Genome divergence in two *Prochlorococcus* ecotypes reflects Oceanic niche differentiation

Gabrielle Rocap*, Frank W. Larimer†‡, Jane Lamerdin‡, Stephanie Malfatti‡, Patrick Chain§‡, Nathan A. Ahlgren*, Andrae Arellano‡, Maureen Coleman‖, Loren Hauser†‡, Wolfgang R. Hess¶#, Zackary I. Johnson‖, Miriam Land†‡, Debbie Lindell‖, Anton F. Post**, Warren Regala‡, Manesh Shah†‡, Stephanie L. Shaw†† ‡‡, Claudia Steglich¶, Matthew B. Sullivan§§, Claire S. Ting‖‖, Andrew Tolonen§§, Eric A. Webb¶¶, Erik R. Zinser‖ and Sallie W. Chisholm‖‖‖

** School of Oceanography, University Of Washington, Seattle, Washington 98195 USA*

† Computational Biology, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831 USA

‡ Joint Genome Institute, Walnut Creek, California 94598 USA

§ Lawrence Livermore National Laboratory, Livermore, California 94550 USA

‖ Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139 USA

¶ Institute of Biology, Humboldt-University, D-10115 Berlin, GERMANY

*** Interuniversity Institute of Marine Science, 88103 Eilat, ISRAEL*

†† Department of Earth, Atmospheric and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139 USA

[§§] *Joint Program in Biological Oceanography, Massachusetts Institute of Technology and Woods Hole Oceanographic Institution, Cambridge, Massachusetts 02139 USA*

[‖‖] *Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139 USA*

[¶¶] *Biology Department, Woods Hole Oceanographic Institution, Woods Hole, Massachusetts 02543 USA*

[#] Present Address: Ocean Genome Legacy, Beverly MA 01915 USA

[‡‡] Present Address: Department of Environmental Science Policy and Management, University of California, Berkeley, California 94720 USA

The marine unicellular cyanobacterium *Prochlorococcus* is the smallest known oxygen-evolving autotroph[1]. It numerically dominates the phytoplankton in the tropical and sub-tropical oceans[2,3], and is responsible for a significant fraction of global photosynthesis. Here we compare the genomes of two *Prochlorococcus* strains that span the largest evolutionary distance within the *Prochlorococcus* lineage[4] and have different minimum, maximum and optimal light intensities for growth[5]. The high light adapted ecotype has the smallest genome (1,657,990 bp, 1716 genes) of any known oxygenic phototroph; the genome of its low light counterpart is significantly larger, at 2,410,873 bp (2275 genes). The comparative architectures of these two strains reveal dynamic genomes which are constantly changing in response to myriad selection pressures. Although the two strains have 1350 genes in common, a significant number are not shared, which have either been differentially retained from the common ancestor, or acquired through duplication or lateral transfer. Some of these genes play obvious roles in determining the relative fitness of the ecotypes in response to key environmental variables, and hence in regulating their distribution and abundance in the oceans.

As an oxyphototroph, *Prochlorococcus* requires only light, CO_2 and inorganic nutrients, thus the opportunities for extensive niche differentiation are not immediately obvious — particularly in view of the high mixing potential in the marine environment (Fig. 1a). Yet co-occurring *Prochlorococcus* cells that differ in their rDNA sequence by less than 3% have different optimal light intensities for growth[6], pigment contents[7], light harvesting efficiencies[5], sensitivities to trace metals[8], nitrogen utilization abilities[9] and cyanophage specificities[10] (Fig. 1b,c). These "ecotypes"— i.e. distinct genetic lineages with ecologically relevant physiological differences — would be lumped as a single species on the basis of their rDNA similarity[11], yet they have strikingly different distributions within a stratified oceanic water column, with high light adapted ecotypes most abundant in surface waters, and their low light adapted counterparts dominating deeper waters[12] (Fig. 1a). The detailed comparison between the genomes of two *Prochlorococcus* ecotypes we report here reveals many of the genetic foundations for the observed differences in their physiologies and vertical niche partitioning and together with the genome of their close relative *Synechococcus*[13], help elucidate the key factors that regulate species diversity, and the resulting biogeochemical cycles, in today's oceans.

The genome of *Prochlorococcus* MED4, a high light adapted strain, is 1,657,990 bp. This is the smallest of any oxygenic phototroph — significantly smaller than that of the low light adapted strain, MIT 9313 (2,410,873 bp) (Table 1). The genomes of MED4 and MIT9313 consist of a single circular chromosome (Supp. Fig. 1), and encode 1716 and 2275 genes respectively, roughly 65% of which can be assigned a functional category (Supp. Fig. 2). Both genomes have undergone numerous large and small-scale rearrangements but they retain conservation of local gene order (Fig. 2). Break points between the orthologous gene clusters are commonly flanked by tRNAs suggesting that these genes serve as loci for rearrangements caused by internal homologous recombination or phage integration events.

The strains have 1352 genes in common, all but 38 of which are also shared with *Synechococcus* WH 8102[13]. Many of the 38 "*Prochlorococcus*-specific" genes encode proteins involved in the atypical light harvesting complex of *Prochlorococcus*, which contains divinyl chlorophylls *a* and *b* rather than the phycobilisomes that characterize most cyanobacteria. They include genes encoding the chlorophyll *a*/*b*-binding proteins (*pcb*)[14], a putative chlorophyll *a* oxygenase which could synthesize (divinyl) chlorophyll *b* from (divinyl) chlorophyll *a*[15], and a lycopene epsilon cyclase involved in the synthesis of alpha carotene[16]. This remarkably low number of "genera defining" genes illustrates how differences in a few gene families can translate into significant niche differentiation among closely related microbes.

MED4 has 364 genes without an ortholog in MIT 9313, while MIT 9313 has 923 that are not present in MED4. These “strain-specific” genes, which are dispersed throughout the chromosome (Fig 2), clearly hold clues about the relative fitness of the two strains under different environmental conditions. Almost half of the 923 MIT 9313-specific genes are in fact present in *Synechococcus* WH 8102, suggesting they have been lost from MED4 in the course of genome reduction. Lateral transfer events, perhaps mediated by phage[10] may also be a source of some of the strain-specific genes (Supp. Fig. 3, 4, 5, 6).

Gene loss has played a major role in defining the *Prochlorococcus* photosynthetic apparatus. MED4 and MIT 9313 are missing many of the genes encoding phycobilisome structural proteins and enzymes involved in phycobilin biosynthesis[15]. Although some of these genes remain, and are functional[17], others appear to be evolving rapidly within the *Prochlorococcus* lineage[18]. Selective genome reduction can also be seen in the photosynthetic reaction center of *Prochlorococcus*. Light acclimation in cyanobacteria often involves differential expression of multiple, but distinct, copies of genes encoding Photosystem II D1 and D2 reaction center proteins (*psbA* and *psbD* respectively)[19]. However, MED4 has a single *psbA* gene, MIT 9313 has two that encode identical PSII D1 polypeptides, and both possess only one *psbD* gene, suggesting a diminished ability to photoacclimate. MED4 has also lost the gene encoding cytochrome c550 (*psbV*), which plays a crucial role in the oxygen evolving complex in *Synechocystis* PCC 6803[20].

There are several differences between the genomes that help account for the different light optima of the two strains. For example, the smaller MED4 genome has more than twice as many genes (22 vs 9) encoding putative High Light Inducible Proteins (HLIPs), which appear to have arisen at least in part via duplication events[15]. MED4 also possesses a photolyase gene that has been lost in MIT9313, likely because there is little selective pressure to retain UV damage repair in low light habitats. Regarding differences in light harvesting efficiencies, it is noteworthy that MED4 contains only a single gene encoding the chlorophyll a/b-binding antenna protein Pcb, while MIT 9313 possesses two copies. The second type has been found exclusively in low light adapted strains[21], and may form an antenna capable of binding more chlorophyll pigments.

Both strains have a low proportion of genes involved in regulatory functions. Compared to the freshwater cyanobacterium *Thermosynechococcus elongatus* (genome size < 2.6 Mbp)[22], MIT 9313 has fewer sigma factors, transcriptional regulators and two component sensor-kinase systems, and MED4 is even more reduced (Supp. Table 1). The circadian clock genes provide an example of this reduction as both genomes lack several components (*pex*, *kaiA*) found in the model *Synechococcus* PCC 7942[23]. However genes for the core clock proteins (*kaiB*, *kaiC*) remain in both genomes, and *Prochlorococcus* cell division is tightly synchronized to the diel light-dark cycle[24]. Thus, loss of some circadian components may imply an alternate signalling pathway for circadian control.

Gene loss may also play a role in the lower % G+C content of MED4 (30.8 %) compared to that of MIT9313 (50.74%) which is more typical of marine *Synechococcus*. MED4 lacks genes for several DNA repair pathways including recombinational repair (*recJ*, *recQ*), and damage reversal (*mutT*). Particularly, the loss of the base excision repair gene *mut*Y, which removes adenosines incorrectly paired with oxidatively damaged guanine residues may imply an increased rate of G·C --> T·A transversions[25]. The tRNA complement of MED4 is largely identical to MIT 9313 and not optimized for a low % G+C genome suggesting it is not evolving as fast as codon usage.

Analysis of the nitrogen acquisition capabilities of the two strains points to a sequential decay in the capacity to utilize nitrate and nitrite during the evolution of the *Prochlorococcus* lineage (Fig. 3a). In *Synechococcus* WH 8102 — representing the presumed ancestral state — many nitrogen acquisition and assimilation genes are grouped together (Fig. 3a). MIT 9313 has lost a 25-gene cluster which includes genes encoding the nitrate/nitrite transporter and nitrate reductase. The nitrite reductase gene has been retained in MIT 9313, but it is flanked by a proteobacterial-like nitrite transporter rather than a typical cyanobacterial nitrate/nitrite permease (Supp. Fig. 4), suggesting acquisition by lateral gene transfer. An additional deletion event occurred in MED4, in which the nitrite reductase gene was also lost (Fig. 3a). As a result of these serial deletion events MIT 9313 cannot utilize nitrate, and MED4 cannot utilize nitrate or nitrite[9]. Thus each *Prochlorococcus* ecotype uses the N-species that is most prevalent at the light levels to which they are best adapted: ammonium in the surface waters, and nitrite at depth (Fig. 1a). *Synechococcus* — which is the only one of the three that has nitrate reductase — is able to bloom when nitrate is upwelled (Fig 1a), as occurs in the spring in the N. Atlantic[3] and the N. Red Sea[26].

The two *Prochlorococcus* strains are also less versatile in their organic N utilization capabilities than *Synechococcus* WH 8102[13]. MED4 contains the genes necessary for utilization of urea, cyanate and oligopeptides, but no monomeric amino acid transporters have been identified. In contrast, MIT 9313 contains transporters for urea, amino acids and oligopeptides but lacks the genes necessary for cyanate utilization (cyanate transporter and cyanate lyase) (Fig. 3a). As expected, both genomes contain the high affinity ammonium transporter *amt1* and both lack the nitrogenase genes essential for nitrogen fixation. Finally, both contain the nitrogen transcriptional regulator encoded by *ntcA* and there are numerous genes in both genomes, including *ntcA, amt1*, the urea transport and GS/GOGAT genes (glutamine synthetase and glutamate synthase, both involved in ammonia assimilation), with an upstream NtcA binding site consensus sequence.

The genomes also have differences in genes involved in phosphorus utilization that have obvious ecological implications. MED4, but not MIT9313, is capable of growth on organic P sources (L. R. Moore and S. W. Chisholm, unpublished data), and organic P can be the prevalent form of P in high light surface waters[27]. This difference may be due to the acquisition of an alkaline phosphatase like gene in MED4 (Supp. Fig. 5). Both genomes contain the high affinity phosphate transport system encoded by *pstS* and *pstABC*[28], but MIT 9313 contains an additional copy of the phosphate binding component *pstS*, perhaps reflecting an increased reliance on orthophosphate in deeper waters. MED4 contains several P related regulatory genes including the *phoB*, *phoR* two component system and the transcriptional activator *ptrA*. In MIT 9313, however, *phoR* is interrupted by 2 frameshifts and *ptrA* is further degenerated, suggesting that this strain has lost the ability to regulate gene expression in response to changing P levels.

Both *Prochlorococcus* strains have iron-related genes missing in *Synechococcus* WH 8102, which may explain its dominance in the iron limited Equatorial Pacific[2]. These genes include flavodoxin (*isiB*) a Fe-free electron transfer protein capable of replacing ferredoxin, and ferritin (located with the ATPase component of an iron ABC transporter), an iron binding molecule implicated in iron storage. Additional characteristics of the iron acquisition system in these genomes include: an Fe-induced transcriptional regulator (Fur) that represses iron uptake genes; numerous genes with an upstream putative *fur* box motif that are candidates for a high affinity iron scavenging system; and absence of genes involved in Fe:siderophore complexes.

Prochlorococcus does not utilize typical cyanobacterial genes for inorganic carbon concentration or fixation. Both genomes contain a sodium/bicarbonate symporter but lack homologs to known families of carbonic anhydrases, suggesting an as yet unidentified gene is fulfilling this function. One of the two carbonic anhydrases in *Synechococcus* WH 8102 was lost in the deletion event that led to the loss of the nitrate reductase (Fig. 3a); the other is located next to a tRNA and appears to have been lost during a genome rearrangement event. Like other *Prochlorococcus* and marine *Synechococcus,* MED4 and MIT9313 possess a form IA Rubisco, rather than the typical cyanobacterial form IB. The Rubisco genes are adjacent to genes encoding structural carboxysome shell proteins and all have phylogenetic affinity to genes in the gamma-proteobacterium *Acidithiobacillus ferroxidans*[15], suggesting lateral transfer of the extended operon.

Prochlorococcus has been identified in deep sub-oxic zones where it is unlikely they can sustain themselves by photosynthesis alone[29], thus we looked for genomic evidence of heterotrophic capability. Indeed, the presence of oligopeptide transporters in both genomes, and the larger proportion of transporters (including some sugar transporters) in the MIT 9313 strain specific genes (Supp. Fig. 2), suggests the potential for partial heterotrophy. Neither genome contains known pathways that would allow for complete heterotrophy, however. They are both missing genes for steps in the tricarboxylic acid cycle, including 2-oxoglutarate dehydrogenase, succinyl-CoA synthetase and succinyl-CoA-acetoacetate-CoA transferase.

Cell surface chemistry plays a major role in phage recognition and grazing by protists and thus is likely to be under intense selective pressure in nature. The two *Prochlorococcus* and *Synechococcus* WH 8102 genomes show evidence of extensive lateral gene transfer and deletion events of genes involved in lipopolysaccharide and/or surface polysaccharide biosynthesis, reinforcing the role of predation pressures in the creation and maintenance of microdiversity. For example, MIT 9313 has a 41.8 kbp cluster of surface polysaccharide genes (Fig. 3b), that has a lower %G+C composition (42 %) than the genome as a whole, implicating acquisition by lateral gene transfer. MED4 has acquired a 74.5 kbp cluster consisting of 67 potential surface polysaccharide genes (Supp. Fig. 6a) and lost another cluster of surface polysaccharide biosynthesis genes shared between MIT 9313 and *Synechococcus* WH 8102 (Supp. Fig. 6b).

The approach we have taken in describing these genomes highlights the known drivers of niche partitioning of these closely related organisms (Fig. 1). Detailed comparisons with the genomes of additional strains, such as *Prochlorococcus* SS120[30], will enrich this story, and the analysis of whole genomes from *in situ* populations will be necessary to understand the full expanse of genomic diversity in this group. The genes of unknown function in all of these genomes hold important clues for undiscovered niche dimensions in the marine pelagic zone. As we unveil their function we will undoubtedly learn that the suite of selective pressures that shape these communities is much larger than we have imagined. Finally, it may be useful to view *Prochlorococcus* and *Synechococcus* as important 'minimal life units', as the information in their roughly 2000 genes is sufficient to create globally abundant biomass from solar energy and inorganic compounds.

Methods

Genome Sequencing and Assembly. DNA was isolated from the clonal, axenic strain MED4 and the clonal strain MIT 9313 essentially as described previously[4]. The two whole genome shotgun libraries were obtained by fragmenting genomic DNA using mechanical shearing, and cloning 2-3 kb fragments into pUC18. Double-ended plasmid sequencing reactions were carried out using PE BigDye Terminator chemistry (Perkin Elmer, Foster City, CA) and sequencing ladders were resolved on PE 377 Automated DNA Sequencers. The whole genome sequence of *Prochlorococcus* MED4 was obtained from 27,065 end sequences (7.3 fold redundancy), while *Prochlorococcus* MIT 9313 was sequenced to 6.2X coverage (33,383 end sequences). For *Prochlorococcus* MIT 9313, supplemental sequencing (0.05X sequence coverage) of a pFos1 fosmid library was used as a scaffold. Sequence assembly was accomplished using PHRAP (P. Green, University of Washington). All gaps were closed by primer walking on gap-spanning library clones or PCR products. The final assembly of *Prochlorococcus* MED4 was verified by long range genomic PCR reactions, while the assembly of *Prochlorococcus* MIT 9313 was confirmed by comparison to the fosmid clones, which were fingerprinted with *Eco*RI.. No plasmids were detected in the course of genome sequencing and insertion sequences, repeated elements, transposons, and prophage are notably absent from both genomes. The likely origin of replication in each genome was identified based on GC skew and base pair 1 was designated adjacent to the *dnaN* gene.

Genome Annotation. The combination of three gene modelers, Critica, Glimmer, and Generation, were used in the determination of potential open reading frames and checked manually. A revised gene/protein set was searched against the KEGG GENES, Pfam, PROSITE, PRINTS, ProDom, COGs and CyanoBase databases, in addition to BLASTP vs. NR. From these results, categorizations were developed using the KEGG and COGs hierarchies, as modified in CyanoBase. Manual annotation of open reading frames was done in conjunction with the *Synechococcus* team. The three way genome comparison was used to refine predicted start sites, add additional open reading frames and standardize the annotation across the three genomes.

Genome Comparisons. The comparative genome architecture of MED4 and MIT 9313 was visualized using the Artemis Comparison Tool (ACT) (http://www.sanger.ac.uk/Software/ACT/). Orthologs were determined by aligning the predicted coding sequences of each gene with the coding sequences of the other genome using BLASTP. Genes were considered orthologs if each was the best hit of the other one and both e-values were less than e^{-10}. In addition, bidirectional best hits with e-values less than e^{-6} and small proteins of conserved function were manually examined and added to the ortholog lists.

Phylogenetic analyses employed PAUP* and used logdet distances and minimum evolution as the objective function. The degree of support at each node was evaluated using 1000 bootstrap resamplings. Ribosomal DNA analyses employed 1160 positions. The gram positive bacterium *Arthrobacter globiformis* was used to root the tree.

1. Chisholm, S. W. et al. A novel free-living prochlorophyte abundant in the oceanic euphotic zone. *Nature* **334**, 340-343 (1988).

2. Campbell, L., Liu, H., Nolla, H. & Vaulot, D. Annual variability of phytoplankton and bacteria in the subtropical North Pacific Ocean at Station ALOHA during the 1991-1994 ENSO event. *Deep Sea Research* **44**, 167-192 (1997).

3. DuRand, M. D., Olson, R. J. & Chisholm, S. W. Phytoplankton population dynamics at the Bermuda Atlantic Time-series Station in the Sargasso Sea. *Deep Sea Research II* **48**, 1983-2003 (2001).

4. Rocap, G., Distel, D. L., Waterbury, J. B. & Chisholm, S. W. Resolution of *Prochlorococcus* and *Synechococcus* ecotypes by using 16S-23S rDNA Internal Transcribed Spacer (ITS) sequences. *Applied and Environmental Microbiology* **68**, 1180-1191 (2002).

5. Moore, L. R. & Chisholm, S. W. Photophysiology of the Marine Cyanobacterium *Prochlorococcus:* Ecotypic differences among cultured isolates. *Limnology and Oceanography* **44**, 628-638 (1999).

6. Moore, L. R., Rocap, G. & Chisholm, S. W. Physiology and Molecular Phylogeny of Coexisting *Prochlorococcus* Ecotypes. *Nature* **393**, 464-467 (1998).

7. Moore, L. R., Goericke, R. E. & Chisholm, S. W. Comparative physiology of *Synechococcus* and *Prochlorococcus*: influence of light and temperature on growth, pigments, fluorescence and absorptive properties. *Marine Ecology Progress Series* **116**, 259-275 (1995).

8. Mann, E. L., Ahlgren, N., Moffett, J. W. & Chisholm, S. W. Copper toxicity and cyanobacteria ecology in the Sargasso Sea. *Limnology and Oceanography* **47**, 976-988 (2002).

9. Moore, L. R., Post, A. F., Rocap, G. & Chisholm, S. W. Utilization of different nitrogen sources by the marine cyanobacteria, *Prochlorococcus* and *Synechococcus*. *Limnology and Oceanography* **47**, 989-996 (2002).

10. Sullivan, M. B., Waterbury, J. B. & Chisholm, S. W. Cyanophage infecting the oceanic cyanobacterium, *Prochlorococcus*. *Nature* **in press** (2003).

11. Hagström, Å. et al. Use of 16S Ribosomal DNA for Delineation of Marine Bacterioplankton Species. *Applied and Environmental Microbiology* **68**, 3628-3633 (2002).

12. West, N. J. et al. Closely related *Prochlorococcus* genotypes show remarkably different depth distributions in two oceanic regions as revealed by in situ hybridization using 16S rRNA-targeted oligonucleotides. *Microbiology* **147**, 1731-1744 (2001).

13. Palenik, B. et al. The genome of a motile marine *Synechococcus*. *Nature* **in press** (2003).

14. Reiter, W. D., Palm, P. & Yeats, S. Transfer RNA genes frequently serve as integration sites for prokaryotic genetic elements. *Nucleic Acids Research* **17** (1989).

15. La Roche, J. et al. Independent evolution of the prochlorophyte and green plant chlorophyll a/b light harvesting proteins. *Proc. Natl. Acad. Sci. USA* **93**, 15244-15248 (1996).

16. Hess, W. et al. The photosynthetic apparatus of *Prochlorococcus*: Insights through comparative genomics. *Photosynthesis Research* **70**, 53-71 (2001).

17. Stickforth, P., Steiger, S., Hess, W. R. & Sandmann, G. A novel type of lycopene ε-cyclase in the marine cyanobacterium *Prochlorococcus marinus* MED4. *Archives of Microbiology* **179**, 407-415 (2003).

18. Frankenberg, N., Mukougawa, K., Kohchi, T. & Lagarias, J. C. Functional genomic analysis of the HY2 family of ferredoxin-dependent bilin reductases from oxygenic photosynthetic organisms. *The Plant Cell* **13**, 965-978 (2001).

19. Ting, C., Rocap, G., King, J. & Chisholm, S. W. Phycobiliprotein genes of the marine prokaryote *Prochlorococcus*: Evidence for rapid evolution of genetic heterogeneity. *Microbiology* **147**, 3171-3182 (2001).

20. Golden, S. S., Brusslan, J. & Haselkorn, R. Expression of a family of *psbA* genes encoding a photosystem II polypeptide in the cyanobacterium *Anacystis nidulans* R2. *EMBO J* **5**, 2789-2798 (1986).

21. Shen, J. R., Qian, M., Inoue, Y. & L., B. R. Functional characterization of *Synechocystis* sp. PCC 6803 delta *psbU* and delta *psbV* mutants reveals important roles of cytochrome c-550 in cyanobacterial oxygen evolution. *Biochemistry* **37**, 1551-1558 (1998).

22. Garczarek, L., van der Staay, G. W. M., Hess, W. R., Le Gall, F. & Partensky, F. Expression and phylogeny of the multiple antenna genes of the low-light -adapted strain *Prochlorococcus marinus* SS120 (Oxyphotobacteria). *Plant Molecular Biology* **46**, 683-693 (2001).

23. Ishiura, M. et al. Expression of a Gene Cluster *kaiABC* as a Circadian Feedback Process in Cyanobateria. *Science* **281**, 1519-1523 (1998).

24. Vaulot, D., Marie, D., Olson, R. J. & Chisholm, S. W. Growth of *Prochlorococcus,* a Photosynthetic Prokaryote, in the Equatorial Pacific Ocean. *Science* **268**, 1480-1482 (1995).

25. Michaels, M. L. & Miller, J. H. The GO system protects organisms from the mutagenic effect of the spontaneous lesion 8-Hydroxyguanine (7,8-Dihydro-8-Oxoguanine). *Journal of Bacteriology* **174**, 6321-6325 (1992).

26. Lindell, D. & Post, A. F. Ultraphytoplankton sucession is triggered by deep winter mixing in the Gulf of Aqaba (Eilat), Red Sea. *Limnology and Oceanography* **40**, 1130-1141 (1995).

27. Karl, D. M., Bidigare, R. R. & Letelier, R. M. Long-term changes in plankton community structure and productivity in the North Pacific Subtropical Gyre: The domain shift hypothesis. *Deep Sea Research Part II* **48**, 1449-1470 (2001).

28. Scanlan, D. J., Mann, N. H. & Carr, N. G. The response of the picoplanktonic marine cyanobacterium *Synechococcus* species WH7803 to phosphate starvation involves a protein homologous to the periplasmic phosphate binding protein of *Escherichia coli*. *Molecular Microbiology* **10**, 181-191 (1993).

29. Johnson, Z. et al. Energetics and growth kinetics of a deep *Prochlorococcus* spp. population in the Arabian Sea. *Deep Sea Research II* **46**, 1719-1743 (1999).

30. Dufresne, A. et al. Genome sequence of the cyanobacterium *Prochlorococcus marinus* SS120, a near minimal oxyphototrophic genome. *PNAS* **submitted** (2003).

Supplementary Information accompanies the paper on Nature's website (http://www.nature.com).

Acknowledgements This research was funded by the Biological and Environmental Research Program of the U.S. Department of Energy's Office of Science. The Joint Genome Institute managed the overall sequencing effort. Genome finishing was carried out under the auspices of the U.S. Department of Energy by the University of California, Lawrence Livermore National Laboratory. Computational annotation was carried out at the Oak Ridge National Laboratory, managed by UT-BATTELLE for the U.S. Department of Energy. Additional support was provided by the DOE and NSF to SWC, the Israel-US Binational Science Foundation to AFP and SWC, and FP5-MARGENES to WRH and AFP.

We thank the *Synechococcus* WH 8102 annotators (B. Palenik, B. Brahamsha, J. McCarren, E. Allen, F. Partensky, A. Dufresne and I. Paulsen) for their help with curating the *Prochlorococcus* genomes and E. V. Armbrust and L. Moore for critical reading of the manuscript.

Competing interests statement. The authors declare they have no competing financial interests.

Correspondence and requests for materials should be addressed to S.W.C. (chisholm@mit.edu). The complete nucleotide sequences and sequences of predicted orfs have been deposited in GenBank under the accession numbers BX548174 (MED4) and BX548175 (MIT 9313).

Table 1 General Features of the *Prochlorococcus* genomes

	MED4	**MIT 9313**
Length (bp)	1657990	2410873
G+C content (%)	30.8	50.7
Protein Coding (%)	88	82
Protein Coding Genes	1716	2275
With Assigned function	1134	1366
Conserved hypothetical	502	709
Hypothetical	80	197
Genes with ortholog in		
Prochlorococcus MED4	----	1352
Prochlorococcus MIT 9313	1352	----
Synechococcus WH8102	1394	1710
Genes without ortholog in		
MED4 and WH 8102	----	527
MIT 9313 and WH 8102	284	----
Transfer RNA	37	43
Ribosomal RNA operons	1	2
Other Structural RNAs	3	3

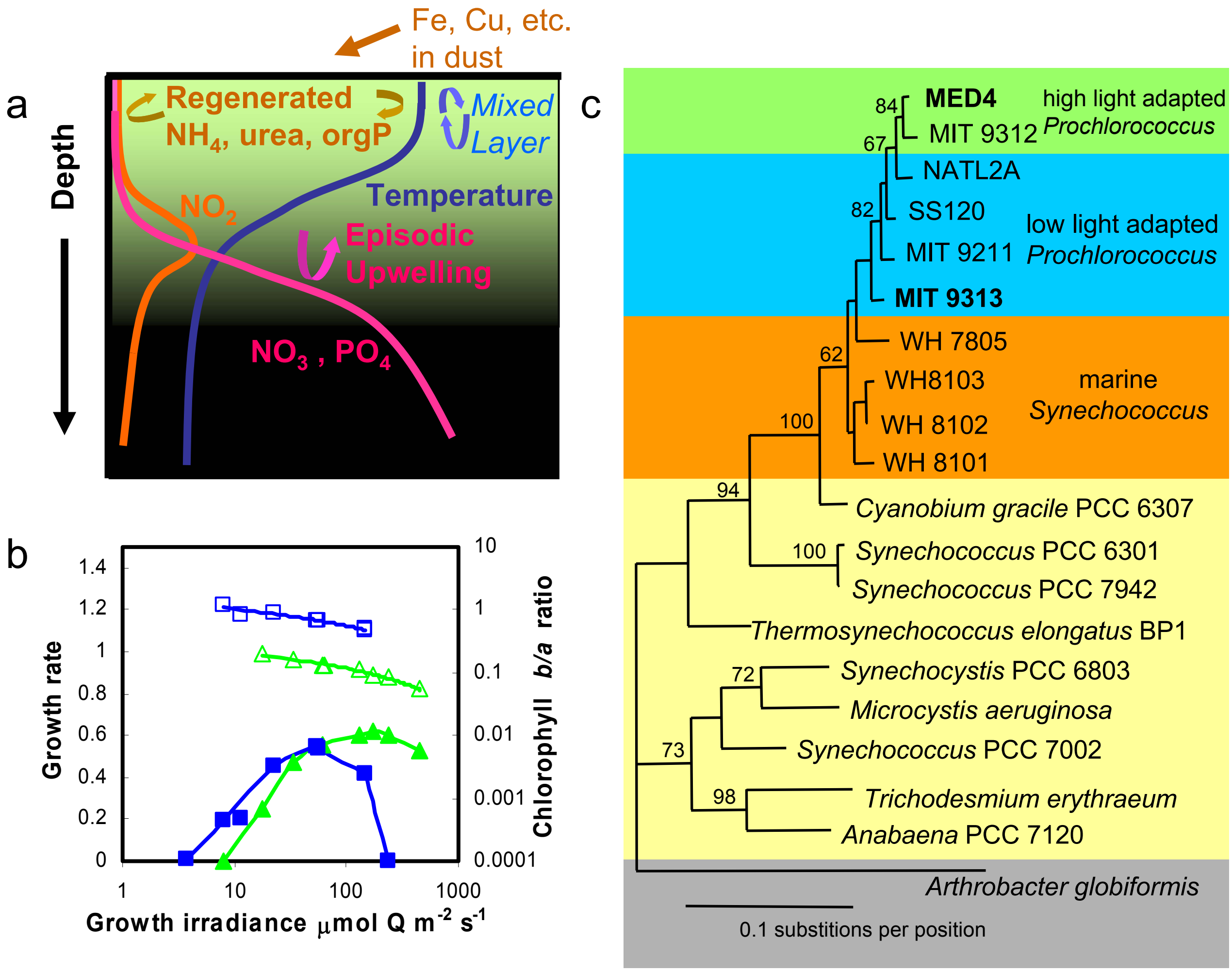
Figure 1 Ecology, physiology and phylogeny of *Prochlorococcus* ecotypes. **a,** Schematic stratified open ocean water column illustrating vertical gradients allowing niche differentiation. Shading represents degree of light penetration. Temperature and salinity gradients provide a mixing barrier, isolating the low nutrient/high light surface layer from the high nutrient/low light deep waters. Photosynthesis in surface waters is driven primarily by rapidly regenerated nutrients, punctuated by episodic upwelling. **b**, Growth rate (solid symbols) and chlorophyll *b/a* ratio (open symbols) as a function of growth irradiance for MED4[7] (green) and MIT 9313[6] (blue). **c**, Relationships between *Prochlorococcus* and other cyanobacteria inferred using 16S rDNA.

Figure 2 Global genome alignment as seen from start positions of orthologous genes. Genes present in one genome but not the other are shown on the axes. The "broken X" pattern has been noted before for closely related bacterial genomes, and is likely due to multiple inversions centered around the origin of replication. Alternating slopes of many adjacent gene clusters indicate multiple smaller scale inversions have also occurred.

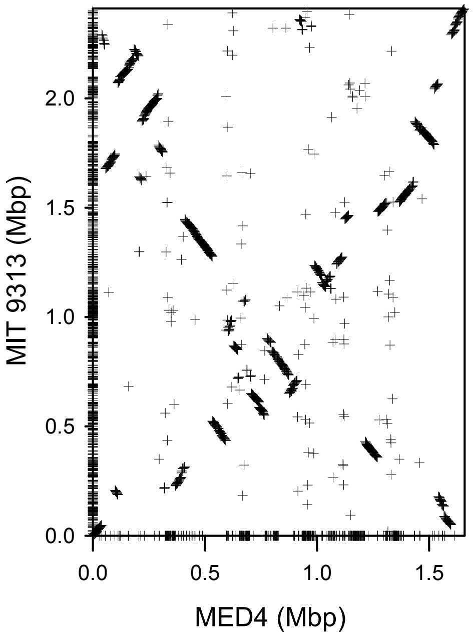
Figure 3 Dynamic architecture of marine cyanobacterial genomes. **a,** Deletion, acquisition and rearrangement of nitrogen utilization genes. In MIT 9313, 25 genes including the nitrate/nitrite transporter (*nrtP*/*napA*), nitrate reductase (*narB*), and carbonic anhydrase have been deleted. The cyanate transporter and cyanate lyase (*cynS*) were likely lost after the divergence of MIT 9313 from the rest of the *Prochlorococcus* lineage, as MED4 possesses these genes. MIT 9313 has retained nitrite reductase (*nirA*) and acquired a nitrite transporter. In MED4 *nirA* has been lost and the urea transporter (*urt* cluster) and urease (*ure* cluster) genes have been rearranged (dotted line). **b,** Lateral transfer of genes involved in LPS biosynthesis including sugar transferases, sugar epimerases,

modifying enzymes, and two pairs of ABC-type transporters. blue, genes in all three genomes; pink, genes hypothesized to have been laterally transferred; red, tRNAs; white, other genes. %G+C content in MIT9313 along this segment is lower (42%) than the whole genome average (horizontal line).

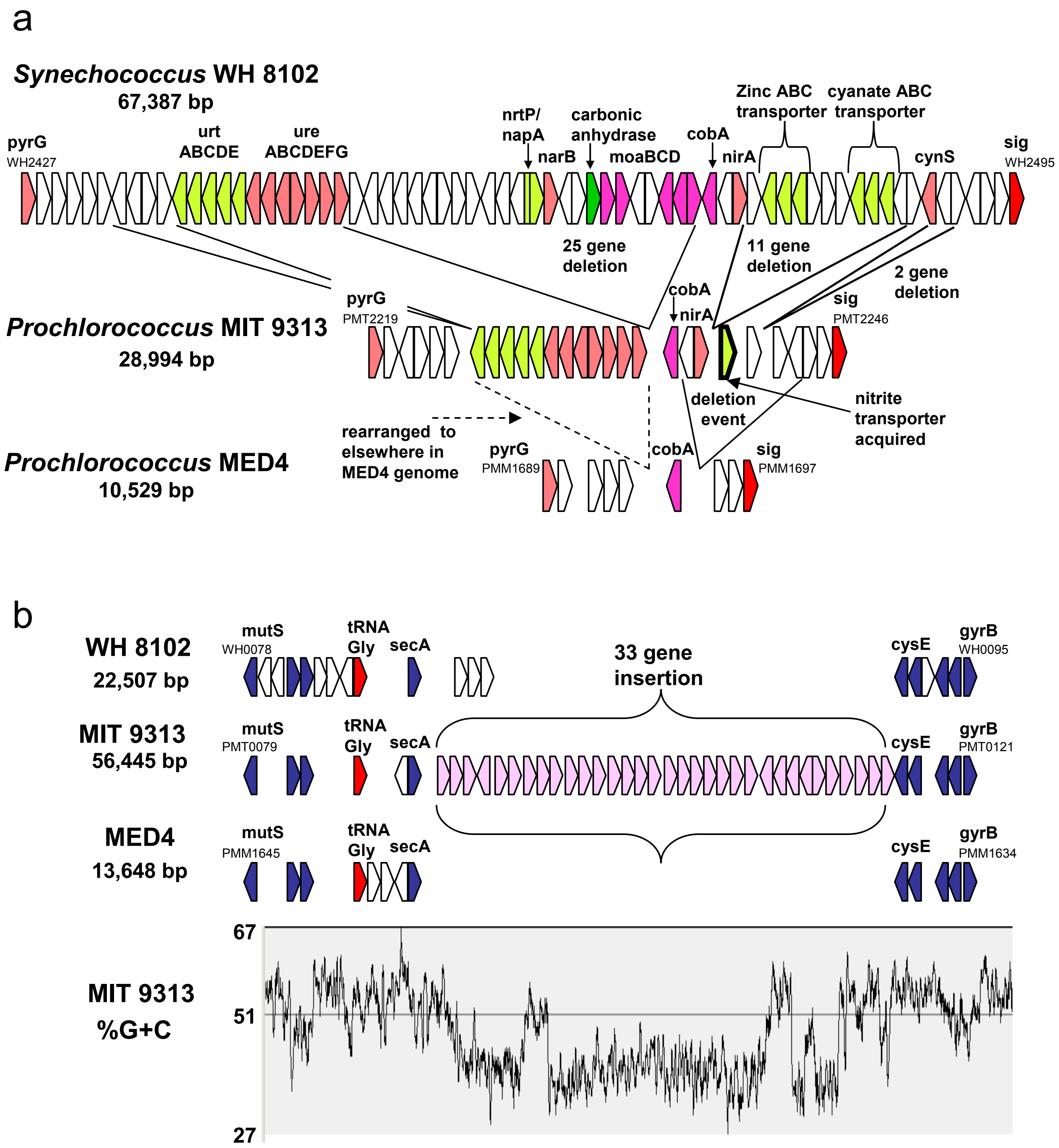
2003-05-04442A-Chisholm Figure 1



2003-05-04442A Chisholm Figure 2



2003-05-04442A-Chisholm Figure 3



Chisholm-2003-05-04442A

Supplementary Figure Legends

Supp. Figure 1. Circular representation of the *Prochlorococcus* genomes. **a,** MED4. **b,** MIT 9313. For both genomes outermost circles (1 and 2) are predicted protein coding regions on the plus and minus strands, respectively. Color coding is as in Supplementary Figure 2. The next two circles show genes not present in the other *Prochlorococcus* genome on the plus (circle 3) and minus (circle 4) strands. Circles 5 and 6 show genes on the plus and minus strands, respectively that contain transmembrane domains. Circle 7 is % G+C content (deviation from average). Innermost circle (8) represents the GC skew curve.

Supp. Figure 2. Functional categorization of predicted open reading frames in the *Prochlorococcus* genomes, following the classification scheme used by CyanoBase. **a,** MED4, entire genome. **b,** MIT 9313, entire genome. **C,** Genes present in both MED4 and MIT 9313. **d,** Genes in MED4 not present in MIT 9313. **e,** Genes in MIT 9313 not present in MED4.

Supp. Figure 3. Comparison of *Prochlorococcus* MED4 and MIT 9313 open reading frames with those of other complete prokaryotic genomes. The predicted coding sequences of each gene in both genomes were aligned with the coding sequences of 90 bacterial genomes using BLASTP. Significant alignments were defined as having an e-value less than 10^{-6}. The bacterial genomes comprised the 89 completed bacterial genomes available from ftp.ncbi.nih.gov/genbank/genomes/Bacteria on 30 October 2002 and *Synechococcus* WH 8102[8]. **a,** MED4, entire genome. **B,** MIT 9313, entire genome. **c,** MED4 genes present in MIT 9313 **c,** MIT 9313 genes present in

MED4 **e,** Genes in MED4 not present in MIT 9313 **f,** Genes in MIT 9313 not present in MED4.

Supp. Figure 4 Alignment of the putative nitrite transporter in *Prochlorococcus* MIT9313 (PMT2240) with its most significant matches in the NR database (all proteobacteria) and with cyanobacterial nitrate/nitrate transporters. The MIT 9313 gene has a formate/nitrite transporter domain (Pfam PF01226) in contrast to the cyanobacterial nitrate transporters which are permeases of the major facilitator superfamily (Pfam PF00083). Furthermore, the MIT 9313 gene has no significant matches (BLASTP evalue < e-2) in the genomes of *Prochlorococcus* MED4, *Synechococcus* WH8102, *Synechocystis* sp. PCC 6803, *Thermosynechococcus elongatus* BP-1, or *Anabaena* sp. PCC 7120 suggesting it may have been acquired via lateral gene transfer. Alignment generated using ClustalW. Shaded residues indicate >50% similarity. Abbreviations and accession numbers as follows: Rhodopseud., *Rhodopseudomonas palustris* (ZP_00012718.1); Bradyrhiz., *Bradyrhizobium japonicum* (NP_769441); Vibrio, *Vibrio vulnificus* (NP_762336.1); Nitros., *Nitrosomonas europaea* (NP_840759); WH 7803, *Synechococcus* WH 7803 napA (AAG45172); PCC 7002, *Synechococcus* PCC 7002 nrtP (AAD45941); WH9601, *Trichodesmium* WH 9601 napA (AAF00917); PCC 73102, *Nostoc punctiforme* PCC 73102 (ZP_00107423).

Supp. Figure 5 Phylogenetic tree showing the relationship of a possible alkaline phosphatase like gene in *Prochlorococcus* MED4 (PMM0708) with the most significant matches in the NR database, which include several proteobacterial sequences, and with the atypical alkaline phosphatase of *Synechococcus* PCC 7942 and related cyanobacterial genes. Accession numbers as follows: *Brucella melitensis* (NP_541633.1), *Agrobacterium tumefaciens* str. C58 (NP_531956.1); *Sinorhizobium meliloti*, (NP_385365.1); *Vibrio vulnificus* (NP_762849.1), *Streptomyces coelicolor* A3(2) (NP_624650.1), Shewanella oneidensis MR-1 (NP_717877.1) *Anabaena* PCC 7102 (NP_489331.1), *Synechocystis* sp. PCC 6803 (NP_440276); *Synechococcus* sp. PCC 7942 (A47026).

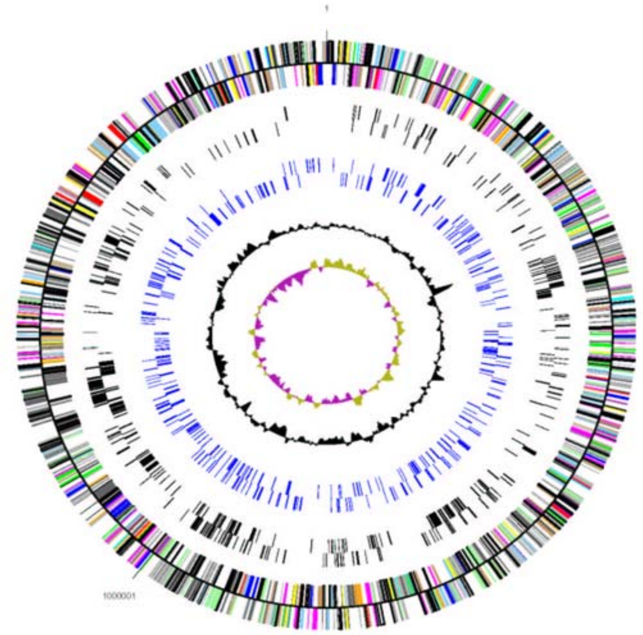
Supp. Figure 6 Insertions, deletions and rearrangements of genes involved in lipopolysaccharide biosynthesis (LPS clusters) in MED4. Color coding is as follows: blue, orthologous genes present in all three genomes; pink, genes hypothesized to be part of lateral transfer events, many have roles in LPS biosynthesis; red, tRNAs; green, orthologous genes present in two genomes, many have roles in LPS biosynthesis; white, other genes. Length in bp represents the size of the region shown for each genome. **a,** Insertion of a 74.5 kbp cluster of LPS genes in MED4, roughly between two tRNAs. The 67 potential surface polysaccharide genes in this cluster include sugar transferases, sugar epimerases, and modifying enzymes such as aminotransferases, methyltransferases, carbamoyltransferases, and acetyltransferases. In MIT 9313 and WH 8102 the genes that flank this insertion are rearranged to other parts of the genome. **b,** Deletion of LPS biosynthesis genes in MED4. LPS related genes present in MIT 9313 and WH 8102, several of which have homologs in the acquired genes shown in part a, have been deleted. In this region a selenophosphate synthase (*selD*) and a tRNA nucleotidyl-transferase in the center of the cluster have been retained suggesting that they are essential genes and separate deletion events have occurred on either side of them.

Supp. Table 1 Number of predicted signal transduction and transcription factors suggests reduced regulatory capacity in *Prochlorococcus*

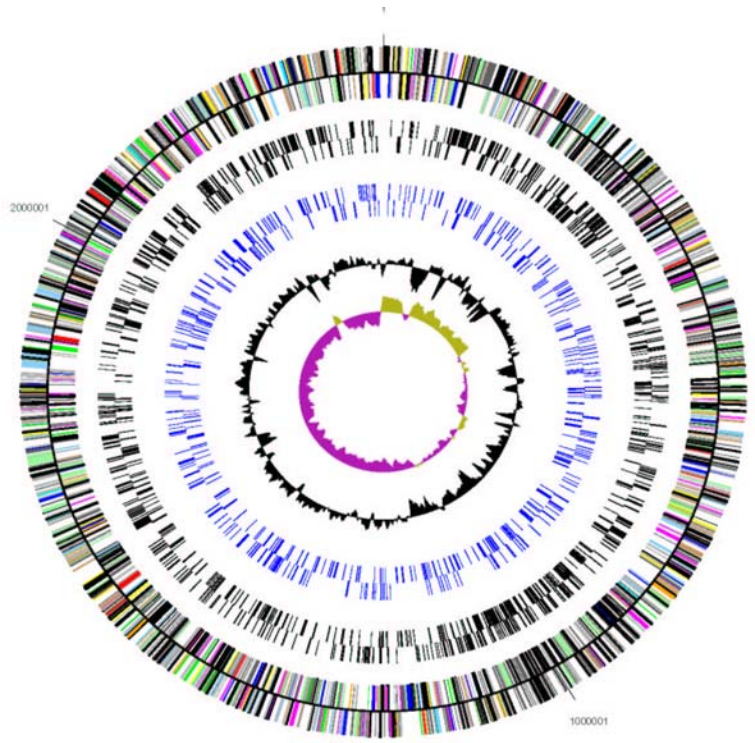
	MED4	MIT 9313	*T. elongatus*
Sigma Factors	5	8	8
Two Component systems			
Histidine Kinases	4	5	17
Response regulators	6	8	27
Ser/Thr protein Kinases	0	1	11
Transcription Factors			
LuxR family	2	5	4
LysR family	1	1	3
CRP family	3	4	3
ArsR family	1	2	2
FUR family	2	3	3
Other	2	3	3
Light sensors/transducers			
Cryptochrome	2	0	2
Bacteriophytochrome	0	0	5
Phototropin	0	0	1

Chisholm Supplementary Figure 1

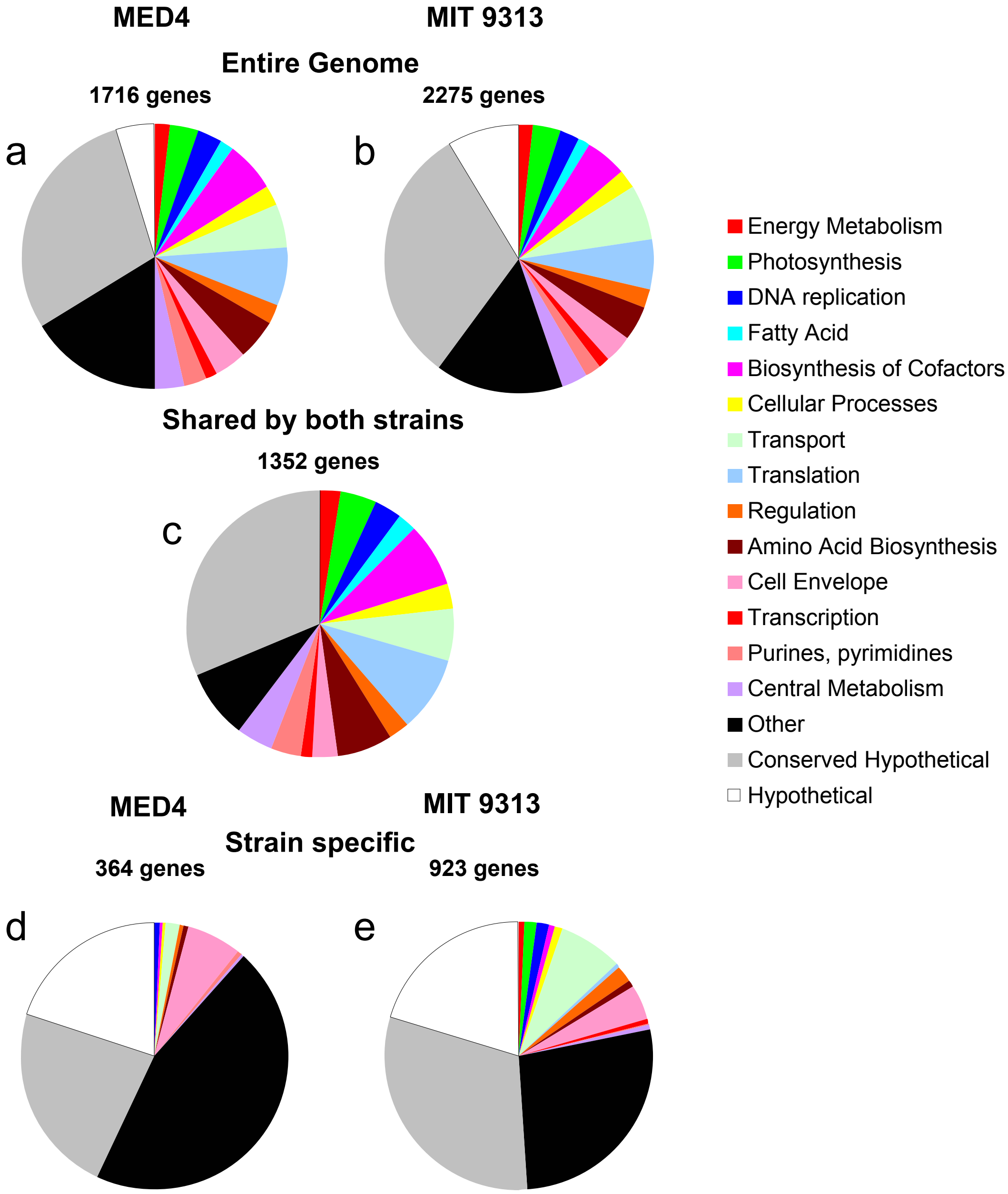
a



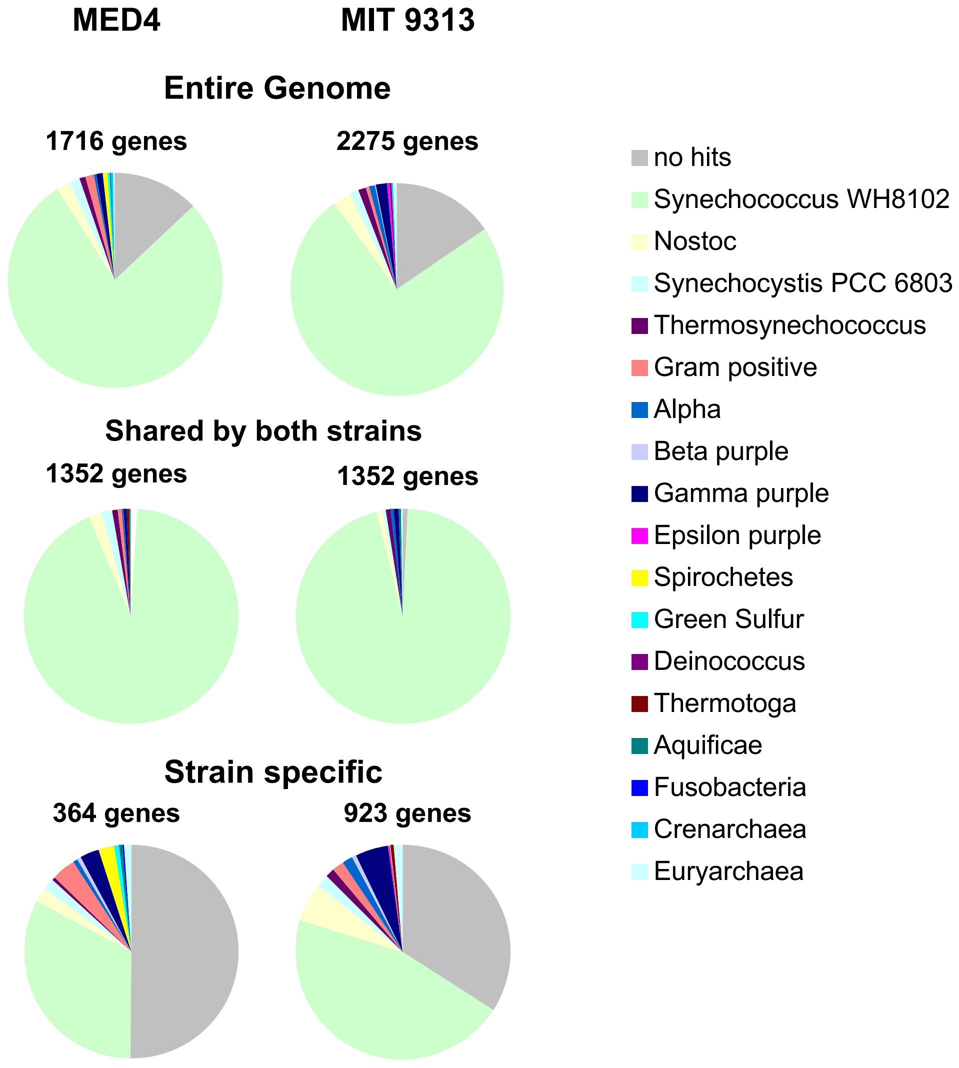
b



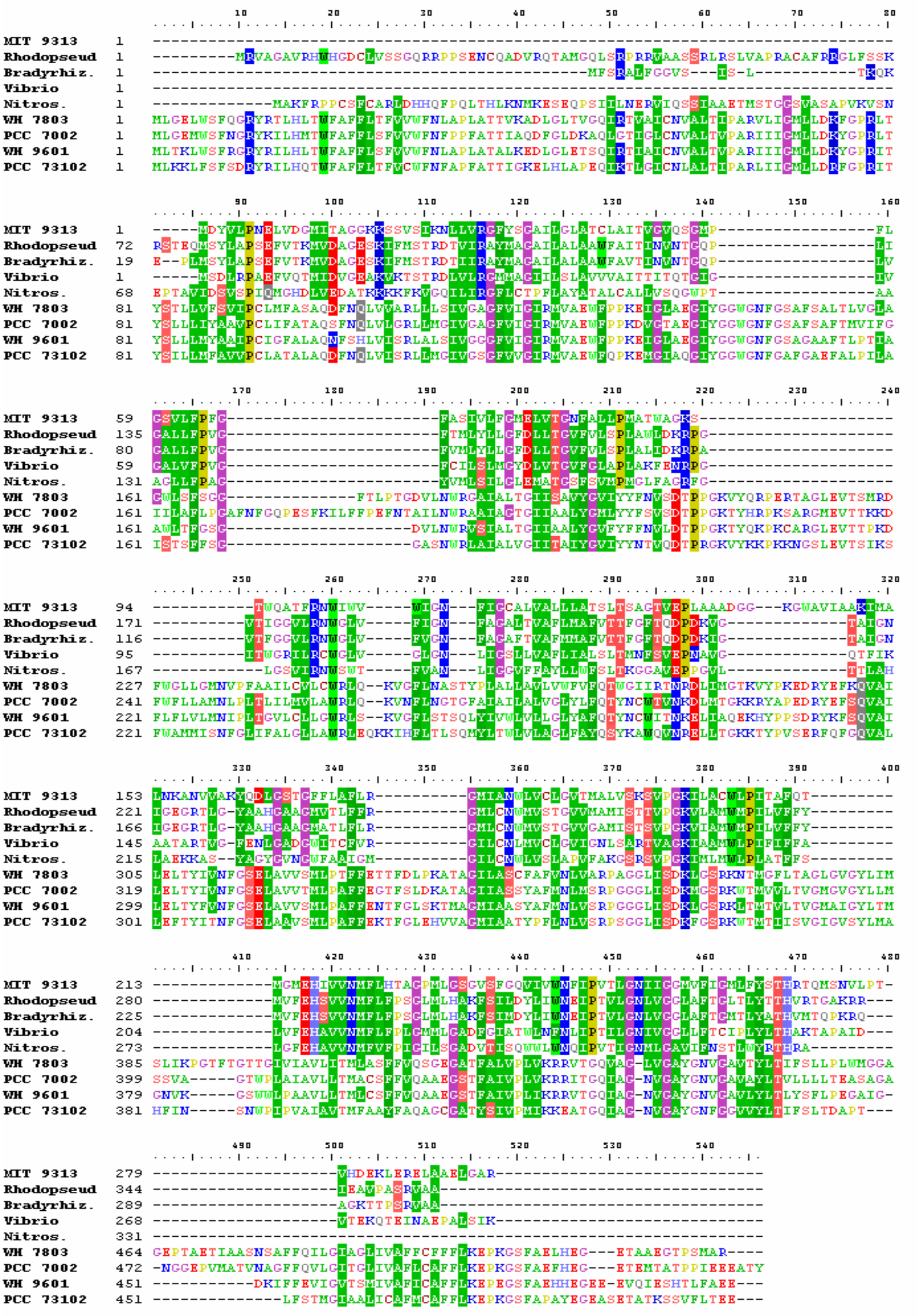
Chisholm Supplementary Figure 2



Chisholm Supplementary Figure 3

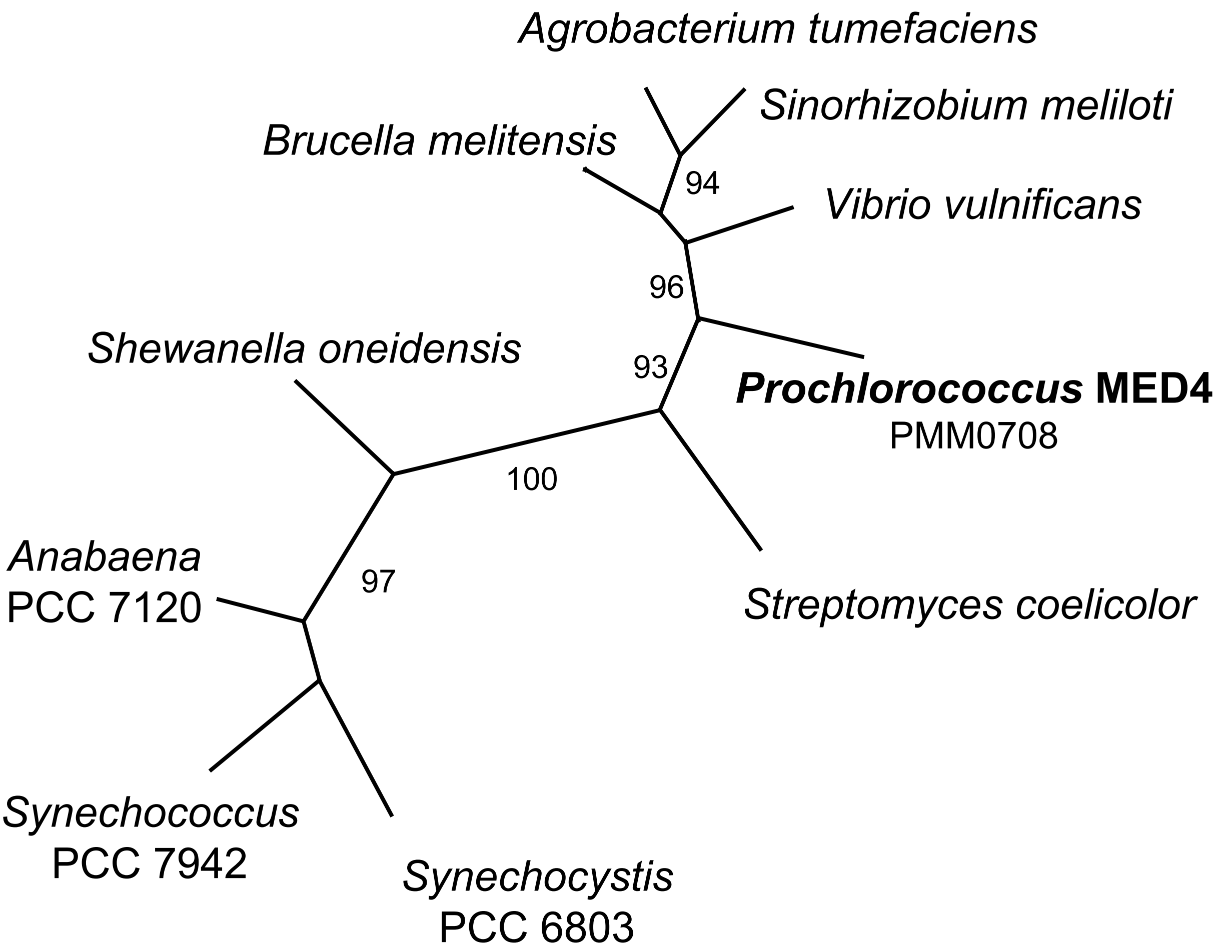


Chisholm Supplementary Figure 4

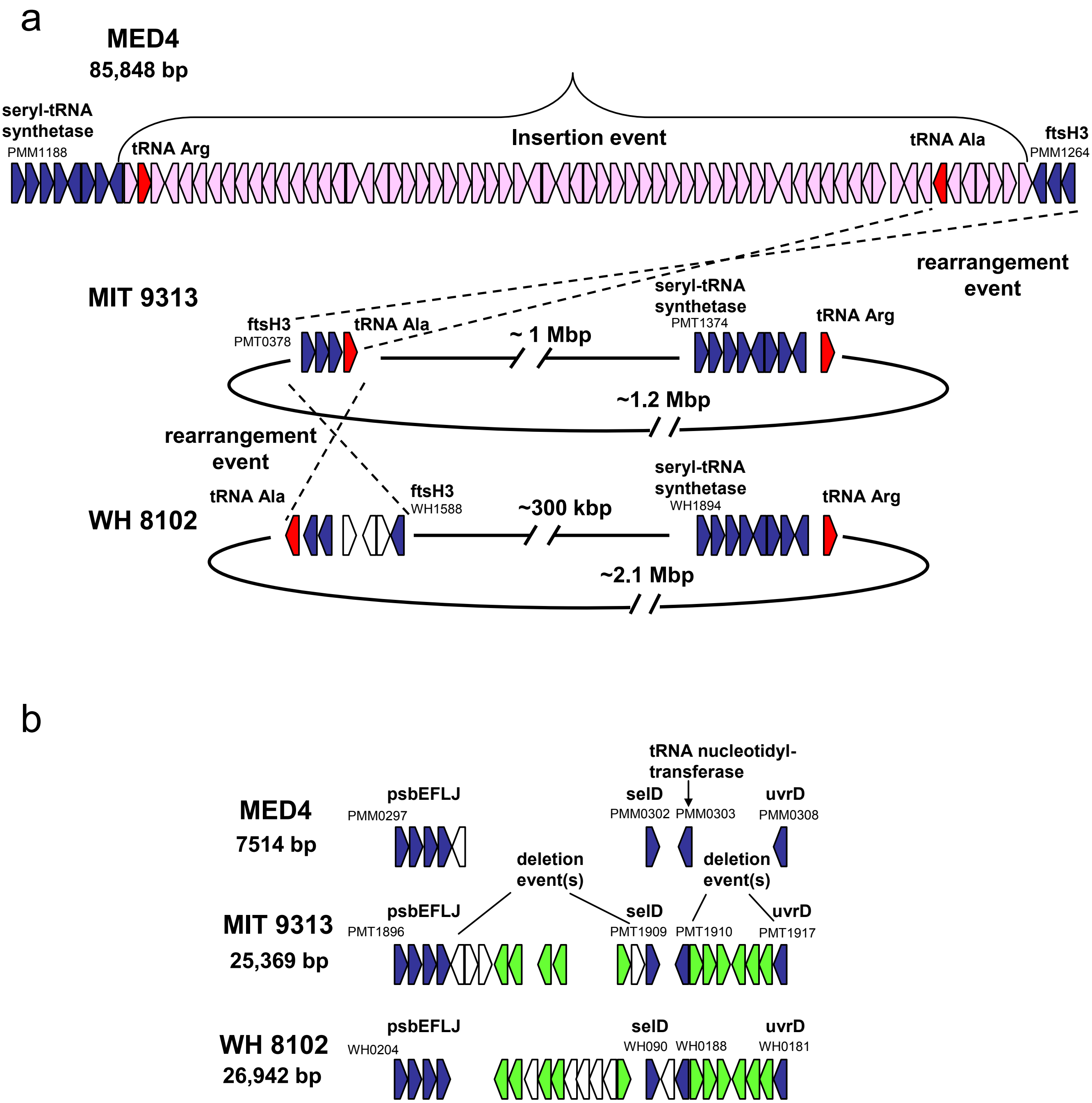


```
                          10        20        30        40        50        60        70        80
                  ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
MIT 9313    1     --------------------------------------------------------------------------------
Rhodopseud  1     ---------MRVAGAVRHWHGDCLVSSGQRRPPSENCQADVRQTAMGQLSRPRRWAASSRLRSLVAPRACAFRRGLFSSK
Bradyrhiz.  1     -----------------------------------------------MFSRALFGGVS---IS-L-----------TKQK
Vibrio      1     --------------------------------------------------------------------------------
Nitros.     1     -------------MAKFRPPCSFCARLDHHQFPQLTHLKNMKESEQPSIILNERVIQSSIAAETMSTGCSVASAPVKVSN
WH 7803     1     MLGELWSFQGRYRTLHLTWFAFFLTFVVWFNLAPLATTVKADLGLTVGQIRTVAICNVALTIPARVLIGMLLDKFGPRLT
PCC 7002    1     MLGEMWSFNGRYKILHMTWFAFFLSFVVWFNFPPFATTIAQDFGLDKAQLGTIGLCNVALTVPARIIIGMLLDKYGPRLT
WH 9601     1     MLTKLWSFRGRYRILHLTWFAFFLSFVVWFNLAPLATALKEDLGLETSQIRTIAICNVALTVPARIIIGMLLDKYGPRIT
PCC 73102   1     MLKKLFSFSDRYRILHQTWFAFFLTFVCWFNFAPFATTIGKELHLAPEQIKTLGICNLALTIPARLIIGMLLDRFGPRIT

                          90       100       110       120       130       140       150       160
                  ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
MIT 9313    1     -----MDYVLPNELVDGMITAGGKKSSVSIKNLLVRGFYSGAILGLATCLAITVGVQSGMP-----------------FL
Rhodopseud  72    RSTEQMSYLAPSEFVTKMVDAGESKIFMSTRDTVIRAYMAGAILALAAWFAITINVNTGQP-----------------LI
Bradyrhiz.  19    E--PLMSYLAPSEFVTKMVDAGESKIFMSTRDTIIRAYMAGAILALAAWFAVTINVNTGQP-----------------LV
Vibrio      1     -----MSDLRPAEFVQTMIDVGEAKVKTSTRDLVLRGMMAGIILSLAVVVAITTITQTGIG-----------------IV
Nitros.     68    EPTAVIDSVSPIQMGHDLVEDATKKKKFKVGQILIRGFLCTPFLAYATALCALLVSQGWPT-----------------AA
WH 7803     81    YSTLLVFSVIPCLMFASAQDFNQLVVARLLLSIVGAGFVIGIRMVAEWFPPKEIGLAEGIYGGWGNFGSAFSALTLVGLA
PCC 7002    81    YSLLLIYAAVPCLIFATAQSFNQLVLGRLLMGIVGAGFVIGIRMVAEWFPPKDVGTAEGIYGGWGNFGSAFSAFTMVIFG
WH 9601     81    YSLLLMYAAIPCIGFALAQNFSHLVISRLALSIVGGGFVIGIRMVAEWFPPKEIGLAEGIYGGWGNFGSAGAAFTLPTIA
PCC 73102   81    YSILLMFAVVPCLATALAQDFNQLVISRLLMGIVGSGFVVGIRMVAEWFQPKEMGIAQGIYGGWGNFGAFGAEFALPILA

                         170       180       190       200       210       220       230       240
                  ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
MIT 9313    59    GSVLFPFG-----------------------FASIVLFGMELVTGNFALLPMATWAGKS---------------------
Rhodopseud  135   GALLFPVG-----------------------FTMLYLLGFDLLTGVFVLSPLAWLDKRPG--------------------
Bradyrhiz.  80    GALLFPVG-----------------------FVMLYLLGFDLLTGVFVLSPLALIDKRPA--------------------
Vibrio      59    GALVFPVG-----------------------FCILSLMGYDLVTGVFGLAPLAKFENRPG--------------------
Nitros.     131   AGLLFPAG-----------------------YVMLSILGLEMATGSFSVMPMGLFAGRFG--------------------
WH 7803     161   GWLSFSGG--------------FTLPTGDVLNWRGAIALTGIISAVYGVIYYFNVSDTPPGKVYQRPERTAGLEVTSMRD
PCC 7002    161   IILAFLPGAFNFGQPESFKILFFPEFNTAILNWRAAIAGTGIIAALYGMLYYFSVSDTPPGKTYHRPKSARGMEVTTKKD
WH 9601     161   AWLTFGSG--------------------DVLNWRVSIALTGIIAALYGVFYFFNVLDTPPGKTYQKPKCARGLEVTTPKD
PCC 73102   161   ISTSFFSG--------------------GASNWRLAIALVGIITAIYGVIYYNTVQDTPRGKVYKKPKKNGSLEVTSIKS

                         250       260       270       280       290       300       310       320
                  ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
MIT 9313    94    -----------TWQATFRNWIWV-----WIGN---FIGCALVALLLATSLTSAGTVEPLAAADGG---KGWAVIAAKIMA
Rhodopseud  171   ----------VTIGGVLRNWGLV-----FIGN---FAGALTVAFLMAFVTTFGFTQDPDKVG-------------TAIGN
Bradyrhiz.  116   ----------VTFGGVLRNWGLV-----FVGN---FAGAFTVAFMMAFVTTFGFTQDPDKIG-------------TAIGN
Vibrio      95    ----------ITWGRILRCWGLV-----GLGN---LIGSLLVAFLIALSLTMNFSVEPNAVG-------------QTFIK
Nitros.     167   ------------LGSVIRNWSWT-----FVAN---LIGGVFFAYLLWFSLTKGGAVEPPGVL-------------TTLAH
WH 7803     227   FWGLLGMNVPFAAILCVLCWRLQ--KVGFLNASTYPLALLAVLVWFVFQTWGIIRTNRDLIMGTKVYPKEDRYEFKQVAI
PCC 7002    241   FWFLLAMNLPLTLILMVLAWRLQ--KVNFLNGTGFAIAILALVGLYLFQTYNCWTVNKDLMTGKKRYAPEDRYEFSQVAI
WH 9601     221   FLFLVLMNIPLTGVLCLLGWRLS--KVGFLSTSQLYIVWLVLLGLYAFQTYNCWITNKELIAQEKHYPPSDRYKFSQVAI
PCC 73102   221   FWAMMISNFGLIFALGLLAWRLEQKKIHFLTLSQMYLTWLVLAGLFAYQSYKAWQVNRELLTGKKTYPVSERFQFGQVAL

                         330       340       350       360       370       380       390       400
                  ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
MIT 9313    153   LNKANVVAKYQDLGSTGFFLAFLR----------GMIANWLVCLGVTMALVSKSVPGKILACWLPITAFQT---------
Rhodopseud  221   IGEGRTLG-YAAHGAAGMVTLFFR----------GMLCNWMVSTGVVMAMISTTVPGKVLAMWMPILVFFY---------
Bradyrhiz.  166   IGEGRTLG-YAAHGAAGMATLFLR----------GMLCNWMVSTGVVGAMISTSVPGKVIAMWMPILVFFY---------
Vibrio      145   AATARTVG-FENLGADGWITCFVR----------GILCNLMVCLGVIGNLSARTVAGKIAAMWLPIFIFFA---------
Nitros.     215   LAEKKAS--YAGYGVNGWFAAIGM----------GILCNWLVSLAPVFAKGSRSVPGKIMLMWLPLATFFS---------
WH 7803     305   LELTYIVNFGSELAVVSMLPTFFETTFDLPKATAGILASCFAFVNLVARPAGGLISDKLGSRKNTMGFLTAGLGVGYLIM
PCC 7002    319   LELTYIVNFGSELAVVTMLPAFFEGTFSLDKATAGIIASSYAFMNLMSRPGGGLISDKMGSRKWTMVVLTVGMGVGYLLM
WH 9601     299   LELTYFVNFGSELAVVSMLPAFFENTFGLSKTMAGMIAASYAFMNLVSRPGGGLISDKLGSRKLTMTVLTVGMAIGYLTM
PCC 73102   301   LEFTYITNFGSELAAVSMLPAFFEKTFGLEHVVAGMIAATYPFLNLVSRPSGGLISDKFGSRKWTMTIISVGIGVSYLMA

                         410       420       430       440       450       460       470       480
                  ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
MIT 9313    213   -------------MGMEHIVVNMFLHTAGPMLGSGVSFGQVIVWNFIPVTLGNIIGGMVFIGMLFYSTHRTQMSNVLPT-
Rhodopseud  280   -------------MVFEHSVVNMFLFPSGLMLHAKFSILDYLIWNEIPTVLGNLVGGLAFTGLTLYTTHVRTGAKRR---
Bradyrhiz.  225   -------------MVFEHSVVNMFLFPSGLMLHAKFSIMDYLIWNEIPTVLGNLVGGLAFTGMTLYATHVMTQPKRQ---
Vibrio      204   -------------LVFEHAVVNMFLFPLGMMLGADFGIATWLNFNLIPTILGNIVGGLLFTCIPLYLTHAKTAPAID---
Nitros.     273   -------------LGFEHAVVNMFVFPIGILSGADVTISQWWLWNQIPVTIGNMLGAVIFNSTLWYRTHRA---------
WH 7803     385   SLIKPGTFTGTTGIVIAVLITMLASFFVQSGEGATFALVPLVKRRVTGQVAG-LVGAYGNVGAVTYLTIFSLLPLWMGGA
PCC 7002    399   SSVA-----GTWPLAIAVLLTMACSFFVQAAEGSTFAIVPLVKRRITGQIAG-NVGAYGNVGAVAYLTVLLLLTEASAGA
WH 9601     379   GNVK-----GSWWLPAAVLLTMLCSFFVQAAEGSTFAIVPLIKRRVTGQIAG-NVGAYGNVGAVLYLTLYSFLPEGAIG-
PCC 73102   381   HFIN-----SNWPIPVAIAVTMFAAYFAQAGCGATYSIVPMIKKEATGQIAG-NVGAYGNFGGVVYLTIFSLTDAPT---

                         490       500       510       520       530       540
                  ....|....|....|....|....|....|....|....|....|....|....|....|....|.
MIT 9313    279   --------------------VHDEKLERELAAELGAR-----------------------------
Rhodopseud  344   --------------------IEAVPASRVAA-----------------------------------
Bradyrhiz.  289   --------------------AGKTTPSRVAA-----------------------------------
Vibrio      268   --------------------VTEKQTEINAEPALSIK-----------------------------
Nitros.     331   ------------------------------------------------------------------
WH 7803     464   GEPTAETIAASNSAFFQILGIAGLIVAFFCFFFLKEPKGSFAELHEG---ETAAEGTPSMAR----
PCC 7002    472   -NGGEPVMATVNAGFFQVLGITGLIVAFLCAFFLKEPKGSFAEFHEG---ETEMTATPPIEEEATY
WH 9601     451   -----------DKIFFEVIGVTSMIVAFICAFFLKEPEGSFAEHHEGEE-EVQIESHTLFAEE---
PCC 73102   451   --------------LFSTMGIAALICAFMCAFFLKEPKGSFAPAYEGEASETATKSSVFLTEE---
```

Chisholm Supplementary Figure 5



Chisholm Supplementary Figure 6


a
MED4
85,848 bp
seryl-tRNA synthetase
PMM1188
tRNA Arg
Insertion event
tRNA Ala
ftsH3
PMM1264
rearrangement event
MIT 9313
ftsH3
PMT0378
tRNA Ala
~ 1 Mbp
seryl-tRNA synthetase
PMT1374
tRNA Arg
~1.2 Mbp
rearrangement event
tRNA Ala
ftsH3
WH1588
~300 kbp
seryl-tRNA synthetase
WH1894
tRNA Arg
WH 8102
~2.1 Mbp
b
tRNA nucleotidyl-transferase
MED4
7514 bp
psbEFLJ
PMM0297
selD
PMM0302
PMM0303
uvrD
PMM0308
deletion event(s)
deletion event(s)
MIT 9313
25,369 bp
psbEFLJ
PMT1896
selD
PMT1909
PMT1910
uvrD
PMT1917
WH 8102
26,942 bp
psbEFLJ
WH0204
selD
WH090
WH0188
uvrD
WH0181