ERNEST ORLANDO LAWRENCE
BERKELEY NATIONAL LABORATORY

Title

Metagenome of a Versatile Chemolithoautotroph from Expanding Oceanic Dead Zones

Author(s)

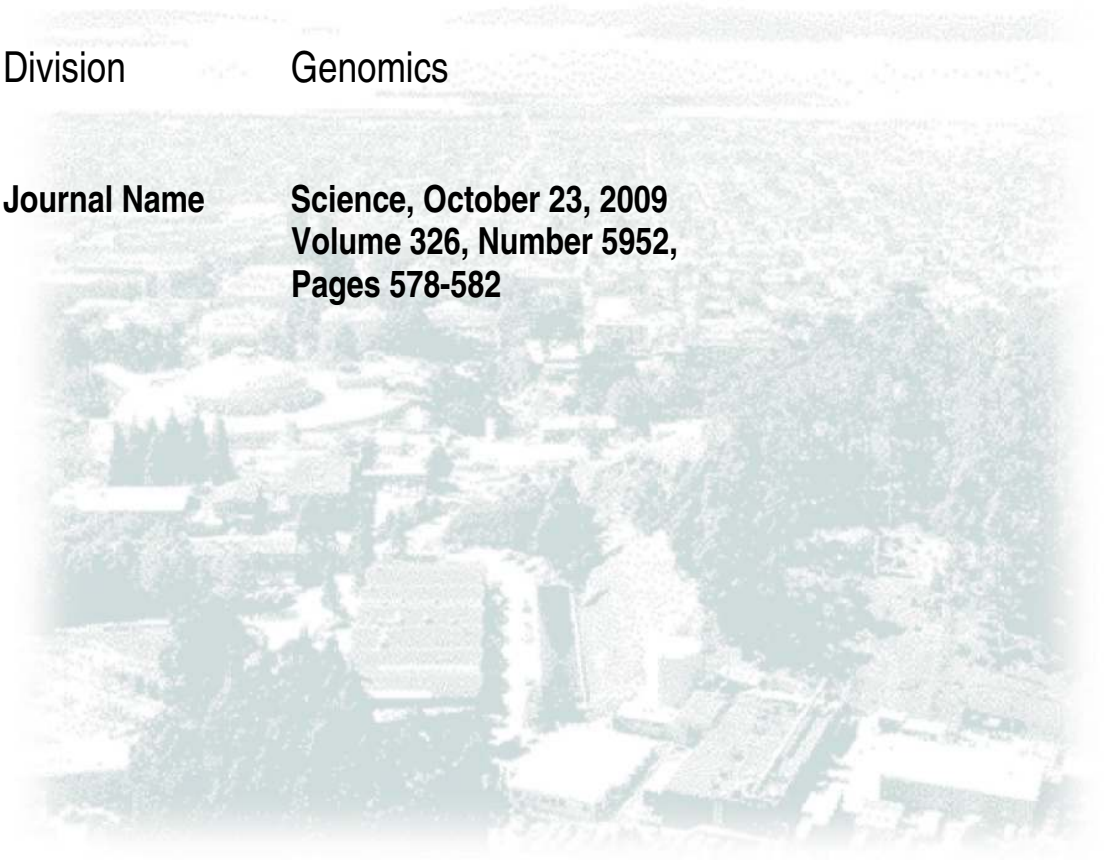
David A. Walsh, Elena Zaikova, Charles L. Howes, Young Song, Jody Wright, Susannah G. Tringe, Philippe D. Tortell, Steven J. Hallam

Division

Genomics

Journal Name

Science, October 23, 2009
Volume 326, Number 5952,
Pages 578-582



Metagenome of a versatile chemolithoautotroph from expanding oceanic dead zones

Summary: Time-resolved metagenomic approaches are used to describe carbon and energy metabolism of an ecologically relevant microbe from expanding oceanic dead zones mediating carbon sequestration, sulfur-detoxification and biological nitrogen loss.

David A. Walsh[1], Elena Zaikova[1], Charles L. Howes[1], Young Song[1], Jody Wright[1], Susannah G. Tringe[5], Philippe D. Tortell[2,3], Steven J. Hallam[1,4*]

[1] Department of Microbiology & Immunology, University of British Columbia

[2] Department of Earth & Ocean Sciences, University of British Columbia

[3] Department of Botany, University of British Columbia

[4] UBC Graduate Program in Bioinformatics

[5] Department of Energy Joint Genome Institute

* To whom correspondence should be addressed:

University of British Columbia, Department of Microbiology & Immunology, Life Sciences Institute, 2552-2350 Health Sciences Mall, Vancouver British Columbia V6T1Z3 Canada

Office: (604) 827-3420 FAX: (604) 822-6041 e-mail: shallam@interchange.ubc.ca

Running Title:

Versatile energy metabolism in a ubiquitous oxygen minimum zone microbe

Abstract

Oxygen minimum zones (OMZs), also known as oceanic “dead zones”, are widespread oceanographic features currently expanding due to global warming and coastal eutrophication. Although inhospitable to metazoan life, OMZs support a thriving but cryptic microbiota whose combined metabolic activity is intimately connected to nutrient and trace gas cycling within the global ocean. Here we report time-resolved metagenomic analyses of a ubiquitous and abundant but uncultivated OMZ microbe (SUP05) closely related to chemoautotrophic gill symbionts of deep-sea clams and mussels. The SUP05 metagenome harbors a versatile repertoire of genes mediating autotrophic carbon assimilation, sulfur-oxidation and nitrate respiration responsive to a wide range of water column redox states. Thus, SUP05 plays integral roles in shaping nutrient and energy flow within oxygen-deficient oceanic waters via carbon sequestration, sulfide detoxification and biological nitrogen loss with important implications for marine productivity and atmospheric greenhouse control.

Dissolved oxygen (O_2) concentration is a critical determinant of marine ecosystem function and food web structure. Water column O_2 deficit results in habitat compression and reduced productivity for aerobic respiring organisms with concomitant expansion of chemolithoautotrophic metabolism (*1*) manifested by biological nitrogen loss and the production of climate active trace gases (*2, 3*). Therefore, OMZ expansion and intensification (*3-5*) represents a global ecological phenomenon with potentially deleterious feedback and forcing effects (*1, 6*). In order to understand, respond to, or mitigate these transitions, studies monitoring and modeling dynamics and systems metabolism of OMZ micobiota in relation to physical and chemical oceanographic parameters are imperative.

Extensive OMZs are found throughout the Eastern North Pacific (ENP), Eastern South Pacific (ESP), Northern Indian Ocean, and Southwest African shelf waters (*i.e.* Namibian upwelling system) (*3*). Although the extent of oxygen-deficiency varies within and between oceanographic provinces, taxonomic surveys have revealed conserved patterns of microbial community composition (*7-11*). In all sites thus far examined, small subunit ribosomal RNA (SSU rRNA) gene libraries are enriched with sequences affiliated with chemoautotrophic gill symbionts of deep-sea clams and mussels (EOSA-1 in the ESP and GSO in African shelf waters) (*9-11*). Phylogenetic analyses indicate that the GSO/EOSA-1 complex consists of two closely related, co-occurring and uncultivated lineages, ARCTIC96BD-19 (*12*) and SUP05 (*13*) with the latter encompassing the symbionts (Fig. 1A-B). From an ecophysiological perspective, blooming SUP05 populations have recently been implicated in chemolithotrophic sulfide (H_2S) oxidation coupled to nitrate (NO_3^-) reduction in the Namibian upwelling system (*10*). Both,

ARCTIC96BD-19 and SUP05 populations are also prevalent in non-sulfidic waters of the ENP and ESP, suggesting alternative or amended modes of energy metabolism. Given the likely importance of ARCTIC96BD-19 and SUP05 to carbon, nitrogen and sulfur cycling in marine OMZs, and their largely unexplored metabolic capabilities, a deeper understanding of both lineages is needed to constrain their respective ecological and biogeochemical roles.

Saanich Inlet, British Columbia is a seasonally anoxic fjord characterized by an annual cycle of stratification and deep water renewal (*14*) that is associated with large water column redox gradients and high rates of trace gas production and consumption (Fig. S1A-C). Time resolved studies identified pelagic SUP05 as a dynamic and numerically abundant denizen of the Saanich Inlet water column, representing up to 37% of total bacteria (Table S2) (*11*). Closer examination of SUP05 SSU rRNA gene copy number during onset and progression of seasonal stratification revealed blooming populations below the oxycline, reaching up to 4.75×10^5 copies ml^{-1} in regions of H_2S and NO_3^- depletion (Fig. S2). Further high-resolution SSU rRNA gene surveys revealed two SUP05 phylotypes, SI-1 and SI-2 (Fig. 1A and 2A), differing by ~4% nucleotide identity. While SI-1 dominated suboxic waters throughout the year, SI-2 was less common, and transiently increased during deep water renewal events, alluding to the presence of ecologically differentiated populations (Fig. 2A). Given these observations we reasoned that Saanich Inlet would provide a natural enrichment amenable to environmental genomic (*i.e.* metagenomic) assembly and metabolic pathway reconstruction of pelagic SUP05 populations. To this end, we analyzed sixteen bi-directionally end sequenced

fosmid libraries constructed from environmental DNA samples spanning oxic to anoxic waters over the seasonal stratification and deep water renewal cycle (Fig. S1D).

Fosmid end sequences were initially screened for putative SUP05 genotypes by fragment recruitment to closely related symbiont reference genomes (*15, 16*) revealing extensive coverage within oxygen-deficient sampling intervals (Fig. S3). To reconstruct and identify SUP05-specific scaffolds, paired-end sequences were assembled and binned based on shared sequence similarity to symbiont reference genomes and analysis of intrinsic oligonucleotide composition patterns (Fig. S4A-B). Nineteen scaffolds encompassing 1.16 million base pairs of SUP05 DNA, herein referred to as the SUP05 metagenome, were identified and taxonomically verified (Table S3). See Table S1 for additional breakdown of important assembly features. Although derived from a heterogeneous population, average polymorphism within the SUP05 metagenome was 0.4%, indicating assembly of closely related sympatric donor genotypes. Consistent with this observation, a single SSU-LSU rRNA operon affiliated with the SI-I phylotype was identified (Fig. S4). Closer examination of individual fosmid library contributions to the assembly revealed a majority of paired-end sequences within SUP05 scaffolds were derived from samples exhibiting elevated SI-1 phylotype abundance, further supporting scaffold assignments (Fig. 2B). Overall coverage efficiency was assessed based on recovery of 83 of 93 information processing genes, including 31 of 32 canonically conserved single copy genes (*17*) (Table S3) suggesting near complete recovery of the core genome, in turn facilitating downstream efforts to reconstruct and interpret metabolic pathways.

The comparative architecture of SUP05 and symbiont reference genomes revealed patterns of genetic relatedness consistent with derivation from a common free-living ancestor. At the same time, differences in gene content and organization provided insight into the process of genome divergence and niche partitioning. Approximately 20% of predicted gene content shared between symbiont reference genomes was absent from the SUP05 metagenome. Genes from this set typically clustered together (Fig. 3), and could represent genomic features mediating symbiont-host interactions or adaptations to hydrothermal vent or cold seep settings (*18*). Of the 861 genes shared between symbiont reference genomes, 80% were also conserved in the SUP05 metagenome (Fig. 3 and S5-6), indicating significant metabolic overlap. Many of these genes are predicted to mediate informational processing steps, particularly translation, although a significant fraction function in carbon, sulfur, amino acid, and coenzyme metabolism (Fig. S6). Approximately 35% of the gene content predicted in the SUP05 metagenome was unique, reflecting characteristic differences between pelagic and symbiotic modes of existence (Fig. 3 and S5-6). These included genes implicated in DNA uptake and repair, denitrification and adaptive or stress responses (See supporting online material for more details).

Given the potential importance of pelagic SUP05 populations on the ecology and biogeochemistry of oxygen-deficient oceanic waters, examination of carbon and energy metabolism within the SUP05 metagenome is of particular interest. Similar to symbiotic counterparts, the SUP05 metagenome harbors genes mediating the Calvin-Benson-Bassham (CBB) cycle for autotrophic carbon assimilation, including a single form II ribulose 1,5-bisphosphate carboxylase-oxygenase (RubisCO) gene, implicating SUP05 in

chemosynthetic carbon fixation within OMZs (*19*). In addition, a gene encoding β-class carbonic anhydrase, encoding a potential CO_2 concentrating mechanism, was also identified. A complete repertoire of genes mediating the conversion of fixed carbon to hexose and ribose sugars via gluconeogenesis and the non-oxidative branch of the pentose phosphate pathway was identified along with the majority of tricarboxylic acid (TCA) cycle components (Fig. S6). However, genes mediating the interconversion of succinyl-CoA and 2-oxoglutarate were not recovered (Fig. S6), indicating the potential for obligate autotrophy as posited earlier for clam symbionts (*15, 16*).

From the standpoint of energy metabolism, the SUP05 metagenome harbors a diverse repertoire of genes mediating chemolithotrophic oxidation of reduced sulfur compounds. Genes encoding flavocytochrome c/sulfide dehydrogenase (*fccAB*) unique to the SUP05 metagenome, and sulfide quinone oxidoreductase (*sqr*) conserved between pelagic SUP05 and symbiont reference genomes mediating the oxidation of H_2S to elemental sulfur (S^0) were identified (Fig. S6). The presence of two enzymatic complexes may facilitate sulfur-based energy metabolism under variable sulfide regimes. For instance, FccAB is thought to be functionally significant at low sulfide concentrations (*20*). In addition, sirohaem dissimilatory sulfite reductase subunits (*dsrAB*), APS reductase (*apr*), ATP sulfurylase (*sat*) mediating the complete oxidation of S^0 to sulfate, and the Sox pathway (*soxABXYZ*) for thiosulfate ($S_2O_3^{2-}$) oxidation (Fig. S6) (*21*) were also conserved between pelagic SUP05 and symbiont reference genomes. The capacity to obtain electrons from $S_2O_3^{2-}$ may be of considerable ecological relevance given that mixing of sulfidic and oxygenated water masses results in $S_2O_3^{2-}$ accumulation due to the chemical oxidation of H_2S (*22, 23*). Moreover, the apparent absence of *soxCD* sulfur

dehydrogenase genes suggests the capacity to store S^0, which can be subsequently oxidized via the reverse DSR pathway thereby provisioning SUP05 in the absence of ambient reductant (*24*). Indeed, *soxCD* homologues are also absent from symbiont reference genomes, and sulfur globule formation has been associated with a subset of clam symbionts (*25*).

Although symbiont reference genomes harbor multiple aerobic respiratory complexes (*15, 16*), none were recovered in the SUP05 metagenome consistent with a facultative or strictly anaerobic lifestyle. Indeed, all the enzymatic machinery needed to reduce NO_3^- to the greenhouse gas nitrous oxide (N_2O) including membrane-bound (*narKK$_2$GHJI)* and periplasmic (*napFBAHGD*) dissimilatory nitrate reductases potentially operating under high and low NO_3^- conditions, respectively (*26, 27*), copper-containing nitrite reductase (*nirK*), and N_2O forming nitric oxide reductase (*norCB*) (Fig. 4 and S6) were identified, mechanistically implicating pelagic SUP05 in biological nitrogen loss from oxygen-deficient oceanic waters. Moreover, the genomic co-localization of sulfur oxidation and denitrification genes suggests a highly integrated and redox-sensitive energy metabolism (Fig. 4). For example, the genomic proximity of Crp/Fnr transcriptional regulators with *fcc* and *nap* gene clusters (Fig. 4) may indicate coordinated gene expression in response to changing redox status (*20, 26, 28*). In both the Namibian upwelling system and Saanich Inlet, blooming SUP05 populations occur in regions of H_2S and NO_3^- depletion (Fig. S2) where coexpression of *nap* and *fcc* genes clusters may become critical for survival as energetic substrates become limiting.

Curiously, we identified more than ten putative toxin-antitoxin (TA) modules unique to the SUP05 metagenome, indicating a highly regulated stress response (Table

S4). TA modules consist of a stable toxin and a labile antitoxin and are commonly associated with environmental bacteria where they control induction of reversible cellular stasis (*29*). Of specific interest is a TA module of the RelE superfamily identified within an operon encoding molybdopterin-guanine dinucleotide synthase (*mobA*) (Fig. 4). The product of MobA, molybdopterin-guanine dinucleotide (MGD), is an essential cofactor for all described classes of nitrate reductase (*30*) and therefore *mobA* expression is integral to denitrification in pelagic SUP05 populations. Severe NO_3^- limitation could limit *mobA* expression leading to co-repression of the embedded TA module. This would result in activation of the RelE toxin, through degradation of the labile antitoxin, and induction of cellular stasis (*31*). In this regard, the integration of a TA system into a denitrification regulon may allow SUP05 to persist during periods of extreme NO_3^- limitation, analogous to other forms of nutritional stress response (*e.g.* amino acid starvation in *E. coli*).

As the number of studies surveying OMZ community structure increases, the ubiquity and abundance of the SUP05 lineage becomes ever more apparent. Analysis of the SUP05 metagenome, and the water column disposition of pelagic SUP05 with respect to H_2S and NO_3^- gradients, resolves a chemolithoautotrophic metabolism based on oxidation of reduced sulfur compounds with NO_3^- through multiple and highly regulated bioenergetic routes. Paradoxically, as "dark" primary producers, blooming SUP05 populations have the potential to sequester large amounts of CO_2 while simultaneously producing N_2O via NO_3^- respiration. Therefore the SUP05 metagenome provides a functional template for analysis of gene expression in relation to climatologically relevant biogeochemical transformations within oxygen-deficient oceanic waters. We anticipate

that this information will become an essential resource for comparative analysis of ecotype diversification within the SUP05 and ARCTIC96BD-19 lineages aiding in the development of monitoring tools to assess and model microbial community responses to OMZ expansion and intensification.

References

1. R. J. Diaz, R. Rosenberg, *Science* **321**, 926 (2008).
2. K. R. Arrigo, *Nature* **437**, 349 (2005).
3. A. Paulmier, D. Ruiz-Pino, *Progress in Oceanography* **In press**, (2008).
4. F. A. Whitney, H. J. Freeland, M. Robert, *Prog. in Oceanog.* **75**, 179 (2007).
5. L. Stramma, G. C. Johnson, J. Sprintall, V. Mohrholz, *Science* **320**, 655 (2008).
6. P. G. Brewer, E. T. Peltzer, *Science* **324**, 347 (2009).
7. B. M. Fuchs, D. Woebken, M. V. Zubkov, P. Burkill, R. Amann, *Aquat Microb Ecol* **39**, 145 (2005).
8. D. Woebken, B. A. Fuchs, M. A. A. Kuypers, R. Amann, *Applied and Environmental Microbiology* **73**, 4648 (2007).
9. H. Stevens, O. Ulloa, *Environ Microbiol* **10**, 1244 (2008).
10. G. Lavik *et al.*, *Nature* **457**, 581 (2009).
11. E. Zaikova *et al.*, *Environ Microbiol* (submitted).
12. N. Bano, J. T. Hollibaugh, *Appl Environ Microbiol* **68**, 505 (2002).
13. M. Sunamura, Y. Higashi, C. Miyako, J. Ishibashi, A. Maruyama, *Appl Environ Microbiol* **70**, 1190 (2004).

14. J. J. Anderson, A. H. Devol, *Estuarine and Coastal Marine Science* **1**, 1 (1973).

15. H. Kuwahara *et al.*, *Curr Biol* **17**, 881 (2007).

16. I. L. G. Newton *et al.*, *Science* **315**, 998 (2007).

17. F. D. Ciccarelli *et al.*, *Science* **311**, 1283 (2006).

18. I. L. Newton, P. R. Girguis, C. M. Cavanaugh, *BMC Genomics* **9**, 585 (2008).

19. G. Jost, M. V. Zubkov, E. Yakushev, M. Labrenz, K. Jurgens, *Limnology and Oceanography* **53**, 14 (2008).

20. M. Mussmann *et al.*, *PLoS Biol* **5**, e230 (2007).

21. C. G. Friedrich, F. Bardischewsky, D. Rother, A. Quentmeier, J. Fischer, *Curr Opin Microbiol* **8**, 253 (2005).

22. J. Zopfi, T. G. Ferdelman, B. B. Jorgensen, A. Teske, B. Thamdrup, *Marine Chemistry* **74**, 29 (2001).

23. R. B. Cardoso *et al.*, *Biotechnology and Bioengineering* **95**, 1148 (2006).

24. C. Dahl *et al.*, *J Bacteriol* **187**, 1392 (2005).

25. R. D. Vetter, *Marine Biology* **88**, 33 (1985).

26. V. Stewart, Y. Lu, A. J. Darwin, *J Bacteriol* **184**, 1314 (2002).

27. H. Wang, C. P. Tseng, R. P. Gunsalus, *J Bacteriol* **181**, 5303 (1999).

28. S. Spiro, J. R. Guest, *FEMS Microbiol Rev* **6**, 399 (1990).

29. D. P. Pandey, K. Gerdes, *Nucleic Acids Res* **33**, 966 (2005).

30. D. J. Richardson, B. C. Berks, D. A. Russell, S. Spiro, C. J. Taylor, *Cell Mol Life Sci* **58**, 165 (2001).

31. S. K. Christensen, M. Mikkelsen, K. Pedersen, K. Gerdes, *Proc Natl Acad Sci U S A* **98**, 14328 (2001).

32. This work was performed under the auspices of the U.S. Department of Energy's Office of Science, Biological, and Environmental Research Program and by the University of California, Lawrence Berkeley National Laboratory, Lawrence Livermore National Laboratory, under contract no. DE-AC02-05CH11231, Lawrence Livermore National Laboratory under contract no. DE-AC52-07NA27344, and Los Alamos National Laboratory under contract no. DE-AC02-06NA25396. This work was also supported by grants from the Natural Sciences and Engineering Research Council (NSERC) of Canada 328256-07 and STPSC 356988, Canada Foundation for Innovation (CFI) 17444; Canadian Institute for Advanced Research (CIFAR), and the Center for Bioinorganic Chemistry (CEBIC). D.A.W. was supported by NSERC, Killam Trust, and the Tula Foundation-funded Centre for Microbial Diversity and Evolution (CMDE). We thank M. Robert (Institute of Ocean Sciences, Sidney, BC, Canada), C. Payne, L. Pakhomova, and J. Granger (UBC) for help in sampling and chemical analyses and the captains and crews of the *CCGS John P. Tulley* and *HMS John Strickland* for logistical support. We thank the Joint Genome Institute, including K. Barry, S. Pitluck, and E. Kirton, for technical assistance and A. Page, K. Mitchell, and S. Lee in the Hallam laboratory for reading the manuscript. This metagenome project has been deposited at the DNA DataBank of Japan and European Molecular Biology Laboratory, and GenBank, under the project accession ACSG00000000. The version described in this paper is the first version, ACSG01000000. SSU rRNA gene sequences were deposited at GenBank under the accession numbers GQ345343-GQ351265, and fosmid sequences were deposited under the accession numbers GQ351266 to GQ351269 and GQ369726.

Figure Titles

Fig. 1 (A) Phylogenetic tree of ARCTIC96BD-19 and SUP05 lineages based on comparative SSU rRNA gene analysis. The tree was inferred using maximum likelihood implemented in PHYML (B) Relative abundance of ARCTIC96BD-19 and SUP05 SSU rRNA sequences recovered from Saanich Inlet (SI), eastern North Pacific (ENP) (this study), eastern South Pacific (ESP) (*9*), and southwest African shelf waters (Namibia) (*10*).

Fig. 2 (A) Phylotype abundance of SUP05 SI-1 (black circles) and SI-2 (white circles) based on recovery of SSU rDNA sequences in PCR generated clone libraries (B) mean depth of coverage of the SUP05 metagenome plotted over the Saanich Inlet nitrate profile during the 2006-2007 season. Sample depths and dates are noted on the axes.

Fig. 3 Gene content comparison between SUP05 metagenome and symbiont reference genomes. Nested circles from outermost to innermost represent the following. (i and ii) COG functional predictions on the forward and reverse strands of the *R. magnifica* reference genome. (Fig. S7 for color designation), (iii) Conservation of gene content (iv) Genes conserved in symbionts, but absent from the SUP05 metagenome. Inset Venn diagram depicts predicted gene distribution among SUP05 metagenome and symbiont reference genomes. Values correspond to the number of shared genes among overlapping genomes, using each genome as the original query. The dotted line represents the open-

genome configuration of the SUP05 metagenome. *The discrepancy in core size when SUP05 metagenome is employed as query (774) compared to symbionts (~683) reflects gene content redundancy in the metagenome assembly.

Fig. 4 Alignment of an ungapped region of a SUP05 metagenomic scaffold, encoding genes involved in nitrate and sulfur metabolism, with the corresponding genomic regions of symbiont reference genomes. The height of red bars corresponds to nucleotide similarity over conserved genomic regions. Proper scaffold assembly across this region was verified by full-length sequencing of two overlapping fosmids.

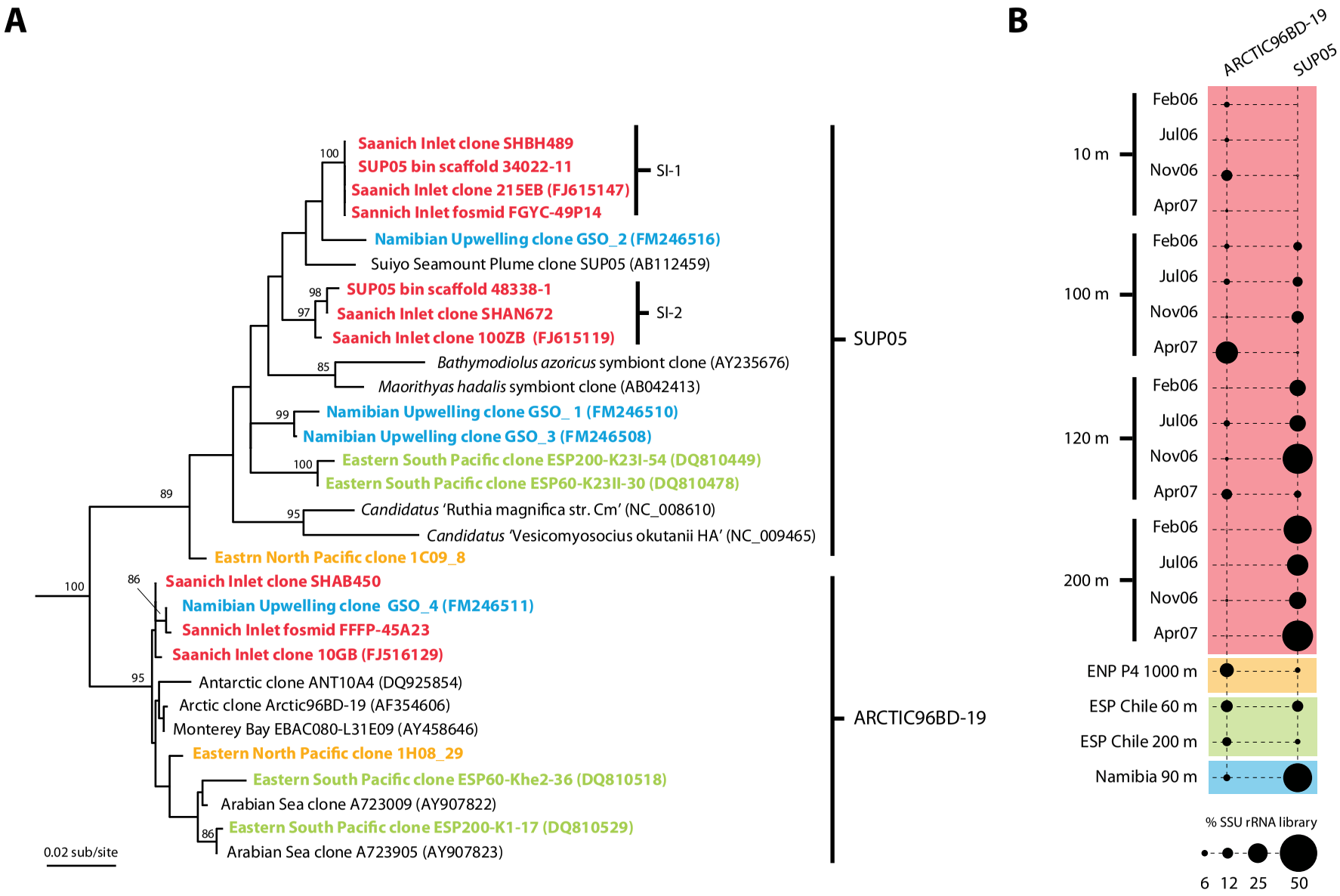

A
Saanich Inlet clone SHBH489
SUP05 bin scaffold 34022-11
Saanich Inlet clone 215EB (FJ615147)
Sannich Inlet fosmid FGYC-49P14
SI-1
Namibian Upwelling clone GSO_2 (FM246516)
Suiyo Seamount Plume clone SUP05 (AB112459)
SUP05 bin scaffold 48338-1
Saanich Inlet clone SHAN672
Saanich Inlet clone 100ZB (FJ615119)
SI-2
Bathymodiolus azoricus symbiont clone (AY235676)
Maorithyas hadalis symbiont clone (AB042413)
Namibian Upwelling clone GSO_ 1 (FM246510)
Namibian Upwelling clone GSO_3 (FM246508)
Eastern South Pacific clone ESP200-K23I-54 (DQ810449)
Eastern South Pacific clone ESP60-K23II-30 (DQ810478)
Candidatus 'Ruthia magnifica str. Cm' (NC_008610)
Candidatus 'Vesicomyosocius okutanii HA' (NC_009465)
SUP05
Eastrn North Pacific clone 1C09_8
Saanich Inlet clone SHAB450
Namibian Upwelling clone GSO_4 (FM246511)
Sannich Inlet fosmid FFFP-45A23
Saanich Inlet clone 10GB (FJ516129)
Antarctic clone ANT10A4 (DQ925854)
Arctic clone Arctic96BD-19 (AF354606)
Monterey Bay EBAC080-L31E09 (AY458646)
Eastern North Pacific clone 1H08_29
Eastern South Pacific clone ESP60-Khe2-36 (DQ810518)
Arabian Sea clone A723009 (AY907822)
Eastern South Pacific clone ESP200-K1-17 (DQ810529)
Arabian Sea clone A723905 (AY907823)
ARCTIC96BD-19
100
98
97
85
99
100
89
95
100
86
95
86
0.02 sub/site
B
ARCTIC96BD-19
SUP05
10 m
100 m
120 m
200 m
Feb06
Jul06
Nov06
Apr07
ENP P4 1000 m
ESP Chile 60 m
ESP Chile 200 m
Namibia 90 m
% SSU rRNA library
6 12 25 50

Fig. 1

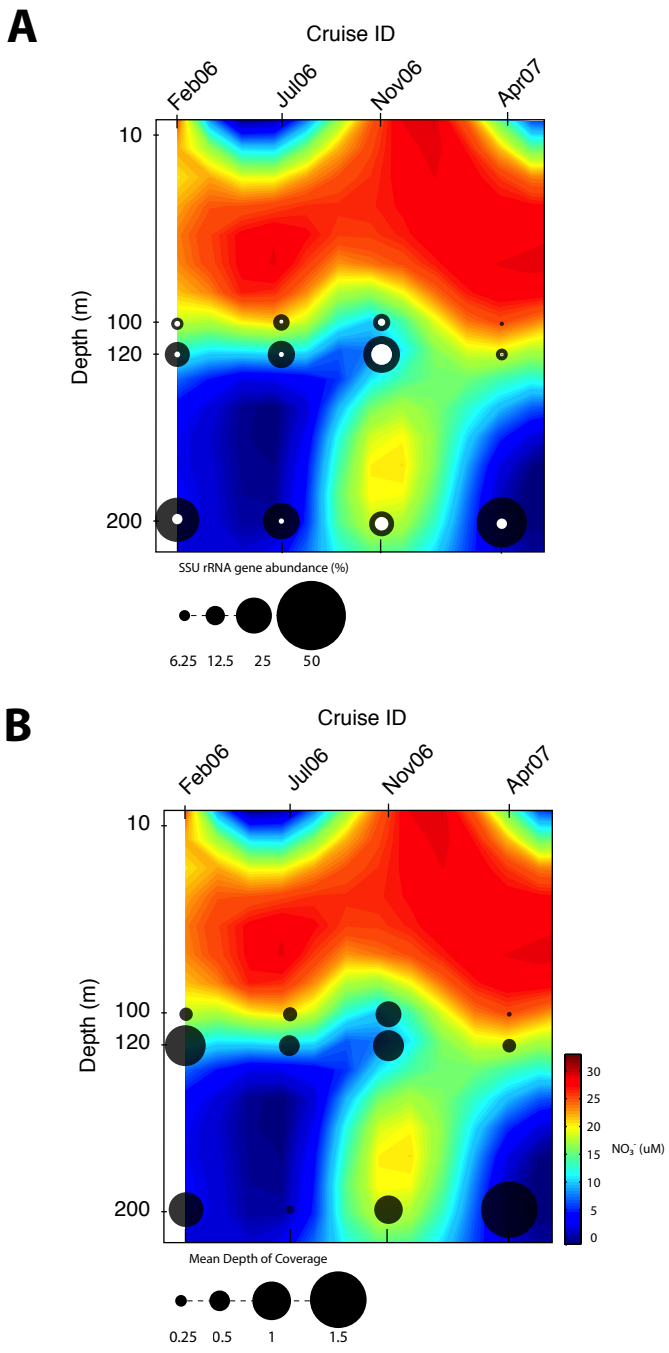

A
Cruise ID
Feb06
Jul06
Nov06
Apr07
10
100
120
200
Depth (m)
SSU rRNA gene abundance (%)
6.25
12.5
25
50
B
Cruise ID
Feb06
Jul06
Nov06
Apr07
10
100
120
200
Depth (m)
30
25
20
15
10
5
0
NO3- (uM)
Mean Depth of Coverage
0.25
0.5
1
1.5

Fig. 2

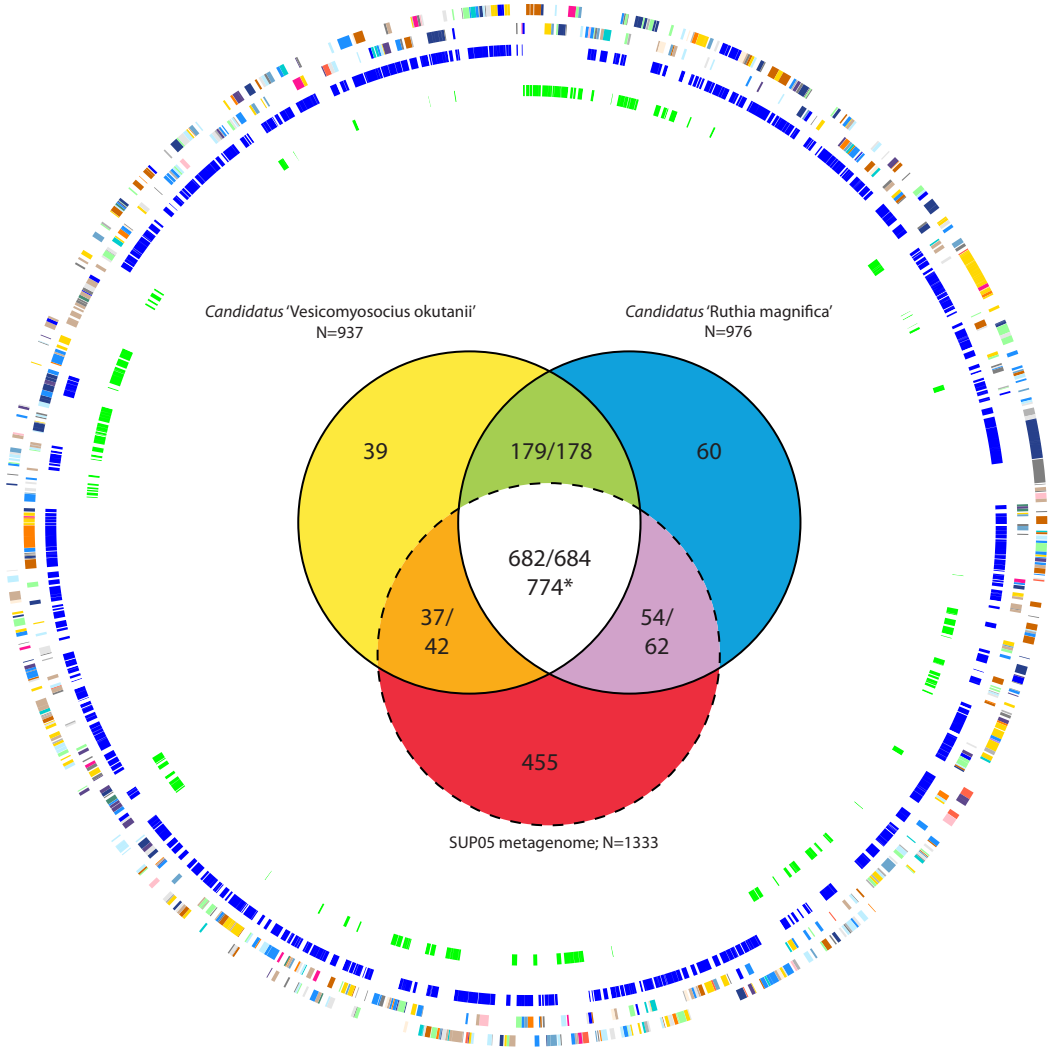

Candidatus 'Vesicomyosocius okutanii'
N=937
Candidatus 'Ruthia magnifica'
N=976
39
179/178
60
682/684
774*
37/
42
54/
62
455
SUP05 metagenome; N=1333

Fig. 3

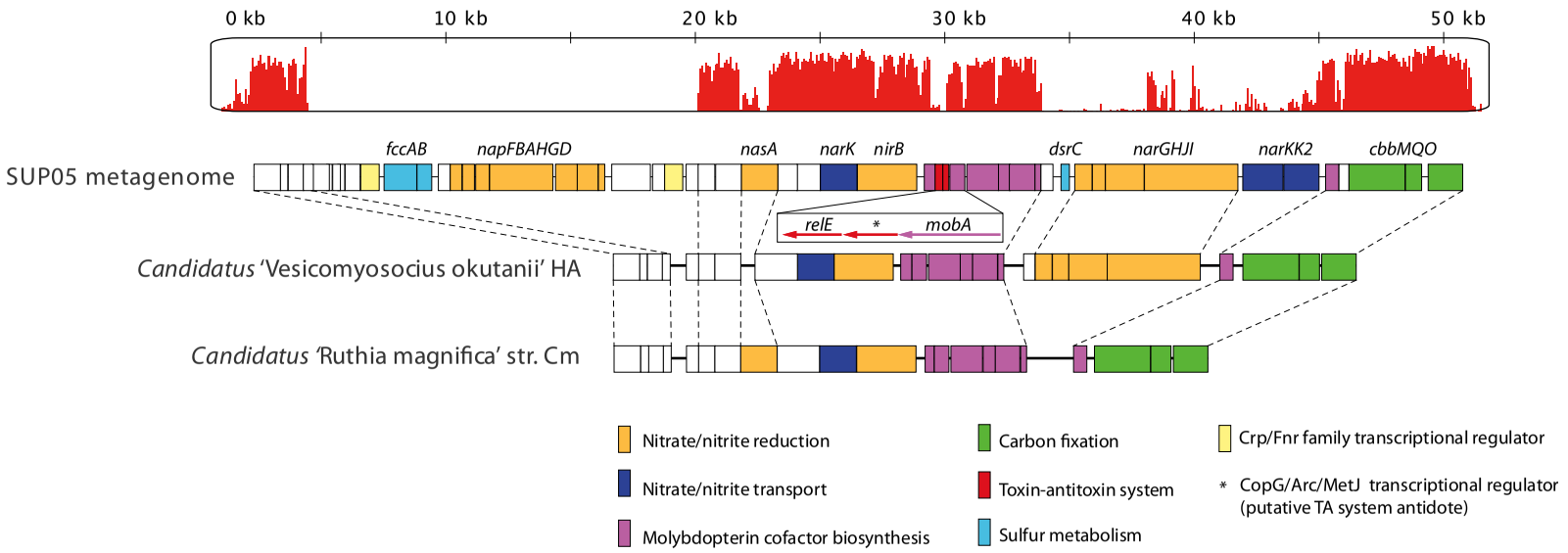

0 kb
10 kb
20 kb
30 kb
40 kb
50 kb
SUP05 metagenome
fccAB
napFBAHGD
nasA
narK
nirB
dsrC
narGHJI
narKK2
cbbMQO
relE
*
mobA
Candidatus 'Vesicomyosocius okutanii' HA
Candidatus 'Ruthia magnifica' str. Cm
Nitrate/nitrite reduction
Nitrate/nitrite transport
Molybdopterin cofactor biosynthesis
Carbon fixation
Toxin-antitoxin system
Sulfur metabolism
Crp/Fnr family transcriptional regulator
* CopG/Arc/MetJ transcriptional regulator (putative TA system antidote)

Fig. 4

Supporting Online Material for

Metagenome of a versatile chemolithoautotroph from expanding oceanic dead zones

David A. Walsh[1], Elena Zaikova[1], Charles L. Howes[1], Young Song[1], Jody Wright[1], Susannah G. Tringe[5], Philippe D. Tortell[2,3], Steven J. Hallam[1,]

The Supporting Online Material includes:

Materials and Methods

Sample collection and processing. Sample collection, O_2 and NO_3^- measurements, microbial biomass concentration, environmental DNA extraction, and quantitative PCR analysis of bacteria and SUP05 SSU rRNA gene sequences were performed as previously described (*1*). Samples for H_2S measurement were fixed with 2% (final concentration) zinc acetate and analyzed by spectrophotometer as described in (*2*).

Saanich Inlet SSU rRNA gene library construction and analysis. For each of sixteen Saanich Inlet environmental DNA samples (see Fig S1), clone libraries were constructed and sequenced from pooled PCR products amplified using the bacterial-specific forward primer 27F (5'-AGAGTTTGATCCTGGCTCAG-3') and reverse primer 1391R (5'-GACGGGCGGTGWGTRCA-3') as described at http://jgi.doe.gov. Chimeric sequences, identified using Bellerophon (*3*) and Mallard (*4*), were excluded from the sequence dataset. A

total of 5,753 partial bacterial SSU rRNA clones passed the quality and chimera check. SSU rRNA gene sequences were aligned using the SILVA aligner (*5*), imported into an ARB database (*6*) with the same alignment, and 1,067 SUP05 sequences were identified based on their position in the SILVA reference tree. The 1,067 SUP05 sequences were extracted from ARB, realigned with MUSCLE (*7*) and two sequence clusters (i.e. the SI-1 and SI-2 phylotypes) were defined after assignment of sequences to operational taxonomic units (OTUs) by applying DOTUR (*8*) analysis to an uncorrected distance matrix generated with dnadist (*9*). The frequency distribution of SI-1 and SI-2 phylotypes across libraries was calculated by dividing phylotype abundance by the total number of library clones. ARCTIC96-BD-19 clones were identified in a similar manner. The phylogenetic tree (Figure 1a) was inferred from a manually refined MUSCLE alignment of partial sequences by PHYML, using an HKY + 4Γ + I model of evolution and estimated values for the α parameter of the Γ distribution, the proportion of invariable sites, and the transition/transversion (*10*). The confidence of each node was determined by assembling a consensus tree of 100 bootstrap replicates.

Eastern North Pacific SSU rRNA gene analysis. An environmental DNA sample extracted from microbial biomass collected from within the OMZ (1000 m) at Station P4 (48^{o} 39.0' N, 126^{o} 40.0' W) of the Line P oceanographic time series (http://www.pac.dfo-mpo.gc.ca/sci/osap/projects/linepdata/default_e.htm) was used to construct a bacterial SSU rRNA gene clone library that was then sequenced as previously described (*1*). In total, 170 nonchimeric sequences were generated and analyzed as describe above.

Fosmid library construction and sequencing. Sixteen fosmid libraries (7,680 clones/libraries) were constructed from DNA samples collected from four depths during four cruises over the 2006-07 seasonal stratification cycle (see Fig. S1). Prior to cloning, ~5 μg of environmental DNA was further purified on a CsCl density gradient as previously described (*11*). Fosmid libraries were prepared using the CopyControl Fosmid Library Production Kit (Epicentre). Briefly, ~1 μg of CsCl-purified DNA was blunt end repaired and separated on a 1% low melt agarose pulse-field gel O/N at 6 V/cm. The 40-50 kb fragment range was excised and gel purified using agarase, followed by concentration using an Amicon Ultracel 10K filter device. DNA was ligated into the pCC1fos vector, packaged using the MaxPlax lambda packaging extract, and used to transfect TransforMax EPI300 *E. coli* cells. Transfected cells were plated on selective agar and fosmid clones picked using the Q-Pix robotic colony picker and grown in selective media for DNA sequencing. Bidirectional end sequencing of fosmids was performed with standard M13 -28 or -40 primers and the BigDye sequencing kit (Applied Biosystems). The reactions were purified by a magnetic bead protocol and run on an ABI PRISM 3730 (Applied Biosystems) capillary DNA sequencer (for research protocols, see http://jgi.doe.gov).

SUP05-focused assembly of fosmid end sequence data. All 243,264 fosmid end sequences (see Table S1) were initially assembled with Phrap using parameters (minmatch 30 maxmatch 55 minscore 55 max_subclone_size 50000 revise_greedy vector_bound 20) reported in (*12*). Phrap-generated contigs were linked into "ambiguous" scaffolds (*i.e.* allowed contig overlap) with Bambus (*13*), using the default settings, based on fosmid paired-end information. The initial assembly was comprised of 31,766 scaffolds with a total length (excluding gaps) and span of 35.6 Mb and 85.2 Mb respectively, and included 192 scaffolds ≥ 5 kb in length. Of the 192

scaffolds, those that exhibited highest similarity to the genomes of either of two symbiont reference species, *Candidatus* 'Ruthia magnifica' or *Candidatus* 'Vesicomyosocius okutanii', upon a BLASTN search of the NCBI refseq database were flagged as potential SUP05 scaffolds and targeted for further assembly as follows. Overlapping contigs within these ambiguous scaffolds were identified in Sequencher using a 90% identity cut-off and 100 bp minimum overlap. Those contigs that cross-assembled in the same order and orientation described by Bambus were then reassembled using the miniassembly tool in consed (*14*) with the following parameters: minscore 55 minmatch 30 forcelevel 10. Addition of the forcelevel parameter decreases assembly stringency and permits assembly between polymorphic reads, leading to the collapse of short non-polymorphic contigs into longer polymorphic scaffold. These next generation contigs were then re-ordered and oriented into ambiguous scaffolds with Bambus and the whole process was repeated until contig extension no longer occurred. Upon final scaffolding of contigs, the untangle script included with Bambus was used to break ambiguous scaffold into single linear scaffolds (*i.e.* disallowed contig overlap). In addition, the stringency of the final assembly was increased by requiring a minimum of four read pairs to link contig into scaffolds. The final assembly was comprised of 33,630 scaffolds with a total length (excluding gaps) and span of 34.9 Mb and 38.8 Mb respectively, and included 86 scaffolds ≥ 5 kb in length (Figure S4a). Within the assembly, 5,703 polymorphic sites, defined by ≥ 2 reads exhibiting high quality discrepancies with the scaffold consensus sequence, were identified.

Identification of putative SUP05 scaffolds. Predicted protein-encoding genes from all scaffolds ≥ 5 kb were searched against the two clam symbiont genomes using BLASTP. A subset of scaffold enriched in SUP05 was identified using the following criterion: ≥ 3 syntenic open

reading frames (ORFs) of high sequence similarity (blast score ratio ≥ 0.6) with either symbiont reference genome. The initial bin was comprised of 38 scaffolds (total length and span of 1.36 Mb and 2.54 Mb, respectively), contained two rRNA operons, and the corresponding SSU rRNA genes were SI-1 and SI-2 phylotypes, respectively (Figure 1a). Principal component analysis (PCA) of intrinsic DNA signatures was then used to explore the oligonucleotide frequencies of the 38 scaffolds. The frequencies of all tri-, tetra-, and penta-nucleotides were summed for each scaffold and their over- and under- representations were evaluated against their expected frequencies using a maximal-order Markov model as previously described (*15*). In addition to scaffolds, four full length overlapping fosmids (~40 kb in length each) known to originate from SUP05 based on an SI-1 SSU rRNA gene sequence were included in order to identify the expected oligonucleotide patterns of SUP05 sequences. The resulting *Z* scores (*15*) were normalized by length, imported into the statistical package PC-ORD 5.10 (*16*) and subjected to PCA using a Correlation Cross-products matrix. Visualization of the first two axes of the PCA analysis revealed many scaffolds that were uniquely positioned in the ordination space, but also revealed a single cluster consisting of 19 scaffolds that exhibited very similar nucleotide pattern compositions (Figure S4b, red circles). We interpret this subset of related scaffolds as arising from the same closely related population of SUP05 cells in Saanich. In support of this interpretation, the SI-I phylotype was observed in this scaffold subset (Figure 1a). Nucleotide composition patterns for the four fully sequenced and overlapping fosmids (spanning ~140 Kb) affiliated with SI-1 on the basis of SSU rRNA gene linkage were indistinguishable from these scaffolds further reinforcing the accuracy and resolution of our binning method (Figure S4b, green circles). The remaining scaffolds (Figure S4b, blue circles), may represent less abundant SUP05 genotypes, as evident from the presence of an SI-2 SSU rRNA gene sequence, or

unrelated bacterial taxa sharing genomic similarity with clam symbionts. As their taxonomic identity is questionable, they were removed prior to further analysis and the remaining scaffolds are herein referred to as the SUP05 metagenome (Table 1). An automated phylogenetic approach using Phylogenie (*17*) was used on a set of 76 typically conserved genes, to assess allelic variation and to verify the absence of non-SUP05 scaffolds in the SUP05 metagenome (see Table S2).

SUP05 metagenome annotation and comparative genome analysis. Scaffolds were annotated using the FGENESB gene calling pipeline from Softberry (www.softberry.com/berry.phtml, Mount Kisco, NY). Ribosomal RNA and tRNA genes were identified using BLASTN and tRNA-Scan (*18*). Functional annotation and classification of the predicted proteome was performed by using BLASTP homology searches against COG, KEGG, and NCBI nr public database, and domain analysis with InterProScan (http://www.ebi.ac.uk/Tools/InterProScan) (*19*). Metabolic pathways were constructed based on KEGG and MetaCyc (*20*). Comparison of gene content between the SUP05 metagenome and the reference symbiont genomes was performed by BLAST Score Ratio (BSR) analysis with a BSR cut off = 0.4 (*21*). Mapping of shared gene content and COG categories onto the *R. magnifica* reference genome was performed with GenomeViz (www.uniklinikum-giessen.de/genome)(*22*). Multi genome alignment of SUP05 scaffolds and symbiont genomes was performed with Mauve (http://asap.ahabs.wisc.edu/mauve/) (*23*).

Supplementary Text

Quantification of total bacterial and SUP05 SSU rRNA gene sequences during the 2008 seasonal stratification cycle in Saanich Inlet. To assess SUP05 distribution with respect to vertical profiles of oxygen (O_2), nitrate (NO_3^-), and sulfide (H_2S), we quantified SUP05 SSU rRNA gene sequences at three time intervals representing the main stages of the 2008 seasonal stratification cycle (Fig. S2). On April 9th, an extensive suboxic zone, defined by the absence of detectable O_2 or H_2S, was observed between ~135 and 185 m, NO_3^- was depleted below 185 m and H_2S was detected at 200 m. SUP05 was detected at $\geq 1.4x10^5$ copies ml^{-1} below 150 m and peaked sharply at the base of the suboxic zone, coinciding with the point of NO_3^- depletion. By August 11th, the suboxic zone had narrowed to within ~150 and 165 m and sulfidic water was detected at 185 m. SUP05 increased in abundance from a depth of 100 to 150 m and then remained at ~ 2.5 $x10^5$ copies ml^{-1} through the suboxic zone and into the sulfidic waters below. Although the absolute abundance of SUP05 decreased at 185 m from April 9th to August 11th, as a proportion of the bacterial community SUP05 abundance increased from 20 to 30% of total bacteria. On October 15th, deep water renewal was underway and a large suboxic zone between 90 and 135 m was observed that was bracketed by oxygenated water above and below. H_2S was not detected at any depths, while NO_3^- was between 20-30 uM above and below the suboxic zone and decreased rapidly within the suboxic zone, reaching the detection limit at 135 m. Overall, SUP05 abundance was lower during deepwater renewal than during the earlier stages of stratification however the SUP05 SSU rRNA gene sequences were still detected at all depths below 100 m and ranged between ~ 3.0 $x10^4$ and 1.5 $x10^5$ copies ml^{-1}. The highest bacterial

abundance was observed within the suboxic zone during renewal (4 x10^6 copies ml^{-1}), but SUP05 was only a minor component of this deep water fall bloom as it was consistently below 10% of total detected bacteria. The abundance of SUP05 within water characterized by very little to no O_2 situated above sulfidic water suggests reliance on reduced sulfur compounds for energy conservation and also suggests the ability to do so anaerobically, most likely through the reduction of NO_3^-, in support of previous findings (*24*).

Alignment of unassembled fosmid end reads to symbiont reference genomes. We investigated the overall genome sequence conservation amongst the Saanich Inlet SUP05 populations and the symbiont reference genomes using fragment recruitment plots generated using nucmer (*25*). Nucmer plots comparing Saanich Inlet metagenomic data to the *R. magnifica* genome were constructed with the following parameters and cut-offs: breaklength = 1000, maximum gap length = 200, and minimum match length = 10. Alignment of all the unassembled fosmid end reads to the *R. magnifica* genome resulted in greater than 6000 reads (3% of total), averaging 79% nucleotide identity, aligning to the reference genome (Fig. S3). Similar results were obtained for *V. okutanii* as the reference, as the two symbionts have highly conserved genome architecture (*26*). The distribution of aligned reads over the length of the reference genome varied considerably between different Saanich Inlet samples, yet a similar pattern to the SUP05 SSU rRNA gene distribution was observed. Coverage was greatest in samples from 200 m collected in the winter and spring, however significant coverage was detected below 100 m depth throughout the year. Such high coverage is striking and suggests significant conservation of gene content between the symbiont genomes and pelagic SUP05. Moreover, only 28% of the aligning sequences were derived from fosmid mate pairs, suggesting a large proportion of genes

present in the Saanich Inlet SUP05 population are either absent or poorly conserved within the reference genome.

Insights into symbiosis. Until now, the closest genome-sequenced relative of the clam symbionts was the sulfur-oxidizing chemoautotroph *Thiomicrospira crunogena* (*27*). Although *T. crunogena* has proved useful in comparative investigation of symbiont gene content (*28*), the availability of metagenomic data from a pelagic member of the symbiont lineage allows for further study of the differences between symbiont and pelagic members of SUP05. The largest category of genes specific to the SUP05 metagenome were those involved in DNA replication recombination and repair, and have likely been lost from symbionts as is typical for obligate intracellular bacteria (*29*). Moreover, ribonucleotide biosynthesis in pelagic SUP05 can proceed via the alpha and beta subunits of ribonucleotide reductase (*nrdAB*) that is conserved with the symbionts. However, the SUP05 metagenome specifically encodes a second, oxygen sensitive reductase (*nrdD*), further supporting the facultative or strict anaerobic lifestyle of the pelagic SUP05 population in Saanich Inlet. Other genes involved in DNA metabolism specific to the SUP05 metagenome include DNA internalization and competence genes as well as a Type II secretion/type IV pilus system that may be involved in DNA uptake or perhaps protein secretion or twitching motility (*30*). The SUP05 metagenome also has many genes involved in inorganic ion transport and metabolism, reflecting the variable external environment of pelagic SUP05. These include high and low affinity sulfate transporter, nitrate/nitrite transporters, TonB dependent receptors, ferrous iron transport proteins, bacterioferritin, and alkylphosphonate uptake proteins.

Recently, a comparative analysis of two symbiont genomes lead to the hypothetical reconstruction of the last common symbiotic ancestor's (LCSA's) carbon and energy metabolism (*28*). By in large, our analysis of the SUP05 metagenome supports the findings that LCSA had a complex sulfur metabolism including both the Sox system and reverse DSR pathways of sulfur compound oxidation, fixed CO_2 via the CBB cycle, and had the ability to respire nitrate anaerobically via a membrane bound dissimilatory nitrate reductase. In addition, it was previously determined that the larger *R. magnifica* genome had many genes involved in the biosynthesis of polysaccharides and peptidoglycan that were not present in the slightly smaller *V. okutanii* genome. Many of these genes were also absent from the SUP05 metagenome (Figure S5) and hence appear specific to the *R magnifica* symbiont genome where they may play an interactive role with the clam host. A complete functional breakdown of genes shared among and between symbionts and the SUP05 metagenome is available from the authors upon request.

Supplementary Figure Legends

Fig. S1 The seasonal stratification cycle in Saanich Inlet. (A) Oxygen (O_2) (B) nitrate (NO_3^-) and (C) sulfide (H_2S) concentrations plotted from February 2006 to December 2008. H_2S data collection began in April 07. The inset at the bottom left of the Fig. (D) outlines the depths and dates of samples employed in environmental genomic analysis

Fig. S2 Quantification of SUP05 and bacterial SSU rRNA genes during the main stages of the 2008 seasonal stratification and renewal cycle. The grey box denotes the non-sulfidic suboxic zone.

Fig. S3 Identification of metagenomic reads originating from SUP05 in Saanich Inlet. For each water sample, fosmid end reads were aligned to the *R magnifica* symbiont reference genome using nucmer (*25*).

Fig. S4 Overview of scaffolds identified during the SUP05 assembly and binning process (A) Net scaffold length versus sequence depth of the assembled Saanich Inlet scaffolds, highlighting the set of scaffolds enriched with SUP05. The complete scaffold set comprises 13,628 scaffolds ≥ 1kb comprised two or more reads, with a combined net length of 21.6 Mb. (B) Visualization of the first two components of a principal component analysis of the SUP05-enriched scaffolds and known fully-sequenced SUP05 fosmids, in which Z-scores for all possible 64 tri-mers, 256 tetra-mers, and 1024 penta-mers were calculated with TETRA (*31*) and normalized by length. Scaffolds binned to SUP05 are in red while scaffolds that were included in the enriched scaffold

set but were not binned to SUP05 are in blue. The complete scaffold set is in grey. Fully sequenced SUP05 fosmids are in green. Scaffolds and fosmids containing SSU rRNA genes are indicated.

Fig. S5 Functional analysis of the genes shared between the SUP05 members based on (A) COG and (B) KEGG categories. Column colors correspond to the same sectors of the Venn diagram in Fig. 3. The number of genes is normalized to the nonredundant size of each sector of the Venn diagram, which is included in the column headers in brackets. Therefore, one can compare the functional distribution of genes within any one sector (across columns) but not between sectors (across rows).

Fig. S6 Pathways of central carbon, nitrogen, and sulfur metabolism in the SUP05 metagenome. Genes are color coded by their distribution amongst the three SUP05 representatives as revealed in the Venn diagram presented in Fig. 3.

Fig. S7 Color coding system for COG categories used in Fig. 3.

References

1. E. Zaikova *et al.*, *Environ Microbiol*, (submitted).
2. J. D. Cline, *Limnol. Oceanogr.* 14, 288 (1969).
3. T. Huber, G. Faulkner, P. Hugenholtz, *Bioinformatics* 20, 2317 (2004).
4. K. E. Ashelford, N. A. Chuzhanova, J. C. Fry, A. J. Jones, A. J. Weightman, *Appl Environ Microbiol* 72, 5734 (2006).
5. E. Pruesse *et al.*, *Nucleic Acids Research* 35, 7188 (2007).
6. W. Ludwig *et al.*, *Nucleic Acids Research* 32, 1363 (2004).
7. R. C. Edgar, *Nucleic Acids Res* 32, 1792 (2004).
8. P. D. Schloss, J. Handelsman, *Appl Environ Microbiol* 71, 1501 (2005).
9. Felsenstein, PHYLIP (Phylogeny Inference Package) version 3 (2004).
10. S. Guindon, F. Lethiec, P. Duroux, O. Gascuel, *Nucleic Acids Res* 33, W557 (2005).
11. S. J. Hallam *et al.*, *Science* 305, 1457 (2004).
12. K. Mavromatis *et al.*, *Nat Methods* 4, 495 (2007).
13. M. Pop, D. S. Kosack, S. L. Salzberg, *Genome Res* 14, 149 (2004).
14. D. Gordon, C. Abajian, P. Green, *Genome Res* 8, 195 (1998).
15. H. Teeling, A. Meyerdierks, M. Bauer, R. Amann, F. O. Glockner, *Environ Microbiol* 6, 938 (2004).
16. B. McCune, J. B. Grace, D. L. Urban. Mjm Software Design (2002).
17. T. Frickey, A. N. Lupas, *Nucleic Acids Res* 32, 5231 (2004).
18. T. M. Lowe, S. R. Eddy, *Nucleic Acids Research* 25, 955 (1997).
19. S. Hunter *et al.*, *Nucleic Acids Research* 37, D211 (2009).
20. R. Caspi *et al.*, *Nucleic Acids Research* 36, D623 (2008).

21. D. A. Rasko, G. S. Myers, J. Ravel, *BMC Bioinformatics* 6, 2 (2005).

22. R. Ghai, T. Hain, T. Chakraborty, *Bmc Bioinformatics* 5, (2004).

23. A. C. E. Darling, B. Mau, F. R. Blattner, N. T. Perna, *Genome Research* 14, 1394 (2004).

24. G. Lavik *et al.*, *Nature* 457, 581 (2009).

25. A. L. Delcher, A. Phillippy, J. Carlton, S. L. Salzberg, *Nucleic Acids Research* 30, 2478 (2002).

26. H. Kuwahara *et al.*, *Extremophiles* 12, 365 (2008).

27. K. M. Scott *et al.*, *PLoS Biol* 4, e383 (2006).

28. I. L. Newton, P. R. Girguis, C. M. Cavanaugh, *BMC Genomics* 9, 585 (2008).

29. N. A. Moran, *Curr Opin Microbiol* 6, 512 (2003).

30. F. F. Evans, S. Egan, S. Kjelleberg, *Environ Microbiol* 10, 1101 (2008).

31. H. Teeling, J. Waldmann, T. Lombardot, M. Bauer, F. O. Glockner, *BMC Bioinformatics* 5, 163 (2004).

Table S1. SUP05 metagenome features

Specifications	
No. of scaffolds, contigs	19, 90
Total length (Mb)	1.16
Total span (Mb)	1.91
Average depth of coverage	7.25
No. of assembled end reads, % of total	10449, 4.3
Average polymorphism (%)	0.4
Average G+C content (%)	40
ORF content	
No. of ORFs	1333
Predicted functional in COGs[1]	1033
Predicted functional in KEGG[1]	771
Conserved hypothetical	216
Hypothetical[2]	74
rRNA genes	
16S-5S-23S operon	1
tRNAs	29

[1]Expectation-value cut off 10E-10

[2]No similarity at an expectation value cut off 10E-5

Table S2: Quantification of SUP05 SSU rRNA genes in Saanich Inlet and summary of fosmid end sequence data

Cruise ID	Sample Date (mm/dd/yy)	Sample Depth (m)	O_2 (μM)	NO_3^- (μM)	% bacterial SSU rRNA gene library[1]	SUP05 SSU rRNA copies/ml (s.d.)[2]	No. of fosmid end reads	No. of fosmid end reads in SUP05 metagenome (% of total)
Feb06	02/18/06	10	212.0	26.7	0/0	$1.22x10^3$ ($1.1 x10^3$)	13339	4 (< 1)
Feb06	02/18/06	100	51.1	21.0	7/2	$1.66x10^4$ ($2.24 x10^3$)	12709	483 (4)
Feb06	02/18/06	125	05	9.7	17/2	$6.24x10^4$ ($2.15 x10^4$)	13860	1631 (12)
Feb06	02/18/06	215	1.4	1.8	31/6	$5.48x10^4$ ($1.33 x10^4$)	13453	1400 (10)
Jul06	07/06/06	10	381.6	0.3	0/0	0 (0)	13754	1 (< 1)
Jul06	07/06/06	100	22.9	20.4	10/1	$5.57x10^4$ ($2.32 x10^4$)	12921	513 (4)
Jul06	07/06/06	120	6.5	20.2	19/2	$2.48x10^5$ ($7.89 x10^4$)	13321	835 (6)
Jul06	07/06/06	200	0	0.1	25/2	$9.65x10^4$ ($2.46x10^4$)	12443	265 (2)
Nov06	11/14/06	10	249.0	25.5	1/0	0 (0)	12933	1 (< 1)
Nov06	11/14/06	100	15.4	13.0	11/4	$1.05x10^5$ ($3.88x10^4$)	13151	824 (6)
Nov06	11/14/06	120	9.8	8.9	26/14	$7.06x10^4$ ($5.74 x10^3$)	14810	1014 (7)
Nov06	11/14/06	200	54.0	19.8	17/8	$1.07x10^5$ ($2.07 x10^4$)	13627	990 (7)
Apr07	04/24/07	10	316.1	18.2	0/0	0 (0)	13811	3 (< 1)
Apr07	04/24/07	100	67.3	26.5	1/0	$6.35x10^3$ ($9 x10^2$)	13286	70 (< 1)
Apr07	04/24/07	120	26.9	20.6	7/1	$1.58x10^4$ ($2.37 x10^3$)	14522	390 (3)
Apr07	04/24/07	200	1.1	0.0	36/6	$3.70x10^5$ ($5.86 x10^4$)	14066	2026 (15)

[1]Report as SUP05 SI-1/SI-2 phylotypes

[2]Quantification of SUP05 SSU rRNA gene copies during Feb06, Jul06, and Nov06 were first reported in (Zaikova et al.)

Table S3. Identification and phylogenetic identity of typically conserved genes in the SUP05 metagenome[1]

COG family	Conserved genes	No. of genes (% id)[2]	SUP05 metagenome locus tag[3]	*R. magnifica* accession no.	SUP05 monophyly[4]
	Large subunit ribosomal proteins				
COG0080	**L11**	**1**	**Sup05_0527**	**YP_904007**	**+**
COG0081	**L1**	**1**	**Sup05_0526**	**YP_904006**	**+**
COG0087	**L3**	**1**	**Sup05_0550**	**YP_903428**	**+**
COG0088	L4	1	Sup05_0551	YP_903429	+
COG0089	L23	1	Sup05_0552	YP_903430	+
COG0090	L2	1	Sup05_0553	YP_903431	+
COG0091	**L22**	**1**	**Sup05_0555**	**YP_903433**	**+**
COG0093	**L14**	**1**	**Sup05_0559**	**YP_903438**	**+**
COG0094	**L5**	**1**	**Sup05_0561**	**YP_903440**	**+**
COG0097	**L6P/L9E**	**1**	**Sup05_0277**	**YP_903443**	**+**
COG0102	**L13**	**1**	**Sup05_1311**	**YP_904120**	**+**
COG0197	**L16/L10E**	**1**	**Sup05_0557**	**YP_903435**	**+**
COG0198	L24	1	Sup05_0560	YP_903439	+
COG0200	**L15**	**1**	**Sup05_0281**	**YP_903447**	**+**
COG0203	L17	2 (99)	Sup05_0288 Sup05_0652	YP_903454	+ +
COG0211	L27	1	Sup05_1289	YP_904102	+
COG0222	L7/L12	1	Sup05_0524	YP_904004	+
COG0227	L28	2 (100)	Sup05_1278 Sup05_0596	YP_904171	+ +
COG0230	L34	1	NA	YP_904212	ND
COG0244	L10	1	Sup05_0525	YP_904005	+
COG0254	L31	1	Sup05_1222	YP_904203	+
COG0255	L29	1	NA	YP_903436	ND
COG0256	**L18**	**1**	**Sup05_0278**	**YP_903444**	**+**
COG0257	L36	1	NA	YP_903449	ND
COG0261	L21	1	Sup05_1290	YP_904103	+
COG0267	L33	2	NA	YP_904170	ND
COG0291	L35	2	NA	YP_903862	ND
COG0292	L20	1	Sup05_1135	YP_903861	+
COG0333	L32	1	Sup05_0983	YP_903712	+
COG0335	L19	0	No hits found	YP_904059	ND
COG0359	L9	1	Sup05_0377	YP_903879	+
COG1825	L25	1	Sup05_0668	YP_904242	+
COG1841	L30/L7E	1	Sup05_0280	YP_903446	+
	Small subunit ribosomal proteins				
COG0048	**S12**	**1**	**Sup05_0234**	**YP_903423**	**+**
COG0049	**S7**	**1**	**Sup05_0235**	**YP_903424**	**+**
COG0051	S10	2 (100)	Sup05_0549 Sup05_0238	YP_903427	+ +
COG0052	**S2**	**1**	**Sup05_1214**	**YP_904247**	**+**
COG0092	**S3**	**1**	**Sup05_0556**	**YP_903434**	**+**
COG0096	**S8**	**1**	**Sup05_0276**	**YP_903442**	**+**
COG0098	**S5**	**1**	**Sup05_0279**	**YP_903445**	**+**
COG0099	**S13**	**1**	**Sup05_0283**	**YP_903450**	**+**
COG0100	**S11**	**1**	**Sup05_0284**	**YP_903451**	**+**

COG0103	**S9**	**1**	**Sup05_1310**	**YP_904119**	**+**
COG0184	**S15P/S13E**	**1**	**Sup05_0244**	**YP_903544**	**+**
COG0185	S19	1	Sup05_0554	YP_903432	+
COG0186	**S17**	**1**	**Sup05_0558**	**YP_903437**	**+**
COG0199	S14	2	NA	YP_903441	ND
COG0228	S16	1	Sup05_0593	YP_904167	+
COG0238	S18	1	Sup05_0376	YP_903878	+
COG0268	S20	1	Sup05_1171	YP_903947	+
COG0360	S6	1	Sup05_0375	YP_903877	+
COG0522	**S4**	**1**	**Sup05_0285**	**YP_903452**	**+**
COG0539	S1	1	Sup05_1121	YP_903814	+
COG0828	S21	1	Sup05_0653	YP_903455	+
	tRNA synthetases				
COG0008	Glutaminyl-tRNA synthetase	0	No hits found	YP_904093	ND
COG0008	Glutamyl-tRNA synthetase	1	Sup05_0971	YP_903701	+
COG0013	Alanyl-tRNA synthetase	0	No hits found	YP_904068	ND
COG0016	Phenylalanyl-tRNA synthetase alpha subunit	1	Sup05_1136	YP_903859	+
COG0017	Aspartyl/asparaginyl-tRNA synthetases	0	No hits found	No hits found	
COG0018	Arginyl-tRNA synthetase	1	Sup05_0153	YP_903350	+
COG0060	Isoleucyl-tRNA synthetase	1	Sup05_0784	YP_903587	+
COG0072	Phenylalanyl-tRNA synthetase beta subunit	1	Sup05_1137	YP_903858	+
COG0124	Histidyl-tRNA synthetase	1	Sup05_0384	YP_903884	+
COG0162	Tyrosyl-tRNA synthetase	1	Sup05_0054	YP_903397	+
COG0172	Seryl-tRNA synthetase	1	Sup05_1269	YP_904185	+
COG0173	Aspartyl-tRNA synthetase	1	Sup05_0825	YP_903634	+
COG0180	Tryptophanyl-tRNA synthetase	1	Sup05_0782	YP_903585	+
COG0215	Cysteinyl-tRNA synthetase	2 (91)	Sup05_0188 Sup05_0135	YP_903366	+ +
COG0441	Threonyl-tRNA synthetase	2 (98)	Sup05_0348 Sup05_1133	YP_903864	+ +
COG0442	Prolyl-tRNA synthetase	1	Sup05_0877	YP_903595	+
COG0495	**Leucyl-tRNA synthetase**	**0**	**No hits found**	**YP_903280**	**ND**
COG0525	Valyl-tRNA synthetase	1	Sup05_0929	YP_903692	+
COG0751	Glycyl-tRNA synthetase, beta subunit	1	Sup05_0457	YP_903923	+
COG0752	Glycyl-tRNA synthetase, alpha subunit	1	Sup05_0477	YP_903939	+
COG1190	Lysyl-tRNA synthetase	2 (99)	Sup05_1123 Sup05_0305	YP_903816	+ +
	Methionyl-tRNA synthetase	1	Sup05_1115	YP_903792	+
COG0201	**Preprotein translocase subunit SecY**	**1**	**Sup05_0282**	**YP_903448**	**+**
COG0341	Preprotein translocase subunit SecF	1	Sup05_1265	YP_904182	+
COG0342	Preprotein translocase subunit SecD	1	Sup05_1267	YP_904183	+
COG0653	Preprotein translocase subunit SecA	0	No hits found	YP_903288	ND
COG0690	Preprotein translocase subunit SecE	0	No hits found	YP_904009	ND
COG0706	Preprotein translocase subunit YidC	1	Sup05_0697	YP_904215	+
COG1314	Preprotein translocase subunit SecG	0	No hits found	No hits found	
COG1862	Preprotein translocase subunit YajC	1	Sup05_1268	YP_904184	+
COG1952	Preprotein translocase subunit SecB	0	No hits found	No hits found	
	RNA polymerase subunits				
COG0085	**DNA-directed RNA polymerase, beta subunit/140 kD subunit**	**1**	**Sup05_0523**	**YP_904003**	**+**
COG0086	DNA-directed RNA polymerase, beta' subunit/160 kD subunit	2	Sup05_0521 Sup05_0522 (potential split gene)	YP_904002	+ +

COG0202	**DNA-directed RNA polymerase, alpha subunit/40 kD subunit**	**2**	**Sup05_0287** **Sup05_0286** (potential split gene)	**YP_903453**	+ +
COG0568	DNA-directed RNA polymerase, sigma 32	0	No hits found	YP_903783	+
COG0568	DNA-directed RNA polymerase, sigma 70	1	Sup05_0106	YP_903375	+
COG1758	DNA-directed RNA polymerase, subunit K/omega	1	NA	YP_903640	ND
COG0012	**Predicted GTPase**	**1**	**Sup05_0666**	**YP_904244**	+
COG0533	**Metal-dependent protease**	**1**	**Sup05_0585**	**YP_904157**	+

[1] Genes outlined in bold are the 32 universally conserved single copy genes reported in Ciccarelli et al 2006. Genes missing from the SUP05 metagenome, but present in the *R. magnifica* reference genome are highlighted in green.

[2] For those genes present in multiple copies, the pairwise amino acid sequence similarity is presented in brackets.

[3] NA refers to genes that were present in the SUP05 metagenome but not automatically annotated by fgenesb gene-calling software.

[4] + refers to genes that formed a phylogenetically coherent cluster with the symbiont reference genomes upon automated phylogenetic reconstruction.

Table S4. Toxin-antitoxin systems identified in the SUP05 metagenome.

Toxin family[1]	Locus tag	Putative antidote locus tag	Genes in operon with TA system
relE/*parE*	Sup05_0405	Sup05_0404	MGD synthase
	Sup05_0483	Sup05_0482	-
	Sup05_0538	-	-
	Sup05_0866	-	-
	Sup05_0615	-	-
	Sup05_0057	Sup05_0056	-
	Sup05_0104	Sup05_0103	PRPP synthetase
	Sup05_0829	Sup05_0830	-
	Sup05_0293	-	-
	Sup05_1093	Sup05_1092	Selenophosphate synthase
hicB	Sup05_1327	Sup05_1326	OmpA-like protein, PIN domain proteins
	Sup05_0962	Sup05_0961	Integrases, transposases

[1]Family assignment is based on conserved domain search at NCBI

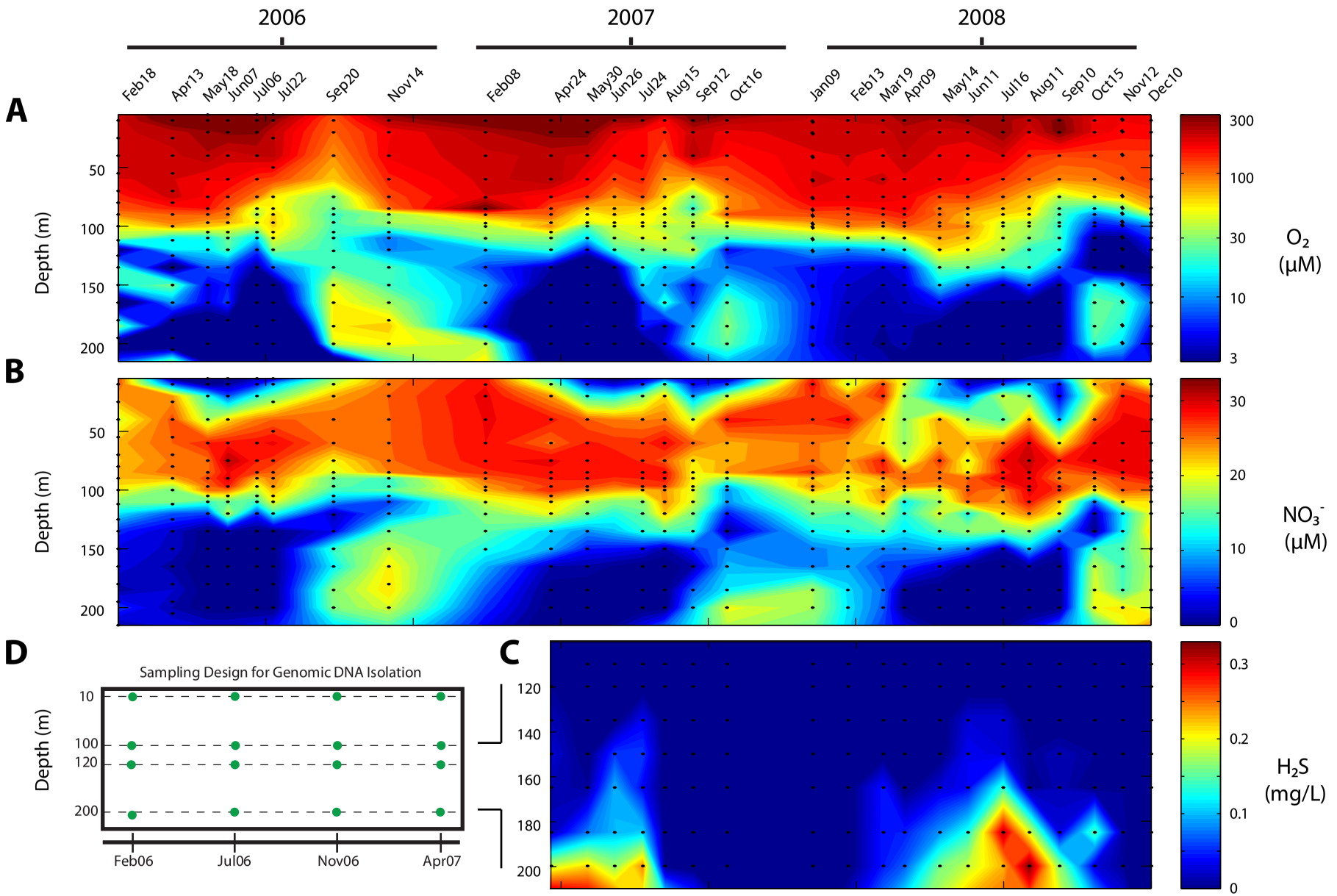

2006
2007
2008
Feb18
Apr13
May18
Jun07
Jul06
Jul22
Sep20
Nov14
Feb08
Apr24
May30
Jun26
Jul24
Aug15
Sep12
Oct16
Jan09
Feb13
Mar19
Apr09
May14
Jun11
Jul16
Aug11
Sep10
Oct15
Nov12
Dec10
A
B
C
D
Depth (m)
50
100
150
200
300
100
30
10
3
O_2 (μM)
30
20
10
0
NO_3^- (μM)
120
140
160
180
200
0.3
0.2
0.1
0
H_2S (mg/L)
Sampling Design for Genomic DNA Isolation
10
100
120
200
Feb06
Jul06
Nov06
Apr07

Fig. S1

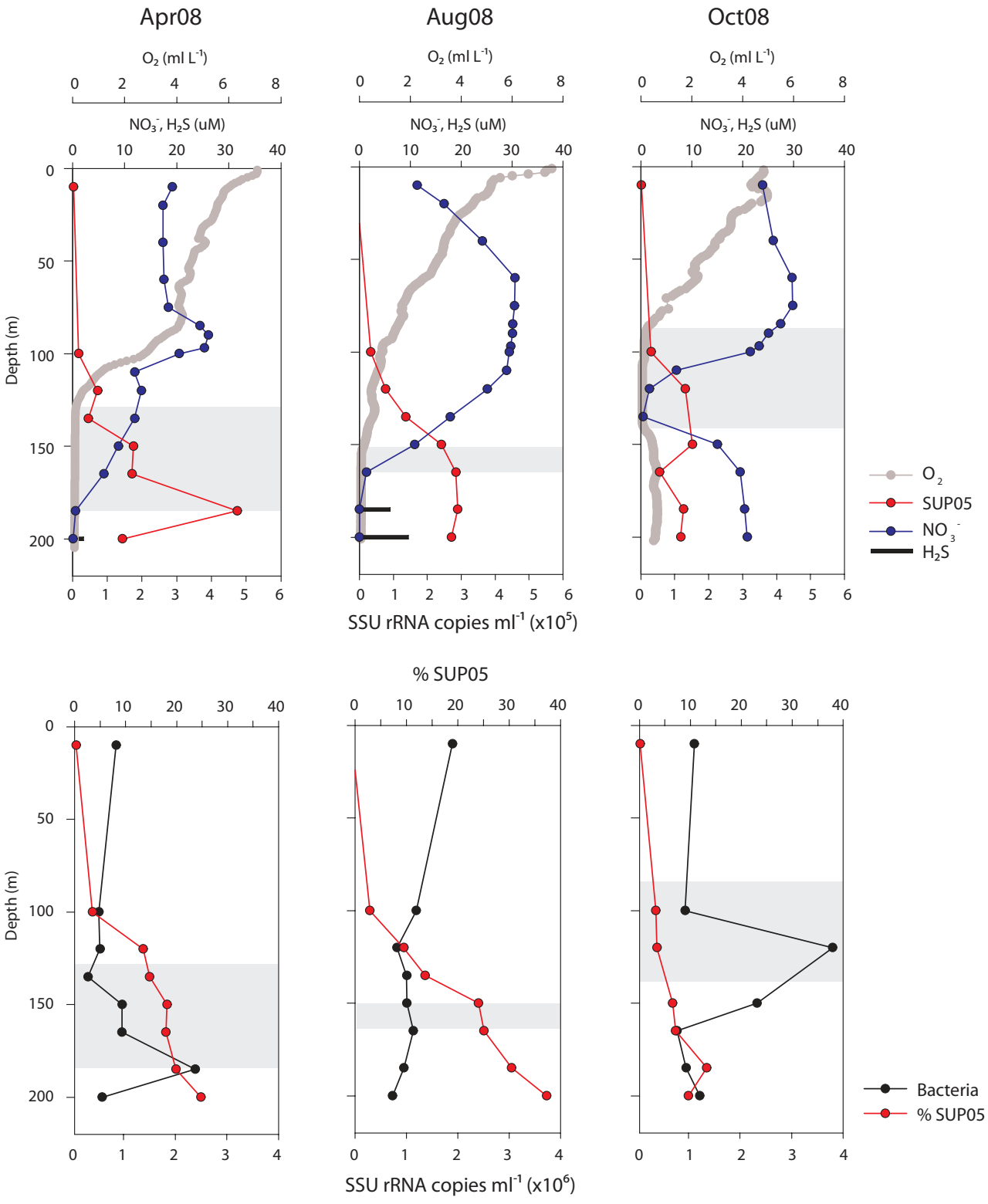

Apr08
Aug08
Oct08
O2 (ml L-1)
NO3-, H2S (uM)
Depth (m)
SSU rRNA copies ml-1 (x10^5)
% SUP05
SSU rRNA copies ml-1 (x10^6)
O2
SUP05
NO3-
H2S
Bacteria
% SUP05

Fig. S2

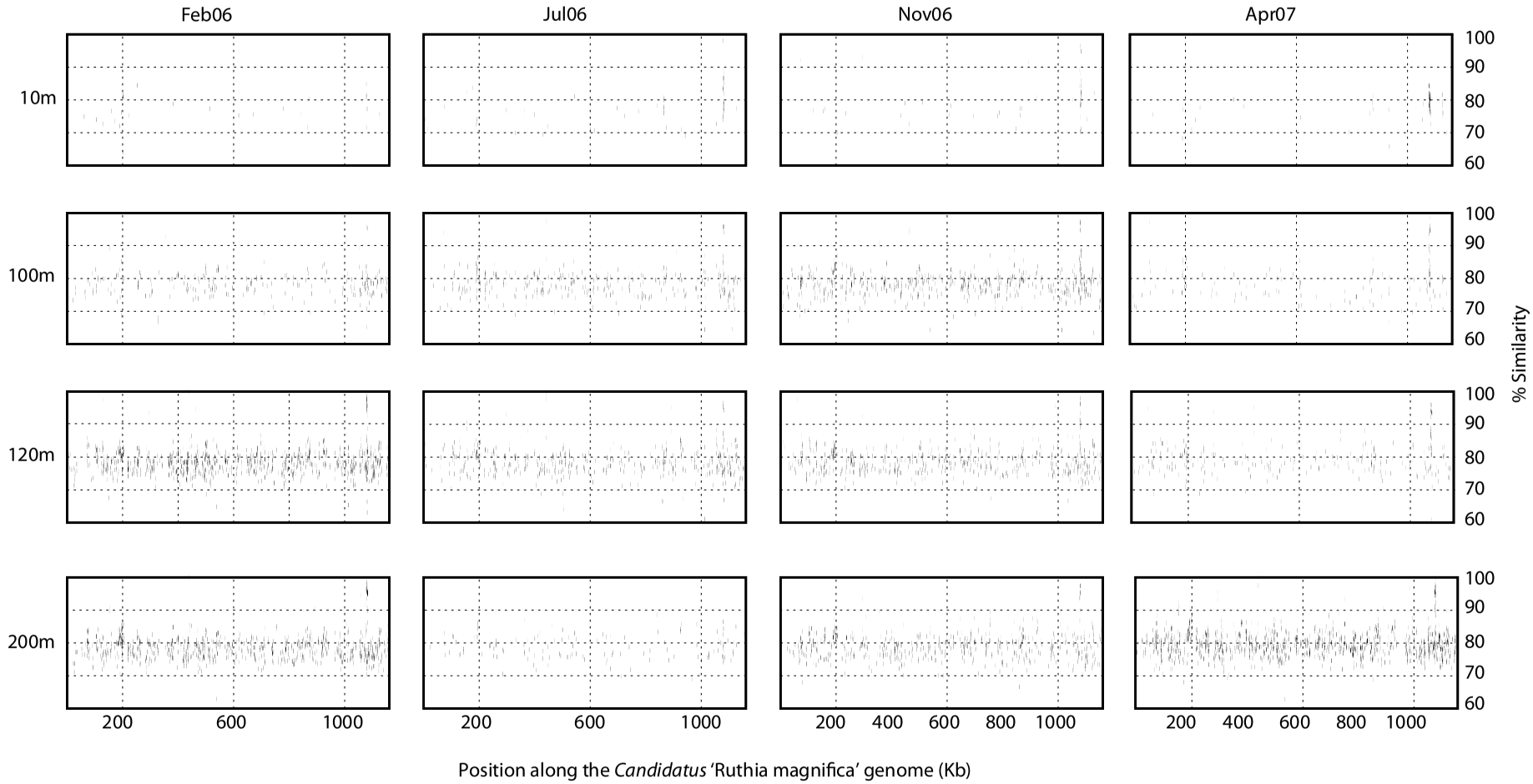

Feb06
Jul06
Nov06
Apr07
10m
100m
120m
200m
100
90
80
70
60
% Similarity
200
600
1000
200
600
1000
200
400
600
800
1000
200
400
600
800
1000
Position along the *Candidatus* 'Ruthia magnifica' genome (Kb)

Fig. S3

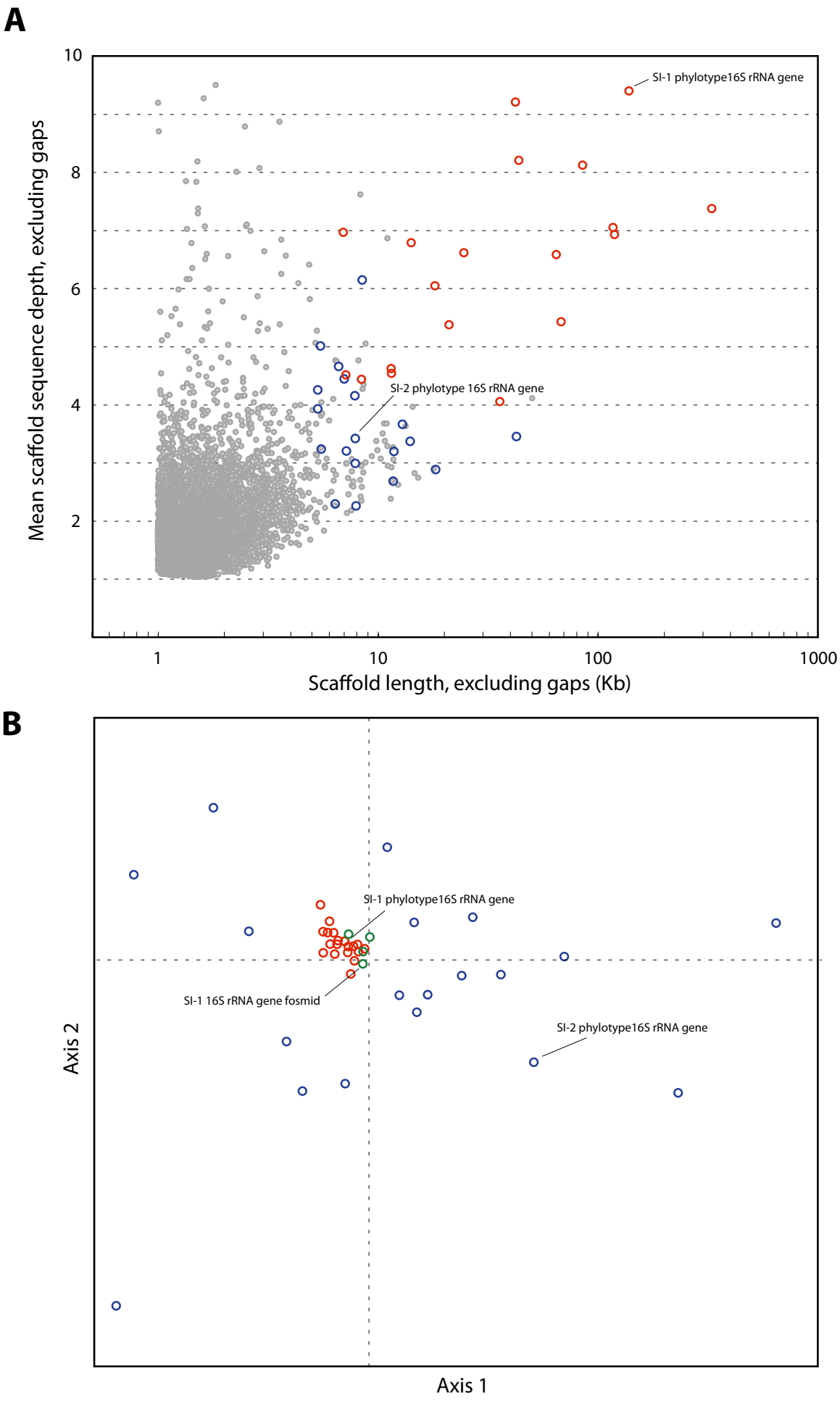

A
10
8
6
4
2
Mean scaffold sequence depth, excluding gaps
SI-1 phylotype16S rRNA gene
SI-2 phylotype 16S rRNA gene
1
10
100
1000
Scaffold length, excluding gaps (Kb)
B
SI-1 phylotype16S rRNA gene
SI-1 16S rRNA gene fosmid
SI-2 phylotype16S rRNA gene
Axis 2
Axis 1

Fig. S4

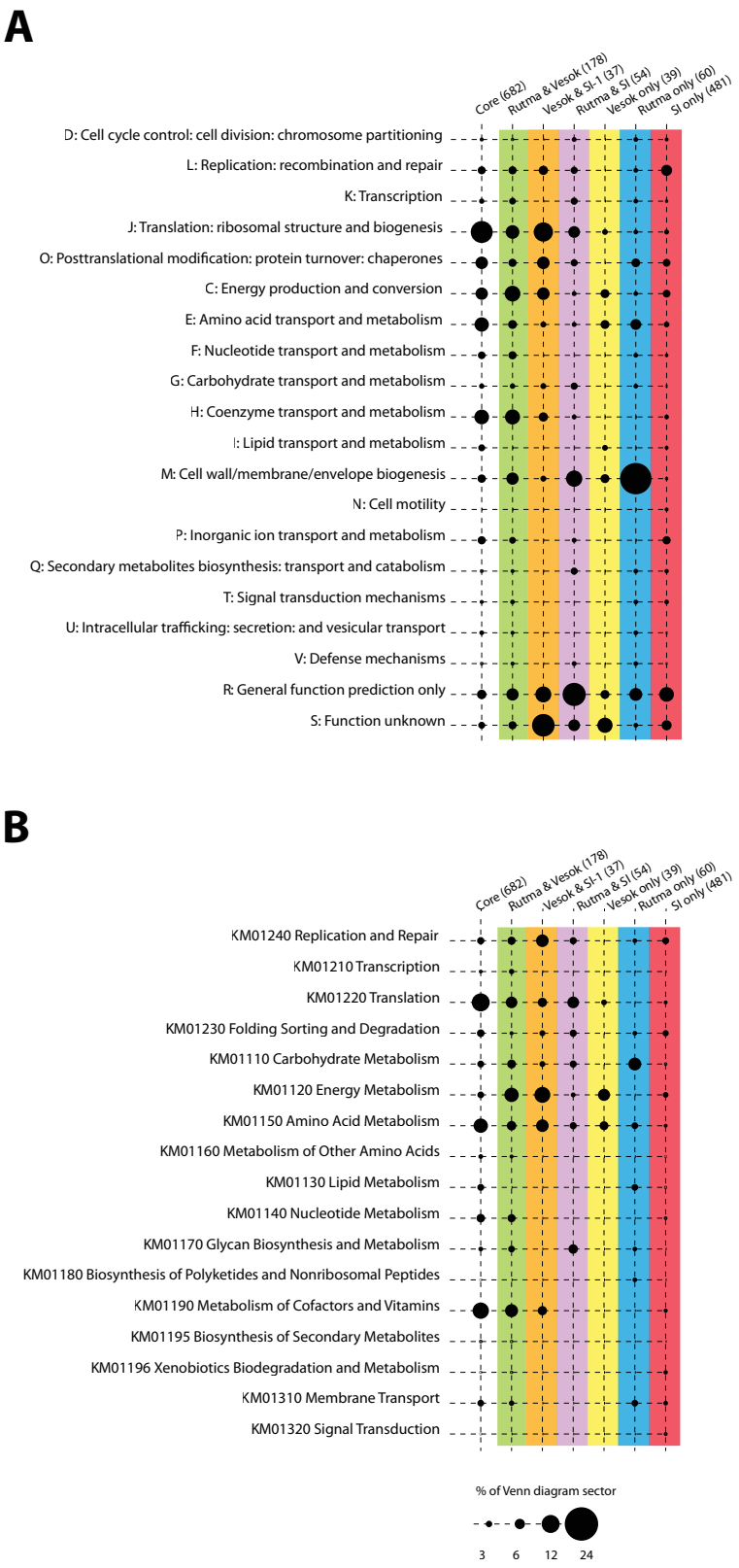

A
Core (682)
Rutma & Vesok (178)
Vesok & SI-1 (37)
Rutma & SI (54)
Vesok only (39)
Rutma only (60)
SI only (481)
D: Cell cycle control: cell division: chromosome partitioning
L: Replication: recombination and repair
K: Transcription
J: Translation: ribosomal structure and biogenesis
O: Posttranslational modification: protein turnover: chaperones
C: Energy production and conversion
E: Amino acid transport and metabolism
F: Nucleotide transport and metabolism
G: Carbohydrate transport and metabolism
H: Coenzyme transport and metabolism
I: Lipid transport and metabolism
M: Cell wall/membrane/envelope biogenesis
N: Cell motility
P: Inorganic ion transport and metabolism
Q: Secondary metabolites biosynthesis: transport and catabolism
T: Signal transduction mechanisms
U: Intracellular trafficking: secretion: and vesicular transport
V: Defense mechanisms
R: General function prediction only
S: Function unknown
B
Core (682)
Rutma & Vesok (178)
Vesok & SI-1 (37)
Rutma & SI (54)
Vesok only (39)
Rutma only (60)
SI only (481)
KM01240 Replication and Repair
KM01210 Transcription
KM01220 Translation
KM01230 Folding Sorting and Degradation
KM01110 Carbohydrate Metabolism
KM01120 Energy Metabolism
KM01150 Amino Acid Metabolism
KM01160 Metabolism of Other Amino Acids
KM01130 Lipid Metabolism
KM01140 Nucleotide Metabolism
KM01170 Glycan Biosynthesis and Metabolism
KM01180 Biosynthesis of Polyketides and Nonribosomal Peptides
KM01190 Metabolism of Cofactors and Vitamins
KM01195 Biosynthesis of Secondary Metabolites
KM01196 Xenobiotics Biodegradation and Metabolism
KM01310 Membrane Transport
KM01320 Signal Transduction
% of Venn diagram sector
3
6
12
24

Fig. S5

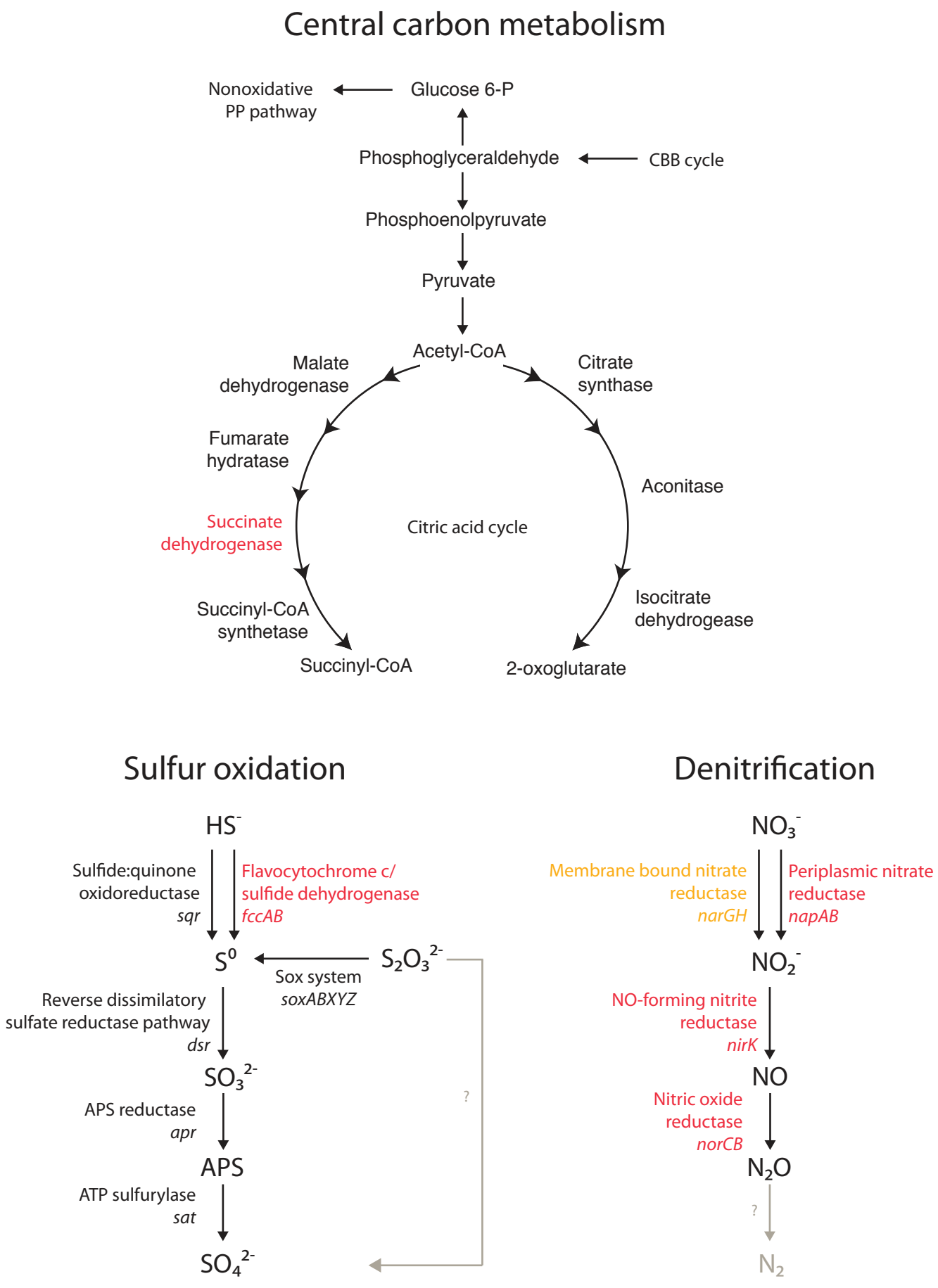

Central carbon metabolism
Nonoxidative
PP pathway
Glucose 6-P
Phosphoglyceraldehyde
CBB cycle
Phosphoenolpyruvate
Pyruvate
Acetyl-CoA
Malate
dehydrogenase
Citrate
synthase
Fumarate
hydratase
Aconitase
Succinate
dehydrogenase
Citric acid cycle
Succinyl-CoA
synthetase
Isocitrate
dehydrogease
Succinyl-CoA
2-oxoglutarate
Sulfur oxidation
HS-
Sulfide:quinone
oxidoreductase
sqr
Flavocytochrome c/
sulfide dehydrogenase
fccAB
S0
S2O32-
Sox system
soxABXYZ
Reverse dissimilatory
sulfate reductase pathway
dsr
SO32-
APS reductase
apr
APS
ATP sulfurylase
sat
SO42-
?
Denitrification
NO3-
Membrane bound nitrate
reductase
narGH
Periplasmic nitrate
reductase
napAB
NO2-
NO-forming nitrite
reductase
nirK
NO
Nitric oxide
reductase
norCB
N2O
?
N2

Fig. S6

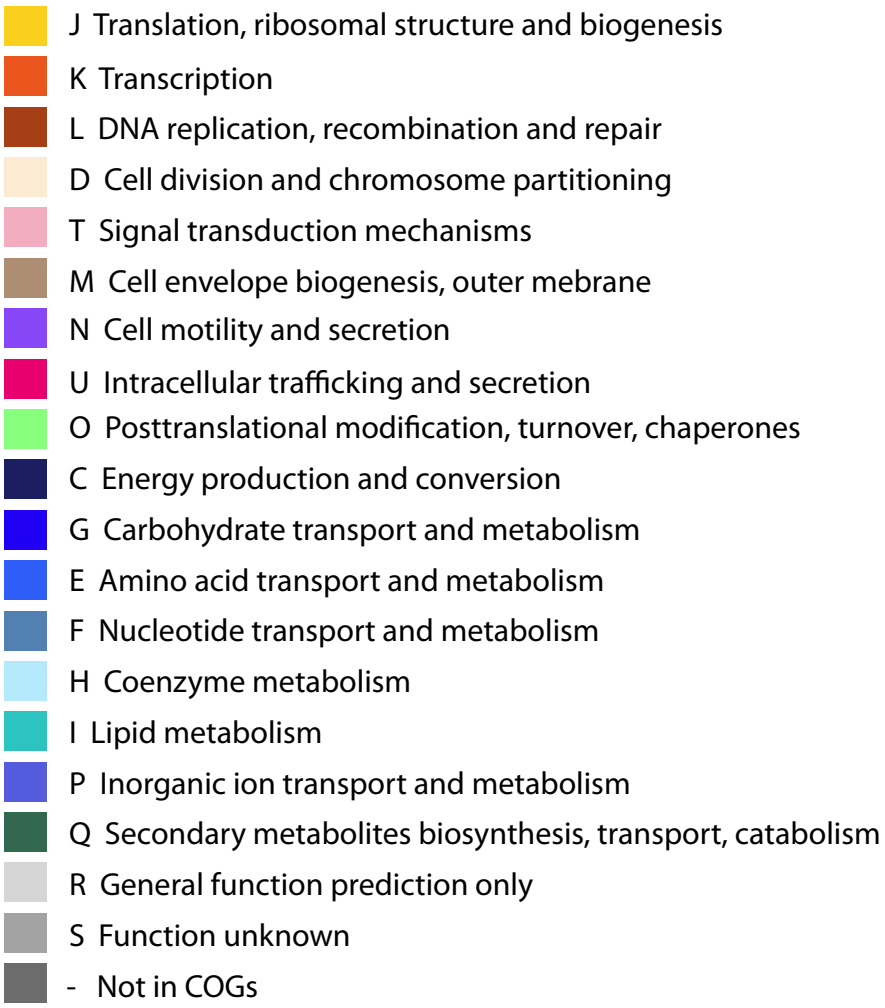

J Translation, ribosomal structure and biogenesis
K Transcription
L DNA replication, recombination and repair
D Cell division and chromosome partitioning
T Signal transduction mechanisms
M Cell envelope biogenesis, outer mebrane
N Cell motility and secretion
U Intracellular trafficking and secretion
O Posttranslational modification, turnover, chaperones
C Energy production and conversion
G Carbohydrate transport and metabolism
E Amino acid transport and metabolism
F Nucleotide transport and metabolism
H Coenzyme metabolism
I Lipid metabolism
P Inorganic ion transport and metabolism
Q Secondary metabolites biosynthesis, transport, catabolism
R General function prediction only
S Function unknown
- Not in COGs

Fig. S7