A single-cell genome for *Thiovulum* sp.

Ian P.G. Marshall¹*, Paul C. Blainey²*, Alfred M. Spormann¹,³,#, Stephen R. Quake²,⁴,#

1) Department of Civil and Environmental Engineering
2) Department of Bioengineering
3) Department of Chemical Engineering
Stanford University, Stanford, CA 94305
4) Howard Hughes Medical Institute, Stanford, CA 94305

* These authors contributed equally to this work
+ Present Address: Broad Institute, Cambridge, MA
# Corresponding Authors

Corresponding Author E-mail: spormann@stanford.edu (A.M. Spormann), quake@stanford.edu (S.R. Quake)

Running Title: *Thiovulum* sp. Single-Cell Genome
Abstract

We determined a significant fraction of the genome sequence of a representative of *Thiovulum*, the uncultivated genus of colorless sulfur epsilon-Proteobacteria, by analyzing the genome sequences of four individual cells collected from phototrophic mats from Elkhorn Slough, California. These cells were isolated utilizing a microfluidic/laser-tweezing system, and their genomes were amplified by multiple-displacement amplification prior to sequencing. *Thiovulum* is a gradient bacterium found at oxic/anoxic marine interfaces noted for its distinctive morphology and rapid swimming motility. The genomic sequences of the four individual cells were assembled into a composite genome consisting of 221 contigs covering 2.083 megabases encoding 2,162 genes. This single-cell genome represents a genomic view of the physiological capabilities of isolated *Thiovulum* cells. *Thiovulum* is the second fastest bacterium ever observed, swimming at 615 μm/s, and this genome shows that this rapid swimming motility is a result of a standard flagellar machinery that has been extensively characterized in other bacteria. This suggests that standard flagella are capable of propelling bacterial cells at speeds much faster than typically thought. Analysis of the genome suggests that naturally occurring *Thiovulum* populations are more diverse than previously recognized, and that studies performed in the past probably address a wide range of unrecognized genotypic and phenotypic diversity of *Thiovulum*. The genome presented in this study provides a basis for future isolation-independent studies of *Thiovulum*, where single-cell and metagenomic tools can be used to differentiate between different *Thiovulum* genotypes.
**Introduction**

*Thiovulum* is a genus of as yet uncultivated colorless sulfur bacteria thought to conserve energy through the aerobic oxidation of reduced sulfur compounds at the oxic/sulfidic interface in marine environments (47). The *Thiovulum* genus has primarily been defined observationally by its distinct large, egg-shaped cell morphology (Figure 1) and unusual rapid swimming motility (60). A species within the genus *Thiovulum* was first described as *Volvox punctum* (40), then renamed *Monas mülleri* (60). Two similar species were described in 1913 and given the names *Thiovulum majus* and *Thiovulum minus* (17, 47). *Monas mülleri* was then renamed *Thiovulum mülleri* after its similarities to Hinze’s new species were recognized (31, 60). These *Thiovulum* species were classified according to their size and motility, with *T. mülleri* 4.9-10.2µm in diameter and moving in an undirected zig-zag manner, while *T. minus* and *T. majus* cells revealed a more unidirectional motility and were 7.2-9µm across and 9-17µm across respectively (5, 17, 40, 63).

Despite many isolation attempts since its discovery (28, 59-61), no *Thiovulum* pure culture has been reported so far. In spite of evading isolation in pure culture, *Thiovulum*’s distinctive morphology has enabled a number of important ecological and physiological studies in enrichments and laboratory microcosms. *Thiovulum* has been described as chemolithoautotrophic (61) and microaerophilic (13, 22). It is one of the fastest-swimming bacteria ever observed, swimming at 615 µm/s (14). Large numbers of *Thiovulum* cells naturally form conspicuous veils at marine oxic-anoxic interfaces (17). These veils consist of a polymeric matrix where *Thiovulum* are reversibly attached and can rapidly traverse within a 50-100 µm wide, diffusion controlled, overlapping sulfide/O2 gradient (23). This steep sulfide/O2 gradient is self-generated as the calculated mean residence time of both substrates within this overlapping layer is 0.1 to 0.6 seconds. *Thiovulum* cells demonstrated chemotaxis towards O2 in seeking out these anoxic-oxic interfaces (56), and veil-associated cells were postulated to use their flagella to generate advective water flow within the veil.
Thiovulum is the only large-celled (cell diameter >5µm) colorless sulfur bacterium classified as an epsilon-Proteobacterium, while other large colorless sulfur bacteria such as Beggiatoa, Thioploca, and Thiomargarita are classified within a single cluster of the gamma-Proteobacteria (Figure 2) (24, 51). Thiovulum’s large cells with their distinctive morphology, 226-year history of observations, and uncultivated status make this microbe an ideal candidate for applying the recently developed method of single-cell genome sequencing using a microfluidic/laser-tweezer system (6, 25, 35, 66). In this method a mixture of microorganisms from an environmental sample (or an enrichment culture) is introduced into a poly(dimethylsiloxane) (PDMS) microfluidic device. Single cells are selected upon microscopic inspection and then guided with a laser trap into separate chambers, where each sorted cell is lysed individually, and then subjected to whole-genome amplification by multiple displacement amplification (MDA) (30). Amplified genomic DNA from each cell is then separately recovered from the chip and sequenced.

In 2007, students of the Hopkins Microbiology Course in Pacific Grove, California, USA (http://hmc.stanford.edu) observed Thiovulum veils in phototrophic microbial mats from Elkhorn Slough, California. In these phototrophic mats, the circadian changes in flux of light result in well-oxygenated surface layers during the day and anoxic, fermenting cyanobacterial layers during the night (9, 62), thus pushing the oxic-anoxic transition zone of Thiovulum’s habitat into the mat during the day, and out of the mat during night, respectively. In 2009 we set out to sequence the genome of cells from a Thiovulum population recovered from these mats in order to reveal the molecular basis of its known ecophysiological characteristics and to form hypotheses about as-yet unknown physiological capabilities based on genomic data and comparisons with its closest sequenced relatives.

Materials and Methods

Thiovulum Collection and Enrichment
Phototrophic microbial mats were collected from Elkhorn Slough in California in fall 2009 as previously described (9, 62). A 10 cm x 10 cm section of mat overlain with approximately 2 cm of seawater was incubated at room temperature overnight in the laboratory. By morning a white cloudy veil had formed. This veil was confirmed by phase contrast microscopy to contain cells matching the morphological and motile characteristics detailed in previous descriptions of *Thiovulum*.

**Microfluidic Sorting and Whole-Genome Amplification**
Sorting in a microfluidic laser-tweezer device and whole-genome amplification were carried out as previously described (6).

**PCR and Quantitative PCR**
16S rRNA genes were amplified from MDA-amplified genomic DNA using primers GM3F (5’- AGAGTTTGATCMTGGC -3’) (39) and Uni1392R (5’- ACGGGCGGTGTGTRC -3’) (29). PCR was performed with 0.5µM of each primer in 50µL volumes using DreamTaq Green MasterMix (Fermentas, Vilnius, Lithuania) with a 5-minute 95ºC initial denaturation step, followed by 25 cycles of 1 minute at 95ºC, 1 minute annealing at 48ºC, then 1 minute extension at 72ºC. PCR products were cloned using the TOPO TA cloning kit (Invitrogen, Carlsbad, CA, USA) and subsequently sequenced.

Quantitative PCR primers were designed using Geneious (Biomatters, Auckland, New Zealand) based on the 16S rRNA gene sequences obtained from the single cells. The forward primer was Thio_121F (5’- GGATAGTTACATGCCCCTTGGAG-3’), the reverse primer was Thio_221R (5’- TAGGAGATAGCCCCAATCCCTTG-3’). Total bacterial DNA was quantified using universal bacterial primers Eub341F (5’-CTGCTGCCTCCCGGCTGAG-3’) and Eub534R (5’-ATTACCGCGGCCTGCTGCG-3’) (39). The fraction of each MDA product that was from *Thiovulum* was estimated through qPCR using the method previously described (4), by quantifying total bacterial 16S rRNA gene copies and *Thiovulum* 16S rRNA gene copies with a plasmid standard created using the 16S rRNA gene cloning protocol described above.
**Genome Sequencing and Assembly**

454 shotgun libraries were prepared from the amplified genomes of each individual cell by a variation of the manufacturer's protocol (7). 454 sequencing was carried out according to the manufacturer's instructions for Titanium chemistry (Roche, Branford, Connecticut, USA) to obtain 192 Mb raw sequence data. The data were filtered for read quality using the MOTHUR package (50) according to the following criteria: no homopolymers greater than 10, no ambiguous bases, average quality score greater than 26. This yielded reads totaling 174 Mb. Individual assemblies were generated with the Newbler assembler (Roche, default parameters, except for specifying an increased expected read depth in excess of the actual value) and found to share large regions of homology exceeding 99% by blastn, thus, we took a coassembly approach, pooling the data from multiple single cells to generate a consensus coassembly. The raw reads were co-assembled in an iterative ‘jackknife’ procedure using Newbler (v2.6, default parameters, except for specifying an increased expected read depth in excess of the actual value) to exclude the chimeric sequences generated by the MDA process without the aid of a reference sequence (for details see the supplemental methods section). At this stage, the ~2 Mb coassembly was contained in approximately 240 contigs. Seeking to further improve the co-assembly, we implemented two high-throughput gap-closing strategies, one based on PCR amplification of gap regions, and the other taking advantage of a third-generation sequencing platform (for details see the supplemental methods section).

**GC Content Bias Analysis**

To check the correlation between GC content and coverage, the GC content for every assembled base more than 50 bases from a contig end was calculated based on a 100 base pair window. Coverage was set equal to the “Total Depth” statistic generated by the Newbler assembler. The correlation was tested using the cor.test() function in R version 2.13.1 with the Spearman method and other parameters set to the defaults. GC bias at contig ends was found by comparing the GC content of the
first and last 100 base pairs of each contig to the GC content of the entire contig for
all contigs 1000 base pairs or greater in length. The binomial test (binom.test() in R
version 2.13.1) was used to determine whether the results of this analysis differed
from results expected through random changes at contig ends (25% high GC ends,
25% low GC ends, and 50% mixed low/high GC ends).

Genome Purity Analysis
Nucleotide sequences of genes expected to be present in single copies (44) from the
assembled four-cell genome were used with BLAST (2) as query sequences against
four databases, each containing contigs from one of the four individual-cell
assemblies. Results of this BLAST were used to determine whether the genomes for
all of the four cells were non-identical.

Genome Annotation
Assembled contigs were submitted to the Integrated Microbial Genomes database
annotation pipeline (36) in late 2011 (IMG version 3.4). Some computationally
assigned annotations were manually changed based on inspection of evidence for
the IMG-assigned annotations, orthologs in related genomes and gene
neighborhoods. Pathways were predicted using the PathoLogic tool in Pathway
Tools version 15.5.

Sequence Submission
This Whole Genome Shotgun project has been deposited at DDBJ/EMBL/GenBank
under the accession AKKQ00000000. The version described in this paper is the first
version, AKKQ01000000.

Phylogenetic Analysis
A 16S rRNA gene tree was constructed by uploading the full *Thiovulum* 16S rRNA
gene sequence from the assembled genome to the online Silva SINA aligner
(http://www.arb-silva.de/) (43) then downloading the aligned *Thiovulum* sp. ES
sequence along with the closest relatives for which a genome exists, some
representative epsilon-Proteobacteria and other colorless sulfur bacteria from within the gamma-Proteobacteria. This alignment was used to construct a 100X bootstrapped phylogenetic tree in Geneious (Biomatters) using the PhyML tool with default settings (16).

A coding sequence phylogeny was created by taking single-copy genes from Thiovulum sp. ES and related genomes (Supplemental Table ST1), performing translation alignments of the nucleic acid sequences using the MUSCLE alignment feature within Geneious, and then concatenating those alignments together. The concatenated alignments were used as input for PhyML to generate a 1000X bootstrapped tree with other parameters at default settings. Campylobacter jejuni subsp. jejuni M1 was used as the outgroup genome.

Ortholog Analysis
Orthologs were determined by constructing a BLAST (2) database containing protein sequences from the Thiovulum genome plus five related genomes obtained from IMG version 3.4. All 6 genomes were used as query sequences against this database. BLAST hits with an e-value of less than $10^{-5}$ and 35% or greater sequence identity were considered homologs, with only the lowest e-value hit counting where multiple hits met this criteria for a single query sequence. These homolog pairs were then considered ortholog pairs if the same criteria applied for the reverse BLAST with query and subject sequences exchanged.

Results and Discussion
Genome Completeness
Six single Thiovulum cells from Elkhorn Slough mat material were sorted into single chambers on a microfluidic device, and whole-genome amplification was performed subsequently on the lysed cells. 16S rRNA gene sequences obtained from the sorted Thiovulum DNA by PCR were identical to each other. Of the six Thiovulum whole-genome amplification products obtained, four were estimated to have 100% of their 16S rRNA gene copies representing Thiovulum sp. by qPCR (data not shown). These
four whole-genome amplification products were thus thought to contain if at all minimal non-Thiovulum DNA contamination. A genome was reconstructed by assembling reads sequenced from the amplification of those four cells. Single-copy genes in the four-cell assembly almost exclusively shared 100% identity with each of the four single-cell assemblies, with the exceptions either being single-base insertions (and thus more likely to be a sequencing error than a frameshift mutation) or single nucleotide polymorphisms within 7 bases of the end of a contig (and thus probably sequencing error) (Supplemental Table ST2). The cells used are pictured in Supplemental Figure SF1. For the purpose of convenience we have named the strain represented by this assembled genome Thiovulum sp. ES, with ES being an abbreviation for Elkhorn Slough.

The genome was assembled into 221 contigs ranging in size from 100 bp to 81,617 bp. 82% of the contigs were longer than 500 bp, 50% were longer than 2,592 bp, and 28% longer than 10,000 bp. The N50 (37) was calculated as 28,163 bp. GC content was investigated as a potential source of bias in MDA and sequencing, and found to be slightly positively correlated with coverage (Spearman’s rho = 0.08, p < 2.2e-16). It appears that this bias affected where contigs were terminated, as contigs were more likely to have two 100bp ends of low GC content (42% of contigs 1000bp or greater, p=5.054x10^-6) than two 100bp ends of high GC content (12% of contigs 1000bp or greater, p=9.091x10^-8). A similar positive GC-coverage correlation has been observed in the past while using the RepliG kit for MDA (65).

Despite being unfinished, the assembled genome appears to provide a reasonable representation of Thiovulum sp. ES gene content, with a complete ribosomal RNA gene operon, 36 tRNA genes enabling translational insertion of all amino acids from all codons (Supplemental Table ST3), and genes encoding all proteins necessary for replication, transcription, and translation with the exception of two ribosomal proteins (S14 and L33) considered universal in bacteria (67) (Supplemental Table ST4). The total number of genes identified was 2,162 and the total genome length 2.083 Mbp, which are comparable to its closest sequenced relatives from the genera.
Sulfurimonas, Sulfurovum, and Sulfuricurvum (Table 1). Although it cannot be ruled out that genes may be missing from the genome in its present partially finished state, we believe that coverage is sufficient to draw some significant yet cautious conclusions about the microorganism’s genetic inventory and resultant physiology.

**Genome Phylogenetic Identification**

Prior to this study there were only two 16S rRNA gene sequences from a *Thiovulum* enrichment in Genbank (accession numbers M92323.1 and M92324.1) in the form of two non-overlapping fragments from the 5’ and 3’ end of the gene (49), as well as a complete 5S rRNA gene sequence from the same source (accession M35570.1). These sequences were derived from *Thiovulum* cells enriched from sediment from Eel Pond in Woods Hole, Massachusetts, USA (David Stahl, personal communication 2012). These cells were about 9-10 µm in size (55). Interestingly, the 16S rRNA gene in the *Thiovulum* sp. ES genome (identical to the 16S rRNA sequences obtained from the MDA products by PCR) was 88.2% and 88.7% identical to the two 16S rRNA sequences in Genbank, and the 5S rRNA gene sequence was 81.9% identical. The top BLAST hit for the full gene sequence was from an uncultured bacterium (accession EF467529.1) with 87.2% identity to the *Thiovulum* sp. ES 16S rRNA gene. Notably, these identities are well below the typical 94.9%±0.4 minimum identity observed between members of a genus and the type strain of a genus (64). This suggests that in nature there may in fact be multiple genera within the group defined phenotypically as *Thiovulum*. However, the genus *Thiovulum* has been traditionally defined phenotypically rather than based on sequence identity, as there exists a paucity of available *Thiovulum* 16S rRNA gene sequences. For these reasons we have chosen to call the microorganism whose genome was sequenced in this study a member of the genus *Thiovulum*. The phylogenetic positioning of the genus *Thiovulum* as a relative of other colorless sulfur bacteria within the epsilon-Proteobacteria was confirmed by its position in phylogenetic trees based on 16S rRNA genes (Figure 2) and a concatenated alignment of 35 single-copy genes highly conserved amongst bacteria (Figure 3).
Catabolic Gene Content

Physiological studies of *Thiovulum* have revealed that it is an aerobic colorless sulfur bacterium, capable of conserving energy by oxidizing H$_2$S to elemental sulfur with molecular oxygen as terminal electron acceptor (11, 23, 61). This observation is supported by evidence from the genome, with genes encoding sulfide:quinone reductase (*sqr*), ubiquinol cytochrome bc1 complex and cytochrome c oxidase (Figure 4, Supplemental Table ST4). *Sqr* has been shown to couple the oxidation of sulfide to sulfur to the reduction of quinones (53). There is also genomic evidence that *Thiovulum* sp. may be able to use formate as an electron donor, with genes encoding a cytoplasmic formate dehydrogenase and cytochrome c553 (Figure 4, Supplemental Table ST4) that may enable the oxidation of formate in a similar manner as in *Desulfovibrio vulgaris* (52).

Genes encoding enzymes for the reduction of various terminal electron acceptors were also identified. For aerobic respiration, a cytochrome c oxidase was identified (Figure 4, Supplemental Table ST4), which is of the high-affinity cbb3 type, adapted for low O$_2$ concentrations (41, 42). This matches earlier observations that *Thiovulum* cells are typically found in microoxic (0-10µM) environments (13, 22, 23). The genome also contains a gene predicted to encode a cytochrome c peroxidase and desulfoferredoxin, possibly involved in oxidative stress response.

Anaerobic growth using nitrate as an electron acceptor is a common trait amongst colorless sulfur bacteria (48) especially amongst the epsilon-*Proteobacteria* (15, 18, 27). Other large colorless sulfur bacteria including *Beggiatoa*, *Thiomargarita*, and *Thioploca* profit from their large cell size by storing up to 0.3 M nitrate in vacuoles comprising the majority of the cell volume and then respiring that nitrate under anoxic conditions (24). Moreover, the epsilon-*Proteobacterium Wolinella succinogenes* has been shown to grow anaerobically by coupling the oxidation of formate to the reduction of elemental sulfur (21, 34). In the absence of a pure culture, the capacity for anaerobic growth by *Thiovulum* has been more difficult to
evaluate. Wirsen and Jannasch (61) were unsuccessful in maintaining *Thiovulum*
enrichments under anoxic conditions, but hypothesized that *Thiovulum* could
persist in a dormant state in anoxic environments. On the other hand, using an O\textsubscript{2}
microsensor Jørgensen and Revsbech (23) observed the majority of cells in
*Thiovulum* veils to be on the anoxic side of the oxic-anoxic interface, which is
possibly evidence of anaerobic energy conservation. Fenchel (13) refuted this
finding, with observations of *Thiovulum* veils showing no cells present on the anoxic
side of an oxic-anoxic interface.

The *Thiovulum* sp. ES genome contains two genes encoding the \(\alpha\)- and \(\beta\)-subunits of
periplasmic nitrate reductase (NarGH), as well as a membrane-associated
polysulfide reductase (NrfD)(Figure 2, Supplemental Table ST4), suggesting that
this particular population of *Thiovulum* may be capable of anaerobic growth either
through the oxidation of reduced sulfur compounds coupled to the reduction of
nitrate or the oxidation of formate coupled to the reduction of sulfur. The above
mentioned contrasting conclusions on the capacity of *Thiovulum* for anaerobic
growth over the last several decades could be explained in light of the genomic
insights if multiple species of *Thiovulum* existed, including some facultative
anaerobic species and some obligate aerobic species. A similarly varying capacity for
denitrification has been observed within the genus *Sulfurimonas*, with species that
are facultative anaerobes (18), obligate aerobes (20), and obligate anaerobes (15).
An alternative explanation for the presence of genes encoding a nitrate reductase in
the *Thiovulum* sp. ES genome is that the nitrate reductases are involved in anabolic
nitrate reduction.

In contrast to nitrate reductase, no nitrite reductase was identified, either indicating
that it is present but unsequenced or that *Thiovulum* sp. ES is an incomplete
denitrifying bacterium. Also absent were homologs of the hydroxylamine oxidase
and reductase hypothesized to play a role in nitrite reduction in the
epsilonproteobacterium *Nautilia profundicola* (10). Other incomplete-denitrifying
colorless sulfur bacteria have been identified (27, 48) suggesting that the latter possibility may be true.

One puzzling finding from the genome was that no gene encoding an enzyme to further oxidize elemental sulfur was identified. While genes encoding enzymes for sulfite oxidation via adenosylphosphosulfate as intermediate (26), including sulfate adenylyltransferase (ThvES_00004190-4200), were found to be present, these enzymes are also required for sulfate activation for sulfonation reactions or sulfate assimilation, and are thus not evidence of the capacity to oxidize elemental sulfur to sulfate. Since elemental sulfur forms intracellularly in *Thiovulum* cells (12, 17, 28) and no obvious set of enzymes for the complete oxidation of elemental sulfur was found, *Thiovulum* cells somehow need to escape ever-accumulating intracellular sulfur. One possible mechanism is that *Thiovulum* has to oscillate between an aerobic mode of energy conservation where elemental sulfur accumulates in the cell and an anaerobic mode of energy conservation where intracellular sulfur serves as an electron acceptor, perhaps with formate acting as electron donor or via anaerobic sulfur disproportionation. This explanation would be consistent with *Thiovulum*’s ecophysiology, as microbial mats go through diel oxic/anoxic fluctuations, where fermentation products including formate are formed during the night in the anoxic cyanobacterial layer (9). This would also explain why attempts at continuous aerobic or anaerobic *Thiovulum* cultivation have failed thus far, as obligatory oxic/anoxic cycling (possibly providing a second electron acceptor, such as formate, during the anoxic phase) would be critical for sustained cultivation. A second possible explanation is that the genes encoding an elemental sulfur oxidation pathway await discovery or further sequencing of *Thiovulum* sp. ES. Both explanations of how elemental sulfur leaves the cell are intriguing possibilities that warrant further investigation.

**Anabolic Gene Content**

*Thiovulum* has been shown to grow autotrophically (61). Other autotrophic colorless sulfur bacteria, including epsilon-Proteobacteria such as *Sulfurimonas*
denitrificans and Candidatus Arcobacter sulfidicus, fix carbon dioxide via the reverse tricarboxylic acid (TCA) cycle (19). We have identified genes encoding all enzymes necessary for the reverse TCA cycle, including ATP-citrate lyase, fumarate reductase, and 2-oxoglutarate synthase, specific to the reductive (reverse) rather than the oxidative TCA cycle (Figure 5, Supplemental Table ST4). No gene encoding citrate synthase has been identified, suggesting that the TCA cycle of the Thiovulum sp. ES is only used for CO2 fixation and anabolism. An exclusively reductive TCA cycle is consistent with observations by Wirsen & Jannasch (61), who did not observe any significant uptake of acetate or other organic compounds.

It is, however, unclear how Thiovulum sp. ES obtains the reduced ferredoxin the microbe needs as reductant for the 2-oxoglutarate synthase reaction in reverse TCA cycle as well as for the pyruvate synthase reaction. The oxidation of sulfide to elemental sulfur has a redox potential of only -270 mV under standard state conditions (58), which is insufficient for the reduction of ferredoxin at ~-500 mV. One possible explanation is that electrons are bifurcated via the flavin-containing EtfBC complex in a similar process to that which occurs in Clostridium kluyveri (8). However, no genes homologous to the C. kluyveri genes encoding EtfBC or matching the ‘ETF’ pfam family were identified in the Thiovulum sp. ES genome. Indeed, these genes are also absent from the genome of Sulfurimonas denitrificans, the first colorless sulfur bacterium in the epsilon-Proteobacteria shown to fix CO2 through the reverse TCA cycle (19). An inspection of pfam (http://pfam.sanger.ac.uk) reveals that of the 665 proteobacterial species shown to have genes encoding proteins in the ‘ETF’ family, these are all in the alpha-, beta-, gamma-, and delta-Proteobacteria, with no ETF proteins in epsilon-proteobacterial genomes. It appears that there may be an as yet undiscovered mechanism in Thiovulum sp. ES and other epsilon-Proteobacteria that enables the reduction of ferredoxin coupled to sulfide oxidation.

In addition to genes encoding CO2 fixation, genes encoding enzymes necessary for the biosynthesis of amino acids (tryptophan, leucine, isoleucine, histidine, alanine, tyrosine, lysine, glutamate, cysteine, methionine, arginine, threonine, proline,
glycine), purine and pyrimidine nucleotides, peptidoglycan, lipids, and various cofactors (tetrahydrofolate, NAD, flavins, coenzyme A, thiamin) and glucose-6-phosphate (gluconeogenesis) were identified (Supplemental Table ST5, pathway hole report is provided as Supplemental Table ST6). This relatively complete repertoire of biosynthetic capabilities supports the hypothesis that *Thiovulum* is capable of autotrophic growth, as well as that this genome is reasonably complete. No gene encoding nitrogenase was identified in the *Thiovulum* sp. ES genome, but a gene encoding the ammonium transporter (AmtB) was identified adjacent to a gene encoding the nitrogen regulatory protein PII (Supplemental Table ST4), implicated in the regulation of the ammonium transporter in *Azospirillum brasilense* and *Synechococcus* sp (3). This suggests that *Thiovulum* sp. ES is unable to fix N₂, but is capable of importing ammonia from the environment.

**Chemotaxis and Motility**

*Thiovulum* has been noted to be highly motile (14, 23) and chemotactic (13, 56). Flagella have been observed via staining (17) and electron microscopy (11, 46). The *Thiovulum* sp. ES genome contains all genes necessary for flagellum biosynthesis (Supplemental Table ST4), including two genes encoding FliC (flagellin) - these correspond to *flaA* and *flaB* in *Helicobacter pylori* (33). However, sequence identity with the *H. pylori* orthologs is not sufficiently high to determine which gene encodes FlaA and which encodes FlaB. There are also two paralogs each of *flgE* and *flgG*, which encode structural proteins of the flagellar hook and proximal rod, respectively. One interesting property of the *Thiovulum* strain ES genes encoding the flagellar motor proteins, MotA and MotB, is that in the related *Sulfurimonas* and *Sulfuricurvum* genomes these genes are positioned adjacent to one another suggesting they are part of a single operon, while in *Thiovulum* sp. ES *motA* and *motB* appear to be genetically unlinked and are present on different contigs (Supplemental Figure SF2). Apart from this anomaly, at the genetic level the *Thiovulum* flagellar machinery appears to be similar to well-characterized flagella in other bacteria, including *H. pylori* and *Escherichia coli*.
The genome also contains genes encoding the core chemotaxis proteins CheA, CheW, and CheY plus 13 genes encoding methyl-accepting chemotaxis proteins (MCPs) (Supplemental Table ST4). Of the auxiliary chemotaxis genes that have been defined (63), *Thiovulum* sp. ES was found to possess genes coding for CheC, CheD, CheV and the *Helicobacter pylori*-type CheZ (32). However, the *Thiovulum* sp. ES genome does not contain clear homologs for CheX, or CheB and CheR. No bacterial genome containing cheCD but missing both cheB and cheR has been identified to date (63), suggesting that genes enabling methylation-dependent chemotaxis have not yet been sequenced in this unfinished genome. On the other hand, methylation-independent chemotaxis is not unprecedented in knockout strains (5). As studies in *Bacillus subtilis* have shown CheC, CheD, and CheV are involved in sensory adaptation to a wide range of substrate concentrations by a feedback mechanism from phosphorylated CheY to the MCP (CheC and CheD) (38, 45), or by modulating the efficiency of CheY phosphorylation by CheA (CheV) (1). CheZ accelerates the dephosphorylation of CheY, enabling a rapid response to changing environmental conditions (32).

Thirteen MCPs can be reasonably considered to constitute a complete suite of MCP proteins given the number of MCPs found in genomes of closely-related sequenced organisms (15). Of the 13 MCPs, putative sensory domains could be identified in five, with nitrate/nitrite sensory domains (54) in three, a CACHE domain (associated with small-molecule sensing) in one and a PAS domain (associated with O₂ sensing as well as other substrates) in another. O₂ is the only chemotaxis substrate reported for *Thiovulum* sp. to date, but the genes thought to encode MCPs in the genome suggest that *Thiovulum* may be capable of chemotaxis towards substrates other than O₂, possibly including nitrate/nitrite and small organic molecules.
Comparisons between the *Thiovulum* genome and its five closest sequenced relatives revealed the 16S rRNA gene identity, as described above, and the percentage of orthologs shared between each genome pair (Table 1). *Thiovulum* is more distantly related from all of its sequenced close relatives than any of its closest sequenced relatives are from each other (Figure 2, Table 2). This means that although *Thiovulum* sp. ES is phylogenetically an epsilon-*Proteobacterium*, it is only distantly related to epsilon-*Proteobacteria* with sequenced genomes, with less than half of its genes having orthologs in the most closely related genomes.

**Conclusions**

The sequencing and annotation of a genome from four single *Thiovulum* cells revealed genes encoding many traits previously observed in *Thiovulum* like chemolithoautotrophy, motility and chemotaxis, and unusual cell shape. However, the 88% sequence identity between the genome's 16S rRNA gene and the only other *Thiovulum* 16S rRNA gene existing in public databases, the presence of genes encoding for nitrate and polysulfide reduction while the literature record is ambivalent about *Thiovulum*’s capacity to grow anaerobically, and previous observations of different *Thiovulum* populations having different cell sizes all suggest that the *Thiovulum* species are more genotypically and phenotypically diverse than previously thought. We believe that future use of a combination of metagenomic sequencing and single-cell sequencing from *Thiovulum* populations containing different cell sizes and from different geographic locations will lead to a better understanding to the true extent of *Thiovulum*’s genotypic diversity.

**Acknowledgements**

We thank the students, faculty, and teaching assistants of the Hopkins Microbiology Course from 2007 to 2009 for initial experiments to single-cell sort *Thiovulum* cells from Elkhorn Slough as part of their course project. We also thank Drs. Dagmar Woebken and Luke C. Burow for providing microbial mat material, to Lolita Penland for 454 sequencing, Veronica Brand for microfluidic PCR, and Dr. Norma F. Neff and Gary Mantalas for PacBio library preparation and sequencing.
IPGM was supported by a Bio-X graduate fellowship. The Hopkins Microbiology Course was supported by funds from NSF, DOE (BER), the Gordon and Betty Moore Foundation, and the Stanford Bio-X program.

References


40. Müller OF. 1786. Animalcula infusoria fluvia tilia et marina.


60. Warming E. 1876. Om nogle ved Danmarks Kyster levende Bakterier.


novel cyanobacterial group as active diazotrophs in a coastal microbial mat using NanoSIMS analysis. ISME J.


Figure Legends

Figure 1 – Phase contrast micrographs of Thiovulum sp. cells from an enrichment from Elkhorn Slough, California. A) shows a typical Thiovulum cell from this environment, B) shows a cell undergoing division. Note the highly refractile polar sulfur granules in both cells. Scale bars indicate 5 μm. Micrographs are courtesy of Marie B. Lund.

Figure 2 - 16S rRNA phylogenetic tree of Thiovulum sp. ES, other colorless sulfur bacteria, and other epsilon-Proteobacteria. Maximum likelihood tree bootstrapped 100X, branch labels show number of trees with displayed branching pattern.

Figure 3 - Phylogenetic tree of Thiovulum sp. ES and related epsilon-Proteobacteria based on concatenated alignment of 35 single-copy genes (Supplemental Table ST1). Maximum likelihood tree bootstrapped 1000X, branch labels show number of trees with displayed branching pattern.

Figure 4 – Genomically predicted catabolic electron flow network in Thiovulum sp. ES. Locus tags are sulfide: quinone oxidoreductase: ThvES_00003240; formate dehydrogenase: ThvES_00014980; ubiquinol cytochrome bc1 complex: ThvES_00002990-3010; cytochrome c553: ThvES_00003480, ThvES_00009980, ThvES_00017420; nitrate reductase: ThvES_00000060-70; cytochrome c peroxidase: ThvES_00008770, ThvES_00012320; cbb3-type cytochrome c oxidase: ThvES_00014560-14590; polysulfide reductase: ThvES_0006260.

Figure 5 – Genomically identified enzymes for the operation of a complete reverse TCA cycle and pyruvate synthase for autotrophic CO2 fixation in Thiovulum sp. ES. The cofactors are indicated as predicted based on genome annotation and are not experimentally verified. Locus tags are isocitrate dehydrogenase: ThvES_00010990; aconitate hydratase: ThvES_00009840; ATP-citrate lyase: ThvES_00016940;
Thiovulum sp. ES
506
1000
1000
1000
1000
to
outgroup
Helicobacter pylori 26695
(637000132)
Sulfurovum sp. NBC37-1
(640753056)
Sulfurimonas denitrificans DSM 1251
(637000326)
Sulfurimonas autotrophica OK10, DSM 16294
(648028058)
Sulfuricurvum kujiense YK-1, DSM 16994
(649633097)
Sulfurimonas gotlandica GD 1
(64753121)
<table>
<thead>
<tr>
<th></th>
<th>Thiovulum sp. ES</th>
<th>Sulfurimonas autotrophica OK10</th>
<th>Sulfurimonas denitrificans</th>
<th>Sulfuricurvum kujiense YK-1</th>
<th>Sulfurimonas gotlandica GD1</th>
<th>Sulfurovum sp. NBC37-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Genome Size</td>
<td>2.083 Megabases</td>
<td>2.153 Megabases</td>
<td>2.201 Megabases</td>
<td>2.819 Megabases</td>
<td>2.946 Megabases</td>
<td>2.562 Megabases</td>
</tr>
<tr>
<td>Contigs</td>
<td>221</td>
<td>1</td>
<td>1</td>
<td>5</td>
<td>12</td>
<td>1</td>
</tr>
<tr>
<td>% Coding DNA</td>
<td>91.83%</td>
<td>94.88%</td>
<td>93.72%</td>
<td>93.03%</td>
<td>90.38%</td>
<td>90.70%</td>
</tr>
<tr>
<td>GC Content</td>
<td>33%</td>
<td>35%</td>
<td>34%</td>
<td>45%</td>
<td>33.59%</td>
<td>43.87%</td>
</tr>
<tr>
<td>Genes total number</td>
<td>2162</td>
<td>2220</td>
<td>2164</td>
<td>2879</td>
<td>2804</td>
<td>2525</td>
</tr>
<tr>
<td>Protein coding genes</td>
<td>2120</td>
<td>2165</td>
<td>2104</td>
<td>2818</td>
<td>2745</td>
<td>2465</td>
</tr>
<tr>
<td>Orthologs shared with Thiovulum</td>
<td>2120</td>
<td>1005</td>
<td>1016</td>
<td>1147</td>
<td>1182</td>
<td>920</td>
</tr>
<tr>
<td>Protein coding genes with functional prediction</td>
<td>1438</td>
<td>1525</td>
<td>1499</td>
<td>1954</td>
<td>1908</td>
<td>1333</td>
</tr>
<tr>
<td>5S rRNA</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>16S rRNA</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>23S rRNA</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>tRNA genes</td>
<td>36</td>
<td>43</td>
<td>44</td>
<td>50</td>
<td>47</td>
<td>44</td>
</tr>
</tbody>
</table>
**Table 1** – Statistics for the genome of *Thiovulum* sp. ES and of its five closest sequenced relatives as obtained from the Integrated Microbial Genomes Database in January 2012 (IMG 3.5). Ortholog numbers were calculated as described in Experimental Procedures.
Table 2 - 16S rRNA gene identities and ortholog fractions for each genome pair for *Thiovulum* sp. ES and its five closest relatives. The denominator (total number of protein-encoding genes) in ortholog fraction calculations is always the genome listed along the horizontal axis of the table.

<table>
<thead>
<tr>
<th></th>
<th>Thiovulum sp. ES</th>
<th>Sulfurimonas autotrophica OK10</th>
<th>Sulfurimonas denitrificans</th>
<th>Sulfuricurvum kujiense YK-1</th>
<th>Sulfurimonas gotlandica GD1</th>
<th>Sulfurovum sp. NBC37-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thiovulum sp. ES</td>
<td>100.0%/100.0%</td>
<td>85.1%/41.0%</td>
<td>85.1%/41.3%</td>
<td>85.0%/32.0%</td>
<td>86.2%/31.5%</td>
<td>85.2%/32.2%</td>
</tr>
<tr>
<td>Sulfurimonas autotrophica OK10</td>
<td>85.1%/41.0%</td>
<td>100.0%/100.0%</td>
<td>93.2%/68.9%</td>
<td>93.2%/48.8%</td>
<td>88.6%/56.6%</td>
<td>87.1%/50.6%</td>
</tr>
<tr>
<td>Sulfurimonas denitrificans</td>
<td>85.1%/41.0%</td>
<td>93.2%/66.9%</td>
<td>100.0%/100.0%</td>
<td>93.9%/48.9%</td>
<td>88.0%/56.0%</td>
<td>85.9%/45.9%</td>
</tr>
<tr>
<td>Sulfuricurvum kujiense YK-1</td>
<td>85.0%/42.5%</td>
<td>93.2%/63.5%</td>
<td>93.9%/65.5%</td>
<td>100.0%/100.0%</td>
<td>87.7%/52.1%</td>
<td>86.4%/46.6%</td>
</tr>
<tr>
<td>Sulfurimonas gotlandica GD1</td>
<td>86.2%/40.8%</td>
<td>88.6%/71.8%</td>
<td>88.0%/73.1%</td>
<td>87.7%/50.7%</td>
<td>100.0%/100.0%</td>
<td>86.7%/48.0%</td>
</tr>
<tr>
<td>Sulfurovum sp. NBC37-1</td>
<td>85.2%/37.4%</td>
<td>87.1%/57.6%</td>
<td>85.9%/53.8%</td>
<td>86.4%/40.7%</td>
<td>86.7%/43.1%</td>
<td>100.0%/100.0%</td>
</tr>
</tbody>
</table>