translocated effector proteins of the *Mesorhizobium loti* strain R7A VirB/D4 type IV secretion system. *Mol Microbiol* 2004;54:561–574.

71. **Farrand SK, van Berkum PB, Oger P.** *Agrobacterium* is a definable genus of the family *Rhizobiaceae*. *Int J Syst Evol Microbiol* 2003;53: 1681–1687.

72. **van Veen RJM, den Dulk-Ras H, Schilperoort RA, Hooykaas PJJ.** Ti plasmid containing Rhizobium meliloti are non-tumorigenic on plants, despite proper virulence gene induction and T-strand formation. *Arch Microbiol* 1989;153:85–89.

73. **Hooykaas PJJ, Klapwijk PM, Nuti MP, Schilperoort RA, Rorsch A.** Transfer of the *Agrobacterium tumefaciens* TI plasmid to avirulent agrobacteria and to Rhizobium *ex planta*. *J Gen Microbiol* 1977;98: 477–484.

74. **Paulus F, Otten L.** Functional and mutated agrocinopine synthase genes on octopine T-DNAs. *Mol Plant Microbe Interact* 1993;6:393–402.