JOURNAL OF BACTERIOLOGY, Mar. 2007, p. 1914–1921 Vol. 189, No. 5
0021-9193/07/$08.00+0 doi:10.1128/JB.01498-06

Genome Analyses of Three Strains of *Rhodobacter sphaeroides*: Evidence of Rapid Evolution of Chromosome II▽

M. Choudhary,[1] Xie Zanhua,[2] Y. X. Fu,[2] and S. Kaplan[1]*

Department of Microbiology and Molecular Genetics[1] and Computational Genomic Section, Human Genetics Center,[2] The University of Texas Health Science Center, Houston, Texas 77030

Received 22 September 2006/Accepted 11 December 2006

Three strains of *Rhodobacter sphaeroides* of diverse origin have been under investigation in our laboratory for their genome complexities, including the presence of multiple chromosomes and the distribution of essential genes within their genomes. The genome of *R. sphaeroides* 2.4.1 has been completely sequenced and fully annotated, and now two additional strains (ATCC 17019 and ATCC 17025) of *R. sphaeroides* have been sequenced. Thus, genome comparisons have become a useful approach in determining the evolutionary relationships among different strains of *R. sphaeroides*. In this study, the concatenated chromosomal sequences from the three strains of *R. sphaeroides* were aligned, using Mauve, to examine the extent of shared DNA regions and the degree of relatedness among their chromosome-specific DNA sequences. In addition, the exact intra- and interchromosomal DNA duplications were analyzed using Mummer. Genome analyses employing these two independent approaches revealed that strain ATCC 17025 diverged considerably from the other two strains, 2.4.1 and ATCC 17029, and that the two latter strains are more closely related to one another. Results further demonstrated that chromosome II (CII)-specific DNA sequences of *R. sphaeroides* have rapidly evolved, while CI-specific DNA sequences have remained highly conserved. Aside from the size variation of CII of *R. sphaeroides*, variation in sequence lengths of the CII-shared DNA regions and their high sequence divergence among strains of *R. sphaeroides* suggest the involvement of CII in the evolution of strain-specific genomic rearrangements, perhaps requiring strains to adapt in specialized niches.

Rhodobacter sphaeroides, a purple nonsulfur phototrophic bacterium, belongs to the α-3 subgroup of the *Proteobacteria* (38). A number of strains with common morphological, physiological, and biochemical characteristics have been identified as representatives of this species (36); these strains were originally collected from Delft, Holland, and California from a variety of enrichment cultures. Together with the other α subgroup of the *Proteobacteria* (12, 18), members of *R. sphaeroides* exhibit substantial metabolic versatility (6) and genomic complexity, including the existence of two chromosomes (23, 31, 32).

Twenty-five strains of *R. sphaeroides*, including the three strains examined in the present study, have been previously investigated for their macro-restriction length polymorphisms by using pulsed-field gel electrophoresis. Multiple genetic markers derived from *R. sphaeroides* 2.4.1 were used to examine the genome complexity of the different strains of *R. sphaeroides*. Identification of diagnostic gene loci located on specific macro-restriction fragments belonging to either chromosome I (CI) or CII demonstrated the wide divergence between the two chromosomes among the different strains of *R. sphaeroides*. For example, the number of *rrn* operons varied from two to five among the different strains (23). The genome of *R. sphaeroides* 2.4.1 has been completely sequenced and "fully" annotated. Preliminary analysis of the genome of *R. sphaeroides* 2.4.1 facilitated our understanding of its genome organization (19). DNA duplication analysis revealed an abundance of exact DNA sequence duplications in the genome of *R. sphaeroides* 2.4.1, which further demonstrated that both CI and CII have coexisted in the *R. sphaeroides* genome (3), and this long association may have been established prior to the derivation of *R. sphaeroides* as a species. Optical mapping data (reference 41 and unpublished data) obtained from three strains, 2.4.1, ATCC 17029, and ATCC 17025, indicated that the sizes of CI of these stains were similar while the sizes of CII varied. In addition to that of *R. sphaeroides* 2.4.1, genomic sequences of strains ATCC 17029 and ATCC 17025 are now available, and therefore genome analyses of these strains provide a powerful approach to the examination of the differential evolution of CI and CII.

In this study, two independent approaches, global DNA sequence alignment and exact DNA duplication analysis, were used to identify the extent of DNA sequence conservation among the three *R. sphaeroides* genomes. Chromosomal sequences of the three strains were aligned in order to identify the common backbone regions and the degree of their DNA sequence similarities. Also, we examined the distribution of exact DNA sequence duplications within and between the genome(s) of these three strains. Results of both the global sequence alignment and the exact DNA duplication analysis revealed that the genome of ATCC 17025 is more highly diverged from those of the other two strains (2.4.1 and ATCC 17029), whose genomes are more closely related. Comparison of CI- and CII-specific DNA sequences from these three strains provided evidence that CII has evolved at a higher rate than CI and suggested that the rapid evolution of CII of *R. sphaeroides* is mediated by either acquiring new genetic material through horizontal DNA transfer or rapidly generating

* Corresponding author. Mailing address: Department of Microbiology and Molecular Genetics, University of Texas Medical School at Houston, Houston, Texas 77030. Phone: (713) 500-5502. Fax: (713) 500-5499. E-mail: Samuel.Kaplan@uth.tmc.edu.

▽ Published ahead of print on 15 December 2006.

genetic rearrangements. Thus, the resulting new genetic variants may play differential metabolic roles in allowing the different strains to utilize different niches of diverse nutritional resources.

MATERIALS AND METHODS

The genome of *R. sphaeroides* 2.4.1 contains two chromosomes and five endogenous plasmids and has been completely sequenced and fully annotated (NCBI accession no. NC_007488 and NC_007494). In addition, genomes of the additional two strains (ATCC 17029 and ATCC 17025) of *R. sphaeroides* have been sequenced, and genomic sequences are available at http://genome.ornl.gov. These two strains were chosen for DNA sequencing out of many strains available in our laboratory because the two strains varied in their macro-restriction fragment length polymorphisms and genome sizes as well as numbers of rRNA operons. Also, optical maps of the genomes of these two strains were available. The genomic sequence assemblies of *R. sphaeroides* ATCC 17029 and ATC17025 contain 20 and 88 contigs, respectively. DNA sequences of all contigs were subjected to Blast analysis against the reference genome of *R. sphaeroides* 2.4.1, which was completely assembled. The locations of these contigs were reliably mapped on CI or CII of *R. sphaeroides* 2.4.1. Contigs that did not map to any chromosome were removed from the analysis because those contigs belonged to either plasmids or some unique chromosomal fragments.

Global DNA sequence alignment. Concatenated chromosomal DNA sequences of the three strains of *R. sphaeroides* were aligned using Mauve 1.0 (4). This method utilizes pairwise or multiple alignments of conserved genomic sequences of whole genomes, with modest computational requirements without compromising the alignment quality (35). Local alignments were performed in order to identify multiple maximal unique matches (multi-MUMs), which were subsequently used to calculate a guide for phylogenetic tree constructions. A subset of multi-MUMs were then used as anchors, which were divided into local collinear blocks (LCBs). Using multi-MUMs lowers anchoring sensitivity in conserved repetitive regions, such as rRNA operons and prophages. Each LCB is a homologous DNA region of multi-MUMs, which lacks any sequence rearrangements and which is shared by two or more of the genomes under analysis. The sequence alignment identifies the number of common LCBs by using the length of the total conserved regions and the overall nucleotide identity between chromosomal sequences for each pair of strains. The conserved backbone sequence was extracted from the alignment.

DNA sequences in any given genome are aligned once to each of the other genomes, and therefore Mauve detects only orthologous sequences. The number of matching nucleotides and the number of insertions and deletions within each LCB over the complete aligned regions were extracted from the alignment output. The identity between DNA sequences was calculated as a ratio of the number of matching nucleotides to the number of total nucleotides spanning all aligned regions interrupted with many unaligned DNA strings. For example, a nucleotide identity of 0.95 means that 95 out of the total 100 nucleotides of the matching regions are identical.

DNA duplication analysis. Since the Mauve global alignment identifies only orthologous DNA sequences, MUMmer 3.0 (5) was used to identify the exact DNA sequence duplications within and between the genomes of the three strains of *R. sphaeroides*. This method also includes paralogous duplicated regions. We used 20 nucleotides as the minimum cutoff length of the DNA sequence that perfectly matches elsewhere within or between genomes. The MUMmer output results show the coordinates of each sequence pair of identical matches and the length of each exactly duplicated DNA sequence.

RESULTS

DNA sequence conservation in *Rhodobacter sphaeroides*. Since organisms evolve through several types of genetic rearrangements in their genomes, such as random mutation, DNA duplication (1, 3, 7, 13, 16, 17, 20, 26, 37), gene loss (16, 28, 29), repeated inversions and translocation (30), and integration of new genetic elements through horizontal DNA transfer (11, 24), genome comparison of the strains of *R. sphaeroides* provides a complete depiction of the rates and patterns of each of these evolutionary changes. Thus, the identification of the chromosomal region with the most diverged DNA sequence among genomes of different strains of *R. sphaeroides* would possibly provide the roles of these diverged DNA sequences in strain differentiation.

A global alignment of the concatenated chromosomal DNA sequences of the three strains of *Rhodobacter sphaeroides* is shown in Fig. 1, and the results of the alignment are described in Table 1. One of the important criteria for the sequence alignment is the minimum weight of LCBs, which measures the confidence in distinguishing between a real genome rearrangement and a false match. For example, a minimum weigh of 45 refers to three times the sequence length of 15 nucleotides, which was used during the initial search for the multi-MUMs. Multiple alignments of concatenated chromosomal sequences at a minimum weight of 45 displayed 382 common LCBs comprised of ~3.35 Mb of shared DNA sequence with 76.1% overall nucleotide identity. The total numbers of LCBs identified in CI and CII were 263 and 119, respectively. The CIs of the three strains of *R. sphaeroides* shared ~90% of their DNA sequences, while the CIIs of these strains shared only ~50% of their DNA sequences. In addition, the nucleotide identity between CII-specific sequences was lower (up to 5%) than the nucleotide identity observed between CI-specific sequences of the three strains of *R. sphaeroides*, as shown in Table 1. Approximately 10% of CI- and ~50% of CII-specific DNA sequences of the three stains of *R. sphaeroides* remain unaligned; these are often repetitive, paralogous, or strain-specific genomic regions acquired by horizontal DNA transfer from other organisms. The role of the unaligned DNA sequences with no homology among the three strains will be discussed below.

Pattern of nucleotide insertions and deletions. The distribution of the number and size of nucleotide insertions and deletions is one of the major forces that affect the evolution of genome size (10, 21, 29). If insertions occur more frequently and are longer than deletions, this would allow the genome size to expand. In contrast, more frequent deletions that are longer would contract the genome. Thus, the varied spectrum of nucleotide insertions and deletions among strains of *R. sphaeroides* would provide a force for diversification of genome size, including the sizes of CI and CII.

The numbers of insertion sites in strains ATCC 17029 and ATCC 17025 were calculated as the number of locations where the nucleotides were missing from the reference strain 2.4.1 but the insertion of a nucleotide(s) was present in either of the two strains. Conversely, the number of nucleotide deletions in either of the two strains was calculated as the number of locations where the nucleotides were present in the reference strain 2.4.1 and the nucleotides were missing from either of the two strains. The genomes of the three strains of *R. sphaeroides*, as shown in Table 1, demonstrated a varied number of deletions and insertions. The total amounts of nucleotide deletions and insertions in 2.4.1, ATCC 17029, and ATCC 17025 were 135,867, 145,540, and 248,296 base pairs of DNA, respectively. A similar DNA deletion pattern was observed in CI in all three strains of *R. sphaeroides*; however, the pattern of nucleotide deletions in CII of the three strains differed significantly. The numbers of insertions and deletions and their corresponding nucleotide lengths in the *R. sphaeroides* strains are provided in Table 2. The numbers of total DNA deletions and insertions occurring in strain ATCC 17029 were 990 and 837, respec-

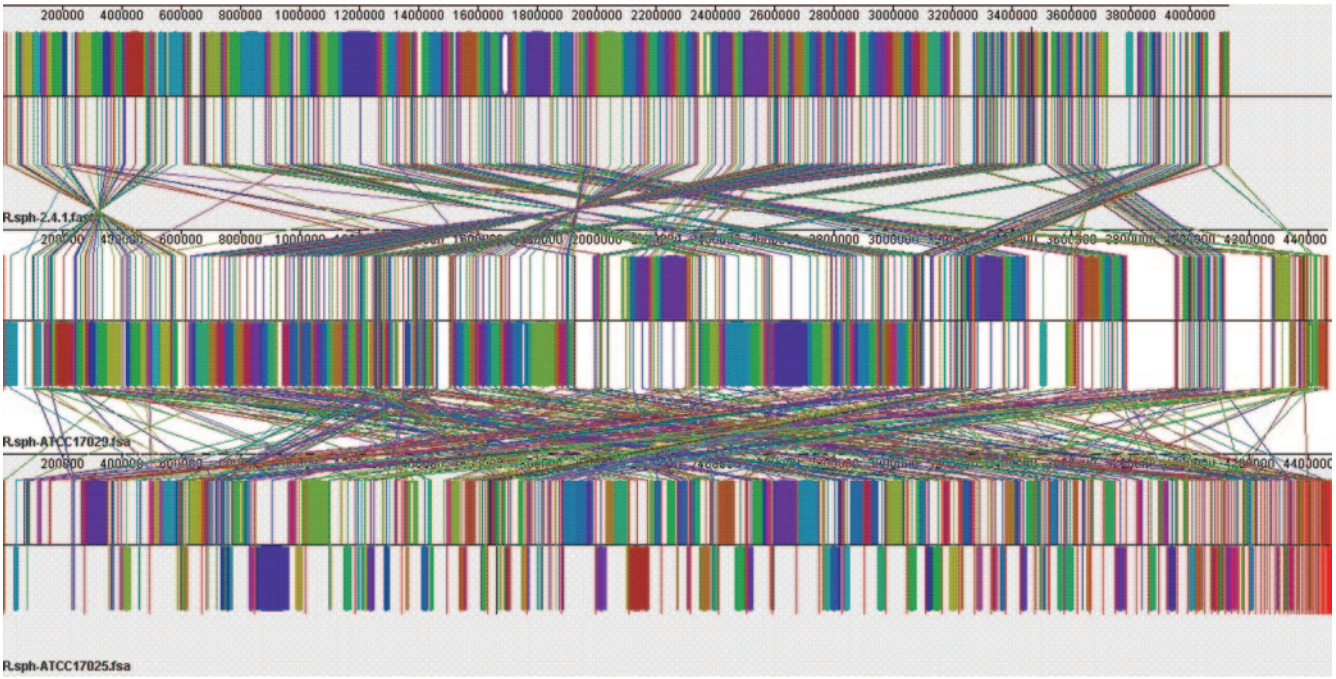

FIG. 1. Mauve representation of the total of 382 LCBs observed between concatenated chromosomal sequences of the three strains of *Rhodobacter sphaeroides*, 2.4.1, ATCC 17029, and ATCC 17025, at a minimum weight of 45. Black vertical bars indicate concatenated chromosomal boundaries. The *R. sphaeroides* 2.4.1 DNA sequence given on the forward strand is the reference against which the sequences of the other two strains were aligned and compared. LCBs placed under the vertical bars represent the reverse complement of the reference DNA sequence. The Mauve display window provides viewers the ability to zoom in on regions of interest and examine the local rearrangements of DNA sequences, where connecting lines between genomes identify the locations of each orthologous LCB in all three genomes. Unmatched regions within an LCB indicate the presence of strain-specific sequence. Each sequential colored block represents homologous backbone DNA sequence without rearrangements. ATCC 17025 has undergone more chromosomal rearrangements than the other two strains.

tively, representing ~62 and ~42 kb of DNA, respectively. In contrast, the numbers of total DNA deletions and insertions in strain ATCC 17025 were 7506 and 5671, respectively, comprising ~255 and ~105 kb of DNA, respectively. The high frequencies of both nucleotide insertions and deletions found in strain ATCC 17025 reflect the older separation of strain ATCC 17025 as opposed to the separation of ATCC 17029 from the reference strain 2.4.1, and therefore the earlier separation of strain ATCC 17025 allowed this strain to accumulate more insertions and deletions.

Overall, the number of deletions was higher than the number of insertions in both strains ATCC 17029 and ATCC 17025. Similarly, the pattern of a higher number of deletions was also found in both CI and CII of ATCC 17029. However, the numbers of deletions and insertions were about the same in CII of strain ATCC 17025, as shown in Table 2, but CI harbored more deletions (>4-fold) than insertions. The relative reduction of nucleotide deletions in CII of strain ATCC 17025 indicates that CII lost some of its original DNA content and its size was later increased by the insertion of new genetic elements by horizontal gene transfer. The majority of the DNA content of nucleotide deletions and insertions of these three strains of *R. sphaeroides* represented longer DNA sequences (>100 nucleotides). To understand the nature of the longer DNA insertions, some of these insertion sequences were analyzed for their percent G+C composition and di- and trinucleotide repeats, which were subsequently compared to the overall *R. sphaeroides* genomic pattern comprising these features. The results indicated a lower percent G+C composition of many DNA insertions than the average percent G+C composition of the total *R. sphaeroides* genomes. Also, the di- and trinucleotide frequencies of the insertion sequence differed from the average di- and trinucleotide repeat patterns of the complete genome. In a number of cases, insertion sequences contained many phage-related protein genes, including those for phage-related site-specific integrase, restriction endonuclease, and *recA* family type ATPase (data not shown).

DNA sequence divergence between CI and CII. Aside from the variation in size of CI and CII, the number of LCBs identified in the two chromosomes is related to the extent of their DNA sequence conservation; while a low number of LCBs reflects the sequence conservation over a longer DNA length, a higher number of LCBs indicates sequence conservation over small segments of DNA. All pairwise comparisons revealed that CI contained fewer LCBs than expected based on its relative size, as shown in Table 1. In contrast, CII displayed more than the expected number of LCBs, as estimated from the varied sizes of CII as determined by optical mapping (Tim Donohue, personal communication) of the three different strains.

As mentioned earlier, the total numbers of contigs in the sequence assemblies of ATCC 17029 and ATCC 17025 are 20 and 88, respectively. The numbers of contigs present in the final genome assemblies of these two strains affect the numbers

of LCBs in their genome comparisons. However, most of the DNA sequences remained in longer contigs in both ATCC 17029 and ATCC 17025, with a majority of smaller contigs from the ATCC 17025 genome assembly appearing as one LCB when aligned to another genome. Thus, the analysis of the genome comparisons and the validity of the results are not compromised. Moreover, when the number of existing contigs was included in the analysis, the numbers of LCBs identified in the genome alignments of strains 2.4.1 and ATCC 17029; 2.4.1 and ATCC 17025; ATCC 17029 and ATCC 17025; and 2.4.1, ATCC 17029, and ATCC 17025 were 110, 354, 372, and 274, respectively. Therefore, the patterns observed before taking the number of contigs into consideration are not different than those seen when these were not considered. This finding suggested that CI of *R. sphaeroides* has been more conserved than CII. Seemingly, at the same minimum weight of 45, the pairwise alignment of the chromosomal sequences of 2.4.1 and ATCC 17029 identified fewer LCBs than were identified in sequence alignments of either of the two pairs of strains 2.4.1 and ATCC 17025 or ATCC 17029 and ATCC 17025.

In addition, the data revealed that the genomes of strains 2.4.1 and ATCC 17029 shared extensive DNA homology (~3.96 Mb), with a very high degree of sequence conservation (95.6% nucleotide identity), whereas the sequence alignments of either of the two other pairs of strains, ATCC 17029 and ATCC 17025 or 2.4.1 and ATCC 17025, revealed less DNA homology (~3.4 Mb) as well as a lower level of sequence conservation (~77% nucleotide identity).

The DNA sequence divergence between each pair of the three strains of *R. sphaeroides* was separately measured for CI and CII, as shown in Table 3. Nucleotide identities were determined over all LCBs extracted from the various pairwise alignments, which were performed at different minimum weights, such as 30, 45, 60, 500, and 5,000. The minimum weight refers to three times the size of the initial DNA sequence length used for searching the multi-MUMs. A high minimum weight identifies genome rearrangements that are more likely to exist at a large scale, while a low minimum weight deals with sensitivity to smaller genome rearrangements. Alignments at low minimum weight (such as 30 and 45) identified higher numbers of LCBs with a greater degree of DNA sequence conservation. In contrast, sequence alignments at high minimum weight (such as 60, 500, and 5,000) revealed lower numbers of LCBs with a lesser degree of DNA sequence conservation. However, all pairwise genome comparisons revealed on average low nucleotide identity between CII-specific orthologous sequences, and therefore a high DNA sequence divergence, as shown in Table 3, and the difference in sequence divergence between CI and CII was highly significant (chi-square test, $P = 10^{-6}$). Thus, the CII sequences of each strain have diverged more than the CI sequences.

The coding capabilities and the average intergenic lengths in CI and CII of the three strains of *R. sphaeroides* are shown in Table 4. While CI of these strains revealed higher coding abilities and shorter average intergenic lengths, CII revealed ~2% lower coding capabilities and relatively longer intergenic lengths. The longer intergenic lengths in CII may possibly allow more intrachromosomal recombination and access to horizontal gene transfer. These two factors will influence the genetic divergence of CII among the strains of *R. sphaeroides*.

TABLE 1. Comparison of genomic alignments among three strains of *R. sphaeroides*

Chromosome	No. of LCBs[a]			Length (nucleotides) of aligned LCBs			Nucleotide identity			Length (bp) of insertions/deletions		
	2.4.1/ATCC 17029	2.4.1/ATCC 17025	ATCC 17029/ATCC 17025	2.4.1/ATCC 17029	2.4.1/ATCC 17025	ATCC 17029/ATCC 17025	2.4.1/ATCC 17029	2.4.1/ATCC 17025	ATCC 17029/ATCC 17025	2.4.1/ATCC 17029/ATCC 17025	2.4.1/ATCC 17029	2.4.1/ATCC 17025	ATCC 17029/ATCC 17025
I	80 (97)	293 (321)	335 (394)	3,067,730	2,835,876	2,610,141	0.957	0.776	0.782	0.768	95,024	109,052	219,179
II	50 (33)	149 (121)	145 (86)	836,697	438,499	680,043	0.950	0.731	0.759	0.714	40,843	36,488	29,117
Total	130	442	480	3,956,220	3,379,790	3,401,341	0.956	0.770	0.777	0.761	135,867	145,540	248,296
				263	119	382							

[a] Numbers in parentheses indicate the expected number of LCBs based on different sizes of CI and CII.

TABLE 2. Numbers of insertions and deletions

Insertion or deletion[a]	No. (total nucleotides)					
	2.4.1/ATCC 17029			2.4.1/ATCC 17025		
	CI	CII	Total	CI	CII	Total
Insertions						
<25	555 (2,140)	207 (812)	762 (2,942)	4164 (18,304)	813 (3,845)	4977 (22,149)
26–50	20 (712)	9 (329)	29 (1,041)	249 (8,784)	57 (2,027)	306 (10,811)
51–100	20 (1,421)	7 (548)	27 (1,969)	178 (12,340)	37 (2,735)	215 (15,075)
101–1000	8 (1,677)	8 (1,162)	16 (2,939)	128 (28,813)	34 (9,414)	162 (38,227)
>1 kb	2 (27,953)	1 (5,091)	3 (33,044)	2 (2,266)	9 (16,887)	11 (19,153)
Total	605 (33,903)	232 (7,942)	837 (41,845)	4721 (70,507)	950 (34,908)	5671 (105,415)
Deletions						
<25	625 (2,393)	220 (960)	845 (3,353)	5110 (27,651)	942 (5,559)	6052 (33,210)
26–50	33 (1,256)	15 (524)	48 (1,780)	547 (19,544)	98 (3,395)	645 (22,939)
51–100	20 (1,493)	22 (1,642)	42 (3,135)	327 (23,240)	63 (4,425)	390 (27,665)
101–1000	32 (8,613)	18 (6,165)	50 (14,778)	340 (80,136)	57 (15,411)	397 (95,547)
>1 kb	5 (31,038)	3 (7,423)	8 (38,461)	20 (71,740)	2 (4,190)	22 (75,930)
Total	715 (44,793)	275 (16,714)	990 (61,507)	6,344 (222,311)	1,162 (32,980)	7,506 (255,291)

[a] Length of insertion or deletion size in nucleotides.

Prevalence of exact DNA duplications. Gene duplication plays a major role in the evolution of biological novelty and biodiversity (7, 9, 26). A comparison of exact DNA sequence duplications within and between the genomes of the three strains of *R. sphaeroides* is shown in Table 5. The genome of *R. sphaeroides* 2.4.1 contains 4,763 DNA duplications, while the numbers of exact intragenomic DNA duplications found in the genomes of the other two strains of *R. sphaeroides* (ATCC 17029 and ATCC 17025) were 3,805 and 3,777, respectively, which are ~8% lower than the number of DNA sequence duplications shown for strain 2.4.1. The genome of *R. sphaeroides* 2.4.1 revealed ~158 kb (~3.4% of its genome) of exactly duplicated DNA sequences. The genome of *R. sphaeroides* ATCC 17029 similarly displayed ~100 kb (~2.8% of its genome) of duplicated DNA regions. However, *R. sphaeroides* ATCC 17025 contained ~205 kb (~5.9% of its genome) of exactly duplicated DNA sequences, which was approximately twice the amount of exactly duplicated DNA sequences found in the other two strains, 2.4.1 and ATCC 17029. Although the majority of DNA sequence duplications in all three strains were small (<100 nucleotides), as shown in Table 5, the genome of *R. sphaeroides* ATCC 17025 possesses a higher frequency of DNA duplications for relatively longer DNA sequences (100 to 1,000 nucleotides). The genome of ATCC 17025 showed twice the number of longer duplications (100 to 1,000 nucleotides) than identified for the genome of either 2.4.1 or ATCC 17029. Frequent gene duplications have also been reported to exist in many bacterial species, including *Enterococcus faecalis* and *Lactobacillus johnsonii*, which represent relatively large and small genomes, respectively (2). Thus, different selective constraints might explain the varied degree of sequence amplification among strains or closely related bacterial species.

TABLE 3. Nucleotide identity among three strains of *R. sphaeroides*

Strains	LCB wt	No. of LCBs	Identity (%)		
			Total	CI	CII
2.4.1/ATCC 17029	30	146	95.66	95.77	95.26
	45	130	95.60	95.75	95.04
	60	118	95.57	95.75	94.88
2.4.1/ATCC 17025	30	603	77.74	78.32	74.21
	45	442	77.01	77.64	73.14
	60	374	76.77	77.53	72.15
ATCC 17029/ATCC 17025	30	656	78.64	79.24	76.32
	45	480	77.77	78.24	75.96
	60	421	77.45	77.81	76.04
2.4.1/ATCC 17029/ATCC 17025	30	521	76.89	77.35	73.83
	45	382	76.15	76.88	71.47
	60	341	76.08	76.80	71.46
	500	219	75.50	76.31	70.04
	5000	106	75.48	76.33	67.95

The intergenomic DNA duplications represent orthologues which are shared through common ancestry. The total number of orthologous DNA sequences between any two of the three strains of *R. sphaeroides* was ~10 times more than the number for the intragenomic DNA duplications found in each of the three strains alone, as shown in Table 5. In addition, the total number of DNA duplications identified between strains 2.4.1 and ATCC 17029 was ~7% higher than the number of DNA

TABLE 4. Percent coding capabilities and average intergenic lengths in CI and CII

Chromosome	% Coding nucleotides			Avg intergenic length (nucleotides)		
	2.4.1	ATCC 17029	ATCC 17025	2.4.1	ATCC 17029	ATCC 17025
I	89.51	88.97	89.72	108	120	106
II	88.08	86.84	87.52	130	151	132
I + II	89.19	88.25	89.38	113	130	110

TABLE 5. Distribution of intra- and intergenomic DNA duplications in *R. sphaeroides*

Duplication type and strain(s)	No. of duplications of length (nucleotides):							Total no. of duplications	Content (bp)
	20–25	26–50	51–100	101–200	201–500	501–1000	>1 kb		
Intragenomic									
2.4.1	3,572	887	204	54	25	18	3	4,763	157,914
ATCC 17029	3,320	403	36	16	19	7	4	3,805	100,438
ATCC 17025	2,611	777	168	71	43	91	16	3,777	204,721
Intergenomic									
2.4.1/ATCC 17029	9,787	12,924	11,164	7,684	3,604	316	21	45,550	3,819,471
2.4.1/ATCC 17025	19,066	16,412	2478	230	20	1	0	38,209	1,153,925
ATCC 17029/ATCC 17025	18,009	16,121	2416	244	23	2	0	36,815	1,116,281

duplications between either the genomes of strains 2.4.1 and ATCC 17025 or those of strains ATCC 17029 and ATCC 17025.

The total content of exact DNA duplications between strains 2.4.1 and ATCC 17029 was 3.82 Mb, which was ~7% higher than the amount of exact DNA duplications found between either strains 2.4.1 and ATCC 17025 or strains ATCC 17029 and ATCC 17025. These results revealed that the majority of the genomic DNA duplications between any two strains were small, but ~50% of the total DNA duplications between 2.4.1 and ATCC 17029 were of longer DNA sequences than the exact DNA duplications identified between either 2.4.1 and ATCC 17025 or ATCC 17029 and ATCC 17025. Thus, the frequency and the amount of total intergenomic DNA duplications demonstrated that the genome of strain ATCC 17025 diverged more from each of the other two strains and possibly separated before the separation of the latter two strains, 2.4.1 and ATCC 17029, which share ~85% of exactly duplicated DNA sequences.

DISCUSSION

Evolutionary relationships among the three strains of *R. sphaeroides*. The degree of alignment between genomes varies depending on the evolutionary distance of the organisms being compared. Multiple alignments of chromosomal sequences demonstrated that *R. sphaeroides* 2.4.1 and ATCC 17029 have more extensive DNA homology (~3.95 Mb of DNA), with a high degree of conservation (95.6% nucleotide identity), than that (~3.4 Mb of DNA) between either of the other two pairs of strains, 2.4.1 and ATCC 17025 or ATCC 17029 and ATCC 17025. Also, the DNA homologies of the latter two pairs of strains revealed an average low level of sequence conservation (~77% nucleotide identity). Similarly, DNA duplication analysis revealed more exact DNA sequence duplications between the genomes of 2.4.1 and ATCC 17029 than between the genomes of either 2.4.1 and ATCC 17025 or ATCC 17029 and ATCC 17025. Thus, two different and independent analyses yielded similar results, which demonstrated a closer phylogenetic relationship between the genomes of *R. sphaeroides* 2.4.1 and ATCC 17029 and that the genome of ATCC 17025 was more diverged from either of the other two strains, namely, 2.4.1 and ATCC 17029. The high degree of sequence conservation among the three strains of *R. sphaeroides* is in agreement with the level of DNA sequence conservation found among genomes of other bacterial species. Recently, a genome comparison of two strains of *Francisella tularensis* showed that these two strains share 97.6% of their genomes and that the nucleotides within these regions are 98.9% identical. However, the major difference between the two strains is the level of genomic rearrangements in gene order (27). Similarly, among nine different representative strains of *Vibrio cholerae*, only 1% difference exists in their gene contents (34), but there was an extensive level of genomic rearrangements. Thus, the high degree of intraspecies DNA sequence conservation suggests that the differences in lifestyles or virulence types are not due to large differences in the genetic contents of the strains.

R. sphaeroides ATCC 17025 harbors ~3 times more nucleotide deletions and insertions than either 2.4.1 or ATCC 17029 (Tables 1 and 2). If the rates of deletion formation were the same in all three strains of *R. sphaeroides*, the cumulative number of deletions would correlate with the relative time of separation of each strain from its common lineage. The higher number of deletions and insertions in ATCC 17025 suggests its earlier separation from the common lineage. In addition to the divergence of the CI-specific sequences of the three strains, the number of deletions and insertions in CI of these strains also correlates with the relative separation times of the three strains. However, CII conserved regions reflected about the same number of deletions in all three strains. Thus, the divergence of CI-specific DNA sequences among the strains is the best indicator of a phylogenetic relationship among different strains of *R. sphaeroides*.

The use of nonfunctional DNA sequences, such as pseudogenes and intergenic sequences, is essential for evaluating the role of spontaneous mutations, including insertions and deletions. The nonfunctional DNA sequences accumulate random mutations, and the occurrence of such mutations would not be affected by natural selection (15). The spectrum of insertions/deletions was first used as a parameter of genome size evolution in the study of mammalian pseudogenes, and it was subsequently shown that DNA loss was estimated to be faster in rodents than in humans (8), possibly resulting in smaller rodent genomes. The pattern of DNA insertions and deletions in strains of *R. sphaeroides* revealed more nucleotide deletions than insertions, which corroborates the earlier finding of a deletional bias as a major force that shapes bacterial genomes (21). Mutational analyses of the genomes of the three strains is currently in progress, and the results of such analyses would further our understanding of the genome size variation in *R. sphaeroides*.

Rapid divergence of CII and evolution of strain-specific genomic rearrangements. Genome analysis using a combination of criteria, such as the analyses of the restriction patterns with AseI- and CeuI-generated DNA fragments and the localization of various genes on these restriction fragments by using strain 2.4.1 as the prototype, demonstrated the similar size of CI (~3.0 Mb) in many *R. sphaeroides* strains, but the size of CII of *R. sphaeroides* varies (23). Also, optical mapping revealed that CII of *R. sphaeroides* 2.4.1, ATCC 17029, and ATCC 17025 consisted of 0.94, 1.23, and 0.91 Mb of DNA, respectively (41; T. J. Donohue, personal communication).

Genome comparison of *Vibrio cholerae* and *Vibrio parahaemolyticus* revealed a similar observation, i.e., that chromosomes I of these two species do not differ greatly in size (3.0 and 3.3 Mb, respectively) but chromosome II is much larger in *V. parahaemolyticus* than in *V. cholerae* (33). Furthermore, the global transcriptional pattern of in vivo-grown cells of *V. cholerae* shows the highest levels of expression for genes located on CI, while bacterial growth in the intestine has the highest levels of expression for genes that are located on CII (39). Thus, the diverse size, genetic content, and the pattern of expression of genes of CII suggest that CII maintains the genetic reservoir required for species adaptation in specialized environments (34).

The rapid divergence of CII of *R. sphaeroides* includes a high degree of nucleotide differences between orthologues, rearranged DNA sequences, duplicated genes, and/or newly acquired genetic elements on CII. These genetic divergences revealed a faster evolution of CII, and that could be attributed to different evolutionary forces. Since the two chromosomes (CI and CII) of *R. sphaeroides* appear to have coexisted, possibly prior to the formation of *R. sphaeroides* as a species (3), the origin of the two-chromosome genome architecture must have occurred before the diversification of *R. sphaeroides* strains. An ancient association of the two chromosomes in *R. sphaeroides* 2.4.1 is further supported by the fact that there is no difference in DNA parameters, such as percent G+C content, di- and trinucleotide frequencies, and codon preferences of CI and CII, in all three strains of *R. sphaeroides* (reference 19 and unpublished observations). However, some of the CII DNA sequences of *R. sphaeroides* might have been recently acquired from other bacteria with similar or different genetic backgrounds. In a number of instances, longer DNA insertions identified in *R. sphaeroides* differ slightly in percent G+C composition (~2% low G+C composition) as well as in nucleotide repeat patterns from the complete chromosome or genome and encode many phage-related functions. The presence of a large number of prophages (lambda like) has also been found in *Escherichia coli* genomes (25) and are suspected to be involved in the diversification of different strains of *E. coli*. These newly acquired genetic variations of CII were selected for strain-specific adaptations, but their DNA had not yet drifted towards the genome average. Therefore, the rapid evolution of CII could be attributed to more recent horizontal DNA transfers from bacteria with similar genetic backgrounds, as these newly acquired insertions do not reflect a drastic difference from their overall genome composition.

The susceptibility of CII for rapid chromosomal changes may be due to the fact that CII has a relatively low coding capacity and relatively long intergenic sequences (Table 4), which together may make CII more prone to accumulating genetic variants. Although the division of the genome into two replicons would be advantageous for rapid DNA replication as observed in *V. parahaemolyticus* (40), the difference in their copy numbers might amplify the level of gene expression in changing environmental conditions. However, the existence of a single chromosome in species closely related to *R. sphaeroides*, such as *Rhodobacter capsulatus* and *Rhodopseudomonas palustris* (14), does not appear to be disadvantageous. Seemingly, a genome analysis of *R. sphaeroides* suggests that the closest relative of *R. sphaeroides* is *Silicibacter pomeroyi*, a member of the marine *Roseobacter* clade (Chris Mackenzie, personal communication), and members of the *Roseobacter* clade are widely distributed over diverse hydrographical regions of the ocean. The genome of *S. pomeroyi* consists of a main chromosome and a megaplasmid, and its genome sequence is fully equipped to take advantage of transient high-nutrient niches within a low-nutrient marine environment (22). Thus, a detailed analysis of CII-specific sequences is required to substantiate the hypothesis that the possession of multiple chromosomes in bacteria has some adaptive advantages.

ACKNOWLEDGMENTS

This work was supported by Department of Energy grant DOE ER63232-1018220-0007203 and NIH grant GM15590-37 to Samuel Kaplan.

REFERENCES

1. **Blanc, G., A. Barakat, R. Guyot, and M. Delseny.** 2000. Extensive duplication and reshuffling in the *Arabidopsis* genome. Plant Cell **12:**1093–1102.
2. **Canchaya, C., M. J. Claesson, G. F. Fitzgerald, D. Sinderen, and P. W. O'Toole.** 2006. Diversity of the genus *Lactobacillus* revealed by comparative genomics of five species. Microbiology **152:**3185–3196.
3. **Choudhary, M., Y.-X. Fu, C. Mackenzie, and S. Kaplan.** 2004. DNA sequence duplication in *R. sphaeroides* 2.4.1: evidence of an ancient partnership between chromosome I and II. J. Bacteriol. **186:**2019–2027.
4. **Darling, Aaron, C. E., B. Mau, F. R. Blattner, and N. T. Perna.** 2004. Mauve: multiple alignment of conserved genomic sequences with rearrangements. Genome Res. **14:**1394–1403.
5. **Delcher, A. L., S. Kasif, R. D. Fleischmann, J. Peterson, O. White, and S. L. Salzberg.** 1999. Alignment of whole genomes. Nucleic Acids Res. **27:**2369–2376.
6. **Gest, H.** 1972. Energy conservation and generation of reducing power in bacterial photosynthesis. Adv. Microb. Physiol. **7:**243–282.
7. **Gevers, D., K. Vandepoele, C. Simillion, and Y. V. Deeper.** 2004. Gene duplication and biased functional retention of paralogs in bacterial genomes. Trends Microbiol. **12:**148–154.
8. **Graur, D., Y. Shuali, and W. H. Li.** 1989. Deletions in processed pseudogenes accumulate faster in rodents than in humans. J. Mol. Evol. **28:**279–285.
9. **Hancock, J. M.** 2005. Gene factories, microfunctionalization and the evolution of gene families. Trends Genet. **11:**591–595.
10. **Herdman, M.** 1985. The evolution of bacterial genomes, p. 37–68. *In* T. Cavalier-Smith (ed.), The evolution of genome size. John Wiley and Sons, Chichester, United Kingdom.
11. **Inagaki, Y., W. F. Doolittle, S. L. Baldauf, and A. J. Roger.** 2002. Lateral transfer of an EF-1 alpha gene: origin and evolution of large subunit of ATP sulfurylase in eubacteria. Curr. Biol. **12:**772–776.
12. **Jumas-Bilak, E., S. Michaux-Charachon, G. Bourg, M. Ramuz, and A. Allardet-Servent.** 1998. Unconventional genomic organization in the alpha subgroup of the *Proteobacteria*. J. Bacteriol. **180:**2749–2755.
13. **Langkjaer, R., P. F. Cliften, M. Johnston, and J. Piskur.** 2003. Yeast genome duplication was followed by asynchronous differentiation of duplicated genes. Nature **42:**848–852.
14. **Larimer, F. W., P. Chain, L. Hauser, J. Lamerdin, S. Malfatti, L. Do, M. L. Land, D. A. Pelletier, J. T. Beatty, A. S. Lang, F. R. Tabita, J. L. Gibson, T. E. Hanson, C. Bobst, J. L. Torresy Torres, C. Peres, F. H. Harrison, J. Gibson, and C. S. Harwood.** 2004. Complete genome sequence of the metabolically versatile photosynthetic bacterium *Rhopseudomonas palutris*. Nat. Biotechnol. **22:**55–61.
15. **Li, W. H., T. Gojobori, and M. Nei.** 1981. Pseudogenes as a paradigm of neutral evolution. Nature **292:**237–239.
16. **Lynch, M., and J. S. Conery.** 2000. The evolutionary fate and consequences of duplicated genes. Science **290:**1151–1155.

17. **Lynch, M., and A. G. Force.** 2000. Gene duplication and the origin of interspecific genomic incompatibility. Am. Nat. **156:**590–605.
18. **Mackenzie, C., S. Kaplan, and M. Choudhary.** 2004. Multiple chromosomes: intracellular mechanism for generating sequence diversity, p. 82–101. *In* R. V. Miller and M. J. Day (ed.), Microbial evolution. ASM Press, Washington, DC.
19. **Mackenzie, C., M. Choudhary, F. W. Larimer, P. F. Predki, S. Stilwagen, J. P. Armitage, R. D. Barber, T. J. Donohue, J. P. Hosler, J. E. Newman, J. P. Shapleigh, R. E. Sockett, J. Zeilstra-Ryalls, and S. Kaplan.** 2001. The home stretch, a first analysis of the nearly completed genome of *Rhodobacter sphaeroides* 2.4.1. Photosynthesis Res. **70:**19–41.
20. **McLysaght, A., K. Hokamp, and K. H. Wolfe.** 2002. Extensive genomic duplications during early chordate evolution. Nat. Genet. **31:**200–204.
21. **Mira, A., H. Ochman, and N. A. Moran.** 2001. Deletional bias and the evolution of bacterial genomes. Trends Genet. **17:**589–596.
22. **Moran, M. A., A. Buchan, J. M. Gonzalez, J. F. Heidelberg, W. B. Whitman, R. P. Kiene, J. R. Henriksen, G. M. King, R. Belas, C. Fuqua, L. Brinkac, M. Lewis, S. Johri, B. Weaver, G. Pai, J. A. Eisen, E. Rahe, W. M. Sheldon, W. Ye, T. R. Miller, J. Cartton, D. A. Rasko, I. T. Paulsen, Q. Ren, S. C. Daugherty, R. T. Deboy, R. J. Dodson, A. S. Durkin, R. Madupu, W. C. Nelson, S. A. Sullivan, M. J. Rosovitz, D. H. Haft, J. Selengut, and N. Ward.** 2004. Genome sequence of *Silicibacter pomeroyi* reveals adaptations to the marine environment. Nature **432:**910–913.
23. **Nereng, K., and S. Kaplan.** 1999. Genomic complexity among strains of the facultative phototrophic bacterium *Rhodobacter sphaeroides*. J. Bacteriol. **181:**1684–1688.
24. **Ochman, H., J. G. Lawrence, and E. A. Groisman.** 2000. Lateral gene transfer and the nature of bacterial innovation. Nature **405:**299–304.
25. **Ohnishi, M., K. Kurokawa, and T. Hayashi.** 2001. Diversification of *Escherichia coli* genomes: are bacteriophages the major contributors? Trends Microbiol. **9:**481–485.
26. **Ohno, S.** 1970. Evolution by gene duplication. Springer-Verlag, Heidelberg, Germany.
27. **Petrosino, J. F., Q. Xiang, S. E. Karpathy, H. Jiang, S. Yerrapragada, Y. Liu, J. Gioia, L. Hemphill, A. Gonzalez, T. M. Raghavan, A. Uzman, G. E. Fox, S. Highlander, M. Reichard, R. J. Morton, K. D. Clinkenbeard, and G. M. Weinstock.** 2006. Chromosome rearrangement and diversification of *Francisella tularensis* revealed by the type B (OSU18) genome sequence. J. Bacteriol. **188:**6977–6985.
28. **Petrov, D. A., and D. L. Hartl.** 1997. Trash DNA is what gets thrown away: high rate of DNA loss in *Drosophila*. Gene **205:**279–289.
29. **Petrov, D. A., T. A. Sangster, J. S. Johnston, D. L. Hartl, and K. L. Shaw.** 2000. Evidence of DNA loss as a determinant of genome size. Science **287:**1060–1062.
30. **Sankoff, D.** 2003. Rearrangements and chromosomal evolution. Curr. Opin. Genet. Dev. **13:**1–5.
31. **Suwanto, A., and S. Kaplan.** 1989. Physical and genetic mapping of the *Rhodobacter sphaeroides* 2.4.1 genome: presence of two unique circular chromosomes. J. Bacteriol. **171:**5850–5859.
32. **Suwanto, A., and S. Kaplan.** 1992. Chromosome transfer in *Rhodobacter sphaeroides*: Hfr formation and genetic evidence for two circular chromosomes. J. Bacteriol. **174:**1135–1145.
33. **Tagomori, K., T. Iida, and T. Honda.** 2002. Comparison of genome structure of vibrios, bacteria possessing two chromosomes. J. Bacteriol. **184:**4351–4358.
34. **Thompson, F. L., T. Iida, and J. Swings.** 2004. Biodiversity of vibrios. Microbiol. Mol. Biol. Rev. **68:**403–431.
35. **Ureta-Vidal, A., L. Ettwiller, and E. Birnay.** 2003. Comparative genomics: genome wide analysis in metazoans eukaryotes. Nat. Rev. Genet. **4:**251–262.
36. **Van Niel, C. B.** 1944. The culture, general physiology, morphology, and classification of the nonsulfur purple and brown bacteria. Bacteriol. Rev. **8:**1–118.
37. **Vision, T. J., D. G. Brown, and S. D. Tanksley.** 2000. The origin of genomic duplication in *Arabidopsis*. Science **290:**2114–2117.
38. **Woese, C. R., E. Stachebrandt, W. G. Weisburg, B. J. Paster, M. T. Madigan, C. R. M. Fowler, C. M. Hahn, P. Blanz, R. Gupta, K. H. Nealson, and G. E. Fox.** 1984. The phylogeny of the purple bacteria: the α subdivision. Syst. Appl. Microbiol. **5:**315–326.
39. **Xu, Q., M. Dziejman, and J. J. Mekalanos.** 2003. Determination of the transcriptome of *Vibrio cholerae* during intraintestinal growth and mid-exponential phase *in vitro*. Proc. Natl. Acad. Sci. USA **100:**1286–1291.
40. **Yamaichi, Y., T. Iida, K. S. Park, K. Yamamoto, and T. Honda.** 1999. Physical and genetic map of the genome of *Vibrio parahaemolyticus*: presence of two chromosomes in *Vibrio* species. Mol. Microbiol. **31:**1513–1521.
41. **Zhou, S., E. Kvikstad, A. Kile, J. Severin, D. Forrest, R. Runnheim, C. Churas, T. S. Anantharaman, J. W. Hickman, C. Mackenzie, M. Choudhary, T. J. Donohue, S. Kaplan, and D. C. Schwartz.** 2003. Whole-genome shotgun optical mapping of *Rhodobacter sphaeroides* 2.4.1 and its use for whole-genome shotgun sequence assembly. Genome Res. **13:**2142–2151.