Data in Brief 38 (2021) 107437

Contents lists available at ScienceDirect

Data in Brief

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

Data Article

Whole-genome sequence data of the proteolytic and bacteriocin producing strain *Enterococcus faecalis* PK23 isolated from the traditional Halitzia cheese produced in Cyprus

Konstantinos Papadimitriou[a,b,*], Anastasia Venieraki[c], Markella Tsigkrimani[a], Panagiotis Katinakis[d], Panagiotis N. Skandamis[a,*]

[a] *Laboratory of Food Quality Control and Hygiene, Department of Food Science and Human Nutrition, Agricultural University of Athens, Iera Odos 75, Athens 11855, Greece*
[b] *Department of Food Science and Technology, University of Peloponnese, Antikalamos 24100, Greece*
[c] *Laboratory of Plant Pathology, Crop Science Department, Agricultural University of Athens, Iera Odos 75, Athens 118 55, Greece*
[d] *General and Agricultural Microbiology Laboratory, Crop Science Department, Agricultural University of Athens, Iera Odos 75, Athens 118 55, Greece*

ARTICLE INFO

Article history:
Received 13 July 2021
Accepted 27 September 2021
Available online 30 September 2021

Keywords:
Lactic acid bacteria
Genomics
Enterococcus
Bacteriocin
Plasmid
Proteolysis
Cheese
Adaptation

ABSTRACT

Halitzia is a traditional white-brined cheese produced by a limited number of producers in Cyprus. During a survey of the microbiome of a number of different Halitzia samples, we identified a bacterial strain that exhibited enhanced proteolytic activity compared to the other isolates. The strain was further studied, and it was assigned as *Enterococcus faecalis* PK23. We proceeded with sequencing of its whole genome using Illumina technology. Initial sequencing and assembly produced 116 scaffolds with a length of 3,149,036 bp. Comparison with the available *E. faecalis* genomes revealed that the strain PK23 exhibited high levels of identity to the genome sequence of *E. faecalis* isolate 26975_2#180 deposited in GenBank as a single complete contig. From the 116 scaffolds 106 could be aligned to the genome of isolate 26975_2#180 leading to a chromosomal length of 3,132,784

* Corresponding authors at: Laboratory of Food Quality Control and Hygiene, Department of Food Science and Human Nutrition, Agricultural University of Athens, Iera Odos 75, Athens 11855, Greece.
E-mail addresses: k.papadimitriou@uop.gr (K. Papadimitriou), pskan@aua.gr (P.N. Skandamis).

https://doi.org/10.1016/j.dib.2021.107437

bp with a GC content of 37.3%. From the remaining 10 scaffolds, five showed similarity to plasmid sequences. More specifically, scaffold 54 showed high identity with most part of plasmid pEF1071 of *E. faecalis* strain BFE 1071, which carries the gene cluster involved in the biosynthesis of enterocins 1071A and 1071B, while scaffold 77 showed high identity with the entire sequence of the unnamed_5 cryptic plasmid of *Enterococcus faecium* strain PR05720-3. The other three scaffolds were only short parts of larger plasmids. The remaining five scaffolds which could not be related to any plasmid sequence most probably constitute chromosomal sequences present in strain PK23 but absent from isolate 26975_2#180. Their total length was around 2.7 kb, which does not affect the sequence of the PK23 pseudochromosome in a major way. The whole-genome sequence annotation of strain PK23 identified 3161 coding sequences and 62 RNA sequences. The results from the Rapid Annotation using Subsystem Technology (RAST) version 2.0 server indicated the presence of seven putative genes which were related to the subsystem of Protein Degradation. This dataset provides a first overview of the proteolytic and bacteriocin producing properties of *E. faecalis* PK23. The dataset may also be used in future experiments which could shed light on the adaptation of the strain in the dairy environment and its role in cheese production.

Specifications Table

Subject	Food Science: Food Microbiology
Specific subject area	Genome sequencing and in silico analysis
Type of data	Genome sequence; Tables and Figures
How data were acquired	Genome sequencing: Illumina HiSeq 2000 platform, Trimming of adaptors: BBDuk, De novo sequence assembly: SPAdes-3.12.0, Chromosomal alignments: r2cat and MAUVE, Sequence annotation: Rapid Annotation using Subsystem Technology (RAST) version 2.0 server, Prokaryotic-genome Analysis Tool (PGAT), Additional bioinformatics analysis: Average Nucleotide Identity (ANI) calculator, Genome-to-Genome Distance Calculator 2.1 (GGDC), DNAPlotter, nucleotide Basic Local Alignment Search Tool (BLASTn)
Data format	Raw Illumina paired-end sequence reads in fastq files, fasta and genbank file formats of annotated sequence scaffolds
Parameters for data collection	*Enterococcus faecalis* strain PK23 was isolated from traditional Halitzia cheese produced in Cyprus and the genomic DNA sequenced was isolated from the pure culture
Description of data collection	Genomic DNA was isolated from the pure culture of *E. faecalis* PK23 and it was sequenced by Illumina HiSeq 2000 platform resulting in two fastq files of raw paired-end reads. Trimming of adaptors and de novo assembly were performed with BBDuk and SPAdes-3.12.0, respectively. A reference genome was identified with the ANI calculator and the GGDC. Scaffolds were aligned against the reference sequence with the r2cat and MAUVE tools. Scaffolds were annotated with the RAST version 2.0 server and PGAT. DNAPlotter was employed to draw the pseudochromosome map of stain PK23. Scaffolds were also analysed by BLASTn to identify putative chromosomal or plasmid sequences
Data source location	Laboratory of Food Quality Control and Hygiene, Department of Food Science and Human Nutrition, Agricultural University of Athens, 11855 Athens, Greece

(*continued on next page*)

Data accessibility	Data are deposited in the respective databases and are publicly available. Raw sequence data of strain Enterococcus faecalis PK23 was deposited in the Sequence Read Archive (SRA) under accession number SRX11406112 (https://trace.ncbi.nlm.nih.gov/Traces/sra/?run=SRR15096257). The whole-genome sequence of *Enterococcus faecalis* PK23 has been deposited in GenBank under accession number JAHBBU000000000 (https://www.ncbi.nlm.nih.gov/nuccore/JAHBBU000000000). In GenBank the sequence can be found annotated by PGAT. The BioProject ID in GenBank is: PRJNA729111 The annotated genome is also available through the Rast version 2.0 server for anyone logging in with the guest account under the genome ID 1351.5033.

Value of the Data

- *Enterococcus faecalis* has been identified very frequently as part of the microbiome of artisanal cheeses. The genome sequence of strain PK23 may aid understanding the adaptation mechanisms which underpin its growth in the dairy environment.
- Data included in this manuscript may be useful for researchers in the field of Dairy Microbiology.
- Data presented here may be useful to understand the differences and similarities between commensal/pathogenic and dairy *E. faecalis* strains through comparative and evolutionary genomics.
- *E. faecalis* PK23 presents potentially important technological properties which warranty further investigation, including its proteolytic potential and bacteriocin production.

1. Data Description

Halitzia is a rare white-brined cheese produced in Cyprus. It is named after its shape which resembles stones or pebbles [1]. During a recent screening of the microbial ecosystem of a number of Halitzia cheese samples, we isolated a strain exhibiting enhanced proteolytic and bacteriocin producing activity compared to other isolates (our unpublished results). The strain was identified and assigned as *Enterococcus faecalis* PK23. The proteolytic activity of enterococci has been suggested to be more pronounced than other lactic acid bacteria (LAB) [2]. It is an important technological property since it contributes to the cheese ripening process, the generation of bioactive peptides or even the reduction of the allergenicity of bovine milk proteins [2–4]. Nevertheless, enterococcal proteolytic enzymes have also been related to the pathogenicity of clinical strains [5]. Furthermore, the ability of enterococci to produce bacteriocins against foodborne pathogens and food spoilage microorganisms is also an important technological trait [6].

The initial assembly of the paired-end Ilumina sequencing reads resulted in 116 scaffolds. After manual BLASTn searches [7], a number of closely related *E. faecalis* genomes were identified and downloaded from GenBank. Further analysis with the Average Nucleotide Identity (ANI) calculator [8] and the Genome-to-Genome Distance Calculator (GGDC) 2.1 [9] revealed that the closest related genome currently available for strain PK23 was the chromosomal sequences of *E. faecalis* isolate 26975_2#180. The two strains exhibited an ANI value of 99.88 and a dDDH value of 99.20. Alignment of the 116 PK23 scaffolds against the sequence of isolate 26975_2#180 resulted in the ordering of 106 of them (Fig. 1A). The ordered scaffolds had a length of 3,132,784 bp and a GC content of 37.3% (Table 1). The circular map of the sequence indicated that the organization of the PK23 pseudochromosome had a typical bacterial GC skew from the replication initiation region towards the replication termination region (Fig. 1B). Five additional scaffolds of approx. 2.7 kb could be correlated to chromosomal *E. faecalis* regions according to BLASTn

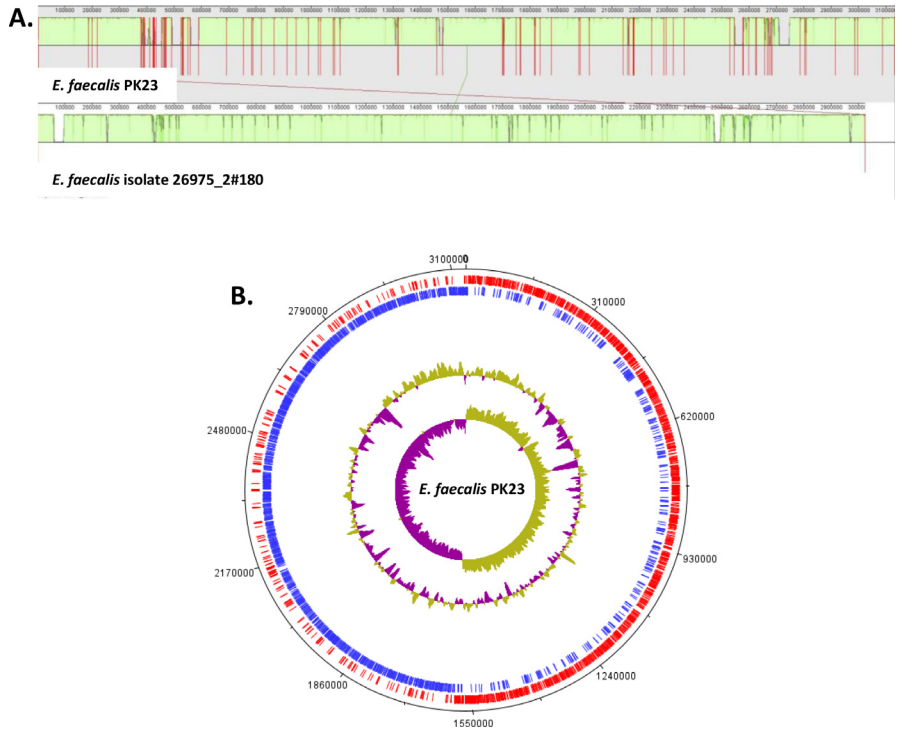

Fig. 1. **(A)** Alignment of the *Enterococcus faecalis* PK23 pseudochromosome (i.e. concatenated ordered scaffolds) against the chromosome sequence of *E. faecalis* isolate 26975_2#180 using MAUVE. MAUVE uses local collinear blocks to align regions of high identity. **(B)** Circular map of the PK23 pseudochromosome generated using DNAPlotter. Genomic features drawn from the periphery to the centre of the map: 1. Forward CDSs (red); 2. Reverse CDSs (blue); 3. %GC plot; 4. GC skew.

Table 1
Statistics of the assembled sequence of *Enterococcus faecalis* PK23 according to the RAST version 2.0 server.

Sequence trait	Value
Size (bp)	3,149,036
GC Content (%)	37.3
Number of Scaffolds	116
Number of Subsystems	253
Number of Coding Sequences	3,161
Number of RNAs	62

searches. These scaffolds could not be ordered with the rest of the PK23 scaffolds as they were probably absent from the chromosome of isolate 26975_2#180. Finally, five scaffolds exhibited high identity to plasmid sequences. More specifically, scaffold 54 showed high identity with the entire length of plasmid pEF1071 of *E. faecalis* strain BFE 1071 [10] while scaffold 77 showed high identity with the entire sequence of the cryptic plasmid unnamed_5 of *Enterococcus faecium* strain PR05720-3. The final three scaffolds (i.e. scaffold 88, 105 and 106) represented short fragments of larger plasmids.

Annotation with Rapid Annotation using Subsystem Technology (RAST) version 2.0 server [11] of the whole genome sequence of *E. faecalis* PK23 identified 3161 coding sequences and 62 RNA sequences (Table 1). From the total number of proteins encoded in the genome of strain PK23 27% and 73% could or could not be assigned in subsystem categories, respectively. The most

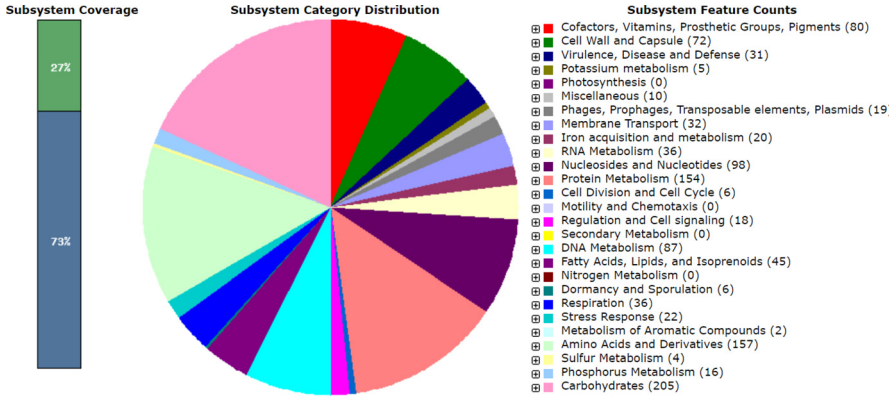

Fig. 2. Analysis of the protein encoding genes (pegs) of the *Enterococcus faecalis* PK23 whole-genome sequence assigned to subsystems categories according to the RAST version 2.0 server. The bar on the left presents the percentage of pegs assigned to subsystems (green) and the pegs which could not be placed into any subsystem (blue). The pie chart in the centre depicts the subsystem category distribution. The coloured categories on the right indicate the subsystem feature counts.

Table 2
Protein encoding genes (pegs) of *Enterococcus faecalis* PK23 assigned to the subsystems category of Protein Degradation according to the RAST version 2.0 server.

Subsystem	Role	peg
Aminopeptidases	Aminopeptidase S	2688
Metallocarboxypeptidases	D-alanyl-D-alanine carboxypeptidase	652
		2225
	Thermostable carboxypeptidase 1	304
Protein degradation	Aminopeptidase YpdF	1428
	Aminopeptidase C	2160
Omega peptidases	Pyrrolidone-carboxylate peptidase	2119

abundant subsystem feature counts were Carbohydrates (205 counts), Amino Acids and Derivatives (157 counts) and Protein Metabolism (154 counts) (Fig. 2). Given the proteolytic activity of strain PK23, preliminary analysis identified seven putative proteolytic enzymes in subsystems counts for Protein Degradation (Table 2). Additionally, annotation of scaffold 54 showed that it carries the biosynthetic genes for enterocins 1071A and 1071B described previously for plasmid pEF1071 [10]. This observation may explain the antimicrobial activity of the strain against certain foodborne pathogens (our unpublished results).

2. Experimental Design, Materials and Methods

E. faecalis PK23 was isolated from traditional Halitzia cheese. The culture was routinely grown in M17 broth (Oxoid) at 37 °C statically. High quality genomic DNA was extracted from the PK23 strain with the PureLink® Genomic DNA Mini Kit (Invitrogen, Life Technologies) according to the manufacturer's instructions. Genome sequencing was performed by SNPsaurus (Eugene, OR) using an Illumina HiSeq 2000 platform (Illumina, CA). Standard workflows were followed for library preparation, sequencing, read trimming and assembly. In detail, a Nextera kit (Illumina) was used for library preparation, followed by 2 × 150-bp paired-end read sequencing, trimming of adaptors with BBDuk (https://sourceforge.net/projects/bbmap) and scaffold assembly with SPAdes-3.12.0 using default parameters [12]. This analysis generated total sequence corresponding to > 200x coverage of the PK23 genome. As mentioned above, the closest *E. faecalis* genomes to the PK23 scaffolds were identified by manual BLASTn searches [7]. These genomes were then

analysed against the PK23 scaffolds with the ANI calculator [8] and the GGDC 2.1 [9]. PK23 scaffolds were ordered against the chromosomal sequence of *E. faecalis* isolate 26975_2#180 using the r2cat [13] and MAUVE [14] tools. DNAPlotter was employed for drawing the circular map of the PK23 pseudochromosome [15]. The scaffolds which could not be aligned to the reference genome were analysed with BLASTn to determine whether they derived from chromosomal or plasmid sequences. The entire whole-genome sequence of strain PK23 was annotated with the RAST version 2.0 server [11] and the Prokaryotic-genome Analysis Tool (PGAT) [16].

3. Nucleotide Sequence Accession Number

Raw sequence data of strain *Enterococcus faecalis* PK23 was deposited in the Sequence Read Archive (SRA) under accession number SRX11406112. The Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the accession JAHBBU000000000. The version described in this paper is version JAHBBU010000000.

Ethics Statement

N/A.

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.

CRediT Author Statement

Konstantinos Papadimitriou: Supervision, Conceptualization, Methodology, Data curation, Writing – original draft; **Anastasia Venieraki:** Methodology, Writing – original draft; **Markella Tsigkrimani:** Formal analysis, Writing – original draft; **Panagiotis Katinakis:** Conceptualization, Methodology, Data curation, Writing – original draft; **Panagiotis N. Skandamis:** Supervision, Conceptualization, Methodology, Data curation, Writing – original draft.

Acknowledgments

This research has been co-financed by the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH–CREATE–INNOVATE: "*Improvement of the added value and competitiveness of the Greek artisanal cheeses by exploiting the multi-omics approach and bioinformatics tools*" (T1EΔK-02087)

References

[1] P. Papademas, M. Aspri, M. Mariou, S.E. Dowd, M. Kazou, E. Tsakalidou, Conventional and omics approaches shed light on Halitzia cheese, a long-forgotten white-brined cheese from Cyprus, Int. Dairy J. 98 (2019) 72–83, doi:10.1016/j.idairyj.2019.06.010.

[2] A. Bhardwaj, R.K. Malik, P. Chauhan, Functional and safety aspects of enterococci in dairy foods, Indian J. Microbiol. 48 (2008) 317–325, doi:10.1007/s12088-008-0041-2.

[3] V. Biscola, F.L. Tulini, Y. Choiset, H. Rabesona, I. Ivanova, J.M. Chobert, S.D. Todorov, T. Haertlé, B. Franco, Proteolytic activity of *Enterococcus faecalis* VB63F for reduction of allergenicity of bovine milk proteins, J. Dairy Sci. 99 (2016) 5144–5154, doi:10.3168/jds.2016-11036.

[4] F.L. Tulini, V. Bíscola, Y. Choiset, N. Hymery, G. Le Blay, E.C.P. De Martinis, J.M. Chobert, T. Haertlé, Evaluation of the proteolytic activity of *Enterococcus faecalis* FT132 and *Lactobacillus paracasei* FT700, isolated from dairy products in Brazil, using milk proteins as substrates, Eur. Food Res. Technol. 241 (2015) 385–392, doi:10.1007/s00217-015-2470-6.
[5] C.M. Waters, M.H. Antiporta, B.E. Murray, G.M. Dunny, Role of the *Enterococcus faecalis* GelE protease in determination of cellular chain length, supernatant pheromone levels, and degradation of fibrin and misfolded surface proteins, J. Bacteriol. 185 (2003) 3613–3623, doi:10.1128/JB.185.12.3613-3623.2003.
[6] H. Khan, S. Flint, P.L. Yu, Enterocins in food preservation, Int. J. Food Microbiol. 141 (2010) 1–10, doi:10.1016/j.ijfoodmicro.2010.03.005.
[7] D.W. Mount, Using the basic local alignment search tool (BLAST), CSH Protoc. 2007 (2007) pdb.top17, doi:10.1101/pdb.top17.
[8] S.H. Yoon, S.M. Ha, J. Lim, S. Kwon, J. Chun, A large-scale evaluation of algorithms to calculate average nucleotide identity, Antonie Van Leeuwenhoek 110 (2017) 1281–1286, doi:10.1007/s10482-017-0844-4.
[9] A.F. Auch, H.P. Klenk, M. Göker, Standard operating procedure for calculating genome-to-genome distances based on high-scoring segment pairs, Stand. Genom. Sci. 2 (2010) 142–148, doi:10.4056/sigs.541628.
[10] E. Balla, L.M. Dicks, Molecular analysis of the gene cluster involved in the production and secretion of enterocins 1071A and 1071B and of the genes responsible for the replication and transfer of plasmid pEF1071, Int. J. Food Microbiol. 99 (2005) 33–45, doi:10.1016/j.ijfoodmicro.2004.08.008.
[11] R.K. Aziz, D. Bartels, A.A. Best, M. DeJongh, T. Disz, R.A. Edwards, K. Formsma, S. Gerdes, E.M. Glass, M. Kubal, et al., The RAST Server: rapid annotations using subsystems technology, BMC Genom. 9 (2008) 75, doi:10.1186/1471-2164-9-75.
[12] A. Bankevich, S. Nurk, D. Antipov, A.A. Gurevich, M. Dvorkin, A.S. Kulikov, V.M. Lesin, S.I. Nikolenko, S. Pham, A.D. Prjibelski, et al., SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing, J. Comput. Biol. 19 (2012) 455–477, doi:10.1089/cmb.2012.0021.
[13] P. Husemann, J. Stoye, r2cat: synteny plots and comparative assembly, Bioinformatics 26 (2010) 570–571, doi:10.1093/bioinformatics/btp690.
[14] A.C.E. Darling, B. Mau, F.R. Blattner, N.T. Perna, Mauve: multiple alignment of conserved genomic sequence with rearrangements, Genome Res. 14 (2004) 1394–1403, doi:10.1101/gr.2289704.
[15] 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.
[16] M.J. Brittnacher, C. Fong, H.S. Hayden, M.A. Jacobs, M. Radey, L. Rohmer, PGAT: a multistrain analysis resource for microbial genomes, Bioinformatics 27 (2011) 2429–2430, doi:10.1093/bioinformatics/btr418.