Complete Genome Sequence of the Type Strain *Tepidimonas taiwanensis* LMG 22826^T, a Thermophilic Alkaline Protease and Polyhydroxyalkanoate Producer

Kristyna Hermankova[1], Xenie Kourilova[2], Iva Pernicova[2], Matej Bezdicek[3], Martina Lengerova[3], Stanislav Obruca[2], and Karel Sedlar [1,4,*]

[1]Department of Biomedical Engineering, Faculty of Electrical Engineering and Communication, Brno University of Technology, Czech Republic
[2]Department of Food Chemistry and Biotechnology, Faculty of Chemistry, Brno University of Technology, Czech Republic
[3]Department of Internal Medicine—Hematology and Oncology, University Hospital Brno, Czech Republic
[4]Department of Informatics, Institute of Bioinformatics, Ludwig-Maximilians-Universität München, Germany

*Corresponding author: E-mail: sedlar@vut.cz.

Accepted: 9 December 2021

Abstract

Tepidimonas taiwanensis is a moderately thermophilic, Gram-negative, rod-shaped, chemoorganoheterotrophic, motile bacterium. The alkaline protease producing type strain *T. taiwanensis* LMG 22826^T was recently reported to also be a promising producer of polyhydroxyalkanoates (PHAs)—renewable and biodegradable polymers representing an alternative to conventional plastics. Here, we present its first complete genome sequence which is also the first complete genome sequence of the whole species. The genome consists of a single 2,915,587-bp-long circular chromosome with GC content of 68.75%. Genome annotation identified 2,764 genes in total while 2,634 open reading frames belonged to protein-coding genes. Although functional annotation of the genome and division of genes into Clusters of Orthologous Groups (COGs) revealed a relatively high number of 694 genes with unknown function or unknown COG, the majority of genes were assigned a function. Most of the genes, 406 in total, were involved in energy production and conversion, and amino acid transport and metabolism. Moreover, particular key genes involved in the metabolism of PHA were identified. Knowledge of the genome in connection with the recently reported ability to produce bioplastics from the waste stream of wine production makes *T. taiwanensis* LMG 22826^T, an ideal candidate for further genome engineering as a bacterium with high biotechnological potential.

Key words: *Tepidimonas taiwanensis*, polyhydroxyalkanoates, hybrid assembly, Oxford Nanopore Technologies, functional annotation, alkaline protease.

Introduction

The majority of current plastics, for example, polyethylene, polyvinyl chloride, polystyrene, and nylon, are made from petroleum. Although their production is cheap, the environmental burden and the resources for their production will be depleted in the future. Therefore, a search for alternatives is needed. A solution has re-emerged in bio-based plastics (Kawashima et al. 2019). A promising group of bioplastics is now presented by polyesters of hydroxyalkanoic acids, that is, polyhydroxyalkanoates (PHA). These environmentally friendly alternatives to petroleum-based polymers are accumulated naturally by numerous prokaryotic microorganisms (Muhammadi et al. 2015; Sabapathy et al. 2020). Unfortunately, less than 1% of the total plastic production comes from the bioplastics industry (Shogren et al. 2019). The main obstacle preventing wider utilization of PHA in viable industrial processes is the cost of the carbon resources and the cost of the fermentation and downstream processing. A promising strategy which might help to reduce the cost of PHA is the use of inexpensive or waste carbon substrates that do not compete with the human food chain (Koller 2018) as well as the employment of extremophilic microorganisms

Significance

Next-generation industrial biotechnology is a concept that relies on the biosynthetic potential of extremophilic microorganisms, and which benefits from the robustness of such processes against contamination by common microflora. In this context, *Tepidimonas taiwanensis* is a very interesting, moderately thermophilic bacterium with great biotechnological potential. It has been reported that it is a potent producer of alkaline proteases, which are enzymes used in large quantities in numerous fields, including but not limited to the detergent industry. Furthermore, it was recently reported that *T. taiwanensis* is also capable of producing polyhydroxyalkanoates, a "green" alternative to petrochemical polymers, from inexpensive waste substrates such as grape pomace. In this work, we describe the complete genome sequence of the bacterium, which is an important step on the route toward complete utilization of the biosynthetic potential of the bacterium that can be further expanded by approaches of genetic engineering and synthetic biology.

(Obruca et al. 2018). Although some pivotal work has been completed and the fact that microorganisms use PHA to store unused energy and carbon in the cytoplasm in the form of intracellular granules is known (Obruca et al. 2018), additional knowledge that can be mined from various genomes of PHA producers is of high importance.

The type strain *Tepidimonas taiwanensis* LGM 22826^T (=BCRC 17406^T, I1-1^T) is a thermophilic, Gram-negative bacterium that was isolated from a hot spring in the Pingtung area in southern Taiwan (Chen et al. 2006). The rod-shaped cells are motile via a single polar flagellum. The bacterium was originally investigated for its strong alkaline protease activity, which is usable in different industries (Gupta et al. 2002). Nevertheless, other important features of the strain remained hidden, which may be due to the missing high-quality complete genome assembly and functional annotation of the genome. Only recently was its ability to utilize glucose and fructose to produce PHA reported (Kourilova et al. 2021). As it is a thermophile, PHA production takes place within the temperature range 45–55 °C, which reduces the risk of microbial contamination. Therefore, the strain presents an ideal organism for utilization under unsterile conditions, known as the "next-generation industrial biotechnology" concept (Chen and Jiang 2018). In this article, we present its first high-quality complete genome sequence. We annotated the genome, identified key genes in PHA metabolism and in coding extracellular proteases, and searched for prophage DNA and CRISPR arrays.

Results and Discussion

Genome Assembly and Properties

The complete genome sequence of *T. taiwanensis* LMG 22826^T was reconstructed using more than 415,000 Oxford Nanopore Technologies (ONT) reads with average length of 17 kb and polished by an additional 2.3 million high-quality (average Phred score $Q \approx 35$) Illumina PE reads. The overall coverage of the final assembly consisting of a single circular chromosome was 2785x. The genome has been deposited at the DDBJ/EMBL/GenBank under accession No CP083911.1.

More than 2.9 Mb long, the genome of *T. taiwanensis* consists of 2,764 genes, some of them organized into 569 predicted operons that comprise two or more structural genes. From 2,700 coding genes, 66 were marked as pseudogenes, which in most cases are made of incomplete gene sequences according to NCBI Prokaryotic Genome Annotation Pipeline (PGAP). All rRNA genes are present in three copies and 16S rRNA copies have similarity >99%. Further analysis of tRNAs encoded in the genome with tRNA-scanSE (Chan and Lowe 2019) revealed the differences between numbers of different isoaceptor tRNAs which can correlate with codon usage bias as has been reported (Rocha 2004). For example, three of four possible types of alanine amino acid isoaceptors are encoded in the genome in the ratio 1:1:3, so the abundance of the codon corresponding to the third isoaceptor is expected to be higher. In addition, the tRNA analysis revealed a high number, precisely 42, of tRNA isoaceptors that can be affected by relatively high GC content (Kanaya et al. 1999), in this case, almost 69%. All genome features of the *T. taiwanensis* genome are summarized in Table 1. Using the complete genome sequence, *Tepidimonas thermarum* was found to be the closest species to *T. taiwanensis*. Whole-genome sequence-based phylogeny of the ten most closely related species is available under supplementary figure S1, Supplementary Material online. Genome similarity of *T. taiwanensis* to these ten species

Table 1
Genomic Features of *Tepidimonas taiwanensis* LMG 22826^T

Feature	Chromosome
Length [bp]	2,915,587
GC content [%]	68.75
Genes	2,764
Operons	569
CDSs	2,700
Pseudogenes	66
ncRNAs	3
rRNAs (5S, 16S, 23S)	3, 3, 3
tRNAs	52

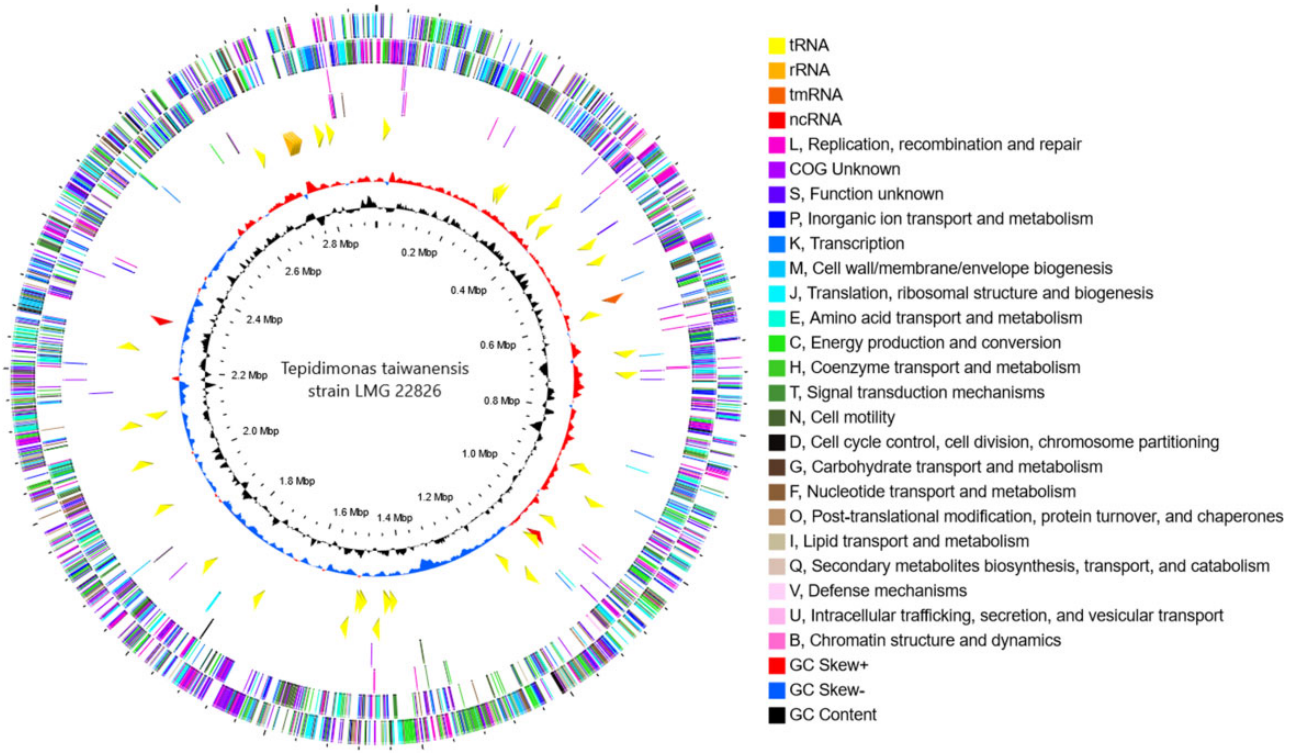

Fig. 1.—A genome map of the *T. taiwanensis* LMG 22826^T. The first two outermost circles represent CDS on the forward and backward strands, respectively. The next two circles contain pseudogenes on the forward and backward strands. Colors represent the functional classification of a COG. The fifth and the sixth circles consist of various types of RNA genes. Two inner circles show GC skew and GC content, respectively.

expressed as digital DNA to DNA hybridization reached values from 19.8% to 25.6%.

Functional Annotation

Protein-coding sequences (CDSs) were divided into 21 categories according to Cluster of Orthologous Groups (COGs). The most abundant known gene function is the "Energy production and conversion group of genes (C)" with 7.59% from all CDSs. Additionally, many genes belong to "Amino acid transport and metabolism group (E)," which makes up 7.44%. The high number of genes in these two groups corresponded to housekeeping functions of cells but was also related to industrially utilizable features, for example, the ability to produce PHA. Its production by the strain *T. taiwanensis* LMG 22826^T was proved recently, and the presence of *phaC* gene was confirmed by PCR (Kourilova et al. 2021). Here, we identified loci of *phaC* (LCC91_05560) and neighboring *phaR* (LCC91_04215) genes that are necessary for PHA production as well as a locus of *phaZ* (LCC91_05500) coding PHA depolymerase.

Unfortunately, almost 17% of coding genes were assigned to "Unknown function (S)" category and 8.85% genes were not recognized at all. All categories with gene counts are available under supplementary table S1, Supplementary Material online.

The arrangement of all genes in the *T. taiwanensis* LMG 22826^T genome is shown in figure 1 where every COG and every type of RNA is distinguished by a different color.

Although the strain was reported to produce extracellular alkaline proteases, due to the lack of genome sequence the enzymes were never identified. KEGG searches revealed 21 orthologues for alkaline proteases, see supplementary table S2, Supplementary Material online. Three of them, *sppA* (LCC91_01535), *degP/htrA* (LCC91_09805), and *prpL* (LCC91_10460), coded for extracellular proteases. Moreover, their predicted molecular weights 36.4, 52.8, and 69.9 kDa matched experimental evidence provided by zymography (Chen et al. 2006). These enzymes might be of great industrial importance. For example, the optimum enzymatic activity of protease IV coded by *sppA* was reported to be at pH 10 and temperature of 45 °C (Engel et al. 1998), which makes this enzyme utilizable in the detergent industry for production of washing powder.

The genome was searched for CRISPR arrays and, as a consensus from three prediction methods, four large arrays were reported in the *T. taiwanensis* LMG 22826^T genome. The largest array consisted of 42 spacers and was 2,590 bp long. Moreover, *cas* genes, such as *cas1* or *cas2*, whose proteins are responsible for spacers acquisition into CRISPR arrays (Yosef et al. 2012), were found in the genome. Unfortunately, neither of these genes was the gene that encodes the Cas9

protein well known for its high utilization in genetic engineering. The summary of CRISPR arrays is included under supplementary in table S3, Supplementary Material online.

Finally, the presence of prophage DNA was checked. PHASTER found three prophages: two of them were labeled as incomplete and one as intact prophage. The sequence that was labeled as intact prophage consists of 70 proteins, and 46 of them match to phage proteins such as phage tail protein or phage virion protein. Eight of these proteins correspond to *Escherichia* phage *vB_EcoM_ECO1230-10*, which has not been reported as a phage able to survive life conditions of thermophilic bacteria such as *T. taiwanensis*. Although Prophage Hunter did not label any of the phages as active, the previously mentioned phage sequence achieved the highest score and corresponded to *Acidithiobacillus* phage *Aca*ML1, which has been found in thermophilic, acidophilic bacterium *Acidithiobacillus caldus* (Covarrubias et al. 2018), so it is possible to presume this phage has the ability to survive the life conditions of *T. taiwanensis*. Overall, the reliable statement of whether the prophage is active would need further analysis. The summary table of present prophage DNA from PHASTER tool is available under supplementary table S4, Supplementary Material online.

Materials and Methods

Growth Conditions, DNA Extraction, and Sequencing

Bacterial cultures of *T. taiwanensis* LMG 22826^T were purchased from Belgian-coordinated collections of microorganisms. First, the bacterial culture was cultivated in complex medium (nutrient broth with 1% peptone, HiMedia) for 24 h at 50 °C with constant shaking (180 rpm). Subsequently, *T. taiwanensis* LMG 22826^T was cultivated in a mineral salt medium composed of $Na_2HPO_4 \cdot 12\ H_2O$ (9.0 g/l), KH_2PO_4 (1.5 g/l), NH_4Cl (1.0 g/l), $MgSO_4 \cdot 7\ H_2O$ (0.2 g/l), $CaCl_2 \cdot 2\ H_2O$ (0.02 g/l), $Fe^{(III)}NH_4$ citrate (0.0012 g/l), yeast extract (0.5 g/l), 1 ml/l of microelements solution containing EDTA (50.0 g/l), $FeCl_3 \cdot 6\ H_2O$ (13.8 g/l), $ZnCl_2$ (0.84 g/l), $CuCl_2 \cdot 2\ H_2O$ (0.13 g/l), $CoCl_2 \cdot 6\ H_2O$ (0.1 g/l), $MnCl_2 \cdot 6\ H_2O$ (0.016 g/l) and H_3BO_3 (0.1 g/l) dissolved in distilled water, and a glucose (20.0 g/l) as a carbon substrate. The parameters of cultivation conditions on mineral salt medium corresponded to the cultivation conditions on complex medium.

Genomic DNA was extracted using MagAttract HMW DNA kit (Qiagene, NL). The DNA purity was checked using NanoDrop (Thermo Fisher Scientific, USA), the concentration was measured using Qubit 2.0 Fluorometr (Thermo Fisher Scientific, USA), and the proper length of the isolated DNA was confirmed using Agilent 4200 TapeStation (Agilent Technologies, USA). The sequencing library for Oxford Nanopore sequencing was prepared using Ligation Sequencing 1D Kit (Oxford Nanopore Technologies, UK). The sequencing was performed using the R9.4.1 Flow Cell and the MinION platform (Oxford Nanopore Technologies, UK). The sequencing library for short read sequencing was prepared using KAPA HyperPlus kit and was carried out using Miseq Reagent kit v2 (500 cycles) and Illumina MiSeq platform (Illumina, USA).

Genome Assembly

Long ONT reads were basecalled with Guppy basecaller v3.4.4 (https://nanoporetech.com/, last accessed November 8, 2021) and the quality was checked with MinIONQC v1.4.1 (Lanfear et al. 2019). Similarly, Illumina paired-end reads quality was checked with FastQC v0.11.5 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/, last accessed November 17, 2021) and low-quality ends and adapters were trimmed with Trimmomatic v0.39 (Bolger et al. 2014).

Nanopore reads were assembled using Flye v2.8.1 (Kolmogorov et al. 2019). The final assembly was then polished with four rounds of Racon v1.4.20 (Vaser et al. 2017) using Minimap2 v2.17 (Li 2018) to map long reads. After that, two rounds of polishing with Medaka v1.2.5 (https://nanoporetech.github.io/medaka/, last accessed November 9, 2021) were applied. In addition, Illumina reads were mapped to the polished assembly with BWA aligner v0.7.17 (Li and Durbin 2009) and additional corrections were performed using three rounds of Pilon v1.24 (Walker et al. 2014) polishing. The final step was to rearrange the genome according to the predicted replication origin (*oriC*) with Ori-Finder 2 (Luo et al. 2014) so that the genome starts with the *dnaA* gene.

Genome Annotation and Analysis

The genome of *T. taiwanensis* LMG 22826^T was annotated with the PGAP v5.3 (Tatusova et al. 2016). The prediction of operons was performed using Operon-mapper (Taboada et al. 2018). Protein-coding genes were assigned to COGs according to eggNOG database (Huerta-Cepas et al. 2019) through eggNOG mapper (Cantalapiedra et al. 2021). Genome-based phylogeny and comparison to the most closely related species was done with the type strain genome server (Meier-Kolthoff and Göker 2019). Furthermore, the genome sequence was searched for prophage DNA with the PHASTER tool (Arndt et al. 2016), Prophage Hunter (Song et al. 2019), and the occurrences of CRISPR arrays and *cas* genes were inspected through CRISPRDetect (Biswas et al. 2016) and CRISPRCasFinder (Couvin et al. 2018). The results were compared with PGAP annotation and only arrays predicted by at least two tools were reported. Finally, the genome sequence was processed with CGView (Stothard and Wishart 2005) to construct the circular genome map. Identification of orthologous genes coding key enzymes was made by manual BLAST (Altschul et al. 1990) searches and by browsing the KEGG database (Kanehisa and Goto 2000). Molecular weights were predicted from the primary structure of protein sequence using ExPASy (Gasteiger et al. 2003).

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgment

This study was funded by the Czech Science Foundation (GACR) (Project No. GA19-20697S).

Data Availability

The whole-genome sequence has been deposited at DDBJ/ENA/GenBank under the accession number CP083911.1. The NCBI BioProject and BioSample accession numbers are PRJNA764859 and SAMN21531155, respectively. The raw reads have been deposited at NCBI SRA database under the accession numbers SRR16956434 (paired-end Illumina) and SRR16956433 (Oxford Nanopore Technologies).

Literature Cited

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol. 215(3):403–410.

Arndt D, et al. 2016. PHASTER: a better, faster version of the PHAST phage search tool. Nucleic Acids Res. 44(W1):W16–W21.

Biswas A, Staals RHJ, Morales SE, Fineran PC, Brown CM. 2016. CRISPRDetect: a flexible algorithm to define CRISPR arrays. BMC Genomics 17(1):356.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30(15):2114–2120.

Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J. 2021. eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. Mol Biol Evol. 38(12):5825–5829.

Chan PP, Lowe TM. 2019. tRNAscan-SE: searching for tRNA genes in genomic sequences. Methods Mol Biol. 1962:1.

Chen GQ, Jiang XR. 2018. Next generation industrial biotechnology based on extremophilic bacteria. Curr Opin Biotechnol. 50:94–100.

Chen TL, Chou YJ, Chen WM, Arun B, Young CC. 2006. *Tepidimonas taiwanensis* sp. nov., a novel alkaline-protease-producing bacterium isolated from a hot spring. Extremophiles 10(1):35–40.

Couvin D, et al. 2018. CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. Nucleic Acids Res. 46(W1):W246–W251.

Covarrubias PC, et al. 2018. Occurrence, integrity and functionality of AcaML1–like viruses infecting extreme acidophiles of the *Acidithiobacillus* species complex. Res Microbiol. 169(10):628–637.

Engel LS, Hill JM, Caballero AR, Green LC, O'Callaghan RJ. 1998. Protease IV, a unique extracellular protease and virulence factor from *Pseudomonas aeruginosa*. J Biol Chem. 273(27):16792–16797.

Gasteiger E, et al. 2003. ExPASy: the proteomics server for in-depth protein knowledge and analysis. Nucleic Acids Res. 31(13):3784–3788.

Gupta R, Beg QK, Khan S, Chauhan B. 2002. An overview on fermentation, downstream processing and properties of microbial alkaline proteases. Appl Microbiol Biotechnol. 60:4. 60:381–395.

Huerta-Cepas J, et al. 2019. EggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res. 47(D1):D309–D314.

Kanaya S, Yamada Y, Kudo Y, Ikemura T. 1999. Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of *Bacillus subtilis* tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis. Gene 238(1):143–155.

Kanehisa M, Goto S. 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 28(1):27–30.

Kawashima N, Yagi T, Kojima K. 2019. How do bioplastics and fossil-based plastics play in a circular economy? Macromol Mater Eng. 304(9):1900383.

Koller M. 2018. A review on established and emerging fermentation schemes for microbial production of polyhydroxyalkanoate (PHA) biopolyesters. Fermentation 4(2):30.

Kolmogorov M, Yuan J, Lin Y, Pevzner PA. 2019. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol. 37(5):540–546.

Kourilova X, et al. 2021. Biotechnological conversion of grape pomace to poly(3-hydroxybutyrate) by moderately thermophilic bacterium *Tepidimonas taiwanensis*. Bioengineering 8(10):141.

Lanfear R, Schalamun M, Kainer D, Wang W, Schwessinger B. 2019. MinIONQC: fast and simple quality control for MinION sequencing data. Bioinformatics 35(3):523–525.

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34(18):3094–3100.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25(14):1754–1760.

Luo H, Zhang CT, Gao F. 2014. Ori-Finder 2, an integrated tool to predict replication origins in the archaeal genomes. Front Microbiol. 5:482.

Meier-Kolthoff JP, Göker M. 2019. TYGS is an automated high-throughput platform for state-of-the-art genome-based taxonomy. Nat Commun. 10:1–10.

Muhammadi S, Afzal M, Hameed S. 2015. Bacterial polyhydroxyalkanoates-eco-friendly next generation plastic: production, biocompatibility, biodegradation, physical properties and applications. Green Chem Lett Rev.8:56–77.

Obruca S, Sedlacek P, Koller M, Kucera D, Pernicova I. 2018. Involvement of polyhydroxyalkanoates in stress resistance of microbial cells: biotechnological consequences and applications. Biotechnol Adv. 36(3):856–870.

Rocha EPC. 2004. Codon usage bias from tRNA's point of view: redundancy, specialization, and efficient decoding for translation optimization. Genome Res. 14(11):2279–2286.

Sabapathy PC, et al. 2020. Recent developments in polyhydroxyalkanoates (PHAs) production – A review. Bioresour Technol. 306:123132.

Shogren R, Wood D, Orts W, Glenn G. 2019. Plant-based materials and transitioning to a circular economy. Sust Prod Consump. 19:194–215.

Song W, et al. 2019. Prophage Hunter: an integrative hunting tool for active prophages. Nucleic Acids Res. 47(W1):W74–W80.

Stothard P, Wishart DS. 2005. Circular genome visualization and exploration using CGView. Bioinformatics 21(4):537–539.

Taboada B, Estrada K, Ciria R, Merino E. 2018. Operon-mapper: a web server for precise operon identification in bacterial and archaeal genomes. Bioinformatics 34(23):4118–4120.

Tatusova T, et al. 2016. NCBI prokaryotic genome annotation pipeline. Nucleic Acids Res. 44(14):6614–6624.

Vaser R, Sović I, Nagarajan N, Šikić M. 2017. Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res. 27(5):737–746.

Walker BJ, et al. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One 9(11):e112963.

Yosef I, Goren MG, Qimron U. 2012. Proteins and DNA elements essential for the CRISPR adaptation process in *Escherichia coli*. Nucleic Acids Res. 40(12):5569–5576.

Associate editor: Howard Ochman