Data in Brief 40 (2022) 107764

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Complete genome sequence data of tropical thermophilic bacterium *Parageobacillus caldoxylosilyticus* ER4B

Xin Jie Ching[a], Nazalan Najimudin[b], Yoke Kqueen Cheah[c], Clemente Michael Vui Ling Wong[a,*]

[a] *Biotechnology Research Institute, Universiti Malaysia Sabah, Jalan UMS, 88400 Kota Kinabalu, Sabah, Malaysia*
[b] *School of Biological Science, Universiti Sains Malaysia, Persiaran Bukit Jambul, Bayan Lepas, Penang 11900, Malaysia*
[c] *Department of Biomedical Science, Faculty of Medicine and Health Sciences, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor Darul Ehsan, Malaysia*

ARTICLE INFO

Article history:
Received 25 August 2021
Revised 11 December 2021
Accepted 22 December 2021
Available online 25 December 2021

Keywords:
Parageobacillus caldoxylosilyticus
Complete whole genome sequence
Thermophile
Thermal stress

ABSTRACT

Parageobacillus caldoxylosilyticus, or previously identified as *Geobacillus caldoxylosilyticus*, is a thermophilic Gram-positive bacterium which can easily withstand growth temperatures ranging from 40 °C to 70 °C. Here, we present the first complete genome sequence of *Parageobacillus caldoxylosilyticus* ER4B which was isolated from an empty oil palm fruit bunch compost in Malaysia. Whole genome sequencing was performed using the PacBio RSII platform. The genome size of strain ER4B was around 3.9Mbp, with GC content of 44.31%. The genome consists of two contigs, in which the larger contig (3,909,276bp) represents the chromosome, while the smaller one (54,250bp) represents the plasmid. A total of 4,164 genes were successfully predicted, including 3,972 protein coding sequences, 26 rRNAs, 91 tRNAs, 74 miscRNA, and 1 tmRNA. The genome sequence data of strain ER4B reported here may contribute to the current molecular information of the species. It may also facilitate the discovery of molecular traits related to thermal stress, thus, expanding our

* Corresponding author.
E-mail address: michaelw@ums.edu.my (C.M.V.L. Wong).

https://doi.org/10.1016/j.dib.2021.107764

understanding in the acclimation or adaptation towards extreme temperature in bacteria.

Specifications Table

Subject	Biology
Specific subject area	Microbiology and Genomics
Type of data	Table Image Figure
How data were acquired	Whole genome sequence of *Parageobacillus caldoxylosilyticus* ER4B was obtained using PacBio RSII
Data format	Raw and Analyzed
Parameters for data collection	Pure culture of strain ER4B was grown in Lennox Broth (LB) at its' optimal growth temperature 64 °C and the genomic DNA was extracted when the culture reaches mid log phase.
Description of data collection	The genomic DNA was sequenced using PacBio RSII, while subsequent genome assembly and annotation was done using Canu (V1.6) and Prokka (v1.12) respectively.
Data source location	*Parageobacillus caldoxylosilyticus* strain ER4B was previously isolated from an oil palm empty fruit bunch compost in Malaysia on 11th November, 2005, and it was provided by Prof. Dr. Clemente Michael Wong Vui Ling group from Biotechnology Research Institute, Universiti Malaysia Sabah.
Data accessibility	The complete genome sequence of *Parageobacillus caldoxylosilyticus* ER4B has been deposited in NCBI GenBank under accession number CP040553-CP040554.

Value of the Data

- The data from this work represents the first complete and gapless genome of *Parageobacillus caldoxylosilyticus* as four other genome sequences from the same species deposited in NCBI GenBank are draft genomes.
- Complete whole genome sequence of *Parageobacillus caldoxylosilyticus* ER4B could provide valuable information about thermal adaptation in the bacterium, particularly at high growth temperature.
- Comparative genomics can also be carried out using this genomic data against the genome of different strains, or even different species, and this will definitely contribute in the further development and understanding molecular basis of thermal adaptation in different bacteria.
- The data can be very useful for scientists and students working in the field of microbiology, genomics, and biotechnology in extremophiles, especially thermophiles.

1. Data Description

We present the whole genome sequence of *P. caldoxylosilyticus* ER4B that was obtained from PacBio RSII. *P. caldoxylosilyticus* ER4B was previously isolated from an oil palm empty fruit bunch compost, and it would grow optimally at 64 °C in Lennox broth (LB). The genome features of strain ER4B were summarized in Table 1. The assembled genome is approximately 3.96Mbp in size, and comprises of two contigs, where the larger contig represents the chromosome and the smaller contig represents the plasmid. Subsequent genome annotation revealed a total of 4,164 genes were successfully predicted from both chromosome and plasmid, including 3,972 protein

Table 1
Genomic features of *P. caldoxylosilyticus* ER4B.

Features	Value
Contigs no.	2
Genome size (bp)	3,963,526
Chromosome size (bp)	3,909,276
Plasmid size (bp)	54,250
GC content (%)	44.32
Total number of genes	4,164 *(C: 4,106 + P: 58)
Protein coding sequences (CDS)	3,972 *(C: 3,914 + P: 58)
Genes with predicted function	2,933 *(C: 2,917 + P: 16)
Hypothetical genes	1,039 *(C: 997 + P: 42)
rRNA	26
tRNA	91
miscRNA	74
tmRNA	1

* C represents chromosome; P represents plasmid.

coding genes and 192 non-coding RNAs. Among all the predicted CDS, 2,933 of them are genes with known functions, whereas 1,039 are categorized as hypothetical genes. The position of each CDS and RNA genes can be better visualized in the genome map in Fig. 1.

The whole genome sequence of strain ER4B was utilized to construct a clearer and more accurate evolutionary relationship with other bacterial whole genomes closely related to *Para-*

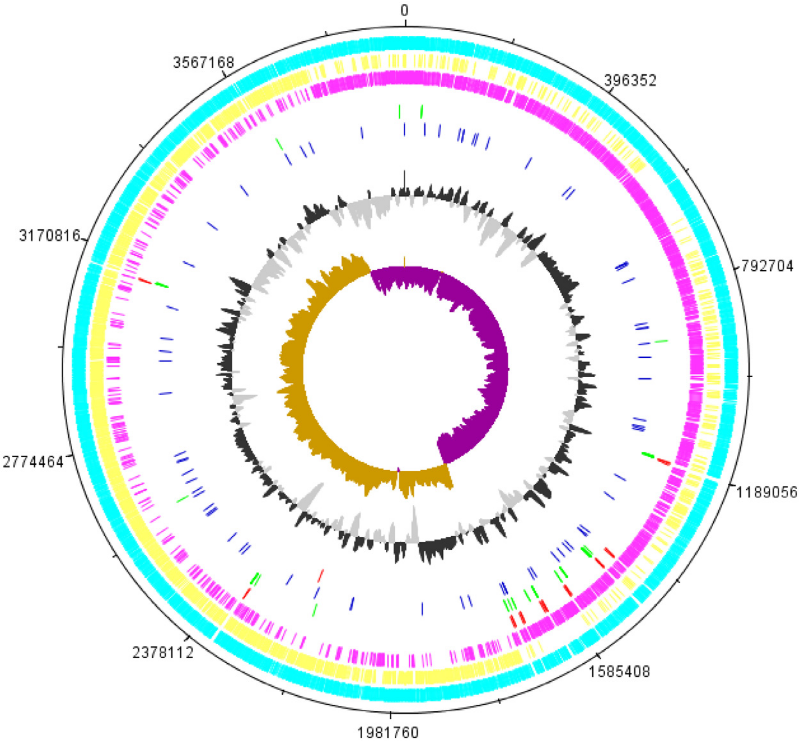

Fig. 1. Genome map of ER4B was constructed using DNAPlotter. From the outer track: 1st track represents total annotated genes, 2nd track represents forward CDS, 3rd track represents reverse CDS, 4th track represents rRNA, 5th track represents tRNA, 6th track represents miscellaneous RNA (miscRNA), 7th track represents tmRNA, 8th track represents GC plot, and the last track represents GC skew. Major tick marks interval was set at 1/10th of the overall genome size, which is 396,352bp, so 0 represents both the beginning and the ending of the sequence.

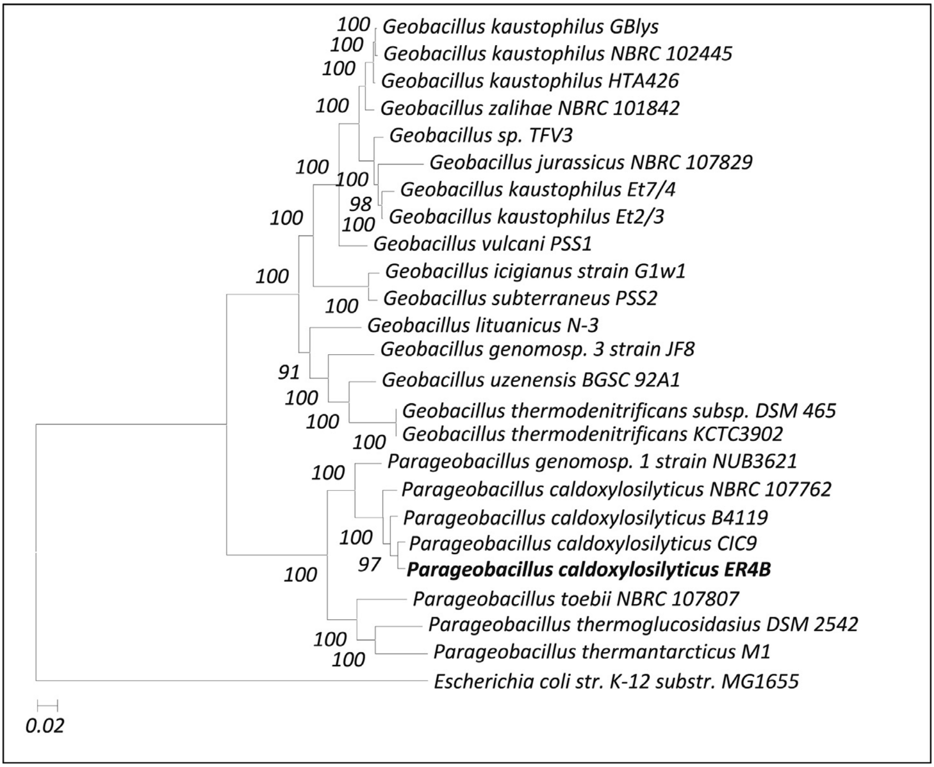
Geobacillus kaustophilus GBlys
Geobacillus kaustophilus NBRC 102445
Geobacillus kaustophilus HTA426
Geobacillus zalihae NBRC 101842
Geobacillus sp. TFV3
Geobacillus jurassicus NBRC 107829
Geobacillus kaustophilus Et7/4
Geobacillus kaustophilus Et2/3
Geobacillus vulcani PSS1
Geobacillus icigianus strain G1w1
Geobacillus subterraneus PSS2
Geobacillus lituanicus N-3
Geobacillus genomosp. 3 strain JF8
Geobacillus uzenensis BGSC 92A1
Geobacillus thermodenitrificans subsp. DSM 465
Geobacillus thermodenitrificans KCTC3902
Parageobacillus genomosp. 1 strain NUB3621
Parageobacillus caldoxylosilyticus NBRC 107762
Parageobacillus caldoxylosilyticus B4119
Parageobacillus caldoxylosilyticus CIC9
Parageobacillus caldoxylosilyticus ER4B
Parageobacillus toebii NBRC 107807
Parageobacillus thermoglucosidasius DSM 2542
Parageobacillus thermantarcticus M1
Escherichia coli str. K-12 substr. MG1655
100
98
97
91
0.02

Fig. 2. Whole genome phylogenetic tree constructed by PhyloSift, using Maximum Likelihood method based on Generalised Time-Reversible (GTR) model. The tree shows the close relationship between *P. caldoxylosilyticus* ER4B with the close related species, while *E. coli* K-12 substr. MG1655 is included to serve as an outgroup.

geobacillus and *Geobacillus* through PhyloSift. Fig. 2 clearly depicted that *P. caldoxylosilyticus* CIC9 is the closest strain to ER4B, followed by the other strains from the same species. This reconfirmed the identity of strain ER4B as it is strongly affiliated with other *P. caldoxylosilyticus* strains. It is also noteworthy that *Parageobacillus genomosp.* appeared to be closer to *P. caldoxylosilyticus* as compared to the other three species *Parageobacillus toebii, Parageobacillus thermoglucosidans* and *Parageobacillus thermantarcticus*. Besides, Fig. 2 also showed that *Parageobacillus* and *Geobacillus* were separated at the main node, forming two distinct clades between the two genera as proposed in previous study [1]. This clustering suggests that PhyloSift is able to provide higher phylogenetic resolution and better taxonomy assignment in phylogenetic analysis as compared to the more congurent single-gene phylogenetic analysis.

The annotated genome was further classified into orthologous group based on their function. 3,819 of the annotated genes were successfully classified into any one of the COG categories. As depicted in Fig. 3, 35.41% of the annotated genes were classified into “Metabolism” major category, followed by “Poorly characterized” with 28.37%, “Cellular processes and signaling” with 19.44%, and “Information storage and processing” with 16.79%. While there was a total of 26 functional COG categories, no eggNOG-annotated genes were found to be categorized under the RNA processing and modification (A), general function prediction only (R), extracellular structures (W), and nuclear structure (Y).

From Fig. 3, it was also clear that most of the annotated genes (28.37%) fell in category S, as these genes encode for hypothetical or novel proteins in which their functions were not readily assigned. Among the remaining genes with assigned functions, 8.96% of the genes which have

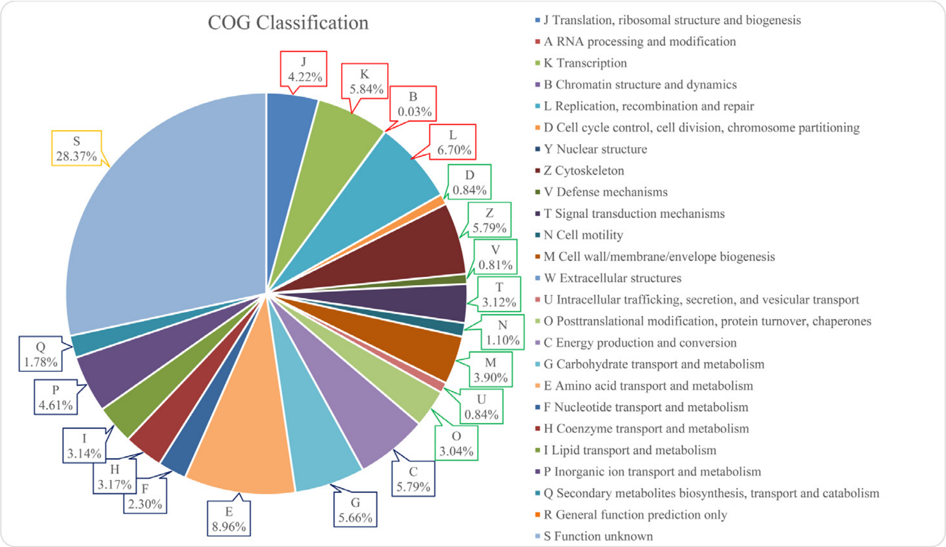

Fig. 3. Functional distribution of genes within the *P. caldoxylosilyticus* ER4B genome classified by clusters of orthologous groups (COG). COG in red box refers to major category “information storage and processing”; green box refers to “cellular processes and signaling”; blue box refers to “metabolism”; and yellow bow refers to “poorly characterized”.

functions related to amino acid transport and metabolism were classified into category E, making E a category with second highest abundance of genes. The third highest category is L with 6.70%, followed by K (5.84%), C (5.79%), Z (5.79%), and G (5.66%).

Similar to other thermophilic bacteria, strain ER4B is constantly exposed to high growth temperatures. Although group O occupied only 3.04%, it is important for the survival of strain ER4B at high temperatures as many of the heat stress related proteins were categorized in this group. The genome of strain ER4B was found to harbour several genes encoding for heat-related proteins, including GrpE, GroEL, GroES, DnaJ, DnaK, and ClpB [2,3,4] as shown in Table 2. These heat shock proteins work in conjunction with one another to prevent protein aggregation at high temperature. Besides, various features responding to heat-induced stress, such as general stress proteins, DNA SOS response proteins, and oxidative stress proteins (Table 2), can also be found in the genome of this bacterium. These proteins would trigger stress responses to prevent or mitigate the cellular damage caused by heat stress, thus crucial in contributing to the thermophilicity in strain ER4B. Interestingly, several copies of cold shock protein B (CspB) was also found in the genome of this thermophilic bacterium [5].

2. Experimental Design, Materials and Methods

The genomic DNA of strain ER4B was prepared from cells in the exponential growth phase. DNA extraction was then carried out using Qiagen DNeasy® Blood & Tissue kit (Qiagen, Valencia, CA, USA) according to manufacturer’s protocols, with several optimization (personal communication, Yong Sheau Ting) to maximize both quality and quantity of the genomic DNA extracted.

The complete genome of strain ER4B was sequenced using the PacBio RSII instrument (Pacific Biosciences, Menlo Park, CA, USA). The Single Molecule Real Time (SMRT) sequencing was conducted using 20kb SMRT bell templates and DNA Polymerase Binding kit P6-V2 on top of PacBio RSII system. The raw sequencing data obtained was then proceeded with reads correction, trimming, and *de novo* assembly using Canu v1.6 [6]. Subsequently, the assembled genome was annotated using Prokka v1.12 [7], and the complete genome map was constructed using DNAPlotter

Table 2
Number of gene copies for thermal stress related proteins in *P. Caldoxylosilyticus* ER4B.

Thermal stress related proteins	Number of copies
Cold shock proteins	
Cold shock protein CspB	3
Heat shock proteins	
Chaperone protein DnaJ	1
Chaperone protein DnaK	1
Chaperone protein ClpB	1
Heat shock protein 60 co-chaperone GroES	1
Heat shock protein 60 family chaperone GroEL	1
Heat shock protein GrpE	1
Heat-inducible transcription repressor HrcA	1
small heat shock protein*	6
Stress proteins	
General stress protein	9
Universal stress protein	1
DNA SOS response proteins	
Putative SOS response-associated peptidase YedK	1
Recombinase A (RecA)	1
LexA repressor	1
Oxidative stress proteins	
Catalases	2
Peroxiredoxin and Peroxidase	7
Superoxide dismutase	3
Thioredoxins	9

* Small heat shock proteins include HSP15, HSP18, HSP20, HSP31, and HSP33.

[8]. The annotated genome of strain ER4B was then used for the construction of the phylogenetic tree using PhyloSift [9]. Furthermore, the annotated genome was further distributed into clusters of orthologous groups (COGs) based on functional annotation using eggNOG-mapper [10].

Ethics Statement

This work did not involve any animals or human subjects. The manuscript represents the author’s original work which has not been published elsewhere.

CRediT Author Statement

Clemente Michael Vui Ling Wong, Nazalan Najimudin, Yoke Kqueen Cheah and Xin Jie Ching: Conceptualization, Methodology; **Xin Jie Ching:** Data curation, Writing- Original draft preparation; **Clemente Michael Vui Ling Wong:** Supervision; **Xin Jie Ching and Clemente Michael Vui Ling Wong:** Writing - Reviewing and Editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships which have or could be perceived to have influenced the work reported in this article.

Acknowledgments

The funding support from the Ministry of Science, Technology, and Innovation (MOSTI), Malaysia, under the Antarctica Flagship Programme (Sub-Project 1: Grant no. FP1213E036) is gratefully acknowledged.

Supplementary Materials

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.dib.2021.107764.

References

[1] H. Aliyu, P. Lebre, J. Blom, D. Cowan, P. De Maayer, Phylogenomic re-assessment of the thermophilic genus *Geobacillus*, Syst. Appl. Microbiol. 39 (2016) 527–533, doi:10.1016/j.syapm.2016.09.004.

[2] F. Anglès, M. Castanié-Cornet, N. Slama, M. Dinclaux, A. Cirinesi, J. Portais, F. Létisse, P. Genevaux, Multilevel interaction of the DnaK/DnaJ (HSP70/HS40) stress-responsive chaperone machine with the central metabolism, Sci. Rep. 7 (2017) 41341, doi:10.1038/srep41341.

[3] M. Hayer-Hartl, A. Bracher, F.U. Hartl, The GroEL-GroES chaperonin machine: a nano-cage for protein folding, Trends Biochem. Sci. 41 (2015) 62–67, doi:10.1016/j.tibs.2015.07.009.

[4] D. Whitley, S.P. Goldberg, W.D. Jordan, Heat shock proteins: a review of the molecular chaperones, J. Vasc. Surg. 29 (1999) 748–751, doi:10.1016/S0741-5214(99)70329-0.

[5] R. Keto-Timonen, N. Hietala, E. Palonen, A. Hakakorpi, M. Lindström, H. Korkeala, Cold shock proteins: a mini review with special emphasis on Csp-family of enteropathogenic *Yersinia*, Front. Microbiol. 7 (2016) 1151, doi:10.3389/fmicb.2016.01151.

[6] S. Koren, B.P. Walenz, K. Berlin, J.R. Miller, N.H. Bergman, A.M. Phillippy, Canu: scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation, Genome Res. 27 (2017) 722–736, doi:10.1101/gr.215087.116.

[7] T. Seemann, Prokka: rapid prokaryotic genome annotation, Bioinformatics 30 (2014) 2068–2069, doi:10.1093/bioinformatics/btu153.

[8] T. Carver, N. Thomson, A. Bleasby, M. Berriman, J. Parkhill, DNAPlotter: circular and linear interactive genome visualization, Bioinformatics 25 (2009) 119–120, doi:10.1093/bioinformatics/btn578.

[9] A.E. Darling, G. Jospin, E. Lowe, F.A. Matsen IV, H.M. Bik, J.A. Eisen, PhyloSift: phylogenetic analysis of genomes and metagenomes, PeerJ 2 (2014) e243, doi:10.7717/peerj.243.

[10] J. Huerta-Cepas, K. Forslund, L.P. Coelho, D. Szklarczyk, L.J. Jensen, C. van Meing, P. Bork, Fast genome-wide functional annotation through orthology assignment by eggnog-mapper, Mol. Biol. Evol. 34 (2017) 2115–2122, doi:10.1093/molbev/msx148.