Standards in Genomic Sciences

Permanent draft genome of Thiobacillus thioparus DSM 505T, an obligately chemolithoautotrophic member of the Betaproteobacteria

--Manuscript Draft--

Manuscript Number: SIGS-D-16-00109R3

Full Title: Permanent draft genome of Thiobacillus thioparus DSM 505T, an obligately chemolithoautotrophic member of the Betaproteobacteria

Article Type: Short genome report

Funding Information:
- U.S. Department of Energy (DE-AC02-05CH11231)
  Dr Nikos C Kyrpides
- Royal Society (RG120444)
  Dr Rich Boden

Abstract:
Thiobacillus thioparus DSM 505T is one of the first two isolated strains of inorganic sulfur-oxidising Bacteria. The original strain of T. thioparus was lost almost 100 years ago and the working type strain is Culture CT (=DSM 505T = ATCC 8158T) isolated by Starkey in 1934 from agricultural soil at Rutgers University, New Jersey, USA. It is an obligate chemolithoautotroph that conserves energy from the oxidation of reduced inorganic sulfur compounds using the Kelly-Trudinger pathway and uses it to fix carbon dioxide. It is not capable of heterotrophic or mixotrophic growth. The strain has a genome size of 3,201,518 bp. Here we report the genome sequence, annotation and characteristics. The genome contains 3,135 protein coding and 62 RNA coding genes. Genes encoding the transaldolase variant of the Calvin-Benson-Bassham cycle were also identified and an operon encoding carboxysomes, along with Smith's biosynthetic horseshoe in lieu of Krebs' cycle sensu stricto. Terminal oxidases were identified, viz. cytochrome c oxidase (cbb3, EC 1.9.3.1) and ubiquinol oxidase (bd, EC 1.10.3.10). There is a partial sox operon of the Kelly-Friedrich pathway of inorganic sulfur-oxidation that contains soxXYZAB genes but lacking soxCDEF, there is also a lack of the DUF302 gene previously noted in the sox operon of other members of the Proteobacteria that can use trithionate as an energy source. In spite of apparently not growing anaerobically with denitrification, the nar, nir, nor and nos operons encoding enzymes of denitrification are found in the T. thioparus genome, in the same arrangements as in the true denitrifier T. denitrificans.

Corresponding Author: Rich Boden, Ph.D B.Sc
University of Plymouth
Plymouth, Devon UNITED KINGDOM

Corresponding Author Secondary Information:

Corresponding Author's Institution: University of Plymouth

Corresponding Author's Secondary Institution:

First Author: Lee P Hutt

First Author Secondary Information:

Order of Authors:
- Lee P Hutt
- Marcel Huntemann
- Alicia Clum
- Manoj Pillay
- Krishnaveni Palaniappan
- Neha Varghese
- Natalia Mikhailova
- Natalia Mikhailova
- Dimitrios Stamatis
<table>
<thead>
<tr>
<th>Order of Authors Secondary Information:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Response to Reviewers:</td>
</tr>
<tr>
<td>In response to request to add 2 percentages to one of the tables (which could have been added in proof tbh), they are done. Hopefully this is now final.</td>
</tr>
</tbody>
</table>
Permanent draft genome of *Thiobacillus thioparus* DSM 505<sup>T</sup>, an obligately chemolithoautotrophic member of the *Betaproteobacteria*.

Lee P. Hutt<sup>1,2</sup>, Marcel Huntemann<sup>3</sup>, Alicia Clum<sup>3</sup>, Manoj Pillay<sup>3</sup>, Krishnaveni Palaniappan<sup>3</sup>, Neha Varghese<sup>3</sup>, Natalia Mikhailova<sup>3</sup>, Dimitrios Stamatis<sup>3</sup>, Tatiparthi Reddy<sup>3</sup>, Chris Daum<sup>3</sup>, Nicole Shapiro<sup>3</sup>, Natalia Ivanova<sup>3</sup>, Nikos Kyrpides<sup>3</sup>, Tanja Woyke<sup>3</sup>, Rich Boden<sup>1,2,*</sup>

[1] School of Biological and Marine Sciences, University of Plymouth, Drake Circus, Plymouth, PL4 8AA, UK.

[2] Sustainable Earth Institute, University of Plymouth, Drake Circus, Plymouth, PL4 8AA, UK.

[3] DOE Joint Genome Institute, Walnut Creek, CA 94598, USA.

* Corresponding author (rich.boden@plymouth.ac.uk)
**Abstract**

*Thiobacillus thioparus* DSM 505ᵀ is one of first two isolated strains of inorganic sulfur-oxidising bacteria. The original strain of *T. thioparus* was lost almost 100 years ago and the working type strain is Culture Cᵀ (≡DSM 505ᵀ = ATCC 8158ᵀ) isolated by Starkey in 1934 from agricultural soil at Rutgers University, New Jersey, USA. It is an obligate chemolithoautotroph that conserves energy from the oxidation of reduced inorganic sulfur compounds using the Kelly-Trudinger pathway and uses it to fix carbon dioxide. It is not capable of heterotrophic or mixotrophic growth. The strain has a genome size of 3,201,518 bp. Here we report the genome sequence, annotation and characteristics. The genome contains 3,135 protein coding and 62 RNA coding genes. Genes encoding the transaldolase variant of the Calvin-Benson-Bassham cycle were also identified and an operon encoding carboxysomes, along with Smith’s biosynthetic horseshoe in lieu of Krebs’ cycle *sensu stricto*. Terminal oxidases were identified, viz. cytochrome *c* oxidase (*cbb*3, EC 1.9.3.1) and ubiquinol oxidase (*bd*, EC 1.10.3.10). There is a partial *sox* operon of the Kelly-Friedrich pathway of inorganic sulfur-oxidation that contains *soxXYZAB* genes but lacking *soxCDEF*, there is also a lack of the DUF302 gene previously noted in the *sox* operon of other members of the ‘Proteobacteria’ that can use trithionate as an energy source. In spite of apparently not growing anaerobically with denitrification, the *nar, nir, nor* and *nos* operons encoding enzymes of denitrification are found in the *T. thioparus* genome, in the same arrangements as in the true denitrifier *T. denitrificans*. 
Keywords:

Thiobacillus thioparus, Betaproteobacteria, sulfur oxidation, chemolithoautotroph, carboxysome, denitrification

Abbreviations

ATCC – American Type Culture Collection
BLAST – Basic Linear Alignment Search Tool
CIP – Collection de L’institut Pasteur
COG – Clusters of Orthologous Groups
DSMZ - Leibniz-Institut DSMZ - Deutsche Sammlung von Mikroorganismen und Zellkulturen GmbH
GEBA – Genomic Encyclopedia of Bacteria and Archaea.
IMG – Integrated Microbial Genomes database.
JCM – Japan Collection of Microorganisms
JGI – Joint Genome Institute
KMG – 1,000 Microbial Genomes project
NBRC – NITE Biological Resource Center
NCBI – National Center for Biotechnology Information
RuBisCO - ribulose-1,5-bisphosphate carboxylase/oxygenase
Introduction

In 1902, Thiobacillus thioparus was one of the first two obligately chemolithoautrophic sulfur-oxidizing bacteria to be isolated, (along with what is now Halothiobacillus neapolitanus), and was named in 1904 [1, 2]. The original isolates were lost, but Thiobacillus thioparus is now the type species of the genus (a member of the Beta proteobacteria) and the type strain is an isolate from Starkey (1934) (= Culture C\textsuperscript{T} = DSM 505\textsuperscript{T} = ATCC 8158\textsuperscript{T} = CIP 3859\textsuperscript{T} = NBRC 103402\textsuperscript{T} = JCM 3859\textsuperscript{T} = NBRC 103402) [3, 4]. Originally, the characteristic of utilizing inorganic sulfur compounds as an energy source was thought to be a taxonomic trait unique to Thiobacillus and at its height the genus contained in at least 32 different 'species' [4]. With phylogenetic methods however, many of these strains have since been reassigned to different, often new genera with T. thioparus DSM 505\textsuperscript{T} being one of four species with validly published names left in the genus. Compared to other inorganic sulfur-oxidisers, surprisingly little research has been conducted on T. thioparus DSM 505\textsuperscript{T} in terms of physiology and biochemistry or genetics, possibly due to the low growth yields of this species when compared to T. denitrificans and T. aquaesulis [5, 6] making it more challenging to study. It may, however, give extended and contrasted insights into autotrophic sulfur-oxidation in the Bacteria. It was selected for genome sequencing as part of the Department of Energy DOE-CSP 2012 initiative – as type species of a genus.

Organism Information

Classification and features

Thiobacillus thioparus DSM 505\textsuperscript{T} was isolated from sandy loam soil from the New Jersey Agricultural Experimental Station planted with unspecified vegetable crops using 20mM thiosulfate as sole energy source in basal medium at pH 8.5 by Starkey (1934) [3] and is currently one of four species with validly published names within the genus. It forms small white colonies of 1-3 mm diameter after 2-3 days that turn pink or brown with age and which become coated with yellowish elementary sulfur. When grown in liquid media, finely divided (white) elementary sulfur is formed during early stages of growth,
particularly when thiosulfate is used as the energy source. This disappears when growth approaches stationary phase. Thiosulfate is oxidized stoichiometrically to tetrathionate after 24 h accompanied by a rise in pH, characteristic of the Kelly-Trudinger pathway. Tetrathionate is subsequently oxidized to sulfate with culture pH falling to pH 4.8 by stationary phase. During continuous culture, no intermediates are detected in the medium once steady-state has been established. If the dilution rate of a thiosulfate limited chemostat is increased, a large production of elementary sulfur is observed, which disappears as a new steady-state is established at the faster dilution rate. General features of T. thioparus DSM 505T are summarized in Table 1. A phylogenetic tree based on the 16S rRNA gene sequence showing this organisms position within the Betaproteobacteria and rooted with Thermithiobacillus tepidarius is given in Figure 1.

Cells are 1.5 – 2.0 by 0.6 to 0.8 μm and stain Gram negative. They are motile by means of a single polar flagellum between 6 - 10 μm in length, as shown in Figure 2. The dominant respiratory quinone is ubiquione-8 [7-10] and fix carbon dioxide using the Calvin-Benson-Bassham cycle at the expense of inorganic sulfur oxidation. Cells accumulate polyphosphate (‘volutin’) granules in batch culture and to a lesser extent in an energy-limited chemostat [8]. Carboxysomes are present in cells regardless of growth conditions employed (Fig. 1), but are apparently not found in T. denitrificans cells, at least not under growth conditions employed in previous studies [9]. T. thioparus can only use oxygen as a terminal electron acceptor – nitrate, nitrite, thiosulfate, sulfate, elementary sulfur and ferric iron do not support growth as terminal electron acceptors. The genomic DNA G+C content has been estimated using the thermal denaturation [11] at 61 – 62 mol% [7, 10]. T. thioparus DSM 505T does not grow on any organic carbon compound tested, including sugars (glucose, ribose, fructose, sucrose), intermediates of Krebs’ cycle (citrate, succinate, fumarate, malate, oxaloacetate), carboxylates (glycolate, formate, acetate, propionate, pyruvate), one-carbon compounds (monomethylamine, dimethylamine, trimethylamine, methanol, methane), structural amino acids (all 20), substituted thiophenes (thiophene-2-carboxylate, thiophene-3-carboxylate) or complex media (yeast extract, nutrient broth, brain-heart infusion, Columbia
sheep or horse blood agar, chocolate agar). Energy sources that support autotrophic growth of DSM 505T include thiosulfate, trithionate, tetrathionate, pentathionate, hexathionate, thiocyanate and dithionate. Some bone fide strains of T. thioparus (Tk-m and E6) grow autotrophically on carbon disulfide, dimethylsulfide, dimethyldisulfide and Admidate (O,O-dimethylphosphoramidothioate) [4], but it is not known if the type strain DSM 505T is capable of this. Autotrophic growth is not supported by Fe(II), Mn(II), Cu(I), U(IV), sulfite, dimethylsulfoxide, dimethylsulfone, pyrite or formate. During batch growth on thiosulfate the intermediate production of tetrathionate is observed during early stages of growth, indicative of the Kelly-Trudinger pathway [12]. Compared to the other two members of the genus [5, 6], T. thioparus DSM 505T has relatively low growth yields on thiosulfate (Hutt and Boden, manuscript in preparation) which may give insight into the physiological variances of Kelly-Trudinger pathway organisms even within one genus.

**Genome sequencing information**

**Genome project history**

This organism was selected for sequencing on the basis of its role in sulfur cycling, physiological, biochemical, evolutionary and biogeochemical importance, and is part of the GEBA-KMG project at the U.S. Department of Energy JGI. The genome project is deposited in the Genomes OnLine Database [13] and a high-quality permanent draft genome sequence in IMG (the annotated genome is publically available in IMG under Genome ID 2515154076) [14]. Sequencing, finishing and annotation were performed by the JGI using state of the art sequencing technology [15]. A summary of the project information is given in Table 2.

**Growth conditions and genomic DNA preparation**

T. thioparus DSM 505T DNA was obtained from Dr Hans-Peter Klenk at the DSMZ, having been grown on basal salts medium pH 6.6, supplemented with 40 mM thiosulfate as the sole energy source (DSM
Medium 36), under air at 26 °C for 72 h. DNA was extracted using the JETFLEX Genomic DNA Purification Kit from Genomed (Löhne, Germany) into TE Buffer. Quality was checked by agarose gel electrophoresis.

**Genome sequencing and assembly**

The draft genome of *Thiobacillus thioparus* DSM 503T was generated at the DOE JGI using the Illumina technology [16]. An Illumina standard shotgun library was constructed and sequenced using the Illumina HiSeq 2000 platform which generated 11,161,382 reads totalling 1,674.2 Mbp. All general aspects of library construction and sequencing performed at the JGI can be found at http://www.jgi.doe.gov. All raw Illumina sequence data was passed through DUKE, a filtering program developed at JGI, which removes known Illumina sequencing and library preparation artifacts [17]. Following steps were then performed for assembly: (1) filtered Illumina reads were assembled using Velvet (version 1.1.04) [18], (2) 1–3 Kbp simulated paired end reads were created from Velvet contigs using wgsim [19], (3) Illumina reads were assembled with simulated read pairs using Allpaths-LG (version r41043) [20]. Parameters for assembly steps were: 1) Velvet (velveth: 63 -shortPaired and velvetg: -very clean yes -export-Filtered yes -min contig lgth 500 -scaffolding no -cov cutoff 10) 2) wgsim (-e 0 -l 100 -2 100 -r 0 -R 0 -X 0) 3) Allpaths-LG (PrepareAllpathsInputs: PHRED 64=1 PLOIDY=1 FRAG COVERAGE=125 JUMP COVERAGE=25 LONG JUMP COV=50, RunAllpathsLG: THREADS=8 RUN=std shredpairs TARGETS=standard VAPI WARN ONLY=True OVERWRITE= OVERWRITE=True). The final draft assembly contained 20 contigs in 20 scaffolds. The total size of the genome is 3.2 Mbp and the final assembly is based on 392.6 Mbp of Illumina data, which provides an average 122.7x coverage of the genome.

**Genome annotation**

Genes were identified using Prodigal [21], followed by a round of manual curation using GenePRIMP [22] for finished genomes and Draft genomes in fewer than 10 scaffolds. The predicted CDSs were
translated and used to search the NCBI nonredundant database, UniProt, TIGRFam, Pfam, KEGG, COG, and InterPro databases. The tRNAscanSE tool [23] was used to find tRNA genes, whereas ribosomal RNA genes were found by searches against models of the ribosomal RNA genes built from SILVA [24]. Other non-coding RNAs such as the RNA components of the protein secretion complex and the RNase P were identified by searching the genome for the corresponding Rfam profiles using INFERNAL [25]. Additional gene prediction analysis and manual functional annotation was performed within the IMG platform [26] developed by the JGI, Walnut Creek, CA, USA [27, 28].

**Genome Properties**

The genome of *T. thioparus* DSM 505\(^T\) is 3,201,518 bp-long with a 62.3 mol% G+C content (Table 3). Of the 3,197 predicted genes, 3,135 were protein-coding genes and 62 were RNA genes. A total of 2,597 genes (81.23 %) have predicted function. A total of 538 (16.83 %) were identified as pseudogenes – the remainder annotated as hypothetical proteins. The properties and the statistics of the genome are given in Table 3, the distribution of genes into COG functional categories is given in Table 4. The genome is the second largest genome of obligate chemolithoautotrophs sequenced to date and is 89 % (bp/bp) or 90 % (protein coding genes/protein coding genes) of the size of that of *T. denitrificans* DSM 12475\(^T\) [12].

**Insights from the genome sequence**

As an obligate autotroph, it would be expected for a complete Calvin-Benson-Bassham cycle and, in lieu of Krebs’ cycle, Smith’s biosynthetic horseshoe [12, 29, 30, 33]. Smith’s horseshoe representing a very near-complete Krebs’ cycle was identified, in which citrate synthase (EC 2.3.3.16), aconitase (EC 4.2.1.3), isocitrate dehydrogenase (NADP\(^+\), EC 1.1.1.42), succinyl coenzyme A synthase (ADP-forming, EC 2.6.1.5), succinate dehydrogenase (EC 1.3.5.1) and malate dehydrogenase (oxaloacetate decarboxylating, NADP\(^+\), EC 1.1.1.40) genes are present, but fumarase (EC 4.2.1.2) and the E3 subunit of α-ketoglutarate dehydrogenase (NAD\(^+\), EC 1.2.4.2) genes are absent – E1 and E2 are present. This is a comparatively
large version of Smith’s horseshoe [33], which varies between Classes of the ‘Proteobacteria’, for example, genes for fumarase, succinate dehydrogenase and the E1 subunit of α-ketoglutarate dehydrogenase were recently found to be missing from *Thermithiobacillus tepidarius DSM 3134T* [12, 31, 32] of the *Acidithiobacillia*. Activities of Krebs’ cycle enzymes assayed by Smith *et al.* [33] in cell-free extracts of *T. thioparus* found activities of all enzymes except for α-ketoglutarate dehydrogenase, indicating the E1 and E2 subunits alone were not sufficient for activity. Interestingly, Smith detected fumarase activity; however, this strain of *T. thioparus* was isolated by the authors themselves and was not DSM 505T, so there may be further variation of Smith’s horseshoe at strain level. With many species having been erroneously classified as strains of *T. thioparus* in the past that have subsequently been proven to belong to other genera [4] it is also plausible that Smith’s strain was from another species, genus or even Class. The E1, E2 and E3 subunit genes for α-ketoglutarate dehydrogenase were identified in the genome of *T. denitrificans ATCC 25259T* [34], though enzyme activity was also absent in this strain [33, 35]. It is important to stress that the full suite of enzymes of Krebs’ cycle or Smith’s horseshoe have never been assayed in *T. thioparus* DSM 505T - clearly this is needed in order to identify if genomic data reflect true activities *in vivo* [29].

A complete Calvin-Benson-Bassham cycle is present, with a single copy of the large (*cbbL*) and small (*cbbS*) form I RuBisCO (EC 4.1.1.39) subunits and, owing to the presence of a transaldolase (EC 2.2.1.2) and absence of a sedoheptulose-1,7-bisphosphatase (EC 3.1.3.37) gene, we can conclude that it uses the transaldolase-variant Calvin-Benson-Bassham cycle [36]. Adjacent to these genes are *cbbO* and *cbbQ*, consistent with form IAq RuBisCO [37], which is canonically cytoplasmic, rather than carboxysomal. A cluster of 11 genes that encode carboxysome shell proteins and a carboxysome carbonic anhydrase were also identified. This cluster is located on the forward strand while on the reverse strand the RuBisCO cluster is located c.9 kb upstream, between which is a divergently transcribed transcriptional regulator (*cbbR*) [34]. Further evidence of carboxysome expression can be seen in transmission electron micrographs (Fig 1). The carboxysome gene cluster (experimental data would be required to demonstrate
if it is an operon or not) in *T. thioparus* DSM 505\textsuperscript{T} does not start with *cbbLS*, (as would be found in *Halothiobacillus*, *Acidithiobacillus* and *Thermithiobacillus* spp. [38]) - the lack of these genes has also been observed in *T. denitrificans* ATCC 25259\textsuperscript{T} [34] and may be a diagnostic property of this genus.

Whilst *T. thioparus* DSM 505\textsuperscript{T} cannot grow anaerobically with denitrification but *T. denitrificans* does, putative operons encoding the nitrate reductase (*nar*), nitrite reductase (*nir*), nitric oxide reductase (*nor*) and nitrous oxide reductase (*nos*) proteins of canonical denitrification are found in both *T. denitrificans* and *T. thioparas* – this may indicate a potential for the latter to grow with denitrification albeit not under conditions previously employed – for example, it may occur only under micro-oxic conditions rather than fully anoxic conditions, alternatively, some unknown factor may prevent detection of nitrate and/or the expression of these genes. No evidence for any alternative denitrification pathways [56] was found.

With regard to respiration, five cytochromes *c*\textsubscript{553}, one cytochrome *c*\textsubscript{556} and three cytochrome *b* were present. The two high-affinity terminal oxidases were both found – with multiple gene copies of cytochrome *c* oxidase (*cbb*, EC 1.9.3.1) and a single copy of cytochrome *bd*-type quinol oxidase (EC 1.10.3.10). It is worth noting that *T. denitrificans* also possess the *aa*\textsubscript{3} variant of cytochrome *c* oxidase (EC 1.9.3.1), which may permit it greater metabolic diversity, for example, growth under more variable oxygen partial pressures - this could potentially explain why *T. thioparas* cannot denitrify under anoxic conditions, even though it possesses the operons encoding enzymes of denitrification. It is known that microaerophilic organisms usually employ *cbb* or *bd*-type high-affinity terminal oxidases [56] and this may further evidence that *T. thioparas* and *T. denitrificans* may grow under such conditions, potentially with denitrification, though we cannot find any studies that demonstrate this *in vivo* thus far.

**Extended insights**

Enzymes of the Kelly-Trudinger pathway remain poorly understood and many of the genes are yet to be identified [12]. The oxidation of thiosulfate to tetrathionate takes place via a cytochrome *c*-linked thiosulfate dehydrogenase (EC 1.8.2.2), one gene (*tsdA*) for which was identified in *Allochromatium*
vinosum, a member of the Gammaproteobacteria [39]. However, tsdA is not present in T. thioparus DSM 505T, supporting the hypothesis that more than one thiosulfate dehydrogenase is present in the ‘Proteobacteria’ or could vary at Class level. Numerous Kelly-Friedrich pathway genes (soxXYZAB) were present but the remaining of the conserved soxTRS-VW-XYZABCDEFGH genes being absent [40-42]. In Paracoccus spp. (Alphaproteobacteria) soxYZ encodes a protein complex that binds thiosulfate via a cysteine residue in the initial stage of the Kelly-Friedrich pathway while soxA encode cytochromes c351 and c352.5 which capture two electrons from thiosulfate oxidation. Finally soxB encodes a hydrolase that removes the terminal sulfone group as sulfate. Missing the SoxCD or sulfur dehydrogenase protein from the multi-enzyme system would leave a sulfur atom attached to the SoxYZ residue, preventing the action of SoxB to liberate this residue to act further with another thiosulfate molecule. An unidentified protein may participate in the release of these sulfur atoms and potentially may explain the deposits of sulfur seen during initial stages of growth. This does not explain the production of tetrathionate which still would require the activity of thiosulfate dehydrogenase. If soxXYZAB genes are being expressed then they may be required as a functional part of the Kelly-Trudinger pathway. The action of Sox proteins (if any) in T. thioparus DSM 505T in conjunction and potentially collaboration with additional Kelly-Trudinger pathway proteins would undoubtedly be essential in resolving its chemolithoautotrophic metabolism, which remains poorly understood. An unidentified gene encoding a putative DUF302-family protein is present between soxA and soxB genes of Thermitiobacillus tepidarius DSM 3134T, Acidithiobacillus thiooxidans ATCC 19377T and Acidithiobacillus caldus ATCC 51756T, and Thiohalorhabdus denitrificans DSM 15699T, the function of which may be important in the Kelly-Trudinger pathway [12]; however, DUF302 is not present on the sox operon of T. thioparus DSM 505T, although six unidentified genes annotated as DUF302 family proteins are present elsewhere in the genome. The soxEF genes encode a flavocytochrome c sulfide dehydrogenase (EC 1.8.2.3) and are both found in the T. denitrificans genome, separate from the main sox gene cluster, but they are not found in T. thioparus, whereas dissimilatory sulfite reductase (dsr) genes are found in both genomes, as are adenylyl sulfate
reductase (aprAB, EC 1.8.99.2) genes, the presence of which has also been confirmed in *T. aquaesulis* [57].

It was previously noted that 5.9% of the genome (178 genes) of *Thermithiobacillus tepidarius* DSM 3134T were potential horizontally transferred genes from *Thiobacillus* *thioparus*, *Thiobacillus denitrificans* and *Sulfuricella denitrificans* of the Betaproteobacteria [12]; 96 of these genes came from the two *Thiobacillus* spp. However, very little gene transfer has taken place from members of *Acidithiobacillia* to *T. thioparus* DSM 505T with only 6 genes from *Ttb. tepidarius* DSM 3134T, 4 from *Acidithiobacillus* spp. This is perhaps not surprising due to the thermophilic and acidophilic nature of these three *Acidithiobacillia* compared to the mesophilic requirements of *T. thioparus* and the unlikelihood of these species co-inhabiting the same environments. A far larger portion (143 genes; 4.82%) of genes were attributed to transfer from members of Gammaproteobacteria, many of which grow at more neutral pH and mesophilic temperatures. There was no distinct pattern of any particular metabolism pathways or resistances *etc* being encoded by these potentially transferred genes.

**Conclusions**

The genome of *Thiobacillus thioparus* DSM 505T gives insights into many aspects of its physiology, biochemistry and evolution. This organism uses the transaldolase variant of the Calvin-Benson-Bassham cycle and produces carboxysomes for carbon dioxide fixation, evident from both the genome and transmission electron microscopy. The expression of both carboxysomes and RuBisCO may be regulated by the same divergently transcribed transcriptional regulator. Smith’s biosynthetic horseshoe is present *in lieu* of Krebs’ cycle *sensu stricto*, but this is unusually large as only 2 genes are missing, though *T. denitrificans* [34] is only missing 1 – this may in part explain the heterotrophic growth of *T. aquaesulis* since just one additional gene would convert Smith’s horseshoe into a functional version of Krebs’ cycle [55]. Many inorganic sulfur-oxidation genes of the *sox* cluster were found but *soxC*, *soxD*, *soxE* and *soxF* are absent. The *tsdA* gene for a thiosulfate dehydrogenase identified in *Allochromatium vinosum* is absent.
and confirmation of the presence of a different thiosulfate dehydrogenase enzyme and gene will require further study. The genome sequence will enable evolutionary studies into the nature of *Thiobacillus* and chemolithoautotrophs in general, in particular reigniting the obligate versus facultative and the autotrophic versus mixotrophic debates that have been largely absent from the literature in recent years, but genome sequences becoming available will now answer many questions proposed over 10 years ago [29,33].

**Competing interests**

The authors have no competing interests.

**Funding**

The sequencing and annotation was performed under the auspices of the United States Department of Energy JGI, a DOE Office of Science User Facility and is supported by the Office of Science of the United States Department of Energy under Contract Number DE-AC02-05CH11231. The authors wish to acknowledge the School of Biological and Marine Sciences, University of Plymouth for studentship funding to LH that supported the analysis of the genome and to the Royal Society Research Grant RG120444 awarded to RB to support the analysis of this genome.

**Authors' contributions**

LPH and RB analysed and mined the genome data in public databases for genes of interest and performed BLASTn/BLASTp searches to verify and validate the annotation etc and made comparisons of the sulfur oxidation, denitrification etc operons with those in other organisms. RB constructed the phylogenetic tree. LPH grew the organism and performed electron microscopy at the Electron Microscopy Unit, University of Plymouth. All other authors contributed to the sequencing, assembly and annotation of the genome sequence. All authors read and approved the final manuscript.
Acknowledgements

We acknowledge Dr Hans-Peter Klenk at the DSMZ for the provision of genomic DNA.

References


[26] INFERNAL. Inference of RNA alignments. [http://infernal.janelia.org]

[27] The Integrated Microbial Genomes (IMG) platform. [http://img.jgi.doe.gov]


Table 1

Table 1. Classification and general features of Thiobacillus thioparus DSM 505\textsuperscript{T} according to MIGS recommendations [43]

<table>
<thead>
<tr>
<th>MIGS ID</th>
<th>Property</th>
<th>Term</th>
<th>Evidence code*</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Classification</td>
<td>Domain</td>
<td>Bacteria</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Phylum</td>
<td>'Proteobacteria'</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Class</td>
<td>Betaproteobacteria</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Order</td>
<td>Hydrogenophilales</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Family</td>
<td>Hydrogenophilaceae</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Genus</td>
<td>Thiobacillus</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Species</td>
<td>Thiobacillus thioparus</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(Type) strain:</td>
<td>DSM 505\textsuperscript{T}</td>
</tr>
<tr>
<td></td>
<td>Gram stain</td>
<td></td>
<td>Negative</td>
</tr>
<tr>
<td></td>
<td>Cell shape</td>
<td></td>
<td>Rod</td>
</tr>
<tr>
<td></td>
<td>Motility</td>
<td></td>
<td>Motile</td>
</tr>
<tr>
<td></td>
<td>Sporulation</td>
<td></td>
<td>None</td>
</tr>
<tr>
<td></td>
<td>Temperature range</td>
<td></td>
<td>N.D.</td>
</tr>
<tr>
<td></td>
<td>Optimum temperature</td>
<td></td>
<td>28 °C</td>
</tr>
<tr>
<td></td>
<td>pH range; Optimum</td>
<td></td>
<td>N.D.; 7.0</td>
</tr>
<tr>
<td></td>
<td>Carbon source</td>
<td></td>
<td>Carbon dioxide</td>
</tr>
<tr>
<td>MIGS-6</td>
<td>Habitat</td>
<td></td>
<td>Agricultural soil (sandy loam)</td>
</tr>
<tr>
<td>MIGS-6.3</td>
<td>Salinity</td>
<td></td>
<td>N.D.</td>
</tr>
<tr>
<td>MIGS-22</td>
<td>Oxygen requirement</td>
<td></td>
<td>Aerobic</td>
</tr>
<tr>
<td>MIGS-15</td>
<td>Biotic relationship</td>
<td></td>
<td>Free-living</td>
</tr>
<tr>
<td>MIGS-14</td>
<td>Pathogenicity</td>
<td></td>
<td>Non-pathogen</td>
</tr>
<tr>
<td>MIGS-4</td>
<td>Geographic location</td>
<td></td>
<td>New Jersey, United States of America</td>
</tr>
<tr>
<td>MIGS-5</td>
<td>Sample collection</td>
<td></td>
<td>1934</td>
</tr>
<tr>
<td>MIGS-4.1</td>
<td>Latitude</td>
<td></td>
<td>40° 28' 57&quot; N</td>
</tr>
<tr>
<td>MIGS-4.2</td>
<td>Longitude</td>
<td></td>
<td>74° 26' 14&quot; E</td>
</tr>
<tr>
<td>MIGS-4.4</td>
<td>Altitude</td>
<td></td>
<td>28 m</td>
</tr>
</tbody>
</table>

* Evidence codes - IDA: Inferred from Direct Assay; TAS: Traceable Author Statement (i.e., a direct report exists in the literature); NAS: Non-traceable Author Statement (i.e., not directly observed for the living, isolated sample, but based on a generally accepted property for the species, or anecdotal evidence). These evidence codes are from the Gene Ontology project [44,45].
# Table 2

Table 2. Project information.

<table>
<thead>
<tr>
<th>MIGS ID</th>
<th>Property</th>
<th>Term</th>
</tr>
</thead>
<tbody>
<tr>
<td>MIGS 31</td>
<td>Finishing quality</td>
<td>Improved High-Quality Draft</td>
</tr>
<tr>
<td>MIGS-28</td>
<td>Libraries used</td>
<td>Illumina Standard PE</td>
</tr>
<tr>
<td>MIGS 29</td>
<td>Sequencing platforms</td>
<td>Illumina</td>
</tr>
<tr>
<td>MIGS 31.2</td>
<td>Fold coverage</td>
<td>122.7</td>
</tr>
<tr>
<td>MIGS 30</td>
<td>Assemblers</td>
<td>Allpaths/Velvet</td>
</tr>
<tr>
<td>MIGS 32</td>
<td>Gene calling method</td>
<td>NCBI Prokaryotic Genome Annotation Pipeline</td>
</tr>
<tr>
<td></td>
<td>Locus Tag</td>
<td>B058</td>
</tr>
<tr>
<td></td>
<td>Genbank ID</td>
<td>ARDU000000000</td>
</tr>
<tr>
<td></td>
<td>GenBank Date of Release</td>
<td>April 16th, 2013</td>
</tr>
<tr>
<td></td>
<td>GOLD ID</td>
<td>Ga0025551</td>
</tr>
<tr>
<td></td>
<td>BIOPROJECT</td>
<td>PRJNA169730</td>
</tr>
<tr>
<td>MIGS 13</td>
<td>Source Material Identifier</td>
<td>DSM 505(^T)</td>
</tr>
<tr>
<td></td>
<td>Project relevance</td>
<td>GEBA-KMG</td>
</tr>
</tbody>
</table>
### Table 3

**Table 3. Genome statistics.**

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Value</th>
<th>% of Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genome size (bp)</td>
<td>3,201,518</td>
<td>100.00</td>
</tr>
<tr>
<td>DNA coding (bp)</td>
<td>2,937,381</td>
<td>91.75</td>
</tr>
<tr>
<td>DNA G+C (bp)</td>
<td>1,994,510</td>
<td>62.80</td>
</tr>
<tr>
<td>DNA scaffolds</td>
<td>18</td>
<td>100.00</td>
</tr>
<tr>
<td>Total genes</td>
<td>3,197</td>
<td>100.00</td>
</tr>
<tr>
<td>Protein coding genes</td>
<td>3,135</td>
<td>98.06</td>
</tr>
<tr>
<td>RNA genes</td>
<td>62</td>
<td>19.94</td>
</tr>
<tr>
<td>Pseudo genes</td>
<td>538</td>
<td>16.83</td>
</tr>
<tr>
<td>Genes in internal clusters</td>
<td>267</td>
<td>8.35</td>
</tr>
<tr>
<td>Genes with function prediction</td>
<td>2,597</td>
<td>84.23</td>
</tr>
<tr>
<td>Genes assigned to COGs</td>
<td>2,258</td>
<td>70.63</td>
</tr>
<tr>
<td>Genes with Pfam domains</td>
<td>2,700</td>
<td>84.55</td>
</tr>
<tr>
<td>Genes with signal peptides</td>
<td>376</td>
<td>11.76</td>
</tr>
<tr>
<td>Genes with transmembrane helices</td>
<td>767</td>
<td>2.959</td>
</tr>
<tr>
<td>CRISPR repeats</td>
<td>2</td>
<td>0.04</td>
</tr>
</tbody>
</table>

*Commented [BSBCL2]: Please include the percentage of the total number of bp that this genome attribute accounts for.*

*Commented [BSBCL3]: Please add the percentage of the total number of bp that this genome attribute accounts for to the adjacent column.*
### Table 4

Number of genes associated with general COG functional categories.

<table>
<thead>
<tr>
<th>Code</th>
<th>Value</th>
<th>%age</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>J</td>
<td>205</td>
<td>8.2</td>
<td>Translation, ribosomal structure and biogenesis</td>
</tr>
<tr>
<td>A</td>
<td>1</td>
<td>0.0</td>
<td>RNA processing and modification</td>
</tr>
<tr>
<td>K</td>
<td>120</td>
<td>4.8</td>
<td>Transcription</td>
</tr>
<tr>
<td>L</td>
<td>83</td>
<td>3.3</td>
<td>Replication, recombination and repair</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>0.0</td>
<td>Chromatin structure and dynamics</td>
</tr>
<tr>
<td>D</td>
<td>38</td>
<td>1.5</td>
<td>Cell cycle control, Cell division, chromosome partitioning</td>
</tr>
<tr>
<td>V</td>
<td>61</td>
<td>2.4</td>
<td>Defense mechanisms</td>
</tr>
<tr>
<td>T</td>
<td>156</td>
<td>6.2</td>
<td>Signal transduction mechanisms</td>
</tr>
<tr>
<td>M</td>
<td>225</td>
<td>9.0</td>
<td>Cell wall/membrane biogenesis</td>
</tr>
<tr>
<td>N</td>
<td>101</td>
<td>4.0</td>
<td>Cell motility</td>
</tr>
<tr>
<td>U</td>
<td>56</td>
<td>2.2</td>
<td>Intracellular trafficking and secretion</td>
</tr>
<tr>
<td>O</td>
<td>144</td>
<td>5.8</td>
<td>Posttranslational modification, protein turnover, chaperones</td>
</tr>
<tr>
<td>C</td>
<td>207</td>
<td>8.3</td>
<td>Energy production and conversion</td>
</tr>
<tr>
<td>G</td>
<td>86</td>
<td>3.4</td>
<td>Carbohydrate transport and metabolism</td>
</tr>
<tr>
<td>E</td>
<td>153</td>
<td>6.1</td>
<td>Amino acid transport and metabolism</td>
</tr>
<tr>
<td>F</td>
<td>63</td>
<td>2.5</td>
<td>Nucleotide transport and metabolism</td>
</tr>
<tr>
<td>H</td>
<td>144</td>
<td>5.8</td>
<td>Coenzyme transport and metabolism</td>
</tr>
<tr>
<td>I</td>
<td>76</td>
<td>3.0</td>
<td>Lipid transport and metabolism</td>
</tr>
<tr>
<td>P</td>
<td>206</td>
<td>8.2</td>
<td>Inorganic ion transport and metabolism</td>
</tr>
<tr>
<td>Q</td>
<td>37</td>
<td>1.5</td>
<td>Secondary metabolites biosynthesis, transport and catabolism</td>
</tr>
<tr>
<td>R</td>
<td>165</td>
<td>6.6</td>
<td>General function prediction only</td>
</tr>
<tr>
<td>S</td>
<td>143</td>
<td>5.7</td>
<td>Function unknown</td>
</tr>
<tr>
<td>-</td>
<td>939</td>
<td>29.4</td>
<td>Not in COGs</td>
</tr>
</tbody>
</table>

The total is based on the total number of protein coding genes in the genome.
**Figure legends**

**Figure 1.** Maximum-likelihood phylogenetic tree based on MUSCLE alignment of 16S rRNA gene sequences of the genus *Thiobacillus* and the closely related members of the *Betaproteobacteria*. Type strains of each species are used and only species with validly published names are shown.

Sequences pertaining to organisms for which a publicly available genome sequence exists are underlined. Accession numbers for the GenBank database are in parentheses. Alignment and tree were constructed in MEGA 6 [46]. Tree was drawn using the Tamura-Nei model for maximum-likelihood trees [47]. Values at nodes are based on 5,000 bootstrap replicates, with values <70% omitted. Scale-bar indicates 2 substitutions per 100. *Thermithiobacillus tepidarius* DSM 3134T from the *Acidithiobacillia* is used as the outgroup.

**Figure 2.** Transmission electron micrographs of *T. thioparus* DSM 505T cells obtained from a thiosulfate-limited chemostat (20 mM, $D = 0.07$ h$^{-1}$) visualized in a JEOL JEM-1400Plus transmission electron microscope, operating at 120 kV.

(A) Negatively stained cells. Cells were applied to Formvar® and carbon coated copper grid before washing with saline and staining in 50 mM uranyl acetate for 5 mins and washing again.

(B) Sectioned cells showing the presence of an electron dense polyphosphate (‘volutin’) granule and numerous polyhedral carboxysomes that are paler in comparison.
Figure 1

![New tree in TIF](new tree.tif)