Identification and Characterization of Gene Duplication in *Rhodobacter sphaeroides* 2.4.1

Anne Peters, Kristen Schroeder, Lin Lin, and Madhusudan Choudhary
Department of Biological Sciences, Sam Houston State University, Huntsville, Texas 77341

**INTRODUCTION**

*Rhodobacter sphaeroides* 2.4.1 belongs to α-3 subgroup of proteobacteria, and its genome consists of two circular chromosomes, Chromosome I (CI, 3.0 Mb), Chromosome II (CII, 0.9 Mb), and five endogenous plasmids as shown in Figure 1. *R. sphaeroides* along with other anoxygenic photosynthetic bacteria have provided a wealth of physiological, biochemical and biophysical information which has contributed to the understanding of energy balance, physiological diversity, and metabolic capabilities of these organisms (2). Members of this subgroup are considered to be progenitors of present day mitochondria, and also represent an earlier assemblage of organisms from which chloroplasts evolved, later giving rise to the chloroplast. The genome of the *R. sphaeroides* 2.4.1 has been sequenced, assembled and fully annotated. It provides an ideal prokaryotic model system for the study of genome complexity since it possesses multiple chromosomes and has been studied at the molecular level. It was the first organism whose genome was fully sequenced (1).

**RESULTS AND DISCUSSION**

A total of 1440 genes (34% of the total protein coding gene functions) were found to be present in varying number of copies in the genome of *R. sphaeroides*, as shown in Figure 1. Of 1440 genes, 7 times, 4 copies were present in 55 genes, 3 copies in 98 genes and 2 copies in 217 genes. This result indicates that the majority of gene duplications (~80%) were chromosomal. There were 191 and 26 gene duplications, where all copies were dispersed within Chromosome I (~65%) and Chromosome II (~35%), respectively. In addition, 153 gene duplications (~13%) were inter-chromosomal. In addition, 73 gene duplications (~16%) were dispersed between one of the two chromosomes and one of the plasmids. This study demonstrated that inter-chromosomal DNA sequence duplications were found to be as abundant as the number of duplications found within each of the two chromosomes separately, which further validates the hypothesis that both chromosomes have been ancient partners in the genome of *R. sphaeroides* (3).

The data revealed that gene duplication in *R. sphaeroides* represented a variety of metabolic functions, as described in Figure 3. Results showed that the gene duplications represent the COG functions in the following categories: metabolism (54%), information processing (14%), cellular processes (20%), and poorly characterized functions (12%). When compared to the total genes found in each of these categories in the whole genome, metabolism, information processing, cellular processes, and poorly characterized were present in 45%, 16%, 20%, and 10%, respectively. This suggests that the genes involved only in metabolism were overrepresented while genes involved in information processing, cellular processes, and in poorly characterized functional categories were not significantly different from the percentages of the organism's genome present in the entire genome. Thus, the mechanism of gene duplication for the evolution of gene functions is more prevalent for the genes that are involved in metabolism, which may play important roles in adaptation of *R. sphaeroides* in different ecological niches (4, 5).

Although the gene duplication of individual genes is prevalent in the genome, many genes are found to be duplicated as a result of segmental duplications of the different portion of the chromosomes. Chemotaxis-related genes are located at four major loci, chemotaxis operon I (RSP0642-85), chemotaxis operon II (RSP0924-18), chemotaxis operon III (RSP0342-36), and chemotaxis operon IV (RSP0126-21). These genes are all present on the same plasmid and are conserved in all genomes of *R. sphaeroides*. In addition, the chemotaxis operon III is a part of a 36 kb flagellar biosynthesis gene cluster (3). The chemotaxis genes are involved in chemotaxis, and their conservation in all genomes of *R. sphaeroides* suggests that these genes play an important role in adaptation to their environment.

In addition, the data revealed that gene duplication has occurred in varying number of copies in the genome of *R. sphaeroides*. The number of copies for each gene duplication, their locations on the different replicons, and the metabolic functions, Clusters of Orthologous Groups (COGs) they were involved in were identified.

**METHODS**

In order to identify the amount of gene duplication in the *R. sphaeroides*, the complete set of protein coding ORFs and their corresponding amino acid sequences of its genome were first downloaded from the *R. sphaeroides* website (6). Each ORF was then used as a query to search the local protein database of *R. sphaeroides* 2.4.1 using the Gapped BLASTP (12). The program included gap penalties, and it is therefore more conservative in database searches. All gene loci, which have more than one copy in the *R. sphaeroides*, were aligned to their full length protein. Gene homologous with an alignment score of 100 or more with an E-value < 10^-10 were retained. The score and the amino acid identity reflect the strength of the relationship between homologues. Alignment score is set at 100% amino acid identity which defines the level above which gene duplication can be reliably identified. The number of copies for each gene duplication, their locations on the different replicons, and the metabolic functions, Clusters of Orthologous Groups (COGs) they were involved in were identified.

**REFERENCES**