BMC Research Notes

DATA NOTE

Open Access

Genomic sequence data of *Thiohalocapsa marina*: a sulfur-oxidizing bacterium prevalent in treated municipal wastewater and commercial shrimp hatchery effluents

Guillermo Reyes[1*], Irma Betancourt[1], Betsy Andrade[1], Yessenia Pozo[1], Lita Sorroza[2], Luis E. Trujillo[3] and Bonny Bayot[1,4]

Abstract

Objectives This study highlighted the gap in the genetic characterization of marine bacteria, specifically within the genus *Thiohalocapsa*. This genus thrives in contaminated environments with high concentrations of sulfide, such as treated municipal wastewater. *Thc. marina* is a phototrophic purple bacterium known for its role in sulfur oxidation and bioremediation in marine aquaculture systems. To date, only one *Thc. marina* genome has been published in the GenBank database. This study enhances the understanding of the ecological adaptation and bioremediation capabilities of *Thc. marina* in treated municipal wastewater effluents.

Data description We present a draft genome of *Thc. marina* LNA26 recovered from treated municipal wastewater effluents using shotgun metagenomic sequencing. The genome of *Thc. marina* LNA26 harbors 4,356,720 bp and contains 4,032 genes (3,936 CDSs, 50 RNA genes, and 46 pseudogenes), some of them involved in sporulation, siderophores biosynthesis, arsenate bioremediation, sulfide metabolism, capacity for nitrogen fixation, the biosynthesis of PHA, and NHPL bacteriocins. *Thc. marina* LNA26 exhibits 3 CRISPR Arrays and a high abundance of COGs in signal transduction, energy production, and cell wall biogenesis, indicating advanced environmental responsiveness, energy efficiency, and cellular robustness.

Keywords Purple sulfur bacteria, Sewage, White pacific shrimp, Whole genome sequencing, Marine bacteria, Sulfide oxidation, Bioremediation

Objective

Phototrophic purple bacteria play a critical role in maintaining ecosystem health, particularly in those environments contaminated with high concentrations of sulfide, such as treated municipal wastewater effluents [1]. Despite their importance, there has been limited genetic characterization of marine bacteria from these environments, particularly within the *Thiohalocapsa* genus [2].

Thc. marina is a phototrophic purple bacterium first isolated from anoxic sediment and water [3], where it plays an important role in carbon and nitrogen cycling,

*Correspondence:
Guillermo Reyes
guianrey@espol.edu.ec
[1] Centro Nacional de Acuicultura E Investigaciones Marinas, CENAIM, Escuela Superior Politécnica del Litoral, ESPOL, Guayaquil, Ecuador
[2] Facultad de Ciencias Agropecuarias, Universidad Técnica de Machala, 5.5 Av Panamericana, Machala, Ecuador
[3] Life Science Department, Universidad de Las Fuerzas Armadas, ESPE, CENCINAT, Sangolquí, Ecuador
[4] Facultad de Ingeniería Marítima y Ciencias del Mar, FIMCM, Escuela Superior Politécnica del Litoral, ESPOL, Guayaquil, Ecuador

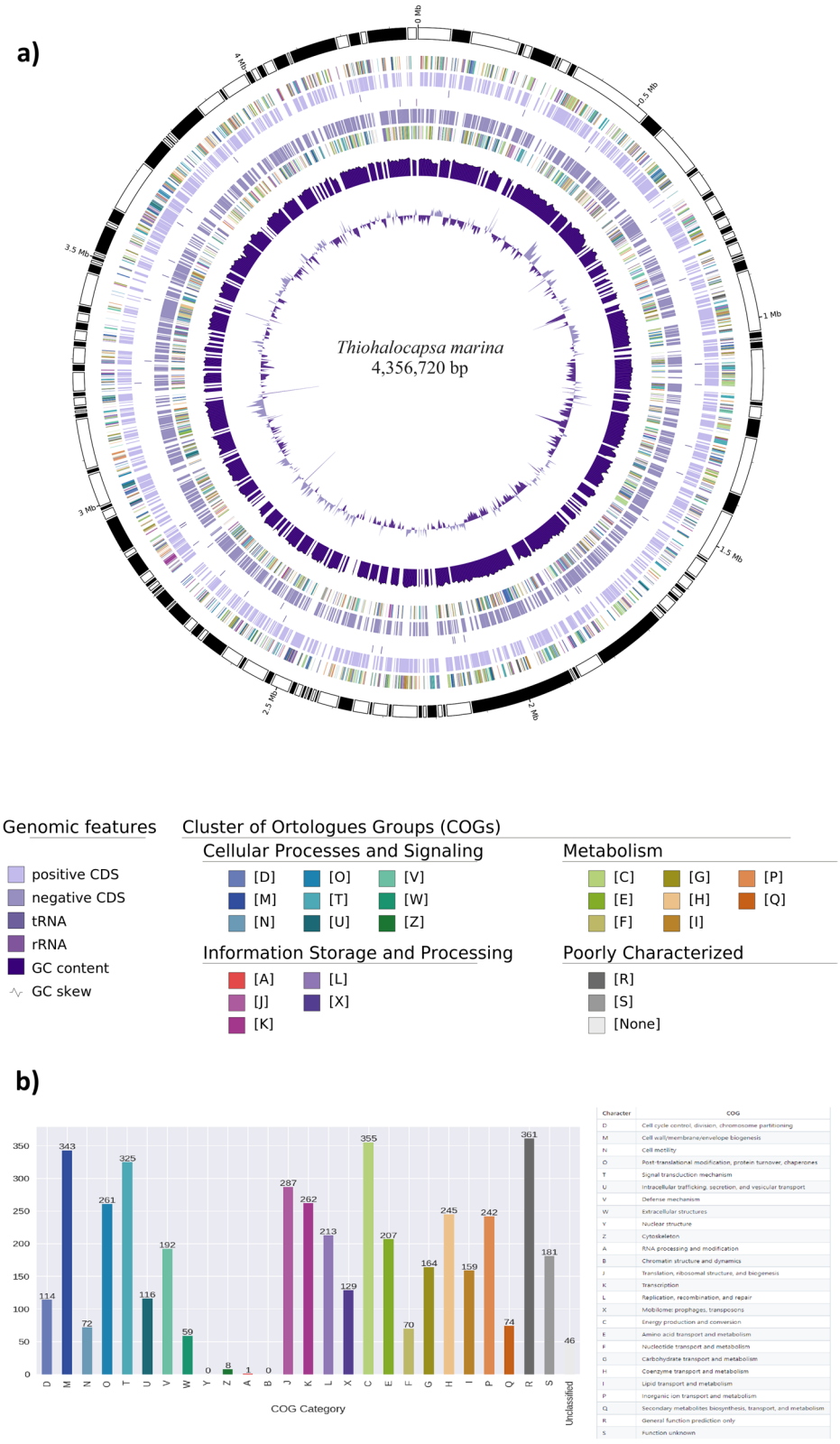

Fig. 1 Genome representation of *Thiohalocapsa marina* LNA26. a) Genome map: circles from inside to outside: GC skew, GC content, rRNA, tRNA, negative CDS, and positive CDS. Genes are colored according to their function. b) Distribution of clusters of orthologous groups of proteins (COGs) of *Thc. marina* LNA26. Each COG, grouping the coding sequences (CDSs), was involved in a category from A to Z

contributing to efficient bioremediation in shrimp culture ponds [4]. Unpublished data from our ongoing research indicates that *Thc. marina* is also prevalent in *P. vannamei* postlarvae and hatchery effluents. This fact suggests that *Thc. marina* could be playing a crucial role in improving water quality and nutrient management in aquaculture environments.

In addition, *Thc. marina* is a promising candidate for energy capture and resource recovery in the biological oxidation of sulfur in wastewater treatments, particularly due to the production of PHA [1]. These biological characteristics make it of great interest to search for genes with a wide range of biotechnological applications. However, to date, only one *Thc. marina* bacterial genome has been published [5]. We present herein the draft genome sequence of *Thc. marina* LNA26, which is the first genomic report of *Thc. marina* recovered from treated municipal wastewater using shotgun metagenomics. The data collected from the *Thc. marina* LNA26 genome enhances the understanding of the ecological adaptation, evolution, phylogenetics, and biological mechanisms for wastewater bioremediation.

Data description

A water microbiome study along a channel where effluents from treated municipal wastewater and commercial *P. vannamei* hatcheries converge was conducted. Approximately 500 mL of each effluent and the channel was filtered through 0.45-micron and 0.22-micron membranes, and an additional 500 mL was used for water quality analysis. *Thc. marina* LNA26 was identified in the treated municipal wastewater, prompting the measurement of key parameters, including salinity (5.00 g/L), dissolved oxygen concentration (0.26 $mg{\cdot}L^{-1}$), and pH (7.92) were measured using a portable multi-meter (Hach HQ Series, Hach, Germany). In addition, the concentrations of nitrate (NO_3^-—N = 0.1 $mg{\cdot}L^{-1}$), nitrite (NO_2^-—N = 0.3 $mg{\cdot}L^{-1}$), and ammonia (NH_4^+—N = 2.2 $mg{\cdot}L^{-1}$) were determined according to a published protocol [6]. The membranes were pulverized in liquid nitrogen for subsequent DNA extraction. The gDNA was extracted using a Quick-DNA Fecal/Soil Microbe Miniprep Kit (Zymo Research, USA), and the quality and concentration were evaluated using a Denovix DS-11 spectrophotometer (Denovix Inc., USA). Shotgun library preparation and sequencing were performed by Novogene Inc. (Sacramento, USA) using the Illumina NovaSeq PE150 platform. The generated data (raw sequence reads) have been deposited at SRA-NCBI under accession number SRR30726425 (Table 1, Data set 1).

Metagenome analysis was conducted using the SqueezeMeta pipeline v1.6.4 [7] in coassembly mode with Megahit v1.2.9 [8]. After binning with DAS tool [9], the genome of *Thc. marina* LNA26 was reconstructed from the municipal wastewater samples. In addition, LNA26 was identified as one of the most abundant MAGs in the microbiome of treated wastewater effluent compared to hatchery effluent and channel. The quality assessment genome was performed using CheckM v1.1.6 [10] and annotation with the NCBI PGAP [11]. A circular genomic map (Data file 1, Fig. 1a) was constructed using Genovi [12]. The bacterium was identified as *Thc. marina* based on the taxonomic classification system of Kraken v2.0 [13] and 16S rRNA gene sequence homology using BLAST [14]. *Thc. marina* LNA26 showed an average nucleotide identity of 99.62% with the reference genome of *Thc. marina* DSM 19078 (Genbank GCA_008632335.1) using fastANI v1.3.3 [15].

The *Thc. marina* LNA26 genome possesses 4,356,720 bp in 139 contigs (Data file 1, Fig. 1a), with a completeness of 96.92%, a GC content of 66.38%, and an average read coverage of 139 x (Data set 2). Homology-based gene prediction identified a total of 4,032 genes, with 3,936 CDSs (3 CRISPR Arrays), 50 RNA genes (45 tRNA; 1 rRNA; 4 ncRNA), and 46 pseudogenes. Several genes encoding enzymes involved in sporulation, siderophores biosynthesis, arsenate bioremediation, sulfide metabolism, capacity for nitrogen fixation, and the biosynthesis of NHPL bacteriocins were detected. *Thc. marina* LNA26 shows high COG abundance in signal transduction, energy production, and cell wall biogenesis (Data file 1, Fig. 1b).

A genome-based phylogenomic analysis was performed (Data file 2, Fig. 2) using TYGS [16]. Nomenclature information was obtained from LPSN (available at https://lpsn.dsmz.de) [17]. The *Thc. marina* LNA26 genome was matched to all type strain genomes available in the TYGS database using the MASH algorithm [18], and the ten strains with the smallest MASH distances were chosen and calculated using the GBDP method.

Limitations

The genomic data of *Thiohalocapsa marina* LNA26 was obtained through metagenomic shotgun sequencing from municipal wastewater samples, which are inherently complex and contaminated environments. These conditions may introduce challenges in accurately delineating the genomic characteristics and ecological roles of bacterium due to potential contamination and the presence of diverse microbial communities. Additionally, the complexity of the wastewater environment could complicate the functional annotation and interpretation of the genomic data, requiring cautious analysis

Fig. 2 Phylogenetic tree of *Thiohalocapsa marina* genomes, including the *Thc. marina* LNA26 and related species

Table 1 Overview of data files/data sets

Label	Name of data file/data set	File types (file extension)	Data repository and identifier (DOI or accession number)
Data file 1	Figure 1. Genome representation of *Thiohalocapsa marina* LNA26	*Picture file (.tiff)*	*Figshare* https://doi.org/10.6084/m9.figshare.27074959.v1 *[19]*
Data file 2	Figure 2. Phylogenetic tree of the genomes of *Thiohalocapsa marina*	*Picture file (.tiff)*	*Figshare* https://doi.org/10.6084/m9.figshare.27074884.v1 *[20]*
Data set 1	Raw sequence reads from water microbiome	*Fastq files (.fastq)*	*Sequence Read Archive from NCBI repository under accession number* SRR30726425 (https://identifiers.org/ncbi/insdc.sra:SRP533655) [21]
Data set 2	Whole genome assembly and annotation of *Thiohalocapsa marina* LNA26	*Fasta file (.fasta)*	*GenBank from NCBI repository under accession number JBFUOH000000000* (https://identifiers.org/ncbi/insdc:JBFUOH000000000) [22]

and validation of inferred metabolic pathways and interactions.

Abbreviations

ANI	Average nucleotide identity
BLAST	Basic Local Alignment Search Tool
CDS	Coding sequence
COG	Clusters of Orthologous Genes
CRISPR	Clustered regularly interspaced short palindromic repeats
DNA	Deoxyribonucleic acid
GBDP	Genome BLAST distance phylogeny
gDNA	Genomic DNA
LPSN	List of Prokaryotic names with Standing in Nomenclature
MAGs	Metagenome-Assembled Genomes
MASH	Multiple alignment from sequence homologies
NHPL	Non-Homologous Peptide Linear
PGAP	Prokaryotic genome annotation pipeline
PHA	Polyhydroxyalkanoate
RNA	Ribonucleic acid
TYGS	Type (strain) genome server

Acknowledgements
We thank Stanislaus Sonnenholzner for his kind review of the manuscript and suggestions.

Author contributions
GR, IB, and BB conceived the study. GR, BA, and YP performed the sampling and water quality analysis, GR and IB carried out the DNA extraction. GR and BB bioinformatic analysis and evaluated the genome sequencing data. GR, LS, LT, and BB supervision and obtaining funding. The manuscript was prepared by GR. The final manuscript was reviewed and approved by all authors.

Funding
This study was supported by the Corporación Ecuatoriana para el Desarrollo de la Investigación y la Academia (CEDIA) through its CEPRA program, project CEDIA N°66 "Microbiomas de un sistema marino costero integrado acuacultura-turismo-humedal: implicaciones para la salud animal y pública". Additional collaboration was provided by the project "Secuenciación de ácidos nucleicos de la biodiversidad de Ecuador continental" managed by National Biodiversity Institute and CENAIM.

Availability of data and materials
The raw sequence reads are available in the NCBI database in sequence read archive (SRA) format with the accession number SRP533655 (https://identifiers.org/ncbi/insdc.sra:SRP533655) (Table 1, Data set 1). The whole genome assembly and annotation described here has been uploaded to the GenBank database under the accession number JBFUOH000000000 (https://identifiers.org/ncbi/insdc:JBFUOH000000000) (Table 1, Data set 2).

Declarations

Ethics approval and consent to participate
Sampling was conducted on private land with the explicit consent of the landowner. The sampling process adhered to national legislation and was conducted under the authorization granted by permit MAATE-DBI-CM-2024-0413.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 6 August 2024 Accepted: 20 February 2025
Published online: 04 March 2025

References

1. Lin S, Mackey HR, Hao T, Guo G, van Loosdrecht MCM, Chen G. Biological sulfur oxidation in wastewater treatment: a review of emerging opportunities. Water Res. 2018;143:399–415. https://doi.org/10.1016/j.watres.2018.06.051.
2. Koblížek M. The purple phototrophic bacteria, vol. 28. Dordrecht: Springer, Netherlands; 2009.
3. Anil Kumar P, Srinivas TNR, Thiel V, Tank M, Sasikala C, Ramana CV, Imhoff JF. Thiohalocapsa Marina Sp. Nov., from an Indian Marine Aquaculture Pond. Int J Syst Evol Microbiol. 2009;59:2333–8. https://doi.org/10.1099/ijs.0.003053-0.
4. Joseph V, Chellappan G, Aparajitha S, Ramya RN, Vrinda S, Rejish Kumar VJ, Bright Singh IS. Molecular characterization of bacteria and archaea in a bioaugmented zero-water exchange shrimp pond. SN Appl Sci. 2021;3:458. https://doi.org/10.1007/s42452-021-04392-z.
5. NCBI GenBank. https://www.ncbi.nlm.nih.gov/Datasets/Genome/GCF_008632335.1/. Accessed 17 Jul 2024
6. American Public Health Association. Standard Methods for the Examination of Water and Wastewater. 1926; 6.
7. Tamames J, Puente-Sánchez F. SqueezeMeta, a highly portable, fully automatic metagenomic analysis pipeline. Front Microbiol. 2019. https://doi.org/10.3389/fmicb.2018.03349.
8. Li D, Liu C-M, Luo R, Sadakane K, Lam T-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct *de Bruijn* graph. Bioinformatics. 2015;31:1674–6. https://doi.org/10.1093/bioinformatics/btv033.
9. Sieber CMK, Probst AJ, Sharrar A, Thomas BC, Hess M, Tringe SG, Banfield JF. Recovery of genomes from metagenomes via a dereplication aggregation and scoring strategy. Nat Microbiol. 2018;3:836–43. https://doi.org/10.1038/s41564-018-0171-1.
10. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 2015;25:1043–55. https://doi.org/10.1101/gr.186072.114.
11. Tatusova T, DiCuccio M, Badretdin A, Chetvernin V, Nawrocki EP, Zaslavsky L, Lomsadze A, Pruitt KD, Borodovsky M, Ostell J. NCBI prokaryotic genome annotation pipeline. Nucleic Acids Res. 2016;44:6614–24. https://doi.org/10.1093/nar/gkw569.
12. Cumsille A, Durán RE, Rodríguez-Delherbe A, Saona-Urmeneta V, Cámara B, Seeger M, Araya M, Jara N, Buil-Aranda C. GenoVi, an open-source automated circular genome visualizer for bacteria and archaea. PLoS Comput Biol. 2023;19: e1010998. https://doi.org/10.1371/journal.pcbi.1010998.
13. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. Genome Biol. 2019;20:257. https://doi.org/10.1186/s13059-019-1891-0.
14. Johnson M, Zaretskaya I, Raytselis Y, Merezhuk Y, McGinnis S, Madden TL. NCBI BLAST: a better web interface. Nucleic Acids Res. 2008;36:W5–9. https://doi.org/10.1093/nar/gkn201.
15. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun. 2018;9:5114. https://doi.org/10.1038/s41467-018-07641-9.
16. Meier-Kolthoff JP, Göker M. TYGS is an automated high-throughput platform for state-of-the-art genome-based taxonomy. Nat Commun. 2019;10:2182. https://doi.org/10.1038/s41467-019-10210-3.
17. Meier-Kolthoff JP, Carbasse JS, Peinado-Olarte RL, Göker M. TYGS and LPSN: a database tandem for fast and reliable genome-based classification and nomenclature of prokaryotes. Nucleic Acids Res. 2022;50:D801–7. https://doi.org/10.1093/nar/gkab902.
18. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, Phillippy AM. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol. 2016;17:132. https://doi.org/10.1186/s13059-016-0997-x.
19. Reyes G, Betancourt I, Andrade B, Pozo Y, Sorroza L, Trujillo LE, Bayot B. Genome representation of *Thiohalocapsa marina* LNA26. **2024**. Figshare. https://doi.org/10.6084/m9.figshare.27074959.v1
20. Reyes G, Betancourt I, Andrade B, Pozo Y, Sorroza L, Trujillo LE, Bayot B. Phylogenetic tree of the genomes of *Thiohalocapsa marina*. **2024**. figshare. https://doi.org/10.6084/m9.figshare.27074884.v1
21. NCBI Sequence Read Archive https://identifiers.org/ncbi/insdc.sra:SRP533655 (2024). Accessed 09 Oct 2024
22. Reyes G. *Thiohalocapsa marina* strain LNA26, whole genome shotgun sequencing project. GenBank. 2024. https://identifiers.org/ncbi/insdc:JBFUOH000000000. Accessed 09 Oct 2024.

Publisher's Note