Title: Metagenomic Analysis of Microbial Symbionts in a Gutless Worm

Author(s): Tanja Woyke[1,2], Hanno Teeling[3], Natalia N. Ivanova[1], Marcel Huntemann[3], Michael Richter[3], Frank Oliver Gloeckner[3,4], Dario Boffelli[1,2], Iain J. Anderson[1], Kerrie W. Barry[1], Harris J. Shapiro[1], Ernest Szeto[1], Nikos C. Kyrpides[1], Marc Mussmann[3], Rudolf Amann[3], Claudia Bergin[3], Caroline Ruehland[3], Edward M. Rubin[1,2] & Nicole Dubilier[3]

Author Affiliations:

1. DOE Joint Genome Institute, Walnut Creek, California 94598, USA
2. Lawrence Berkeley National Laboratory, Genomics Division, Berkeley, California 94720, USA
3. Max Planck Institute for Marine Microbiology, 28359 Bremen, Germany
4. International University Bremen, 28759 Bremen, Germany

Date: 10/26/06

Funding: This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Berkeley 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.

Metagenomic analysis of microbial symbionts in a gutless worm

Tanja Woyke[1,2], Hanno Teeling[3], Natalia N. Ivanova[1], Marcel Hunteman[3], Michael Richter[3], Frank Oliver Gloeckner[3,4], Dario Boffelli[1,2], Kerrie W. Barry[1], Harris J. Shapiro[1], Iain J. Anderson[1], Ernest Szeto[1], Nikos C. Kyrpides[1], Marc Mussmann[3], Rudolf Amann[3], Claudia Bergin[3], Caroline Ruehland[3], Edward M. Rubin[1,2,†] & Nicole Dubilier[3,†]

[1]DOE Joint Genome Institute, Walnut Creek, California 94598, USA.

[2]Lawrence Berkeley National Laboratory, Genomics Division, Berkeley, California 94720, USA.

[3]Max Planck Institute for Marine Microbiology, 28359 Bremen, Germany.

[4]International University Bremen, 28759 Bremen, Germany.

† Correspondence should be addressed to: N. D. (ndubilie@mpi-bremen.de) and E. M. R. (EMRubin@lbl.gov).

All eukaryotes are inhabited in some way by commensal, mutualistic, or parasitic bacteria, yet our understanding of the intricate interactions that drive these associations is hampered by our inability to cultivate most host-associated microbes. Here, we use a metagenomic approach for a comprehensive analysis of the symbiotic microbial community in the eukaryotic host *Olavius algarvensis*. This gutless oligochaete belongs to an unusual group of marine worms harboring a highly specific consortium of multiple phylogenetically diverse bacteria, a symbiosis that has lead to the complete reduction of the worm's digestive and excretory systems. Shotgun sequencing of a bacteria-enriched sample combined with nucleotide-signature based binning enabled us to assemble two nearly complete and two partial genomes of the oligochaetes' predominant symbionts. Metabolic pathway reconstruction from the sequenced genomes revealed that these symbionts are two sulfur-oxidizing and two sulfate-reducing bacteria, all four of which are capable of autotrophic carbon fixation, thus providing their host with multiple sources of organic carbon. Molecular evidence for the uptake and recycling of worm waste products by the symbionts explains how these worms could afford to eliminate their excretory system, an adaptation unique among free-living animals. We propose a model which describes how this complex bacterial consortium provides *O. algarvensis* with an optimal energy supply as it shuttles between the upper oxic and lower anoxic coastal sediments which it inhabits.

Symbiosis plays a major role in shaping the evolution and diversity of eukaryotic organisms[1]. Remarkably, only recently has there been an emerging recognition that most eukaryotic organisms are intimately associated with a complex community of beneficial microbes that are essential for their development, health, and interactions with the environment[2]. This renaissance in symbiosis research stems largely from advances in molecular approaches that have enabled the study of natural microbial consortia using cultivation-independent methods[3-6]. To date, genomic analyses of symbiotic microbes from eukaryotes have been confined to individual strains, limiting our ability to understand the intricate interactions involving communication, competition, and resource partitioning that shape symbiotic microbial communities. Metagenomic analyses have revolutionized the study of community organization and metabolism in natural microbial communities[7-11], but have not as yet been used to describe a consortium of microbial symbionts from a eukaryotic host.

Here, we used random shotgun sequencing of the symbiotic community in the gutless marine worm *Olavius algarvensis* to obtain the first comprehensive reconstruction of multiple genomes from a eukaryotic host. The symbiotic associations in gutless marine oligochaetes are ideal for studying microbe-eukaryote symbiotic interactions because they represent a diverse yet species-limited ecosystem exposed to fluctuating environmental gradients as the host migrates between the oxic and anoxic sediment layers. The oligochaete worms are defined by the absence of mouth, gut, and anus, and are unique among free-living animals in having reduced their nephridia, excretory organs used for the removal of nitrogenous waste compounds and osmoregulation[12-14]. They live in obligate, stable, and species-specific associations with multiple bacterial symbionts

located just below the worm cuticle between extensions of the epidermal cells. The bacteria are endosymbiotic but extracellular and are separated from the environment by a thin cuticle that is freely permeable for most dissolved inorganic and organic compounds[14]. Since the obligate symbionts have yet to be grown in culture, their phylogeny has only been accessible through 16S rRNA analysis and fluorescence in situ hybridization (FISH)[15-17].

The gutless oligochaete *O. algarvensis* lives in the pore waters of coastal Mediterranean sediments[18]. These worms harbor a chemoautotrophic sulfur-oxidizing Gammaproteobacterium, called the γ1 symbiont, which is related to free-living sulfur oxidizers. This symbiont co-occurs with a deltaproteobacterial sulfate reducer, called δ1, and we recently showed that these two symbionts are engaged in a novel type of endosymbiotic sulfur cycling[16]. *O. algarvensis* harbors other gamma- and deltaproteobacterial symbionts (called γ3 and δ4), and in some individuals a spirochete has been observed as a minor part of the symbiotic consortium[14]. The metabolic capabilities of these additional symbionts and their relationships with their host *O. algarvensis* are unclear.

Given that most chemosynthetic invertebrate-microbe symbioses involve only one or two bacterial symbionts, the associations in gutless oligochaetes like *O. algarvensis* with multiple bacterial symbionts raise a series of questions about how the multiple symbionts interact with each other, their host, and their environment. What is the selective advantage for *O. algarvensis* in harboring multiple symbiotic partners, what do the various partners gain from this relationship, and is it mutually obligate? How do the interactions within the community and with the host outweigh the negative effects of

competition for space and resources between the symbionts and what role does the environment play in this symbiosis? In addition, how does the symbiosis compensate for the loss of digestive and excretory systems in the host? Here we use a metagenome shotgun approach and metabolic reconstructions to address these questions. Our analyses show how resource partitioning between the phylogenetically diverse endosymbionts benefits the worm as it shuttles between the oxidized and reduced layers of coastal marine sediments. We present how the breadth of metabolic pathways carried out between the symbionts enable *O. algarvensis* to exploit the diverse energy sources available in its heterogeneous environment. Finally, we propose a model on how the bacterial consortium meets the energy and waste management needs of its oligochaete host.

Metagenomic sequence processing and analysis

Sample characterization

O. algarvensis specimens were collected from shallow subtidal sediments (6 m water depth) off the island of Elba, Italy. *O. algarvensis* is the dominant gutless oligochaete species at the collection site, where a second gutless oligochaete species, *O. ilvae* co-occurs in low abundance[19]. To facilitate the sequencing of the endosymbiotic bacteria, we enriched for the symbionts via nycodenz density gradient centrifugation of 200 pooled, gently homogenized fresh worm specimens. DNA obtained from the bacteria-enriched sample was used for the construction of a 3 kilobase (kb) library. Due to the unavailability of large amounts of fresh sample, fosmid libraries were constructed from pooled frozen specimens that could not be gradient separated, resulting in considerable

amounts of host contamination (see Supplementary Information). We generated 185 million bases (Mb) of 3 kb library end sequence and 19 Mb of fosmid end sequence, for a total of 204 Mb of high-quality shotgun sequence data. BLAST analysis[20] of the unassembled sequence reads clearly illustrated the improved bacterial enrichment in the sample used for the 3 kb library relative to the non-fractionated sample used for the fosmid libraries, achieved by density gradient centrifugation of fresh worms (Fig. 1a, Supplementary Fig. S1a, Supplementary Information).

To estimate the relative abundance of the *O. algarvensis* symbionts, we analyzed 16S rRNA gene sequences found within the metagenomic 3 kb library as well as within a 16S rRNA PCR library obtained from the DNA used to construct the 3 kb library. Among the forty 16S rRNA genes found within the metagenomic reads, sequences derived from the *O. algarvensis* symbionts were highly represented, with the δ1 symbiont 16S rRNA genes most dominant (Fig. 1b). We found a lower abundance of sequences from *O. ilvae* γ1, from an unknown deltaproteobacterial phylotype closely related to *O. algarvensis* δ1, and from other bacterial species of largely unknown origin (Fig. 1b). 16S rRNA gene sequences from the 16S rRNA PCR library showed a similar distribution of phylotypes (Fig. 1c). A comparable analysis was carried out for the fosmid libraries (Supplementary Fig. S1b, Supplementary Information). This analysis showed that the 3 kb library largely consists of *O. algarvensis* symbiont DNA. This and the low level of eukaryotic DNA support our choice of the 3 kb library for in-depth sequencing.

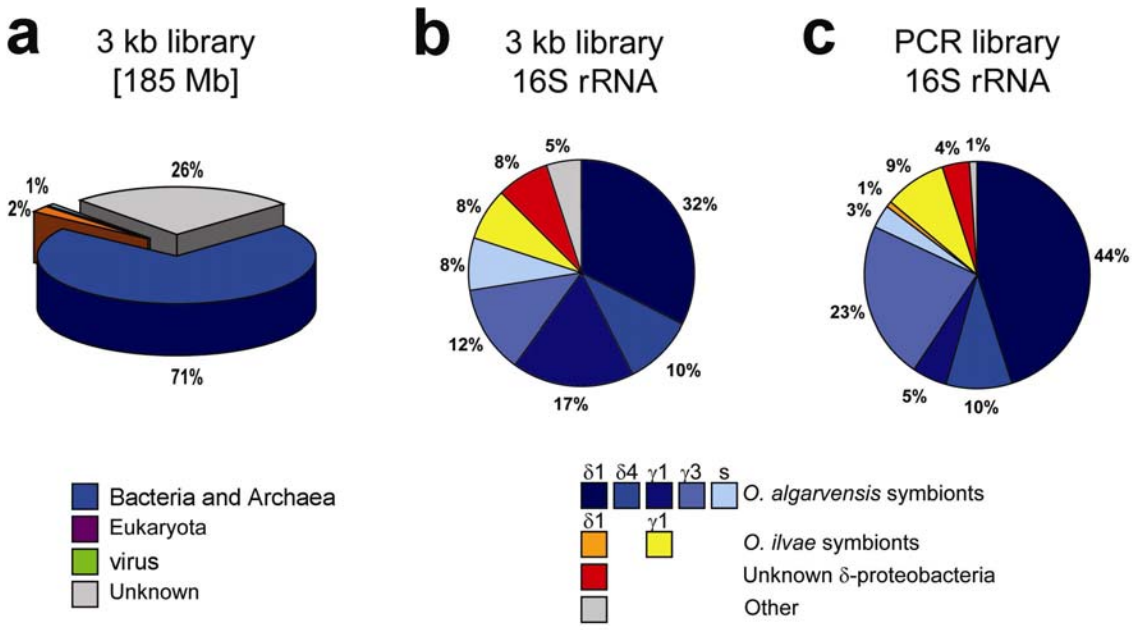


Figure 1: Characterization of the 3 kb library and 16S rRNA PCR library. (a) Percentage of 3 kb library reads (unassembled) with similarities to proteins of bacterial and archaeal, eukaryotic or viral origin (BLASTx, e-value 1e-3, NCBI nr). Unknown = reads with no similarity to proteins in the public databases. (b) Relative phylotype abundance based on 16S rRNA gene representation within the metagenomic 3 kb library reads. (c) Relative phylotype abundance in the DNA used for the 3 kb library based on 16S rRNA PCR library sequences. s = spirochete. Other includes 16S rRNA phylotypes of low abundance species.

Metagenomic data assembly and binning

To obtain the most complete and cohesive gene content for each *O. algarvensis* symbiont, metagenomic sequence reads were assembled into scaffolds. Genome assemblies of individual bacteria were reconstructed by binning scaffolds into species-specific clusters based on intrinsic DNA signatures and then assigned to one of four bacterial phylotypes (see below).

The metagenomic shotgun sequences were assembled using the whole-genome shotgun assembler JAZZ[21]. After removing short (< 1 kb of contig sequence), redundant scaffolds, we obtained a total of 2,286 scaffolds with a combined total length of 39.3 Mb (gapped length; net length of scaffolds not including gaps was 23.7 Mb; Supplementary

Fig. S2, Supplementary Information). The longest scaffold was 1.9 Mb (total length). Approximately 61% of the reads fell into this filtered assembly with 50% of the total scaffold sequence captured within the largest 106 scaffolds. The unassembled reads and short scaffolds likely represent the lower abundance species as well as worm DNA.

To assign metagenomic scaffolds to their likely phylotype origin, we used a combinatory binning approach based on GC-content, dinucleotide relative abundance[22], Markov model-based statistical evaluations of tri-, tetra and pentamer over- and under-representation[23,24] and normalized chaos game representations for tri- to hexamers[25,26]. Binning of the *Olavius* spp. symbionts' metagenome resulted in 511 scaffolds with a combined total length of 29.9 Mb (net length was 20.1 Mb; Supplementary Fig. S2) forming four distinct clusters. The presence of the corresponding rRNA operons enabled us to identify them as the *O. algarvensis* symbionts δ1, δ4, γ1, and γ3 (Fig. 2, Table 1).

This study illustrates the usefulness of our nucleotide frequency binning for the assignment of metagenomic scaffolds from a natural microbial community to a phylotype when using shotgun sequencing. It would not have been possible to separate these four genomes based only on their GC contents and scaffold read depth, as GC contents were very similar in all four bins. Moreover, no closely related, fully sequenced reference genomes for any of the symbionts was available for homology analyses that would have assisted in genome reconstruction[9].

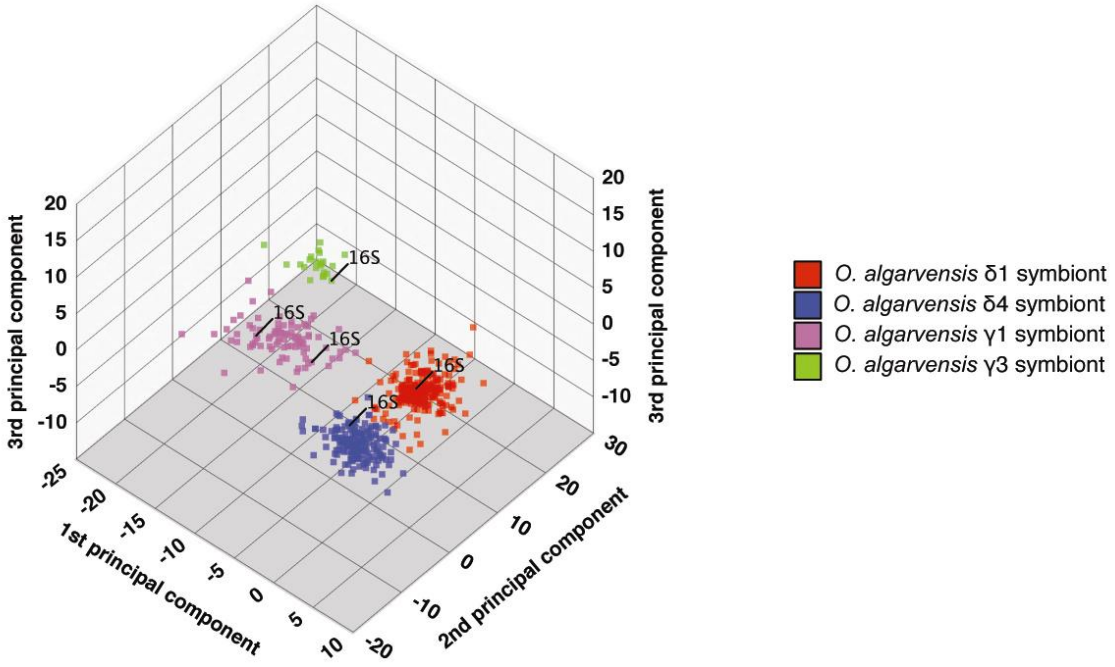


Figure 2: Clustering of the symbiont scaffolds. Visualization of the first three components of a principal component analysis (PCA), in which GC-content, net read depth, z-scores for all possible 64 trinucleotides and 256 tetranucleotides were incorporated with equal weight (z-scores calculated with TETRA and normalized by length). The colors represent the four clusters of binned scaffolds (calculated with MetaClust); sequences < 5 kb are not represented. Scaffolds containing the 16S rRNA genes are tagged.

Table 1. General features of the *O. algarvensis* symbiont bins				
	δ1	**δ4**	**γ1**	**γ3**
Assembly statistics				
Genome bin size [bp]	13,536,737	6,382,161	5,317,000	4,647,793
Gaps filled with N's [%]	16	52	73	12
Number of scaffolds	226	172	91	22
GC content [%]	49.2	54.6	57.5	55.7
Mean total read depth*	7.1	1.6	0.8	4.6
Mean net read depth*	8.4	3.3	3.0	5.2
Mean total length [bp]	59,897	37,106	58,429	211,263
Mean net length [bp]	50,593	17,887	16,059	185,783
Gene predictions				
Protein coding genes	12,084	3,012	1,872	4,154
Genes with similarity to nr	7,505	2,399	1,302	3,778
Genes with similarity to COG	5,340	1,919	831	3,083
Number of rRNA operons	1	1	2**	1
Number of tRNA genes	49	23	17	33
Number of tRNA synthetases	26	22	12	26

*normalized with respect to length.

**one complete rRNA operon and one partial 16S rRNA gene.

nr, non-redundant Genbank.

COG, clusters of orthologous groups of proteins.

Symbiont bin assignments based on 16S rRNA genes were confirmed by phylogenetic analysis of predicted proteins within each cluster of scaffolds (Supplementary Fig. S3). No phylogenetically inconsistent sub-clusters emerged. The clusters were furthermore supported by the distribution of 49 single-copy genes. Only one gene, *secG* encoding the preprotein translocase SecG subunit, was found in duplicate and only in the δ1 bin (sequence similarities between the two copies of *secG* were 95% at the nucleotide level and 98% at the amino acid level). This could indicate the presence of more than one strain in this bin and might explain its large size of 13.5 Mb (gapped length), although genomes of comparable size are known from the Deltaproteobacteria (e.g. *Polyangium cellulosum*) (http://genomesonline.org)[27]. The single occurrence of 48 out of 49 single-copy genes in the δ1 bin and the presence in single copy of other key genes, such as most

ribosomal proteins, cell division genes, flagellum genes and amino acyl tRNA synthetases (Supplementary Table S1), is however indicative of the presence of a single dominant strain. Although the populations are not clonal, the frequencies of polymorphic sites found in the four symbiont bins, ranging from 0.01-0.1% (Supplementary Table S2), were rather low compared to environmental microbes such as those from the acid mine drainage[9]. In the following discussion, we describe the metabolism of the symbionts based on the genes found in each symbiont bin. To capture these core metabolic pathways, it is not important if the genes within each bin originated from a single strain or represent a pan-genome of several very closely related strains[28].

Metagenomic insights into a symbiotic lifestyle

Life at the interface

Symbiotic associations involving chemosynthesis are inseparably linked to the geochemical properties of the environment and thrive along chemoclines, needing access to both reduced and oxidized compounds for chemosynthesis. Found in shallow Mediterranean ocean sediments at 2–15 cm depth, *O. algarvensis* inhabits the oxic-anoxic interface of this marine ecosystem, with an upper oxygenated layer and a lower anoxic zone characterized by reduced compounds such as sulfide. Most abundantly found in the deeper anoxic sediment layers at 5-15 cm depth however[29], *O. algarvensis* is believed to migrate to the upper oxygenated sediments when oxygen becomes limiting, as described for other gutless oligochaetes[30]. Here we propose that *O. algarvensis* functions as a shuttle for its inhabitants. Through its vertical migration along the oxygen-sulfide chemocline, the worm accommodates the physiological needs of its bacterial symbionts

by facilitating spatial access to the substrates necessary for their metabolism. The bacterial endosymbionts exploit the energy substrates available in the different environments to fix carbon, which is used as nutritional source by their oligochaete host, and recycle the worm's nitrogenous waste.

Microbial physiology

The availability of large metagenomic datasets for the *O. algarvensis* symbionts enabled us to explore the molecular basis of their complex interactions with each other, with the oligochaete host, and with their environmental surroundings. We analyzed and annotated the gene inventory for the four symbiont bins. Cell metabolic capabilities were reconstructed from the annotation of 4,154 protein-coding genes for the γ3 bin and 12,084 for the δ1 bin, as well as 3,012 and 1,872 protein-coding genes for the δ4 and γ1 bins, respectively (Table 1).

Carbon and energy metabolism of the gammaproteobacterial symbionts. Chemoautotrophic symbionts feed their hosts by providing them with organic carbon from autotrophic CO_2 fixation driven by oxidation of reduced inorganic compounds such as sulfide. In agreement with previous studies indicating the chemoautotrophic, sulfur-oxidizing nature of the *O. algarvensis* γ1 symbiont[14], our analysis of the γ1 bin revealed the presence of genes required for autotrophic CO_2 fixation via the Calvin-Benson-Bassham cycle using type I ribulose 1,5-bisphosphate carboxylase-oxygenase (RubisCO) (*cbb*), the oxidation of reduced sulfur compounds (such as *dsr*, *fcc* and *sox*), and the storage of sulfur in globules (*sgpB* encoding one of three known sulfur globule proteins).

Unexpectedly, our metagenomic analyses also revealed the presence of a second sulfur-oxidizing chemoautotroph in *O. algarvensis*, the γ3 symbiont. Several gutless oligochaete species are known to harbor γ3 symbionts but the metabolism of these bacteria was previously unknown and the benefit of harboring additional Gammaproteobacteria is unclear. The nearly complete genomic sequence for the *O. algarvensis* γ3 symbiont obtained in this study is notable, as this is the first sequenced genome from a chemoautotrophic symbiont. The γ3 bin carries all the genes required for a thiotrophic (sulfur-oxidizing) metabolism including those needed for the oxidation of reduced sulfur compounds (including *dsr*, *apr*, *sat*, *fcc*, and *sox*) as well as autotrophic CO_2 fixation by means of genes closely related to but phylogenetically distinct from the γ1 symbiont (Fig. 3). The absence of sulfur globule proteins in the near complete γ3 genome bin suggests that these symbionts do not store sulfur and confirms transmission electron microscopy analyses showing that only the γ1 symbionts contain sulfur globules (Giere, pers. communication). Differences in electron donor storage capabilities between the two gammaproteobacterial symbionts suggest their functional non-redundancy, indicative of resource partitioning and increased physiological flexibility when the worm experiences environmental changes. In addition to using oxygen as an electron acceptor, the presence of *nap* and *nir* gene clusters suggests that the γ3 symbionts couple oxidation of reduced sulfur compounds to dissimilatory nitrate reduction under oxygen-limiting conditions (Fig. 3).

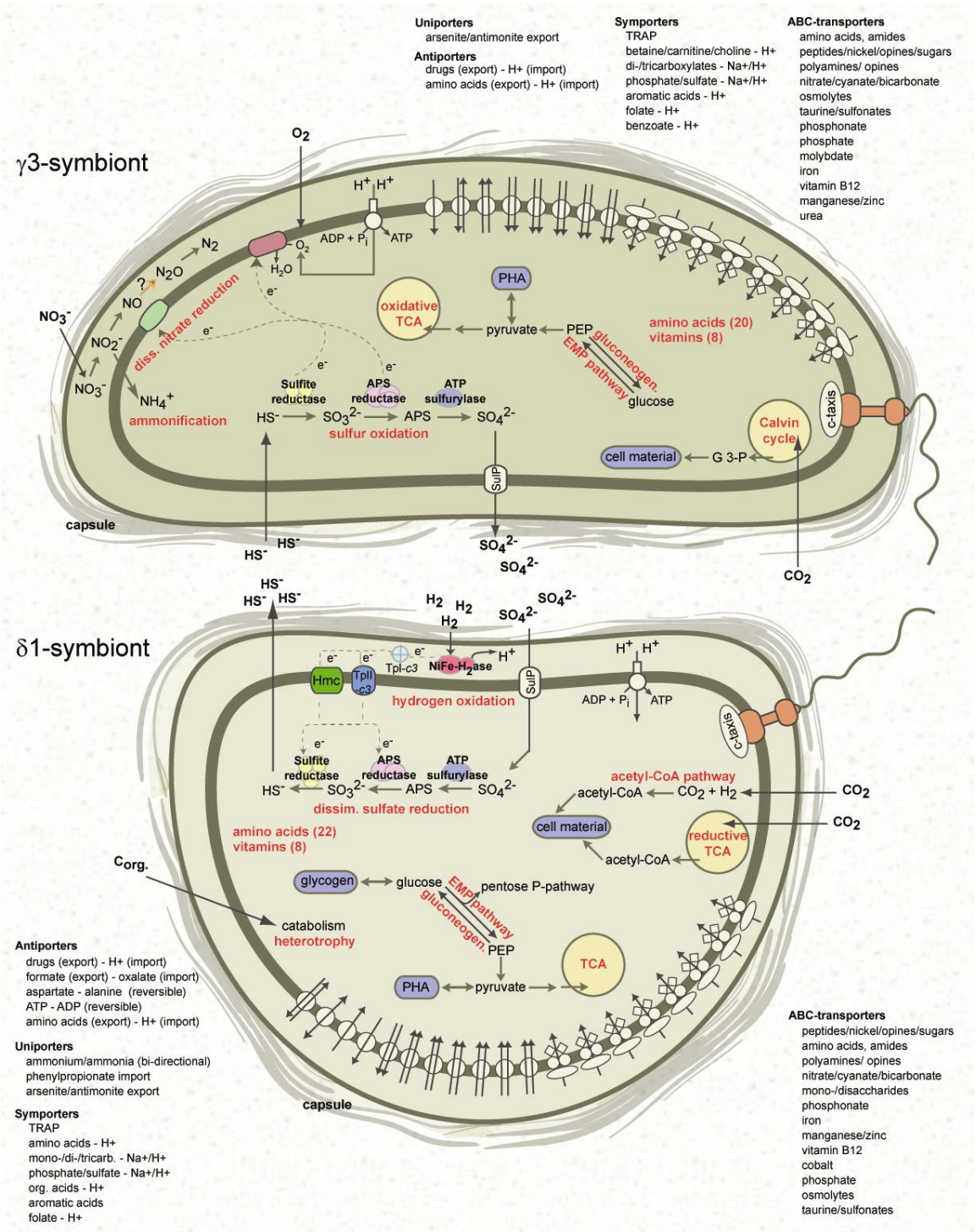


Figure 3: Reconstruction of the symbionts' physiology. PHA, polyhydroxyalkanoates. EMP, Embden-Meyerhof pathway. TCA, tricyclic acid. C-taxis, chemotaxis. APS, adenosine 5'-phosphosufate. G 3-P, glyceraldehyde 3-phosphate. Hmc, high-molecular-weight cytochrome c. TpI/II-c_3, tetrahaem type I/II tetrahaem cytochrome c_3. H_2ase, hydrogenase. PEP, phosphoenolpyruvatye. CoA, coenzymeA. ? indicates the lack of nitric oxide reductase in the γ3 bin. TRAP, tripartite ATP-independent periplasmic. Numbers in

parenthesis indicate the numbers of amino acids/vitamin synthesis pathways found (Supplementary Table S3).

Carbon and energy metabolism of the deltaproteobacterial symbionts. The presence of genes characteristic of dissimilatory sulfate reduction (such as *dsr*, *qmo*, and *apr*) in both the δ1 and δ4 bins suggests that these symbionts are sulfate-reducing bacteria that use oxidized sulfur compounds such as sulfate as an electron acceptor, thereby producing sulfide (Fig. 3). In addition to the syntrophic cycling of sulfate and sulfide between the gamma- and deltaproteobacterial *O. algarvensis* symbionts[16], intermediate sulfur compounds such as tetrathionate and thiosulfate may be cycled between the symbionts. The δ4 symbiont appears to be able to reduce sulfur compounds of intermediate oxidation states, as suggested by the presence of a multi-heme cytochrome most closely related to tetrathionate reductase of *Shewanella oneidensis*[31] and located in a chromosomal cluster with molybdopterin-dependent dehydrogenase related to thiosulfate reductase of *Wolinella succinogenes*. Cycling of intermediate sulfur compounds is energetically more favorable than the exchange of sulfide and sulfate, as shown previously in experiments with free-living sulfate reducers and sulfur oxidizers[32].

Heterotrophy is important in sulfate-reducing bacteria and correspondingly we found in the δ1 bin genes for the transport and utilization of a large variety of carbohydrate substrates, including uronic acids (glucuronate, galacturonate and fructuronate), xylose, fructose, dihydroxyacetone, and polyols (mannitol, sorbitol and glycerol). While all sulfate-reducing bacteria are heterotrophic, only some are autotrophic and it is intriguing that both deltaproteobacterial symbionts are capable of autotrophic CO_2 fixation via the reductive acetyl-coenzymeA (CoA) pathway. Moreover, both

Deltaproteobacteria likely fix carbon via the reductive tricarboxylic acid (TCA) cycle (see Supplementary Information). Thus, *O. algarvensis* has established an association with four symbionts that are all capable of providing it with organic carbon compounds through three different autotrophic pathways.

One of the most common electron donors for autotrophic sulfate-reducing bacteria is hydrogen. We found gene clusters for periplasmic Ni-Fe hydrogenases, transmembrane high-molecular-weight cytochrome c (hmc) complex, and tetraheam type II tetrahaem cytochrome c_3 (TpII-*c*$_3$) in the bins of both sulfate-reducing symbionts, as well as TpI-*c3* in the δ1 bin. This is a compelling indication for the uptake and oxidation of molecular hydrogen using sulfate as an electron acceptor[33]. It is not clear if hydrogen is provided by the gammaproteobacterial symbionts. Within the γ3 bin, we found genes encoding a pyruvate ferredoxin oxidoreductase (POR), typically used in an alternative route for pyruvate oxidation[34] and indicative of hydrogen release from low-potential ferredoxins. Released hydrogen could subsequently be taken up by the sulfate-reducing symbionts leading to hydrogen syntrophy within the microbial consortium. Alternative electron donors to hydrogen include glycerol, lactate, proline and betaine, and potentially glycolate and other 2-hydroxy acids, as well as succinate, acetate and propionate.

Symbiont host interactions

The thick layer of bacterial symbionts in the gutless oligochaetes of almost 1 million cells per host individual, representing approximately 25% of the host's body volume, highlights the importance of the symbionts for the worm[13]. The complete reduction of both the digestive and excretory systems of the worm clearly indicates the obligate nature

of this symbiosis for the host and suggests that the bacteria provide nutrients for the host as well as recycling of its waste products.

"Feeding" of the host. In other chemosynthetic associations, the symbionts provide their hosts with nutrition using either reduced sulfur compounds or methane as their energy source[35-38]. In the *O. algarvensis* symbiosis, both reduced sulfur compounds and hydrogen can be used as energy sources, and all four symbionts have the potential to fix CO_2 into organic carbon via autotrophy. In addition, the sulfate-reducing symbionts can feed the oligochaete host through heterotrophy by taking up dissolved organic carbon compounds from the environment. Largely all amino acids and a variety of vitamins can be synthesized by the symbionts to provide their host with these required nutrients (Supplementary Table S3).

It is unlikely that the symbionts feed *O. algarvensis* by excretion and transfer of metabolites. The number of genes encoding amino acid exporters was not elevated in the symbionts when compared to those of free-living bacteria. The only known family of sugar exporters[39] was not encoded in any of the symbiont bins. Furthermore, the extracellular position of the symbionts and their close physical co-occurrence might lead to competition between the host and its symbionts for excreted metabolites. It is therefore more probable that intracellular uptake and digestion of the symbionts themselves provides the host with most of its nutrition. Morphological analyses show symbionts in different stages of lysis in the basal region of the worm's epidermis[13,19].

Host waste recycling. The reduction of nephridia in the oligochaete host, used for the excretion of nitrogenous waste compounds and osmoregulation, suggests that its symbionts have taken over these functions as well. Most aquatic organisms excrete

ammonium. Urea, which also functions as a common organic osmolyte, has however been found in the annelid hydrothermal vent worm *Riftia pachyptila*[40]. We found a likely urea ABC transporter for urea uptake in the γ3 bin, adjacent to a urease operon encoding genes involved in urea hydrolysis (Fig. 4). This suggests active urea import and recycling by the microbes, which would not only aid the host in the removal of this toxic waste product, but also lead to the conservation of valuable nitrogen by the symbionts.

Marine invertebrates are typical osmoconformers, maintaining their cell volume largely with organic osmoregulatory compounds such as amino acids, taurine, glycine betaine, trimethylamine *N*-oxide (TMAO), and urea, as well as polyols and sugars[41]. The presence of gene clusters encoding proteins used for taurine and glycine betaine import and catabolism in the γ3 bin may indicate the use of these osmolytes as both a carbon and nitrogen source[42] (Fig. 4). Furthermore, we found pathways for TMAO degradation in the *O. algarvensis* symbionts. Microbial TMAO degradation includes its conversion to trimethylamine by TMAO reductase and further demethylation of trimethylamine either by trimethylamine and dimethylamine dehydrogenases found in Bacteria[43] or by trimethylamine, dimethylamine and monomethylamine methyltransferases found mostly in Archaea[44]. Genes coding for the key enzymes in both pathways were present in the γ3 and δ1 symbionts, suggestive of their TMAO catabolic activity (see Supplementary Information). The availability of the osmolyte TMAO would furthermore be particularly advantageous for the sulfur-oxidizing symbionts that could use this organic carbon compound as an alternate electron acceptor in the absence of oxygen or nitrate. As indicated, the γ3 bin encodes enzymes likely to be involved in the reduction of TMAO, although their specificity for this substrate could not be clearly identified.

Polyamines are essential in all organisms for DNA stabilization, DNA replication, and cell proliferation[45] and represent additional products of host protein breakdown. We found abundant gene clusters encoding ABC transporters for the uptake of the polyamines putrescine and spermidine in the δ1 bin, and putrescine in the γ3 bin.

Finally, evidence is provided for the recycling of host fermentation waste products such the dicarboxylate succinate, as well as the monocarboxylates acetate and propionate. Pathways for their utilization and a variety of potential dicarboxylate transporters, and likely monocarboxylate transporters were found in all four symbiont bins. The δ1 bin encodes 23 tripartite ATP-independent periplasmic (TRAP)-T family dicarboxylate transporters[46], some of which are likely involved in monocarboxylate and dicarboxylate transport.

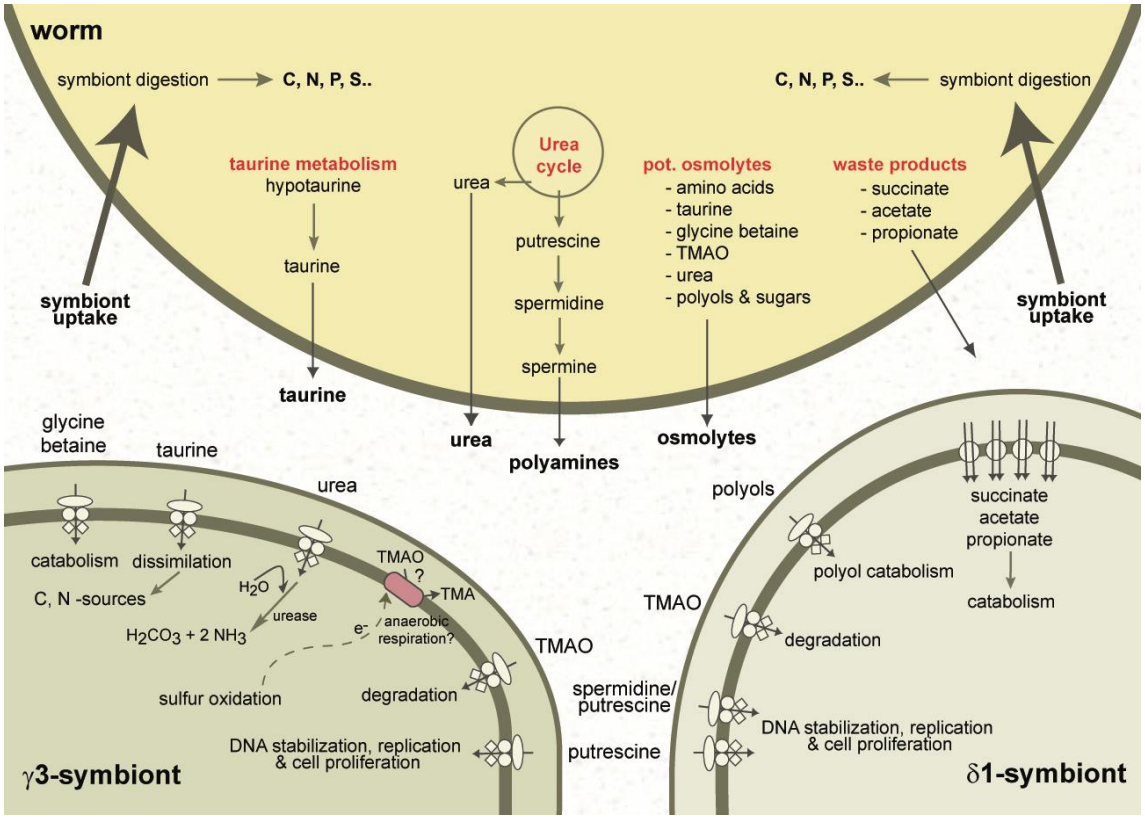


Figure 4: Reconstruction of the proposed symbiont host interactions. TMAO, trimethylamine *N*-oxide. TMA, trimethylamine. The metagenomic data uncovered pathways for the uptake and recycling of organic

osmolytes and excretion products of the worm by its symbionts.

The worm as an environmental shuttle

The analysis of the *O. algarvensis* microbial genomes provides insights on how resources are used and shared among the different symbionts and with their host and how different metabolic pathways are used by the symbionts to generate energy as the worm migrates through the chemocline (Fig. 5). In the upper oxygenated sediment layers of the *O. algarvensis* habitat where no reduced inorganic compounds are present, the γ1 symbiont can use the large supply of sulfur stored in its cytoplasm as an energy source for autotrophy with oxygen as an electron acceptor. As the worm migrates downwards it encounters sediment layers in which oxygen is no longer present but other electron acceptors are still available. Under these conditions the γ3 symbiont can use nitrate or TMAO for the oxidation of reduced sulfur compounds. Given the extremely low concentrations of sulfide in the Elba habitat (in the low nM range[16]), it is likely that the sulfate-reducing symbionts provide part of the reduced sulfur compounds for the gammaproteobacterial symbionts in a syntrophic sulfur cycle. In even deeper sediment layers characterized by reducing conditions and the absence of oxygen or nitrate, hydrogen oxidation by the sulfate-reducing symbionts may occur, in which hydrogen is used as an energy source for the autotrophic fixation of inorganic carbon. Finally, heterotrophic pathways should play an important role for the deltaproteobacterial symbionts under these reducing conditions. Degradation products from the decaying sea grass beds below the worms provide a rich source of organic matter and the worm provides a steady flux of anaerobic metabolites such as succinate, acetate, and propionate

under anoxic conditions which would further increase the net carbon gain within the symbiosis.

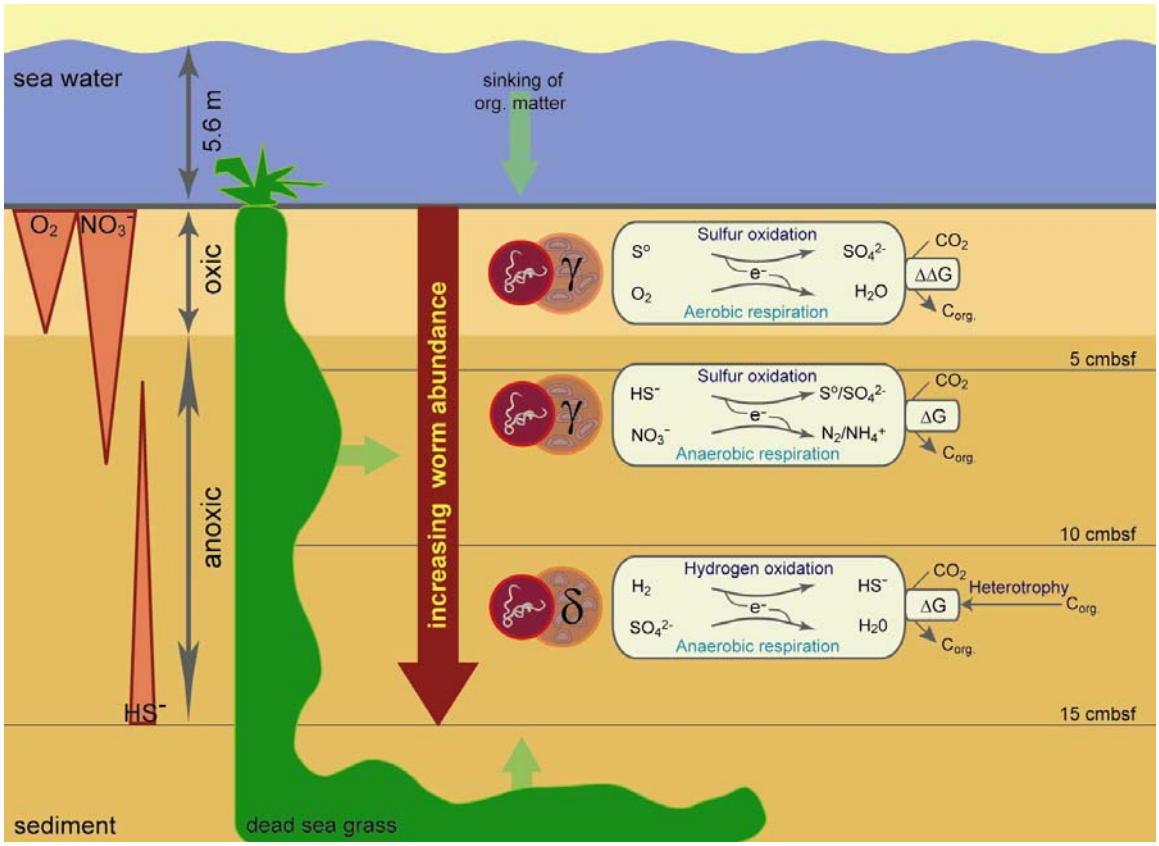


Figure 5: Model for energy metabolism in the symbiosis as the worm moves between oxic and anoxic sediment layers. Specimen abundance increases with sediment depth reaching a maximum at 10-15 cm depth, as indicated by the red arrow. cmbsf = cm below sea floor.

A mutually obligate relationship?

The lack of a digestive and excretory system in *O. algarvensis* means that its symbionts are crucial for its survival. But is this relationship mutually obligate or can the symbionts survive outside of the host in a free-living stage? We did not find any signs of obligate host-dependence of any of the symbionts. The genomes of the extracellular bacteria do not show AT-bias or genome size reduction (Table 1) and there was no obvious loss of essential metabolic pathways in the δ1 or γ3 bins suggestive of host-dependence, as is the

case for many obligate host-associated bacteria[47,48]. Finally, we found genes required for cell motility via a flagellum in the δ1, δ4, and γ3 bins. While this evidence suggests that the symbionts associated with the worm are capable of a free-living stage, the presence of a remarkable number of transposases in the γ3 and γ1 symbiont bins (7.5% and 20.5% respectively; Supplementary Tables S2, S4) suggests that these symbionts may be in transition to an obligate stage. Bacteria which have recently evolved into obligate symbionts show an increase in frequency of mobile elements, representing a source for chromosomal rearrangements and gene inactivation[49].

Symbiotic bacteria without a free-living stage are transmitted vertically from one generation to the next. In gutless oligochaetes, at least some of the symbionts may be transmitted vertically in a smear-like infection as the eggs exit the worm and pass genital pads packed with the symbiotic bacteria[50]. However, the deposition of the eggs directly into the surrounding sediments would also offer free-living bacteria from the environment an opportunity to invade the egg. It is therefore possible that some of the symbionts are inherited vertically from the parents and some horizontally from the environment. It is also possible that the same symbiont is transmitted vertically as a rule, but can also be acquired from the environment as shown for the *Wolbachia* symbionts in arthropods[51,52]. The low levels of polymorphism in all four symbiont bins of *O. algarvensis* do not exclude horizontal transmission, as selection by the host for a single bacterial strain is well known from other marine symbioses in which the symbionts are acquired from the environment, such as the luminescence symbioses between squid and *Vibrio fischeri*[53,54].

Summary of a multiple partner symbiosis

While most chemosynthetic hosts described to date harbor only one or two bacterial symbionts, gutless oligochaetes are engaged with multiple bacterial partners. What favors this symbiont diversity and how does the presence of multiple symbionts out-compete rivalry for resources and space? This study indicates that, as the worm shuttles through the sediment chemocline, it derives different benefits from each symbiont depending on the energy sources and electron acceptors present in the oxidized and reduced sediment layers. Moreover, the symbionts can use different pathways for recycling the worm's waste products. Thus, this comprehensive metagenomic analysis shows that these highly integrated synergistic assemblages of multiple bacterial partners provide their eukaryotic host with an optimal energy supply and waste management through resource partitioning as well as cooperation during sulfur syntrophy.

Methods

Sample collection and preparation

Specimens were collected in the spring and fall of 2004 at 5.6 m water depth from silicate sediments in a bay off Capo di Sant' Andrea, Elba, Italy (42°48'26"N, 010°08'28"E). The worms were removed from the sediment via decantation and washed in seawater. Symbionts from approximately 200 fresh worms were highly enriched on a discontinuous nycodenz density gradient of 1.146-1.083 g/ml. All remaining worm samples were snap-frozen and kept at –80˚C until further processing.

DNA extraction and library construction

A small insert library was constructed from the nycodenz-separated, symbiont-enriched cells. Briefly, metagenomic DNA was extracted from cells recovered from three combined nycodenz fractions using Bactozol (Molecular Research Center), which yielded 400 ng of total DNA. Approximately 200 ng of DNA was sheared to 2-4 kb and gel-extracted to construct a pMCL200 library. High-molecular weight (HMW) DNA was isolated from 200 frozen worm specimens (corresponding to ~180 mg of worm containing ~2x 10^8 bacteria) for each fosmid library using the Qiagen genomic tip procedure. DNA samples were sheared to 35 kb and gel purified after pulsed-field gel electrophoresis separation. Two pCC1Fos fosmid libraries, fosmid library-1 and fosmid library-2, were then constructed using the CopyControl^{TM} Fosmid Library Construction Kit (Epicentre).

16S rRNA PCR libraries were created from DNA sources used for each of the three metagenomic libraries to compare potential discrepancies in the symbionts'

community structure. PCR amplification was carried out using 27f and 1492r primers at 20 cycles. To minimize the accumulation of heteroduplex molecules, amplification was followed by a three cycle long 'reconditioning step'[55]. PCR products of five replicate reactions were pooled, gel-purified, and cloned into pCR4-TOPO vector (Invitrogen). For each DNA source, the 16S rRNA sequences of approximately 384 clones were sequenced and analyzed.

Sequencing, assembly and binning

End-sequencing of the shotgun libraries was carried out using PE BigDye sequencing chemistry on an ABI PRISM 3730 capillary DNA sequencer. We generated 281,448 reads totaling 204 Mb of vector-and quality-trimmed shotgun sequence. Unassembled shotgun sequences were evaluated as described in the Supplementary Information. The data was assembled using the WGS assembler JAZZ, and redundant and short scaffolds removed. This filtering left a set of 2,286 *Olavius* spp. symbionts' metagenome scaffolds (total length of 39 Mb of sequence), which were binned using a combinatory approach based on dimer to pentamer frequencies using the newly developed software system MetaClust[56]. Final clusters with 511 scaffolds were verified by phylogenetic affiliation of each scaffold based on the most common phylogeny of its predicted proteins, by a Bayesian classifier, and by checking for paralogs of 49 genes that typically occur with only one copy per genome (Kunin *et al*, unpublished).

Gene prediction and annotation

Potential open reading frames (ORFs) were identified using the in house meta gene prediction software "mORFind" (Waldmann, unpublished), a system with enhanced sensitivity and specificity which analyzes and combines the output of the gene-finders CRITICA[57], GLIMMER3[58] and ZCURVE[59]. Annotation was performed with the GenDB v2.2 system[60] and our in house software MicHanThi[61]. The annotated *Olavius* spp. symbionts' metagenome was loaded into the metagenomics version of Integrated Microbial Genomes/M (IMG/M)[62] (http://img.jgi.doe.gov/m), a data management and analysis platform for metagenomic data.

Community heterogeneity

To assess nucleotide sequence variation within the binned scaffolds or bins, we analyzed the multiple alignment of the JAZZ assembly. A site was considered polymorphic, if at least two reads showed at least two different nucleotides (or gaps) in regions covered by 4-20 reads. Frequencies of polymorphic sites (the total number of polymorphic sites divided by the total number of nucleotide sites at 4-20X read depth) were averaged over all contigs assigned to a given bin.

Details for all methods used are provided in the Supplementary Information.

1. Margulis, L. *Symbiosis in Cell Evolution* (W. H. Freeman, New York, 1993).
2. Ruby, E. G., Henderson, B. & McFall-Ngai, M. Microbiology. We get by with a little help from our (little) friends. *Science* **303**, 1305-7 (2004).
3. DeLong, E. F. Microbial population genomics and ecology. *Curr. Opin. Microbiol.* **5**, 520-4 (2002).
4. Handelsman, J. Metagenomics: application of genomics to uncultured microorganisms. *Microbiol. Mol. Biol. Rev.* **68**, 669-85 (2004).
5. Riesenfeld, C. S., Schloss, P. D. & Handelsman, J. Metagenomics: genomic analysis of microbial communities. *Annu. Rev. Genet.* **38**, 525-52 (2004).
6. Olsen, G. J., Lane, D. J., Giovannoni, S. J., Pace, N. R. & Stahl, D. A. Microbial ecology and evolution: a ribosomal RNA approach. *Annu. Rev. Microbiol.* **40**, 337-65 (1986).
7. DeLong, E. F. *et al*. Community genomics among stratified microbial assemblages in the ocean's interior. *Science* **311**, 496-503 (2006).
8. Hallam, S. J. *et al*. Reverse methanogenesis: testing the hypothesis with environmental genomics. *Science* **305**, 1457-62 (2004).
9. Tyson, G. W. *et al*. Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**, 37-43 (2004).
10. Venter, J. C. *et al*. Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**, 66-74 (2004).
11. Tringe, S. G. *et al*. Comparative metagenomics of microbial communities. *Science* **308**, 554-7 (2005).
12. Dubilier, N. Multiple bacterial endosymbionts in gutless marine worms: competitors or partners? *Nova Acta Leopoldina NF* **88**, 107-112 (2004).
13. Bright, M. & Giere, O. Microbial symbiosis in Annelida. *Symbiosis* **38**, 1-45 (2005).
14. Dubilier, N., Blazejak, A. & Ruhland, C. Symbioses between bacteria and gutless marine oligochaetes. *Prog. Mol. Subcell. Biol.* **41**, 251-75 (2006).
15. Blazejak, A., Erseus, C., Amann, R. & Dubilier, N. Coexistence of bacterial sulfide oxidizers, sulfate reducers, and spirochetes in a gutless worm (Oligochaeta) from the Peru margin. *Appl. Environ. Microbiol.* **71**, 1553-61 (2005).
16. Dubilier, N. *et al*. Endosymbiotic sulphate-reducing and sulphide-oxidizing bacteria in an oligochaete worm. *Nature* **411**, 298-302 (2001).
17. Dubilier, N., Giere, O., Distel, D. L. & Cavanaugh, C. M. Characterization of chemoautotrophic bacterial symbionts in a gutless marine worm (Oligochaeta, Annelida) by phylogenetic 16S rRNA sequence analysis and in situ hybridization. *Appl. Environ. Microbiol.* **61**, 2346-50 (1995).
18. Giere, O., Erseus, C. & Stuhlmacher, F. A new species of *Olavius* (Tubificidae, Phallodrilinae) from the Algarve Coast in Portugal, the first East Atlantic gutless oligochaete with symbiotic bacteria. *Zool. Anzeiger* **237**, 209-214 (1998).
19. Giere, O. & Erseus, C. Taxonomy and new bacterial symbioses of gutless marine Tubificidae (Annelida, Oligochaete) from the Island of Elba (Italy). *Org. Divers. Evol.* **2**, 289-297 (2002).
20. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403-10 (1990).

21. Aparicio, S. *et al*. Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science* **297**, 1301-10 (2002).
22. Karlin, S. & Burge, C. Dinucleotide relative abundance extremes: a genomic signature. *Trends Genet.* **11**, 283-90 (1995).
23. Teeling, H., Waldmann, J., Lombardot, T., Bauer, M. & Glockner, F. O. TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. *BMC Bioinformatics* **5**, 163 (2004).
24. Teeling, H., Meyerdierks, A., Bauer, M., Amann, R. & Glockner, F. O. Application of tetranucleotide frequencies for the assignment of genomic fragments. *Environ. Microbiol.* **6**, 938-47 (2004).
25. Deschavanne, P. J., Giron, A., Vilain, J., Fagot, G. & Fertil, B. Genomic signature: characterization and classification of species assessed by chaos game representation of sequences. *Mol. Biol. Evol.* **16**, 1391-9 (1999).
26. Wang, Y., Hill, K., Singh, S. & Kari, L. The spectrum of genomic signatures: from dinucleotides to chaos game representation. *Gene* **346**, 173-85 (2005).
27. Liolios, K., Tavernarakis, N., Hugenholtz, P. & Kyrpides, N. C. The Genomes On Line Database (GOLD) v.2: a monitor of genome projects worldwide. *Nucleic Acids Res.* **34**, D332-4 (2006).
28. Medini, D., Donati, C., Tettelin, H., Masignani, V. & Rappuoli, R. The microbial pan-genome. *Curr. Opin. Genet. Dev.* **15**, 589-94 (2005).
29. Perner, M. *Biogeochemische und mikrobiologische Charakterisierung mariner Sedimente vor Elba - ein Beitrag zur oekosystemaren Analyse bakteriensymbiontischer Oligochaeten*. Thesis, University Hamburg (2003).
30. Giere, O., Conway, N. M., Gastrock, G. & Schmidt, C. 'Regulation' of gutless annelid ecology by endosymbiotic bacteria. *Mar. Ecol. Prog. Ser.* **68**, 287-299 (1991).
31. Mowat, C. G. *et al*. Octaheme tetrathionate reductase is a respiratory enzyme with novel heme ligation. *Nat. Struct. Mol. Biol.* **11**, 1023-4 (2004).
32. van den Ende, F. P., Meier, J. & van Gemerden, H. Syntrophic growth of sulfate-reducing bacteria and colorless sulfur bacteria during oxygen limitation. *FEMS Microbiol. Ecol.* **23**, 65-80 (1997).
33. Matias, P. M., Pereira, I. A., Soares, C. M. & Carrondo, M. A. Sulphate respiration from hydrogen in *Desulfovibrio* bacteria: a structural biology overview. *Prog. Biophys. Mol. Biol.* **89**, 292-329 (2005).
34. Kletzin, A. & Adams, M. W. Molecular and phylogenetic characterization of pyruvate and 2-ketoisovalerate ferredoxin oxidoreductases from *Pyrococcus furiosus* and pyruvate ferredoxin oxidoreductase from *Thermotoga maritima*. *J. Bacteriol.* **178**, 248-57 (1996).
35. Cavanaugh, C. M., Gardiner, S. L., Jones, M. L., Jannasch, H. W. & Waterbury, J. B. Prokaryotic cells in the hydrothermal vent tube worm *Riftia pachyptila* Jones: possible chemoautotrophic symbionts. *Science* **213**, 340-342 (1981).
36. Cavanaugh, C. M., McKiness, Z. P., Newton, I. L. G. & Stewart, F. J. *Marine chemosynthetic symbioses.* In *The Prokaryotes: A handbook on the biology of bacteria* (eds. Dworkin, M. *et al*) (Springer Verlag, New York, 2006).

37. Giere, O. The gutless marine oligochaete *Phallodrilus leukodermatus*. Structural studies on an abberant tubificid associated with bacteria. *Mar. Ecol. Prog. Ser.* **5**, 353-357 (1981).
38. Felbeck, H., Liebezeit, G., Dawson, R. & Giere, O. CO2 fixation in tissues of marine oligochaetes (*Phallodrilus leukodermatus* and *P. planus*) containing symbiotic, chemoautotrophic bacteria. *Mar. Biol.* **75**, 187-191 (1983).
39. Liu, J. Y., Miller, P. F., Gosink, M. & Olson, E. R. The identification of a new family of sugar efflux pumps in *Escherichia coli*. *Mol. Microbiol.* **31**, 1845-51 (1999).
40. De Cian, M., Regnault, M. & Lallier, F. H. Nitrogen metabolites and related enzymatic activities in the body fluids and tissues of the hydrothermal vent tubeworm *Riftia pachyptila*. *J. Exp. Biol.* **203**, 2907-20 (2000).
41. Yancey, P. H., Blake, W. R. & Conley, J. Unusual organic osmolytes in deep-sea animals: adaptations to hydrostatic pressure and other perturbants. *Comp. Biochem. Physiol. A Mol. Integr. Physiol.* **133**, 667-76 (2002).
42. Denger, K., Ruff, J., Schleheck, D. & Cook, A. M. *Rhodococcus opacus* expresses the xsc gene to utilize taurine as a carbon source or as a nitrogen source but not as a sulfur source. *Microbiology* **150**, 1859-67 (2004).
43. Yang, C. C., Packman, L. C. & Scrutton, N. S. The primary structure of *Hyphomicrobium* X dimethylamine dehydrogenase. Relationship to trimethylamine dehydrogenase and implications for substrate recognition. *Eur. J. Biochem.* **232**, 264-71 (1995).
44. Paul, L., Ferguson, D. J., Jr. & Krzycki, J. A. The trimethylamine methyltransferase gene and multiple dimethylamine methyltransferase genes of *Methanosarcina barkeri* contain in-frame and read-through amber codons. *J. Bacteriol.* **182**, 2520-9 (2000).
45. Cohen, S. S. *A Guide to the Polyamines* (Oxford University Press, New York, 1998).
46. Kelly, D. J. & Thomas, G. H. The tripartite ATP-independent periplasmic (TRAP) transporters of bacteria and archaea. *FEMS Microbiol. Rev.* **25**, 405-24 (2001).
47. Moran, N. A. Microbial minimalism: genome reduction in bacterial pathogens. *Cell* **108**, 583-6 (2002).
48. Moran, N. A. Tracing the evolution of gene loss in obligate bacterial symbionts. *Curr. Opin. Microbiol.* **6**, 512-8 (2003).
49. Moran, N. A. & Plague, G. R. Genomic changes following host restriction in bacteria. *Curr. Opin. Genet. Dev.* **14**, 627-33 (2004).
50. Giere, O. & Langheld, C. Structural organisation, transfer and biological fate of endosymbiotic bacteria in gutless oligochaetes. *Mar. Biol.* **93**, 641-650 (1987).
51. Vavre, F., Fleury, F., Lepetit, D., Fouillet, P. & Bouletreau, M. Phylogenetic evidence for horizontal transmission of *Wolbachia* in host-parasitoid associations. *Mol. Biol. Evol.* **16**, 1711-23 (1999).
52. Huigens, M. E., de Almeida, R. P., Boons, P. A., Luck, R. F. & Stouthamer, R. Natural interspecific and intraspecific horizontal transfer of parthenogenesis-inducing *Wolbachia* in Trichogramma wasps. *Proc. Biol. Sci.* **271**, 509-15 (2004).

53. Nishiguchi, M. K., Ruby, E. G. & McFall-Ngai, M. J. Competitive dominance among strains of luminous bacteria provides an unusual form of evidence for parallel evolution in Sepiolid squid-vibrio symbioses. *Appl. Environ. Microbiol.* **64**, 3209-13 (1998).
54. Visick, K. L. & McFall-Ngai, M. J. An exclusive contract: specificity in the *Vibrio fischeri-Euprymna scolopes* partnership. *J. Bacteriol.* **182**, 1779-87 (2000).
55. Thompson, J. R., Marcelino, L. A. & Polz, M. F. Heteroduplexes in mixed-template amplifications: formation, consequence and elimination by 'reconditioning PCR'. *Nucleic Acids Res.* **30**, 2083-8 (2002).
56. Huntemann, M. *MetaClust - Entwicklung eines modularen Programms zum Clustern von Metagenomfragmenten anhand verschiedener intrinsischer DNA-Signaturen*. Thesis, University Bremen (2006).
57. Badger, J. H. & Olsen, G. J. CRITICA: coding region identification tool invoking comparative analysis. *Mol. Biol. Evol.* **16**, 512-24 (1999).
58. Delcher, A. L., Harmon, D., Kasif, S., White, O. & Salzberg, S. L. Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.* **27**, 4636-41 (1999).
59. Guo, F. B., Ou, H. Y. & Zhang, C. T. ZCURVE: a new system for recognizing protein-coding genes in bacterial and archaeal genomes. *Nucleic Acids Res.* **31**, 1780-9 (2003).
60. Meyer, F. *et al*. GenDB--an open source genome annotation system for prokaryote genomes. *Nucleic Acids Res.* **31**, 2187-95 (2003).
61. Quast, C. *MicHanThi - Design and Implementation of a System for the Prediction of Gene Functions in Genome Annotation Projects*. Thesis, University Bremen (2006).
62. Markowitz, V. M. *et al*. The integrated microbial genomes (IMG) system. *Nucleic Acids Res.* **34**, D344-8 (2006).

Acknowledgements

This work was performed under the auspices of the DOE's Office of Science, Biological and Environmental Research Program; the University of California, Lawrence Livermore National Laboratory, under contract no. W-7405-Eng-48; Lawrence Berkeley National Laboratory under contract no. DE-AC03-76SF00098; and Los Alamos National Laboratory under contract no. W-7405-ENG-36, and the Max Planck Society. We thank members of the Rubin, Jim Bristow and Phil Hugenholtz labs as well as the JGI for their contributions. We furthermore thank Victor Markowitz, Eileen Dalin, Nik Putnam, Asaf Salamov, Art Kobayashi, and Kennan Kellaris for their assistance. At the Max Planck Institute for Marine Microbiology we thank Silke Wetzel for excellent technical assistance. We are grateful to Christian Lott and the staff of the HYDRA field station at Elba for their generous support and help in sampling the worms.

Author information. Correspondence and requests for materials should be addressed to: N. D. (ndubilie@mpi-bremen.de) and E. M. R. (EMRubin@lbl.gov).

1. Supplementary Methods

Specimen collection

Juvenile and adult *Olavius algarvensis* specimens were collected in May and September 2004 from 5.6 m water depth in silicate sediments around sea grass beds of *Posidonia oceanica* in a bay off Capo di Sant' Andrea, Elba, Italy (42°48'26"N, 010°08'28"E). The worms were removed from the sediment via decantation with seawater and identified under a dissection scope. Fresh samples collected in September 2004 were kept in the original sediment and seawater for 3 days until preparation. All other specimens were cleaned by successive washes in sterile seawater, snap-frozen on dry ice and stored at –80ºC until further processing. *O. algarvensis* is the dominant species at the collection site, however an additional gutless oligochaete species named *O. ilvae* co-occurs[1]. Identification at the species level requires careful examination of individual sexually mature worms by an expert and was thus not possible for this study.

Bacterial symbiont enrichment for the 3 kb library

Bacterial cells from approximately 200 live worms (~180 mg of sample containing ~2x 10^8 bacteria) were enriched using discontinuous nycodenz density gradient centrifugation. Briefly, fresh worms were removed from the sediment via decantation, cleaned by successive washes in sterile seawater and gently homogenized in 2 ml phosphate-buffered saline (PBS), pH 7.4 using a glass dounce homogenizer. A step gradient of 1.146-1.083 g/ml density was prepared with Histodenz™ (Sigma, St. Louis, MO) in 5 ml OpiSeal Polyallomer tubes (Beckman Coulter, Fullerton, CA) and the cell suspension loaded on top of the gradient. The overlain gradient was then centrifuged at 10,000 g in a Beckman L8-M ultracentrifuge and SW 55Ti swing rotor for 1 h at 4ºC. Following centrifugation, 200-300 µl fractions were withdrawn from the bottom of the gradient tube and diluted with 10 vol of PBS to remove excess nycodenz. Cells were pelleted and resuspended in PBS and fractions evaluated semi-quantitatively for the enrichment of bacterial cells using real-time PCR amplification of 16S and 18S rRNA genes. As expected, the best enrichment of bacterial cells was found in higher density fractions, which were subsequently used for DNA extraction.

DNA extraction

For the fosmid libraries, metagenomic high molecular weight (HMW) DNA was extracted from approximately 200 pooled frozen worms for each library. The frozen worms were ground into fine powder under liquid nitrogen with mortar and pestle, transferred to a screw-cap tube and DNA extracted using the Qiagen Genomic-tip (Qiagen, Valencia, CA) according to the manufacturer's instructions. Approximately 60 µg of HMW metagenomic DNA was purified per 200 frozen specimens. For the 3 kb library, metagenomic DNA was extracted from nycodenz gradient enriched bacterial cells using Bactozol™ (Molecular Research Center, Inc., Cincinnati, OH) according to the manufacturer. Fresh cells recovered from three combined nycodenz fractions yielded ~400 ng of DNA. DNA concentrations and purities were assessed by agarose gel electrophoresis and spectrophotometric analysis.

Shotgun library construction & end-sequencing
Three metagenomic shotgun libraries were constructed for this study: one 3 kb library from live worms collected in September 2004 and two fosmid libraries from frozen worms collected in April 2004 and September 2004.

3 kb library. A small insert library was constructed from DNA derived from the nycodenz-enriched sample collected in September 2004. Briefly, 300 ng of metagenomic DNA was randomly sheared to 2-4 kb fragments using a HydroShear (GeneMachines, San Carlos, CA). The sheared DNA was separated on an agarose gel, gel-purified using the QIAquick Gel Extraction Kit and end-repaired using the End-it™ DNA End-Repair kit (Epicentre, Madison, WI) according to the manufacturer's instructions. After an additional agarose gel separation, 2-4 kb DNA fragments were gel-purified once more. The entire DNA extract was blunt-end ligated into 100 ng of pMCL200 vector O/N at 16ºC using T4 DNA ligase (Roche Applied Science, Indianapolis, IN) and 10% (vol/vol) polyethylene glycol (Sigma). The ligation was phenol-chloroform extracted, ethanol precipitated and resuspended in 15 µl TE. According to the manufacturer's instructions, 1 µl of ligation product was electroporated into ElectroMAX DH10B™ Cells (Invitrogen, Carlsbad, CA) and plated on selective agar plates. Positive library clones were picked using the Q-Bot multitasking robot (Genetix, Dorset, U.K.) and grown in selective media for sequencing.

Fosmid libraries. Fosmid libraries were constructed using the CopyControl™ Fosmid Library Production Kit (Epicentre). Briefly, ~20 µg of metagenomic DNA derived from the frozen samples was randomly sheared using a HydroShear, blunt-end repaired as described above and separated on an agarose pulse-field gel O/N at 4.5 V/cm. The 35 kb fragments were excised, gel-purified using AgarACE™ (Promega, Madison, WI) digestion, followed by phenol-chloroform extraction, and ethanol precipitation. DNA fragments were ligated into the pCC1Fos™ Vector. The ligation was packaged using MaxPlax™ Lambda Packaging Extract and used to transfect TransforMax™ EPI300 *Escherichia coli*. Transfected cells were plated on selective agar plates and fosmid clones picked using the Q-Bot multitasking robot and grown in selective media for sequencing.

End-sequencing. Plasmids were amplified using the TempliPhi™ DNA Sequencing Template Amplification Kit (Amersham Biosciences, Piscataway, NJ) and sequenced using standard M13 –28 or –40 primers and the BigDye sequencing kit (Applied Biosystems, Foster City, CA) according to the manufacturer's instructions. The reactions were purified using magnetic beads and run on an ABI PRISM 3730 (Applied Biosystems) capillary DNA sequencer (for research protocols, see www.jgi.doe.gov).

16S rRNA libraries & phylogenetic analysis
16S rRNA PCR libraries were created from DNA sources used for all three metagenomic libraries. Amplification of 16S rRNA genes was performed using the bacteria-specific universal primers 27f (5'-AGAGTTTGATCCTGGCTCAG-3') and 1492r (5'-GGTTACCTTGTTACGACTT-3')[2]. The following cycling conditions were used: 94°C for 5 min, followed by 20 cycles of 94°C for 30 sec, 55°C for 25 sec, 72°C for 90 sec, and an extension at 72°C for 7 min. To minimize heteroduplex formation, a

reconditioning step was applied[3]. Briefly, PCR reactions were diluted 10-fold into fresh reaction mixtures of the same composition and cycled three more times using the above parameters. PCR products of five replicate reactions were combined, gel-extracted as described above and ligated into the pCR4-TOPO vector using the TOPO TA Cloning Kit (Invitrogen). Ligations were then electroporated into One Shot TOP10 Electrocomp™ *E. coli* cells and plated on selective media agar plates. Approximately 384 clones per library were picked and grown in selective media for sequencing (see above).

The bi-directional 16S rRNA gene sequence reads were end-paired and trimmed for PCR primer sequence and quality. Approximately 3% of the sequences were removed as putative chimeras by identification with Bellerophon[4]. The resulting chimera-free sequences were evaluated using BLAST analysis[5] against sequences in the NCBI database and the 16S rRNA sequences of the *O. ilvae* and *O. algarvensis* symbionts (unpublished data). Phylogenetic trees were calculated by neighbor joining analyses using the ARB software package (www.arb-home.de)[6]. Only sequences ≥ 1,400 bp were used for tree construction.

Processing, analysis & assembly of shotgun data

Initial data set. The initial data set was derived from the three shotgun libraries described above. We sequenced 279,157 reads from the 3 kb library, containing 279 Mb of raw sequence. 36,095 reads were sequenced from the two 35 kb libraries, containing 37 Mb of raw sequence. The reads were screened for vector using cross_match, then trimmed for vector and quality[7]. Reads < 100 bases after trimming were excluded. This reduced the amount of data to 250,034 reads (185 Mb) of 3 kb library end-sequence, and 31,414 reads (19 Mb) of 35 kb library end-sequence.

Analysis of unassembled shotgun sequences. Unassembled shotgun sequence reads (trimmed for vector sequence and quality) were evaluated using BLAST analysis[5] against the NCBI nr database (BLASTx, e-value 1e-3) and NCBI nt database (BLASTn), as well as the 16S rRNA gene sequences of the *O. ilvae* and *O. algarvensis* symbionts (BLASTn).

Jazz assembly parameters. The data was assembled using release 2.8 of JAZZ, a WGS assembler developed at the JGI[7,8]. A word size of 13 was used for seeding alignments between reads. The unhashability threshold was set to 50, preventing words present in more than 50 copies in the data set from being used to seed alignments. A mismatch penalty of -30.0 was used, which will generally assemble sequences that are more than about 97% identical. As the different organisms in the data set were expected to be present at different sequence depths, the usual depth-based bonus/penalty system was turned off.

Post-assembly analysis. The initial assembly contained 5,016 scaffolds, with 42 Mb of sequence, of which 37% were gaps. As JAZZ links contigs into scaffolds based on fosmid paired-end information and pads these gaps with N's based on the known approximate insert fosmid size, many scaffolds are gapped. The scaffold N/L50 was 122/73 kb, while the contig N/L50 was 741/8.2 kb. If the scaffolds are sorted by total length in descending order, the scaffold N50 value is equal to the number of scaffolds one

needs to go down the list before one has encompassed half of the total scaffold sequence in the set. The scaffold L50 value is then the total length of the smallest scaffold in the "top half" of this list. The Contig N50 and L50 are values analogous to the scaffold N50 and L50 values with the difference that the contig values are calculated using net, instead of total, scaffold lengths. Redundant scaffolds were identified by aligning all scaffolds with less than 5 kb of contig sequence against those with more than 5 kb of contig sequence using BLAST-like alignment tool[9]. Any scaffolds from the former set that matched any of the larger over more than 80% of their length were excluded. Short scaffolds (< 1 kb of contig sequence) were also excluded. This filtering left 2,286 scaffolds, with 39 Mb of sequence, of which 40% was gap. The scaffold N/L50 was 106/80 kb, while the contig N/L50 was 606/9.6 kb. This filtered scaffold set served as the starting point of all downstream analyses

Binning

Scaffolds from the *Olavius* spp. symbionts' metagenome were binned by a combinatory approach based on the following intrinsic DNA signatures: (a) GC-content (b) dinucleotide relative abundance[10] (c) Markov model-based statistical evaluations of tri-, tetra and pentamer over- and under-representation[11] and (d) normalized chaos game representations for tri- to hexamers Deschavanne[12,13]. Values for (c) and (d) were computed by ocount and cgr, two self-written C-programs that are available from www.megx.net/tetra.

A self-written Java program (MetaClust[14]) was used to automatically trigger the individual calculations and subsequently store them in a MySQL database. Seven different combinations of subsets of the individual methods were built for all scaffolds exceeding 50 kb and imported into Cluster 3.0[15]. The data was normalized and a hierarchical clustering was computed using complete linkage and the Euclidian distance as distance measure. The corresponding result files were analyzed using Java TreeView (http://genetics.stanford.edu/~alok/TreeView/) and merged into consensus clusters in a semi-automatic manner by parsing the Java TreeView result files. This procedure was repeated for all scaffolds exceeding 15 kb and thereafter for all scaffolds exceeding 5 kb. After each of these steps, the newer and the former clusters were compared and ambiguous scaffolds were sorted out. A Bayesian classifier[16] was trained with all scaffolds ≥ 50 kb and subsequently used to assign some of the much shorter scaffolds with ambiguous classifications.

Final clusters were verified threefold, (a) by phylogenetic affiliation of each scaffold based on the most common phylogeny of its predicted proteins (BLASTp, e-value ≥ 1e-5, NCBI nr) (b) by a Bayesian classifier and (c) by checking for paralogs of 49 genes that typically occur with only one copy per genome (Kunin *et al*, unpublished).

Gene prediction and annotation

Potential open reading frames (ORFs) were identified using the meta gene prediction software "mORFind" (Waldmann, unpublished) developed at the MPI-Bremen. This system analyzes and combines the output of the three commonly used gene-finders CRITICA[17], GLIMMER3[18] and ZCURVE[19] to enhance sensitivity and specificity. To resolve conflicts, an iterative post-processing algorithm is used taking into account signal peptide[20] and transmembrane[21] predictions, ORF-length, and the number of gene-finders

by which an ORF has been predicted. The system was adapted to deal with typical problems of community sequencing projects like ambiguities, stretches of Ns, and fragmented genes. Annotation was performed with the GenDB v2.2 system[22], seeking for each predicted ORF observations from similarity searches against sequence databases (nr, Swiss-Prot, Kegg-Genes, release December 2005) and protein family databases (Pfam (release 19.0), InterPro (release 12.0, InterProScan version 4.2)), and from predictive signal peptide- (SignalP v3.0[20]) and transmembrane helix-analysis (TMHMM v2.0[21]). tRNA genes were identified using tRNAScan-SE[23] and rRNA genes were detected by standard similarity searches (BLAST[5]) against public nucleotide databases and the 16S rRNA sequences of the *O. ilvae* and *O. algarvensis* symbionts. Predicted protein coding sequences were automatically annotated with the software MicHanThi[24] developed at the MPI Bremen. The system simulates the human annotation process using fuzzy logic. First, informative BLAST observations are selected taking into account several BLAST parameters. The gene product is then assembled by functional clustering of observations and by selection of the most supported one. Each annotation is labelled by the corresponding reliability value to support further human inspection. Once the gene product is set, MicHanThi adds further information like gene name, EC, and GO numbers to each protein coding gene based on Swiss-Prot and InterPro observations, respectively. A functional classification was performed with similarity searches against COG v2[25]. All ORFs described in this publication were manually refined. All binned scaffolds were furthermore analyzed for the presence of clustered regularly interspaced short palindromic repeats (CRISPRs) using the CRISPR PILER-CR v1.0[26].
The annotated *Olavius* spp. symbionts' bins were incorporated into the metagenomics version of the U.S. Department of Energy Joint Genome Institute Integrated Microbial Genomes (IMG)[27], IMG/M (http://img.jgi.doe.gov/m), a data management and analysis platform for metagenomic data. This facilitates public access and visualization and comparative analyses of the data in the context of other metagenomic datasets and all publicly available complete microbial genomes.

Community heterogeneity
To assess nucleotide sequence variation within the binned scaffolds or bins, we analyzed the multiple alignment of the JAZZ assembly. A site was considered polymorphic, if at least two reads showed at least two different nucleotides (or gaps) in regions covered by 4-20 reads. Frequencies of polymorphic sites (the total number of polymorphic sites divided by the total number of nucleotide sites at 4-20X read depth) were averaged over all contigs assigned to a given bin.

Nucleotide sequence accession numbers
The unassembled raw sequence reads, as well as the annotated sequences from the *Olavius* spp. symbionts' metagenome have been deposited with the public databases under the project accessions xxxxxx.

2. Supplementary Text

Sample characterization

To characterize the origin of the *Olavius* spp. shotgun sequences, we used BLAST[5] analysis to assign unassembled reads to the category of likely bacterial and archaeal, eukaryotic or unknown origin. Approximately 71% of the metagenomic 3 kb library reads showed BLASTx-based similarity to bacterial and archaeal (vastly proteobacterial) proteins, whereas only 2% showed similarity to eukaryotic proteins and 26% exhibited no similarities to any proteins in the database (Fig. 1a). The fosmid libraries had a greatly decreased ratio of bacterial to host sequences (Supplementary Fig. S1a).

We performed 16S rRNA PCR amplification of each source DNA using parallel PCR reaction aliquots, low PCR cycle numbers and a reconditioning step to minimize PCR bias as well as chimera and heteroduplex formation[3]. The 16S rRNA gene analysis of the DNA source used for the 3 kb library construction revealed that sequences derived from the *O. algarvensis* δ1 and γ3 symbionts were highly represented (Fig. 1c). Sequences from *O. algarvensis* δ4, γ1 and spirochete symbionts, as well as an unknown Deltaproteobacteria closely related to *O. algarvensis* δ1 and *O. ilvae* γ1 and other contaminating bacterial species were present at much lower abundance. 16S rRNA gene analyses of DNA used for the fosmid libraries showed that all three samples differed in their symbiont species distribution and in contamination with *O. ilvae* symbionts (Supplementary Fig. S1b, Fig. 1c). This may be due to variations in oligochaete species distribution patterns.

Correlation of bins and 16S rRNA gene representations

The bin of the sulfur oxidizing symbiont γ3 contained 22 scaffolds with a mean net read depth of 5.3X, while the δ1 bin comprised 226 scaffolds at 8.4X mean net read depth (Table 1). These read depths were consistent with the symbionts' sequence representations based on the 16S rRNA PCR analysis of the metagenomic DNA used for the shotgun library construction (Fig. 1c, Supplementary Fig. S1b). The highly abundant δ1 16S rRNA gene sequences in the DNA used for the 3 kb library construction were reflected in the high sequence read depth for this genome bin, whereas the less abundant γ3 16S rRNA gene sequences mirrored a lower read depth within the γ3 bin. The high abundance of γ3 16S rRNA gene sequences in the DNA used for fosmid construction moreover was mirrored in the great length of the γ3 bin scaffolds (with the longest scaffold comprising 1.9 Mb total length). The mean net read depth for the scaffolds in the δ4 and γ1 bins was ~3X and, as expected, scaffolds contained a significant number of gaps (52% and 73% of N's, respectively). We found 172 scaffolds contained in the bin for δ4 and 91 scaffolds for γ1.

Microbial Physiology

Carbon metabolism

All four symbionts encode glycolysis represented by Embden-Meyerhof pathway. We found furthermore gluconeogenesis encoded in the symbiont genome bins δ1 and γ3, and an oxidative branch of pentose-phosphate pathway encoded in the δ1 and γ1 bins. Methylglyoxal bypass pathway is found in the δ4 genome bin. Both

gammaproteobacterial bins possess the genes required for the oxidative tricyclic acid cycle.

Carbon fixation within both deltaproteobacterial symbionts is possible via the reductive acetyl-CoenzymeA (CoA) pathway and also likely via the reductive tricarboxylic acid (TCA) cycle. Both bins encode citrate lyase and oxoglutarate-ferrodoxin oxidoreductase, which catalyze two out of three potentially irreversible steps in the TCA. The third potentially irreversible step is catalyzed by succinate dehydrogenase (SDH). Depending on the specific implementation of this enzyme, it could be either reversible (as in *Bacillus subtilis*) or irreversible (as in *E. coli*, which has a separate succinate dehydrogenase and fumarate reductase). Succinate dehydrogenase is encoded in both proteobacterial symbionts, yet we are unable to determine in which direction their SDH catalyzes its reaction and whether it is reversible. The proteins highest percent identity is to the actinobacterial succinate dehydrogenase. In actinobacteria, this enzyme strictly catalyzes succinate oxidation, while a second enzyme catalyzes fumarate reduction. However, in general the direction of reaction of SDH/QFR is dependent on the structure and type of cytochrome subunit; yet we were unable to find any cytochrome subunit in deltaproteobacterial bins (and no hydrophobic anchor subunit for that matter). The lack of subunit may either be due to the incompleteness of these genome bins, or the symbionts have a very unusual form of this enzyme, a soluble form. A soluble fumarate reductase is known in methanogenic archaea, where a soluble donor (coenzyme M) is used instead of a quinone.

For the storage of carbon, genes for glycogen synthesis were found in the γ1 bin and for polyhydroxyalkanoates (PHA) synthesis in the γ3 symbiont. In the bins of both deltaproteobacterial symbionts, we found genes for glycogen and PHA synthesis.
The presence of the fatty acid oxidation complex in the δ1 and γ3 bin indicates their ability for complete oxidation of fatty acids.

Nitrogen metabolism
In addition to denitrification, ammonification is encoded by the γ3 symbiont.

Cell wall and capsule
Genes required for peptidoglycan and lipopolysaccharide biosynthesis pathways are encoded in all four symbiont bins. All symbiont bins furthermore encode biosynthesis of capsule polysaccharides, which may be used by the symbionts for innate immunity protection from the host.

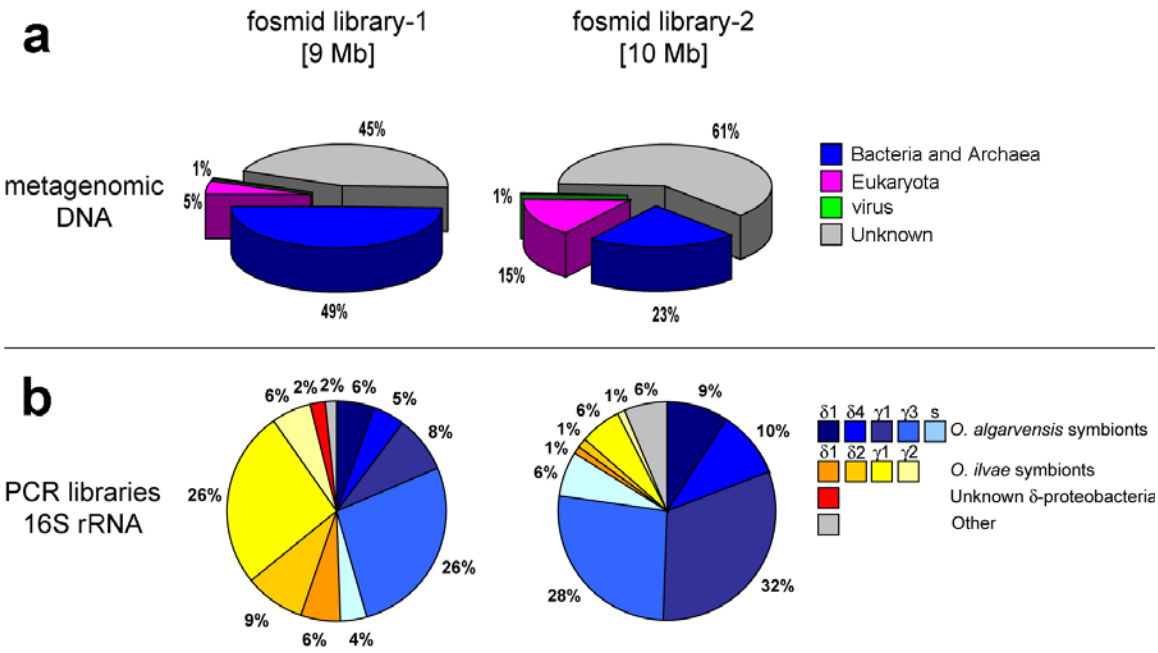
Osmolyte breakdown
The microbial degradation of trimethylamine N-oxide (TMAO) includes its conversion to trimethylamine by TMAO reductase and further demethylation of trimethylamine either by trimethylamine and dimethylamine dehydrogenases found in Bacteria[28] or by trimethylamine, dimethylamine and monomethylamine methyltransferases found mostly in Archaea[29]. Both pathways are present in the *O. algarvensis* symbionts: several homologs of trimethylamine/dimethylamine dehydrogenase are found in the γ3 and δ1 symbiont bins and 22 proteins from the trimethylamine:corrinoid methyltransferase family were found in γ3, δ1 and δ4 symbionts, as well as on unassigned scaffolds. Six genes coding for dimethylamine:corrinoid methyltransferase and two genes encoding

monomethylamine:corrinoid methyltransferase are present in the δ1 bin. Like their archaeal counterparts, at least two proteins from the δ1 symbiont coding for trimethylamine methyltransferases and all proteins encoding dimethylamine methyltransferases are interrupted by an amber stop codon UAG, which is most likely translated to pyrrolysine. This is supported by the presence of pyrrolysine-specific aminoacyl-tRNA synthetase and PylS-associated genes in the δ1 symbiont. Tetrahydrofolate appears to be the most likely acceptor of methyl groups in the symbionts due to the presence of at least three homologs of methylcobalamin:tetrahydrofolate methyltransferase MtvA from the vanillate demethylase complex of *Moorella thermoacetica*[30]. Some of the methylamine methyltransferase genes are found in chromosomal clusters with the genes encoding putative enzymes from the glycine oxidase/dimethylglycine dehydrogenase family and the molybdopterin dehydrogenase family, suggesting the existence of branched and possibly novel pathways for degradation of trimethylamine N-oxide.

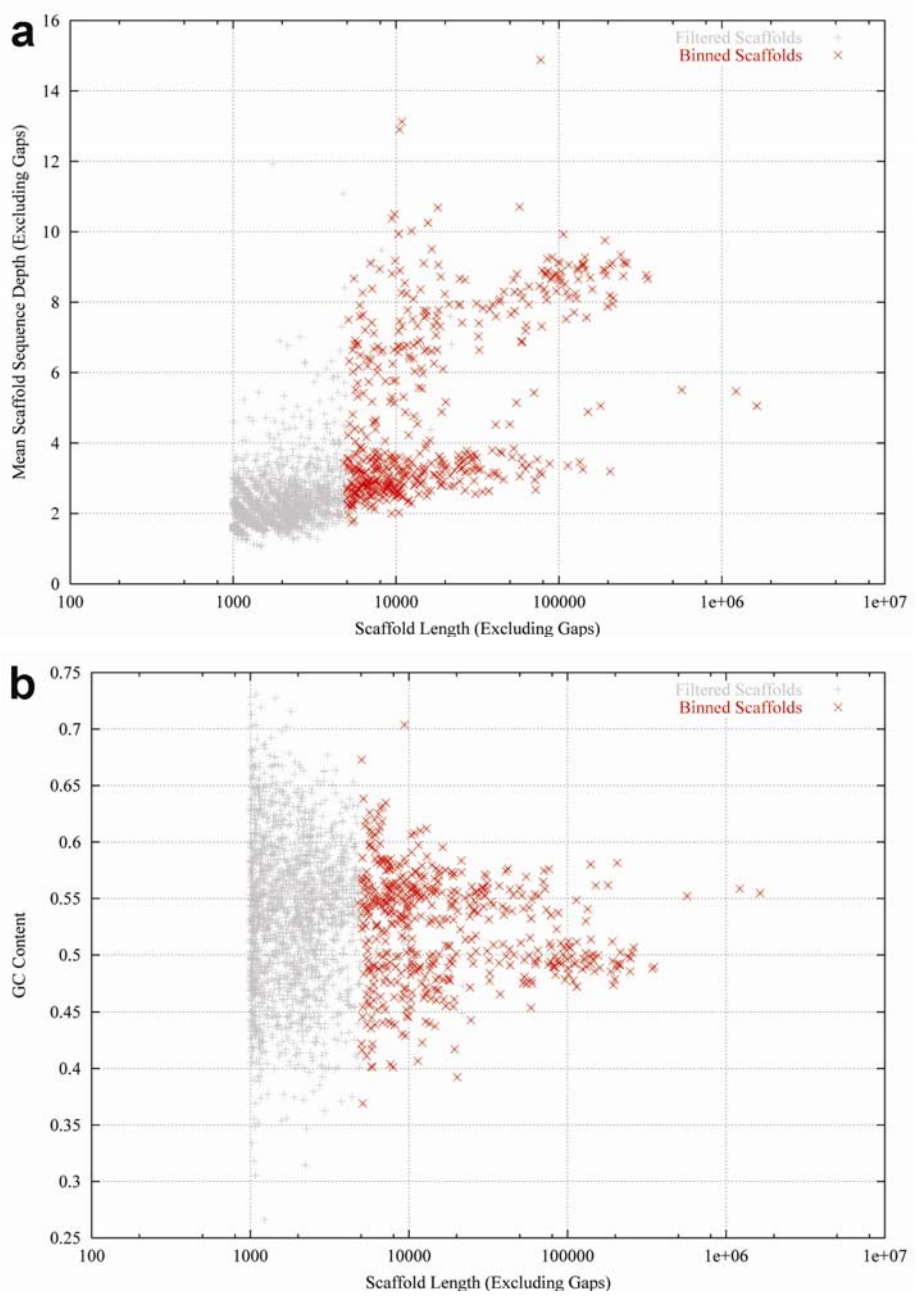
COG analysis

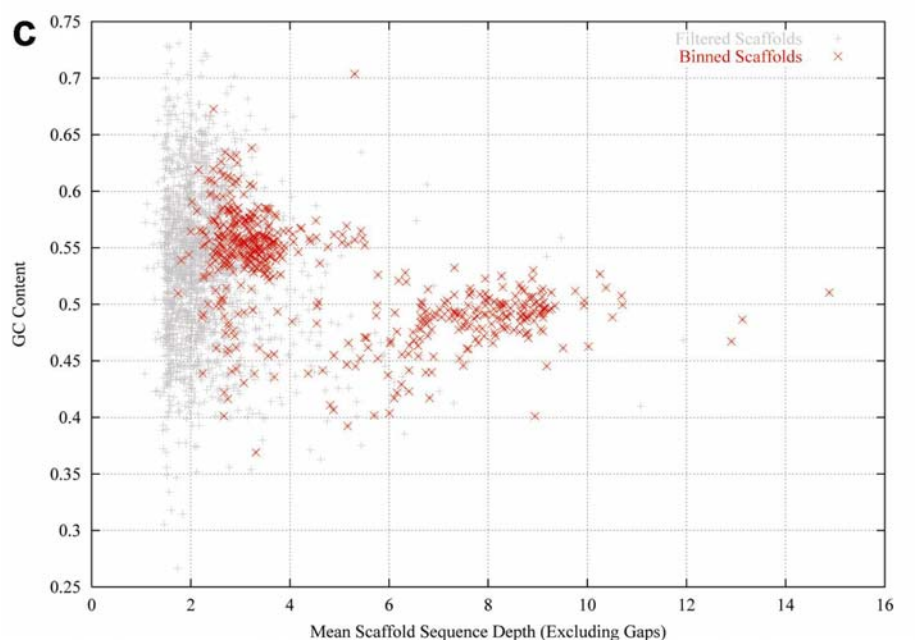
For distribution of the genes identified for δ1 and γ3 bins by broad functional category of clusters of orthologous groups of proteins (COGs) see Supplementary Fig. S4.

3. Supplementary Figures and Legends

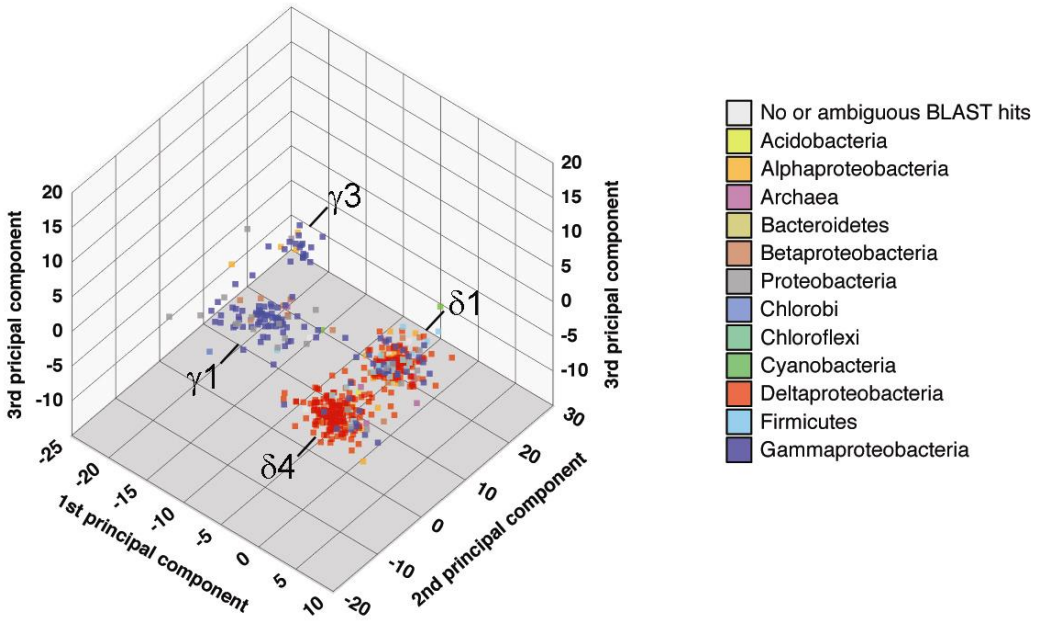


Supplementary Figure S1. Characterization of the metagenomic fosmid libraries and the 16S rRNA PCR libraries. (a) Percentage of fosmid library end reads (unassembled) with similarities to proteins of bacterial and archaeal, eukaryotic or viral origin (BLASTx, e-value 1e-3, NCBI nr). Unknown = reads with no similarity to proteins in the public databases. (b) Relative phylotype abundance in the DNA used for fosmid library-1 (left) and fosmid library-2 (right) based on 16S rRNA PCR library sequences. s = spirochete. Other includes 16S rRNA phylotypes of low abundance species.

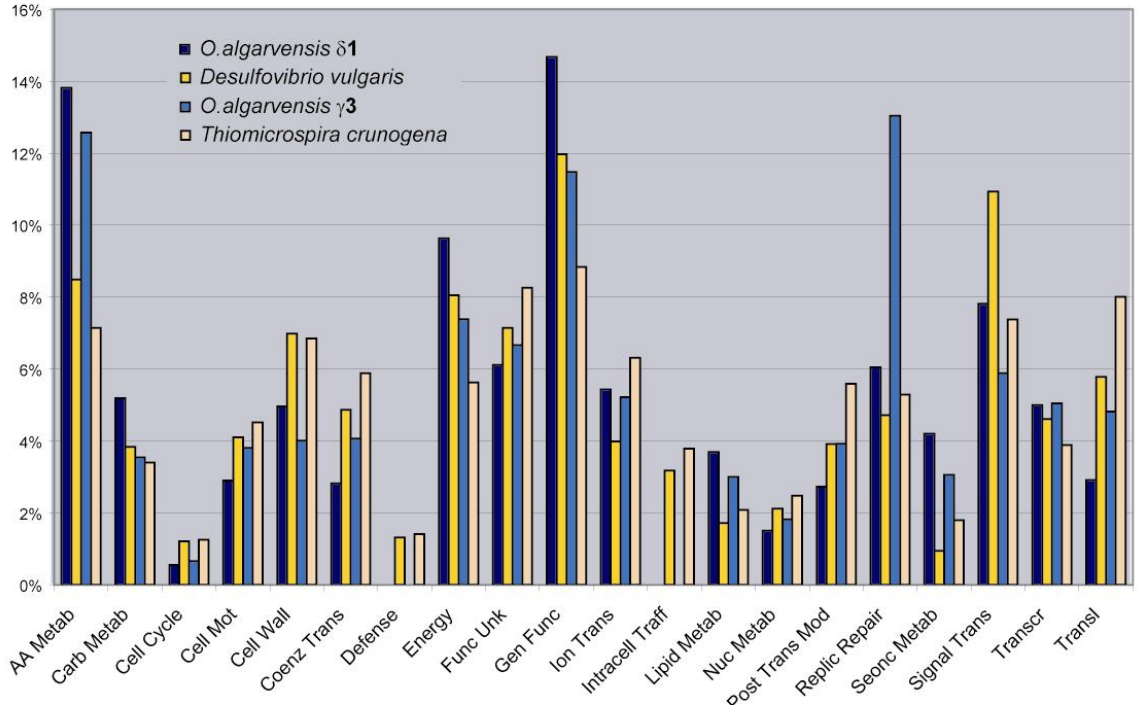

a
Filtered Scaffolds
Binned Scaffolds
Mean Scaffold Sequence Depth (Excluding Gaps)
Scaffold Length (Excluding Gaps)
b
Filtered Scaffolds
Binned Scaffolds
GC Content
Scaffold Length (Excluding Gaps)



Supplementary Figure S2. Length, sequence depth and GC content distributions of the JAZZ assembled, filtered scaffold set. (a) Net scaffold length vs. mean scaffold sequence depth (excluding gaps). (b) Net scaffold length vs. GC content. (c) Mean scaffold sequence depth (excluding gaps) vs. GC content. The filtered set comprises 2,286 scaffolds with a combined net length of 23.7 Mb. The 511 scaffolds that were binned based on nucleotide signatures are shown in red (comprising a net length of 20.1 Mb).



Supplementary Figure S3. Phylogenetic scaffold affiliations within the nucleotide signature based symbiont clusters. Visualization of the first three components of a principal component analysis (PCA), in which GC-content, net read depth, z-scores for all possible 64 trinucleotides and 256 tetranucleotides were incorporated with equal weight (z-scores calculated with TETRA and normalized by length); sequences < 5 kb are not represented. Phylogenetic affiliation of each scaffold was based on the most common phylogeny of its predicted proteins and is indicated by color.



Supplementary Figure S4. Distribution of the genes identified for δ1 and γ3 bins by broad functional category of clusters of orthologous groups of proteins (COGs) (e-value 1e-5), as compared to the complete genomes of *Desulfovibrio vulgaris* Hildenborough and *Thiomicrospira crunogena* XCL-2. Categories which show gene representations below 0.2 % are excluded. Both symbiont genome bins show a higher incidence of genes involved in amino acid as well as lipid transport and metabolism as compared to the non-symbiotic bacteria, while genes involved in translation, ribosomal structure and biogenesis show a lower relative abundance. Genes involved in replication, recombination and repair are furthermore highly represented within the γ3 bin.

4. Supplementary Tables

Supplementary Table S1. Amino acyl tRNA synthetase genes in the *O. algarvensis* symbiont bins

	δ1	δ4	γ1	γ3
Glutamyl- and glutaminyl-tRNA synthetase	1	0	3	3
Alanyl--tRNA synthetase	1	1	0	1
Phenylalanyl-tRNA synthetase	1	1	0	1
Aspartyl/asparaginyl-tRNA synthetase	1	0	0	0
Arginyl-tRNA synthetase	1	1	1	1
Isoleucyl-tRNA synthetase	1	1	0	1
Phenylalanyl-tRNA synthetase beta subunit	1	1	0	2
Histidyl-tRNA synthetase	1	0	1	2
Tyrosyl-tRNA synthetase	0	0	0	1
Seryl-tRNA synthetase	1	1	0	2
Aspartyl-tRNA synthetase	1	1	1	1
Tryptophanyl-tRNA synthetase	1	0	1	0
Cysteinyl-tRNA synthetase	1	3	0	1
Threonyl-tRNA synthetase	2	1	0	1
Prolyl-tRNA synthetase	1	1	1	1
Leucyl-tRNA synthetase	1	1	2	1
Valyl-tRNA synthetase	1	2	0	1
Glycyl-tRNA synthetase beta subunit	1	0	0	2
Glycyl-tRNA synthetase alpha subunit	1	0	0	1
Lysyl-tRNA synthetase class II	2	2	0	2
Lysyl-tRNA synthetase class I	0	0	0	0
Pseudouridine-tRNA synthetase	3	2	0	1
Methionyl-tRNA synthetase	2	2	1	0
tRNA-dihydrouridine synthetase	0	1	0	0
tRNA(IIe)-lysidine synthetase	0	0	1	0
Total	26	22	12	26

Supplementary Table S2. Repeats and polymorphic sites in the *O. algarvensis* symbiont bins

	δ1	δ4	γ1	γ3
Number of transposases	276	17	389	313
Percent of transposases	2.3	0.6	20.5	7.5
Number of integrases	30	3	28	78
CRISPR elements	yes	no	no	no
Number of polymorphic sites*	7565	144	422	1280
Frequencies of polymorphic sites** [%]	0.08	0.01	0.1	0.04

*a site was considered polymorphic, if at least two reads showed at least two different nucleotides (or gaps) in regions covered by 4-20 reads.

**averaged over all contigs assigned to a given bin.

CRISPR, clustered regularly interspaced short palindromic repeats.

Supplementary Table S3. Amino acid and vitamin biosynthesis genes in the *O. algarvensis* symbiont bins

	δ1	δ4	γ1	γ3
amino acid biosynthesis				
histidine	+	+	±	+
phenylalanine	+	+	±	+
tyrosine	+	+	±	+
leucine	+	+	±	+
isoleucine	+	+	±	+
valine	+	+	±	+
tryptophan	+	+	-	+
arginine	+	+	±	+
lysine	+	+	±	+
methionine	+	+	±	+
threonine	±	+	±	±
serine	+	+	±	+
proline	+	+	±	+
glycine	+	+	±	+
cycteine	+	+	±	±
asparagine	+	+	±	±
glutamine	+	+	+	+
alanine	+	+	+	+
aspartic acid	+	+	+	+
glutamic acid	+	+	-	+
selenocysteine	+	+	-	-
pyrrolysine	+	+	-	-
coenzymes and cofactor biosynthesis				
biotin	-	+	+	+
cobalamin	+	+	+	-
coenzyme A	+	+	+	+
riboflavin and FAD	+	+	+	+
heme	+	+	+	+
NAD	+	+	+	+
pyridoxal phosphate	+	+	+	+
thiamin	+	+	+	+
ubiquinone	+	+	-	+

+, at least 80% of the required synthesis genes are present.
±, less than 80% of the required synthesis genes are present.
-, none of the required synthesis genes are present.

Supplementary Table S4. Transposase genes in the *O. algarvensis* symbiont bins				
	δ1	δ4	γ1	γ3
IS1	-	-	48	-
IS106	-	-	3	-
IS110	-	-	8	-
IS111A	-	-	3	-
IS111A/IS1328/IS1533	-	1	-	-
IS116	-	-	5	-
IS116/IS110/IS902	-	1	-	-
IS1249	10	-	-	-
IS1479	-	-	-	1
IS1480	-	-	1	-
IS1595	1	-	-	-
IS1663	-	-	1	-
IS180	-	-	3	-
IS186	13	-	-	-
IS200	-	-	5	2
IS21	2	-	-	-
IS298	-	-	1	-
IS3	-	-	-	4
IS3/IS911	1	-	12	59
IS3231	28	-	-	-
IS4	60	-	52	77
IS5	-	-	4	-
IS630	1	-	73	-
IS641	-	-	-	27
IS642	-	-	2	-
IS643	1	-	-	-
IS653	2	-	-	-
IS66	1	-	-	9
ISChy4	2	-	-	-
ISCps6	-	-	-	1
ISDvu4	2	6	-	-
ISEcp1	-	-	13	-
ISGsu4	5	-	-	7
ISGsu6	1	-	-	-
Iso-IS1	-	-	20	-
ISPpu8	3	-	-	-
ISPsy19	-	-	1	-
ISPsy5	-	-	-	1
ISR013	5	-	-	-
ISRm	-	-	4	-
ISRm22	-	-	-	18
ISRPsy14	5	-	-	-
ISSod13	5	8	-	-
ISSpo2	-	-	1	-
ISSpo3	1	-	-	-
ISSpo8	99	-	-	-
ISxac3	1	-	-	-
Tnp	3	-	-	-
TnpA	-	-	16	2
Transposase	10	-	75	13
Transposase & inactivated derivates	-	1	19	81
Transposase, putative	14	-	19	11
Total	276	17	389	313

5. Supplementary References

1. Giere, O. & Erseus, C. Taxonomy and new bacterial symbioses of gutless marine Tubificidae (Annelida, Oligochaete) from the Island of Elba (Italy). *Org. Divers. Evol.* **2**, 289-297 (2002).
2. Lane, D. J. *16S/23S rRNA sequencing.* In: *Nucleic Acid Techniques in Bacterial Systematics* (eds. Stachebrandt, E., Goodfellow, M.) (Wiley, Chichester, New York, 1991).
3. Thompson, J. R., Marcelino, L. A. & Polz, M. F. Heteroduplexes in mixed-template amplifications: formation, consequence and elimination by 'reconditioning PCR'. *Nucleic Acids Res.* **30**, 2083-8 (2002).
4. Huber, T., Faulkner, G. & Hugenholtz, P. Bellerophon: a program to detect chimeric sequences in multiple sequence alignments. *Bioinformatics* **20**, 2317-9 (2004).
5. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403-10 (1990).
6. Ludwig, W. *et al*. ARB: a software environment for sequence data. *Nucleic Acids Res* **32**, 1363-71 (2004).
7. Chapman, J., Putnam, N., Ho, I. & Rokhsar, D. *JAZZ, a whole genome shotgun assembler.* Unpublished.
8. Aparicio, S. *et al*. Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science* **297**, 1301-10 (2002).
9. Kent, W. J. BLAT--the BLAST-like alignment tool. *Genome Res.* **12**, 656-64 (2002).
10. Karlin, S. & Burge, C. Dinucleotide relative abundance extremes: a genomic signature. *Trends Genet.* **11**, 283-90 (1995).
11. Teeling, H., Meyerdierks, A., Bauer, M., Amann, R. & Glockner, F. O. Application of tetranucleotide frequencies for the assignment of genomic fragments. *Environ. Microbiol.* **6**, 938-47 (2004).
12. Deschavanne, P. J., Giron, A., Vilain, J., Fagot, G. & Fertil, B. Genomic signature: characterization and classification of species assessed by chaos game representation of sequences. *Mol. Biol. Evol.* **16**, 1391-9 (1999).
13. Wang, Y., Hill, K., Singh, S. & Kari, L. The spectrum of genomic signatures: from dinucleotides to chaos game representation. *Gene* **346**, 173-85 (2005).
14. Huntemann, M. *MetaClust - Entwicklung eines modularen Programms zum Clustern von Metagenomfragmenten anhand verschiedener intrinsischer DNA-Signaturen*. Thesis, University Bremen (2006).
15. de Hoon, M. J., Imoto, S., Nolan, J. & Miyano, S. Open source clustering software. *Bioinformatics* **20**, 1453-4 (2004).
16. Sandberg, R. *et al*. Capturing whole-genome characteristics in short sequences using a naive Bayesian classifier. *Genome Res.* **11**, 1404-9 (2001).
17. Badger, J. H. & Olsen, G. J. CRITICA: coding region identification tool invoking comparative analysis. *Mol. Biol. Evol.* **16**, 512-24 (1999).
18. Delcher, A. L., Harmon, D., Kasif, S., White, O. & Salzberg, S. L. Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.* **27**, 4636-41 (1999).

19. Guo, F. B., Ou, H. Y. & Zhang, C. T. ZCURVE: a new system for recognizing protein-coding genes in bacterial and archaeal genomes. *Nucleic Acids Res.* **31**, 1780-9 (2003).
20. Bendtsen, J. D., Nielsen, H., von Heijne, G. & Brunak, S. Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol.* **340**, 783-95 (2004).
21. Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. L. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* **305**, 567-80 (2001).
22. Meyer, F. *et al*. GenDB--an open source genome annotation system for prokaryote genomes. *Nucleic Acids Res.* **31**, 2187-95 (2003).
23. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**, 955-64 (1997).
24. Quast, C. *MicHanThi - Design and Implementation of a System for the Prediction of Gene Functions in Genome Annotation Projects.* Thesis, University Bremen (2006).
25. Tatusov, R. L. *et al*. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4**, 41 (2003).
26. Edgar, R. C. & Myers, E. W. PILER: identification and classification of genomic repeats. *Bioinformatics* **21 Suppl 1**, i152-i158 (2005).
27. Markowitz, V. M. *et al*. The integrated microbial genomes (IMG) system. *Nucleic Acids Res.* **34**, D344-8 (2006).
28. Yang, C. C., Packman, L. C. & Scrutton, N. S. The primary structure of *Hyphomicrobium* X dimethylamine dehydrogenase. Relationship to trimethylamine dehydrogenase and implications for substrate recognition. *Eur. J. Biochem.* **232**, 264-71 (1995).
29. Paul, L., Ferguson, D. J., Jr. & Krzycki, J. A. The trimethylamine methyltransferase gene and multiple dimethylamine methyltransferase genes of *Methanosarcina barkeri* contain in-frame and read-through amber codons. *J. Bacteriol.* **182**, 2520-9 (2000).
30. Naidu, D. & Ragsdale, S. W. Characterization of a three-component vanillate O-demethylase from *Moorella thermoacetica*. *J. Bacteriol.* **183**, 3276-81 (2001).