

Kim Wong · Turlough M. Finan · G. Brian Golding

## Dinucleotide compositional analysis of *Sinorhizobium meliloti* using the genome signature: distinguishing chromosomes and plasmids

Received: 2 April 2002 / Accepted: 7 June 2002 / Published online: 1 August 2002  
© Springer-Verlag 2002

**Abstract** The symbiotic  $N_2$ -fixing  $\alpha$ -proteobacterium *Sinorhizobium meliloti* has three replicons: a circular chromosome (3.7 Mb) and two smaller replicons, pSymA (1.4 Mb) and pSymB (1.7 Mb). Sequence analysis has revealed that an essential tRNA<sup>Arg</sup><sub>CCG</sub> gene is carried on pSymB, which brings into question whether pSymB should be considered a chromosome or a plasmid. Based on the criterion that essential genes define a chromosome, several species have been shown to have multiple chromosomes. Many of these species are part of the  $\alpha$  subdivision of the Proteobacteria family. Here, additional justification is presented for designating the pSymB replicon as a chromosome. It is shown that chromosomes within a species share a more similar dinucleotide composition, or genome signature, than plasmids do with the host chromosome(s). Dinucleotide signatures were determined for each of the *S. meliloti* replicons, and, consistent with the suggestion that pSymB is a chromosome, it is shown that the pSymB signature more closely resembles that of the *S. meliloti* chromosome, while the pSymA signature is typical of other  $\alpha$ -proteobacterial plasmids.

**Keywords** Genome signature · Dinucleotide composition

### Introduction

It is well recognized that the nucleotide composition of a genome is non-random; elements contributing to this heterogeneity include distinct regions high in G + C or A + T (e.g. isochores; Bernardi et al. 1985), dispersed and tandem repeated sequences, and transposable elements. Coding regions also have different compositions to non-coding regions (Aota and Ikemura 1986; Muto

and Osawa 1987), and even strand composition may be biased (Wu and Maeder 1987). Thus, local compositional variations make it difficult to generalize about a whole genome based on analyses of small regions of DNA.

Genomes of organisms representing all domains of life have been sequenced, allowing in-depth compositional and comparative analysis of these genomes. Examination of the frequencies of short oligonucleotides (di-, tri- and tetranucleotides) in prokaryotic, eukaryotic, and mitochondrial DNA sequences has revealed consistent patterns of oligonucleotide biases, some of which are common to all groups. However, when the relative frequencies of all dinucleotides are considered together, the pattern of dinucleotide bias is unique to each species (Nussinov 1984b; Burge et al. 1992; Karlin and Ladunga 1994; Karlin et al. 1994, 1997; Karlin and Mrázek 1997; Campbell et al. 1999). Unlike G + C content, dinucleotide biases tend to be consistent throughout a genome, in both coding and non-coding DNA (Karlin and Mrázek 1996), giving a genome-wide perspective of the patterns of nucleotide composition within a genome.

The preference or avoidance of specific dinucleotides was first quantified by Bird (1980) who observed a CG dinucleotide (or CpG) suppression in vertebrate sequences. The frequency of the CG dinucleotide is up to fivefold lower than the expected frequency based on C and G mononucleotide frequencies. It has been suggested that the high mutability of methylated cytosine to thymine due to deamination is responsible for the CG under-representation in these organisms (Bird 1980). Evidence for this is given by the existence of “CpG islands,” G + C-rich DNA sequences of variable length which are abundant in unmethylated CG dinucleotides (Bird 1986). More recent studies establish that the CG suppression observed in vertebrates is also prevalent in fungi, plants, protists, and some bacteria (Cardon et al. 1994). Considering that mitochondria and bacteria lack the appropriate DNA methylases, the CG suppression observed in these organisms cannot be explained by the methylation/deamination/mutation hypothesis. Additionally, this hypothesis cannot account for other dinucle-

K. Wong · T.M. Finan · G.B. Golding (✉)  
Department of Biology, McMaster University,  
1280 Main Street West, Hamilton, Ontario, L8S 4K1, Canada  
e-mail: golding@mcmaster.ca  
Tel.: +1-905-5259140, Fax: +1-905-5226066

otide biases, such as the TA dinucleotide suppression. Instead, it has been suggested that dinucleotide biases reflect the avoidance of unfavorable base-step conformations and stacking energies (Nussinov 1984a, b). Since dinucleotide biases are a genome-wide property, it has also been suggested that the mutational biases of the modification, replication and repair machinery play a role in the generation and maintenance of species-specific dinucleotide biases (Karlin and Ladunga 1994; Karlin et al. 1997).

Karlin and his colleagues used dinucleotide relative abundance, the observed frequency of a given dinucleotide relative to the expected frequency based on base composition, to compare genomes (Burge et al. 1992; Karlin et al. 1994, 1997; Karlin and Burge 1995). The set of 16 dinucleotide relative abundance values is referred to as the “dinucleotide relative abundance profile” of an organism. The difference between two profiles is the dinucleotide relative abundance distance, or  $\delta^*$ -distance. Because dinucleotide relative abundance profiles are unique to each organism, and within-species  $\delta^*$ -distances between disjoint 50-kb regions of a replicon are more similar than between-species  $\delta^*$ -distances, the dinucleotide relative abundance profile has been termed the “genome signature”. In addition, organisms that are closely related, as determined by 16S rDNA analysis, generally have more similar genome signatures than more distantly related organisms. Analysis of plasmids and chromosomes has shown that  $\delta^*$ -distances tend to be small (but not necessarily the smallest) between plasmids and natural host chromosomes (Campbell et al. 1999). Thus, plasmids generally tend to track host chromosomal signatures.

In this paper, dinucleotide relative abundance profiles and  $\delta^*$ -distances have been used to characterize the genome of *Sinorhizobium meliloti*. *S. meliloti* is a  $N_2$ -fixing  $\alpha$ -proteobacterium that can form an endosymbiotic relationship with leguminous plants. The genome consists of three replicons: megaplasmids pSymA and pSymB, and one chromosome. The pSymA megaplasmid (1.4 Mb) has long been known to carry symbiotic genes essential for symbiotic nitrogen fixation and root nodulation (Banfalvi et al. 1981; Rosenberg et al. 1981). The pSymB megaplasmid (1.7 Mb) also carries genes essential for the establishment of a successful endosymbiotic relationship with host legumes (Finan et al. 1986; Hynes et al. 1986).

The complete genome of *S. meliloti* has recently been sequenced and annotated (Barnett et al. 2001; Capela et al. 2001; Finan et al. 2001; Galibert et al. 2001). Although pSymB and pSymA both play roles in symbiosis, the difference in the G + C contents of the replicons imply that their evolutionary histories are different; the G + C content of pSymA (60.4%) is lower than the chromosome (62.7%) and pSymB (62.4%). Because of this difference, it has been suggested that pSymA had been acquired much later in evolution than pSymB (Galibert et al. 2001). However, the similarity in G + C content alone does not necessarily reflect the degree of related-

ness between sequences. Because dinucleotide frequencies reflect restrictions in DNA conformation and mutational biases of DNA modification, replication, and repair enzymes, the application of genome signatures and  $\delta^*$ -distances in this paper give a more precise picture of the compositional differences and similarities between the replicons than base composition alone. The results presented here demonstrate that the pSymB genome signature is chromosome-like and, in this respect, that the pSymB replicon is atypical of other  $\alpha$ -proteobacterial plasmids. Taken together with previously observed chromosome-like features, genome signatures support the argument that pSymB should be considered a chromosome rather than a plasmid.

## Materials and methods

### Sequence data

Sequences were downloaded from the National Center for Biotechnology Information (NCBI) website at <http://www.ncbi.nlm.nih.gov> as of August, 2001. Organisms were selected based on their close relationship to *S. meliloti* or similar soil habitat. Complete genome sequences from the following organisms were used in the analysis: *Mesorhizobium loti* (chromosome, plasmids pMLa and pMLb), *Agrobacterium tumefaciens* C58 (circular chromosome, linear chromosome, plasmids pAT and pTi; Cereon Genomics), *Bacillus subtilis*, *Pseudomonas aeruginosa*, *Escherichia coli* K12, *Caulobacter crescentus*, *Haemophilus influenzae*, *Deinococcus radiodurans* (chromosomes I and II, and plasmids MP1 and CP1), *Helicobacter pylori*, *Mycobacterium tuberculosis*, *Synechocystis* sp. PCC 6803, *Rickettsia prowazekii*, *Vibrio cholerae* (chromosomes I and II), *Thermotoga maritima*, and *Halobacterium* sp. NRC-1 (chromosome and plasmids pNRC100 and pNRC200). The *Brucella melitensis* genome sequence became available in December 2001. Complete plasmid sequences from the following organisms were used in analysis: *Rhizobium* sp. NGR234 (pNGR234a), *Agrobacterium rhizogenes* (pRi1724), *Lactococcus lactis* (pMRC01), *Yersinia pestis* (pMT1), and *Sphingomonas aromaticivorans* (pNL1).

In addition, preliminary sequence data of *Rhodobacter sphaeroides* was downloaded from the DOE Joint Genome Institute at <http://www.jgi.doe.gov>. Six contigs greater than 100 kb from the *R. sphaeroides* genome were available (February 2001), totaling 807,975 bp. The *R. sphaeroides* genome is composed of two chromosomes. However, because sequencing and assembly of contigs are incomplete, for dinucleotide analysis, the sequences were joined to form one continuous sequence. Complete *S. meliloti* sequences for the chromosome, pSymA, and pSymB were downloaded from the *S. meliloti* sequencing consortium website at: <http://sequence.toulouse.inra.fr/rhime/Complete/doc/Complete.html>.

### Dinucleotide relative abundance values and $\delta^*$ -distance

Dinucleotide extremes were determined by calculating the observed frequency divided by the expected frequency of each dinucleotide. The set of 16 dinucleotide relative frequencies,  $\{\rho^*_{XY}\}$ , for a particular sequence has been termed the “dinucleotide relative abundance profile”, where

$$\rho^*_{XY} = f_{XY}^* / f_X^* f_Y^*$$

for all dinucleotides XY where  $f_X^*$  is the frequency of nucleotide X and  $f_Y^*$  is the frequency of nucleotide Y (Burge et al. 1992). To control for strand differences, the frequencies are computed from the sequence concatenated with its inverted complementary

sequence. Statistical analysis from previous studies (Karlin et al. 1997) has determined the values of dinucleotide relative abundance values that represent statistically significant extremes for double-stranded 50-kb sequences, and they are applied in this work. Over-representation of a dinucleotide is indicated according to the following scheme:  $1.23 \leq \rho^* < 1.30$  (marginally high),  $1.30 \leq \rho^* < 1.50$  (very high), and  $\rho^* \geq 1.50$  (extremely high); under-representation of a dinucleotide is indicated by:  $0.70 < \rho^* \leq 0.78$  (marginally low),  $0.50 < \rho^* \leq 0.70$  (very low), and  $\rho^* \leq 0.50$  (extremely low). These values represent the extremes which would occur in a random, double-stranded 50-kb nucleotide sequence with the probabilities  $P \leq 10^{-3}$ ,  $P \leq 10^{-6}$  and  $P \leq 10^{-9}$  for the marginally high/low, very high/low, and extremely high/low categories, respectively. Values  $0.78 \leq \rho^* < 1.23$  are considered within the “normal” range.

The difference between two profiles was calculated by the following formula:

$$\delta^*(f, g) = (1/16) \sum |\rho_{XY}^*(f) - \rho_{XY}^*(g)| * 1000$$

where  $\delta^*$  is the dinucleotide relative abundance distance,  $X$  and  $Y$  are nucleotides A, T, C and G,  $f$  and  $g$  denote the two sequences, and the sum extends over all nucleotides. This is the average absolute dinucleotide relative abundance difference, referred to as the  $\delta^*$ -distance. Empirical, qualitative rankings of the  $\delta^*$ -distances for 50-kb sequences were modified from Karlin et al. (1999), and an example of the relatedness reflected by the  $\delta^*$ -distance is given in parentheses: closely similar,  $\delta^* \leq 55$  (*E. coli* vs *S. typhimurium*); moderately similar,  $55 < \delta^* \leq 85$  (*E. coli* vs *H. influenzae*); weakly similar,  $85 < \delta^* \leq 115$  (*Sulfolobus* sp. vs *M. jannaschii*); distantly similar,  $115 < \delta^* \leq 145$  (human vs *S. cerevisiae*); distant,  $145 < \delta^* \leq 185$  (*E. coli* vs *H. pylori*); very distant,  $\delta^* > 185$  (human vs *E. coli*).

**Table 1** Dinucleotide relative abundance profiles of  $\alpha$ -proteobacterial replicons. *Italics* indicate chromosomal sequences. *Agrobacterium tumefaciens I* and *A. tumefaciens II* refer to the circular and linear chromosomes, respectively. The plasmid sequences of  $\alpha$ -proteobacteria are as follows: *Rhizobium* sp. NGR234 (*pNGR234a*), *M. loti* (*pMLa*, *pMLb*), *A. tumefaciens C58* (*pTi*, *pAT*), *A. rhizogenes* (*pRi1724*), and *S. aromaticivorans* (*pNL1*). Significant over-representation of a dinucleotide is indicated by the following scheme:  $1.23 \leq \rho^* < 1.30$  (marginally high, green box),  $1.30 \leq \rho^* < 1.50$  (very high, blue box), and  $\rho^* \geq 1.50$  (extremely high, black box); under-representation of a dinucleotide is indicated by:  $0.70 < \rho^* \leq 0.78$  (marginally low, yellow box),  $0.50 < \rho^* \leq 0.70$  (very low, magenta box), and  $\rho^* \leq 0.50$  (extremely low, red box)

Replicon	CG	GC	TA	AT	CC/ GG	TT/ AA	TG/ CA	AG/ CT	AC/ GT	GA/ TC	G+C (%)
<i>S. meliloti</i>	1.29	1.15	0.47	1.39	0.82	1.18	0.93	0.89	0.77	1.28	62.7
pSymB	1.27	1.15	0.48	1.38	0.82	1.16	0.94	0.90	0.78	1.27	62.4
pSymA	1.23	1.15	0.52	1.28	0.84	1.15	1.00	0.91	0.81	1.22	60.4
pNGR234a	1.20	1.18	0.55	1.22	0.84	1.17	1.03	0.91	0.81	1.17	58.5
pMLa	1.20	1.17	0.52	1.27	0.85	1.17	1.04	0.89	0.81	1.17	59.3
pMLb	1.20	1.18	0.50	1.25	0.84	1.20	1.03	0.90	0.80	1.17	59.9
pTi	1.22	1.14	0.56	1.23	0.86	1.17	1.00	0.89	0.80	1.20	56.7
pAT	1.21	1.15	0.53	1.24	0.86	1.18	1.04	0.87	0.81	1.17	57.3
pRi1724	1.23	1.15	0.52	1.22	0.85	1.19	1.00	0.90	0.79	1.21	57.3
pNL1	1.19	1.16	0.47	1.36	0.85	1.16	1.05	0.88	0.81	1.18	62.2
<i>M. loti</i>	1.23	1.21	0.44	1.41	0.81	1.17	1.05	0.87	0.79	1.18	62.8
<i>A. tumefaciens I</i>	1.24	1.21	0.47	1.37	0.86	1.26	1.04	0.82	0.75	1.14	59.4
<i>A. tumefaciens II</i>	1.22	1.20	0.48	1.39	0.87	1.24	1.05	0.82	0.76	1.14	59.3
<i>B. melitensis I</i>	1.20	1.30	0.52	1.36	0.88	1.30	1.08	0.81	0.70	1.06	57.2
<i>B. melitensis II</i>	1.18	1.31	0.53	1.39	0.89	1.29	1.11	0.80	0.70	1.03	57.3
<i>C. crescentus</i>	1.16	1.11	0.45	1.29	0.85	1.09	1.01	0.96	0.86	1.22	67.2
<i>R. sphaeroides</i>	1.17	1.12	0.38	1.57	0.84	0.97	0.99	0.99	0.76	1.31	68.6
<i>R. prowazekii</i>	0.77	1.53	0.98	0.98	1.03	1.05	1.02	1.06	0.86	0.91	29.0

Unless otherwise indicated, whole genome signatures were calculated from complete genome sequences and  $\delta^*$ -distances for within- and between-species comparisons were calculated from non-overlapping 50-kb regions spanning the sequence of the replicon. Two-sample *t*-tests were performed to compare mean  $\delta^*$ -distances of 50-kb regions.

## Results

The *S. meliloti* pSymB replicon is atypical of other  $\alpha$ -proteobacterial plasmids

The dinucleotide relative abundance profile, or signature, was determined for each replicon in *S. meliloti* and completely sequenced chromosomes and plasmids from other  $\alpha$ -proteobacteria (Table 1). Dinucleotide abundances of a sequence are calculated relative to expected values based on the actual nucleotide content of that sequence; therefore, the dinucleotide abundances of two sequences with different G + C contents can be compared. The plasmids have a distinctive pattern of dinucleotide extremes, when compared to those in the chromosomes. The chromosomal sequences, including the *S. meliloti* chromosome, tend to have an over-representation of CG, a common feature of halobacterial and proteobacterial chromosomes (Karlin et al. 1997). The relative abundances of TA and AT are higher in the chro-

**Table 2** Comparison of mean dinucleotide relative abundance distances ( $\delta^*$ -distances) by *Sinorhizobium meliloti* replicons. Chromosomal sequences are indicated in *italics*. *Agrobacterium tumefaciens I* and *A. tumefaciens II* refer to the circular and linear chromosomes, respectively. Plasmid sequences are from the following organisms: *A. rhizogenes* (pRi1724), *S. aromaticovorans* (pNLI), *M. loti* (pMLa, pMLb), *Rhizobium* sp. NGR234 (pNGR234a), *A. tumefaciens* (pTi, pAT), and *Y. pestis* (pMT1). *S. meliloti* replicons are indicated in **bold**. The seven  $\alpha$ -proteobacterial plasmids have smaller  $\delta^*$ -distances to pSymA than pSymB does to pSymA

Chromosome		pSymB		pSymA	
Sequence	$\delta^*$ -distance	Sequence	$\delta^*$ -distance	Sequence	$\delta^*$ -distance
<b>pSymB</b>	29.7	<b><i>S. meliloti</i></b>	29.7	pMLa	28.2
<b>pSymA</b>	49.4	<b>pSymA</b>	42.7	pNGR234a	30.5
pNLI	53.8	pNLI	47.3	pNLI	31.1
pTi	56.7	pTi	51.0	pAT	33.8
<i>M. loti</i>	57.9	pMLa	53.3	pTi	34.4
pRi1724	58.5	pRi1724	54.7	pMLb	34.6
pMLb	59.8	<i>M. loti</i>	55.4	pRi1724	39.8
pMLa	62.7	pNGR234a	57.1	<i>C. crescentus</i>	42.7
pNGR234a	63.6	pMLb	58.2	<b>pSymB</b>	42.7
pAT	65.6	pAT	59.6	<i>M. loti</i>	48.9
<i>C. crescentus</i>	72.4	<i>C. crescentus</i>	64.5	<b><i>S. meliloti</i></b>	49.4
<i>A. tumefaciens I</i>	72.4	<i>A. tumefaciens I</i>	72.2	<i>A. tumefaciens II</i>	65.4
<i>A. tumefaciens II</i>	73.1	<i>A. tumefaciens II</i>	72.3	<i>A. tumefaciens I</i>	66.3
<i>R. sphaeroides</i>	87.3	<i>R. sphaeroides</i>	83.8	<i>P. aeruginosa</i>	79.0
<i>B. melitensis I</i>	111.6	<i>P. aeruginosa</i>	110.8	<i>R. sphaeroides</i>	91.1
<i>B. melitensis II</i>	119.6	<i>B. melitensis I</i>	111.0	pMT1	97.0
<i>P. aeruginosa</i>	120.4	<i>B. melitensis II</i>	119.1	<i>B. melitensis I</i>	100.9
pMT1	139.4	pMT1	129.7	<i>B. melitensis II</i>	110.2

mosomes than the plasmids, and biases are also observed for AC/GT, GA/TC, TT/AA, and GC in the chromosomes. The pSymB replicon clearly tracks the signature of the *S. meliloti* chromosome, as the relative abundance of each dinucleotide is the same for both replicons. Unlike pSymB, the pSymA replicon has a typical plasmid-like signature, marked by “very low” relative abundance of TA, “marginally high” relative abundance of AT, and the absence of a bias in AC/GT or GA/TC dinucleotides.

To quantify the differences in genome signatures among the three *S. meliloti* replicons, mean  $\delta^*$ -distances were calculated from pairwise comparisons of 50-kb regions with other plasmids and chromosomes found in closely related species or species sharing a similar habitat (Table 2). In accordance with the previous observation that plasmids tend to track host chromosome signatures (Campbell et al. 1999), both pSymB and pSymA have  $\delta^*$ -distances which are “closely similar” to the *S. meliloti* chromosome (29.7 and 49.4, respectively). The smallest  $\delta^*$ -distance to the *S. meliloti* chromosome is pSymB, and vice versa. However, the  $\delta^*$ -distances for pSymA contrast those of pSymB; the  $\delta^*$ -distances between pSymA and the chromosome (49.4) and pSymA and pSymB (42.7) are “closely similar”. However, all seven of the plasmids tested have smaller  $\delta^*$ -distances to pSymA than does pSymB or the chromosome. Thus, the pSymB dinucleotide relative abundance profile is atypical of other  $\alpha$ -proteobacterial plasmids, and more similar to that of the *S. meliloti* chromosome.

Comparison of  $\delta^*$ -distances between chromosomes within a species suggest the pSymB replicon is a chromosome rather than a plasmid

Additional evidence for the chromosome-like nature of the pSymB signature is provided by a  $\delta^*$ -distance comparison of genome signatures among and between plasmids and chromosomes. In previous studies it was shown that (1) plasmid signatures tend to track host chromosomal sequences (Campbell et al. 1999), and (2) sequences within the same replicon have smaller  $\delta^*$ -distances than sequences from two different species (Karlin et al. 1994; Karlin and Burge 1995).

In organisms harboring more than one plasmid, comparisons can be made between plasmid-chromosome  $\delta^*$ -distances and plasmid-plasmid  $\delta^*$ -distances (Table 3). These values are listed for various species in Table 3. The standard errors for all mean  $\delta^*$ -distances listed are less than 4.5, with the exception of the pTi within-plasmid  $\delta^*$ -distance which has a standard error of 8.4. The mean  $\delta^*$ -distance between the *M. loti* plasmids (21.4) is significantly smaller ( $P < 0.01$ ) than the mean  $\delta^*$ -distances between either plasmid and the *M. loti* chromosome (40.8 and 43.3 for pMLa and pMLb, respectively). The same relationship is also observed for the chromosome and plasmids of the archaeobacterium *Halobacterium* sp. NRC-1 (Table 3). In contrast, the mean  $\delta^*$ -distance between the *S. meliloti* chromosome and pSymB (29.7) is significantly lower ( $P < 0.01, t = -43.4$ ) than that between

**Table 3** Mean  $\delta^*$ -distances between plasmids and chromosomes. *Bold numbers* indicate mean  $\delta^*$ -distances between 50-kb regions in plasmid and natural host chromosome(s). *Within-chromosome* and *Within-plasmid* refer to comparison of 50-kb regions within the same chromosome and plasmid, respectively. *Between-chromosome* and *Between-plasmid* refers to comparisons between two different chromosomes or plasmids, respectively. All standard er-

rors of mean  $\delta^*$ -distances are less than 4.5 with the exception of the pTi within-plasmid  $\delta^*$ -distance (SE =8.4). A within-plasmid  $\delta^*$ -distance was not determined for CP1, since the plasmid is too short. For between-chromosome  $\delta^*$ -distances, an *asterisk* indicates values significantly different (at the level of 5%) from the *S. meliloti* chromosome-pSymB mean  $\delta^*$ -distance of 29.7. See text for *t*-test results for comparison of means

Chromosome	Plasmid										Within-chromosome	Between-chromosome
	pSymA	pSymB	pMLa	pMLb	pAT	pTi	MP1	CP1	pNCR100	pNCR200		
<i>S. meliloti</i>	<b>49.4</b>	<b>29.7</b>	59.8	62.7	65.6	56.7	162.4	183.1	146.8	147.5	26.6	
<i>M. loti</i>	48.9	55.4	<b>40.8</b>	<b>43.3</b>	43.9	55.8	137.9	157.0	182.7	179.0	31.1	
<i>A. tumefaciens I</i>	66.3	72.2	52.5	48.6	<b>51.9</b>	<b>68.2</b>	138.4	167.0	208.3	202.2	23.7	
<i>A. tumefaciens II</i>	65.4	72.3	50.4	47.4	<b>49.1</b>	<b>66.9</b>	133.8	160.8	208.7	202.8	27.2	27.0*
<i>D. radiodurans I</i>	122.1	151.2	107.7	100.8	106.4	120.9	<b>30.1</b>	<b>81.3</b>	190.1	193.9	22.0	
<i>D. radiodurans II</i>	122.7	150.9	108.9	102.9	108.5	121.3	<b>33.2</b>	<b>81.6</b>	190.0	194.4	37.3	30.2
<i>V. cholerae I</i>	127.7	158.0	110.9	108.3	108.9	121.4	65.9	107.8	199.5	204.7	28.9	
<i>V. cholerae II</i>	135.1	165.1	117.4	114.3	114.1	127.3	70.1	112.2	210.4	214.6	30.4	30.8*
<i>B. melitensis I</i>	101.0	111.0	83.1	79.6	80.8	99.5	136.5	174.6	243.0	237.1	27.2	
<i>B. melitensis II</i>	110.2	119.1	91.8	88.2	89.4	108.1	130.9	174.4	253.3	247.3	27.2	30.1
<i>Halobacterium sp.</i>	176.0	171.6	190.1	198.0	187.6	179.3	231.9	214.9	<b>75.9</b>	<b>66.9</b>	32.9	
Within-plasmid	25.1	30.3	17.3	21.9	22.6	44.7	17.9	n/a	15.0	35.3		
Between-plasmid		42.7		21.4		37.8		66.5		25.2		

pSymA and the *S. meliloti* chromosome (49.4) and between pSymA and pSymB ( $P < 0.01, t = -23.0$ ). Consistent with previous observations (Campbell et al. 1999; Karlin et al. 1994; Karlin and Burge 1995), the within-chromosome  $\delta^*$ -distances are significantly smaller than plasmid-chromosome  $\delta^*$ -distances for all three species.

A similar relationship is observed with organisms consisting of two chromosomes as well as plasmids. In these cases, comparisons can be made between chromosome-chromosome  $\delta^*$ -distances, plasmid-chromosome  $\delta^*$ -distances, and plasmid-plasmid  $\delta^*$ -distances. *Agrobacterium tumefaciens* has one circular chromosome, one linear chromosome, and two plasmids. The  $\delta^*$ -distance of 37.8 between the two plasmids is significantly lower ( $P < 0.01$ ) than the  $\delta^*$ -distances between either plasmid to either chromosome (Table 3). The *Deinococcus radiodurans* genome comprises two chromosomes (2,649 kb and 412 kb), one megaplasmid, and one small plasmid. The signatures of the two chromosomes are very similar, as reflected by the small (30.2)  $\delta^*$ -distance (Table 3). The megaplasmid (MP1) is “closely similar” to both of the chromosomes and the plasmid-chromosome  $\delta^*$ -distances are not significantly different from the between-chromosome  $\delta^*$ -distance at the 5% level. This is analogous to the relationship amongst pSymB, pSymA, and the *S. meliloti* chromosome; however, MP1 does not carry any essential genes and only comprises 5.4% of the whole genome and was not designated a chromosome.

The mean  $\delta^*$ -distance between pSymB and the *S. meliloti* chromosome (29.7) is not significantly different ( $P > 0.05$ ) from the between-chromosome  $\delta^*$ -distance of *D. radiodurans* (30.2) or the  $\alpha$ -proteobacterium *B. melitensis* (30.1), and significantly lower than the *V. cholerae* between-chromosome  $\delta^*$ -distance ( $P = 0.03, t = -2.2$ ). Thus, the difference in relative dinucleotide frequencies between pSymB and the *S. meliloti* chromosome is generally at the same level or lower than that found between two chromosomes in the same organism. Only the *A. tumefaciens* between-chromosome  $\delta^*$ -distance of 27.0 was smaller ( $P < 0.01, t = 7.1$ ). Within-chromosome  $\delta^*$ -distance values are comparable to between-chromosome values, ranging from 22.0 to 37.3, while within-plasmid values range from 15.0 to 44.7 (Table 3).

These data indicate that the application of genome signatures is a good indicator of the degree of relatedness of replicons not only between species, but within-species as well. Among replicons within the same organism, the similarity in genome signatures probably reflects long-term replication by the same polymerase and repair enzymes.

## Discussion

Plasmids are considered “facultative” genetic elements, not essential for cell viability but they often carry genes that allow for adaptation to different environments, life-

styles, or stress conditions (Joset and Guespin-Michel 1994). The pSymB replicon, which carries many genes involved in small molecule transport in addition to those genes involved in the endosymbiotic lifestyle, appears to be specialized for adaptation to different environments. However, most of these genes are non-essential to cell viability, as a large portion of the pSymB replicon can be deleted without loss of viability (Charles and Finan 1991), nor does pSymB carry any genes coding for ribosomal RNA (*rrn*; Finan et al. 2001). Control of the replication initiation and inheritance by plasmid-encoded RepABC proteins is characteristic of  $\alpha$ -proteobacterial plasmids, especially within the *Agrobacterium* and *Rhizobium* genera (Tabata et al. 1989; Palmer et al. 2000). However, there are exceptions; *repABC* is also found on the linear chromosome of *A. tumefaciens* C58, leading to the speculation that this chromosome was derived from a plasmid (Goodner et al. 2001). A *repABC* operon and a replication origin (*oriV*) on pSymB have been shown to allow autonomous plasmid replication and stable inheritance (Chain et al. 2000). There is, however, also evidence that the pSymB replicon has another replication origin that is chromosome-like in nature, AT-rich, containing potential DnaA-binding sites, and a 13-mer sequence similar to the *C. crescentus oriC* (Margolin and Long 1993).

Considering the presence of *repABC* and that the overall role of pSymB appears to be adaptation rather than cell viability, pSymB appears plasmid-like. However, when other characteristics of pSymB are taken into consideration, the task of designating this replicon a chromosome or a plasmid becomes more difficult.

The suggestion that a bacterial genome contains more than one chromosome is not unprecedented. Traditionally, within a genome of multiple replicons, the designation of a replicon as a chromosome has been based on the presence of genes essential to cell viability, such as 16S rRNA genes, or essential housekeeping genes such as *dnaK*. The presence of multiple chromosomes in prokaryotic genomes was first reported for *Rhodobacter sphaeroides* 2.4.1 based on the finding that two rRNA cistrons and the gene coding for glyceraldehyde-3-phosphate, two “chromosomal” loci, are found on a second replicon (Suwanto and Kaplan 1989). Since then, several other bacteria, including *Deinococcus radiodurans* (White et al. 1999), *Vibrio* (Heidelberg et al. 2000; Yamaichi et al. 1999), *Brucella* (Michaux et al. 1993; Cheng and Lessie 1994), and *Burkholderia* (Rodley et al. 1995) species, have been reported to harbor multiple circular chromosomes. Aside from the presence of essential “chromosomal” genes, criteria that are used to differentiate chromosomes from plasmids are significant replicon size, non-self transmissibility, and presence in all strains (Trucksis et al. 1998). It has also been suggested that replication machinery and evolutionary history should also be taken into account (Ng et al. 1998).

The *S. meliloti* pSymB replicon meets each of these criteria, while pSymA has characteristics typical of plasmids. pSymB comprises 25% of the whole genome,

and is present in all *S. meliloti* strains. Unlike pSymA, sequence analysis of pSymB did not reveal characteristics of a self-transmissible plasmid (e.g. it lacks an *oriT* sequence and conjugative transfer genes). The only exception is a single copy of the *traA* gene, also found on pSymA. The codon usage of pSymB is very similar to that of the chromosome, while pSymA codon usage is notably different from the chromosome or pSymB (Galibert et al. 2001). The complete sequence and annotation of pSymB has revealed the presence of a tRNA<sub>CCG</sub><sup>Arg</sup> gene and two loci involved in cell division: one of two genomic copies of *ftsK*, and the single-copy *minCDE* genes (Finan et al. 2001). Two *ftsK* genes and the *minCDE* genes are also found on the *M. loti* chromosome, outside of the transmissible “symbiotic island” (Kaneko et al. 2000). The presence of these genes on pSymB suggests that this replicon plays a central role in control of cell division and chromosome partitioning. No essential genes were identified on pSymA. Additionally, while pSymB has resisted all attempts, the pSymA megaplasmid has been successfully cured from *S. meliloti* strain 2011 (Oresnik et al. 2000).

Here, the application of the dinucleotide relative abundance analysis has shown that, beyond the presence of essential genes, pSymB shares genome-wide characteristics with the chromosome of *S. meliloti* that are atypical of other  $\alpha$ -proteobacterial plasmids, including pSymA. When compared to other  $\delta^*$ -distances observed between chromosomes within the same organism, the difference between pSymB and the *S. meliloti* chromosome falls within the same level of similarity; this is indicative of a high degree of relatedness and/or long-term residence of pSymB in the *S. meliloti* genome.

The genome of the  $\gamma$ -proteobacterium *V. cholerae* is comprised of two circular chromosomes (Trucksis et al. 1998). It has been suggested that the *V. cholerae* chromosome II was derived from a megaplasmid captured in an ancestral *Vibrio* which has acquired essential genes (Heidelberg et al. 2000). The pSymB replicon may also have a similar history, since it contains both plasmid-like features (an *oriV*, *repABC* genes, absence of *rrn* genes), chromosomal-like features (a tRNA gene, *min* and *ftsK* genes and large size), and a signature similar to the *S. meliloti* chromosome rather than other  $\alpha$ -proteobacterial plasmids. Over evolutionary time, pSymB may have acquired genes essential to the organism’s viability as well as a similar signature as the chromosome due to long-term residence in *S. meliloti* and exposure to mutational biases of its replication and repair machinery. Taken together, these functional and compositional characteristics indicate that pSymB is not merely an accessory genetic element, but a chromosome-like replicon.

## Conclusion

It has already been recognized that pSymB has chromosome-like features; pSymB comprises a large proportion

of the genome, carries essential genes, is non-self-transmissible, and strains cured of pSymb cannot be produced. In terms of nucleotide composition, not only is the resemblance of pSymb to the *S. meliloti* chromosome reflected in the similarity of their G + C contents, but more precisely by the dinucleotide relative abundances. It was shown here that pSymb dinucleotide extremes parallel those of  $\alpha$ -proteobacterial chromosomes rather than those of other  $\alpha$ -proteobacterial plasmids, as pSymbA does. In addition, the level of variability in dinucleotide frequencies between chromosomes in the same organism corresponds to the amount of variability between pSymb and the *S. meliloti* chromosome. Collectively, these characteristics of pSymb justify the designation of this replicon as a second chromosome in *S. meliloti*.

**Acknowledgements** This work was supported by Natural Sciences and Engineering Research Council (Canada) Research, Strategic, and Genomics grants to T.M.F. and G.B.G.

## References

- Aota S, Ikemura T (1986) Diversity in G + C content at the third position of codons in vertebrate genes and its cause. *Nucleic Acids Res* 14:6345–6355
- Banfalvi Z, Sakanyan V, Koncz C, Kiss A, Dusha I, Kondorosi A (1981) Location of nodulation and nitrogen fixation genes on a high molecular weight plasmid of *R. meliloti*. *Mol Gen Genet* 184:318–325
- Barnett MJ, Fisher RF, Jones T, Komp C, Abola AP, Barloy-Hubler F, Bowser L, Capela D, Galibert F, Gouzy J, Gurjal M, Hong A, Huizar L, Hyman RW, Kahn D, Kahn ML, Kalman S, Keating DH, Palm C, Peck MC, Surzycki R, Wells DH, Yeh KC, Davis RW, Federspiel NA, Long SR (2001) Nucleotide sequence and predicted functions of the entire *Sinorhizobium meliloti* pSymbA megaplasmid. *Proc Natl Acad Sci USA* 98:9883–9888
- Bernardi G, Olofsson B, Filipski J, Zerial M, Salinas J, Cuny G, Meunier-Rotival M, Rodier F (1985) The mosaic genome of warm-blooded vertebrates. *Science* 228:953–958
- Bird AP (1980) DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Res* 8:1499–1504
- Bird AP (1986) CpG-rich islands and the function of DNA methylation. *Nature* 321:209–213
- Burge C, Campbell AM, Karlin S (1992) Over- and under-representation of short oligonucleotides in DNA sequences. *Proc Natl Acad Sci USA* 89:1358–1362
- Campbell A, Mrázek J, Karlin S (1999) Genome signature comparisons among prokaryote, plasmid, and mitochondrial DNA. *Proc Natl Acad Sci USA* 96:9184–9189
- Capela D, Barloy-Hubler F, Gouzy J, Bothe G, Ampe F, Batut J, Boistard P, Becker A, Boutry M, Cadieu E, Dréano S, Gloux S, Godrie T, Goffeau A, Kahn D, Kiss E, Lelaure V, Masuy D, Pohl T, Portetelle D, Pühler A, Purnelle B, Ramsperger U, Renard C, Thébault P, Vandenbol M, Weidner S, Galibert F (2001) Analysis of the chromosome sequence of the legume symbiont *Sinorhizobium meliloti* strain 1021. *Proc Natl Acad Sci USA* 98:9877–9882
- Cardon LR, Burge C, Clayton DA, Karlin S (1994) Pervasive CpG suppression in animal mitochondrial genomes. *Proc Natl Acad Sci USA* 91:3799–3803
- Chain PS, Hernández-Lucas I, Golding B, Finan TM (2000) oriT-directed cloning of defined large regions from bacterial genomes: identification of the *Sinorhizobium meliloti* pExo megaplasmid replicator region. *J Bacteriol* 182:5486–5494
- Charles TC, Finan TM (1991) Analysis of a 1600-kilobase *Rhizobium meliloti* megaplasmid using defined deletions generated in vivo. *Genetics* 127:5–20
- Cheng HP, Lessie TG (1994) Multiple replicons constituting the genome of *Pseudomonas cepacia* 17616. *J Bacteriol* 176:4034–4042
- Finan TM, Kunkel B, De Vos GF, Signer ER (1986) Second symbiotic megaplasmid in *Rhizobium meliloti* carrying exopolysaccharide and thiamine synthesis genes. *J Bacteriol* 167:66–72
- Finan TM, Weidner S, Wong K, Buhrmester J, Chain P, Vorhölter FJ, Hernandez-Lucas I, Becker A, Cowie A, Gouzy J, Golding B, Pühler A (2001) The complete sequence of the 1,683-kb pSymb megaplasmid from the N<sub>2</sub>-fixing endosymbiont *Sinorhizobium meliloti*. *Proc Natl Acad Sci USA* 98:9889–9894
- Galibert F, Finan TM, Long SR, Pühler A, Abola P, Ampe F, Barloy-Hubler F, Barnett MJ, Becker A, Boistard P, Bothe G, Boutry G, Bowser L, Buhrmester J, Cadieu E, Capela D, Chain P, Cowie A, Davis RW, Dreano S, Federspiel NA, Fisher RF, Gloux S, Godrie T, Goffeau A, Golding B, Gouzy J, Gurjal M, Hernandez-Lucas I, Hong A, Huizar L, Hyman RW, Jones T, Kahn D, Kahn ML, Kalman S, Keating DH, Kiss E, Komp C, Lelaure V, Masuy D, Palm C, Peck MC, Pohl TM, Portetelle D, Purnelle B, Ramsperger U, Surzycki R, Thebault P, Vandenbol M, Vorholter FJ, Weidner S, Wells DH, Wong K, Yeh KC, Batut J (2001) The composite genome of the legume symbiont *Sinorhizobium meliloti*. *Science* 293:668–672
- Goodner B, Hinkle G, Gattung S, Miller N, Blanchard M, Qurollo B, Goldman BS, Cao Y, Askenazi M, Halling C, Mullin L, Houmiel K, Gordon J, Vaudin M, Iartchouk O, Epp A, Liu F, Wollam C, Allinger M, Doughty D, Scott C, Lappas C, Markelz B, Flanagan C, Crowell C, Gurson J, Lomo C, Sear C, Strub G, Cielo C, Slater S (2001) Genome sequence of the plant pathogen and biotechnology agent *Agrobacterium tumefaciens* C58. *Science* 294:2323–2328
- Heidelberg JF, Eisen JA, Nelson WC, Clayton RA, Gwinn ML, Dodson RJ, Haft DL, Hickey EK, Peterson JD, Umayam L, Gill SR, Nelson KE, Read TD, Tettelin H, Richardson D, Ermolaeva MD, Vamathevan J, Bass S, Qin H, Dragoi I, Sellers P, McDonald L, Utterback T, Fleishmann RD, Nierman WC, White O (2000) DNA sequence of both chromosomes of the cholera pathogen *Vibrio cholerae*. *Nature* 406:477–483
- Hynes MF, Simon R, Müller P, Niehaus K, Labes M, Pühler A (1986) The two megaplasmids of *Rhizobium meliloti* are involved in the effective nodulation of alfalfa. *Mol Gen Genet* 202:356–362
- Joset F, Guespin-Michel J (1994) Prokaryotic genetics: genome organization, transfer, and plasticity. Blackwell, Oxford
- Kaneko T, Nakamura Y, Sato S, Asamizu E, Kato T, Sasamoto S, Watanabe A, Idesawa K, Ishikawa A, Kawashima K, Kimura T, Kishida Y, Kiyokawa C, Kohara M, Matsumoto M, Matsuno A, Mochizuki Y, Nakayama S, Nakazaki N, Shimpo S, Sugimoto M, Takeuchi C, Yamada M, Tabata S (2000) Complete genome structure of the nitrogen-fixing symbiotic bacterium *Mesorhizobium loti* (supplement). *DNA Res* 7:381–406
- Karlin S, Burge C (1995) Dinucleotide relative abundance extremes: a genomic signature. *Trends Genet* 11:283–290
- Karlin S, Ladunga I (1994) Comparisons of eukaryotic genomic sequences. *Proc Natl Acad Sci USA* 91:12832–12836
- Karlin S, Mrázek J (1996) What drives codon choices in human genes? *J Mol Biol* 262:459–472
- Karlin S, Mrázek J (1997) Compositional differences within and between eukaryotic genomes. *Proc Natl Acad Sci USA* 94:10227–10232
- Karlin S, Ladunga I, Blaisdell BE (1994) Heterogeneity of genomes: measures and values. *Proc Natl Acad Sci USA* 91:12837–12841
- Karlin S, Mrázek J, Campbell AM (1997) Compositional biases of bacterial genomes and evolutionary implications. *J Bacteriol* 179:3899–3913

- Karlin S, Brocchieri L, Mrazek J, Campbell AM, Spormann AM (1999) A chimeric prokaryotic ancestry of mitochondria and primitive eukaryotes. *Proc Natl Acad Sci USA* 96:9190–9195
- Margolin W, Long SR (1993) Isolation and characterization of a DNA replication origin from the 1,700-kilobase-pair symbiotic megaplasmid pSym-b of *Rhizobium meliloti*. *J Bacteriol* 175:6553–6561
- Michaux S, Paillisson J, Carles-Nurit MJ, Bourg G, Allardet-Servent A, Ramuz M (1993) Presence of two independent chromosomes in the *Brucella melitensis* 16 M genome. *J Bacteriol* 175:701–705
- Muto A, Osawa S (1987) The guanine and cytosine content of genomic DNA and bacterial evolution. *Proc Natl Acad Sci USA* 84:166–169
- Ng WV, Ciufu SA, Smith TM, Bumgarner RE, Baskin D, Faust J, Hall B, Loretz C, Seto J, Slagel J, Hood L, DasSarma S (1998) Snapshot of a large dynamic replicon in a halophilic archaeon: megaplasmid or minichromosome? *Genome Res* 8:1131–1141
- Nussinov R (1984a) Doublet frequencies in evolutionary distinct groups. *Nucleic Acids Res* 12:1749–1763
- Nussinov R (1984b) Strong doublet preferences in nucleotide sequences and DNA geometry. *J Mol Evol* 20:111–119
- OresnikIJ, Liu SL, Yost CK, Hynes MF (2000) Megaplasmid pRme2011a of *Sinorhizobium meliloti* is not required for viability. *J Bacteriol* 182:3582–3586
- Palmer KM, Turner SL, Young JP (2000) Sequence diversity of the plasmid replication gene repC in the Rhizobiaceae. *Plasmid* 44:209–219
- Rodley PD, Romling U, Tummeler B (1995) A physical genome map of the *Burkholderia cepacia* type strain. *Mol Microbiol* 17:57–67
- Rosenberg C, Boistard P, Dénarié J, Casse-Delbart F (1981) Genes controlling early and late functions in symbiosis are located on a megaplasmid in *Rhizobium meliloti*. *Mol Gen Genet* 184:326–333
- Suwanto A, Kaplan S (1989) Physical and genetic mapping of the *Rhodobacter sphaeroides* 2.4.1 genome: presence of two unique circular chromosomes. *J Bacteriol* 171:5850–5859
- Tabata S, Hooykaas PJ, Oka A (1989) Sequence determination and characterization of the replicator region in the tumor-inducing plasmid pTiB6S3. *J Bacteriol* 171:1665–1672
- Trucksis M, Michalski J, Deng YK, Kaper JB (1998) The *Vibrio cholerae* genome contains two unique circular chromosomes. *Proc Natl Acad Sci USA* 95:14464–14469
- White O, Eisen JA, Heidelberg JF, Hickey EK, Peterson JD, Dodson RJ, Haft DH, Gwinn ML, Nelson WC, Richardson DL, Moffat KS, Qin H, Jiang L, Pamphile W, Crosby M, Shen M, Vamathevan JJ, Lam P, McDonald L, Utterback T, Zalewski C, Makarova KS, Aravind L, Daly MJ, Fraser CM, et al (1999) Genome sequence of the radioresistant bacterium *Deinococcus radiodurans* R1. *Science* 286:1571–1577
- Wu CI, Maeda N (1987) Inequality in mutation rates of the two strands of DNA. *Nature* 327:169–170
- Yamaichi Y, Iida T, Park KS, Yamamoto K, Honda T (1999) Physical and genetic map of the genome of *Vibrio parahaemolyticus*: presence of two chromosomes in *Vibrio* species. *Mol Microbiol* 31:1513–1521