AMERICAN SOCIETY FOR MICROBIOLOGY
Microbiology® Resource Announcements

Identification of Putative β-Galactosidase Genes in the Genome of *Lactobacillus helveticus* OSU-PECh-4A

Israel García-Cano,[a] Alejandra Escobar-Zepeda,[b,c] Silvette Ruiz-Ramírez,[a] Diana Rocha-Mendoza,[a] Rafael Jiménez-Flores[a]

[a]Department of Food Science and Technology, The Ohio State University, Columbus, Ohio, USA
[b]Host-Microbiota Interactions Laboratory, Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, United Kingdom
[c]Microbiome Informatics Team, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, United Kingdom

ABSTRACT The *Lactobacillus helveticus* OSU-PECh-4A strain, from the Ohio State University Parker Chair collection, produces exceptional β-galactosidase activity using acid whey as a culture medium, compared with a commercial broth. The strain has a genome sequence of 1,834,843 bp, and its GC content is 36.69%. Using InterProScan v5.50-84.0 software, four genes with putative β-galactosidase function were found.

Lactobacillus helveticus strain OSU-PECh-4A was isolated from commercial fermented milk (Columbus, OH, USA). Twenty-five grams of sample was mixed with 225 mL of sterile saline solution (0.85% NaCl [pH 7.0]). Serial dilutions were performed and plated on MRS agar (BD Difco, USA). The plates were incubated under aerobic conditions for 16 h at 37°C. The colonies were selected based on phenotypic features, i.e., shape, color, and texture. Using a synthetic substrate, the OSU-PECh-4A strain showed 5 times more β-galactosidase activity when it was cultivated in acid whey (AW) as a medium, compared with the commercial broth (MRS broth). The relative expression of the *bgal*-620 gene was 3 times higher in AW than in the MRS medium (1). It has been reported that lactic acid bacteria (LAB) have two overlapping genes (*lacL* and *lacM*) for β-galactosidase production (2). However, LAB can contain one, two, or three genes for the production of β-galactosidase in their genomes. The draft genome sequence of this strain should facilitate the identification of the putative genes encoding prospective β-galactosidase proteins and the understanding of the high activity levels shown by *L. helveticus* OSU-PECh-4A.

For genomic DNA (gDNA) extraction from *L. helveticus* OSU-PECh-4A, a purification kit (Wizard gDNA kit; Promega, USA) was used. Previously, the cells were grown in MRS broth (BD Difco, USA) and recovered by centrifugation at 10,000 × *g* for 10 min. The concentration and quality of the gDNA were measured using the PicoGreen method (catalog number P7589; Life Technologies, USA) and a 2200 TapeStation system (Agilent Technologies, Inc., USA), respectively. The DNA concentration used for the sequencing step was ~50 ng/μL, with a DNA Integrity Number (DIN) value (with the Agilent 2200 TapeStation system and the Agilent gDNA ScreenTape assay) of 9.7.

The gDNA was used for Illumina high-throughput sequencing (NovaSeq 6000 S4 system; Illumina). The library was constructed following the TruSeq DNA PCR-free protocol, and 151 cycles of paired-end sequencing were performed at Psomagen (Rockville, MD, USA). A total of 14.28 million raw reads were processed. Default parameters were used except where otherwise noted. For quality control, we used Fastp v1.14.5 software (3); 98.88% of reads passed quality control and were used for genomic assembly with the SPAdes genome assembler v3.15 in mode --careful (4). We filtered out fragments shorter than 500 bp and computed the assembly statistics using in-house-built scripts (available at https://github.com/Ales-ibt/in_house_scripts). The *L. helveticus* OSU-PECh-4A genome is fragmented in 146 contigs (N_{50}, 20,442 bp; L_{50}, 27; N_{90}, 6,481 bp; L_{90}, 84), likely due to the presence of many repetitive sequences according to the large number of transposases encountered (130 genes).

Editor Irene L. G. Newton, Indiana University, Bloomington

Address correspondence to Rafael Jiménez-Flores, jimenez-flores.1@osu.edu.

The authors declare no conflict of interest.

Received 29 July 2021
Accepted 1 December 2021
Published 6 January 2022

TABLE 1 Identification of genes with putative β-galactosidase function in *L. helveticus* OSU-PECh-4A using Prokka and InterProScan software[a]

Prokka gene ID	Contig ID	Nucleotide position Start	 End	Strand	Gene name	UniProtKB annot	IPS database annot	IPS database ID(s)	IPS annot
NIPFOCJE_01570	NODE_62_length_11113_cov_1011.571856	187	2073	+	*lacL*	β-Galactosidase large subunit	Pfam	PF00703, PF02836, PF02837	Glycosyl hydrolase family 2; glycosyl hydrolase family 2, TIM barrel domain; glycosyl hydrolase family 2, sugar-binding domain
NIPFOCJE_01571	NODE_62_length_11113_cov_1011.571856	2057	3013	+	*lacM*	β-Galactosidase small subunit	Pfam	PF02929	β-Galactosidase small chain
NIPFOCJE_01818	NODE_88_length_5952_cov_940.937021	485	1558	−	*lacZ*	β-Galactosidase LacZ	Pfam	PF02449	β-Galactosidase
NIPFOCJE_01932	NODE_107_length_3404_cov_922.981671	1303	1884	−	*lacG*	6-Phospho-β-galactosidase	Pfam	PF00232	Glycosyl hydrolase family 1

[a] ID, identification; annot, annotation; IPS, InterProScan; TIM, triosephosphateisomerase.

The genome size is 1,834,843 bp, and the GC content is 36.69%. According to CheckM v1.1.2 (5), this genome has 99.03% completeness and 0.00% contamination. Taxonomic assignment to *L. helveticus* was corroborated using GTDB-Tk v1.5.0 (6). Additionally, we computed the average nucleotide identity (ANI) versus 21 complete genome assemblies of *Lactobacillus helveticus* strains from RefSeq using the FastANI tool v1.3 (7). This analysis revealed that the two closest reference strains are *L. helveticus* strain D76 (GenBank accession number CP016827.1) and *L. helveticus* isolate MGYG-HGUT-02384 (GenBank accession number LR698986.1), both with 99.96% ANI.

According to gene prediction and functional annotation by NCBI Prokaryotic Genome Annotation Pipeline (PGAP) v5.2, this assembly has 1,929 total genes, 2 copies of 16S rRNA, and 51 genes encoding tRNAs. Additional annotation of functional domains in the amino acid sequences retrieved by Prokka v1.14.5 (8) was performed using InterProScan v5.50-84.0 (9) for the identification of genes with putative β-galactosidase function (Table 1). Four genes with this putative function were found in the *L. helveticus* OSU-PECh-4A genome. Two genes are contiguous and represent the large and small β-galactosidase subunits. The other two genes encode different proteins. The gene sequences and the amino acid sequences for the four genes detected did not show similarity to each other, as observed by multiple sequence alignment using MUSCLE v3.32.0 (10).

Data availability. The *Lactobacillus helveticus* OSU-PECh-4A draft genome was deposited in the NCBI database under the BioProject and BioSample accessions numbers PRJNA746544 and SAMN20209453, respectively. The Sequence Read Archive (SRA) accession number is SRR15131330. The GenBank accession number for the whole-genome sequence is JAHWBM000000000, and the GenBank accession number for the 16S rRNA gene is MW810614.1.

ACKNOWLEDGMENTS

This study was funded by the Parker Endowment Chair in Dairy Foods at the Ohio State University.

We thank Molly J. Davis (Food Science and Technology Department, The Ohio State University, Columbus, OH, USA) for editing the manuscript.

REFERENCES

1. Kolev P, Rocha-Mendoza D, Ruiz-Ramírez S, Ortega-Anaya J, Jiménez-Flores R, García-Cano I. 2021. Screening and characterization of β-galactosidase activity in lactic acid bacteria for the valorization of acid whey. JDS Commun https://doi.org/10.3168/jdsc.2021-0145.
2. Kittibunchakul S, Pham ML, Tran AM, Nguyen TH. 2019. β-Galactosidase from *Lactobacillus helveticus* DSM 20075: biochemical characterization and recombinant expression for applications in dairy industry. Int J Mol Sci 20:947. https://doi.org/10.3390/ijms20040947.
3. Chen W, Zhou Y, Chen Y, Gu J. 2018. Fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 34:i884–i890. https://doi.org/10.1093/bioinformatics/bty560.
4. Prjibelski A, Antipov D, Meleshko D, Lapidus A, Korobeynikov A. 2020. Using SPAdes de novo assembler. Curr Protoc Bioinformatics 70:e102. https://doi.org/10.1002/cpbi.102.
5. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res 25:1043–1055. https://doi.org/10.1101/gr.186072.114.
6. Chaumeil PA, Mussig AJ, Hugenholtz P, Parks DH. 2019. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. Bioinformatics 36:1925–1927. https://doi.org/10.1093/bioinformatics/btz848.
7. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. 2018. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun 9:5114. https://doi.org/10.1038/s41467-018-07641-9.
8. Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics 30:2068–2069. https://doi.org/10.1093/bioinformatics/btu153.
9. Mitchell AL, Attwood TK, Babbitt PC, Blum M, Bork P, Bridge A, Brown SD, Chang HY, El-Gebali S, Fraser MI, Gough J, Haft DR, Huang H, Letunic I, Lopez R, Luciani A, Madeira F, Marchler-Bauer A, Mi H, Natale DA, Necci M, Nuka G, Orengo C, Pandurangan AP, Paysan-Lafosse T, Pesseat S, Potter SC, Qureshi MA, Rawlings ND, Redaschi N, Richardson LJ, Rivoire C, Salazar GA, Sangrador-Vegas A, Sigrist CJA, Sillitoe I, Sutton GG, Thanki N, Thomas PD, Tosatto SCE, Yong SY, Finn RD. 2019. InterPro in 2019: improving coverage, classification and access to protein sequence annotations. Nucleic Acids Res 47:D351–D360. https://doi.org/10.1093/nar/gky1100.
10. Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797. https://doi.org/10.1093/nar/gkh340.