Photosynthesis Research **70:** 19–41, 2001.

Minireview

The home stretch, a first analysis of the nearly completed genome of *Rhodobacter sphaeroides* 2.4.1

Chris Mackenzie[1], Madhusudan Choudhary[1], Frank W. Larimer[2], Paul F. Predki[3], Stephanie Stilwagen[3], Judith P. Armitage[4], Robert D. Barber[5], Timothy J. Donohue[6], Jonathan P. Hosler[7], Jack E. Newman[6], James P. Shapleigh[8], R. Elizabeth Sockett[9], Jill Zeilstra-Ryalls[10] & Samuel Kaplan[1,*]

[1]*Department of Microbiology and Molecular Genetics, University of Texas-Houston Medical School, 6431 Fannin St., Houston, TX 77030, USA;* [2]*Life Sciences Division, 1060 Commerce Park, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA;* [3]*DOE Joint Genome Institute, 2800 Mitchell Drive, B400, Walnut Creek, CA 94598, USA;* [4]*Department of Biochemistry, University of Oxford, South Parks Road, Oxford OX1 3QU, UK;* [5]*University of Wisconsin-Parkside, Department of Biological Sciences, Greenquist Hall Kenosha, WI 53141, USA;* [6]*Bacteriology Department, University of Wisconsin-Madison, 1550 Linden Drive, Madison, WI 53706, USA;* [7]*Department of Biochemistry, University of Mississippi Medical Center, 2500 N. State St., Jackson, MS 39216, USA;* [8]*Department of Microbiology, Wing Hall, Cornell University, Ithaca, NY 14853-8101, USA;* [9]*Institute of Genetics, University of Nottingham, QMC, Nottingham NG7 2UH, UK;* [10]*Department of Biological Sciences, Oakland University, Rochester, MI 48309, USA; *Author for correspondence (e-mail: Samuel.Kaplan@uth.tmc.edu; fax: +1-713-500-5499)*

Key words: complexity, flagella, gene duplication, genome, heme biosynthesis, nitrogen oxide reductase, scaffold, sequencing, sigma factors, terminal oxidases

Abstract

Rhodobacter sphaeroides 2.4.1 is an α-3 purple nonsulfur eubacterium with an extensive metabolic repertoire. Under anaerobic conditions, it is able to grow by photosynthesis, respiration and fermentation. Photosynthesis may be photoheterotrophic using organic compounds as both a carbon and a reducing source, or photoautotrophic using carbon dioxide as the sole carbon source and hydrogen as the source of reducing power. In addition, *R. sphaeroides* can grow both chemoheterotrophically and chemoautotrophically. The structural components of this metabolically diverse organism and their modes of integrated regulation are encoded by a genome of ~4.5 Mb in size. The genome comprises two chromosomes CI and CII (2.9 and 0.9 Mb, respectively) and five other replicons. Sequencing of the genome has been carried out by two groups, the Joint Genome Institute, which carried out shotgun-sequencing of the entire genome and The University of Texas-Houston Medical School, which carried out a targeted sequencing strategy of CII. Here we describe our current understanding of the genome when data from both of these groups are combined. Previous work had suggested that the two chromosomes are equal partners sharing responsibilities for fundamental cellular processes. This view has been reinforced by our preliminary analysis of the virtually completed genome sequence. We also have some evidence to suggest that two of the plasmids, pRS241a and pRS241b encode chromosomal type functions and their role may be more than that of accessory elements, perhaps representing replicons in a transition state.

Introduction

Since first described by van Neil in 1944 (van Neil 1944), *Rhodobacter sphaeroides* (originally called *Rhodopseudomonas spheroides*) has become the bioenergetics focus of many laboratories around the world. It is capable of most methods of energy acquisition. It can grow; photoheterotrophically using

a variety of organic compounds as both a source of carbon and reducing power, photoautotrophically in the absence of oxygen using carbon dioxide as the sole carbon source and hydrogen as a source of reducing power, chemoheterotrophically in the dark in the presence of oxygen using a variety of reduced organic compounds as both a source of carbon and reducing power and chemoautotrophically in the dark in the presence of oxygen using carbon dioxide as the sole carbon source and hydrogen as a source of reducing power. Fermentative growth occurs without light in the absence of oxygen.

The metabolic diversity of this organism is further illustrated by its ability to oxidize a broad spectrum of organic compounds including organic acids, sugars, polyols, methanol and toxic compounds like formaldehyde (Clayton and Sistrom 1978; Barber and Donohue 1998). It is also able to reduce both inorganic and organic compounds such as sulfate, toxic metal oxides/oxyanions and thymine (Moore and Kaplan 1992; Mouncey et al. 1997).

Besides bioenergetic capacity, *R. sphaeroides* has been examined over the years for a number of other reasons. Being a member of the α-subgroup of eubacteria places it within a diverse group of organisms that include the *Rhizobia*, *Agrobacteria* and *Rickettsia* (Woese et al. 1984; Woese 1987). It has been forcefully argued that a member of this group may have been the ancestor of the mitochondria (Yang et al. 1985). *R. sphaeroides* in particular has both structural features, e.g. photosynthetic membranes and biochemical features, e.g. aminolevulinic acid synthases (Neidle and Kaplan 1993a, b) and benzodiazepine receptors (Yeliseev and Kaplan 1995), that reinforce this view.

The work of Antonius Suwanto demonstrated that in *R. sphaeroides* strain 2.4.1 all of these processes are encoded by a genome of ~4.5 Mb in size (Suwanto and Kaplan 1989a). But surprisingly and somewhat controversial at the time, was that genome did not consist of a single circular chromosome but rather of two circular chromosomes, CI (2.9 Mb) and CII (0.9 Mb) plus five other replicons (Suwanto and Kaplan 1989b). Four of these replicons were shown to range in size from ~114 to 100 kb (pRS241a–d). A fifth replicon, known as the S-factor (pRS241e) was the smallest at ~42 kb in size (Fornari et al. 1984). Two of these replicons, pRS241d and pRS241e were also shown to be self-transmissible (Suwanto and Kaplan 1992).

The chromosomal status of CII has not always been clear. Some in the scientific community considered CII a megaplasmid such as those found in *Rhizobium meliloti*. We felt that key to giving CII chromosome status rested with the demonstration of essentiality under physiologic conditions. That is, if essential functions could be demonstrated to be encoded by this replicon then we should in good conscience designate CII a chromosome and not a megaplasmid.

After the creation of the first physical map of *R. sphaeroides*, it was demonstrated that CII encoded two ribosomal RNA operons (*rrnB* and *rrnC*). A third ribosomal rRNA operon (*rrnA*) mapped to CI (Dryden and Kaplan 1990, 1993). With time, other clearly important genes were found to reside on CII. However, like the *rrn* operons, these genes were also found to have duplicates on CI. Such duplicate gene pairs included; $cbbA_{I}$/$cbbA_{II}$ and $cbbP_{I}$/$cbbP_{II}$ (enzymes of the reductive Calvin Cycle (Hallenbeck et al. 1990a, b)), *hemA*/*hemT* (5-aminolevulinic acid synthase (Neidle and Kaplan 1993a, b)), *rdxB*/*rdxA* (ferredoxin-like (Neidle and Kaplan 1992)), $rpoN_{I}$/$rpoN_{II}$ (alternative sigma factors, σ^{54} (Meijer and Tabita 1992)), and $groEL_{I}$/$groEL_{II}$ (chaperones (Lee et al. 1997)), to name but a few. While these genes, as well as the *rrn* operons, clearly demonstrated that CII encoded important functions they were not sufficient to address the essentiality question because of their duplication.

To determine the true nature of this replicon, we subjected the genome to random transposon (Tn*5*) mutagenesis (Mackenzie et al. 1995). We then looked for auxotrophic strains carrying Tn*5* insertions that mapped to CII. Five strains, auxotrophic for histidine, thymine, serine, uracil and tryptophan were found (Choudhary et al. 1994). Sequencing near the site of Tn*5* insertion in the Trp^{-} strains indicated that the insertion was located within *trpB*, with *trpF* just a short distance upstream. In this region, we also found *rpsA1*, a gene that encodes the largest protein of the small ribosomal subunit and *cmkA* which encodes cytidylate monophosphate kinase (Mackenzie et al. 1999). In *E. coli*, these genes are considered essential for translation and chromosome replication, respectively. By encoding essential genes involved in central metabolic processes the genomic status of CII was unambiguously demonstrated to be that of a chromosome (Choudhary et al. 1994).

Buoyed by these findings we embarked upon a low redundancy approach to sequencing CII. The findings of this work reinforced the notion that CII encoded typical chromosomal functions and did not appear to

encode genes specialized for any particular mode of growth (Choudhary et al. 1997, 1999).

In July 2000, the Department of Energy Joint Genome Institute (JGI) offered to include *R. sphaeroides* in its microbial genome sequencing initiative which involved subjecting the genome to a whole genome shotgun sequencing strategy. We then combined the DNA sequences from the CII specific project with those generated by the JGI. This resulted in the closure of many gaps in CII and gave us a scaffold of order and orientation of the contigs that we believe is representative of the *in vivo* location of DNA islands within the genome. With the caveat that our maps are accurate, we have decided in this review to broadly examine the distribution of the genes between the two chromosomes and the other replicons. In addition, we compare the genome of *R. sphaeroides* to what is considered a close relative, *Rhodobacter capsulatus*. We also note genes or families of genes that we think are points of interest to the traveler passing around the genome. All sequence information described here can be accessed through our web-site at www.rhodobacter.org.

Methodology

Sequencing CII

Genomic DNA libraries of *R. sphaeroides* strain 2.4.1 were constructed using the cosmid vectors pLA2197 and pJRD215 (Allen and Hanson 1985; Heusterspreute et al. 1985). These were ordered into a minimal set (46 cosmids) around CII (Choudhary et al. 1994). Each cosmid was digested with restriction enzymes that cut infrequently or not at all within the vector, but a few times (4–10 times) within the insert. The restriction fragments were subcloned into pBluescript SK(+) then templates were prepared. The subclone insert ends were sequenced using the T3 and T7 primers and the sequences assembled using Phred/Phrap (Choudhary et al. 1997). The assemblies were then used to design primers to continue the sequencing by primer walking. Subsequent rounds of primer design, sequencing, and assembly were then carried out. All sequencing was carried out using Big Dye Terminator cycle-sequencing chemistry on Applied Biosystems 373 Stretch and 377 machines. A total of 4526 CII specific reads were generated and these assembled into 176 contigs that ranged in size from 15 to 80 kb.

Whole genome sequencing

Genomic DNA of *R. sphaeroides* strain 2.4.1 was prepared using 1% CTAB (cetyltrimethylammonium bromide) and 0.7 M NaCl (Wilson 1989). Readers should refer to the 'Methods' section of the Minireview 'An overview of the genome of *Nostoc punctiforme*, a multicellular, symbiotic cynobacterium' in this issue for details of the methodologies used in data processing (Meeks et al. 2001).

Assembly of 61 000 reads (∼7.0 × coverage) resulted in 195 contigs. The total contig size was ∼4.6 Mb, 200 kb larger than the estimated genome size of ∼4.5 Mb. The largest contig was ∼180 Kb. Contigs of 20 reads or greater (160 contigs with a total contig size of 4 578 469 bp) were 'first pass' annotated and can be accessed at http://genome.ornl.gov/microbial/rsph/.

Reassembly and scaffold data

Using Phrap, the 176 CII specific contigs were assembled with FASTA files from the 195 whole genome contigs. The outcome of this assembly was that 52 whole genome contigs were assigned to CII. Subsequently, CII was reduced from 176 small contigs to being described by 10 large contigs. These were generated by the merging of the CII specific sequences and the 52 whole genome sequences. By default, the remaining 143 whole genome contigs were assigned collectively to CI and the other replicons.

The JGI also provided scaffolding information for the Phrap assembly of the whole genome shotgun sequence data. A scaffold is a set of contigs linked together by plasmid paired end sequence data with the contigs in the correct order and orientation. The ideology used to determine the contig order and orientation can be further described as follows. The insert of a subclone X is sequenced at both ends with universal forward and reverse primers. These reads are then assembled with other 'paired-end' reads from other subclones. If one end of subclone X assembled into left end of contig A and the other end assembled into right end of contig B, this result would imply that contig A and B are adjacent to one another. It also implies that the left end of contig A is linked to the right end of contig B, which gives information not only on contig order but also on contig orientation. By examining the scaffold data, 115 whole genome contigs were assigned to CI. The scaffold data also allowed us to verify the accuracy of assembly of the 10 CII contigs. With these data we generated contig maps (see Figures 1a, b). They represent our current

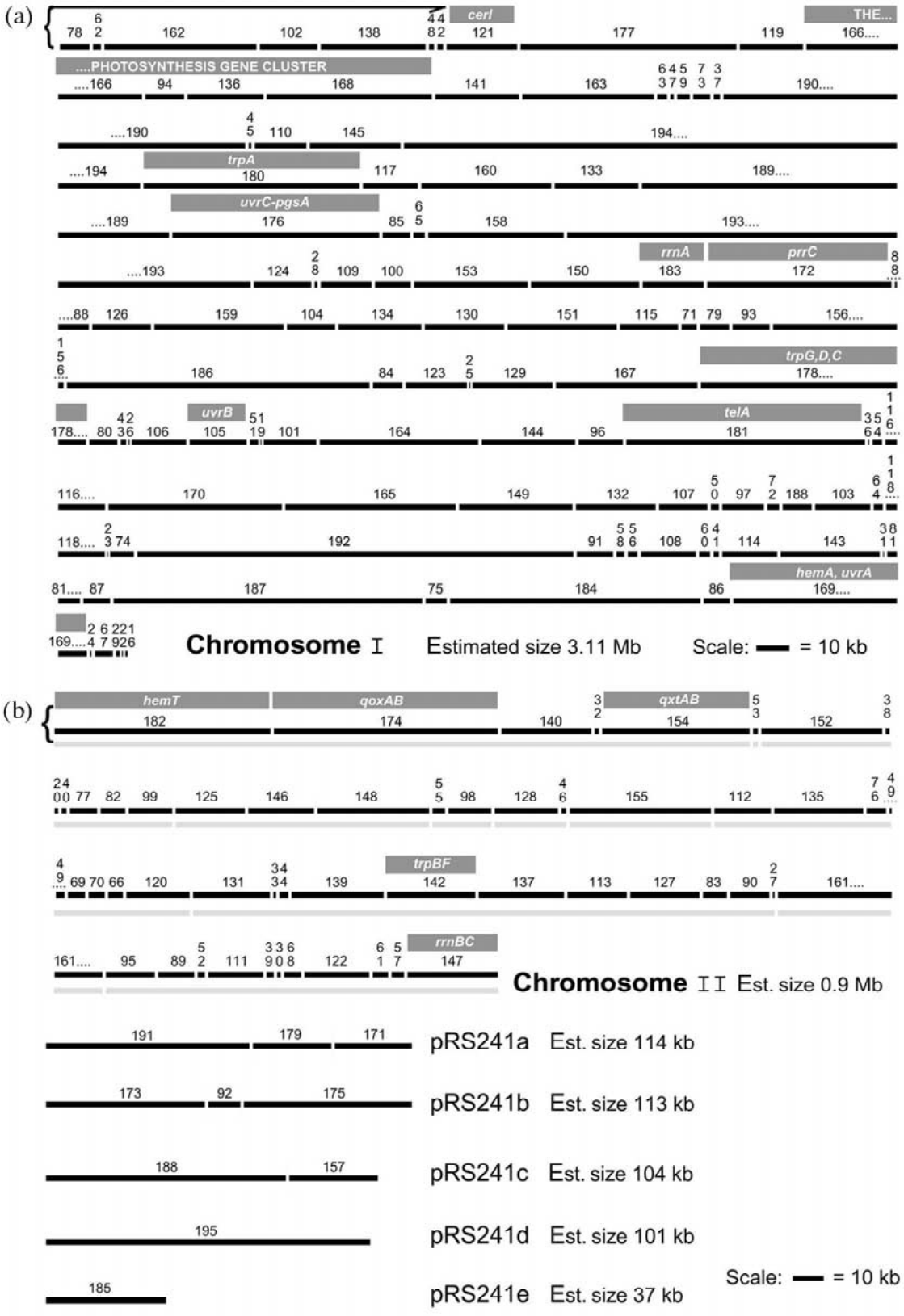
(a)
cerI
THE...
....PHOTOSYNTHESIS GENE CLUSTER
trpA
uvrC-pgsA
rrnA
prrC
trpG,D,C
uvrB
telA
hemA, uvrA
Chromosome I
Estimated size 3.11 Mb
Scale: = 10 kb
(b)
hemT
qoxAB
qxtAB
trpBF
rrnBC
Chromosome II Est. size 0.9 Mb
pRS241a Est. size 114 kb
pRS241b Est. size 113 kb
pRS241c Est. size 104 kb
pRS241d Est. size 101 kb
pRS241e Est. size 37 kb
Scale: = 10 kb

Figure 1.

working model of the genome and fulfill all of the assembly, scaffold and independently generated physical and genetic data.

It is important to point out that we have a high degree of confidence in the maps for CI, CII, and pRS241d. However, we had very little independent (physical/genetic) data for the plasmids a, b, c and e and their maps were deduced from scaffold data alone. Because of this, we are somewhat less confident in them and feel that they may change as more data becomes available. Supplemental scaffold data, which includes contig orientation, has been provided at www.rhodobacter.org/psresreview.

The maps account for the location of 175 contigs from the whole genome project. Twenty contigs, eight of which were single reads, could not be placed on the map. The largest unplaced contig was 1.5 kb. The sum of the unplaced contigs was 15 kb. The predicted sizes of CI and CII and other replicons are within <10% of the sizes predicted by pulsed-field gel electrophoresis.

Genome overview

The estimated sizes of CI, CII and plasmids pRS241a–e are 3110, 901, 114, 113, 104 and 37 kb, respectively, with % G + C compositions of 68.9%, 69.1%, 69.3%, 70.1%, 64.1%, 63.9% and 67.6%, respectively, see Table 1. The overall genome % G + C composition is 68.81%.

The di- and trinucleotide frequencies were examined and the four most frequent and four least frequent are presented in Table 1. These results indicate that the CI, CII, pRS241a, pRS241b and pRS241e share similar frequencies of both di and trinucleotides, whereas plasmids pRS241c and pRS241d are somewhat different in composition to the other replicons. This can be exemplified by examining the most common dinucleotide (CG) that makes up the following percentages: 13.9, 13.9, 13.5, 14.5, 11.6, 11.6 and 13.2% of CI, CII and pRS241a–e, respectively, i.e. the CG dinucleotide is ~2% less frequent in pRS241c and pRS241d compared to what is found in the other replicons. This notion is reinforced by looking at the most common of the trinucleotides (GCG) that make up 4.9, 5.0, 5.0, 5.3, 3.6, 3.8 and 4.7% of CI, CII, pRS241a–e, respectively. This difference in frequency between plasmids pRS241c and d and the rest of the genome is consistent for all the di- and trinucleotides not just those shown in the table. This result suggests that these replicons may have been a recent horizontal acquisition from a slightly lower %G + C organism. The difference in their composition compared to the rest of the genome probably results from their DNA having had less time to drift towards the genome average.

Codon usage

The 4152 predicted gene models were analyzed for their codon usage, the results being presented in Table 2. All of the 64 possible codons in the genetic code were used. It was also noted that an adenine or thymine deoxynucleotide at the third position correlated with a greatly reduced frequency of usage of that codon. This is perhaps most striking in the case of the codons for isoleucine (Ile) where the codon ATC is used 91% of the time but codons ATT and ATA are used just 3 and 6% of the time, respectively.

Exceptions to this scenario were found in tyrosine (Tyr), histidine (His) and termination codons. In the case of Tyr and His, the frequency of codons with an A or T at the third position were only marginally lower than those with a G or C, 47% vs 53% for Tyr and

←

Figure 1. (a, b) These maps represent our current working model of the order of the DOE contigs around the two chromosomes. The black horizontal lines represent, to scale, the size of the contigs. Above each black line is the contig number. Small contigs have their numbers written vertically, e.g. the second contig on the chromosome I map is 62. Contigs that wrap onto a second line have ellipsis (. . .) next to the contig number. The contig number can be used to download DNA sequence and annotation information for that contig from the web site www.rhodobcter.org. A brace ( ) at the top left of each chromosome map indicates which lines of the map should be read together e.g. the top left line of the chromosome I has contigs 78, 62, 162, 102, 138, 48, 42 and 121 with the gene *cerI* placed on contig 121. At the right end of the top line is the start of contig 166 and the photosynthesis gene cluster. Note that both of these wraps onto the next line. Some pulse field gel mapped and sequenced marker genes have been placed *above* the contigs in dark gray boxes. Note that the size of the box does not represent the size of the gene, it simply indicates that, for example, *trpA* maps to contig 180. On chromosome II there is light gray line beneath the contig lines. This indicates the contigs that were generated by amalgamating the data from the CII specific project and the DOE whole genome project. This sequence can be downloaded from our web site. Note that the size of chromosome I is larger than on previously published maps, 3.11 vs 2.97 Mb. The size may increase (or decrease) as assembly continues. At the current level the difference is <10% of previously published maps. Chromosome II came out to be close to the previously published 0.9 Mb. We estimate that the sum of the gaps in the DOE CII sequence is 11 kb.

Table 1. Variations in replicon di- and trinucleotide composition

	Replicon						
	CI	CII	pRS241a	pRS241b	pRS241c	pRS241d	pRS241e
%G + C	68.9	69.1	69.3	70.1	64.1	63.9	67.6
dinucleotide composition (%)							
CG	13.9	13.9	13.5	14.5	11.6	11.6	13.2
GC	13.3	13.4	13.4	14.0	11.3	11.5	12.9
GG	10.2	10.1	10.9	10.1	8.8	9.3	10.0
CC	10.0	10.0	10.3	10.3	9.5	8.2	9.6
AT	3.8	3.8	3.7	3.8	4.3	4.3	4.0
AA	2.4	2.3	2.4	2.0	3.8	3.7	2.6
TT	2.4	2.3	2.4	2.0	3.3	3.3	2.8
TA	0.9	0.9	1.0	0.8	1.5	1.6	1.2
trinucleotide composition (%)							
GCG	4.9	5.0	5.0	5.3	3.6	3.8	4.7
CGC	4.9	5.0	4.7	5.3	3.8	3.8	4.4
GGC	4.6	4.6	4.8	4.6	3.5	3.8	4.3
GCC	4.5	4.5	4.5	4.9	3.7	3.4	4.2
TAC	0.3	0.3	0.2	0.2	0.5	0.6	0.3
GTA	0.2	0.2	0.3	0.2	0.5	0.4	0.3
TAA	0.1	0.1	0.1	0.1	0.2	0.2	0.2
TTA	0.1	0.1	0.1	0.1	0.2	0.3	0.2

This table shows the %G + C, dinucleotide and trinucleotide compositions of the seven replicons or *R. sphaeroides*. The four most and four least used di- and tri-nucleotides expressed as a percentage of all di- or trinucleotides are shown. The most common are separated from the least common by a dotted line.

46% vs 54% for His. The biological meaning of this observation is not clear. In the case of the termination codons, TGA was used 82% of the time but TAG was used only 10% of the time, just the reverse of those observed for the sense codons. However, if we use the second position of these two codons, they follow the same rule. Perhaps this is due to the fact that in the second position of sense codons (with the exception of the serine codons) the base is fixed, while in the termination codons it is variable.

The most commonly used sense codon i.e. coding for an amino acid rather than for polypeptide termination was GCC (coding for alanine, 86 807 instances) and least used TTA (coding for leucine, 421 instances). It might have been expected that GCC, being the most commonly used codon would also have been the most common trinucleotide (it was the fourth most common) but another alanine codon (GCG) held this title. The least frequent trinucleotide, TTA, was also the least used codon.

Although the 61 possible sense codons were found, the same cannot be said (despite and exhaustive search) of their cognate tRNA genes. It is plausible that some tRNA genes do not exist and codon wobble counteracts their absence. This might be the case for threonine where a tRNA carrying the AGT anticodon (the ACT codon is the least used, 3% usage) has not been found but all other tRNA genes for threonine are present. However, it seems unlikely in other cases, e.g. the most frequently used codon for arginine, CGC, which is used 49% of the time has not as yet a cognate tRNA gene. It seems likely in this case and in others, that the tRNA gene lies in a sequencing gap, or lies in a region of poor quality DNA sequence, making it impossible for the tRNA calling software to detect the gene.

Like many genes in *R. sphaeroides*, the tRNA genes are often found in multiple copies, see Tables 2 and 3. Copies may be on the same chromosome e.g. the aspartyl-tRNA, or they may be distributed between

Table 2. Codon usage

Amino acid	Codon	number[a]	/1000[b]	Fraction[c]	tRNA[d]
Gly	GGG	29640	21.05	0.22	CI
Gly	GGA	9616	6.83	0.07	CII
Gly	GGT	10586	7.52	0.08	–
Gly	GGC	83799	59.5	0.63	2 CI
Glu	GAG	66636	47.31	0.77	–
Glu	GAA	19965	14.18	0.23	2 CI
Asp	GAT	26611	18.89	0.35	–
Asp	GAC	49144	34.89	0.65	2 CI
Val	GTG	51859	36.82	0.52	CI
Val	GTA	2308	1.64	0.02	CII
Val	GTT	6095	4.33	0.06	–
Val	GTC	38672	27.46	0.39	CI
Ala[e]	GCG	81346	57.76	0.43	CII
Ala[e]	GCA	9945	7.06	0.05	CI
Ala[e]	GCT	9607	6.82	0.05	–
Ala[e]	GCC	86807	61.64	0.46	CI
Arg	AGG	5193	3.69	0.04	CI
Arg	AGA	2052	1.46	0.02	CI
Ser	AGT	1734	1.23	0.03	–
Ser	AGC	17715	12.58	0.26	CI
Lys	AAG	27939	19.84	0.86	–
Lys	AAA	4398	3.12	0.14	CI
Asn	AAT	6021	4.28	0.23	–
Asn	AAC	20375	14.47	0.77	CI
Met	ATG	32711	23.23	1	3 CII, 2 CI
Ile[e]	ATA	1782	1.27	0.03	–
Ile[e]	ATT	3752	2.66	0.06	–
Ile[e]	ATC	54529	38.72	0.91	CI
Thr	ACG	26914	19.11	0.41	CI
Thr	ACA	2434	1.73	0.04	CI
Thr	ACT	1876	1.33	0.03	–
Thr	ACC	34652	24.6	0.53	CI
Trp	TGG	17509	12.43	1	CI
End	TGA	3394	2.41	0.82	N/A
Cys	TGT	1762	1.25	0.13	–
Cys	TGC	12004	8.52	0.87	CI
End	TAG	411	0.29	0.10	N/A
End	TAA	327	0.23	0.08	N/A
Tyr	TAT	11398	8.09	0.47	–
Tyr	TAC	12727	9.04	0.53	CI

Table 2. Continued

Amino acid	Codon	number[a]	/1000[b]	Fraction[c]	tRNA[d]
Leu	TTG	4310	3.06	0.03	CI
Leu	TTA	421	0.30	0	CI
Phe	TTT	4307	3.06	0.10	–
Phe	TTC	40896	29.04	0.90	CI
Ser	TCG	29723	21.10	0.44	1 CI, 1 CII
Ser	TCA	2047	1.45	0.03	CII
Ser	TCT	2910	2.07	0.04	–
Ser	TCC	13627	9.68	0.20	CI
Arg	CGG	40734	28.92	0.32	CI
Arg	CGA	8454	6.00	0.07	–
Arg	CGT	9153	6.50	0.07	CI
Arg	CGC	62147	44.13	0.49	–
Gln	CAG	39380	27.96	0.91	–
Gln	CAA	3789	2.69	0.09	2 CI
His	CAT	15680	11.13	0.46	–
His	CAC	18353	13.03	0.54	CI
Leu	CTG	66477	47.20	0.45	CI
Leu	CTA	1405	1.00	0.01	CI
Leu	CTT	13230	9.39	0.09	–
Leu	CTC	61096	43.38	0.42	CI
Pro	CCG	42049	29.86	0.5	CI
Pro	CCA	4236	3.01	0.05	CI
Pro	CCT	5764	4.09	0.07	–
Pro	CCC	31946	22.68	0.38	CI

[a]The number of occurrences of that codon in the genome.
[b]The number of occurrences of that codon per 1000 codons.
[c]The fractional value of one type of codon of the total codons for a particular residue.
[d]Forty-nine tRNAs and one pseudo-tRNA (codon GCA, but not shown in the table) were predicted from the DOE sequence (see footnote [e] below). tRNAs are listed with their cognate codons. The number of copies and their chromosomal location are also provided, e.g. there are two copies of the serine tRNA carrying the CGA anticodon (TCG codon), one located on CI the other on CII.
[e]The *R. sphaeroides* genome contains three ribosomal RNA operons, *rrnA* on CI and *rrnB*, *C* on CII. Within each of the operons, lying between the 16S and 23S genes, lies an isoleucine and an alanine tRNA (Dryden and Kaplan 1990). To date, the DOE sequence suggests that there is a single Ile tRNA that maps to *rrnA* on CI (where as you would predict 1 CI, 2 CII) and three Ala tRNAs (although only one of these maps to *rrnA* on CI). Probably due to misassembly of the *rrn* regions within the DOE sequence, Ile and Ala tRNAs are under represented within the genome. We estimate that in reality there are 3–4 Ile and 5–7 Ala tRNAs.

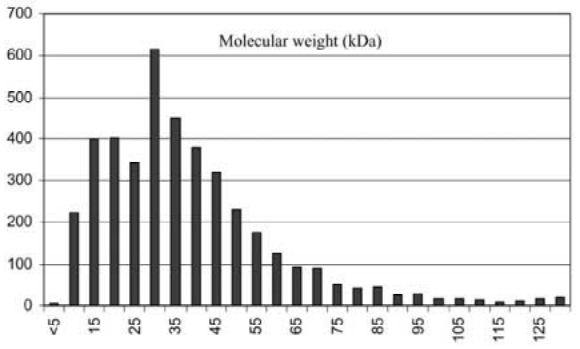

Figure 2. The distribution of *R. sphaeroides* proteins according to their molecular weight. Proteins greater than 130 kDa were few in number and are not shown.

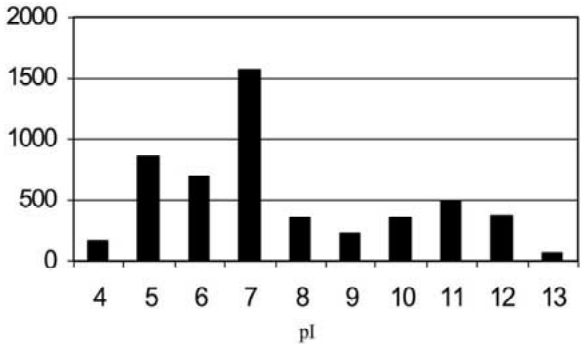

Figure 3. The distribution of *R. sphaeroides* proteins according to their isoelectric points (pI).

the two chromosomes e.g. glycyl-tRNA. To date, the most highly duplicated of the tRNAs is that of methionine. In this case, there are five methionyl-tRNAs, two copies on CI and three copies on CII. We know that three of these are formyl-Met tRNAs which lie downstream of the ribosomal RNA operons (Dryden and Kaplan 1990). The biological significance of this level of duplication and distribution remains unknown.

Polypeptide overview

As with other bacteria, the vast majority of the deduced proteins (∼85%) encoded by the genome fall in the 10–60 kDa size range, see Figure 2. However, there were some very large proteins. The largest weighed in at 299 kDa and was encoded by a homolog of *ndvB*, a gene required for *Rhizobium meliloti* root nodule development (Ielpi et al. 1990). Overall the proteins tended to be neutral to slightly acidic in nature (Figure 3) with a modal isoelectric point value of seven. This distribution of protein pIs is perhaps to be expected from a high %G + C organism where the GC bias of the codons would lead to a higher usage of neutral amino acids such as glycine and alanine.

Gene duplication

Gene duplication, followed by sequence divergence, has been suggested to be a major mechanism of molecular evolution. It was already known that gene duplication in *R. sphaeroides* was more common than in other organisms. Consequently, we undertook a systematic approach to study sequence duplication within the *R. sphaeroides* genome itself. We randomly sampled 25% (1038) of the total predicted genes and compared them to the genome as a whole to determine the level of gene duplication. We found that not only were many genes duplicated, but many were present at greater multiplicities (see Table 3). Of the genes sampled, 18% (189 genes) were found to be present in multiple copies. To us this seemed like an unusually high level of gene duplication. This was particularly true in the case of the genes involved in flagellar biosynthesis. It was known that planktonic *R. sphaeroides* have a single, sub-equatorial flagellum, therefore the finding that two copies of many of the structural genes for the formation of this structure are present on CI was somewhat surprising.

Previous mutational and sequence analysis of *R. sphaeroides* WS8 flagellar and motility genes had established that operons homologous with those on contigs 162, 102, 138 and 166 of the 2.4.1 genome are responsible for the synthesis and rotation of the subequatorial flagellum. The similarity between WS8 and 2.4.1 flagellar and motility proteins encoded by genes on these operons is very high ranging from 89 to 100% identity. Interestingly the FliC, flagellar filament proteins of 2.4.1 and WS8 are greater than 99% identical, despite the external selective pressures of predatory phage and protozoa which target flagellar filaments and which might have been expected to have selected for sequence diversity in FliCs between strains. In addition to the high levels of identity between 2.4.1 and WS8 sequences, the organization of the flagellar and motility operons on contigs 162, 102, 138 and 166 are virtually identical to those characterized in WS8. σ^{28} and two types of σ^{54} dependent promoters are found upstream of WS8 flagellar and motility genes (Sockett et al. 1999; Shah et al. 2000a); these consensus sequences are conserved in 2.4.1 and in addition the genome sequence has revealed that one of four *rpoN* (σ^{54}) genes lies downstream of the *fliC* gene in con-

Table 3. Gene duplication in *R. sphaeroides*

Gene	Copies[a]	Function	% Identity[b]	% Similarity[c]	E value[d]
		Flagellum biosynthesis			
flgB	2	Flagellar basal body protein	31	49	2.00E-04
flgC	2	Flagellar basal body rod protein	28	43	4.00E-04
flgE	2	Flagellar hook protein	24	54	3.00E-27
flgF	2	Basal body proximal rod protein	30	42	1.00E-04
flgG	2	Basal body rod protein	42	63	2.00E-52
flgH	2	Basal body L-ring protein	30	45	2.00E-14
flgI	2	Flagellar P-ring protein	39	54	2.00E-34
fliI	2	Flagellar specific ATP synthase	36	50	8.00E-60
fliF	2	Basal ring-M protein	29	42	3.00E-23
fliG	2	Flagellar motor switch protein	21	42	1.00E-09
fliN	2	Flagellar motor switch protein	35	58	4.00E-05
fliP	2	Flagellar biogenesis	41	67	2.00E-26
fliQ	2	Flagellar biogenesis/export	47	66	3.00E-11
fliR	2	Flagellar biogenesis	33	46	1.00E-20
flhA	2	Flagellar assembly	36	55	6.00E-95
flhB	2	Flagellar assembly	42	57	1.00E-20
		Carbon metabolism			
pucB	2	Light harvesting complex	94	95	4.00E-22
pucA	2	Light harvesting complex	58	72	8.00E-12
crtI	2	Phytoene desaturase	21	38	8.00E-14
dxs	2	Deoxyxylulose-5-phosphate synthase	66	78	0
cbbA	2	Fructose bisphosphate aldolase	79	87	1.00E-149
cbbP	2	Phosphoribulokinase	88	93	1.00E-128
cbbF	2	Fructose bisphosphatase	68	78	1.00E-116
cbbM	2	RuBP carboxylase	30	44	7.00E-30
cbbG	3	Glyceraldehyde-3P-dehydrogenase	46	63	1.00E-71
cbbT	2	Transketolase	58	69	0
		Oxidoreductases			
qoxA	2	Quinol oxidase	45	60	1.00E-179
nuoA	2	NADH dehyrdogenase I, A subunit	35	54	8.00E-10
nuoB	2	NADH dehyrdogenase I, B subunit	47	59	9.00E-30
nuoD	2	NADH dehyrdogenase I, D subunit	41	58	4.00E-79
nuoF	2	NADH dehyrdogenase I, F subunit	37	54	1.00E-68
nuoH	2	NADH quinone oxidireductase	41	60	1.00E-56
nuoI	2	NADH ubiquinone oxidoreductase	43	55	4.00E-24
nuoM	2	NADH dehyrdogenase I, M subunit	36	49	4.00E-09
qor	2	NAD(P)H quinone oxidoreductase	22	37	0.086
rdxA,B	2	RDX protein	67	79	0
		Chemotaxis			
cheA	4	Chemotaxis histidine kinase	34	47	4.00E-77
cheB	2	MCP-glutamate methylesterase	40	52	9.00E-56
cheR	3	MCP-glutamate methyltransferase	35	52	2.00E-33
cheW	4	Chemotaxis scaffold protein	33	55	9.00E-13
cheY	7	Chemotaxis response regulator	69	82	1.00E-42
tlp	4	Transducer like protein	24	39	5.00E-26
mcp	9	Chemorecptor protein	42	62	3.00E-67

Table 3. Continued

Gene	Copies[a]	Function	% Identity[b]	% Similarity[c]	E value[d]
		Replication & partitioning			
repA	5	Replication protein	29	48	4.00E-39
repB	3	Replication protein	34	49	5.00E-31
parA	2	Partitioning protein	29	47	1.00E-13
parB	4	Plasmid stabilization protein	97	98	1.00E-147
dnaE	2	DNA polymerase III, alpha subunit	31	48	1.00E-87
		Transport systems			
smoM	2	Mannitol binding protein	25	41	2.00E-12
expE1	4	Calcium binding protein	32	44	1.00E-27
dctP	2	Dicarboxylate binding protein	26	47	9.00E-15
		Regulators			
fnrL	2	Anaerobic regulator	32	51	4.00E-26
rpoN	4	Sigma factor	41	56	4.00E-69
dctD	3	Dicarboxylate transport regulator	35	47	4.00E-61
dctB	3	Dicarboxylate sensor kinase	30	50	3.00E-17
		Lipid metabolism			
fadB	3	Enoyl -CoA hydratase	33	44	3.00E-15
		Amino acid metabolism			
glnA	5	Glutamine synthase	27	39	2.00E-26
serA	4	Phosphoglycerate dehydrogenase	39	55	1.00E-44
trpB	2	Tryptophan synthase,beta subunit	45	62	2.00E-97
		Transposons			
y4bF	6	Putative transposase	78	79	2.00E-42
		Heat shock proteins			
groES	2	HSP10 (GroES)	37	57	8.00E-04
groEL	3	HSP60 (GroEL)	45	63	1.00E-121
cspA	2	Cold shock protein	88	95	3.00E-29
cspB	3	Cold shock protein	89	96	2.00E-32
		Others			
atpA	2	ATP synthase, alpha subunit	31	46	1.00E-16
atpB	2	ATP synthase, beta subunit	48	64	1.00E-121
atpC	2	Vacuolar/archael ATP synthase, K subunit	48	61	6.00E-04
hemA,T	2	5-Aminolevulinic acid synthase	54	70	1.00E-119
hemN	3	Coproporphyrinogen III oxidase	24	46	3.00E-13

[a]The number of copies of the gene found in the genome using BLASTP.
[b]Percentage of identical amino acids between the query protein and its matching duplicate. In instances where more than one duplicate copy of a gene was found, the amino acid identity given is for the subject that gave the lowest match with the deduced query.
[c]As in [b] above but includes similar plus identical residues.
[d]The expect value of the match. In instances where more than one duplicate copy of a gene was found, the expect value given is for the subject that gave the lowest match with the deduced query.

tig 102. These data lead us to the conclusion that the flagellar and motility genes (Set 1 genes) on contigs 162, 102, 138 and 166 of the 2.4.1 genome encode the sub-equatorial flagellum of that strain.

As mentioned above, in addition to the Set 1 genes another partial set of flagellar structural genes (Set 2) is evident on contigs 109, 124 and 181 of the 2.4.1 genome, still located on CI. Set 2 is not complete and thus could not encode an entire flagellum without using some Set 1 gene products, also there is a distinct absence of the σ^{54} and σ^{28} dependent promoter sequences at these loci suggesting that regulation of their expression differs to Set 1. The level of identity between Set 1 and Set 2 open reading frames is low (Table 3) ranging from 21% for FliG motor-switch proteins to 47% for FliQ export proteins, indicating that they likely did not arise as a result of a recent gene duplication event. Despite these low identities, in many cases key amino acids known experimentally to be required for flagellar assembly or function are conserved between Set 1 and Set 2 genes. The Set 2 genes had not previously been cloned by mutagenesis and complementation studies in *R. sphaeroides* WS8 and work is now underway to test their functionality.

Limited evidence that some Set 2 genes may be active, and may also be present in the WS8 genome, comes from studies of a *fliI* (Set 1 homologue) mutant in WS8. FliI is a flagellar specific ATPase which is essential to the FliC export process during external filament synthesis in *E. coli*, without it those bacteria are totally aflagellate and non motile. Goodfellow and co-workers (Goodfellow et al. 1996) found that in *R. sphaeroides* WS8, deletion of *fliI* (Set 1 homologue) resulted in a population of planktonic cells where approximately one in one thousand cells were flagellate and motile and the vast majority were aflagellate and non motile as predicted. This low percentage motility phenotype could be due to the activity of a Set 2 *fliI* gene in *R. sphaeroides* WS8, and experiments are currently underway to test this hypothesis and to establish what might produce such sporadic expression of such a gene in a population. A similar story has been found for FliF in WS8 also (unpublished data).

Interestingly some regulatory genes, including *flhCD* (the master regulator of flagellar synthesis in *E. coli*) and apparently *flgM* (another regulator of FliC synthesis that is an anti-sigma factor for σ^{28}) that are important in controlling the synthesis of flagella in enteric bacteria are not present in the draft genome sequence of 2.4.1. They may yet be discovered in the finishing process, or their absence may indicate some novel features in the flagellar regulatory hierarchy of *Rhodobacter.* We have only just scratched the surface in our understanding of the diverse lifestyles of *R. sphaeroides*.

Truly a 'Walter Mitty' amongst bacteria, the planktonic, free-swimming life style of *R. sphaeroides* may not be the only motile option open to it. The Set 2 flagellar genes could be required for surface translocation using lateral flagella produced during biofilm growth. The highest identity to the Set 2 motor gene products MotAB is to those of the lateral flagella of *Rhodospirillum centenum* (a phototroph that crawls on solid surfaces), so now genomic information prompts us to inquire more deeply into the expression of flagella during alternate growth modes of *R. sphaeroides*.

Basal transcription apparatus

In *R. sphaeroides* as in other eubacteria, RNA polymerase holoenzyme contains core RNA polymerase subunits ($\alpha_2\beta\beta'\omega$) and one of several sigma (σ) factors that allows it to recognize promoter DNA, interact with activator proteins, and initiate transcription. The *R. sphaeroides* genome sequence predicts that 18 of the 21 genes that encode the components of this basal transcription apparatus map to CI. CI contains the genes that encode all the core RNA polymerase subunits and the genes for 14 of the sigma factors; two sigma factors map to CII and one sigma factor gene lies on pRS241e.

Sigma factors

The genome sequence predicts the existence of 17 σ factors in *R. sphaeroides*; more than the 7 present in *E. coli* but fewer than the number predicted from other sequenced eubacterial genomes (http://www.ncbi.nlm.nih.gov/COG/). The two major superfamilies of bacterial sigma factors are distinguished by amino acid sequence similarity, the promoter elements they recognize, and how they influence transcription (Busby and Ebright 1994). Proteins in the σ^{70} superfamily recognize promoters containing 'hexamer' DNA sequences centered at -35 and -10 relative to the start of transcription. In contrast, promoters recognized by members of the σ^{54} superfamily recognize hexamers at -24 and -12 and require a specific type of activator to stimulate transcription.

Members of the σ^{54} superfamily

The *R. sphaeroides* genome sequence predicts the existence of four σ^{54} homologues. One of these ($RpoN_I$) is within the *nif* cluster (CI) that contains genes required for N_2 assimilation (Suwanto and Kaplan 1989a; Meijer and Tabita 1992). A gene for a second σ^{54} homologue ($RpoN_{II}$) is present on CII (Choudhary et al. 1997); two other σ^{54} homologues are encoded by CI. The loss of $RpoN_I$ either alone, or in combination with $RpoN_{II}$, shows that this protein is required for both diazotrophic growth and the synthesis of detectable nitrogenase activity, suggesting that $RpoN_{II}$ has no function under these conditions (Smith and Kaplan, unpublished). Similarly mutant analysis suggests that neither $RpoN_I$ nor $RpoN_{II}$ is involved in motility. On the other hand, the gene for $RpoN_{III}$ lies ~100 bp downstream of *fliC* and might control motility gene expression especially since it could be in the same operon. Consistent with this notion, an unknown σ^{54} family member has been suggested to play a role expression of *R. sphaeroides* flagellar biosynthesis genes (Poggio et al. 2000). In contrast, the $RpoN_{II}$ and $RpoN_{IV}$ structural genes are not located in regions of the genome that would give a hint as to their possible roles. Thus, it will be interesting to see if these σ^{54} homologues have unique regulons.

Members of the σ^{70} superfamily

The other 13 sigma factors predicted to exist in *R. sphaeroides* are members of the σ^{70} superfamily (Lonetto et al. 1992). Proteins in group 1 of the σ^{70} superfamily are 'housekeeping' sigma factors that are required for cell growth. *R. sphaeroides* was known to contain a protein with amino acid and functional similarity to the housekeeping σ^{70} protein (Kansy and Kaplan 1989; Karls et al. 1993, 1999; Gruber and Bryant 1997); the gene for this protein is on CI.

A well-studied member of group 2 proteins in the σ^{70} superfamily is *E. coli* σ^S, whose activity increases as cells enter stationary phase or respond to oxidative/osmotic stress (Lonetto et al. 1992). Many proteins in the group 2 σ^{70} superfamily are non-essential, so it is not surprising that *R. sphaeroides*, as well as other eubacteria (http://www.ncbi.nlm.nih.gov/COG/) do not contain a σ^S homologue.

Proteins in the group 3 σ^{70} superfamily have been placed in sub-divisions based on amino acid sequence similarity (Lonetto et al. 1992). These sub-divisions contain proteins (related to FliA) that control the expression of flagellar genes; homologues of RpoH that regulate the heat shock response, and sigma factors (members of the *E*xtra-*C*ytoplasmic *F*unction or ECF sub-family) that direct transcription of genes whose products function outside of the cytoplasm (Lonetto et al. 1994). *R. sphaeroides* is motile (Armitage and Schmitt 1997), so it is not surprising that the genome encodes a sigma factor related to FliA (chromosome I). CI also encodes two sigma factors, $RpoH_I$ and $RpoH_{II}$ each of which recognize heat shock gene promoters and the P1 promoter for the cytochrome c_2 gene (Karls et al. 1998; MacGregor et al. 1998).

The remaining 9 predicted σ factors are all members of the ECF subfamily, making this the largest group of sigma factors in *R. sphaeroides*. Six genes for ECF sigma factors are located on CI, one is encoded by CII and one maps to pRS241e. Only one of these ECF sigma factors, RpoE, has been characterized; the promoters known to be recognized by RpoE are the P3 promoter for the *cycA* gene and the P1 promoter for the *rpoEchrR* operon (Newman et al. 1999, 2001). It is common for ECF sigma factor activity to be controlled by the product of the gene that lies immediately downstream of the one that encodes the σ factor (Lonetto et al. 1994), so it is not surprising that ChrR inhibits RpoE activity *in vivo* and *in vitro* (Newman et al. 1999, 2001). The predicted amino acid sequence of the remaining *R. sphaeroides* ECF sigma factors and their downstream gene products suggests that *R. sphaeroides* uses these proteins to control iron acquisition (FecI/FecR homologues are encoded by pRS241e, and a PvdS homologue is encoded by CI) and the expression of periplasmic proteases (homologues of PrtRI are encoded on CII). The other 6 predicted ECF sigma factors and their potential inhibitors are not homologues of well-studied proteins in other eubacteria, so it is not clear what genes they might control in *R. sphaeroides*.

Contrasts along the pathway of protein expression

Given the vast metabolic repertoire of *R. sphaeroides*, it is interesting to consider the regulatory mechanisms that govern cell physiology and metabolism. An overview of gene products involved in transcription, translation and post-translational processes presents a contrasting view of potential complexity. For example, the content of the *R. sphaeroides* genome exhibits a fairly diverse group of predicted bacterial helix-turn-helix transcriptional regulatory proteins; a profile on par with what has been observed in other

bacteria such as *Bacillus subtilis*. However, this content is dwarfed when compared with the genomes of *Mesorhizobium loti* and *Pseudomonas aeruginosa*, which contain four to eight times as many proteins of any given helix-turn-helix regulator family.

One possible explanation for this relative reduction in 'complexity' could be subtle alterations in the profile of proteins involved in translation. The *R. sphaeroides* genome contains multiple genes encoding N-formylmethionyl-tRNA deformylase and translation initiation factor eIF5-a. Furthermore, a gene encoding translation initiation factor eIF-2B alpha subunit, found only in a limited number of bacteria, is present in *R. sphaeroides*. These observations suggest a potential for distinct regulatory inputs at the level of translation initiation and formation of initial peptide bonds. Further survey of potential gene products involved in *R. sphaeroides* protein translation reveals an interesting parallel with archaeal species. The *R. sphaeroides* genome contains a class I lysyl-tRNA synthetase that is found primarily among archaeal species but also included in some bacterial genomes. Surprisingly, the *R. sphaeroides* genome also contains a bacterial class II lysyl-tRNA synthetase, making along with the methanoarchaeon *Methanosarcina acetivorans*, the only microbes to contain genes encoding both classes of lysyl-tRNA synthetase (Galagan et al. 2001). In addition, the *R. sphaeroides* genome is among only a few bacterial genomes to contain a predicted metal dependent hydrolase related to alanyl-tRNA synthetase that is ubiquitous within Archaea.

The novel features of *R. sphaeroides* translational machinery are contrasted by the conventional nature of post-translational machinery. For example, the usual complement of proteases and chaperonins routinely found among bacteria is present, with the exceptions of a prophage-encoded ClpP homolog, and multiple copies of *groEL*. An apparent dearth in protein secretion machinery is perhaps the most surprising facet of the *R. sphaeroides* genome. *R. sphaeroides* has gene products for the classic Sec-dependent protein secretory pathway, but apparently contains an incomplete Tat secretion system and appears to lack type II or IV protein secretion machinery that is prevalent in the bacterial world. Since certain proteins lacking signal sequences are properly processed in *R. sphaeroides*, this finding is quite surprising (Brandner and Donohue 1994). Perhaps the type III secretory machinery encoded by the flagellar loci has additional physiological roles, similar to those characterized in other bacteria (Young et al. 1999). Naturally, true assessment of the complexity of protein expression in *R. sphaeroides* awaits subsequent proteomic, genetic and biochemical dissection of these relevant regulatory and metabolic networks revealed from the genomic sequence.

Clusters of Orthologous Groups, overview

Of the 4152 candidate protein encoding genes found in the genome, 2587 were assigned to Clusters of Ortholgous Groups (COGs). It might be expected that given the size of the replicons ($\sim$3.0 Mb for CI, $\sim$0.9 Mb CII and 0.45 Mb for the sum of the plasmids) the numbers of genes distributed between them may be proportional to their size. That is 68.2%, 21.2% and 10.6% of the genes found would be predicted to be on CI, CII and the plasmids (taken cumulatively), respectively. This approximates what was found, (see Table 4), with perhaps a slight under representation (indicated in the text and table by a minus sign (−)) of genes on CII (under representation of −3.8%) and the plasmids (under representation of −4.3%).

With the exception of translation (COG J) and nucleotide transport and metabolism (COG F), which were not found on the plasmids, all of the other orthologous groups were distributed between the chromosomes and plasmids. The most under represented COGs (by greater than 10%) on CII were those involved in translation, −15.7% (COG J), cell division −17.8% (COG D), cell envelope biogenesis −16.9% (COG M) and cell motility and secretion −10.18%, (COG N). The genes involved in amino acid transport and metabolism (COG E) contained 341 genes (13.2% of all COG assigned genes) the largest number of genes assigned to a COG. This is nearly double that found for the next largest group 208 genes assigned to energy production and conversion (COG C). However, even with this wealth of information, and as noted previously we have not been able to assign a particular role or set of functions which may explain the existence of two chromosomes rather than a single chromosome.

While analyzing the COGs, we noted a few interesting features. For example, with the exception of glutaminyl-tRNA synthetases, all other amino acyl tRNA synthetases are present. In *Bacillus subtilis* (and some other Gram-positive bacteria), glutaminyl-tRNA synthetase is also absent. In these organisms, the $tRNA^{Gln}$ is charged with glutamate which is then converted to glutamine by an amidotransferase (Himmelreich et al. 1996). Interestingly, genes encoding

Table 4. Distribution of Clusters of Orthologous Groups between replicons

COG[a]	CI (68.2%)[b]	CII (21.2%)[c]	Plasmids (10.6%)[b]	Total[c]
Translation (J)	137 (94, +25.8%)	8 (5.5, −15.7%)	0 (0.0, −10.6%)	145 (5.6%)
Transcription (K)	123 (68.7, +0.5%)	33 (18.4, −2.8%)	23 (12.8, −2.2%)	179 (6.9%)
DNA replication and repair (L)	91 (84.2, +16%)	14 (12.96, −8.24%)	3 (2.77, −7.83%)	108 (4.2%)
Cell division (D)	24 (82.7, +14.5%)	1 (3.4, −17.8%)	4 (13.7, +3.1%)	29 (1.2%)
Posttranslational modification, protein turnover, chaperones (O)	76 (82.6, +14.4%)	12 (13.04, −8.16%)	4 (4.3, −6.3%)	92 (3.5%)
Cell envelope biogenesis, outer membrane (M)	101 (88.5, +20.3%)	5 (4.3, −16.9%)	8 (7.0, −3.6%)	114 (4.4%)
Cell motility and secretion (N)	111 (87.4, +19.2%)	14 (11.02, −10.18%)	2 (1.86, −8.74%)	127 (4.9%)
Inorganic ion transport and metabolism (P)	82 (65.6, −2.6%)	31 (24.8, +3.6%)	12 (9.6, −1.0%)	125 (4.9%)
Signal transduction mechanisms (T)	102 (67, −1.2%)	32 (21.0, −0.2%)	17 (11.2, −0.6%)	151 (5.8%)
Energy production and conversion (C)	152 (73.0, −4.8%)	41 (19.7, −1.5%)	15 (7.2, −3.4%)	208 (8.0%)
Carbohydrate transport and metabolism (G)	133 (69.6, −1.4%)	42 (21.9, +0.7%)	16 (8.3, −2.3%)	191 (7.3%)
Amino acid transport and metabolism (E)	220 (64.5, −3.7%)	90 (26.4, +5.2%)	31 (9.1, −1.5%)	341 (13.2%)
Nucleotide transport and metabolism (F)	60 (88.2, +20%)	8 (11.76, −9.44%)	0 (0.0, −10.6%)	68 (2.6%)
Coenzyme metabolism (H)	100 (80.6, +12.4%)	19 (15.3, −5.9%)	5 (4.0, −6.6%)	124 (4.7%)
Lipid metabolism (I)	75 (82.4, +14.2%)	15 (16.4, −4.8%)	1 (1.09, −9.51%)	91 (3.5%)
General function prediction only (R)	283 (77.5, +9.3%)	67 (18.35, −2.85%)	15 (4.1, −6.5%)	365 (14.1%)
Function unknown (S)	104 (80.6, +12.4%)	19 (14.7, −6.5%)	6 (4.65, −5.95%)	129 (5.0%)
Total COGs	1974 (76, +7.8%)	451 (17.4, −3.8%)	162 (6.3, −4.3%)	2587

[a]The genes encoding Clusters of Orthologous Groups of proteins have been assigned to replicons. The single letter code for the COGs is provided after the COG name. More information on COGs can be obtained at: http://www.ncbi.nlm.nih.gov/COG.
[b]The percentage of the genome that is encoded by replicons CI, CII and plasmids. In the columns below, the number of genes that encode COGs on a replicon are presented. In parentheses are the number of genes that were expected to be found based on the proportion of the genome that is encoded in that replicon. This is followed by the percentage over (+) or under (−) representation of that COG on the replicon. For example, on CI which encodes 68.2% of the genome, 137 genes for translation were found. It would be expected that only 94 translation genes would be found. Therefore, translation is over represented on CI by +25.8%.
[c]The total number of genes (deduced proteins) in the genome for a particular COG. In parenthesis is the percentage value that specific COG forms of the total COGs (2587).

both subunits of Glu-tRNAGln amidotransferase were found in *R. sphaeroides*, suggesting that a similar $tRNA^{Gln}$ charging mechanism may also occur.

Surprisingly, we found two copies of *dnaE* that encodes the DNA polymerase III, alpha chain. One copy is on CI, the other copy resides on CII. It would be interesting to know whether both copies are functional and if they are both capable of replicating each of the chromosomes or if the copy on CI is specific for the replication of CI.

Previous work had shown that a variety of genes encoding the enzymes for tryptophan biosynthesis are distributed between the two chromosomes (Mackenzie et al. 1999). The final step in the tryptophan pathway requires the heterodimeric enzyme tryptophan synthetase. In *R. sphaeroides*, the alpha and beta subunits of this enzyme are encoded on different chromosomes, *trpA* on CI and *trpB* on CII. We were, therefore, somewhat surprised to find a second copy of *trpB* on CI (E = 2.00E-97). Previous work clearly demonstrated that under all conditions tested, a disruption of *trpB* on CII resulted in an auxotrophic phenotype. We therefore wonder if this second *trpB* is functional under some rare condition or if it could be a non-functional pseudogene.

We were curious to see if other genes for amino acid biosynthesis were distributed around the genome like those of the tryptophan pathway. We found that all other amino acid biosynthetic enzymes are distributed between the two chromosomes. Furthermore, in each of the biosynthetic pathways, except for lysine biosynthesis, some of the enzymes are also encoded by plamids (pRS241a and pRS241b). It should be noted, however, that in most cases a second or third copy of a plasmid encoded gene is generally found on one or both of the two chromosomes. Interestingly all of the genes involved in proline biosynthesis, i.e. those that encode gamma-

glutamyl phosphate reductase, glutamate 5-kinase and pyrroline-5-carboxylate reductase, reside on CII.

Encoding amino acid biosynthetic genes on a plasmid could be a way of regulating expression through copy number control. Such a scenario has been observed in the aphid symbiont *Buchnera* sp. (Lai et al. 1994). A more radical possibility is that some of these plasmids are in a transitional state and evolving into small chromosomes.

Unlike amino acid biosynthesis, only the chromosomes encode genes for nucleotide transport and metabolism, i.e. purine, pyrimidine biosynthesis and salvage (COG F). Two essential genes, *cmk*, which encodes cytidylate kinase and *deoC*, which encodes deoxyribose-phosphate aldolase are only found in single copies on CII. All other CII genes in COG F are duplicated and distributed between the two chromosomes.

In conclusion, it appears that during evolution the genome has delegated responsibilities between the chromosomes for day to day house keeping functions and at the same time created a level of gene duplication that resulted in increased gene redundancy. It is interesting to consider the possibility that the evolution of the eukaryotic lifestyle i.e. diploidy and multiple chromosomes may have had humble origins in an ancestral *R. sphaeroides* like cell.

Significant genomic findings with respect to terminal oxidases

The complexity of oxidative phosphorylation in bacteria is commonly enhanced by the presence of more than one terminal oxidase. The genome of *R. sphaeroides* 2.4.1 encodes no less than five terminal oxidases, suggesting a remarkable degree of physiologic versatility. Single copies of the genes for the four structural subunits of the mitochondrial-like cytochrome aa_3-type cytochrome *c* oxidase (*coxI, coxII, coxIII* and *coxIV*) are organized into three separate loci. In addition to these genes, the 2.4.1 genome contains clear homologs of five genes encoding proteins required for the assembly of cytochrome oxidase in yeast and humans (*cox10, cox11, cox15, sco1, surf-1*). The aa_3-type oxidase is expressed in vigorously aerated cells. A more ancient member of the heme-Cu oxidase family, a cytochrome cbb_3-type cytochrome *c* oxidase, is encoded by a single operon that contains the genes for all of its structural subunits (*ccoN, ccoO, ccoP* and *ccoQ*). Immediately downstream of this operon is another that encodes the genes for three proteins necessary for the expression/assembly of the cbb_3-type oxidase (*rdxH, rdxI, rdxS*) (Koch et al. 2000; Roh and Kaplan 2000). The cbb_3-type oxidase is present in both aerated cells and those grown under microaerophilic conditions. Such constitutive expression appears to be necessary for the role that this enzyme plays as a redox sensor for the cell; electron flow through cytochrome cbb_3 modulates the *prrA*/*prrB* regulon, presumably *via ccoQ* (Oh and Kaplan 2000). A quinol oxidase belonging to the *bd*-type family of terminal oxidases is encoded by two genes (*qxtA, qxtB*) that are highly similar to genes encoding the *bd*-type oxidase of *Pseudomonas aeruginosa* (Mouncey et al. 2000). The expression of this oxidase in 2.4.1 cells is enhanced by low O_2 and by genetic manipulations that interrupt the flow of electrons to the cytochrome *c* oxidases.

Genes for two unusual terminal oxidases, neither of which have been previously identified by biochemical or spectroscopic analyses, have been revealed by analysis of the 2.4.1 genome. The first of these is a putative cytochrome caa_3-type oxidase since the gene for its subunit II predicts that it should contain both Cu_A and a cytochrome *c*. The presence of a caa_3-type enzyme in *R. sphaeroides* is remarkable, since these enzymes are generally found in gram-positive bacteria that lack soluble, periplasmic, *c*-type cytochromes. Even more remarkable is the predicted subunit I for this enzyme. The gene predicts that this polypeptide contains a fusion of a typical subunit I domain (the large subunit that contains the heme-Cu active site) to an integral membrane protein that contains seven transmembrane helices. The predicted organization of this latter domain is similar to subunit III of the Cu_A-type and quinol oxidases of the heme-Cu oxidase family and its predicted amino acid sequence shows weak but recognizable homology to subunit III of the aa_3-type oxidase of *R. sphaeroides*. Such gene fusions (both the subunit II-cytochrome *c* and the subunit I–subunit III fusion) are reminiscent to what is found in bacteria that grow in extreme environments. Two genes for a fifth potential oxidase (*qoxA, qoxB*) appear to encode a heme-Cu oxidase with the ability to oxidize quinol. Like the caa_3-type oxidase mentioned above, the gene for the largest subunit of this enzyme appears to be a fusion of the genes for a heme-Cu-type subunit I with an aa_3-type subunit III. The notion that the *Qox* oxidase accepts electrons from quinol is derived from the observation that its predicted subunit

II is similar to those of other quinol oxidases in the heme-Cu family (Mouncey et al. 2000).

Significant genomic findings with respect to nitrogen oxide reductases

The genome of 2.4.1 contains genes for a single nitrate reductase. The deduced sequence indicates it is a periplasmic, dissimilatory enzyme. There is no evidence for an assimilatory nitrate reductase. The genes for the nitrate reductase are located on plasmid pRS241c. This is the only terminal oxidoreductase not encoded on either CI or CII.

The genome lacks a gene for either assimilatory or dissimilatory nitrite reductase. Most strains of *R. sphaeroides* lack nitrite reductase and as a consequence accumulate nitrite when cultured in medium containing nitrate. However, some strains are capable of dissimilatory nitrite reduction and have been shown to encode a copper-containing nitrite reductase that reduces nitrite to nitric oxide (Zumft 1997). Reduction of nitrite to nitric oxide is part of the denitrification pathway. Interestingly, the *R. sphaeroides* 2.4.1 genome does encode a pseudoazurin on CI. Pseudoazurin has been shown to be an electron donor to nitrite reductase in other denitrifying bacteria (Zumft 1997). A denitrifying strain of *R. sphaeroides*, strain 2.4.3, also encodes pseudoazurin (Jain and Shapleigh, in press). The gene order upstream and downstream of pseudoazurin is identical in both the *R. sphaeroides* 2.4.1 and 2.4.3 strains except that the 2.4.3 strain contains two extra genes downstream of the pseudoazurin gene. These genes are the structural gene for nitrite reductase and a second gene, cotranscribed with the structural gene, but of unknown function. The equivalent region in the *R. sphaeroides* 2.4.1 genome is apparently noncoding. This similarity in gene organization suggests that the parent strain of 2.4.1 and 2.4.3 was capable of nitrite reduction but as the strains diverged 2.4.1 lost the genes for nitrite reductase.

Even though nitrite cannot be reduced by *R. sphaeroides* 2.4.1, the genome contains the entire complement of genes required for nitric oxide reduction. This means that *R. sphaeroides* 2.4.1 can detect and respire nitric oxide but only under conditions where nitric oxide is exogenously produced. The nitric oxide reductase (*nor*) genes are located on CI and lie within the photosynthesis gene cluster. The location of the *nor* genes within the photosynthesis gene cluster is so far unique to *R. sphaeroides*. One gene frequently associated with the *nor* cluster that is missing in *R. sphaeroides* is a gene encoding a cytochrome oxide subunit III-like protein. This gene has not been shown to be essential for nitric oxide reduction but is nearly always found in the region of the genome encoding *nor* in other denitrifiers (Zumft 1997).

Located within the *nor* cluster is the gene encoding *nnrR*, which regulates the nitrite reductase and *nor* genes. NnrR is related to FnrL but unlike FnrL is present in only a single copy in the genome. The putative binding site of NnrR has been determined and shown to be similar to the binding site of FnrL (Tosques et al. 1996). A search of the genome indicates that there is only one other gene with an obvious NnrR-like binding motif within its promoter region. This indicates that the NnrR regulon is small, consisting principally of those genes whose products are directly involved in nitric oxide production and reduction.

Genes for the terminal step in denitrification, nitrous oxide reduction, are not found in the *R. sphaeroides* 2.4.1 genome. Nitrous oxide reductase has been isolated from other denitrifying strains of *R. sphaeroides* (Zumft 1997). It is possible, that, similar to nitrite reductase, the genes for nitrous oxide reduction were also lost as the 2.4.1 strain evolved.

Examination of the maps of the contigs reveals that the genes for denitrification in *R. sphaeroides* 2.4.1, and by extension, all strains of *R. sphaeroides* are scattered about the genome (Schwitner et al. 1998). The *nor* genes and the gene encoding pseudoazurin are both on CI but are $>$ 1 Mb apart. The nitrate reductase genes are not on either chromosome but on a plasmid. This fragmentation of denitrification genes is somewhat unusual. In most other denitrifiers, the genes encoding nitrogen oxide reductases are tightly clustered (Zumft 1997).

Significant genomic findings with respect to tetrapyrrole biosynthesis

Early in the studies of *R. sphaeroides*, the presence of heme, bacteriochlorophyll and vitamin B_{12} was recognized as an important indicator of the metabolic versatility of this organism. The presence of two *a*mino*l*evulinic *a*cid (ALA) synthases already distinguished *R. sphaeroides* from all other prokaryotes. Now, the genomic DNA sequences provide intriguing evidence for the presence of both the C4 and C5 pathways for ALA formation. If true, this would

make *R. sphaeroides* the only organism known to have this capability, other than *Euglena gracilis* in which the two pathways are strictly compartmentalized (see Jordan 1991 for details). The DNA sequences predict a glutamyl-tRNA reductase that is 23% identical to that of *E. coli*, and a glutamate 1-semialdehyde aminotransferase that is 32% identical. The presence of the appropriate tRNA synthetases (two *gltX* genes on CI) completes the enzymatic repertoire required to form ALA from $tRNA^{Glu(UUC)}$.

Another complexity associated with heme formation is the presence of sequences predicting a total of four coproporphyrinogen III oxidases, including three oxygen-independent enzymes and one oxygen-dependent enzyme. On the other hand, DNA sequences for well-described uroporphyrinogen III synthase and protoporphyrinogen IX oxidase genes are not evident. However, there is presently a limited amount of information correlating these enzyme activities with genes at the DNA sequence level, and so the ability to identify the sequences in *R. sphaeroides* is constrained by the lack of clear-cut knowledge about the enzymes from a diverse group of other organisms.

The *R. sphaeroides* DNA sequences predict the presence of a number of vitamin B_{12}-requiring enzymes, including *metH* (2 copies on CI, one on CII; and no *metE* sequences) and B_{12}-dependent ribonucleotide reductase. Perhaps the most striking example of a B_{12} requirement under anaerobic conditions is the *bchE* gene product that has recently been shown to require vitamin B_{12} to catalyze the formation of an intermediate in bacteriochlorophyll synthesis (Gough et al. 2000). Yet *R. sphaeroides* appears to possess several characteristic enzyme components of the aerobic pathway for vitamin B_{12} formation, for an overview of both pathways (see Roth et al. 1996), including sequences predicted to code for CobS, T and N proteins that catalyze cobalt insertion late in the biosynthetic pathway. These findings correspond to the situation present in *Paracoccus denitrificans*, and the question as to how the need for vitamin B_{12} is met under anaerobic conditions has been raised for that organism (Shearer et al. 1999). The answer suggested by Shearer et al. (1999) is that vitamin B_{12} biosynthesis might occur via a composite pathway that would not require oxygen as is true of the anaerobic pathway, but also be insensitive to oxygen, as is true of the aerobic pathway.

The DNA sequences predict that *R. sphaeroides* lacks the multifunctional CysG enzyme that is present in other Gram-negative bacteria deploying the anaerobic pathway for vitamin B_{12} formation, and which is also responsible for siroheme biosynthesis (Roth et al. 1996). On the other hand, there are two putative *cobA* genes, both of which could code for proteins having the methyltransferase activity of CysG, analogous to the aerobic pathway. These findings also indicate that *R. sphaeroides* is unable to make siroheme. However, the DNA sequences also predict the presence of a siroheme-dependent sulfite reductase, suggesting that this cofactor may be synthesized by a novel pathway in *R. sphaeroides*.

With respect to regulation of tetrapyrrole biosynthesis genes, it has already been determined that *hemA*, *hemN* and *Z*, and *bchE* have upstream FNR consensus-like sequences which correlate with the known participation of the *R. sphaeroides* FNR homologue *fnrL* in regulation of their expression (Zeilstra-Ryalls and Kaplan 1995; Oh et al. 2000). These remain the only tetrapyrrole genes for which upstream FNR consensus-like sequences are readily apparent. However, at least 5 members of the CRP-FNR family of proteins are predicted to be encoded by the *R. sphaeroides* genome, and the conservation of amino acid sequences in their DNA-binding domains suggests previously unsuspected complexity with respect to regulation of expression of all genes having upstream FNR consensus-like sequences. Other potential regulators of tetrapyrrole biosynthesis gene expression predicted by the *R. sphaeroides* DNA sequences include homologs of both Fur and Irr corresponding to the situation in *Bradyrhizobium japonicum* (Hamza et al. 2000). Undoubtedly there are unique features to their activities in *R. sphaeroides* since, for example, the HRM (heme regulatory motif; (Qi et al. 1999)) of the putative *irr* gene product is absent.

Significant genomic findings with respect to chemosensory genes

The majority of information on the chemotactic behavior of *R. sphaeroides* comes from strain WS8, as this is consistently more motile than strain 2.4.1. However, comparison of the genome sequences from 2.4.1 and those of WS8 show that the organization of the genes is identical, the protein sequence ~100% identical and the gene sequence about 98.5% identical. We already had data showing that *R. sphaeroides* has multiple copies of many of the chemosensory genes, located primarily at two loci, one on contig 132 and the other on contig 159 of CI (Hamblin et al. 1997). The gen-

ome sequence identified a third locus on contig 162 associated with many flagellar encoding genes and encoding more copies of predicted chemosensory genes. Overall there are genes encoding 4 copies of CheA, 4 copies of CheW, 2 copies of CheB, 3 copies of CheR, 7 putative CheY genes and 12 genes encoding chemoreceptors (Table 3). Four putative chemoreceptor genes do not have regions that would encode membrane spanning domains and are, therefore, probably encoding soluble receptors (called *T*ransducer *L*ike *P*roteins, Tlps). There is one gene encoding a Tlp with each *che* gene cluster and one with genes encoding the flagellar motor proteins, *motAB*. Only 2 of the 9 genes encoding 'classical' membrane spanning chemoreceptors, MCPs, are located with the *che* genes. The other *mcp* genes are scattered around the genome, with 6 *mcp* genes located on CII. One *mcp* on CII is downstream of the only *che*gene found on CII, a gene encoding a CheY (Shah et al. 2000b; Wadhams et al. 2000).

In addition to multiple copies of the characteristic *che* genes found in other species, there is a gene predicted to encode a membrane bound PAS domain containing CheB-CheR-HPK fusion protein on contig 181 of CI. Genes encoding similar fusions are present in the genomes of *Rhodopseudomonas palustris* and *Sinorhizobum meliloti*, but not *R. capsulatus*. Indeed comparison of the putative chemosensory genes and the arrangements of the genes with other α-subgroup species suggests that the chemosensory organization is most closely related to that of *R. palustris*, which also has 3 loci encoding the multiple copies of Che proteins. *R. capsulatus* only has clusters of genes equivalent to the 2 clusters found on contig 132 and contig 159 and completely lacks the third set of genes and has a CheB–CheR encoding fusion without the kinase.

Mutagenesis studies in *R. sphaeroides* combined with expression of the individual genes in either wild type or mutant strains of *E. coli* deleted for the appropriate *che* gene suggest that the gene products do function in chemotaxis (Shah et al. 2000b). Deletion studies also show that the duplicate genes on contig 159 and contig 162 are not redundant (unpublished data).

Comparison to *Rhodobacter capsulatus* and Genbank

The deduced amino acid sequences of 4,113 predicted *R. sphaeroides* proteins were examined to assess the extent to which their homologs were present in other genomes i.e., the GENBANK database and specifically to the closely related species, *R. capsulatus*. BLASTP was used for both comparisons, in the latter case, the sequences were compared to the ERGO database kindly provided by Integrated Genomics (www.integratedgenomics.com). The cutoff score used to assess matching was a bit score above 100 and a non-stringent Expect (E) value of 10^{-5}. The results of this comparison are summarized in Table 5.

For the two chromosomes and plasmids pRS241a and b, ~80% of the predicted proteins had counterparts in the GENBANK database. Plasmids pRS241c and d had fewer counterparts, ~50%, with a slightly higher number of matches being encoded on pRS241e (68.2%). Interestingly, this finding correlates with the differences found in the relative usage of di- and trinucleotides on the different replicons.

It had been known for some time that the genomes of *R. sphaeroides* and *R. capsulatus* were very different in terms of their genome architecture, the former having two chromosomes, the latter one. It was also known that their methods of gene regulation, especially the regulation of the genes of the photosynthesis gene cluster were very different. However, ribosomal RNA analysis suggested that these organisms were 'kissing cousins' (Woese et al. 1984a; Woese 1987) and indeed they share many other aspects of their physiology. Therefore, it was somewhat surprising to find that only 64% of *R. sphaeroides* genes could be found in *R. capsulatus*. This difference is no doubt in part due to the 18% smaller genome size of *R. capsulatus*, but that in itself suggests that these organisms are very different in their genomic composition and perhaps not quite as closely related as was once thought. This was further emphasized when we looked at gene order and found that only in a few cases the order of the genes were congruent in both organisms.

It is also of interest to note that many of the genes found in *R. sphaeroides*, but not found in *R. capsulatus*, are duplicated within *R. sphaeroides*. Indeed, it is in part the level of gene duplication in *R. sphaeroides* that results in its genome being much larger in size than that of *R. capsulatus*.

Of equal note is that 707 *R. sphaeroides* genes (~17% of total genes) are not present in the GENBANK or ERGO databases and therefore are currently unique to *R. sphaeroides*. As more genomes are sequenced there is a clearly narrowing gap between unique genes and 'previously discovered' genes (in *R. sphaeroides* ~83%). Perhaps it will not be long before

Table 5. Comparison of the genomes of *R. sphaeroides* to Genbank and ERGO

	Replicon							
	CI	CII	pRS241a	pRS241b	pRS241c	pRS241d	pRS241e	Total genes
Number of genes compared	2725	801	99	221	112	111	44	4113
Genes found in Genbank	2267	600	80	189	57	64	30	3287
	(83.2%)	(74.9%)	(80.8%)	(85.5%)	(50.9%)	(57.7%)	(68.2%)	(80.0%)
Genes found in ERGO	1947	467	60	59	34	38	18	2622
	(71.4%)	(58.3%)	(60.6%)	(71.9%)	(30.4%)	(34.2%)	(45.9%)	(64%)
Genes not found in either	369	177	18	28	54	47	14	707
Genbank *or* ERGO	(13.5%)	(22.1%)	(18.2%)	(12.7%)	(48.2%)	(42.3%)	(31.8%)	(17.2%)

Shown are the number of genes from each replicon that were compared to Genbank and ERGO (the *R. capsulatus* database). The numbers of genes from the relevant replicon that were found (or not found) in the database(s) are also presented as a number and a percentage of the total genes on the replicon. The expect value for the match cutoff was E-05 with a bit score of greater than 100.

a genome is sequenced and few, if any, new orthologs are found?

We also noted that less than 35% of the genes on plasmids pRS241b and c matched *R. capsulatus* genes. This is particularly low compared to the other large replicons where ∼60–70% of genes give matches. Interestingly, many of the matching genes (∼40%) encoded on plasmids pRS241a and b match to chromosomally encoded genes in *R. capsulatus*.

Size does not matter?

For a long time, it was believed that bacteria had a single, circular chromosome. This notion was originally shown to be erroneous with the discovery of two chromosomes in *R. sphaeroides* and more recently in other bacteria e.g. *Vibrio cholerae* (Heidelberg et al. 2000). For a while, chromosomes were big and plasmids were small. Therefore, anything smaller than the chromosome was considered, by default, a plasmid. Earlier, we suggested that what was more important than size was the functions that a replicon encodes. If it encodes essential functions, then even though it is smaller than the largest replicon it can still be called a chromosome. How far can we extend this argument?

Four of the small replicons in *R. sphaeroides* are ∼100 kb in size and they all encode functions that would be considered plasmid in nature e.g. genes involved in heavy metal transport, antibiotic resistance and opine catabolism (similar to those found on the Ti plasmid of *Agrobacterium tumefaciens*). However, pRS241a and b also carry functions that may be considered chromosomal, e.g. genes encoding the beta and epsilon subunits of an F-type ATPase, pyruvate dehydrogenase, acetolactate synthase, fumarate hydratase and the *mrr* (adenine) restriction system. With the exception of *mrr* all other chromosomal type functions are duplicated on CI or CII. At this stage, we do not know if the plasmid, chromosomal or both copies of the genes are functional. Can we call pRS241 a and b chromosomes? The jury is still out, but we cannot envisage any barrier to having a 100 kb chromosome encoding the only copy of an essential gene. Provided such a chromosome is maintained and inherited then genomic integrity should continue.

What evolutionary advantage the possession of multiple copies of genes and multiple chromosomes confers on this bacterium is unclear. In the eukaryotic world, multiple copies of genes and multiple chromosomes are the norm. Perhaps *R. sphaeroides* really can be considered a living fossil, one of the early bridges between the prokaryotic and eukaryotic world.

Cherry picking the genome

Every genome has its surprises (it is a reason we sequence them!) and *R. sphaeroides* was not an exception. One such surprise, of special interest to those concerned with photosynthesis, is the presence of urease genes within the photosynthesis gene cluster. This has been described in detail elsewhere (Choudhary and Kaplan 2000).

Two other sets of genes of note are the presence of the circadian clock genes *kaiC1* and *kaiB1* and gas vesicle operons. The existence of clock genes has been reported in cyanobacteria but we believe this is the first report in the purple non-sulfur bacteria. Interestingly the genes are encoded on plasmid pRS241b. Given

that the earth goes around the sun, and this is a photosynthetic organism, time keeping would appear to be a necessity. Perhaps this replicon is the timepiece of the cell and encodes the most important function of all. The gas vesicle genes in many respects are equally important to photosynthetic bacteria. They can raise the organism into the light and lower it when the light becomes too strong. We have found ample sequence evidence that *R. sphaeroides* possess such genes (*gvp* operons on CII). With the genome nearly complete, we are undertaking array analysis. Will the photosynthesis, circadian and gas vesicle genes be linked through expression? We await the results.

As the sequencing of the genome moves towards closure, we look forward to the use of this reagent by the wider scientific community. We have learned a lot from *R. sphaeroides,* but to paraphrase Claire Fraser "The more genomes we sequence the less we understand about biology." So. . . let the biology begin!

Acknowledgements

As a group, we would like to thank Antonius Suwanto who started us down the path of sequencing *R. sphaeroides*. He is also the first scientist in history to demonstrate that bacteria have two chromosomes (Suwanto and Kaplan 1989a, b). C. M. would like to acknowledge his friend and colleague Jesus Eraso for his helpful criticism of this work. This work was supported by: NIH grant GM15590 to S.K.; NIH grant GM56824 to J.P.H.; National Institutes of General Medical Sciences grant GM37509 and US Department of Energy, Microbial Cell Program grant ER63232-1018220-0007203 to T.J.D.; J.N was supported by Biotechnology Training Grant, GM08349, from the National Institutes of General Medical Sciences. J.P.S was supported by Department of Energy grant 95ER20206. J.P.A and R.E.S would like to thank the UK BBSRC Royal Society & Leverhulme Trust for funding their research. R.D.B. would like to thank John Cairo and Melissa Meland, who were supported by Abbott Laboratories Fund. The work of J.Z-R was supported by NSF grant #9805556 and the Michigan Life Sciences Corridor Fund.

References

Allen LN and Hanson RS (1985) Construction of broad-host range cosmid cloning vectors: Identification of genes necessary for growth of *Methylobacterium organophilum* on methanol. J Bacteriol 161: 955–962

Armitage JP and Schmitt R (1997) Bacterial chemotaxis: *Rhodobacter sphaeroides* and *Sinorhizobium meliloti* – variations on a theme? Microbiology 143: 3671–3682

Barber RD and Donohue TJ (1998) Function of a glutathione-dependent formaldehyde dehydrogenase in *Rhodobacter sphaeroides*, formaldehyde oxidation and assimilation. Biochemistry 37: 530–537

Brandner JP and Donohue TJ (1994) The *Rhodobacter sphaeroides* cytochrome *c*2 signal peptide is not necessary for export and heme attachment. J Bacteriol 176: 602–609

Busby S and Ebright RH (1994) Promoter structure, promoter recognition and transcription activation in prokaryotes. Cell 79: 743–746

Choudhary M and Kaplan S (2000) DNA sequence analysis of the photosynthesis region of *Rhodobacter sphaeroides* 2.4.1. Nucleic Acids Res 28: 862–867

Choudhary M, Mackenzie C, Nereng KS, Sodergren EJ, Weinstock G and Kaplan S (1994) Multiple chromosomes in bacteria: Structure and functions of chromosome II of *Rhodobacter sphaeroides* $2.4.1^{T}$. J Bacteriol 176: 7694–7702

Choudhary M, Mackenzie C, Nereng K, Sodergren E, Weinstock GM and Kaplan S (1997) Low-resolution sequencing of *Rhodobacter sphaeroides* $2.4.1^{T}$: Chromosome II is a true chromosome. Microbiology 143: 3085–99

Choudhary M, Mackenzie C, Mouncey NJ and Kaplan S (1999) RsGDB, the *Rhodobacter sphaeroides* Genome Database. Nucleic Acids Res 27: 61–62

Clayton RK and Sistrom WR (1978) The Photosynthetic Bacteria. Plenum Press, New York

Dryden S and Kaplan S (1990) Localization and structural analysis of the ribosomal RNA operons of *Rhodobacter sphaeroides*. Nucleic Acids Res 18: 7267–7277

Dryden SC and Kaplan S (1993) Identification of cis-acting regulatory regions upstream of the rRNA operons of *Rhodobacter sphaeroides*. J Bacteriol 175: 6392–6402

Fornari CS, Watkins M and Kaplan S (1984) Plasmid distribution and analysis in *Rhodopseudomonas sphaeroides*. Plasmid 11: 39–47

Galagan JE, Nusbaum C, Roy A, Endrizzi M, Macdonald P, FitzHugh W, Calvo S, Engels R, Smirnov S, Atnoor D, Brown A, Allen N, Naylor J, Stang-Thomann N, DeArellano K, Johnson R, Linton L, McEwan P, McKernan K, Talamas J, Tirrell A, Ye W, Zimmer A, Barber R, Cann I, Graham DE, Grahame DA, Guss A, Hedderich R, Ingram-Smith C, Kuettner HC, Krzycki JA, Leigh JA, Li W, Liu J, Mukhopadhyay B, Reeve JN, Smith K, Springer T, Umayam LA, White O, White RH, Conway de Macario E, Ferry JG, Jarrell KF, Jing H, Macario AJL, Paulsen I, Pritchett M, Sowers KR, Swanson RV, Zinder SH, Lander E, Metcalf WW and Birren B (2001) The complete genome sequence *of Methanosarcina acetivor*ans C2A. Nature (submitted)

Goodfellow IG, Pollitt CE and Sockett RE (1996) Cloning of the *fliI* gene from *Rhodobacter sphaeroides* WS8 by analysis of a transposon mutant with impaired motility. FEMS Microbiol Lett 142: 111–116

Gough S, Petersen B and Duus J (2000) Anaerobic chlorophyll isocyclic ring formation in *Rhodobacter capsulatus* requires a cobalamin cofactor. Proc Natl Acad Sci USA 97: 6908–6913

Gruber TM and Bryant DA (1997) Molecular systematic studies of eubacteria, using σ^{70}-type sigma factors of group 1 and group 2. J Bacteriol 179: 1734–1747

Hallenbeck PL, Lerchen R, Hessler P and Kaplan S (1990a) Phosphoribulokinase activity and the regulation of CO_2 fixation

critical for photosynthetic growth of *Rhodobacter sphaeroides*. J Bacteriol 172: 1749–1761

Hallenbeck PL, Lerchen R, Hessler P and Kaplan S (1990b) The role of CFXA, CFXB, and external electron acceptors in the regulation of ribulose 1,5-bisphosphate carboxylase/oxygenase expression in *Rhodobacter sphaeroides*. J Bacteriol 172: 1736–1748

Hamblin PA, Maguire BA, Grishanin RN and Armitage JP (1997) Evidence for two chemosensory pathways in *Rhodobacter sphaeroides*. Mol Microbiol 26: 1083–1096

Hamza I, Qi Z, King N and O'Brian M (2000) Fur-independent regulation of iron metabolism by Irr in *Bradyrhizobium japonicum*. Microbiology 146: 669–676

Heidelberg JF, Eisen JA, Nelson WC, Clayton RA, Gwinn ML, Dodson RJ, Haft DH, 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 and White O (2000) DNA sequence of both chromosomes of the cholera pathogen *Vibrio cholerae*. Nature 406: 477–483

Heusterspreute M, Thai VH, Emery S, Tournis-Gamble S, Kennedy N and Davidson J (1985) Vectors with restriction site banks. IV pJRD184, a 3793-bp plasmid vector having 43 unique cloning sites. Gene 39: 299–304

Himmelreich R, Hilbert H, Plagens H, Pirkl E, Li BC and Herrmann R (1996) Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. Nucleic Acids Res 24: 4420–4449

Ielpi L, Dylan T, Ditta GS, Helinski DR and Stanfield SW (1990) The *ndvB* locus of *Rhizobium meliloti* encodes a 319-kDa protein involved in the production of beta-(1-2)-glucan. J Biol Chem 265: 2843–2851

Jain R and Shapleigh JP 2001 Characterization of *nirV* and a gene encoding a novel pseudoazurin in*Rhodobacter sphaeroides* 2.4.3. Microbiology 47: 2505–2515

Jordan P (ed) (1991) Biosynthesis of Tetrapyrroles. Elsevier, Amsterdam

Kansy JW and Kaplan S (1989) Purification, characterization and transcriptional analyses of RNA polymerases from *Rhodobacter sphaeroides* cells grown chemoheterotrophically and photoheterotrophically. J Biol Chem 264: 13751–13759

Karls RK, Jin DJ and Donohue TJ (1993) Transcription properties of RNA polymerase holoenzymes isolated from the purple nonsulfur bacterium *Rhodobacter sphaeroides*. J Bacteriol 175: 7629–7638

Karls RK, Brooks J, Rossmeissl P, Luedke J and Donohue TJ (1998) Metabolic roles of a *Rhodobacter sphaeroides* member of the σ^{32} family. J Bacteriol 180: 10–19

Karls RK, Wolf JR and Donohue TJ (1999) Activation of the *cycA* P2 promoter for the *Rhodobacter sphaeroides* cytochrome c_2 gene by the photosynthesis response regulator. Mol Microbiol 34: 822–835

Koch H-G, Winterstein C, Saribas AS, Alben JO and Daldal F, (2000) Roles of the *ccoGHIS* gene products in the biogenesis of the cbb_3-type cytochrome *c* oxidase. J Mol Biol 297: 49–65

Lai CY, Baumann L and Baumann P (1994) Amplification of *trpEG*: Adaptation of *Buchnera aphidicola* to an endosymbiotic association with aphids. Proc Natl Acad Sci USA 91: 3819–3823

Lee WT, Terlesky KC and Tabita FR (1997) Cloning and characterization of two *groESL* operons of *Rhodobacter sphaeroides*: Transcriptional regulation of the heat-induced *groESL* operon. J Bacteriol 179: 487–495

Lonetto M, Gribskov M and Gross CA (1992) The σ^{70} family: Sequence conservation and evolutionary relationships. J Bacteriol 174: 3843–3849

Lonetto MA, Brown KL, Rudd KE and Buttner MJ (1994) Analysis of the *Streptomyces coelicolor sigE* gene reveals the existence of a subfamily of eubacterial RNA polymerase sigma factors involved in the regulation of extracytoplasmic functions. Proc Natl Acad Sci USA 91: 7573–7577

MacGregor BJ, Karls RK and Donohue TJ (1998) Transcription of the *Rhodobacter sphaeroides cycA* P1 promoter by alternate RNA polymerase holoenzymes. J Bacteriol 180: 1–9

Mackenzie C, Chidambaram M, Sodergren EJ, Kaplan S and Weinstock G (1995) DNA repair mutants of *Rhodobacter sphaeroides*. J Bacteriol 177: 3027–3035

Mackenzie C, Simmons AE and Kaplan S (1999) Multiple chromosomes in bacteria. The yin and yang of *trp* gene localization in *Rhodobacter sphaeroides* 2.4.1. Genetics 153: 525–38

Meeks JC, Elhai J, Thiel T, Potts M, Larimer F, Lamerdin J, Predki P and Atlas R (2001) An overview of the genome of *Nostoc punctiforme*, a multicellular, symbiotic cyanobacterium. Photosynth Res 70: 85–106 (this issue).

Meijer WG and Tabita R (1992) Isolation and characterization of the *nifUSVW-rpoN* gene cluster from *Rhodobacter sphaeroides*. J Bacteriol 174: 3855–3866

Moore MD and Kaplan S (1992) Identification of intrinsic high-level resistance to rare-earth oxides and oxyanions in members of the class *Proteobacteria*: Characterization of tellurite, selenite and rhodium sesquioxide reduction in *Rhodobacter sphaeroides*. J Bacteriol 174: 1505–1514

Mouncey NJ, Choudhary M and Kaplan S (1997) Characterization of genes encoding dimethyl sulfoxide reductase of *Rhodobacter sphaeroides* 2.4.1^{T}: An essential metabolic gene function encoded on chromosome II. J Bacteriol 179: 7617–7624

Mouncey NJ, Gak E, Choudhary M, Oh J-I and Kaplan S (2000) Respiratory pathways of *Rhodobacter sphaeroides* 2.4.1: Identification and characterization of genes encoding quinol oxidases. FEMS Microbiol Lett 192: 205–210

Neidle E and Kaplan S (1992) *Rhodobacter sphaeroides rdxA*, a Homolog of *Rhizobium meliloti fixG*, Encodes a Membrane Protein Which May Bind Cytoplasmic [4Fe–4S] Clusters. J Bacteriol 74: 6444–6454

Neidle E and Kaplan S (1993a) Expression of the *Rhodobacter sphaeroides hemA* and *hemT* genes encoding two aminolevulinic acid synthase isozymes. J Bacteriol 175: 2292–2303

Neidle E and Kaplan S (1993b) 5-Aminolevulinic acid availability and control of spectral complex formation in HemA and HemT mutants of *Rhodobacter sphaeroides*. J Bacteriol 175: 2304–2313

Newman J, Anthony J and Donohue TJ (2001) The importance of zinc coordination for ChrR function as an anti-sigma factor. J Mol Biol 313: 485–499

Newman JD, Falkowski MJ, Schilke BA, Anthony LC and Donohue TJ (1999) The *Rhodobacter sphaeroides* ECF sigma factor, σ^{E}, and the target promoters *cyc*A P3 and *rpo*E P1. J Mol Biol 294: 307–320

Oh J-I, Eraso J and Kaplan S (2000) Interacting regulatory circuits involved in orderly control of photosynthesis gene expression in *Rhodobacter sphaeroides* 2.4.1. J Bacteriol 182: 3081–3087

Oh J-I and Kaplan S (2000) Redox signaling:globalization of gene expression. EMBO J 19: 4237–4247

Poggio S, Aguilar C, Osorio A, Gonzalez–Pedrajo B, Dreyfus G and Camarena L (2000) σ^{54} promoters control expression of genes encoding the hook and basal body complex in *Rhodobacter sphaeroide*s. J Bacteriol 182: 5785–5792

Qi Z, Hamza I and O'Brian M (1999) Heme is an effector molecule for iron-dependent degradation of the bacterial iron response regulator (Irr) protein. Proc Natl Acad Sci USA 96: 13056–13061

Roh JH and Kaplan S (2000) Genetic and phenotypic analyses of the *rdx* locus of *Rhodobacter sphaeroides* 2.4.1. J Bacteriol 182: 3475–3481

Roth J, Lawrence J and Bobik T (1996) Cobalamin (Coenzyme B_{12}): Synthesis and biological significance. Ann Rev Microbiol 50: 137–181

Schwitner C, Sabaty M, Berna B, Cahors S and Richaud P (1998) Plasmid content and localization of the genes encoding the denitrification enzymes in two strains of *Rhodobacter sphaeroides*. FEMS Microbiol Lett 165: 313–321

Shah DSH, Perehinec T, Stevens SM, Aizawa S-I and Sockett RE (2000a) The flagellar filament of *Rhodobacter sphaeroides*: PH-induced polymorphic transitions and analysis of the *fliC* gene. J Bacteriol 182: 5218–5224

Shah DSH, Porter SL, Harris DC, Wadhams GH, Hamblin PA and Armitage JP (2000b) Identification of a fourth *cheY* gene in *Rhodobacter sphaeroides* and inter-species interaction within the bacterial chemotaxis signal transduction pathway. Mol Microbiol 35: 101–112

Shearer N, Hinsley A, Spanning RV and Spiro S (1999) Anaerobic growth of *Paracoccus denitrificans* requires cobalamin: Characterization of *cobK* and *cobJ* genes. J Bacteriol 181: 6907–6913

Sockett RE, Goodfellow IG, Gunther G, Edge MJ and Shah DSH (1999) Properties of *Rhodobacter sphaeroides* flagellar motor proteins. In: Peschek GA, Loffelhardt W and Schmretterer G (eds) The Phototrophic Prokaryotes, pp 693–699. Plenum, New York

Suwanto A and Kaplan S (1989a) Physical and genetic mapping of the *Rhodobacter sphaeroides* 2.4.1 genome: Genome size, fragment identification and gene localization. J Bacteriol 171: 5840–5849

Suwanto A and Kaplan S (1989b) Physical and genetic mapping of the *Rhodobacter sphaeroides* 2.4.1 genome: Presence of two unique circular chromosomes. J Bacteriol 171: 5850–5859

Suwanto A and Kaplan S (1992) Chromosome transfer in *Rhodobacter sphaeroides*: Hfr formation and genetic evidence for two unique circular chromosomes. J Bacteriol 174: 1135–1145

Tosques IE, Shi J and Shapleigh JP (1996) Cloning and characterization of *nnrR*, whose product is required for the expression of proteins involved in nitric oxide metabolism in *Rhodobacter sphaeroides* 2.4.3. J Bacteriol 178: 4958–4964

van Neil CB (1944) The culture, general physiology, morphology and classification of the non-sulfur purple and brown bacteria. Bacteriol Rev 8: 1

Wadhams GH, Martin AC and Armitage JP (2000) Identification and localisation of a methyl-accepting chemotaxis protein in *Rhodobacter sphaeroides.* Mol Microbiol 36: 1222–1233

Wilson K (1989) Preparation of genomic DNA from bacteria. In: Ausubel FM, Brent R, Kingston RE, Moore DD, Seidman JG, Smith JA and Struhl K (eds) Current Protocols in Molecular Biology, pp 24.1–24.5. Wiley Interscience, New York

Woese CR (1987) Bacterial evolution. Microbiol Rev 51: 221–271

Woese CR, Stackebrandt E, Weisburg WG, Paster BJ, Madigan MT, Fowler VJ, Hahn CM, Blanz P, Gupta R, Nealson KH and Fox GE (1984) The phylogeny of the purple bacteria: the alpha subdivision. Sys Appl Microbiol 5: 315–326

Yang D, Oyaizu Y, Oyaizu H, Olsen GJ and Woese CR (1985) Mitochondrial origins. Proc Nat Acad Sci USA 82: 4443–4447

Yeliseev AA and Kaplan S (1995) A sensory transducer homologous to the mammalian peripheral-type benzodiazepine receptor regulates photosynthetic membrane complex formation in *Rhodobacter sphaeroides* 2.4.1. J Biol Chem 270: 21167–21175

Young GM, Schmiel DH and Miller VL (1999) A new pathway for the secretion of virulence factors by bacteria: The flagellar export apparatus functions as a protein-secretion system. Proc Natl Acad Sci USA 96: 6456–6461

Zeilstra–Ryalls JH and Kaplan S (1995) Aerobic and anaerobic regulation in *Rhodobacter sphaeroides* 2.4.1: The role of the *fnrL* gene. J Bacteriol 177: 6422–6431

Zumft WG (1997) Cell biology and molecular basis of denitrification. Microbiol Mol Biol Rev 61: 533–616