frontiers in Microbiology

ORIGINAL RESEARCH
published: 03 May 2016
doi: 10.3389/fmicb.2016.00603

Single-Cell (Meta-)Genomics of a Dimorphic *Candidatus* Thiomargarita nelsonii Reveals Genomic Plasticity

Beverly E. Flood[1], Palmer Fliss[1†], Daniel S. Jones[1,2], Gregory J. Dick[3], Sunit Jain[3], Anne-Kristin Kaster[4], Matthias Winkel[5], Marc Mußmann[6] and Jake Bailey[1]*

[1] Department of Earth Sciences, University of Minnesota, Minneapolis, MN, USA, [2] Biotechnology Institute, University of Minnesota, St. Paul, MN, USA, [3] Department of Earth and Environmental Sciences, University of Michigan, Ann Arbor, MI, USA, [4] German Collection of Microorganisms and Cell Cultures, Leibniz Institute DSMZ, Braunschweig, Germany, [5] Helmholtz Centre Potsdam, GFZ German Research Centre for Geosciences, Potsdam, Germany, [6] Max Planck Institute for Marine Microbiology, Bremen, Germany

OPEN ACCESS

Edited by:
Andreas Teske, University of North Carolina at Chapel Hill, USA

Reviewed by:
Craig Lee Moyer, Western Washington University, USA
Jeremy Dodsworth, California State University, San Bernardino, USA

***Correspondence:**
Beverly E. Flood
beflood@umn.edu

†Present Address:
Palmer Fliss, Molecular Characterization and Clinical Assay Development Laboratory, Leidos Biomedical Research Inc., and Frederick National Laboratory for Cancer Research, Frederick, Maryland

Specialty section:
This article was submitted to Extreme Microbiology, a section of the journal Frontiers in Microbiology

Received: *29 January 2016*
Accepted: *11 April 2016*
Published: *03 May 2016*

Citation:
Flood BE, Fliss P, Jones DS, Dick GJ, Jain S, Kaster A-K, Winkel M, Mußmann M and Bailey J (2016) Single-Cell (Meta-)Genomics of a Dimorphic Candidatus Thiomargarita nelsonii Reveals Genomic Plasticity. Front. Microbiol. 7:603. doi: 10.3389/fmicb.2016.00603

The genus *Thiomargarita* includes the world's largest bacteria. But as uncultured organisms, their physiology, metabolism, and basis for their gigantism are not well understood. Thus, a genomics approach, applied to a single *Candidatus* Thiomargarita nelsonii cell was employed to explore the genetic potential of one of these enigmatic giant bacteria. The *Thiomargarita* cell was obtained from an assemblage of budding *Ca.* T. nelsonii attached to a provannid gastropod shell from Hydrate Ridge, a methane seep offshore of Oregon, USA. Here we present a manually curated genome of Bud S10 resulting from a hybrid assembly of long Pacific Biosciences and short Illumina sequencing reads. With respect to inorganic carbon fixation and sulfur oxidation pathways, the *Ca.* T. nelsonii Hydrate Ridge Bud S10 genome was similar to marine sister taxa within the family *Beggiatoaceae*. However, the Bud S10 genome contains genes suggestive of the genetic potential for lithotrophic growth on arsenite and perhaps hydrogen. The genome also revealed that Bud S10 likely respires nitrate via two pathways: a complete denitrification pathway and a dissimilatory nitrate reduction to ammonia pathway. Both pathways have been predicted, but not previously fully elucidated, in the genomes of other large, vacuolated, sulfur-oxidizing bacteria. Surprisingly, the genome also had a high number of unusual features for a bacterium to include the largest number of metacaspases and introns ever reported in a bacterium. Also present, are a large number of other mobile genetic elements, such as insertion sequence (IS) transposable elements and miniature inverted-repeat transposable elements (MITEs). In some cases, mobile genetic elements disrupted key genes in metabolic pathways. For example, a MITE interrupts *hupL,* which encodes the large subunit of the hydrogenase in hydrogen oxidation. Moreover, we detected a group I intron in one of the most critical genes in the sulfur oxidation pathway, *dsrA.* The *dsrA* group I intron also carried a MITE sequence that, like the *hupL* MITE family, occurs broadly across the genome. The presence of a high degree of mobile elements in genes central to *Thiomargarita*'s core metabolism has not been previously reported in free-living bacteria and suggests a highly mutable genome.

Keywords: *Thiomargarita*, single-cell genomics, arsenite oxidation, intron, mobile genetic elements, genome instability, miniature inverted-repeat transposable elements, metacaspase

INTRODUCTION

The family *Beggiatoaceae* include the largest known free-living bacteria with some marine *Thiomargarita* spp. achieving millimetric cell diameters (Bailey et al., 2009; Salman et al., 2011). These bacteria are chemolithotrophs that obtain energy for metabolism from the oxidation of reduced sulfur species. *Thiomargarita* spp. are thought to primarily use the oxidation of electron donors available in the sediment pore waters to fuel carbon fixation. The terminal electron acceptors used in these reactions can vary. Besides oxygen, nitrate can be used as a terminal electron acceptor in large, vacuolated sulfur bacteria under anoxic conditions, and *Thiomargarita* spp. allocates up to 90% of its volume for intracellular nitrate storage (Schulz and Jørgensen, 2001). *Thiomargarita* spp. have also been shown to accumulate intracellular elemental sulfur inclusions that serve as intermediates in the oxidation of hydrogen sulfide to sulfate, to provide the cell with electron donors when access to sulfide is limited (Schulz et al., 1999). Prior research has also demonstrated that *Thiomargarita* spp. are capable of accumulating phosphate intracellularly as long polyphosphate (poly-p) polymers. The hydrolysis of this polyphosphate, and concomitant release of phosphate into pore water has been linked to the formation of large phosphorite deposits in the seafloor and the subsurface (Schulz and Schulz, 2005; Bailey et al., 2006, 2013; Goldhammer et al., 2010; Crosby and Bailey, 2012; Dale et al., 2013). However, the stimuli and mechanisms for polyphosphate accumulation and release of inorganic phosphorous have not been fully elucidated (Brock and Schulz-Vogt, 2011).

Marine cold seeps are sites where elevated hydrocarbon fluxes fuel the production of sulfide via the anaerobic oxidation of methane and sulfate reduction (Boetius et al., 2000). Dimorphic *Thiomargarita* ecotypes of *Candidatus* Thiomargarita nelsonii initially discovered at seeps along the Costa Rican Margin were later observed at Hydrate Ridge. These ecotypes are commonly found attached to substrates, particularly the shells of provannid snails, **Figure 1A**. These attached cells appeared to undergo a dimorphism (elongate vs. budding) in their life cycle, wherein *Thiomargarita* elongated almost a millimeter in length, and budded spherical daughter cells from the distal end (**Figure 1B**; Bailey et al., 2011). Several of the gastropod samples collected showed a distinct epibiont community resembling morphotypes similar to *Ca.* T. nelsonii as well as *Marithrix* sp. and *Leucothrix* sp. (Salman et al., 2013). Here we report on a new draft genome assembly of a single cell of *Ca.* T. nelsonii Hydrate Ridge budding from an attached *Thiomargarita* cell that we refer to as "Bud S10." Despite washing the bud, several adherent bacteria were not removed, thus a metagenome was assembled using Illumina and Pacific Biosciences sequencing reads that were binned and manually-curated to produce the genome presented here.

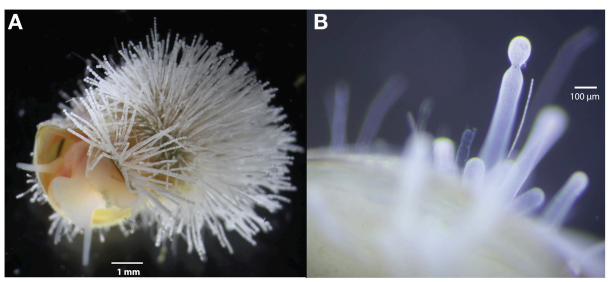

FIGURE 1 | (A) *Provanna* sp. snail with attached *Thiomargarita* epibiont community. **(B)** An elongated *Thiomargarita* morphotype attached to provannid snail that is budding from the distal end. Bud S10 was derived from this population of actively budding daughter cells.

The draft genome revealed many comparable metabolic pathways to those found in sister taxa *Ca.* Maribeggiatoa Guaymas Basin orange filament (BOGUAY) (MacGregor et al., 2013a) and in a *Ca.* T. nelsonii phylotype from Namibia (Winkel et al., in review). However, we found a number of surprising genomic features, including genes that suggest the potential for lithotrophic growth using arsenite as an electron donor. In addition, we found genes encoding two complete nitrate reduction pathways, one that terminates in N_2 and another that produces NH_4^+, as predicted by metabolic studies of giant marine sulfur bacteria (Otte et al., 1999; Høgslund et al., 2009; Prokopenko et al., 2013), but incomplete in previously sequenced genomes (Mußmann et al., 2007; MacGregor et al., 2013a). Surprisingly, the genome displayed a high number of mobile genetic elements that we describe here. These mobile elements include group I and II introns, transposons, and miniature inverted-repeat transposable elements (MITEs). Interestingly, key genes in sulfur and hydrogen oxidation pathways were disrupted by mobile elements. Furthermore, we describe a new form of molecular symbiosis between a group I intron and MITE.

MATERIALS AND METHODS

Site Description and Sample Collection

During a research expedition on board the R/V *Atlantis* (AT18_1) to Hydrate Ridge North (44° 40.02687′ N 125° 5.99969′ W), Cascadian margin, off the coast of Oregon, USA the remotely operated vehicle *Jason* collected methane seep samples containing provannid gastropods. The gastropod exteriors hosted dense biofilms on the posterior surface of their shells. These biofilms included attached *Thiomargarita*-like bacteria similar to those described by Bailey et al. (2011).

These snails were fixed shipboard in a 1:1 mixture of sterile ethanol and Instant Ocean® (Spectrum Brands, USA) and then stored at −20°C for molecular analyses.

Genomic DNA Amplification, Sequencing, and Assembly

An ethanol fixed gastropod and attached community was examined under an Olympus SZX-16 stereo microscope. Large bacterial morphotypes were individually removed from the host via a pipet. The genome was obtained from a single *Ca.* T. nelsonii bud (sample S10) that detached from an elongated attached *Ca.* T. nelsonii cell. The cell was placed in a 40 µm sterile cell strainer (BD Biosciences, San Diego, CA) and rinsed seven times in DNA-free water with 3.5% NaCl and 50% ethanol. Despite our best effort to clean the *Thiomargarita* cell, some adherent bacterial epibionts were not removed. Whole

genome amplification was performed using a RePLI-g midi-kit (Qiagen, USA) according to the manufacturer's instructions. DNA sequencing was performed on a Illumina HiSeq 2000 (CASAVA 1.8), which yielded 85 million 100 base pair (bp) paired end reads with 300 bp inserts (Truseq v1 chemistry). A previous metagenome assembly was performed using these Illumina sequencing reads (Fliss, 2014). Here, we included additional DNA sequencing using 12 Pacific Biosciences RS SMRT cells prepared with P4-C2 chemistry. Seven SMRT cells contained fragmented DNA of ∼600 bases for obtaining high quality circular consensus sequencing (CCS) reads and five SMRT cells for continuous long read (CLR) sequencing of unfragmented DNA. A combination of cutadapt (Martin, 2011), Sickle (v.1.29) (Joshi and Fass, 2011) and Prinseq-lite (v.0.20.4) (Schmieder and Edwards, 2011) were used to error correct the Illumina reads (sliding window q20 at the 3′ end, sequences containing N's removed, average read q30, de-replication of six or more exact duplicates and minimum read length of 50 bases). The PacBio CCS reads were filtered (RS_Subreads.1) using the SMRT Analysis (v.2.20) software package (http://www.pacb.com/devnet/) and concatenated with Illumina unpaired reads generated during the quality filtering and trimming step. Additionally, sequence data from all 12 SMRT cells were filtered and were used for scaffolding and were flagged as "uncorrected PacBio" reads. A hybrid assembly was performed using SPAdes 3.1 with "assembler only" mode (no pre-assembly error correction) with kmer sizes of 21, 35, 55, and 75 and with post-assembly mismatch correction. The final assembled metagenome was 31.68 Mbp (3326 contigs) with an N50 value of 67,259 bp.

Community Assessment, Tetramer-Frequency Binning, and Curation

In preparation for tetranucleotide binning, rRNA gene sequences in the metagenomes were analyzed to identify constituent strains in the dataset. Putative rRNA gene fragments in the metagenome were first identified by comparing unassembled quality-filtered Illumina reads against the Silva small subunit (SSU) rRNA reference database (version 115) (Quast et al., 2013) using BLASTN (McGinnis and Madden, 2004). SSU rRNA gene sequence matches were then extracted from the Illumina dataset and assembled using EMIRGE (Miller et al., 2011). The EMIRGE assembly yielded five 16S rRNA sequences. The best BLASTN matches for these five sequences were *Pseudoalteromonas* sp. M12-11A FN377706 (representing 41% of rRNA genes according to EMIRGE), *Colwellia* sp. BCw110 FJ889596 (26%), an uncultured *Colwellia* AY375054 (18%), *Ca.* T. nelsonii (9%), and *Neptuniibacter* sp. CAR-SF AB086227 (5%). Separately, a 16S rRNA gene and intergenic region with 100% identity to clone *Ca.* T. nelsonii HYR001 (GenBank accession number HF954113) (Salman et al., 2013) was amplified and sequenced from the RePLI-g-processed DNA using *Thiomargarita*-specific PCR primers (VSO233F—ITSReub) (Salman et al., 2011).

A tetramer-frequency based Emergent Self-Organizing Map (tetra-ESOM) was constructed based on the contigs generated, following the protocol outlined in Ultsch and Mörchen (2005), Dick et al. (2009). The genomes of the following strains were included in the tetra-nucleotide training and ESOM binning: *Candidatus* Beggiatoa sp. Orange Guaymas (NCBI project ID: PRJNA19285, Locus Tag BOGUAY), *Colwellia piezophila* BAA-637 (PRJNA182419, F580), *Colwellia psychrerythraea* 34H (PRJNA275, CPS), *Neptuniibacter caesariensis* MED92 (PRJNA13561, MED92), *Pseudoalteromonas* SM9913 (PRJNA39311, PSM), and *Candidatus* Thiomargarita sp. Thio36 (PRJNA79059, Thi036) from Namibian coastal upwelling sediments. Only contigs greater than 2 kb were used in the tetra-ESOM analyses. The network was trained with a K-Batch algorithm, 170 rows and 354 columns (∼60,180 neurons), and a starting radius value of 50. In total, 543 metagenome contigs overlapped or neighbored the genomes of BOGUAY and Thio36 and were selected for the *Thiomargarita* bin. Subsequently, 32 contigs were removed from the bin based on coverage below 20x and/or less than 10% of the contig fell within the bin. Additional analyses using kmer coverage on the binned genome submitted to IMG/ER and a kmer analysis (vs. 4.5) was performed (oligomer size 4, fragment window 5000, fragment step 500) in addition to manual curation and comparison to the *Ca.* T. nelsonii Thio36 genome. These analyses resulted in the removal of 72 additional contigs including high coverage contigs that contains plasmid related genes.

Annotation and Bioinformatics

Annotation was performed by the IMG/ER gene prediction pipeline (Markowitz et al., 2009). Since IMG was the primary tool used for genome analyses, IMG gene and contig notations are used herein. After assessing the genome, we found that several genes were absent from the hybrid assembly (contigs greater than 1500 bp) but present in the original non-hybrid assembly (IMG Gold ID Ga0097846). The genes were located on contigs Ga0097846_10092 (8217 bp), Ga0097846_10099 (2891 bp), Ga0097846_11126 (14,449 bp), Ga0097846_11134 (10,846 bp), Ga0097846_11318 (11,275 bp), (Ga0097846_11205 (7845 bp), and Ga0097846_12056 (5245 bp). These genes were included in the assessment of metabolic pathways below. Alignments and comparison of group I Intron containing sequences was performed using UGENE (Okonechnikov et al., 2012). Repetitive sequences with 15 or more occurrences were identified using the RepeatModeler (Smit and Hubley, 2008-2015). Group II introns were identified using the Database for Bacterial group II Introns (Candales et al., 2011). Repetitive elements identified by RepeatModeler and putative introns where compared with sequences in the RNA families database (Rfam) (Nawrocki et al., 2014) and examined for protein-encoding domains with the NCBI Conserved Domain Database (CDD) (Marchler-Bauer et al., 2014). MITE sequences were compared with IS elements deposited in ISfinder database (https://www-is.biotoul.fr) (Siguier et al., 2006). All BLASTN analyses against Bud S10's genome were performed with a cutoff *e*-value of $10e^{-5}$.

Arsenite oxidoreductase (AioA) peptide sequences were aligned with the Expresso algorithm in T-Coffee using default parameters (Armougom et al., 2006), and the alignment trimmed with the "automated1" option in TrimAL (Capella-Gutiérrez et al., 2009) at the Phylomon2 online workbench (Sánchez et al.,

2011). The final alignment length was 885 amino acid positions. Maximum likelihood analyses were performed with RAxML version v 8.0.24 (Stamatakis, 2006) with 1000 rapid bootstrap replicates. For maximum likelihood analysis, the LG amino acid substitution model (Le and Gascuel, 2008) with proportions of invariant sites, base frequencies, and the alpha parameter estimated from the data, was selected by the AICc in ProtTest v.2.4 (Abascal et al., 2005).

PCR Confirmation of Disrupted Genes

PCR was employed to confirm the sequences of the *dsrA*, *hupL,* and *hupS* genes. We included a second *Ca.* T. nelsonii metagenome sample, obtained from the same gastropod-attached community, in the confirmation screenings for the purposes of replication. PCR primers were designed using Primer3 (Untergasser et al., 2012). To capture most of the *dsrA* gene to include the intron sequence (bases 44-1320) the primers dsrA_TmargS10_F 5′-GAGTGGTCCTTGGCCTAGTT-3′ and dsrA_tmargS10_R 5′-GGGGACAGGGCTTTAGTCAT-3′ were used. Two primer sets were used to cover a portion of the *hupL* gene, the MITE sequence, the frameshift and the *hupS* gene. The primers S10_hupS-F 5′-ATGGATACAAACCGGTGCTT-3′ and S10-hupS-R 5′-GTCATGATGTTCCGCATAGC-3′ covered bases 1296...204 and the primers S10_hupSL-F 5′-ATGGATACAAACCGGTGCTT-3′ and S10_hupSL-R 5′-GTCATGATGTTCCGCATAGC-3′ covered based 1941...2624 in contig Ga0063879_1132. PCRs were performed using GoTaq® polymerase (Promega) with the addition of 10% dimethyl sulfoxide to stabilize hairpin structures.

Nucleotide Sequence Accession Numbers

The draft genome has been deposited with the National Center for Biotechnology Information, BioProject PRJNA266451 (*Candidatus* Thiomargarita nelsonii Hydrate Ridge). Raw sequence reads were deposited in the NCBI Short Read Archive. The binned Bud S10 genome was annotated and is publically available through the Joint Genome Institute's IMG website, Gold ID Ga0097846 (non-hybrid assembly) and Ga0063879 (hybrid assembly). All contigs greater than 1500 bps of the pre-binned metagenome hybrid assembly were annotated by IMG and are publically available through IMG/M, Gold analysis ID Ga0064232.

RESULTS

Comparison to Previous Genome Assembly and Genome Completeness

Prior to producing the hybrid assembly reported on here, an assembly using only the Illumina reads was performed with the metagenome assembler MetaVelvet (Namiki et al., 2012) followed by Mimimus 2 (Sommer et al., 2007). The assembly yielded 144,811 contigs (N50 = 2571), and the *Thiomargarita* bin derived from this initially assembly contained 2497 contigs greater than 2000 bps in size (12.97 MB). Subsequently, the same Illumina reads were co-assembled with PacBio CCS reads and then scaffolded with PacBio CLR reads. This hybrid metagenome assembly resulted in 3326 contigs of which 2370 contigs were greater than 1500 bps long and were used in the ESOM binning (**Supplementary Material 1**). The final *Ca.* Thiomargarita bin reported on here contains 439 contigs (7.71 MB) (N50 scaffold length = 36,326) with an average of 145-fold genome coverage. Only the contents of this *Ca.* Thiomargarita bin are discussed in the remainder of this report. Of the 7525 protein encoding genes, 56.71% had a predicted function.

The IMG pipeline assessed the binned genome as 86.53% complete. We further assessed genome completeness by examining the ribosomal polymerase, rRNA genes, tRNA, tRNA synthase genes and the ribosomal proteins. The binned genome contains *rpoA,* and two copies of *rpoB* (one co-located with *rpoC*) and one set of rRNA genes in a single operon. The contig containing this operon terminates in an incomplete 16S rRNA gene. The 23S rRNA gene is interrupted by three introns at positions 2038..2823, 2838..3111, and 3135..3851. Sister taxa "*Candidatus* Maribeggiatoa Orange Guaymas Basin" (hereafter referred to as BOGUAY) (MacGregor et al., 2013c) and "*Candidatus* Thiomargarita sp. NAM92" also possess introns in their 23S rRNA gene; however, neither have an intron in the third position (3135..3851). The resulting full-length 23S rRNA gene in Bud S10 is 5183 bases long. These introns did not match with introns in the Rfam database and further analyses regarding the phylogeny of the introns and their capacity to be self-catalytic was not determined.

A total of 46 tRNA genes are present and cover the 20 standard amino acids, except as in BOGUAY, tRNA-Arg-TCT and tRNA-Leu-TAA were missing from the annotated genes. MacGregor et al. (2013a) determined that they were present in BOGUAY but interrupted by putative group I introns. BLASTN analyses of these interrupted tRNAs in BOGUAY against the Bud S10 genome revealed similar interrupted tRNAs (tRNA-Arg-TCT contig Ga0063879_1007, bases 79335..79658) (tRNA-Leu-TAA contig Ga0063879_1068, bases 17090..17442). All tRNA synthetase genes were present except threonyl-tRNA synthetases and the glutminyl-tRNA was interrupted by two stop codons (Ga0063879_05998-6000). The binned genome also contains all of the ribosomal proteins with the exception of *rplT* (L20) and *rmpI* (L35). *rplD* (L30) was not annotated by the IMG pipeline, nor were we able to find it in the metagenome assemblies. L30 is a non-essential protein (Akanuma et al., 2012) but is found in most bacteria. However, the absence of L30 has been noted in candidate phyla with small genomes and self-splicing introns in their ribosomal rRNA (Brown et al., 2015), and in some host dependent strains, some cyanobacteria and the *Planctomycetes–Verrucomicrobia–Chlamydiae* superphylum (Lecompte et al., 2002; Yutin et al., 2012; Lagkouvardos et al., 2014).

Major Metabolic Pathways

Carbon Acquisition

The Bud S10 genome contains putative genes for the Calvin Benson Bassham-pathway with RuBisCO Form II (*cbbL*) performing the catalytic step of fixing CO_2. Like the BOGUAY genome, Bud S10 does not possess genes encoding fructose 1,6-bisphosphatase or sedoheptulose 1,7-bisphosphatase. MacGregor et al. (2013a) postulated that these functions may be dependent on a pyrophosphate (PPi)-dependent 6-phosphofructokinase

(*pfkA*) as it has been proposed for some gammaproteobacterial endosymbionts (Kleiner et al., 2012; MacGregor et al., 2013a). As in the BOGUAY genome, there were two putative genes encoding *pfkA* (Ga0063879_01082, Ga0063879_04320). The genome also contains the genes for a complete reverse TCA cycle including the three key genes 2-oxoglutarate ferredoxin oxidoreductase (*korAB*) (Ga0063879_04149-51), pyruvate oxidoreductase (*porABCD*) (Ga0063879_05883), and ATP citrate lyase (*aclAB*) (Ga0063879_05528-9). Bud S10 may not have a complete TCA cycle. The genome has a 2-oxoglutarate dehydrogenase, but a stop codon is present within the gene (Ga0063879_04055-6). Bud S10 also lacks a glyoxylate shunt as well as the genes for carboxysome synthesis. The lack of a viable 2-oxoglutarate dehydrogenase may be indicative of an obligate autotrophic lifestyle (Wood et al., 2004). However, Bud S10 possesses genes for ABC-type amino acid and peptide transporters, in addition to a dicarboxylate transporter which are suggestive of a mixotrophic lifestyle (Schulz and de Beer, 2002). The genetic potential for methylotrophy and alkane degradation (alkane 1-monoxygenase), common in hydrocarbon seep bacteria, was not observed in the genome.

Respiration, Chemotaxic Motility, and Proton Motive Force

Only respiratory pathways involving O_2 and inorganic nitrogen as terminal electron acceptors were detected in the genome. In Bud S10, aerobic respiration terminates with a cbb3-type cytochrome, which has a high affinity for O_2 and is known to be specific to strains adapted to microaerophilic environments (Molinas et al., 2011). A catalase, a key enzyme for handling oxidative stress, was not found, nor did we find it in other marine *Beggiatoaceae*. However, Bud S10 possesses an operon of other genes for reducing oxidative stress including superoxide dismutase, desulfoferrodoxin, and cytochrome c peroxidase. *Thiomargarita* spp. are not known to be motile via flagella or gliding motility; however a rolling movement has been observed in some ecotypes (Salman et al., 2011). The genome of Bud S10 contains genes associated with twitching motility (*pilGHIJTU*) as well as those associated with a chemotactic response to a stimulus such as O_2 (*cheABRWY* and a methyl-accepting chemotaxic protein).

The genome of Bud S10 has the complete Complex I for oxidative phosphorylation (NADH subunits A-N), two operons encoding succinate dehydrogenases, and two operons encoding F-type ATPases. The genes encoding a vacuolar type H^+ translocating ATPase (ntpABCDEFIK) are also present. A V-ATPase has been shown to be responsible for generating a proton-motive force across the membrane of a marine *Beggiatoa*'s nitrate vacuole resulting in intra-vacuolar acidity (Beutler et al., 2012). Additionally, the genome of Bud S10 had two copies of the genes of the rnf complex (*rnfABCDEFG*). The rnf complex is a Na^+-pumping ferredoxin:NAD^+ oxidoreductase that generates a chemiosmotic gradient of Na^+ for ATP production.

Analysis of the Bud S10 dataset revealed two complete pathways for the dissimilatory reduction of nitrate. NO_3^- may be reduced to NO_2^- in the periplasm via a periplasmic nitrate reductase (*napABCDFGH*) and via a membrane bound nitrate reductase (*narGHIJ*), for which there are two complete operons, each containing a putative cytochrome *c*-like gene. Nitrite may then either be reduced to NH_4^+ via a *nirBD*-type nitrate reductase or to N_2 via a *nirS*-type nitrite reductase followed by a nitric oxide reductase (*norBCDQ*), and finally a nitrous oxide reductase (*nosZ*) ($NO_2^- > NO > N_2O > N_2$). The octaheme cytochrome c reductase (BOGUAY_0691) that was shown to have nitrite reductase, hydroxylamine oxidase, and hydrazine oxidase activities in the BOGUAY (MacGregor et al., 2013a,b), was identified in the genome of Bud S10 (Ga0063879_0044), as were other multi-heme cytochromes of unknown function (Ga0063879_04661 and Ga0063879_05965). Several inorganic nitrogen transporters were also identified: a nitrate:nitrite antiporter (*narK*) (Ga0063879_02276), a formate/nitrite transporter (Ga0063879_00790) and two ammonia transporters (Ga0063879_04108, Ga0063879_04110).

Lithotrophy

The genome of Bud S10 indicated that this strain could utilize a number of reduced inorganic sulfur substrates, arsenite (AsO_3^-) and potentially H_2 as electron donors. The sulfide oxidation pathways, operon structure and gene copy numbers were mostly consistent with vacuolated *Beggiatoaceae*. The oxidation of H_2S could occur via a sulfide:quinone oxidoreductase, and/or flavocytochrome c, present as one and two gene copies respectively. Thiosulfate oxidation could occur via the sox system (*soxABXYZ*) co-located on an operon with *qmoABC*. Elemental sulfur (cyclooctasulfur), an intermediate of sulfide oxidation, could be oxidized to SO_3^{2-} via the complete oxidative dissimilatory sulfite reductase system (*dsr*ABCEFHKJLMOP). There is a group II intron that is downstream of *dsrABEF,* which neighbors on the opposite strand *dsrL*, the gene for a sulfur relay protein, **Figure 2**. The other *dsr* genes are on another contig. Another distinctive feature is that *dsrA* (Ga0063879_66054-5)

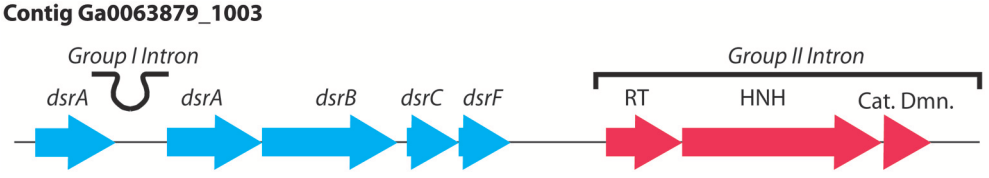

FIGURE 2 | Oxidative dissimilatory sulfite reductase operon with a group I intron interrupting the *dsrA* gene and a group II intron containing accessory genes that facilitate the intron's insertion into the genome.

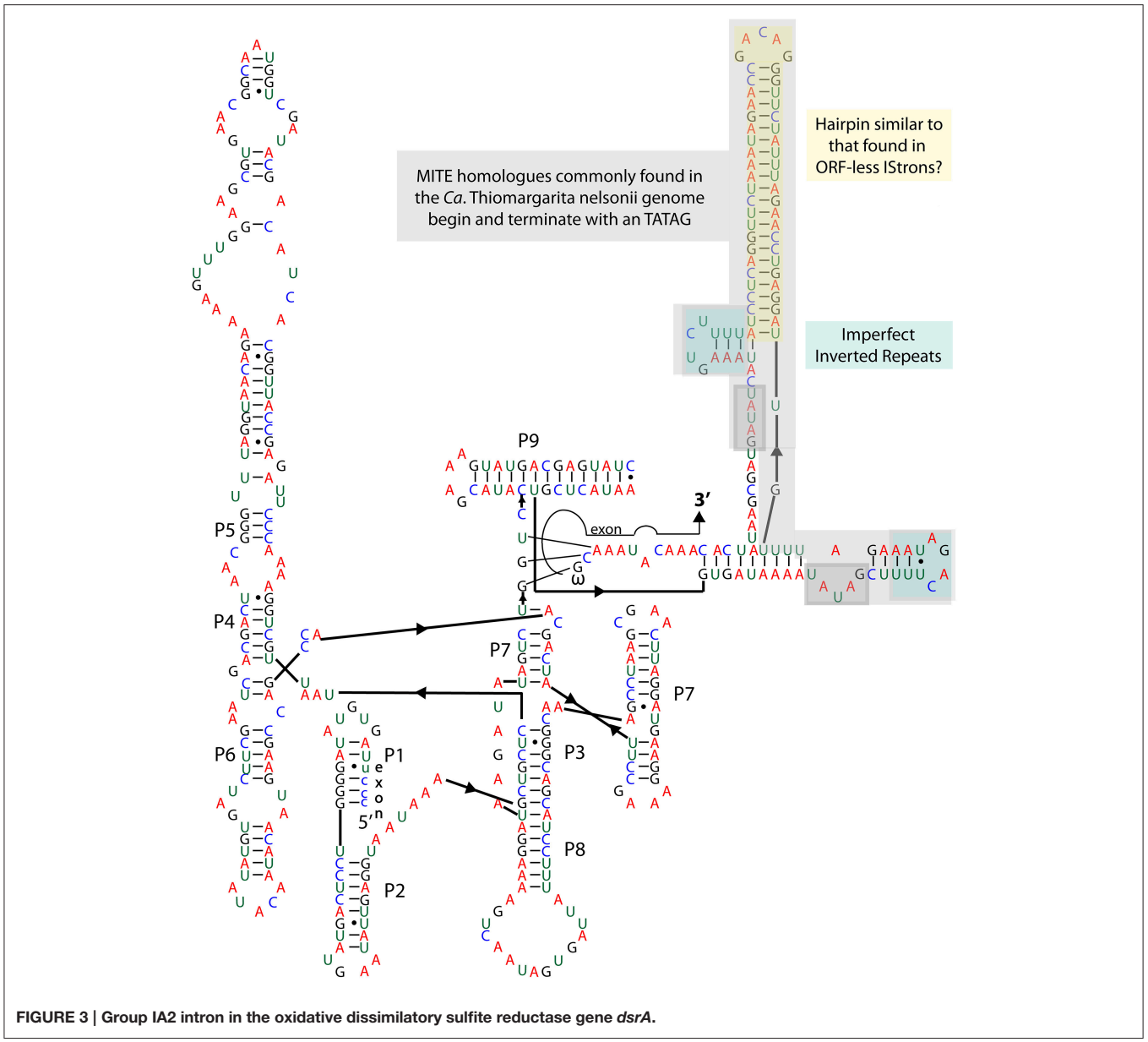

FIGURE 3 | Group IA2 intron in the oxidative dissimilatory sulfite reductase gene *dsrA*.

was found to be interrupted by a group I intron (see analysis below, **Figures 2**, **3**). In addition to cyclooctasulfur, polysulfides were recently identified as intermediates in sulfide oxidation in *Beggiatoa* (Berg et al., 2014). Bud S10 may utilize the polysulfides oxidatively or perhaps reductively via a polysulfide reductase-like operon (*phsABC*) (Eddie and Hanson, 2013; Weissgerber et al., 2013). Bud S10 had the potential to further reduce SO_3^{2-}, either via a sulfite dehydrogenase (*sorA*) or an adenylylsulfate reductase (*aprAB*) and sulfate adenylyltransferase (*sat*). Other operons of unknown function contain genes for putative lithotrophic growth on reduced sulfur. For example, an operon Ga0063879_02957–Ga0063879_02962 contains a gene similar to the polysulfide reductase membrane anchoring protein (*phsC*), along with a number of cytochromes, an ATPase and hypothetical proteins. A number of operons also contain genes for rhodanese-like enzymes which are thought to play a role in lithotrophic growth on certain sulfur species (Weissgerber et al., 2013).

Thiomargarita spp. have not been previously shown to utilize hydrogen as an electron donor. However, Bud S10 appears to possess genes for a Ni-Fe hydrogenase (*hupL* Ga0063879_05378-9; *hupS* Ga0063879_05380) and some hydrogenase accessory proteins *hypABCEF* on a separate contig (Ga0063879_1018). Additionally, 19 putative genes (mostly hypothetical genes) separated *hypE* from the cluster of other accessory genes. However, *hupL* was interrupted by a miniature inverted-repeat transposable element (MITE) (base positions 1357..1561(-), see below and **Supplementary Material 2**). Regardless of the MITE, neither *hupS,* nor *hupL,* appear to be complete genes or they may be fused. This disruption or gene fusion occurred upstream of the MITE sequence. PCR products of the *hupSL*

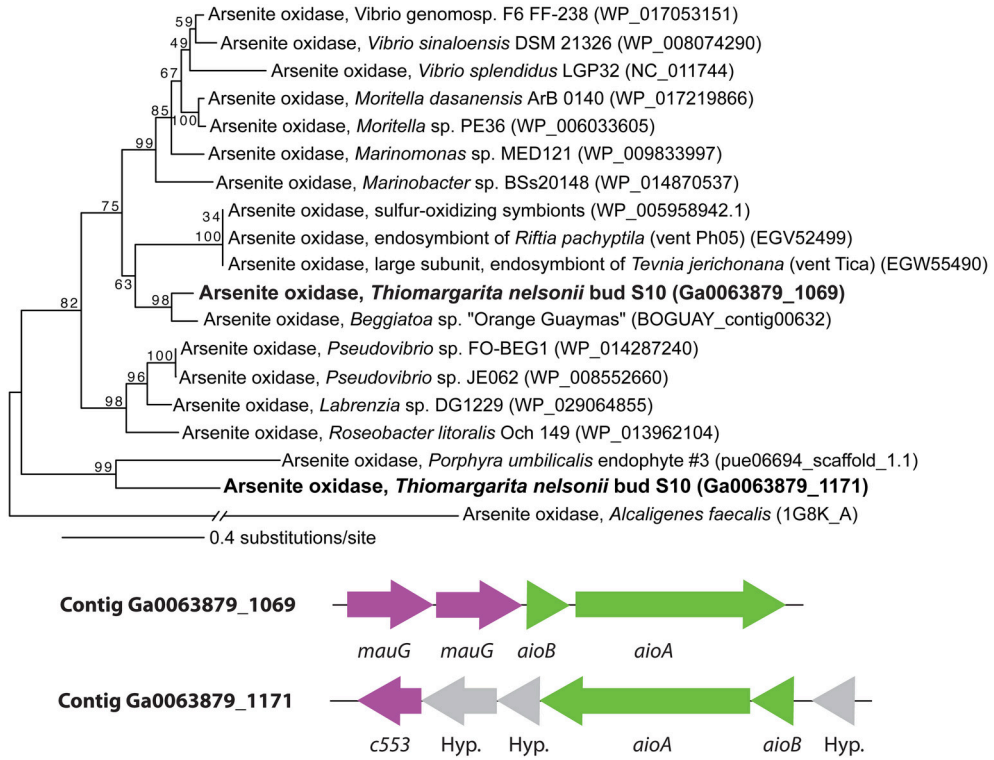

FIGURE 4 | The arsenite oxidoreductase operons and a maximum likelihood tree of arsenite oxidoreductase subunit *aioA*. Numbers indicate bootstrap support for each node.

region of the whole-genome-amplified DNA of both Bud S10 and an additional single cell sample were consistent with the assembled genome. No additional PCR products were detected, thus it appears there are not multiple versions of the region, which could occur in a polyploid organism. A BLASTX analysis against GenBank of the *hupS* and *hupL* sequences with the MITE and frameshifts removed indicated that the likely class of the *hupSL* is the newly described group 2c hydrogenase (top BLASTX score for both genes was *Methylobacter tundripaludum, hupS* WP_031435956 *e*-value $3e^{-79}$, 59% identity, *hupL*, WP_027150633, *e*-value $9e^{-100}$, 49% identity) (Greening et al., 2015). We examined the *hupL* sequence for the two defining metal ligating cysteine residues, L1 and L2 in NiFe hydrogenases (Vignais and Billoud, 2007; Greening et al., 2015). The L1 motif is absent but the L2 motif is most consistent with a group 2c hydrogenase (consensus group 2c: SFDxCLVCTVH; *Thiomargarita*: SHDaCLVCTVH). Group 2 hydrogenases are cytosolic. Group 2a hydrogenases in the Cyanobacteria provide electrons under aerobic conditions and recycle H_2 generated by cellular processes, while group 2b are thought to be hydrogen sensors and thus, perform a role in cellular regulation. Group 2c hydrogenases were, until recently, classified as group 2a. However, group 2c hydrogenases have distinct L1 and L2 motifs and may be co-transcribed with diguanylate cyclases/phosphodiesterases, which are proposed to regulate global transcription based on fluctuating H_2 conditions (Greening et al., 2015). These hydrogenases are uncommon and mostly restricted to aquatic sulfate-reducing bacteria and methylotrophic bacteria. Since *Thiomargarita*'s *hupSL* genes are incomplete perhaps by deletion, gene rearrangement or fusion, we cannot perform full phylogenetic or structural analyses, nor can we confirm that a diguanylate cyclase or a phosphodiersterase may be co-transcribed with these genes since *hupSL* are at the terminal end of a contig.

The genome of Bud S10 also contained two putative arsenite oxidoreductases (*aioBA*) in truncated operons at the terminal ends of contigs, **Figure 4**. One set of *aioBA* genes (Ga0063879_03942–03943) was preceded by two cytochrome c peroxidases (*mauG*) in an operon, an arrangement that is similar to those found in other marine Alpha- and Gammaproteobacteria (Li et al., 2013). The other set of *aioBA* genes (Ga0063879_05976–05977) were distinctly different (**Figure 4**, **Supplementary Material 3**). The large subunit *aioA* contained a large insert that we have not found in other known *aioA* genes; however, this region lacked direct or inverted repetitive elements indicative of a mobile element. Furthermore, a BLASTN against this 379-base region against Bud S10's genome, the Rfam database and NCBI's CDD database produced no hits. The terminal end of the contig would be the expected region for the *mauG-mauG* gene set. However, we found no homology with *mauG,* but rather repetitive and palindromic regions discussed below. It is nearly universal to find co-occurring arsenic resistance genes, phosphate stress response genes and the *aio* genes in "arsenic islands" (Li et al., 2013). However, both of the Bud S10 *aioBA*-containing contigs were short contigs and neither contained the other genes typically found in arsenic

islands. Instead, both contigs contained additional cytochromes. However, genomic positioning may suggest a role in stress response since downstream from one *aioBA* was a universal stress protein gene (*uspA*), and downstream from the other, a carbon starvation protein (*cstA*).

Potential Phosphorus and Carbon Sequestration

In both bacteria and eukaryotes, polyphosphate can be stored in membrane-bound vacuoles, called acidocalcisomes. Polyphosphate inclusions were observed in *Thiomargarita namibiensis* and the hydrolysis of these polyphosphate granules has been linked to phosphorite formation in pore waters inhabited by dense populations of *Thiomargarita* spp. (Schulz and Schulz, 2005). Typically, acidocalcisomes are acidic and contain high levels of divalent cations, most notably calcium and magnesium (Docampo et al., 2005; Seufferheld et al., 2008; Forbes et al., 2009; Rao et al., 2009). The genetics of acidocalcisomes in bacteria and, to lesser extent eukaryotes, is poorly understood. The Bud S10 genome possessed a number of genes for inorganic phosphorus (P_i) acquisition including an ABC-type ATP-binding P_i transporter (*pstAB*), a P_i selective porin, a low-affinity P_i transporter (*PiT*), and a Na^+/P_i co-transporter. Potential storage of P_i in poly-P granules occurs via a polyphosphate kinase I (*ppk1*). Other enzymes connected to poly-P metabolism in the genome included a broad specificity polyphosphate kinase 2 (*ppk2*), an exopolyphosphatase [*surE* but not *Ppx-GppA* (either PFAM 02541 or 02833)], (p)ppGpp synthase/hydrolase and poly-P glucokinase (*ppgK*). Bud S10 possessed several candidate divalent cation transporters, including a K^+ dependent Na^+/Ca^{2+} exchanger and a potential calcium-translocating P-type ATPase, which could contribute to divalent cation accumulation in the acidocalcisomes. A V-type ATPase could be involved in the generation of an acidic polyphosphate compartment. However, a recent study of a marine *Beggiatoa* strain that lacked a nitrate vacuole had unusual membrane bound polyphosphate granules that contained high levels of Ca^{2+} and Mg^{2+} but were not acidic (Brock et al., 2012). Thus, the V-ATPase may be important for other acidic compartments, e.g., the nitrate vacuole vs. polyphosphate granules. Polyphosphate-accumulating bacteria often store organic carbon granules such as glycogen and poly(R)-hydroxyalkanoic acids. Bud S10 could potentially store glycogen synthesized from ADP-glucose via a starch synthase (*glgA*), but not poly(R)-hydroxyalkanoic acids. These genetic findings are consistent with microscopic observations of *Thiomargarita namibiensis* that stained dark brown with iodine indicating the presence of a glucose-containing polymer (Schulz and Schulz, 2005).

Assessment of Mobile and/or Repetitive Elements

In general, our examination of the genome of Bud S10 revealed a number of characteristics that are atypical for bacteria. For example, 18% of the genome was intergenic. Forty two percent of the genes were classified as hypothetical, and introns and/or MITES disrupted numerous ribosomal RNA and protein-encoding genes. Additionally, some of the most abundant gene motifs detected here indicate a complex developmental life cycle and/or genome plasticity. For example, one of the most common motifs detected in the annotation pipeline were genes for metacaspases ($n = 59$) (PFAM00656). In eukaryotes, caspases carry out programed cell death. Metacaspases are uncommon in bacteria (<18%) and their roles are in general, poorly-understood (Asplund-Samuelsson et al., 2012). In Cyanobacteria they are correlated with complex life styles, e.g., multi-cellularity, filamentation, and dimorphism, and have been shown to be restricted to strains that do not possess streamlined genomes (Asplund-Samuelsson et al., 2012) including a filamentous, multicellular endosymbiotic strain that appears to undergo programmed cell death (Zheng et al., 2013). The number of metacaspases in the Bud S10 genome far exceeds those found in other sequenced genomes. Besides the Cyanobacteria, they occur in higher numbers in some symbiotic rhizobia, symbiotic methylotrophs, and *Myxococcus* species. While the highest number ($n = 28$) and the most diverse metacaspases were reported in a Bacteriodetes, *Haliscomenobacter hydrossis* DSM 1100, which has both single cell and filamentous chains that can exhibit branching morphology.

Other common motifs in the genome of Bud S10 were reverse transcriptases ($n = 62$) (PFAM00078), HNH endonucleases ($n = 60$) (PFAM01844) and PIN domain-containing genes ($n = 64$) (PFAM01850). PIN-domain-containing proteins are Mg^{2+} dependent single stranded ribonucleases (Clissold and Ponting, 2000). PIN domain containing proteins have been shown to inhibit transcription (Winther and Gerdes, 2011) and cleave mRNA by recognizing a hairpin in the RNA secondary structure (McKenzie et al., 2012a,b). A few of the other common PFAM motifs were indicative of mobile genetic elements such as group II introns and several types of transposons. Moreover, the GenBank annotation pipeline indicated that 1120 out of 7125 genes are disrupted. Some portion of the genes are likely disrupted by frameshifts or stop codons, whether real or via sequencing errors, and some putatively-disrupted genes are perhaps incorrectly annotated as such. However, we observed some trends in those genes in which GenBank predicted a gene function (**Figure 5**, **Supplementary Material 4**). Most notably, an abundance of ATPases of unknown function that typically contained P-loop motifs and AAA domains. The AAA domain (PFAM13304) was the most common motif in Bud S10 genome with 158 occurrences. The functions of AAA domain ATPases are diverse but include many core functions, including DNA replication, recombination, and repair; transcription; and protein folding (Iyer et al., 2004; Snider et al., 2008). We also observed that many of the disrupted genes were polymerases, transcriptional regulators and transporters. Therefore, we performed a brief assessment of Bud S10's genome stability or plasticity by examining mobile and/or repetitive elements in the genome. In addition to IMG annotation, RepeatModeler was employed to aid in identifying repetitive and palindromic sequences.

RepeatModeler Detected Repetitive Motifs

RepeatModeler identified eight repetitive sequences that occurred greater than 15 times in the genome (ranging from 15 to 55 occurrences) and generated a consensus sequence for each (**Supplementary Material 5**). None of the consensus sequences

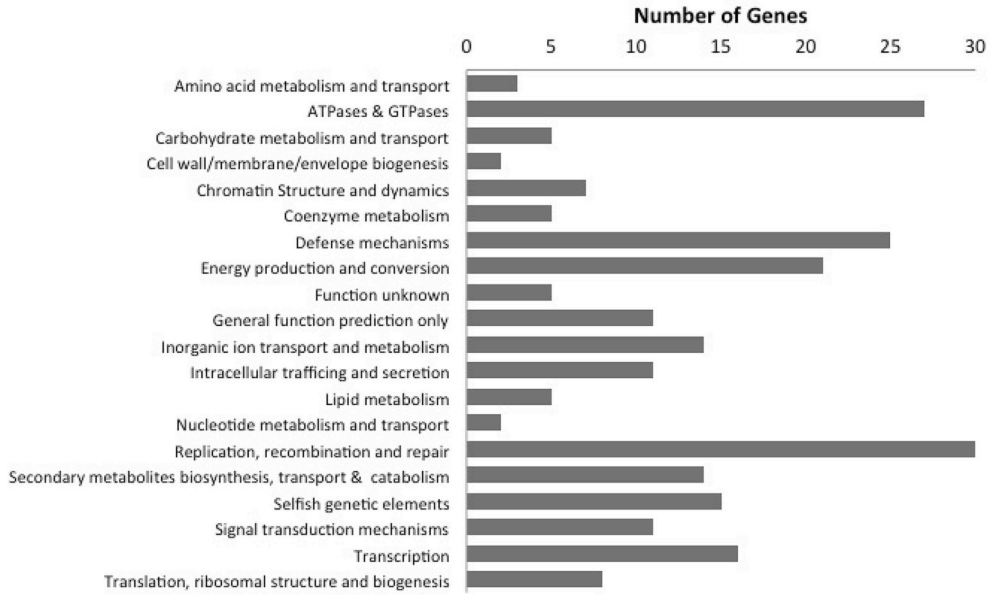

FIGURE 5 | Number of putatively disrupted genes with predicted function by functional category. A total of 1120 out of 7125 genes predicted by GenBank were putatively disrupted and of those only 248 had a predicted function.

contained motifs or homologs to the *dsrA* or rRNA gene introns, the *hupL* MITE, or families and motifs in the Rfam database. The consensus sequences were typically palindromic and a BLASTN of each against Bud S10's genome indicated that motifs in most of the consensus sequences occurred broadly across the genome (**Supplementary Material 6**). Some consensus sequences were group II introns, while others were DDE domain transposons, CRISPR arrays, and IS605 insertion elements, and their remnants as discussed further below. However, the identity and/or function of the other consensus sequences could not be determined. A detailed examination of the location of these sequences in relationship to encoding genes would facilitate greater understanding of these repetitive sequences, but such an analysis would be complicated by the large number of contigs and hypothetical genes, and therefore lies beyond the scope of the analysis reported on here. Recently, a heptamer sequence, TAACTGA, was found to occur as direct and indirect repeats in other marine *Beggiatoaceae* and was proposed to play a role in transcript regulation (MacGregor, 2015). We found 1175 occurrences of this sequence, in some cases as direct repeats, in Bud S10's genome; however, we did not find TAACTGA in any of the Repeat Modeler consensus sequences.

Group I Introns

Group I introns are usually, but not always, self-splicing mobile genetic elements that excise from the flanking exons once transcribed into RNA (Nielsen and Johansen, 2009; Hausner et al., 2014). As noted above, *dsrA* is a key gene in the sulfur oxidation pathway of vacuolate sulfur-oxidizing bacteria, and we anticipated its presence in the genome of Bud S10. However, *dsrA* appears to be interrupted by a group I intron (**Figure 3**). This finding led us to confirm the presence of the intron and determine if other versions of the gene were present by PCR. Amplification of the *dsrA* gene from Bud S10, as well as from a second single-cell amplified genome from the same *Thiomargarita* population, produced a single PCR product, and subsequent Sanger sequencing confirmed the presence of the putative intron. A search for homologs in the Rfam database yielded only one motif hit, a group I intron, *e*-value $5.4e^{-13}$, bit score 298. One key feature of most group I introns is a GU pair at the 5′ splice site. The other is the catalytic domain in the P7 Region where an exogenous guanosine binds and initiates the cleavage of the splice site. A BLASTN of the putative *dsrA* intron against the NCBI database revealed sequence homology with a group IA2 intron found in protein-coding genes of unknown function (Orf142) in the *Staphylococcus* bacteriophage, "Twort." Homology was indicated for the P7 region, as well as for the P3, P8, J6/7, and J3/4 regions (intron orf142-I2, accession number 2RKJ_C, *e*-value 0.011, 60 of 79 identities) (Landthaler and Shub, 1999; Paukstelis et al., 2008). The *dsrA* intron also shares sequence homology with the P4, P5, and partial P6 domain of another intron in a ribonucleotide reductase within the Twort genome (intron nrdE-I2, accession number AF485080 *e*-value $9e^{-10}$, 84 of 111 identities) (Landthaler and Shub, 1999; Landthaler et al., 2002). Sequence and structural similarities in the Twort introns suggest common origins (Landthaler et al., 2002). Within Twort's nrdE-I2's P6 domain is a loop containing a DNA-nicking endonuclease. This feature is absent in the *dsrA* intron. Additionally, the P9 region containing a large (51 bp) hairpin in the *dsrA* intron is not found in the Twort introns. Both

the 5′ and 3′ splice sites, as well as the catalytic regions in the *dsrA* intron, are consistent with other group IA2 introns, thus the *dsrA* intron appears to be a self-splicing ribozyme.

A BLASTN analysis of the *dsrA* intron indicated that the introns in the 23S rRNA gene and the tRNA genes introns were not close homologs. However, there were more than 100 distinct contigs in the genome that contain a portion of the *dsrA* intron that includes the long hairpin in the P9 region (**Figure 3**), but not the regions with sequence similarity with Twort's introns. Most commonly, homologs to the region of the sequence coding for the large P9 hairpin are conserved across the genome, predominantly in intergenic regions. But not uncommonly, a larger section of the *dsrA* intron is also found with the hairpin and these homologs terminate on both ends with the direct repeat TATAG (**Figure 3**). Within this section there is also a one base mismatch inverted repeat sequence. This base arrangement (DR-IR-xx-IR-DR) is a common feature of mobile elements including insertion sequence (IS) elements and MITES. The TATA at the terminal ends of the sequence is a very common feature of a MITE but it is also found in the IS630 family from which they are thought to be derived. If this region constitutes a distinct mobile element, independent of the intron elsewhere in the genome, the lack of a protein-encoding region that would promote its excision (as would occur in a IS element) indicates that it is more likely a MITE. Although a molecular symbiosis between a MITE and a group I intron has not been previously reported, a similar symbiosis has been observed between an IS element and an intron, which is referred to as an IStron (Braun et al., 2000). Interestingly, the long hairpin in the P9 region is reminiscent of a motif found in IStrons. However, in these mobile elements, a transposase is located within the loop of the large hairpin. A recent genome survey found that IStrons were restricted to a few members of the Firmicutes and Fusobacteria (Tourasse et al., 2014). Unlike group I introns, IStrons invade many types of protein-coding genes and are often found without some, or all, of the IS components, making them appear more like a MITE.

Group II Introns

Group II introns are found in bacteria and eukaryotic organelles, and are thought to be the precursors to eukaryotic spliceosomal introns and retrotransposons (Lambowitz and Zimmerly, 2011). They fold into a structure with six domains upon transcription. Many carry a reverse transcriptase and an endonuclease to promote their mobility and integration into the host genome, as well as a maturase within the fourth domain that assists in self-splicing. Some group II introns of the IIB1 family have some distinct alterations from other group II introns to include a LAGLIDADG family homing endonuclease. This feature is present in the group II intron found in *Thiomargarita* 16S genes (Salman et al., 2012). The IMG/ER annotation pipeline identified 57 group II introns, all of which were of the IIB2-type. RepeatModeler identified 55 sequences of these introns aligning to a consensus sequence that encode the 5′ RNA encoding region, a reverse transcriptase, a maturase, a HNH endonuclease and a group II catalytic region. A BLASTN of the consensus sequence revealed elements of these introns on 100 contigs in the genome assembly indicating that remnants and non-protein encoding versions of this intron occur in the genome as well. Many of these group II introns neighbored key genes in cellular replication and repair. Examples include genes for DNA helicase, ribosomal proteins L7, S14, DNA primase, ribosomal recycling factor protein, and putative AAA domain ATPases. In other cases, the group II introns interrupt operons key to *Thiomargarita* spp. metabolism such as the periplasmic nitrate reductase. In other cases they flank operons containing genes for dissimilatory nitrite reductase, dissimilatory sulfite reductase (**Figure 2**), and the nitrous oxide reductase.

RepeatModeler also identified a family of repetitive and palindromic sequences that, in one case, precedes a gene containing group II-like reverse transcriptase, homing endonuclease, and maturase motifs, but that fell below cutoff scoring for a group II intron. This hypothetical gene (Ga0063879_05465) was one of the top scoring BLASTN hits of the Repeat Modeler consensus sequence ("Family 24," BLASTN *e*-value $4e^{-45}$). This sequence does not have motifs found in group II introns but it appears to encode RNA (**Figure 6**). Within the same 16,748 bp contig as Ga0063879_05465 (Ga0063879_1138), there were three additional BLASTN hits with the same consensus sequence, two of which were in intergenic space and the third at the terminal end of the contig. Indeed, a BLASTN analysis indicated that homologs of the consensus sequence occur broadly across the genome. One gene upstream of these pseudo-group II intron genes were two putative restriction endonucleases and two BLASTN hits to a different RepeatModeler consensus sequence ("Family 176," *e*-value $5e^{-23}$; $4e^{-11}$). We found nothing similar to either consensus sequences in our searches of gene and motif databases. But the orientation of Families 24 and 176 to the group II-like protein encoding gene components is suggestive that families 24 and 176 may encode intron-like RNAs.

Insertion Sequence Elements/Transposases

IS elements are autonomous mobile elements that are common in the genomes of bacteria. They have a characteristic structure (DR-IR-transposase-IR-DR) and a promoter for transcription of the transposase in the 5′ inverted repeat (Siguier et al., 2014). There are many families of IS elements and they are primarily classified based on the type of transposase they contain. The IMG/ER pipeline identified 32 IS605-like IS elements (TIGR01766). In addition, the IMG pipeline identified 13 putative IS200-like IS elements, (IPR002686). The IS200/IS605 family of IS elements are unique among known IS elements in that they encode a HUH endonuclease that recognizes a terminal imperfect palindrome and then transposes it into a single strand of genomic DNA (He et al., 2013). They are also unique in that they have been found in IStrons (see Tourasse et al., 2014, and references therein). This IS family also has the potential to transform or decay into a unique MITE-like structures called palindrome-associated transition elements (PATEs) (Dyall-Smith et al., 2011; Siguier et al., 2014). PATEs have been observed in archaea, some cyanobacteria and *Salmonella*. Furthermore, this IS group has been shown to be domesticated by the host to modulate gene expression in a few enterobacteria [bacterial interspersed mosaic elements (BIMES)]. Because RepeatFinder detected the remnants

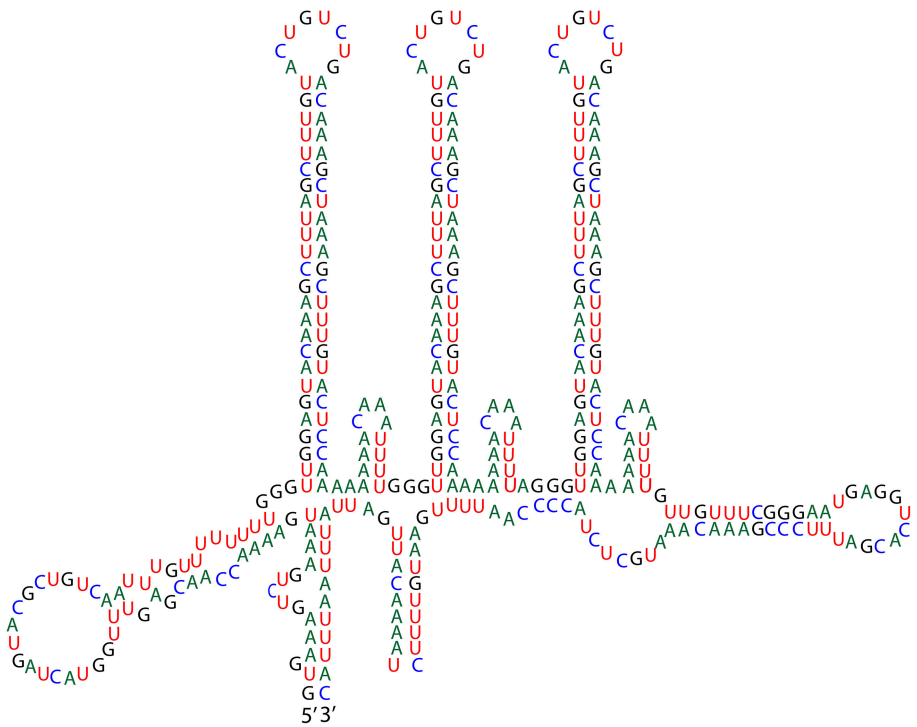

FIGURE 6 | Example of a repetitive element identified by RepeatModeler.

of IS605-like IS elements in the genome of Bud S10, we predict that the genome of also contains MITE-like features derived from these groups of IS elements.

The IMG annotation pipeline also identified 55 other putative transposases/IS elements, either containing the classic DDE transposase domain found in IS elements (Siguier et al., 2014) or a PD-(D/E)XK homing endonuclease domain. Of the DDE domain-containing genes, five are ISXO2-like transposases (PFAM12762), two are in the IS4/5 family (PFAM13340), and 20 other putative transposases contain DDE domains (PFAM13612, PFAM01609, PFAM07592, PFAM13358, PFAM13737, PFAM13586). The remaining 28 putative transposases/invertases (TIGR01784) contain a PD-(D/E)XK_2 homing endonuclease domain (PFAM12784). Many of these genes fall below the trusted bit score and include putative Yhg-A transposases that contain homologs to the *hupL* MITE. PD-(D/E)XK endonucleases have a broad range of functions and include *xisH* discussed below (Zhao et al., 2007). The IS elements that contain PD-(D/E)XK endonucleases are poorly understood.

MITEs

MITEs are small transposable elements that typically encode a recognition sequence, but not a transposase, and are thought to be derived from IS elements (Delihas, 2011; Darmon and Leach, 2014). The mobility of MITEs are likely to be via recognition and excision by endonucleases, e.g., the parental IS element encoded elsewhere in the genome (Bardaji et al., 2011). MITEs are common in eukaryotic genomes, but are less well studied in bacteria even though they appear to have significant impacts on the host. They have been found to carry open reading frames (ORFs), inactivate genes via disruption, affect transcription, create gene fusions, and cause large deletions and gene rearrangements (see Darmon and Leach, 2014, and the references therein). As previously mentioned, a MITE-like sequence disrupts the hydrogenase subunit *hupL* in Bud S10. The *hupL* MITE is a typical MITE sequence. The sequence is capped on either end by a 10 bp direct repeat. Just internal to the direct repeats is a 19 bp inverted repeat (**Supplementary Material 2**). A single bp is shared by the direct and inverted repeats. We did not find a likely candidate for potential parental IS element family to this MITE in the ISfinder database. A BLASTN of the *hupL* MITE against Bud S10's genome resulted in hits on 260/439 contigs. One of the tops hits (*e*-value $3e^{-83}$) is located within a putative Yhg-A transposase (Ga0063879_06094-06095) and immediately precedes the sulfide:quinone oxidoreductase in the ORF. MITEs and MITE-like structures are known to be derived from IS elements that have DDE domain endonucleases and HUH endonucleases, not a PD-(D/E)XK_2 endonuclease, as in putative Yhg-A transposases. Thus, the *hupL* family of MITEs may be the first MITEs detected that originated from a PD-(D/E)XK_2 endonuclease containing IS element.

The MITE found in the *dsrA* group I intron, and across the genome, is unrelated to the *hupL* MITE. The parental origin of MITEs can be difficult to determine, particularly if the IS element is no longer within the genome, and we did not find a strong candidate IS element family in the ISfinder database. MITEs commonly possess TATA at the terminal ends and they can fold into large hairpin structures as in Bud S10. We found

these MITEs sequences most similar in structure to MITEs like Correia elements (Delihas, 2008; Siddique et al., 2011). Correia elements have been shown to encode promoters for transcription. The TATA motifs function as Pribnow box promoters (TATA-box promoter in Eukaryotes). A second box promoter region exists in the inverted repeat region. Correia elements can affect transcription both from the 5′ and 3′ direction. The sequences internal the TATA-box promoters are significantly divergent from that of the Bud S10 MITE. Thus, the function(s) of these MITEs would need to be experimentally determined.

Inteins

Inteins are mobile elements that encode a peptide that splices out from a host protein after translation leaving a functional host protein. The IMG annotation pipeline identified four genes putatively containing an intein domain. As in BOGUAY, inteins were found in a *dnaB* replicative DNA helicase (Ga0063879_01777) and in a *dnaE* DNA polymerase III (Ga0063879_00371). Inteins were also identified in two ribonucleoside-diphosphate reductases- (RNR): one in an aerobic reductase subunit A (TIGR02506) (Ga0063879_01956) and the other in an anaerobic reductase (TIGR02487) (Ga0063879_04057). With the exception of the *dnaE* intein, all inteins contained either a LAGLIDADG-like domain or a homing endonuclease domain, which likely permitted the initial insertion of the intein into the genome. The genes that contain inteins work in concert for DNA replication. Ribonucleoside-diphosphate reductases convert NDPs to dNDPs and their activity is tightly regulated for replication and repair (Herrick and Sclavi, 2007). *dnaB* unwinds the dsDNA and *dnaE* adds the dNDPs generated by the RNR to the ssDNA template. It is not uncommon to find inteins in genes involved with DNA replication (Darmon and Leach, 2014). However, a review of the PFAM database (Finn et al., 2013) showed that the number of bacterial strains with four or more intein domains, as was observed here, (PFAM14890) is very uncommon ($n = 3$).

fdxN Excision Elements

Heterocyst formation in the cyanobacteria, which is a type of cell differentiation, occurs via DNA rearrangements in the genes *hupL*, a nitrogenase gene (*nifD*) and a heterocyst specific ferredoxin (*fdxH*) (Kumar et al., 2010). These rearrangements are performed by site-specific recombinases, of which *xisA* (*hupL*) and *xisC* (*nifJ*) are of the tyrosine or phage recombinase family (Nunes-Düby et al., 1998) and *xisF* is of the large serine recombinase (resolvase or IS605-like) family (Smith and Thorpe, 2002). Serine recombinases often require helper genes. Recently, it was demonstrated that the pair of helper genes, *xisHI,* for *xisF* are likely an endonuclease and a recombination directionality factor, respectively (Hwang et al., 2014). *Beggiatoa* and a few other filamentous or pleomorphic strains have also been shown to possess *xisHI* but not *xisF* (MacGregor et al., 2013c). Bud S10's genome possesses *xisHI* genes: four *xisH* (PFAM08814) and seven *XisI* (PFAM08869). The genome also had eight putative site-specific serine type recombinases (COG2452), many of which are near transposons but none of which appear to be homologs of *xisF*.

DISCUSSION

The genome of Bud S10 revealed a *Candidatus* Thiomargarita nelsonii phylotype that has the genetic potential to oxidize a large variety of sulfur species, hydrogen, and arsenite using oxygen or nitrate as terminal electron acceptors. However, the discovery of a very high number of mobile genetic elements in the draft genome, including some that interrupt genes in the sulfur and hydrogen oxidation pathways, complicates the interpretation of the functionality of some of these pathways. The degree of plasticity seen in the Bud S10 genome is exceedingly rare in bacteria and archaea. In general, a high degree of genome instability is typically seen in host-associated strains (but not ancient symbioses), some Cyanobacteria, and in certain extremophiles (Darmon and Leach, 2014). Genome instability may be a precursor to genome reduction by inactivating non-essential genes, promoting an obligate host-association (Siguier et al., 2014). We hypothesized that polyploidy in *Thiomargarita* spp., (Lane and Martin, 2010) might reduce the deleterious effects of the insertion of certain mobile elements, since multiple versions of the genome may exist in a single cell. However, we did not observe alternative versions of the *dsrA* and *hupSL* genes in the metagenome, nor were we successful in our attempts to produce multiple PCR products for these genes from both Bud S10 or another single *Thiomargarita* cell recovered from the same provannid gastropod. Therefore, we hypothesize that other factors in addition to polyploidy may result in the apparent genome plasticity observed in the Bud S10 genome.

In general, the increased rate of gene rearrangements and protein variants promoted by mobile genetic elements may increase the adaptability of an organism to new and/or extreme habitats (Lin et al., 2011; Darmon and Leach, 2014). Furthermore, mobile elements, as well as short repetitive elements similar to the heptamer recently discovered in the *Beggiatoaceae* (and a few Cyanobacteria and Bacteriodetes), have been demonstrated to promote phase and antigenic variation (see Darmon and Leach, 2014; MacGregor, 2015 and references therein). For example, excision of mobile elements can function as "on" and "off" switches to disrupted genes. Phase variation and antigenic variation have been demonstrated to play roles in numerous processes such as capsule formation, adhesion, nutrient acquisition, biofilm formation and host interactions, e.g., invasion and pathogenicity. *Thiomargarita* spp. requires access to both sulfide and oxygen (or nitrate), substrates that co-occur only in diffusional gradient interfaces, and in locations where sulfidic water is advected into oxygenated waters. Unlike its filamentous sister taxa (e.g., *Beggiatoa*, *Thioploca* etc.) *Thiomargarita* spp. is generally considered to be non-motile. As such, *Thiomargarita* ecotypes are thought to be adapted to habitats that experience temporal changes in geochemical conditions (Grünke et al., 2011). Indeed, some Namibian upwelling sediment ecotypes, such as *Ca.* T. nelsonii Thio36, experience anaerobic conditions for many months and infrequently experience access to oxygen and nitrate via methane eruptions. On the other hand, methane seeps such as Hydrate Ridge are much smaller habitats that are highly heterogenic

both spatially and temporally and they are highly ephemeral. Perhaps the increased rate of genetic change, as well as phase and antigenic variation promoted by mobile genetic elements and metacaspases, increases *Thiomargarita's* adaptability to its environment such as developing a dimorphic lifestyle that includes attached and free-living stages. The attachment to mobile substrates such as marine animals (Bailey et al., 2011), may allow them to transit between oxygenated and sulfidic conditions. Would this lifestyle constitute a fastidious host-association typically seen in other bacteria with a high degree of genome instability? At this time we find it difficult to draw any definitive conclusions regarding this question. Too little is known about the ecology of *Thiomargarita* spp., their association with host organisms and how they themselves may be important hosts to other microorganisms. But the degree of plasticity in the Bud S10 genome does raise many questions about the evolutionary processes involved in keeping the genomes of most bacteria streamlined, while others organisms, such as eukaryotes and Bud S10, abound with mobile and/or repetitive genetic elements.

What is also particularly striking about the mobile genetic elements in Bud S10 genome, beyond the sheer number detected, is the presence of a group I intron in a protein-encoding gene that is thought to be essential to Bud S10's core metabolism. Arguably, group I introns have conserved motifs but typically not conserved sequences, and thus they can be difficult to detect. In eukaryotes, group I introns tend to occur in conserved regions in genes essential to the organism's core metabolism or DNA replication such as NADH dehydrogenase, DNA polymerase, recombinase A, and genes related to photosynthesis (Nielsen and Johansen, 2009; Swithers et al., 2009). Until recently, group I introns in bacteria were thought to be rare (Hausner et al., 2014). However, a recent study found that an estimated 15% or more the bacterial domain contains introns within their rRNA and that these strains likely escaped classic detection techniques in part due to biases in PCR (Brown et al., 2015). Similarly, the Family *Beggiatoaceae* has been difficult to detect via PCR based methods even when visibly present in a sample (Angert et al., 1998; Gillan et al., 1998; Edgcomb et al., 2002; López-García et al., 2003; Sekar et al., 2006; Stevens and Ulloa, 2008; Jones et al., 2015).

Previous work has shown that *Thiomargarita* spp. possess group IA3, ID, and IC1 introns in its 16S rRNA gene (Salman et al., 2012). We have expanded the number of ribosomal genes containing introns to the 23S gene and tRNAs but we have not determined the secondary structure or type of group I introns present in these genes. However, as we report here, a group IA2 intron homologous to introns in *Staphylococcus* bacteriophage genomes occurs in a gene essential to *Thiomargarita's* core metabolism of sulfur oxidation. Group I introns disrupting protein coding genes, particular those not associated with DNA synthesis (i.e., *recA* and *nrdE*), have very rarely been reported in bacteria (Hausner et al., 2014). A group IA2 intron has been found in thermophilic *Firmicutes* flagellin gene (Hayakawa and Ishizuka, 2009). The ribozymal activity was found to be temperature dependent, promoting motility under higher temperatures and thus the intron may function in a regulatory role for flagellin gene expression under lower temperatures (Hayakawa and Ishizuka, 2012). The only other reported examples group I introns are the IStrons found in a few Firmicutes and Fusobacteria (Tourasse et al., 2014). IStrons have been shown to have alternative splicing sites, which could lead to different protein variants (Hasselmayer et al., 2004). Some of these IStrons have been found without transposases (Tourasse et al., 2014). However, we did not find any reports of MITE-like sequences similar to IS elements within IStrons in these genomes. But some of these strains have been reported to possess MITEs (Kristoffersen et al., 2011) and/or to have the "decay products" of IStrons (Siguier et al., 2014). One question that remains to be resolved about the origin of the *dsrA* "MITEtron" is whether it was once an IStron that lost its transposase or did it result from a direct symbiosis between a MITE and an intron. Does the group I intron play a role is the dispersal of the MITE element? And how does the *dsrA* MITEtron affect the regulation and transcription of *dsrA*?

Typically in model sulfur bacteria, *dsrAB* is constitutively expressed, but the level of expression depends on the sulfur source, which indicates that its expression is tightly regulated (Eddie and Hanson, 2013; Weissgerber et al., 2013). Perhaps an endonuclease, not co-located with the intron, removes the entire MITEintron prior to transcription and thus, the MITEintron has no effect on the transcription of *dsrA*. Otherwise, the presence of the MITEintron could affect the transcription of *dsrA*. On the one hand, the long hairpin loop in the MITE region may function to slow down the RNA polymerase to facilitate time for folding and self-excision; but on the other hand, it could be detrimental to polymerase activity (Bikard et al., 2010). Furthermore, these scenarios do not take into account the possibility that the MITE sequence alone is recognized and acted upon by endonucleases, reverse transcriptases, transcriptional promoters etc., and how that could affect the expression of *dsrAB*.

The addition of long Pacific Biosciences reads to our metagenome assembly facilitated the reconstruction of one of the most complete genomes from representatives of the Family *Beggiatoaceae* to date. We suspect that mobile genetic elements in the genomes of sister taxa are primarily responsible for the high number of contigs typically seen in these assembles. Therefore, we suggest that long read sequencing should be employed in (meta-)genome sequencing in these strains. The production of a number of high quality genome assemblies would enable more detailed comparative analyses that would aid in the development and testing of hypotheses concerning their genetic potential and genome instability. Unfortunately, there are no cultivated strains of vacuolated *Beggiatoaceae*, but transcriptomics under *in situ* and in controlled incubation experiments could be employed to test and refine these hypotheses. We also believe that new approaches need to be employed to develop lab cultivars in order to improve our understanding of this clade. After all, members of the *Beggiatoaceae* include the largest bacteria in the world and this genome revealed they are even more enigmatic than previously thought.

AUTHOR CONTRIBUTIONS

JB and BF conceived of this study, BF performed the DNA extraction and PCR screening of the single cell *Thiomargarita* samples. PF performed the Illumina DNA sequencing read assembly and analysis with the assistance of DJ, SJ, and GD in tetranucleotide binning, assembly and analysis. AK performed Pacific Biosciences DNA sequencing. BF performed the hybrid DNA sequencing read metagenome assembly, binning and all analyses reported herein except DJ and JB performed phylogenic analyses. MW and MM performed the assembly of the Thio36 genome. BEF wrote the manuscript (with input from all authors). All authors read and approved the final version of the manuscript.

FUNDING

Portions of this research was supported by a Grant-in-Aid from the University of Minnesota Office of the Vice President for Research (#22025), and Alfred P. Sloan Foundation research fellowship, an Early Career Investigator in Marine Microbial Ecology and Evolution Award from the Simons Foundation and the U.S. National Science Foundation (NSF) grant EAR-1057119. The research cruise was funded by NSF grant OCE-0826254.

ACKNOWLEDGMENTS

We thank Greg Rouse, Lisa Levin, Victoria Orphan and the research scientists and crew of the R/V Atlantis (AT18_1) for collecting the samples. We would like to thank the staff of the University of Minnesota Genomics Center for performing the Illumina DNA sequencing and the Mayo Institute for the Pacific Biosciences DNA sequencing. This research was supported in part by a Grant-in-Aid of Research, Artistry, and Scholarship from the University of Minnesota Office of the Vice President for Research. This work was carried out in part using resources at the University of Minnesota Supercomputing Institute and University of Minnesota Genomics Center. We thank the staff of the Minnesota Supercomputing Institute at the University of Minnesota and Charles Nguyen for computational resources and assistance with bioinformatics. Lastly, we thank Junior Scientist Elizabeth Ricci for performing some of the PCRs for this study.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fmicb.2016.00603

Supplementary Material 1 | ESOM binning map. Contigs within black borders were manually screened post binning.

Supplementary Material 2 | Enriched fasta file of *hupSL* and MITE sequence.

Supplementary Material 3 | Amino acid alignment of arsenite oxidoreductase subunit A.

Supplementary Material 4 | Spreadsheet of disrupted genes (.xls).

Supplementary Material 5 | Consensus sequences generated by RepeatModeler (.docx).

Supplementary Material 6 | Table of RepeatModeler generated consensus sequences (.docx).

REFERENCES

Abascal, F., Zardoya, R., and Posada, D. (2005). ProtTest: selection of best-fit models of protein evolution. *Bioinformatics* 21, 2104–2105. doi: 10.1093/bioinformatics/bti263

Akanuma, G., Nanamiya, H., Natori, Y., Yano, K., Suzuki, S., Omata, S., et al. (2012). Inactivation of ribosomal protein genes in *Bacillus subtilis* reveals importance of each ribosomal protein for cell proliferation and cell differentiation. *J. Bacteriol.* 194, 6282–6291. doi: 10.1128/JB.01544-12

Angert, E. R., Northup, D. E., Reysenbach, A.-L., Peek, A. S., Goebel, B. M., and Pace, N. R. (1998). Molecular phylogenetic analysis of a bacterial community in Sulphur River, Parker Cave, Kentucky. *Am. Miner.* 83, 1583–1592. doi: 10.2138/am-1998-11-1246

Armougom, F., Moretti, S., Poirot, O., Audic, S., Dumas, P., Schaeli, B., et al. (2006). Expresso: automatic incorporation of structural information in multiple sequence alignments using 3D-Coffee. *Nucleic Acids Res.* 34, W604–W608. doi: 10.1093/nar/gkl092

Asplund-Samuelsson, J., Bergman, B., and Larsson, J. (2012). Prokaryotic caspase homologs: phylogenetic patterns and functional characteristics reveal considerable diversity. *PLoS ONE* 7:e49888. doi: 10.1371/journal.pone.0049888

Bailey, J., Corsetti, F., Greene, S., Crosby, C., Liu, P., and Orphan, V. (2013). Filamentous sulfur bacteria preserved in modern and ancient phosphatic sediments: implications for the role of oxygen and bacteria in phosphogenesis. *Geobiology* 11, 397–405. doi: 10.1111/gbi.12046

Bailey, J. V., Joye, S. B., Kalanetra, K. M., Flood, B. E., and Corsetti, F. A. (2006). Evidence of giant sulphur bacteria in Neoproterozoic phosphorites. *Nature* 445, 198–201. doi: 10.1038/nature05457

Bailey, J. V., Orphan, V. J., Joye, S. B., and Corsetti, F. A. (2009). Chemotrophic microbial mats and their potential for preservation in the rock record. *Astrobiology* 9, 843–859. doi: 10.1089/ast.2008.0314

Bailey, J. V., Salman, V., Rouse, G. W., Schulz-Vogt, H. N., Levin, L. A., and Orphan, V. J. (2011). Dimorphism in methane seep-dwelling ecotypes of the largest known bacteria. *ISME J.* 5, 1926–1935. doi: 10.1038/ismej.2011.66

Bardaji, L., Añorga, M., Jackson, R. W., Martínez-Bilbao, A., Yanguas-Casás, N., and Murillo, J. (2011). Miniature transposable sequences are frequently mobilized in the bacterial plant pathogen *Pseudomonas syringae* pv. phaseolicola. *PLoS ONE* 6:e25773. doi: 10.1371/journal.pone.0025773

Berg, J. S., Schwedt, A., Kreutzmann, A.-C., Kuypers, M. M. M., and Milucka, J. (2014). Polysulfides as intermediates in the oxidation of sulfide to sulfate by *Beggiatoa* spp. *Appl. Environ. Microbiol.* 80, 629–636. doi: 10.1128/AEM.02852-13

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. doi: 10.1111/j.1462-2920.2012.02851.x

Bikard, D., Loot, C., Baharoglu, Z., and Mazel, D. (2010). Folded DNA in action: hairpin formation and biological functions in prokaryotes. *Microbiol. Mol. Biol. Rev.* 74, 570–588. doi: 10.1128/MMBR.00026-10

Boetius, A., Ravenschlag, K., Schubert, C. J., Rickert, D., Widdel, F., Gieseke, A., et al. (2000). A marine microbial consortium apparently mediating anaerobic oxidation of methane. *Nature* 407, 623–626. doi: 10.1038/35036572

Braun, V., Mehlig, M., Moos, M., Rupnik, M., Kalt, B., Mahony, D. E., et al. (2000). A chimeric ribozyme in *Clostridium difficile* combines features of

group I introns and insertion elements. *Mol. Microbiol.* 36, 1447–1459. doi: 10.1046/j.1365-2958.2000.01965.x

Brock, J., and Schulz-Vogt, H. N. (2011). Sulfide induces phosphate release from polyphosphate in cultures of a marine *Beggiatoa* strain. *ISME J.* 5, 497–506. doi: 10.1038/ismej.2010.135

Brock, J. R., Rhiel, E., Beutler, M., Salman, V., and Schulz-Vogt, H. N. (2012). Unusual polyphosphate inclusions observed in a marine *Beggiatoa* strain. *Antonie Van Leeuwenhoek* 101, 347–357. doi: 10.1007/s10482-011-9640-8

Brown, C. T., Hug, L. A., Thomas, B. C., Sharon, I., Castelle, C. J., Singh, A., et al. (2015). Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature* 523, 208–211. doi: 10.1038/nature14486

Candales, M. A., Duong, A., Hood, K. S., Li, T., Neufeld, R. A. E., Sun, R., et al. (2011). Database for bacterial group II introns. *Nucleic Acids Res.* 40, D187–D190. doi: 10.1093/nar/gkr1043

Capella-Gutiérrez, S., Silla-Martínez, J. M., and Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25, 1972–1973. doi: 10.1093/bioinformatics/btp348

Clissold, P. M., and Ponting, C. P. (2000). PIN domains in nonsense-mediated mRNA decay and RNAi. *Curr. Biol.* 10, R888–R890. doi: 10.1016/S0960-9822(00)00858-7

Crosby, C. H., and Bailey, J. V. (2012). The role of microbes in the formation of modern and ancient phosphatic mineral deposits. *Front. Microbiol.* 3:241. doi: 10.3389/fmicb.2012.00241

Dale, A. W., Bertics, V. J., Treude, T., Sommer, S., and Wallmann, K. (2013). Modeling benthic–pelagic nutrient exchange processes and porewater distributions in a seasonally hypoxic sediment: evidence for massive phosphate release by *Beggiatoa*? *Biogeosciences* 10, 629–651. doi: 10.5194/bg-10-629-2013

Darmon, E., and Leach, D. R. (2014). Bacterial genome instability. *Microbiol. Mol. Biol. Rev.* 78, 1–39. doi: 10.1128/MMBR.00035-13

Delihas, N. (2008). Small mobile sequences in bacteria display diverse structure/function motifs. *Mol. Microbiol.* 67, 475–481. doi: 10.1111/j.1365-2958.2007.06068.x

Delihas, N. (2011). Impact of small repeat sequences on bacterial genome evolution. *Genome Biol. Evol.* 3, 959–973. doi: 10.1093/gbe/evr077

Dick, G. J., Andersson, A. F., Baker, B. J., Simmons, S. L., Thomas, B. C., Yelton, A. P., et al. (2009). Community-wide analysis of microbial genome sequence signatures. *Genome Biol.* 10:R85. doi: 10.1186/gb-2009-10-8-r85

Docampo, R., De Souza, W., Miranda, K., Rohloff, P., and Moreno, S. N. (2005). Acidocalcisomes ? Conserved from bacteria to man. *Nat. Rev. Microbiol.* 3, 251–261. doi: 10.1038/nrmicro1097

Dyall-Smith, M. L., Pfeiffer, F., Klee, K., Palm, P., Gross, K., Schuster, S. C., et al. (2011). *Haloquadratum walsbyi*: limited diversity in a global pond. *PLoS ONE* 6:e20968. doi: 10.1371/journal.pone.0020968

Eddie, B. J., and Hanson, T. E. (2013). *Chlorobaculum tepidum* TLS displays a complex transcriptional response to sulfide addition. *J. Bacteriol.* 195, 399–408. doi: 10.1128/JB.01342-12

Edgcomb, V. P., Kysela, D. T., Teske, A., De Vera Gomez, A., and Sogin, M. L. (2002). Benthic eukaryotic diversity in the Guaymas Basin hydrothermal vent environment. *Proc. Natl. Acad. Sci. U.S.A.* 99, 7658–7662. doi: 10.1073/pnas.062186399

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., et al. (2013). Pfam: the protein families database. *Nucleic Acids Res.* 42, D222–D230. doi: 10.1093/nar/gkt1223

Fliss, P. S. (2014). *Pearl in the Mud: Genome Assembly and Binning of a Cold Seep Thiomargarita Nelsonii Cell and Associated Epibionts from an Environmental Metagenome*. M.S. thesis, University of Minnesota.

Forbes, C. M., O'leary, N. D., Dobson, A. D., and Marchesi, J. R. (2009). The contribution of 'omic'-based approaches to the study of enhanced biological phosphorus removal microbiology. *FEMS Microbiol. Ecol.* 69, 1–15. doi: 10.1111/j.1574-6941.2009.00698.x

Gillan, D. C., Speksnijder, A. G., Zwart, G., and De Ridder, C. (1998). Genetic diversity of the biofilm covering *Montacuta ferruginosa* (Mollusca, Bivalvia) as evaluated by denaturing gradient gel electrophoresis analysis and cloning of PCR-amplified gene fragments coding for 16S rRNA. *Appl. Environ. Microbiol.* 64, 3464–3472.

Grünke, S., Felden, J., Lichtschlag, A., Girnth, A. -C., De Beer, D., Wenzhöfer, F., et al. (2011). Niche differentiation among mat-forming, sulfide-oxidizing bacteria at cold seeps of the Nile Deep Sea Fan (Eastern Mediterranean Sea). *Geobiology* 9, 330–348. doi: 10.1111/j.1472-4669.2011.00281.x

Goldhammer, T., BrüChert, V., Ferdelman, T. G., and Zabel, M. (2010). Microbial sequestration of phosphorus in anoxic upwelling sediments. *Nat. Geosci.* 3, 557–561. doi: 10.1038/ngeo913

Greening, C., Biswas, A., Carere, C. R., Jackson, C. J., Taylor, M. C., Stott, M. B., et al. (2015). Genomic and metagenomic surveys of hydrogenase distribution indicate H_2 is a widely utilised energy source for microbial growth and survival. *ISME J.* 10, 761–777. doi: 10.1038/ismej.2015.153

Hasselmayer, O., Nitsche, C., Braun, V., and Von Eichel-Streiber, C. (2004). The IStron CdISt1 of *Clostridium difficile*: molecular symbiosis of a group I intron and an insertion element. *Anaerobe* 10, 85–92. doi: 10.1016/j.anaerobe.2003.12.003

Hausner, G., Hafez, M., and Edgell, D. R. (2014). Bacterial group I introns: mobile RNA catalysts. *Mob. DNA* 5:8. doi: 10.1186/1759-8753-5-8

Hayakawa, J., and Ishizuka, M. (2009). A group I self-splicing intron in the flagellin gene of the thermophilic bacterium Geobacillus stearothermophilus. *Biosci. Biotechnol. Biochem.* 73, 2758–2761. doi: 10.1271/bbb.90400

Hayakawa, J., and Ishizuka, M. (2012). Temperature-dependent self-splicing group I introns in the flagellin genes of the thermophilic *Bacillus* species. *Biosci. Biotechnol. Biochem.* 76, 410–413. doi: 10.1271/bbb.110741

He, S., Guynet, C., Siguier, P., Hickman, A. B., Dyda, F., Chandler, M., et al. (2013). IS200/IS605 family single-strand transposition: mechanism of IS608 strand transfer. *Nucleic Acids Res.* 41, 3302–3313. doi: 10.1093/nar/gkt014

Herrick, J., and Sclavi, B. (2007). Ribonucleotide reductase and the regulation of DNA replication: an old story and an ancient heritage. *Mol. Microbiol.* 63, 22–34. doi: 10.1111/j.1365-2958.2006.05493.x

Høgslund, S., Revsbech, N. P., Kuenen, J. G., Jørgensen, B. B., Gallardo, V. A., Vossenberg, J. V. D., et al. (2009). Physiology and behaviour of marine *Thioploca*. *ISME J.* 3, 647–657. doi: 10.1038/ismej.2009.17

Hwang, W. C., Golden, J. W., Pascual, J., Xu, D., Cheltsov, A., and Godzik, A. (2014). Site-specific recombination of nitrogen-fixation genes in cyanobacteria by XisF-XisH-XisI complex: structures and models. *Proteins*. doi: 10.1002/prot.24679. [Epub ahead of print].

Iyer, L. M., Leipe, D. D., Koonin, E. V., and Aravind, L. (2004). Evolutionary history and higher order classification of AAA+ ATPases. *J. Struct. Biol.* 146, 11–31. doi: 10.1016/j.jsb.2003.10.010

Jones, D. S., Flood, B. E., and Bailey, J. V. (2015). Metatranscriptomic analysis of diminutive *Thiomargarita*-like bacteria ("*Candidatus* Thiopilula" spp.) from abyssal cold seeps of the Barbados Accretionary Prism. *Appl. Environ. Microbiol.* 81, 3142–3156. doi: 10.1128/AEM.00039-15

Joshi, N. A., and Fass, J. N. (2011). *Sickle: A Sliding-Window, Adaptive, Quality-Based Trimming Tool for FastQ files*. Available online at: https://github.com/najoshi/sickle

Kleiner, M., Wentrup, C., Lott, C., Teeling, H., Wetzel, S., Young, J., et al. (2012). Metaproteomics of a gutless marine worm and its symbiotic microbial community reveal unusual pathways for carbon and energy use. *Proc. Natl. Acad. Sci. U.S.A.* 109, E1173–E1182. doi: 10.1073/pnas.1121198109

Kristoffersen, S. M., Tourasse, N. J., Kolstø, A.-B., and Økstad, O. A. (2011). Interspersed DNA repeats bcr1–bcr18 of *Bacillus cereus* group bacteria form three distinct groups with different evolutionary and functional patterns. *Mol. Biol. Evol.* 28, 963–983. doi: 10.1093/molbev/msq269

Kumar, K., Mella-Herrera, R. A., and Golden, J. W. (2010). Cyanobacterial heterocysts. *Cold Spring Harb. Perspect. Biol.* 2:a000315. doi: 10.1101/cshperspect.a000315

Lagkouvardos, I., Jehl, M.-A., Rattei, T., and Horn, M. (2014). Signature protein of the PVC superphylum. *Appl. Environ. Microbiol.* 80, 440–445. doi: 10.1128/AEM.02655-13

Lambowitz, A. M., and Zimmerly, S. (2011). Group II introns: mobile ribozymes that invade DNA. *Cold Spring Harb. Perspect. Biol.* 3:a003616. doi: 10.1101/cshperspect.a003616

Landthaler, M., Begley, U., Lau, N. C., and Shub, D. A. (2002). Two self-splicing group I introns in the ribonucleotide reductase large subunit gene of Staphylococcus aureus phage Twort. *Nucleic Acids Res.* 30, 1935–1943. doi: 10.1093/nar/30.9.1935

Landthaler, M., and Shub, D. A. (1999). Unexpected abundance of self-splicing introns in the genome of bacteriophage Twort: introns in multiple genes, a

single gene with three introns, and exon skipping by group I ribozymes. *Proc. Natl. Acad. Sci. U.S.A.* 96, 7005–7010. doi: 10.1073/pnas.96.12.7005

Lane, N., and Martin, W. (2010). The energetics of genome complexity. *Nature* 467, 929–934. doi: 10.1038/nature09486

Le, S. Q., and Gascuel, O. (2008). An improved general amino acid replacement matrix. *Mol. Biol. Evol.* 25, 1307–1320. doi: 10.1093/molbev/msn067

Lecompte, O., Ripp, R., Thierry, J. C., Moras, D., and Poch, O. (2002). Comparative analysis of ribosomal proteins in complete genomes: an example of reductive evolution at the domain scale. *Nucleic Acids Res.* 30, 5382–5390. doi: 10.1093/nar/gkf693

Li, H., Li, M., Huang, Y., Rensing, C., and Wang, G. (2013). *In silico* analysis of bacterial arsenic islands reveals remarkable synteny and functional relatedness between arsenate and phosphate. *Front. Microbiol.* 4:347. doi: 10.3389/fmicb.2013.00347

Lin, S., Haas, S., Zemojtel, T., Xiao, P., Vingron, M., and Li, R. (2011). Genome-wide comparison of cyanobacterial transposable elements, potential genetic diversity indicators. *Gene* 473, 139–149. doi: 10.1016/j.gene.2010.11.011

López-García, P., Duperron, S., Philippot, P., Foriel, J., Susini, J., and Moreira, D. (2003). Bacterial diversity in hydrothermal sediment and epsilonproteobacterial dominance in experimental microcolonizers at the Mid-Atlantic Ridge. *Environ. Microbiol.* 5, 961–976. doi: 10.1046/j.1462-2920.2003.00495.x

MacGregor, B. J. (2015). Abundant intergenic TAACTGA direct repeats and putative alternate RNA polymerase β′ subunits in marine *Beggiatoaceae* genomes: possible regulatory roles and origins. *Front. Microbiol.* 6:1397. doi: 10.3389/fmicb.2015.01397

MacGregor, B. J., Biddle, J. F., Harbort, C., Matthysse, A. G., and 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, 53–65. doi: 10.1016/j.margen.2013.08.001

MacGregor, B. J., Biddle, J. F., Siebert, J. R., Staunton, E., Hegg, E. L., Matthysse, A. G., et al. (2013b). Why orange Guaymas Basin *Beggiatoa* spp. are orange: single-filament-genome-enabled identification of an abundant octaheme cytochrome with hydroxylamine oxidase, hydrazine oxidase, and nitrite reductase activities. *Appl. Environ. Microbiol.* 79, 1183–1190. doi: 10.1128/AEM.02538-12

MacGregor, B. J., Biddle, J. F., and Teskea, A. (2013c). Mobile elements in a single-filament Orange Guaymas Basin *Beggiatoa* ("*Candidatus* Maribeggiatoa") sp. draft genome: evidence for genetic exchange with Cyanobacteria. *Appl. Environ. Microbiol.* 79, 3974–3985. doi: 10.1128/AEM.03821-12

Marchler-Bauer, A., Derbyshire, M. K., Gonzales, N. R., Lu, S., Chitsaz, F., Geer, L. Y., et al. (2014). CDD: NCBI's conserved domain database. *Nucleic Acids Res.* 43, D222–D226. doi: 10.1093/nar/gku1221

Markowitz, V. M., Mavromatis, K., Ivanova, N. N., Chen, I.-M. A., Chu, K., and Kyrpides, N. C. (2009). IMG ER: a system for microbial genome annotation expert review and curation. *Bioinformatics* 25, 2271–2278. doi: 10.1093/bioinformatics/btp393

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. J.* 17, 10–12. doi: 10.14806/ej.17.1.200

McGinnis, S., and Madden, T. L. (2004). BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res.* 32, W20–W25. doi: 10.1093/nar/gkh435

McKenzie, J. L., Duyvestyn, J. M., Smith, T., Bendak, K., Mackay, J., Cursons, R., et al. (2012a). Determination of ribonuclease sequence-specificity using Pentaprobes and mass spectrometry. *RNA* 18, 1267–1278. doi: 10.1261/rna.031229.111

McKenzie, J. L., Robson, J., Berney, M., Smith, T. C., Ruthe, A., Gardner, P. P., et al. (2012b). A VapBC toxin-antitoxin module is a posttranscriptional regulator of metabolic flux in mycobacteria. *J. Bacteriol.* 194, 2189–2204. doi: 10.1128/JB.06790-11

Miller, C. S., Baker, B. J., Thomas, B. C., Singer, S. W., and Banfield, J. F. (2011). EMIRGE: reconstruction of full-length ribosomal genes from microbial community short read sequencing data. *Genome Biol.* 12:R44. doi: 10.1186/gb-2011-12-5-r44

Molinas, M. F., De Candia, A., Szajnman, S. H., Rodriguez, J. B., Marti, M., Pereira, M., et al. (2011). Electron transfer dynamics of *Rhodothermus marinus* caa3 cytochrome c domains on biomimetic films. *Phys. Chem. Chem. Phys.* 13, 18088–18098. doi: 10.1039/c1cp21925a

Mußmann, M., Hu, F., Richter, M., De Beer, D., Preisler, A., Jørgensen, B., et al. (2007). Insights into the genome of large sulfur bacteria revealed by analysis of single filaments. *PLoS Biol.* 5:e230. doi: 10.1371/journal.pbio.0050230

Namiki, T., Hachiya, T., Tanaka, H., and Sakakibara, Y. (2012). MetaVelvet: an extension of Velvet assembler to *de novo* metagenome assembly from short sequence reads. *Nucleic Acids Res.* 40, e155–e155. doi: 10.1093/nar/gks678

Nawrocki, E. P., Burge, S. W., Bateman, A., Daub, J., Eberhardt, R. Y., Eddy, S. R., et al. (2014). Rfam 12.0: updates to the RNA families database. *Nucleic Acids Res.* 43, D130–D137. doi: 10.1093/nar/gku1063

Nielsen, H., and Johansen, S. D. (2009). Group I introns: moving in new directions. *RNA Biol.* 6, 375–383. doi: 10.4161/rna.6.4.9334

Nunes-Düby, S. E., Kwon, H. J., Tirumalai, R. S., Ellenberger, T., and Landy, A. (1998). Similarities and differences among 105 members of the Int family of site-specific recombinases. *Nucleic Acids Res.* 26, 391–406. doi: 10.1093/nar/26.2.391

Okonechnikov, K., Golosova, O., Fursov, M., and Team, T. U. (2012). Unipro UGENE: a unified bioinformatics toolkit. *Bioinformatics* 28, 1166–1167. doi: 10.1093/bioinformatics/bts091

Otte, S., Kuenen, J. G., Nielsen, L. P., Paerl, H. W., Zopfi, J., Schulz, H. N., et al. (1999). Nitrogen, carbon, and sulfur metabolism in natural *Thioploca* samples. *Appl. Environ. Microbiol.* 65, 3148–3157.

Paukstelis, P. J., Chen, J.-H., Chase, E., Lambowitz, A. M., and Golden, B. L. (2008). Structure of a tyrosyl-tRNA synthetase splicing factor bound to a group I intron RNA. *Nature* 451, 94–97. doi: 10.1038/nature06413

Prokopenko, M. G., Hirst, M. B., De Brabandere, L., Lawrence, D., Berelson, W., Granger, J., et al. (2013). Nitrogen losses in anoxic marine sediments driven by *Thioploca*-anammox bacterial consortia. *Nature* 500, 194–198. doi: 10.1038/nature12365

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., et al. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* 41, D590–D596. doi: 10.1093/nar/gks1219

Rao, N. N., G'Omez-Garc'Ia, M. I. R., and Kornberg, A. (2009). Inorganic polyphosphate: essential for growth and survival. *Annu. Rev. Biochem.* 78, 605–647. doi: 10.1146/annurev.biochem.77.083007.093039

Salman, V., Amann, R., Girnth, A.-C., Polerecky, L., Bailey, J. V., Høgslund, S., et al. (2011). A single-cell sequencing approach to the classification of large, vacuolated sulfur bacteria. *Syst. Appl. Microbiol.* 34, 243–259. doi: 10.1016/j.syapm.2011.02.001

Salman, V., Amann, R., Shub, D. A., and Schulz-Vogt, H. N. (2012). Multiple self-splicing introns in the 16S rRNA genes of giant sulfur bacteria. *Proc. Natl. Acad. Sci. U.S.A.* 109, 4203–4208. doi: 10.1073/pnas.1120192109

Salman, V., Bailey, J. V., and Teske, A. (2013). Phylogenetic and morphologic complexity of giant sulphur bacteria. *Antonie Van Leeuwenhoek* 104, 169–186. doi: 10.1007/s10482-013-9952-y

Sánchez, R., Serra, F., Tárraga, J., Medina, I., Carbonell, J., Pulido, L., et al. (2011). Phylemon 2.0: a suite of web-tools for molecular evolution, phylogenetics, phylogenomics and hypotheses testing. *Nucleic Acids Res.* 39, W470–W474. doi: 10.1093/nar/gkr408

Schmieder, R., and Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics* 27, 863–864. doi: 10.1093/bioinformatics/btr026

Schulz, H. N., Brinkhoff, T., Ferdelman, T. G., Marine, M. H., Teske, A., and Jørgensen, B. B. (1999). Dense populations of a giant sulfur bacterium in Namibian shelf sediments. *Science* 284, 493–495. doi: 10.1126/science.284.5413.493

Schulz, H. N., and de Beer, D. (2002). Uptake rates of oxygen and sulfide measured with individual *Thiomargarita namibiensis* cells by using microelectrodes. *Appl. Environ. Microbiol.* 68, 5746–5749. doi: 10.1128/AEM.68.11.5746-5749.2002

Schulz, H. N., and Jørgensen, B. B. (2001). Big bacteria. *Annu. Rev. Microbiol.* 55, 105–137. doi: 10.1146/annurev.micro.55.1.105

Schulz, H. N., and Schulz, H. D. (2005). Large sulfur bacteria and the formation of phosphorite. *Nature* 307, 416–418. doi: 10.1126/science.1103096

Sekar, R., Mills, D. K., Remily, E. R., Voss, J. D., and Richardson, L. L. (2006). Microbial communities in the surface mucopolysaccharide layer and the black band microbial mat of black band-diseased *Siderastrea siderea*. *Appl. Environ. Microbiol.* 72, 5963–5973. doi: 10.1128/AEM.00843-06

Seufferheld, M., Alvarez, H., and Farias, M. (2008). Role of polyphosphates in microbial adaptation to extreme environments. *Appl. Environ. Microbiol.* 74, 5867–5874. doi: 10.1128/AEM.00501-08

Siddique, A., Buisine, N., and Chalmers, R. (2011). The transposon-like Correia elements encode numerous strong promoters and provide a potential new mechanism for phase variation in the meningococcus. *PLoS Genet.* 7:e1001277. doi: 10.1371/journal.pgen.1001277

Siguier, P., Filée, J., and Chandler, M. (2006). Insertion sequences in prokaryotic genomes. *Curr. Opin. Microbiol.* 9, 526–531. doi: 10.1016/j.mib.2006.08.005

Siguier, P., Gourbeyre, E., and Chandler, M. (2014). Bacterial insertion sequences: their genomic impact and diversity. *FEMS Microbiol. Rev.* 38, 865–891. doi: 10.1111/1574-6976.12067

Smit, A., and Hubley, R. (2008-2015). *RepeatModeler Open-1.0, 1.0.7 Edn.* Seattle, WA: Institute for Systems Biology.

Smith, M., and Thorpe, H. M. (2002). Diversity in the serine recombinases. *Mol. Microbiol.* 44, 299–307. doi: 10.1046/j.1365-2958.2002.02891.x

Snider, J., Thibault, G., and Houry, W. A. (2008). The AAA+ superfamily of functionally diverse proteins. *Genome Biol.* 9:216. doi: 10.1186/gb-2008-9-4-216

Sommer, D. D., Delcher, A. L., Salzberg, S. L., and Pop, M. (2007). Minimus: a fast, lightweight genome assembler. *BMC Bioinformatics* 8:64. doi: 10.1186/1471-2105-8-64

Stamatakis, A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22, 2688–2690. doi: 10.1093/bioinformatics/btl446

Stevens, H., and Ulloa, O. (2008). Bacterial diversity in the oxygen minimum zone of the eastern tropical South Pacific. *Environ. Microbiol.* 10, 1244–1259. doi: 10.1111/j.1462-2920.2007.01539.x

Swithers, K. S., Senejani, A. G., Fournier, G. P., and Gogarten, J. P. (2009). Conservation of intron and intein insertion sites: implications for life histories of parasitic genetic elements. *BMC Evol. Biol.* 9:303. doi: 10.1186/1471-2148-9-303

Tourasse, N. J., Stabell, F. B., and Kolstø, A.-B. (2014). Survey of chimeric IStron elements in bacterial genomes: multiple molecular symbioses between group I intron ribozymes and DNA transposons. *Nucleic Acids Res.* 42, 12333–12351. doi: 10.1093/nar/gku939

Ultsch, A., and Mörchen, F. (2005). *ESOM-Maps: Tools for Clustering, Visualization, and Classification with Emergent SOM*. Department of Mathematics and Computer Science, University of Marburg.

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B. C., Remm, M., et al. (2012). Primer3—new capabilities and interfaces. *Nucleic Acids Res.* 40, e115–e115. doi: 10.1093/nar/gks596

Vignais, P. M., and Billoud, B. (2007). Occurrence, classification, and biological function of hydrogenases: an overview. *Chem. Rev.* 107, 4206–4272. doi: 10.1021/cr050196r

Weissgerber, T., Dobler, N., Polen, T., Latus, J., Stockdreher, Y., and Dahl, C. (2013). Genome-wide transcriptional profiling of the purple sulfur bacterium *Allochromatium vinosum* DSM 180T during growth on different reduced sulfur compounds. *J. Bacteriol.* 195, 4231–4245. doi: 10.1128/JB.00154-13

Winther, K. S., and Gerdes, K. (2011). Enteric virulence associated protein VapC inhibits translation by cleavage of initiator tRNA. *Proc. Natl. Acad. Sci. U.S.A.* 108, 7403–7407. doi: 10.1073/pnas.1019587108

Wood, A. P., Aurikko, J. P., and Kelly, D. P. (2004). A challenge for 21st century molecular biology and biochemistry: what are the causes of obligate autotrophy and methanotrophy? *FEMS Microbiol. Rev.* 28, 335–352. doi: 10.1016/j.femsre.2003.12.001

Yutin, N., Puigbò, P., Koonin, E. V., and Wolf, Y. I. (2012). Phylogenomics of prokaryotic ribosomal proteins. *PLoS ONE* 7:e36972. doi: 10.1371/journal.pone.0036972

Zhao, L., Bonocora, R.P., Shub, D. A., and Stoddard, B. L. (2007). The restriction fold turns to the dark side: a bacterial homing endonuclease with a PD-(D/E)-XK motif. *EMBO J.* 26, 2432–2442. doi: 10.1038/sj.emboj.7601672

Zheng, W., Rasmussen, U., Zheng, S., Bao, X., Chen, B., Gao, Y., et al. (2013). Multiple modes of cell death discovered in a prokaryotic (cyanobacterial) endosymbiont. *PLoS ONE* 8:e66147. doi: 10.1371/journal.pone.0066147

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.