The ISME Journal (2015) 9, 1166–1176
© 2015 International Society for Microbial Ecology All rights reserved 1751-7362/15
www.nature.com/ismej

ORIGINAL ARTICLE

Ecophysiology of *Thioploca ingrica* as revealed by the complete genome sequence supplemented with proteomic evidence

Hisaya Kojima[1], Yoshitoshi Ogura[2,3], Nozomi Yamamoto[4], Tomoaki Togashi[5], Hiroshi Mori[5], Tomohiro Watanabe[1], Fumiko Nemoto[1], Ken Kurokawa[4,5], Tetsuya Hayashi[2,3] and Manabu Fukui[1]

[1]*The Institute of Low Temperature Science, Hokkaido University, Sapporo, Japan;* [2]*Division of Microbial Genomics, Department of Genomics and Bioenvironmental Science, Frontier Science Research Center, University of Miyazaki, Miyazaki, Japan;* [3]*Division of Microbiology, Department of Infectious Diseases, Faculty of Medicine, University of Miyazaki, Miyazaki, Japan;* [4]*Earth-Life Science Institute, Tokyo Institute of Technology, Tokyo, Japan and* [5]*Department of Biological Information, Graduate School of Bioscience and Biotechnology, Tokyo Institute of Technology, Tokyo, Japan*

Large sulfur-oxidizing bacteria, which accumulate a high concentration of nitrate, are important constituents of aquatic sediment ecosystems. No representative of this group has been isolated in pure culture, and only fragmented draft genome sequences are available for these microorganisms. In this study, we successfully reconstituted the genome of *Thioploca ingrica* from metagenomic sequences, thereby generating the first complete genome sequence from this group. The *Thioploca* samples for the metagenomic analysis were obtained from a freshwater lake in Japan. A PCR-free paired-end library was constructed from the DNA extracted from the samples and was sequenced on the Illumina MiSeq platform. By closing gaps within and between the scaffolds, we obtained a circular chromosome and a plasmid-like element. The reconstituted chromosome was 4.8 Mbp in length with a 41.2% GC content. A sulfur oxidation pathway identical to that suggested for the closest relatives of *Thioploca* was deduced from the reconstituted genome. A full set of genes required for respiratory nitrate reduction to dinitrogen gas was also identified. We further performed a proteomic analysis of the *Thioploca* sample and detected many enzymes/proteins involved in sulfur oxidation, nitrate respiration and inorganic carbon fixation as major components of the protein extracts from the sample, suggesting that these metabolic activities are strongly associated with the physiology of *T. ingrica* in lake sediment.

The ISME Journal (2015) **9,** 1166–1176; doi:10.1038/ismej.2014.209; published online 24 October 2014

Introduction

Although key microbial players in various biogeochemical processes have been identified, they are often not available in pure culture. One major group of such organisms is large sulfur-oxidizing bacteria that have the capacity to accumulate a large amount of intracellular nitrate. In habitats where a sufficient supply of sulfide is available, they can form dense and widespread bacterial mats. Because of their large biomass and ecophysiological characteristics, they have been regarded as important constituents of aquatic ecosystems. In fact, they have been shown to have a considerable impact on nitrogen and phosphorus dynamics in marine sediments (Zopfi *et al.*, 2001; Schulz and Schulz, 2005; Prokopenko *et al.*, 2013).

Until recently, these organisms were classified into three genera, *Beggiatoa*, *Thioploca* and *Thiomargarita*, on the basis of their morphological features (Salman *et al.*, 2011). Although recent studies have revealed that these microorganisms are much more diverse than previously thought (Salman *et al.*, 2013), all known nitrate-storing sulfur oxidizers belong to a particular lineage within the class *Gammaproteobacteria*. Using a large set of 16S ribosomal RNA (rRNA) gene sequences, the bacteria in this lineage were reclassified in 2011, and the revised family *Beggiatoaceae* was proposed to encompass all of these bacteria (Salman *et al.*, 2011). The extended family contains three genera along with nine candidate genera, two of which were proposed after the reclassification

Correspondence: H Kojima or M Fukui, The Institute of Low Temperature Science, Hokkaido University, Kita-19, Nishi-8, Kita-ku, Sapporo 0600819, Japan.
E-mail: kojimah@pop.lowtem.hokudai.ac.jp or my-fukui@pop.lowtem.hokudai.ac.jp

Received 4 August 2014; revised 17 September 2014; accepted 22 September 2014; published online 24 October 2014

(Salman *et al.*, 2013). From this family, only a few strains of the genus *Beggiatoa* have been isolated in pure culture (Salman *et al.*, 2013).

As a cultured representative of this family, *Beggiatoa alba* B18LD has been subjected to whole-genome sequencing, and its draft genome sequence is now available in public databases; however, this strain cannot accumulate nitrate. The draft genome sequences of nitrate-storing sulfur oxidizers have been obtained for *Candidatus* Isobeggiatoa and *Candidatus* Parabeggiatoa, both of which are from coastal marine sediments (Mußmann *et al.*, 2007), and for *Candidatus* Maribeggiatoa, which is from a deep-sea sediment that is influenced by hydrothermal fluid (MacGregor *et al.*, 2013a, b). These genome sequencing projects all employed whole-genome multiple displacement amplification to obtain sufficient amounts of DNA for sequencing from single bacterial filaments that are expected to consist of clonal cells. The single-filament approach may be effective for coping with genetic diversities among the morphologically indistinguishable organisms inhabiting the same sediment, but risks generating chimeric sequences during the process of amplification. Presumably because of the presence of such chimeric sequences and/or other difficulties (for example, short reads and repeat regions in the genomes), sequence assembly in these studies was not fully successful, as illustrated by large numbers of contigs (>800 contigs). Although these draft genome sequences have provided valuable insights into the physiology and evolution of this group of microorganisms, the availability of complete genome sequences of large sulfur-oxidizing bacteria is highly desirable.

Members of the genus *Thioploca* are gliding filamentous bacteria that have a common sheath surrounding the trichomes. The first description of the genus was made in the early twentieth century, with the type species of *Thioploca schmidlei* obtained from freshwater lake sediment (Lauterborn, 1907). Marine species with much larger cell sizes were once included in this genus, but they have been reclassified within a candidate genus, Candidatus *Marithioploca* (Salman *et al.*, 2011). *Thioploca ingrica* was described as the second species of the genus (Wislouch, 1912) and was listed in the Approved Lists of Bacterial Names after a temporary loss of status as valid name (Maier, 1984). It has been retained in this genus, after the reclassification of the family *Beggiatoaceae* (Salman *et al.*, 2011), as the sole species whose 16S rRNA gene sequences are available. Morphologically, *T. ingrica* was defined as a *Thioploca* species having a trichome 2.0–4.5 μm in diameter (Wislouch, 1912; Maier, 1984). Organisms that fit this description have been found in freshwater and brackish sediments of various localities, and the placement of these organisms in the same species has been generally supported by their 16S rRNA gene sequences (Maier and Murray, 1965; Nishino *et al.*, 1998; Zemskaya *et al.*, 2001, 2009; Dermott and Legner, 2002; Kojima *et al.*, 2003, 2006; Høgslund *et al.*, 2010; Nemoto *et al.*, 2011, 2012; Salman *et al.*, 2011). Nitrate accumulation by *T. ingrica* was reported in previous studies, although the intracellular concentrations were much lower than in relatives with large vacuoles (Zemskaya *et al.*, 2001; Kojima *et al.*, 2007; Høgslund *et al.*, 2010).

In this study, the complete genome sequence of *T. ingrica* was reconstituted from metagenomic sequences, uncovering the metabolic and genetic characteristics of this organism. In addition, proteomic analysis was performed to investigate the physiology of *Thioploca* in lake sediments.

Materials and methods

Sampling

Thioploca samples were obtained at a site near the north shore of Lake Okotanpe, ~200 m from the site of our previous study (Nemoto *et al.*, 2011). Sediment samples were obtained with an Ekman–Birge grab sampler and were immediately sieved at the site with a 0.25-mm mesh in lake water. The materials retained on the mesh were transferred to lake water and kept at 4 °C in the dark until processing in the laboratory. Upon returning to the laboratory, *Thioploca* filaments were individually removed with forceps from the materials collected by sieving and were then repeatedly washed with filter-sterilized lake water. The washed filaments were stored at −30 °C for DNA extraction (the samples obtained in 2013) or at −80 °C for protein extraction (in 2012) or were immediately subjected to protein extraction (in 2011).

Sequencing and genome assembly

Genomic DNA was prepared from the washed *Thioploca* filaments using the Wizard Genomic DNA Purification Kit (Promega, Madison, WI, USA). A PCR-free paired-end library was constructed using the TruSeq DNA PCR-free Sample Prep Kit (Illumina, San Diego, CA, USA) and sequenced on the Illumina MiSeq platform (2 × 300 bp). After trimming low-quality bases (quality score <25) and adapter sequences, sequence assembly using varying numbers of reads (5 000 000, 4 000 000, 3 000 000, 2 000 000 or 1 000 000 reads) was performed using the Platanus assembler (http://platanus.bio.titech.ac.jp; Kajitani *et al.*, 2014) with the default setting to determine the optimal number of input sequences for assembly. Based on the results of this analysis, we used 3 000 000 reads (1 500 000 sequence pairs) for assembly and obtained 38 scaffolds (>1 kb). Of these, 24 were larger than 5 kb. As 23 out of the 24 scaffolds showed similar ranges of GC content and sequence coverage, a single circular superscaffold was manually constructed from the 23 scaffolds by PCR. All gaps within and between scaffolds were closed by sequencing gap-spanning PCR products using an ABI3130xl DNA sequencer (Applied Biosystems, Foster City, CA, USA). Because one scaffold (28.75 kb in size) was derived from a plasmid-like circular DNA molecule, gaps in this scaffold were also closed in the same way. The BWA program (http://bio-bwa.sourceforge.net; Li and Durbin, 2009) was used for mapping analysis of the Illumina reads to the reconstituted genome. The reconstituted sequences have been deposited in the DDBJ/NCBI/EMBL databases (accession no. AP014633).

Annotation

Protein-coding sequences (CDSs) were predicted using MetaGeneAnnotator (Noguchi *et al*., 2008). The transfer RNA (tRNA) and rRNA genes were identified using tRNAScan-SE (Lowe and Eddy, 1997) and RNAmmer (Lagesen *et al*., 2007), respectively. The open reading frame extraction function in the In Silico Molecular Cloning Genomic Edition version 5.2.65 software (In Silico Biology, Inc., Yokohama, Japan) was used for additional gene prediction. Functional annotation of CDSs was performed on the basis of the results of BLASTP searches (Altschul *et al*., 1997) against the NCBI (National Center for Biotechnology Information) nonredundant database, the KEGG (Kyoto Encyclopedia of Genes and Genomes) database (Kanehisa *et al*., 2014) and the NCBI Clusters of Orthologous Groups (COG). CDSs were annotated as hypothetical protein-coding genes when the CDSs met any of the following three criteria in the top hit of the BLASTP analysis: (1) E-value $>1\text{e}-8$, (2) length coverage $<60\%$ against query sequence or (3) sequence identity $<30\%$.

Genomic comparison and phylogenetic analyses

The genome sequences of *Thioploca* relatives were obtained from the following sites. *Beggiatoa* sp. SS (*Candidatus* Parabeggiatoa, henceforth 'SS') and *Beggiatoa* sp. PS (*Candidatus* Isobeggiatoa, 'PS') were taken from the NCBI Bacteria draft genome FTP site (ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria_DRAFT); *Beggiatoa* sp. Orange Guaymas (*Candidatus* Maribeggiatoa, 'Orange Guaymas') and *B. alba* B18LD were taken from DOE-JGI IMG (https://img.jgi.doe.gov/w); and *Thioploca araucae* Tha-CCL (*Candidatus* Marithioploca, 'Tha-CCL') was taken from JCVI (https://moore.jcvi.org/moore/SingleOrganism.do?speciesTag=THACCL). Functional annotation of CDSs was performed as described above. To search for homologous sequences, all-against-all BLASTP searches (E-value <0.001) were performed for all of the CDSs from the six strains (*T. ingrica* and relatives). On the basis of the BLASTP results, the CDSs were clustered by applying OrthoMCL (Li *et al*., 2003) with default parameters.

Homologous gene clusters consisting of members from two or more strains were used to estimate the gene content phylogeny. The Euclidean distances among strains were calculated from the homolog content matrix among these strains, and complete-linkage hierarchical clustering was then conducted using the hclust function in the R software (http://www.r-project.org/).

A maximum-likelihood phylogenetic tree for *T. ingrica* and its relatives was built upon a concatenated protein sequence alignment of the universal single-copy genes (USCGs) and the *gyrB* gene. The genome of SS was excluded from the analysis because the quality of the draft genome was too low, whereas the sequence of *Thioalkalivibrio nitratireducens* DSM 14787 was included in the analysis. Among the 35 USCGs originally proposed (Raes *et al*., 2007), 4 are duplicated in the genomes of Tha-CCL and *T. ingrica* (*rpsG* and *rpsS* in Tha-CCL, *rplM* and *rpsI* in *T. ingrica*). Therefore, these four genes were excluded from the analysis. Because the *valS* and *gyrB* genes of Tha-CCL were both fragmented into two different contigs, they were manually merged. Each homolog cluster was separately aligned using MAFFT software version 7.130b (Katoh and Standley, 2013) with default parameters, and then the 32 alignments (31 USCGs and *gyrB*) were concatenated. Because some homologs were not found in the draft genomes of PS and Tha-CCL (*gyrB* and *ychF* in PS and *pheS* and *rplK* in Tha-CCL), the sequences of these genes were replaced with gaps in all positions. From the alignment of the concatenated sequences, a maximum likelihood tree was constructed using MEGA5 (Tamura *et al*., 2011). The Whelan and Goldman model (Whelan and Goldman, 2001) was selected based on the result of a likelihood ratio test. An initial tree for the heuristic search was obtained by applying the Neighbor-Join and BioNJ algorithms to a matrix of pairwise distances estimated using the JTT (Jones–Thornton–Taylor) model and then selecting the topology with a superior log likelihood value. A discrete gamma distribution was used to model evolutionary rate differences among sites (five categories (+G, parameter $=0.6269$)). All positions with $>50\%$ alignment gaps were eliminated.

Protein analysis

Proteomic analyses were conducted using fresh or stored (at $-80\,^{\circ}$C for 120 days) samples obtained in 2011 and 2012, respectively. Protein extraction, sodium dodecyl sulfate polyacrylamide gel electrophoresis, in-gel trypsin digestion and nano-liquid chromatography tandem mass spectrometry were performed as previously described (Watanabe *et al*., 2012). The reconstituted chromosome sequence was used to generate the database for protein identification with the Mascot search program version 2.4.0 (MS/MS Ion Search; Matrix Science, Boston, MA, USA). Search parameters were set as follows: tryptic digest with a maximum of two missed cleavages; fixed modifications, carbamidomethyl cysteine; variable modifications, methionine oxidation; peptide masses, monoisotopic; positive charge (+1, +2, +3) of peptide; and mass tolerance of 1.2 Da for precursor ions and 0.8 Da for product ions. The false discovery rate was estimated using an automatic decoy search against a randomized database with a significance threshold of $P<0.05$. Protein detection was judged as positive when two or more different peptides were detected, and the exponentially modified protein abundance index was calculated as previously described (Ishihama *et al.*, 2005). The normalized protein content value was calculated as a percentage of each exponentially modified protein abundance index in the summation of all identified proteins.

Results and discussion

Reconstitution of the T. ingrica *genome*

To reconstitute the *T. ingrica* genome, we used the washed *T. ingrica* filaments to generate >100 000 000 paired-end metagenomic sequences using the Illumina MiSeq platform. Because low-redundancy sequences derived from minor contaminating bacteria disturbed sequence assembly, 3 000 000 reads were used as input sequences, in accordance with the results of an optimization procedure (see Materials and methods). As a result, 24 scaffolds of >5 kb were constructed, all of which showed a similar range of GC content (from 39.6 to 44.3%, mostly ~41%). All but one of the scaffolds showed a similar sequence coverage (from 104 × to 116 ×), suggesting that these scaffolds were derived from a single dominant species. Furthermore, all 16S rRNA gene reads belonging to these scaffolds were identical to the published 16S rRNA sequences of *T. ingrica* samples from Lake Ogawara and Lake Okotanpe (Kojima *et al.*, 2006; Nemoto *et al.*, 2011). We manually constructed a single circular super-scaffold from the 23 scaffolds with similar sequence coverage and, finally, reconstituted a complete circular chromosome by closing all of the gaps within and between the scaffolds. We searched the reconstituted chromosome for the USCGs (Raes *et al.*, 2007) and identified a single homolog for each of the USCGs, except for *rplM* and *rpsI* that had been duplicated. This finding confirmed successful reconstitution of the *T. ingrica* genome.

A scaffold that was not included in the reconstituted chromosome showed a GC content (40.3%) similar to the chromosome, but its sequence coverage (381 ×) was much higher. Sequence analysis revealed that this scaffold encoded a putative replication protein, an integrase, a lytic transglycosidase and a response regulator (all other putative genes encoded hypothetical proteins), suggesting that the scaffold represented a plasmid-like DNA molecule. In fact, using PCR examination and gap filling, we reconstituted a single circular sequence (28 669 bp) from this scaffold. Although this plasmid-like element encoded an integrase, we did not find any evidence that it had been integrated into the chromosome of *T. ingrica*. The similarity in GC content suggested that the plasmid-like element may be associated with *T. ingrica*, but this idea needs to be experimentally confirmed.

In a mapping analysis of Illumina reads to the reconstituted sequences, 77.1% and 1.7% of the reads used for assembly mapped to the *T. ingrica* chromosome and the plasmid-like element, respectively. This finding indicated that the DNA from the washed *T. ingrica* filaments represented ca. 79% of the DNA preparation used for genome reconstitution.

General features of the T. ingrica *genome*

The general features of the *T. ingrica* chromosome are shown in Table 1 and Figure 1. Nucleotide position 1 of the chromosome was arbitrarily determined, as the replication origin could not be identified. The genome lacks an identifiable *dnaA* gene, and the GC skew fluctuates irregularly throughout the chromosome. The lack of *dnaA* may be a characteristic shared by nitrate-storing sulfur oxidizers, whereas *dnaA* was identified in *B. alba* B18LD. In all available draft genomes of nitrate-storing sulfur oxidizers, no *dnaA* gene was found by BLASTP or BLASTN searches when using the sequence of *B. alba* as a query.

Approximately 48% of the predicted genes of *T. ingrica* were classified into COG functional categories (Figure 1, Supplementary Table S1). No significant differences were identified among *T. ingrica*, Orange Guaymas and *B. alba* B18LD in their distributions of genes into COG functional categories (Supplementary Table S1).

Genomic comparison and phylogenetic analyses

Homolog distributions among the three strains (*T. ingrica*, Orange Guaymas and *B. alba* B18LD) are summarized in Supplementary Figure S1 and Supplementary Table S2. The number of genes shared by the three strains suggests that the number of core genes in the family *Beggiatoaceae* is <1500.

The phylogenetic trees constructed based on the sequences of USCGs or on gene content are shown in Figure 2. Both trees located *T. ingrica* at a position

Table 1 General features of the *Thioploca ingrica* genome

Features	
Size (bp)	4 810 005
GC content (%)	41.21
Total no. of coding sequences	3964
No. of tRNA genes	44
No. of 23S rRNA	1
No. of 16S rRNA	1
No. of 5S rRNA	1

Abbreviations: rRNA, ribosomal RNA; tRNA, transfer RNA.

between its marine relatives and *B. alba*. This result is consistent with the phylogenetic position of *T. ingrica* in the family *Beggiatoaceae* that was inferred by 16S rRNA sequence analysis (Salman *et al*., 2011).

Proteomic analysis

Metagenomic analysis indicated that 79% of the DNA in the samples originated from *T. ingrica*. Considering that *T. ingrica* has a very large cell size and a normal genome size, *T. ingrica* cells may have a large protein/DNA ratio. If so, protein extracts obtained from the washed *T. ingrica* filaments may contain only a trace amount of proteins from microorganisms contaminating the filament sample. From two samples obtained in different years, we detected 864 and 826 proteins; among these, 563 were detected in both samples (Supplementary Tables S3 and S4). Among the 54 ribosomal proteins encoded in the genome, only 21 were detected in both samples, and 20 were not detected in either sample. This result indicated that only a limited portion of the *T. ingrica* proteome was detected and that a considerable portion of the proteins might have been missed by this analysis. However, the result also suggested that the proteins detected in this analysis were contained in the samples at relatively high concentrations. Therefore, the detected proteins are most likely *T. ingrica* proteins, especially those with high relative abundances. Protein abundance was evaluated based on exponentially modified protein abundance index, and the relative abundances of major proteins are shown in Figure 3. It should be noted that the samples subjected to the protein analysis were mixtures of cells positioned in different layers of sediment. Therefore, the detected proteins might derive from cells experiencing different environmental conditions. For instance, key enzymes for respiration with each of three electron acceptors (oxygen, nitrate and dimethyl sulfoxide) were all detected in both samples, but there is no way to know whether these proteins originated from the same cell or from cells at different positions in the redox gradient. Nevertheless, the detected proteins provide some insight into the ecophysiology of *T. ingrica*, as described below.

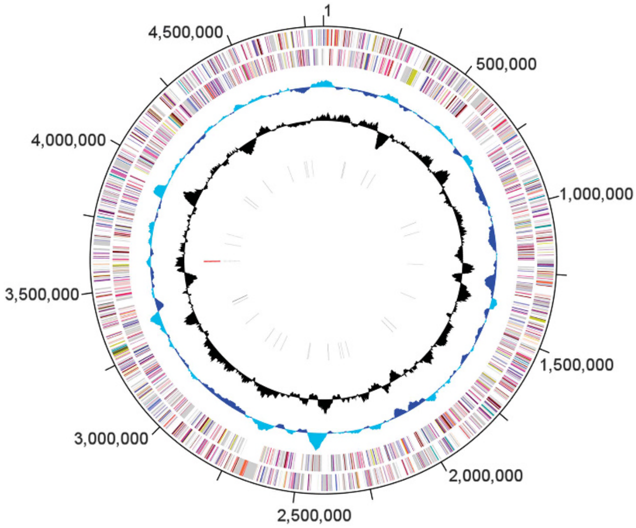

Figure 1 Circular view of the reconstituted *T. ingrica* chromosome. The outermost two circles show predicted CDSs on the plus and minus strands. The CDSs are color coded by COG categories (unclassified genes are shown in gray). The third and fourth circles indicate GC content (mean value =41%) and GC skew, respectively. Genes encoding rRNA (red) and tRNA (gray) are shown in the two innermost circles.

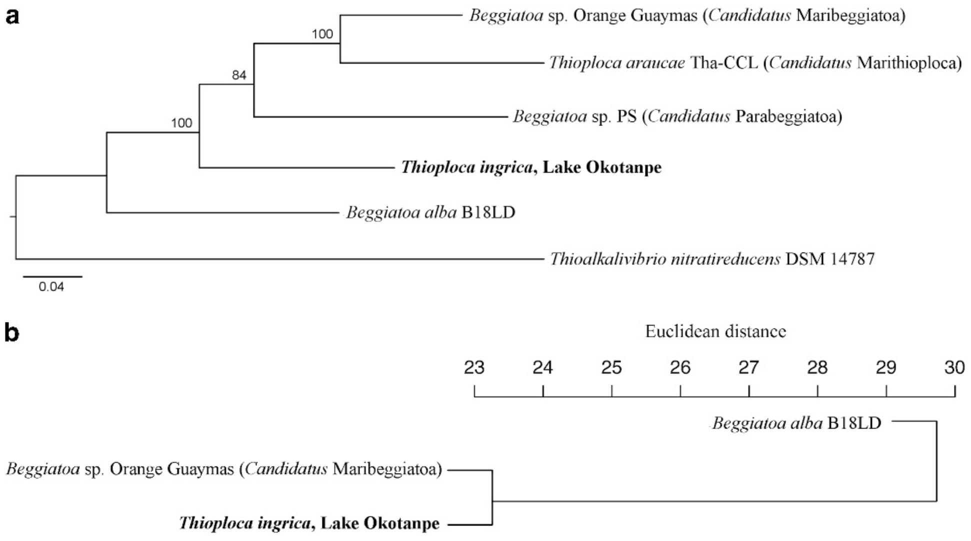
a
100
84
100
Beggiatoa sp. Orange Guaymas (*Candidatus* Maribeggiatoa)
Thioploca araucae Tha-CCL (*Candidatus* Marithioploca)
Beggiatoa sp. PS (*Candidatus* Parabeggiatoa)
***Thioploca ingrica*, Lake Okotanpe**
Beggiatoa alba B18LD
Thioalkalivibrio nitratireducens DSM 14787
0.04
b
Euclidean distance
23
24
25
26
27
28
29
30
Beggiatoa alba B18LD
Beggiatoa sp. Orange Guaymas (*Candidatus* Maribeggiatoa)
***Thioploca ingrica*, Lake Okotanpe**

Figure 2 Phylogenetic trees of *T. ingrica* and its relatives. (**a**) A tree constructed using the concatenated sequences of USCGs. The numbers at each node represent the percentage values from 1000 bootstrap resamplings. (**b**) A tree constructed using the gene contents of each strain.

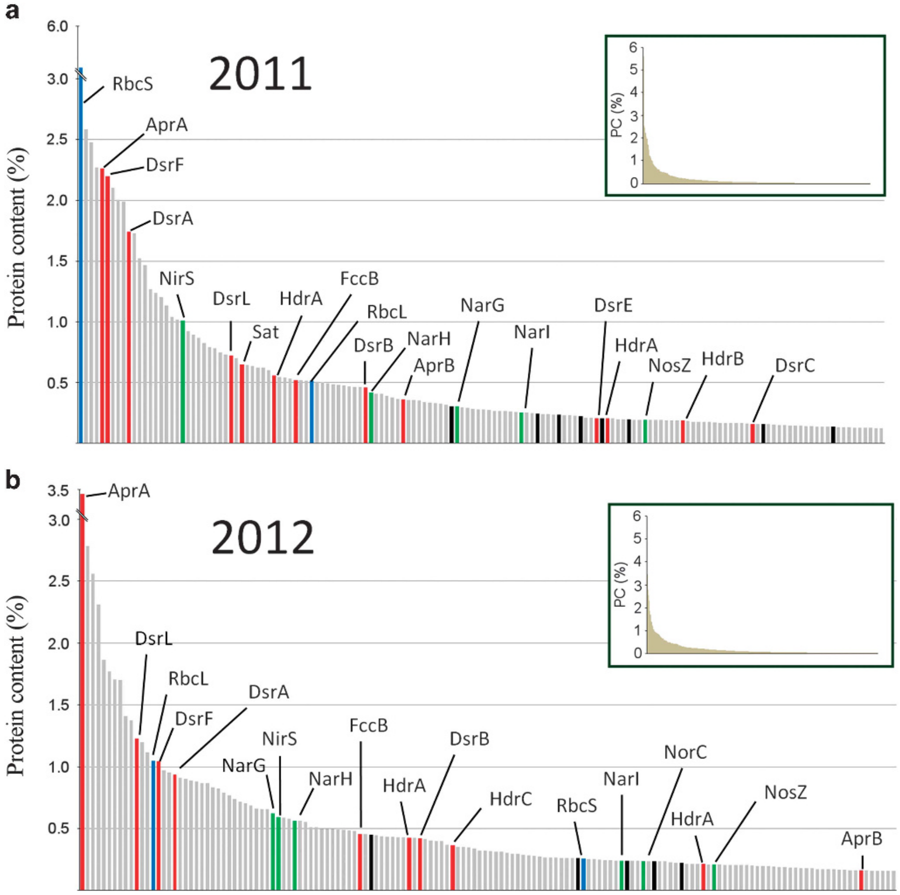

Figure 3 Overviews of the proteomic analysis results from washed *Thioploca* filaments obtained in 2011 (**a**) and 2012 (**b**). The proteins are sorted according to their relative abundances in each sample, and the 150 most abundant proteins are shown. The data for the 500 most abundant proteins are outlined in the small nested boxes. Proteins involved in sulfur oxidation and nitrate respiration are indicated in red and green, respectively. The RuBisCO proteins are indicated by blue bars, and ribosomal proteins are shown in black for comparison.

Sulfur oxidation

Although an array of circumstantial evidence has suggested sulfur oxidation by *Thioploca*, there has been no solid evidence to indicate that *T. ingrica* is truly a sulfur oxidizer. In the *T. ingrica* genome, genes involved in the proposed sulfur oxidation pathway were identified, and their gene products were detected in the protein analysis (Figures 3 and 4). The sulfur oxidation pathway of *T. ingrica* is basically the same as that proposed for its closest relatives, but it is notable for being the first case to confirm the presence of a full set of genes. Furthermore, the sequences of all of the genes are now fully available.

As shown in Figures 3 and 4, it appears that sulfide oxidation by *T. ingrica* is mediated by the flavocytochrome *c*/sulfide dehydrogenase encoded by the *fccAB* (*soxEF*) genes (THII_1691-1692). The gene for another enzyme mediating sulfide oxidation, *sqr*, which encodes a sulfide:quinone oxidoreductase, was also identified in the *T. ingrica* genome (THII_0779).

It appears that *T. ingrica* oxidizes elemental sulfur to sulfite via the dissimilatory sulfite reductase (DSR) system. *T. ingrica* possesses a full set of genes for a DSR system, *dsrABEFHCMKLJOP* (THII_0611-0622), *dsrNR* (THII_3808-3809) and *dsrS* (THII_1535). In general, genes for the DSR system are mutually exclusive with the *soxCD* genes, with an exception in an environmental fosmid sequence (Lenk *et al*., 2012). Accordingly, genes corresponding to *soxCD* were not found in the *T. ingrica* genome.

For sulfite oxidation to sulfate, *T. ingrica* contains the *aprBA* genes, encoding an adenosine-5′-phosphosulfate reductase (THII_3105-3106); the *sat* gene, encoding a sulfate adenylyltransferase (THII_1057); and the *hdrAACB* genes (THII_2277-2280). The proteins encoded by the *hdr* genes are thought to be components of the Qmo complex that interacts with the adenosine-5′-phosphosulfate reductase (Dahl *et al*., 2013). Sulfite oxidation by these enzymes occurs in the cytoplasm, where sulfite is generated by DsrAB (Pott and Dahl, 1998). The *T. ingrica* genome also encodes an enzyme that mediates direct sulfite oxidation to sulfate (the *sorAB* genes; THII_2320-2321), but this sulfite oxidase usually works outside of the cytoplasm (Kappler *et al*., 2000; Kappler and Bailey, 2005).

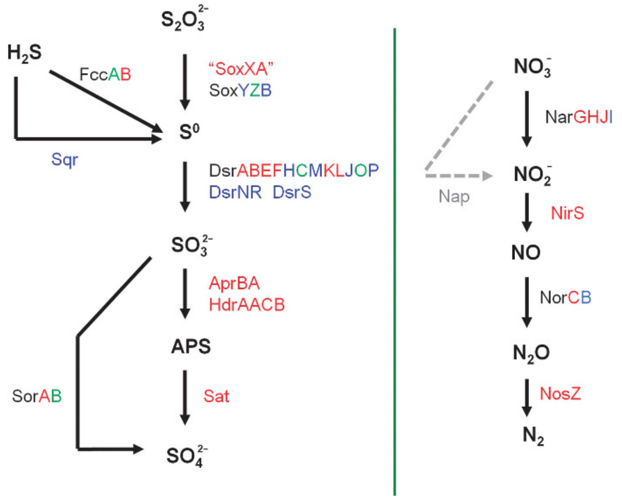

Figure 4 Pathways for sulfur oxidation and dissimilatory nitrate reduction deduced from the *T. ingrica* genome. Proteins in red were detected in both samples subjected to protein analysis, and those in green were detected in one of the samples. Proteins encoded by the genome but not detected in the protein analysis are shown in blue. Gray dashed lines indicate alternative pathways mediated by the enzymes (shown in gray) that are not encoded by the *T. ingrica* genome but are found in its close relatives.

The *T. ingrica* genome sequence further suggested that this microorganism is capable of thiosulfate oxidation by a system referred to as the Sox system (Friedrich *et al.*, 2001). Of the three core components (SoxAX, SoxYZ and SoxB) that are thought to be indispensable for thiosulfate oxidation by this system (Hansen *et al.*, 2006), two are encoded in the *soxYZB* gene cluster (THII_1577–1579) in *T. ingrica* genome. However, genes encoding SoxA and SoxX were not identified. These genes are usually located in a *soxXYZAB* gene cluster (Kappler and Maher, 2013). In *T. ingrica*, the function of the SoxAX complex might be substituted by a single protein, encoded by a gene located adjacent to the *soxYZB* cluster (THII_1580). The protein encoded by this gene is closely related to a protein designated as 'SoxXA' in the PS genome. Although these proteins are larger than conventional SoxA proteins, their sequences exhibit a partial similarity to SoxA.

The lifestyle of *T. ingrica* as a sulfur oxidizer, which was deduced from the genome sequence (Figure 4), is supported by the fact that most of the enzymes described above were among the most abundant proteins identified in the proteomic analysis of the samples directly collected from lake sediment (Figure 3).

Nitrate respiration

A full set of genes required for respiratory reduction of nitrate to N_2 was identified in the *T. ingrica* genome (Figure 4). *T. ingrica* contains a *narGHJI* gene cluster (THII_1673–1676), encoding a membrane-bound nitrate reductase that mediates nitrate reduction to nitrite. In contrast to its closest relatives, the *napAB* genes that encode a periplasmic nitrate reductase were not detected in the *T. ingrica* genome. *T. ingrica* appears to generate N_2 as the end-product of respiration, as it possesses *nirS*, encoding a nitrite reductase (THII_2875); *norCBQ*, encoding a nitric oxide reductase (THII_0334-0336); and *nosZ*, encoding a nitrous oxide reductase (THII_2884). The *norD* gene (THII_0363), which encodes a nitric oxide reductase activation protein, was also identified in the genome. Among these, the *nosZ* gene, which is directly responsible for the generation of N_2, has never been found in the genomes of close relatives of *T. ingrica*. The *nrfA* gene, which encodes an enzyme for dissimilatory nitrite reduction to ammonium, was not found in *T. ingrica*.

The active reduction of nitrate to N_2 by *Thioploca* in sediments has gained strong support from proteomic analysis; enzymes involved in all steps of the successive reduction (Figure 4) were detected, and many of these were major components in the samples (Figure 3). Recently, a large contribution by *Candidatus* Marithioploca to nitrogen loss in anoxic marine sediments was suggested by Prokopenko *et al.* (2013). This hypothesis was premised on nitrate reduction to ammonium by *Candidatus* Marithioploca and on the cooperative involvement of anaerobic ammonia-oxidizing bacteria. Interaction with ammonia oxidizers was also suggested for sulfur oxidizers inhabiting hydrothermal sediments, but the retention of nitrogen is assumed to take place in this case (Winkel *et al.*, 2013). *T. ingrica* may contribute to nitrogen loss in freshwater lake environments in a different way from that of its marine counterparts.

Nitrogen assimilation

Nitrogen (N_2) fixation by *Beggiatoa* species has been previously reported (Nelson *et al.*, 1982), and genes involved in diazotrophy were identified in the *Beggiatoa alba* B18LD genome. In the *T. ingrica* genome, *nifHDKT* (THII_1806-1809), *nifY* (THII_1811), *nifENX* (THII_1813–1815), *nifZW* (THII_3085–3086), *nifV* (THII_3088), *nifQ* (THII_3125), *nifB* (THII_3130), *nifM* (THII_3738) and two copies of *nifA* (THII_2556 and THII_2572) were identified as genes putatively involved in N_2 fixation and its regulation. Among the products of these *nif* genes, the regulatory protein NifA encoded by THII_2556 was detected in the proteomic analysis of both of the samples obtained in 2011 and 2012 (Supplementary Table S4). The others were not detected in the protein analysis, except for NifY that was detected only in the sample from 2011.

In the sediment of Lake Okotanpe, both nitrate and ammonium are available (Nemoto *et al.*, 2011); thus, N_2 fixation would not be advantageous because of the energetically high cost required for the process. The *T. ingrica* genome encodes both an ammonium transporter (THII_2433) and an assimilatory nitrate reductase, although these proteins were not detected in the protein analysis. *T. ingrica* may also use organic compounds as nitrogen sources, as several subunits for amino acid transporters were detected in the proteomic analysis. Taken together, multiple and flexible nitrogen assimilation pathways are deduced for *T. ingrica*.

Nitrate storage

The capacity to store a large amount of nitrate within cells is a unique property of *Thioploca* and closely related sulfur oxidizers. Based on genome sequence information for large marine sulfur oxidizers, a hypothetical model for nitrate accumulation in the vacuole was proposed in analogy to that in plants (Mußmann *et al*., 2007). The model includes two processes: the electrochemical gradient formed by proton pumping and the NO_3^-/H^+ antiporter that depends on the gradient. When the model was proposed, three enzymes were taken into consideration as candidate enzymes responsible for the electrochemical gradient. All three enzymes, namely, a vacuolar-type ATPase, an H^+-pyrophosphatase (THII_0754) and a Ca^{2+}-translocating ATPase (THII_0386), are encoded in the *T. ingrica* genome. The latter two are also present in the *B. alba* genome that lacks nitrate-storing capacity. In the originally proposed model, the vacuolar-type ATPase and H^+-pyrophosphatase were assumed to generate a proton motive force, but experiments using *Candidatus* Allobeggiatoa halophila revealed that these enzymes work in the reverse direction by consuming the proton gradient to generate adenosine triphosphate (ATP) and pyrophosphate (Beutler *et al*., 2012). The originally proposed model also predicted that the NO_3^-/H^+ antiporter is another key protein directly responsible for nitrate accumulation. Whereas BgP0076 and BgP4800 in the genome of PS were assumed to encode NO_3^-/H^+ antiporters, corresponding genes were not identified in *T. ingrica*. This finding indicates that the previously proposed model cannot be fully applicable to *T. ingrica*. Among the proteins mentioned above, an H^+-pyrophophatase and a Ca^{2+}-translocating ATPase were both detected in the proteomic analyses of filament samples collected in both 2011 and 2012 (Supplementary Table S4).

The molecular mechanism of nitrate accumulation in *T. ingrica* and relatives is still not fully understood, and further studies are necessary. The complete genome sequence information of *T. ingrica* obtained in this work will facilitate such studies.

Carbon metabolism

Many genes for the enzymes constituting the tricarboxylic acid cycle have been identified in other members of the family *Beggiatoaceae*, but some of the genes are missing from their draft genomes (MacGregor *et al*., 2013a). In the genome of *T. ingrica*, a full set of genes for tricarboxylic acid cycle enzymes was identified. Most of these enzymes were also detected in the protein analysis, confirming that the cycle was actually operating in *T. ingrica* cells. Similarly, genes encoding enzymes for the glycolytic pathway were also identified in the reconstituted genome, and the enzymes mediating all pathway steps were detected in the protein analysis (Supplementary Table S4).

In previous studies, acetate assimilation by *T. ingrica* was demonstrated by microautoradiography (Kojima *et al*., 2007; Høgslund *et al*., 2010). The gene for acetate uptake, *actP*, which encodes a cation/acetate symporter (THII_3816), was identified in the genome, as was the *acs* gene, encoding an acetyl-CoA synthetase (THII_1554) that converts acetate into acetyl-CoA. In addition to the tricarboxylic acid cycle, enzymes in the glyoxylate cycle (isocitrate lyase and malate synthase, encoded by *aceA* and *aceB*, respectively; THII_0533–0534) are also encoded by the genome; thus, acetyl-CoA can probably be used for both dissimilatory and assimilatory carbon metabolism.

Inorganic carbon fixation by *T. ingrica* has also been demonstrated in a previous study (Høgslund *et al*., 2010), and the key enzymatic activities of the reductive pentose phosphate cycle were detected in some strains of *Beggiatoaceae* (McHatton *et al*., 1996). In the *T. ingrica* genome, the *rbcLS* genes (THII_3311–3312), which encode the large and small chains of form I ribulose bisphosphate carboxylase (RuBisCO), were identified, in addition to eight other genes for the enzymes of Calvin–Benson–Bassham cycle. The gene for form II RuBisCO was not found in the genome.

As described above, the protein analysis suggested that *T. ingrica* cells gain energy from sulfur oxidation. The detection of other key enzymes, namely, the gene products of *rbcLS*, *actP* and *acs*, in both protein samples further supported the notion that acetate and bicarbonate are serving as carbon sources for *T. ingrica* in lake sediment. Methylotrophy in *Beggiatoa alba* has been demonstrated, and genes for key enzymes have been identified (Jewell *et al*., 2008), but such genes were not found in the *T. ingrica* genome.

Oxygen and dimethyl sulfoxide respiration

T. ingrica most likely has the capacity to use oxygen as a terminal electron acceptor for respiration, as it contains the *ccoNOQP* gene cluster that encodes a high-affinity cytochrome *c* bb3-oxidase (THII_1615–1617), and as the proteomic analysis detected the CcoN, CcoO and CcoP proteins (Supplementary Table S4). In the PS genome, the genes for the low-affinity cytochrome *c* aa3-oxidase were also found, but their counterparts were not identified in the *T. ingrica* genome. Notably, all of the subunits of dimethyl sulfoxide reductase (THII_3261–3263) were detected in both samples subjected to protein analysis, suggesting that dimethyl sulfoxide is also utilized by *T. ingrica* for respiration.

Phosphorus metabolism

Accumulation of phosphorus in the form of polyphosphate has been reported for some bacteria of the family *Beggiatoaceae* (Schulz and Schulz, 2005; Brock and Schulz-Vogt, 2011; Brock *et al*., 2012). In previous studies, *T. ingrica* samples were subjected to Toluidine blue staining to visualize intracellular polyphosphate granules, but only negative results were obtained (Høgslund *et al*., 2010; Nemoto *et al*., 2011). In the *T. ingrica* genome, the gene for polyphosphate kinase (*ppk*) was identified (THII_2967). Polyphosphate is a multifunctional molecule, and the presence of this gene may not yield the potential to form polyphosphate granules; however, the *ppgK* gene, encoding a polyphosphate glucokinase, was also identified in the genome (THII_0714). Previously, the phytase-encoding gene found in PS was discussed in relation to the acquisition of inorganic phosphate (Mußmann *et al*., 2007). This gene is also present in the *T. ingrica* (THII_2499) and *B. alba* genomes. In the protein analysis, polyphosphate kinase, polyphosphate glucokinase and phytase were all repeatedly detected.

Osmoregulation by the glycine betaine/proline transporter

Thioploca is unique in habitat preference among nitrate-storing sulfur oxidizers. In contrast to its relatives living in marine sediments, *T. ingrica* inhabits freshwater and brackish environments. To adapt to habitats of differing salinity, osmoregulation systems should play important roles. *T. ingrica* has a full set of genes for the ProU glycine betaine/proline transport system. The system consists of the ATP-binding protein ProV (THII_1898), the permease protein ProW (THII_1897) and the substrate-binding protein ProX (THII_1896). The ProU system is the transport system for various osmolytes, such as glycine betaine, proline and proline betaine, in Gram-negative bacteria (Sleator and Colin, 2001). Some bacteria harboring the ProU system can cope with increased environmental osmolarity by accumulating glycine betaine (Lucht and Bremer, 1994). A full set of genes for this system is also present in the genome of Orange Guaymas, obtained from marine sediment under the influence of hydrothermal fluid in the Guaymas Basin (MacGregor *et al*., 2013a). It was suggested that the habitat of Orange Guaymas is exposed to large fluctuation of temperature (McKay *et al*., 2012), presumably because of changes in supply of the hot water. The osmoregulation systems may also be effective to adapt to frequent changes in the composition of surrounding water, brought about by fluctuating mixing ratio of hydrothermal fluid.

Conclusion

The complete genome sequence of *T. ingrica* was successfully reconstituted from metagenomic sequences, thus representing the first complete genome sequence of a nitrate-storing sulfur oxidizer. We found that *T. ingrica* possesses all of the genes required for complete denitrification. The presence of the denitrification system in *T. ingrica* was further confirmed by protein analysis, as were several other physiologically important functions, such as sulfur oxidation and inorganic carbon fixation. The complete genome sequence of *T. ingrica* will be a valuable genetic basis for a wide range of future studies on nitrate-storing sulfur oxidizers that are important constituents of aquatic ecosystems.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgements

This work was supported by a KAKENHI for Innovative Areas ‘Genome Science’ (No. 221S0002) from the Ministry of Education, Culture, Sports, Science and Technology (MEXT) of Japan. This study was also supported by JSPS KAKENHI Grant Number 22370005.

References

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W *et al*. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleid Acids Res* **25**: 3389–3402.

Beutler M, Milucka J, Hinck S, Schreiber F, Brock J, Mussmann M *et al*. (2012). Vacuolar respiration of nitrate coupled to energy conservation in filamentous *Beggiatoaceae*. *Environ Microbiol* **14**: 2911–2919.

Brock J, Rhiel E, Beutler M, Salman V, Schulz-Vogt HN. (2012). Unusual polyphosphate inclusions observed in a marine *Beggiatoa* strain. *Antonie Van Leeuwenhoek* **101**: 347–357.

Brock J, Schulz-Vogt HN. (2011). Sulfide induces phosphate release from polyphosphate in cultures of a marine *Beggiatoa* strain. *ISME J* **5**: 497–506.

Dahl C, Franz B, Hensen D, Kesselheim A, Zigann R. (2013). Sulfite oxidation in the purple sulfur bacterium *Allochromatium vinosum*: identification of SoeABC as a major player and relevance of SoxYZ in the process. *Microbiology* **159**: 2626–2638.

Dermott R, Legner M. (2002). Dense mat-forming bacterium *Thioploca ingrica* (*Beggiatoaceae*) in eastern Lake Ontario: implications to the benthic food web. *J Great Lakes Res* **28**: 688–697.

Friedrich CG, Rother D, Bardischewsky F, Quentmeier A, Fischer J. (2001). Oxidation of reduced inorganic sulfur compounds by bacteria: emergence of a common mechanism? *Appl Environ Microbiol* **67**: 2873–2882.

Hensen D, Sperling D, Trüper HG, Brune DC, Dahl C. (2006). Thiosulphate oxidation in the phototrophic sulphur bacterium *Allochromatium vinosum*. *Mol Microbiol* **62**: 794–810.

Høgslund S, Nielsen JL, Nielsen LP. (2010). Distribution, ecology and molecular identification of *Thioploca* from Danish brackish water sediments. *FEMS Microbiol Ecol* **73**: 110–120.

Ishihama Y, Oda Y, Tabata T, Sato T, Nagasu T, Rappsilber J *et al.* (2005). Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. *Mol Cell Proteomics* **4**: 1265–1272.

Jewell T, Huston SL, Nelson DC. (2008). Methylotrophy in freshwater *Beggiatoa alba* strains. *Appl Environ Microbiol* **74**: 5575–5578.

Kajitani R, Toshimoto K, Noguchi H, Toyoda A, Ogura Y, Okuno M *et al.* (2014). Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Res* **24**: 1384–1395.

Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M. (2014). Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res* **42**: D199–D205.

Kappler U, Bailey S. (2005). Molecular basis of intramolecular electron transfer in sulfite-oxidizing enzymes is revealed by high resolution structure of a heterodimeric complex of the catalytic molybdopterin subunit and a *c*-type cytochrome subunit. *J Biol Chem* **280**: 24999–25007.

Kappler U, Bennett B, Rethmeier J, Schwarz G, Deutzmann R, McEwan AG *et al.* (2000). Sulfite: Cytochrome *c* oxidoreductase from *Thiobacillus novellus*. Purification, characterization, and molecular biology of a heterodimeric member of the sulfite oxidase family. *J Biol Chem* **275**: 13202–13212.

Kappler U, Maher MJ. (2013). The bacterial SoxAX cytochromes. *Cell Mol Life Sci* **70**: 977–992.

Katoh K, Standley DM. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30**: 772–780.

Kojima H, Koizumi Y, Fukui M. (2006). Community structure of bacteria associated with sheaths of freshwater and brackish *Thioploca* species. *Microb Ecol* **52**: 765–773.

Kojima H, Nakajima T, Fukui M. (2007). Carbon source utilization and accumulation of respiration-related substances by freshwater *Thioploca* species. *FEMS Microbiol Ecol* **59**: 23–31.

Kojima H, Teske A, Fukui M. (2003). Morphological and phylogenetic characterizations of freshwater *Thioploca* species from Lake Biwa, Japan, and Lake Constance, Germany. *Appl Environ Microbiol* **69**: 390–398.

Lagesen K, Hallin PF, Rødland E, Stærfeldt HH, Rognes T, Ussery DW. (2007). RNammer: consistent annotation of rRNA genes in genomic sequences. *Nucleic Acids Res* **35**: 3100–3108.

Lauterborn R. (1907). Eine neue Gattung der Schwefelbakterien (*Thioploca schmidlei* nov. gen. nov. spec.). *Ber Dtsch Bot Ges* **25**: 238–242.

Lenk S, Moraru C, Hahnke S, Arnds J, Richter M, Kube M *et al.* (2012). *Roseobacter* clade bacteria are abundant in coastal sediments and encode a novel combination of sulfur oxidation genes. *ISME J* **6**: 2178–2187.

Li H, Durbin R. (2009). Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics* **25**: 1754–1760.

Li L, Stoeckert CJ Jr, Roos DS. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* **13**: 2178–2189.

Lowe TM, Eddy SR. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* **25**: 955–964.

Lucht JM, Bremer E. (1994). Adaptation of *Escherichia coli* to high osmolarity environments: osmoregulation of the high-affinity glycine betaine transport system ProU. *FEMS Microbiol Rev* **14**: 3–20.

MacGregor BJ, Biddle JF, Harbort C, Matthysse AG, Teske A. (2013a). Sulfide oxidation, nitrate respiration, carbon acquisition, and electron transport pathways suggested by the draft genome of a single orange Guaymas Basin *Beggiatoa* (*Cand*. Maribeggiatoa) sp. filament. *Mar Genomics* **11**: 53–55.

MacGregor BJ, Biddle JF, Teske A. (2013b). Mobile elements in a single-filament orange Guaymas Basin Beggiatoa sp. genome: evidence for genetic exchange with cyanobacteria. *Appl Environ Microbiol* **79**: 3974–3985.

Maier SH, Murray GE. (1965). The fine structure of *Thioploca ingrica* and a comparison with *Beggiatoa*. *Can J Microbiol* **11**: 645–663.

Maier S. (1984). Description of *Thioploca ingrica* sp. nov., nom. rev. *Int J Syst Bacteriol* **34**: 344–345.

McHatton SC, Barry JP, Jannasch HW, Nelson DC. (1996). High nitrate concentrations in vacuolate, autotrophic marine *Beggiatoa* spp. *Appl Environ Microbiol* **62**: 954–958.

McKay LJ, MacGregor BJ, Biddle JF, Albert DB, Mendlovitz HP, Hoer DR *et al.* (2012). Spatial heterogeneity and underlying geochemistry of phylogenetically diverse orange and white *Beggiatoa* mats in Guaymas Basin hydrothermal sediments. *Deep-Sea Res I* **67**: 21–31.

Mußmann M, Hu FZ, Richter M, de Beer D, Preisler A, Jørgensen BB *et al.* (2007). Insights into the genome of large sulfur bacteria revealed by analysis of single filaments. *PLoS Biol* **5**: 1923–1937.

Nelson DC, Waterbury John B, Jannasch Holger W. (1982). Nitrogen fixation and nitrate utilization by marine and freshwater *Beggiatoa*. *Arch Microbiol* **133**: 172–177.

Nemoto F, Kojima H, Fukui M. (2011). Diversity of freshwater *Thioploca* species and their specific association with filamentous bacteria of the phylum *Chloroflexi*. *Microb Ecol* **62**: 753–764.

Nemoto F, Kojima H, Ohtaka A, Fukui M. (2012). Filamentous sulfur-oxidizing bacteria of the genus *Thioploca* from Lake Tonle Sap in Cambodia. *Aquat Microb Ecol* **66**: 295–300.

Nishino M, Fukui M, Nakajima T. (1998). Dense mats of *Thioploca*, gliding filamentous sulfur-oxidizing bacteria in Lake Biwa, central Japan. *Water Res* **32**: 953–957.

Noguchi H, Taniguchi T, Itoh T. (2008). MetaGeneAnnotator: detecting species-specific patterns of ribosomal binding site for precise gene prediction in anonymous prokaryotic and phage genomes. *DNA Res* **15**: 387–396.

Pott AS, Dahl C. (1998). Sirohaem sulfite reductase and other proteins encoded by genes at the *dsr* locus of

Chromatium vinosum are involved in the oxidation of intracellular sulfur. *Microbiology* **144**: 1881–1894.

Prokopenko MG, Hirst MB, De Brabandere L, Lawrence DJ, Berelson WM, Granger J *et al.* (2013). Nitrogen losses in anoxicmarine sediments driven by *Thioploca*–anammox bacterial consortia. *Nature* **500**: 194–198.

Raes J, Korbel JO, Lercher MJ, von Mering C, Bork P. (2007). Prediction of effective genome size in metagenomic samples. *Genome Biol* **8**: R10.

Salman V, Amann R, Girnth AC, Polerecky L, Bailey JV, Høgslund S *et al.* (2011). A single-cell sequencing approach to the classification of large, vacuolated sulfur bacteria. *System Appl Microbiol* **34**: 243–259.

Salman V, Bailey JV, Teske A. (2013). Phylogenetic and morphologic complexity of giant sulphur bacteria. *Antonie Van Leeuwenhoek* **104**: 169–186.

Schulz HN, Schulz HD. (2005). Large sulfur bacteria and the formation of phosphorite. *Science* **307**: 416–418.

Sleator RD, Colin H. (2001). Bacterial osmoadaptation: the role of osmolytes in bacterial stress and virulence. *FEMS Microbiol Rev* **26**: 49–71.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. (2011). MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol* **28**: 2731–2739.

Watanabe T, Kojima H, Fukui M. (2012). Draft genome sequence of a psychrotolerant sulfur-oxidizing bacterium, *Sulfuricella denitrificans* skB26, and proteomic insights into cold adaptation. *Appl Environ Microbiol* **78**: 6545–6549.

Whelan S, Goldman N. (2001). A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol Biol Evol* **18**: 691–699.

Winkel M, de Beer D, Lavik G, Peplies J, Mußmann M. (2013). Close association of active nitrifiers with *Beggiatoa* mats covering deep-sea hydrothermal sediments. *Environ Microbiol* **16**: 1612–1626.

Wislouch SM. (1912). *Thioploca ingrica* nov. sp. *Ber Deutsch Bot Ges* **30**: 470–474.

Zemskaya TI, Namsaraev BB, Dul'tseva NM, Khanaeva TA, Golobokova LP, Dubinina GA *et al.* (2001). Ecophysiological characteristics of the mat-forming bacterium *Thioploca* in bottom sediments of the Frolikha Bay, northern Baikal. *Microbiology* **70**: 335–341.

Zemskaya TI, Chernitsyna SM, Dul'tseva NM, Sergeeva VN, Pogodaeva TV, Namsaraev BB. (2009). Colorless sulfur bacteria Thioploca from different sites in Lake Baikal. *Microbiology* **78**: 117–124.

Zopfi J, Kjær T, Nielsen LP, Jørgensen BB. (2001). Ecology of *Thioploca* spp.: nitrate and sulfur storage in relation to chemical microgradients and influence of *Thioploca* spp. on the sedimentary nitrogen cycle. *Appl Environ Microbiol* **67**: 5530–5537.

Supplementary Information accompanies this paper on The ISME Journal website (http://www.nature.com/ismej)