THE ROYAL SOCIETY

Published online 12 December 2002

Evolution of photosynthetic prokaryotes: a maximum-likelihood mapping approach

Jason Raymond[1], **Olga Zhaxybayeva**[2], **J. Peter Gogarten**[2] **and Robert E. Blankenship**[1*]

[1]*Department of Chemistry and Biochemistry, Arizona State University, Tempe, AZ 85287-1604, USA*
[2]*Department of Molecular and Cell Biology, University of Connecticut, Storrs, CT 06269-3044, USA*

Reconstructing the early evolution of photosynthesis has been guided in part by the geological record, but the complexity and great antiquity of these early events require molecular genetic techniques as the primary tools of inference. Recent genome sequencing efforts have made whole genome data available from representatives of each of the five phyla of bacteria with photosynthetic members, allowing extensive phylogenetic comparisons of these organisms. Here, we have undertaken whole genome comparisons using maximum likelihood to compare 527 unique sets of orthologous genes from all five photosynthetic phyla. Substantiating recent whole genome analyses of other prokaryotes, our results indicate that horizontal gene transfer (HGT) has played a significant part in the evolution of these organisms, resulting in genomes with mosaic evolutionary histories. A small plurality phylogenetic signal was observed, which may be a core of remnant genes not subject to HGT, or may result from a propensity for gene exchange between two or more of the photosynthetic organisms compared.

Keywords: photosynthetic bacteria; evolution; whole genome comparison

1. INTRODUCTION

The origin of photosynthesis was an ancient event that had dramatic ramifications on the subsequent development and diversification of life. The ability to utilize solar energy to drive electron transport and thereby to drive metabolic redox reactions allowed primitive prokaryotes to expand beyond the confines of the resource-limited environments present on early Earth (Des Marais 2000). Anoxygenic photosynthesis, still carried out by most phototrophic bacteria, required electron donors (e.g. H_2S, H_2, Fe^{2+}) characteristic of chemolithotrophic metabolism, necessitating access to geochemically driven, redox-rich environments (Gaidos *et al.* 1999). However, the advent of oxygenic photosynthesis by the cyanobacteria and their progeny in large part bypassed this bottleneck by utilizing ubiquitous liquid H_2O as an electron donor, giving rise to molecular O_2 as a by-product and thereby irrevocably altering the Earth's atmosphere (Blankenship 2002; Wolstencroft & Raven 2002). Although the consequences of these events were momentous, for example in the evolution of eukaryotes, subsequent terrestrial colonization and development of complex organisms, the origin and evolution of photosynthesis has remained enigmatically enshrouded in its own antiquity.

2. MOLECULAR ANALYSIS OF THE FIVE LINEAGES OF PHOTOSYNTHETIC BACTERIA

The widespread use of rRNA sequence evolution as a proxy for organismal speciation has perplexed attempts to explain the distribution of photosynthesis (Olson & Pierson 1987; Oyaizu *et al.* 1987; Gupta *et al.* 1999; Baymann *et al.* 2001). Compounding these difficulties has been the discovery of new phototrophs, including *Chloroflexus aurantiacus*, which, on the basis of 16S rRNA analysis appears to be the earliest branching of the photosynthetic bacteria (Woese 1987), and the heliobacteria, a relatively late-branching Gram-positive family whose members possess the simplest known photosynthetic apparatus (Gest & Favinger 1983). Phototrophy is widely and paraphyletically distributed among different classes of proteobacteria, and this, along with similarities in their reaction centres and light harvesting systems, leads to speculation that the proteobacterial ancestor was a phototroph and that selective loss accounts for the mottled distribution of photosynthesis (Woese 1987). HGT could also be invoked as an explanation for paraphyly. Phototrophic proteobacteria (purple bacteria) so far examined possess photosynthetic gene clusters that might readily be transferred in and among other proteobacteria (Naylor *et al.* 1999; Igarashi *et al.* 2001). Such large gene clusters may signify regions of DNA that are subject to horizontal transfer (Lawrence 1999). However, with the exceptions of the purple bacteria and the heliobacteria (Xiong *et al.* 1998), in all other known examples the genes that code for proteins that are involved in photosynthesis are widely dispersed over the entire genome. *Chlorobium tepidum* and other GSB have long remained enigmatic, both with

[*] Author for correspondence (blankenship@asu.edu).

One contribution of 21 to a Discussion Meeting Issue 'Chloroplasts and mitochondria: functional genomics and evolution'.

DOI 10.1098/rstb.2002.1181

respect to their relationship with other bacteria—typified by long branch lengths separating this phylum from others on 16S rRNA trees (Maidak *et al.* 2001)—and to other phototrophs, as they have very primitive, homodimeric reaction centres and a peripheral antenna complex known as a chlorosome. The cyanobacteria represent the fifth bacterial photosynthetic lineage and, along with the GSB, are a phylum comprised entirely of phototrophs. The cyanobacteria have been of substantial interest not only for the environmental metamorphosis triggered by oxygenic photosynthesis 2.5 billion years ago, but also for their essential role as primary producers and regulators in global carbon cycling (Rye & Holland 1998; Des Marais 2000).

The common photosynthetic components of these organisms—reaction centres, antenna systems and Mg-tetrapyrrole pigments—have evolutionary histories that belie the story told by 16S rRNA (Xiong *et al.* 2000; Xiong & Bauer 2002). The purple bacteria and green filamentous bacteria (*Chloroflexus*) share homologous quinone-type, or type II, reaction centres and core antennas, but differ completely in their peripheral light harvesting systems. The heliobacteria (*Heliobacillus*) and GSB have homodimeric type I reaction centres, typified by FeS centres functioning in light-driven charge separation. While the heliobacteria have no extrinsic light harvesting systems, the GSB possess the novel Fenna–Matthews–Olsen (FMO) protein and a light harvesting chlorosome structure similar to that found in *C. aurantiacus*. These four phyla use different modifications of bacteriochlorophyll as the primary photoactive pigment. With regard to the complexity of their photosynthetic apparatus and evident in their diversity, cyanobacteria represent the pinnacle of bacterial phototrophy. They have evolved the ability to produce chlorophyll in lieu of bacteriochlorophyll, an adaptation energetically required to be able to oxidize water to molecular O_2 (Blankenship & Hartman 1998). They possess both type I and type II reaction centres, although the primary sequence evolution that has taken place from their anoxygenic homologues has been substantial. Their light-gathering phycobilisomes utilize a haem-derived pigment and bear no resemblance to any light harvesting structures found in other bacteria. These irregular similarities and novel differences have obfuscated attempts to explain the evolution of photosynthesis in terms of the proposed origin and diversification of phototrophic lineages. Whole genome analyses, bolstered by the influx of data from genome sequencing projects, have stimulated provocative new ideas on the evolution of prokaryotes, in many cases overwhelmingly demonstrating the substantive role that HGT has played (Aravind *et al.* 1998; Doolittle 1999; Gogarten & Olendzenski 1999; Boucher *et al.* 2001; Zhaxybayeva & Gogarten 2002). Recently, whole genome data have become available for members of each of the five phototrophic phyla, and we have undertaken an extensive analysis of orthologous sequences from representative organisms of each phylum. The five representative species analysed were the firmicute *Heliobacillus mobilis*, the green filamentous bacterium *C. aurantiacus*, the cyanobacterium *Synechocystis* PCC6803, GSB *Chorobium tepidum*, and the α-proteobacterium *Rhodobacter capsulatus*. Here, we describe the comparison of orthologous genes from the complete or near-complete genomes from these five organisms. Rather than supporting a single phylogeny, as might be expected if these organisms had evolved from a common ancestor through vertical descent with modification, orthologous genes analysed by phylogenetic methods support the notion that HGT has been integral to the evolution of photosynthetic prokaryotes and that core genetic elements of the photosynthetic apparatus have been among these laterally transferred genes.

3. WHOLE GENOME ANALYSIS: ORTHOLOGUE SELECTION

The basic analytical method we have used is to identify all orthologous genes that are shared in at least four of the five genomes being analysed and then to establish their interrelationships using tools of phylogenetic inference. The first step of the analysis is the selection of orthologous protein sequences for comparison, which must be distinguished from paralogous and non-homologous sequences so that accurate phylogenies are inferred. This is achieved by performing genome-versus-genome Blast comparisons for all possible pairs of organisms (see Appendix A for details). All putative orthologues must also have Blast expectation values below a preset threshold. Datasets are then compiled from sequences that are reciprocal top-scoring Blast hits across all of the genomes being compared. Therefore, to be included in a dataset for subsequent alignment and phylogenetic analysis, a given sequence must select every other sequence from that dataset as a top-scoring Blast hit.

These methods minimize the number of false positives (paralogous or non-homologous genes), which can skew the results of phylogenetic analysis (for methodology, application and further discussion, see Appendix A; also Tatusov *et al.* 1997; Gogarten & Olendzenski 1999; Zhaxybayeva & Gogarten 2002). Perhaps not surprisingly, the number of orthologues present in all five genomes that fulfil the stringent selection criterion described above is only 188, representing only 5–10% of the genes from these genomes. However, these orthologues represent highly conserved genes with unambiguous homology to one another, and therefore should provide the best phylogenetic signal available for comparison of these genomes. Here, we use an ML method to compare orthologues from these five organisms taken four at a time, thereby allowing the inclusion of genes missing in one of the five genomes and increasing the total number of orthologues compared to 527. Specific results on the 188 orthologues found strictly in all five genomes are presented elsewhere (Raymond *et al.* 2002). Figure 1 shows a Venn diagram of the possible combinations of five genomes and number of orthologues compared for each combination, and illustrates the advantage of analysing only four genomes at a time versus five genomes. In those instances where only one of the genomes is missing a particular orthologue, the other orthologues can still be analysed with respect to their phylogenetic relationship, thereby increasing the number of analysed orthologues.

4. WHOLE GENOME ANALYSIS: ML MAPPING

Utilizing ML mapping as developed by Strimmer & von Haeseler (1997), posterior probabilities of each of the

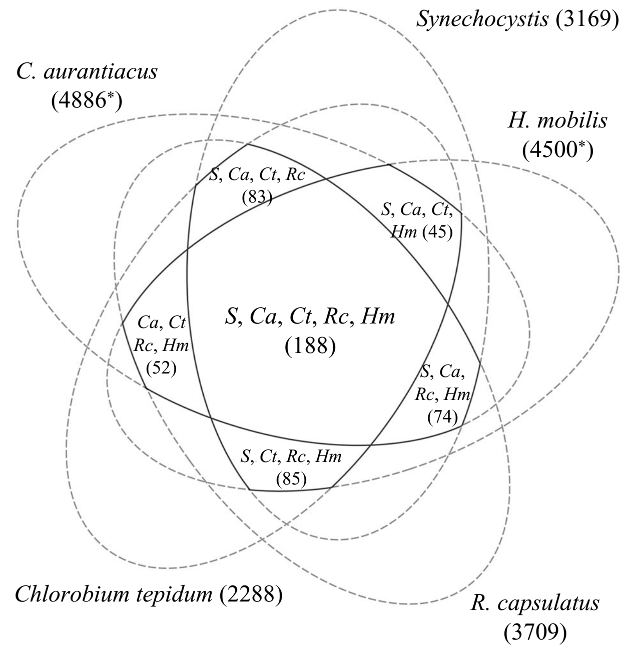


Figure 1. Venn diagram of the intersections of orthologous genes in five genomes, where each ellipse represents one of the five genomes. There are 26 possible intersections between two or more genomes (plus five regions composed of organism-specific genes), and in this paper we focus on the 'phylogenetically informative' cases outlined (unbroken) at the middle of the Venn diagram. The organisms comprising each region are shown along with the number of orthologous sets (in parentheses) from each region found by our orthologue selection criterion. Each of the five ML maps thus combines orthologues from the corresponding four genome intersection and from the relevant genomes from the centre five taxa region (e.g. figure 3 is composed of the 83 orthologues from *S*, *Ca*, *Ct*, *Rc* and 188 orthologues from these genomes from *S*, *Ca*, *Ct*, *Rc*, *Hm*). The corresponding genome and putative number of protein-coding genes is given adjacent to each ellipse (the asterisk indicates preliminary genomes where the number of protein-coding genes might not be finally determined). *S*, *Synechocystis*; *Ct*, *Chlorobium tepidum*; *Ca*, *Chloroflexus aurantiacus*; *Rc*, *Rhodobacter capsulatus*; *Hm*, *Heliobacillus mobilis*.

three possible tree topologies for individual aligned sets of orthologues were plotted as points in a barycentric coordinate system, whose coordinates are specified as posterior probabilities for each of the three unrooted tree topologies possible for a four taxa phylogenetic tree. On these so-called ML maps, points representing orthologues with higher relative support for a specific tree topology are located closer to that particular vertex of the triangle, while orthologues without preference for a particular topology map near the centre of the triangle (for additional methodology see Zhaxybayeva & Gogarten 2002). This method provides a visual way to assess simultaneously the statistical strength and most likely topology for a large number of orthologues, but restricts the number of taxa compared to four. Thus, we have carried out ML mapping for all five possible combinations of four of the five genomes compared here.

The reliability of this technique for detecting close phylogenetic relationships between genomes is demonstrated when two organisms known to be closely related (for example the purple photosynthetic proteobacteria *R. capsulatus* and *Rhodopseudomonas palustris* shown in figure 2) are included. This produces the expected result that nearly all the genes analysed strongly support the tree grouping these two organisms as sister taxa. Most of the points in figure 2 map close to or at the apex of the ML map, representing genes that support that particular tree above 99% posterior probability and indicating a well-resolved topology for those genes.

In stark contrast to figure 2, the five different ML maps for four genomes show a distinct lack of unanimous support for a single topology, even though the majority of individual sets of orthologues have strong posterior probability support. Plurality support is observed for phylogenies clustering *Rhodobacter* and *Chlorobium* against either *Synechocystis*, *Chloroflexus* or *Heliobacillus* (figures 3–5). In ML maps where either *Rhodobacter* or *Chlorobium* is omitted from the analysis, the three possible topologies receive nearly equal support (figures 6 and 7). While these plurality signatures are small, they appear consistently in the different ML maps and indicate either a close relationship between *Rhodobacter* and *Chlorobium*, or perhaps a less resolved but closer relationship between *Synechocystis*, *Heliobacillus* and *Chloroflexus*. While still not completely settled, additional data (discussed below) appear to support the latter hypothesis. Despite the noted plurality signatures, the overwhelming picture from these ML maps indicates that highly conserved genes have complex evolutionary histories, often inconsistent with relationships predicted from 16S-based organismal phylogeny.

Additionally, sets of orthologues were stratified by their putative functional assignments based on COG categories (Tatusov *et al.* 1997, 2000). The distribution of genes among functional categories and number of genes in support of the plurality topology for each of the five ML maps are shown in figure 8. It has been argued that highly interacting genes functioning, for example, in information processing are refractory to HGT and should, as a subset, show preference for a single topology (Jain *et al.* 1999; Woese 2000). However, this trend is not observed when segregating genes in different functional categories by most probable topology. The lack of consensus support for a single topology among these nearly 527 often highly conserved sets of orthologues verifies that the genomes of these organisms are mosaics of genes with often substantially different evolutionary histories. This contradicts what would be observed if these genomes had evolved strictly through vertical descent, and indicates that substantial lateral transfer of genetic information has occurred among these organisms.

5. DISCUSSION

Evidence from even the earliest analyses of complete prokaryote genomes hinted at a remarkable amount of lateral transfer (Aravind *et al.* 1998). Recent work has shown that even organisms virtually indistinguishable on the basis of rRNA sequences have undergone significant evolution on the genomic scale (Bergthorsson & Ochman 1998; Boucher *et al.* 2001; Perna *et al.* 2001). While the representative organisms compared here share photosynthesis as a common thread, many key components of the photosynthetic machinery, especially reaction centre proteins and light harvesting antenna systems, are widely divergent

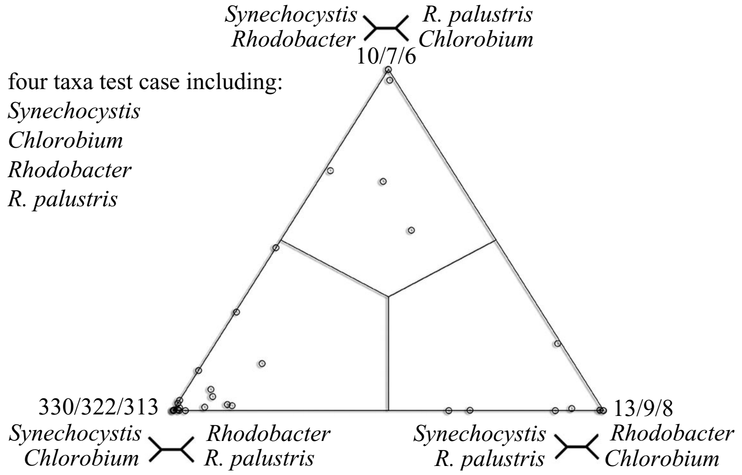


Figure 2. Test case mapping topologies of 353 orthologues from *Synechocystis*, *Chlorobium tepidum* and purple bacteria *Rhodopseudomonas palustris* and *Rhodobacter capsulatus*. Corresponding trees are shown at each vertex, and numbers given at each vertex give: first number, total number of cases supporting that topology; second number, number of cases supporting that topology above 90% posterior probability; and third number, number of cases supporting that topology above 99% posterior probability.

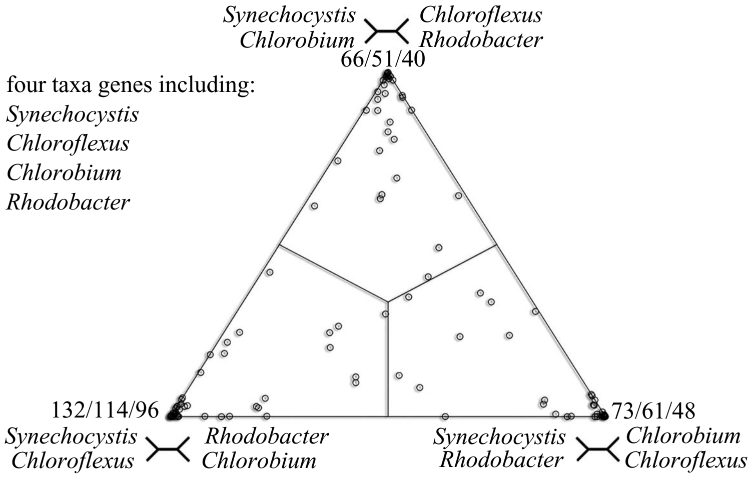


Figure 3. Map of 271 orthologues from *Synechocystis*, *Chloroflexus aurantiacus*, *Chlorobium tepidum* and *Rhodobacter capsulatus*. Numbers at vertices are explained in the caption for figure 2.

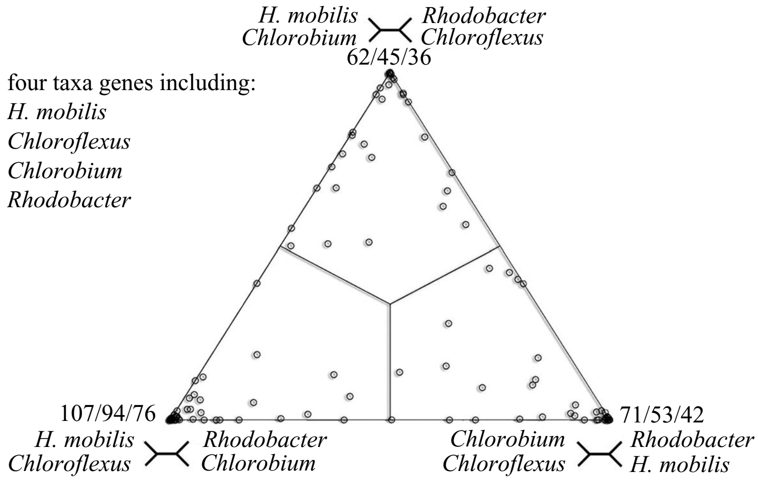


Figure 4. Map of 240 orthologues from *Heliobacillus mobilis*, *Chloroflexus aurantiacus*, *Chlorobium tepidum* and *Rhodobacter capsulatus*. Numbers at vertices are explained in the caption for figure 2.

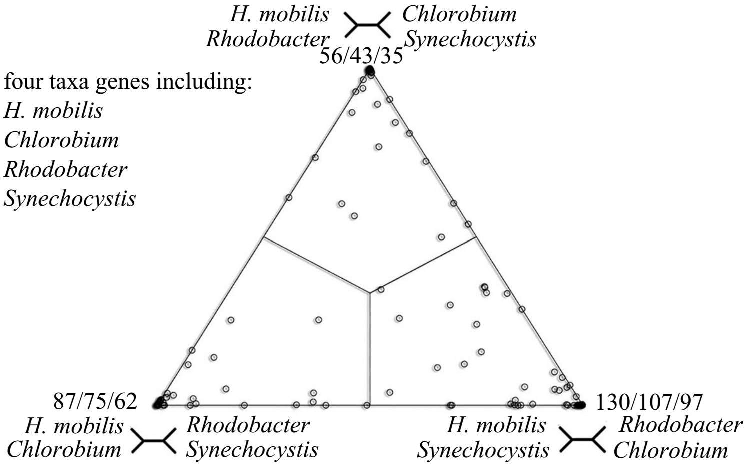


Figure 5. Map of 273 orthologues from *Heliobacillus mobilis*, *Chlorobium tepidum*, *Rhodobacter capsulatus* and *Synechocystis*. Numbers at vertices are explained in the caption for figure 2.

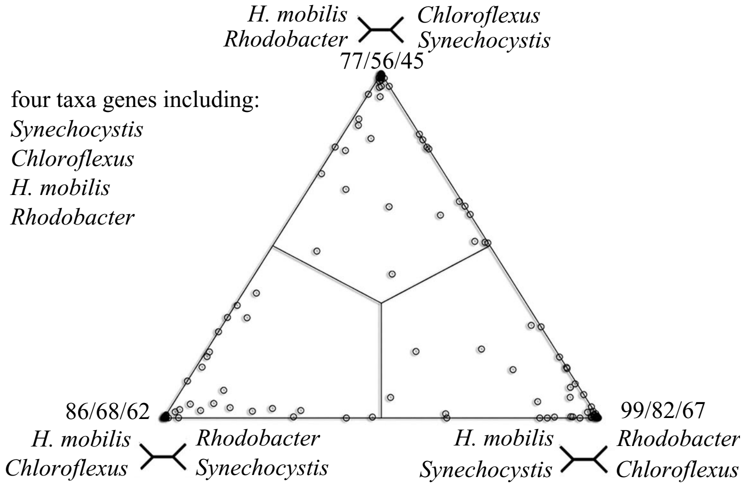


Figure 6. Map of 282 orthologues from *Synechocystis*, *Chloroflexus aurantiacus*, *Heliobacillus mobilis* and *Rhodobacter capsulatus*. Numbers at vertices are explained in the caption for figure 2.

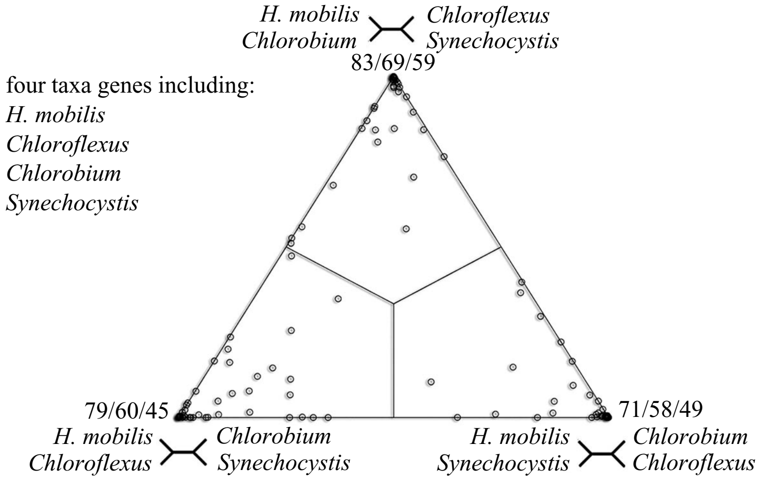
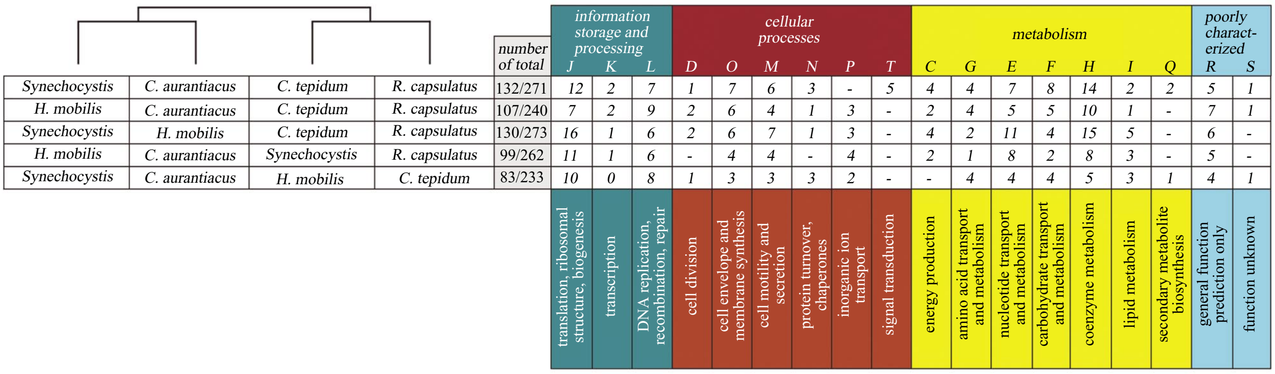


Figure 7. Map of 233 orthologues from *Heliobacillus mobilis*, *Chloroflexus aurantiacus*, *Chlorobium tepidum* and *Synechocystis*. Numbers at vertices are explained in the caption for figure 2.

and therefore not suitable for wide-scale phylogenetic reconstructions. This apparently chimeric composition of these organisms has confounded various attempts to reconcile the evolution of photosynthesis with inferred speciation events.

The plurality signature observed in figures 3–5 might result from two distinct causes. The first is that, in addition to extensive HGT evident between prokaryotes, a subset of two or more of the phototrophs compared here may more frequently exchange genes with one another,



					information storage and processing			cellular processes							metabolism							poorly characterized	
				number of total	*J*	*K*	*L*	*D*	*O*	*M*	*N*	*P*	*T*	*C*	*G*	*E*	*F*	*H*	*I*	*Q*	*R*	*S*
Synechocystis	*C. aurantiacus*	*C. tepidum*	*R. capsulatus*	132/271	12	2	7	1	7	6	3	-	5	4	4	7	8	14	2	2	5	1
H. mobilis	*C. aurantiacus*	*C. tepidum*	*R. capsulatus*	107/240	7	2	9	2	6	4	1	3	-	2	4	5	5	10	1	-	7	1
Synechocystis	*H. mobilis*	*C. tepidum*	*R. capsulatus*	130/273	16	1	6	2	6	7	1	3	-	4	2	11	4	15	5	-	6	-
H. mobilis	*C. aurantiacus*	*Synechocystis*	*R. capsulatus*	99/262	11	1	6	-	4	4	-	4	-	2	1	8	2	8	3	-	5	-
Synechocystis	*C. aurantiacus*	*H. mobilis*	*C. tepidum*	83/233	10	0	8	1	3	3	3	2	-	-	4	4	4	5	3	1	4	1
					translation, ribosomal structure, biogenesis	transcription	DNA replication, recombination, repair	cell division	cell envelope and membrane synthesis	cell motility and secretion	protein turnover, chaperones	inorganic ion transport	signal transduction	energy production	amino acid transport and metabolism	nucleotide transport and metabolism	carbohydrate transport and metabolism	coenzyme metabolism	lipid metabolism	secondary metabolite biosynthesis	general function prediction only	function unknown

Figure 8. Distributions of genes among functional categories for the plurality topologies from figures 3–7. Functional categories (columns) are shown for well-resolved cases receiving greater than 99% posterior probability support, and are given according to the COG database (Tatusov *et al.* 1997, 2000).

for example by virtue of a shared environment. Different genera of these five lineages often inhabit hot springs, and such a close association, for example in bacterial mats composed of *Chloroflexus*-like strains underneath a layer of thermophilic cyanobacteria, could dramatically increase the probability of HGT. Additionally, recent studies of phototrophic consortia (Overmann & Schubert 2002) have illustrated an epibiontic relationship between GSB and proteobacteria, and such juxtaposed microenvironments would strongly facilitate exchange of genetic material. On the basis of branch lengths and estimated distances from protein sequence data (Raymond *et al.* 2002), the association between *Chlorobium tepidum* and *R. capsulatus* appears to be primarily a consequence of relative similarity of *Synechocystis*, *Chloroflexus* and *Heliobacillus*, indicating that *Chlorobium* and *Rhodobacter* are grouped by exclusion rather than by affinity. However, the consistent clustering but relatively long branch lengths found between *Chlorobium* and *Rhodobacter* may result from using taxa that are more distant phylogenetic cousins to symbiotically interacting species that have frequently exchanged genes, as would be consistent with the observations by Overmann & Schubert (2002); so this issue is still unresolved.

The second possible explanation for the plurality signature is that it represents a subset of genes that have not been subject to HGT, a putative 'core' of genes that have been inherited through vertical descent and speciation (for discussion see Gogarten & Olendzenski 1999 and Nesbo *et al.* 2001). While it is interesting to note that the plurality topology is consistent with rDNA analysis from these organisms (Maidak *et al.* 2001), COG categorization of these plurality genes indicates a diverse range of functions not limited to informational genes, contradictory to postulates by Jain *et al.* (1999) and Woese (2000). Furthermore, this phylogeny is also at variance with those of the chlorophyll biosynthetic pathway (Xiong *et al.* 2000), which consistently group *Chlorobium* with *Chloroflexus*. Additionally, many informational genes support alternative topologies, so at the very least the notion of which genes are more amenable or recalcitrant to HGT must be refined. Compelling arguments have been made recently that even highly conserved informational genes, such as rDNA, are not necessarily immune to lateral gene transfer (Asai *et al.* 1999; Yap *et al.* 1999). It has long been shown that functional ribosomes can be reconstituted from parts taken from different organisms (Nomura *et al.* 1968; Bellemare *et al.* 1973; Daya-Grosjean *et al.* 1973; Wrede & Erdmann 1973). Even if the plurality signature does represent a core of vertically transferred genes, the extent of genes with incongruent evolutionary histories must be taken into consideration when extrapolating single-gene analyses to infer, for example, evolutionary relationships between prokaryotes or to estimate environmental metabolic diversity.

From this evidence, what, then, can be said about the origin and distribution of photosynthesis, or, more specifically, where are photosynthesis genes in these maps? Unfortunately, only a small number of genes that are explicitly involved with phototrophy are retained by our orthologue selection criteria. The dichotomous nature of reaction centre genes and light harvesting systems, present only in limited subsets of these five phototrophs, results in all these genes being filtered out of our analysis. However, several pigment biosynthesis genes are present in our sets of orthologues, and almost all support the analysis of Xiong *et al.* (2000) in indicating a close relationship in the bacteriochlorophyll synthesis pathways of *Chloroflexus* and *Chlorobium*. Moreover, the small number of photosynthesis related genes retained by our orthologue selection criterion support the notion of extensive gene recruitment during their evolution. For example, many components of chlorophyll biosynthesis are filtered out because of their paralogous origins in part from nitrogen fixation and cobalamin biosynthesis pathways (Xiong *et al.* 2000). This analysis serves more appropriately as a preliminary map of the evolutionary histories of highly conserved orthologues from these genomes, to which future analyses of individual photosynthetic genes can be compared in an attempt to understand which genes have been horizontally transferred among which organisms, if groups of genes—perhaps clustered in operons—have been subject to simultaneous HGT, and possibly to gain insight into the timing of these complex evolutionary events.

Here, we illustrate that lateral transfer among photosynthetic bacteria has been both frequent and extensive, resulting in mosaic evolutionary histories within their genomes. The difference between the plurality consensus and the phylogenies of key photosynthetic enzymes indicates that portions of the genetic components of photosynthesis have been laterally transferred, perhaps multiple times across different lineages, and subsequently refined during the evolution of these species. In light of the demonstrated propensity for lateral transfer among these groups, we propose that horizontal gene flow emerges as the most plausible explanation for the current distribution of bacterial phototrophy. This result stands to reconcile the multiple lines of typically disparate phylogenetic evidence that have been invoked in recent years to explain the evolution and subsequent distribution of photosynthesis, and substantiates the extensive role that horizontal gene flow has played, even in the transfer of complex metabolic pathways, during the early evolution of prokaryotes.

The NASA Exobiology program and the NASA Astrobiology Institute at ASU are gratefully acknowledged for support. This is publication no. 565 from the Arizona State University Centre for the Study of Early Events in Photosynthesis. Sequence data for *Chlorobium tepidum* were obtained from the Institute for Genomic Research through the website at http://www.tigr.org. Sequencing of *Chlorobium tepidum* was accomplished with support from the DOE. The data for *Chloroflexus aurantiacus* have been provided freely by the US DOE Joint Genome Institute. We also thank Integrated Genomics for collaborative efforts and use of sequence data from *R. capsulatus* and *Heliobacillus mobilis*.

APPENDIX A: METHODS

For *Synechocystis* PCC 6803 and *R. capsulatus*, annotated protein coding gene FASTA-formatted files were downloaded from the NCBI website (www.ncbi.nlm.nih.gov/) and the Integrated Genomics ERGO (ergo.integratedgenomics.com/ergo) website, respectively. Preliminary *C. aurantiacus* coding sequence translations based on the most recently released contig assembly (as of 26 May 2001) were downloaded from the Oak Ridge National Laboratory Computational Biology site (genome.ornl.gov/microbial/caur/). The complete nucleotide genome sequences for *Chlorobium tepidum* and *H. mobilis* were downloaded from TIGR (www.tigr.org) and from Integrated Genomics, respectively, and translated into all six possible ORFs (here used interchangeably with 'gene' with the understanding that this assignment is still putative) greater than 50 amino-acid residues in length, not allowing ORFs to overlap for more than 60 nucleotide bases. While many artefacts are expected from this ORF identification, these are filtered in the course of analysis by the criteria used for orthologue selection. For each ORF in all five genomes, BLAST searches were performed with an *E*-value cutoff of 10^{-4} using the stand-alone BLAST software package (Altschul *et al.* 1990). Reciprocal match orthologues, referred to herein as orthologues, are operationally defined as ORFs from two genomes that are each other's top scoring BLAST hits. Reciprocal match orthologues were extended to multiple genomes, so that each possible pair of ORFs from a given set must fulfil the reciprocal top-scoring-hit relationship. Quartets of orthologues were detected using Structured Query Language queries through the MYSQL, v. 3.23 database (http://www.mysql.com) and also using scripts written in PERL. Multiple sequence alignments of orthologues were carried out using CLUSTALW and further refined manually. TREE-PUZZLE, v. 5.0 (Schmidt *et al.* 2002) was used for ML analysis, using the auto-detection of the substitution model. BIOEDIT (Hall 1999), MEGA2 (Kumar *et al.* 1994) and various programs within the PHYLIP software package (Felsenstein 1989) (v. 3.6) were used to edit sequences and infer distance-based trees. Programs used for gene recovery and automated analyses were written primarily in PERL utilizing scripts from the SEALS package (Walker & Koonin 1997). Modules from the BIOPERL package were implemented and some programs were written using the Java PAL library (Drummond & Strimmer 2001).

REFERENCES

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. 1990 Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410.

Aravind, L., Tatusov, R. L., Wolf, Y. I., Walker, D. R. & Koonin, E. V. 1998 Evidence for massive gene exchange between archaeal and bacterial hyperthermophiles. *Trends Genet.* **14**, 442–444.

Asai, T., Zaporojets, D., Squires, C. & Squires, C. L. 1999 An *Escherichia coli* strain with all chromosomal rRNA operons inactivated: complete exchange of rRNA genes between bacteria. *Proc. Natl Acad. Sci. USA* **96**, 1971–1976.

Baymann, F., Brugna, M., Muhlenhoff, U. & Nitschke, W. 2001 Daddy, where did (PS) I come from? *Biochim. Biophys. Acta Bioenergetics* **1507**, 291–310.

Bellemare, G., Vigne, R. & Jordan, B. R. 1973 Interaction between *Escherichia coli* ribosomal proteins and 5S RNA molecules: recognition of prokaryotic 5S RNAs and rejection of eukaryotic 5S RNAs. *Biochimie* **55**, 29–35.

Bergthorsson, U. & Ochman, H. 1998 Distribution of chromosome length variation in natural isolates of *Escherichia coli*. *Mol. Biol. Evol.* **15**, 6–16.

Blankenship, R. E. 2002 *Molecular mechanisms of photosynthesis*. Oxford: Blackwell Science.

Blankenship, R. E. & Hartman, H. 1998 The origin and evolution of oxygenic photosynthesis. *Trends Biochem. Sci.* **23**, 94–97.

Boucher, Y., Nesbo, C. L. & Doolittle, W. F. 2001 Microbial genomes: dealing with diversity. *Curr. Opin. Microbiol.* **4**, 285–289.

Daya-Grosjean, L., Geisser, M., Stoffler, G. & Garret, R. A. 1973 Heterologous protein-RNA interactions in bacterial ribosomes. *FEBS Lett.* **37**, 17–20.

Des Marais, D. J. 2000 Evolution. When did photosynthesis emerge on Earth? *Science* **289**, 1703–1705.

Doolittle, W. F. 1999 Lateral genomics. *Trends Cell Biol.* **9**, M5–M8.

Drummond, A. & Strimmer, K. 2001 PAL: an object-oriented programming library for molecular evolution and phylogenetics. *Bioinformatics* **17**, 662–663.

Felsenstein, J. 1989 PHYLIP: phylogeny inference package (v. 3.2). *Cladistics* **5**, 164–166.

Gaidos, E. J., Nealson, K. H. & Kirschvink, J. L. 1999 Life in ice-covered oceans. *Science* **284**, 1631–1633.

Gest, H. & Favinger, J. L. 1983 *Heliobacterium chlorum*, an anoxygenic brownish-green photosynthetic bacterium containing a new form of bacteriochlorophyll. *Arch. Microbiol.* **136**, 11–16.

Gogarten, J. P. & Olendzenski, L. 1999 Orthologs, paralogs and genome comparisons. *Curr. Opin. Genet. Dev.* **9**, 630–636.

Gupta, R. S., Mukhtar, T. & Singh, B. 1999 Evolutionary relationships among photosynthetic prokaryotes (*Heliobacterium chlorum*, *Chloroflexus aurantiacus*, cyanobacteria, *Chlorobium tepidum* and proteobacteria): implications regarding the origin of photosynthesis. *Mol. Microbiol.* **32**, 893–906.

Hall, T. A. 1999 BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symp. Ser.* **41**, 95–98.

Igarashi, N., Harada, J., Nagashima, S., Matsuura, K., Shimada, K. & Nagashima, K. V. 2001 Horizontal transfer of the photosynthesis gene cluster and operon rearrangement in purple bacteria. *J. Mol. Evol.* **52**, 333–341.

Jain, R., Rivera, M. C. & Lake, J. A. 1999 Horizontal gene transfer among genomes: the complexity hypothesis. *Proc. Natl Acad. Sci. USA* **96**, 3801–3806.

Kumar, S., Tamura, K. & Nei, M. 1994 Mega: molecular evolutionary genetics analysis software for microcomputers. *Comput. Appl. Biosci.* **10**, 189–191.

Lawrence, J. 1999 Selfish operons: the evolutionary impact of gene clustering in prokaryotes and eukaryotes. *Curr. Opin. Genet. Dev.* **9**, 642–648.

Maidak, B. L., Cole, J. R., Lilburn, T. G., Parker Jr, C. T., Saxman, P. R., Farris, R. J., Garrity, G. M., Olsen, G. J., Schmidt, T. M. & Tiedje, J. M. 2001 The RDP-II (ribosomal database project). *Nucleic Acids Res.* **29**, 173–174.

Naylor, G. W., Addlesee, H. A., Gibson, L. C. D. & Hunter, C. N. 1999 The photosynthesis gene cluster of *Rhodobacter sphaeroides*. *Photosynthesis Res.* **62**, 121–139.

Nesbo, C. L., Boucher, Y. & Doolittle, W. F. 2001 Defining the core of nontransferable prokaryotic genes: the euryarchaeal core. *J. Mol. Evol.* **53**, 340–350.

Nomura, M., Traub, P. & Bechmann, H. 1968 Hybrid 30S ribosomal particles reconstituted from components of different bacterial origins. *Nature* **219**, 793–799.

Olson, J. M. & Pierson, B. K. 1987 Evolution of reaction centers in photosynthetic prokaryotes. *Int. Rev. Cytol.* **108**, 209–248.

Overmann, J. & Schubert, K. 2002 Phototrophic consortia: model systems for symbiotic interrelations between prokaryotes. *Arch. Microbiol.* **177**, 201–208.

Oyaizu, H., Debrunner-Vossbrinck, B., Mandelco, L., Studier, J. A. & Woese, C. R. 1987 The green non-sulfur bacteria: a deep branching in the eubacterial line of descent. *Syst. Appl. Microbiol.* **9**, 47–53.

Perna, N. T. (and 27 others) 2001 Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. *Nature* **409**, 529–533.

Raymond, J., Zhaxybayeva, O., Gogarten, J. P., Gerdes, S. Y. & Blankenship, R. E. 2002 Whole genome analysis of photosynthetic prokaryotes. *Science*. (In the press.)

Rye, R. & Holland, H. D. 1998 Paleosols and the evolution of atmospheric oxygen: a critical review. *Am. J. Sci.* **298**, 621–672.

Schmidt, H. A., Strimmer, K., Vingron, M. & von Haeseler, A. 2002 Tree-Puzzle: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* **18**, 502–504.

Strimmer, K. & von Haeseler, A. 1997 Likelihood-mapping: a simple method to visualize phylogenetic content of a sequence alignment. *Proc. Natl Acad. Sci. USA* **94**, 6815–6819.

Tatusov, R. L., Koonin, E. V. & Lipman, D. J. 1997 A genomic perspective on protein families. *Science* **278**, 631–637.

Tatusov, R. L., Galperin, M. Y., Natale, D. A. & Koonin, E. V. 2000 The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.* **28**, 33–36.

Walker, D. R. & Koonin, E. V. 1997 Seals: a system for easy analysis of lots of sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **5**, 333–339.

Woese, C. R. 1987 Bacterial evolution. *Microbiol. Rev.* **51**, 221–271.

Woese, C. R. 2000 Interpreting the universal phylogenetic tree. *Proc. Natl Acad. Sci. USA* **97**, 8392–8396.

Wolstencroft, R. D. & Raven, J. A. 2002 Photosynthesis: likelihood of occurrence and possibility of detection on Earth-like planets. *Icarus* **157**, 535–548.

Wrede, P. & Erdmann, V. A. 1973 Activities of *B. stearothermophilus* 50 S ribosomes reconstituted with prokaryotic and eukaryotic 5 S RNA. *FEBS Lett.* **33**, 315–319.

Xiong, J. & Bauer, C. E. 2002 Complex evolution of photosynthesis. *A. Rev. Plant Physiol. Plant Mol. Biol.* **53**, 503–521.

Xiong, J., Inoue, K. & Bauer, C. E. 1998 Tracking molecular evolution of photosynthesis by characterization of a major photosynthesis gene cluster from *Heliobacillus mobilis*. *Proc. Natl Acad. Sci. USA* **95**, 14 851–14 856.

Xiong, J., Fischer, W. M., Inoue, K., Nakahara, M. & Bauer, C. E. 2000 Molecular evidence for the early evolution of photosynthesis. *Science* **289**, 1724–1730.

Yap, W. H., Zhang, Z. & Wang, Y. 1999 Distinct types of rRNA operons exist in the genome of the actinomycete *Thermomonospora chromogena* and evidence for horizontal transfer of an entire rRNA operon. *J. Bacteriol.* **181**, 5201–5209.

Zhaxybayeva, O. & Gogarten, J. P. 2002 Bootstrap, Bayesian probability and maximum likelihood mapping: exploring new tools for comparative genome analyses. *BMC Genomics* **3**, 4.

GLOSSARY

COG: clusters of orthologous groups
GSB: green sulphur bacteria
HGT: horizontal gene transfer
ML: maximum likelihood
ORF: open reading frame