Call: HORIZON-CL6-2021-FARM2FORK-01
Topic: HORIZON-CL6-2021-FARM2FORK-01-14
Funding Scheme: HORIZON Research and Innovation Actions

HealthFerm

Deliverable No. 1.3
High-quality genomes from microbes with optimal performance
WP 1

Grant Agreement no.: 101060247

Project Title: Innovative pulse and cereal-based food fermentations for human health and sustainable diets

Contractual Submission Date: 31/08/2025

Actual Submission Date: 29/08/2025

Responsible partner: P6-VUB

Please note that, while deposition of the high-quality genome data is complete, the data is not yet available open access.

Co-funded by the European Union

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Grant agreement no.	101060247
Project full title	Innovative pulse and cereal-based food fermentations for human health and sustainable diets

Deliverable number	**D1.3**
Deliverable title	**High-quality genomes from microbes with optimal performance**
Type[1]	R
Dissemination level[2]	PU
Work package number	1
Work package leader	Nicholas Bokulich
Author(s)	Gabriela Samaniego, Stefan Weckx
Keywords	Genome sequencing, genome annotation, lactic acid bacteria, acetic acid bacteria, yeasts, propionic acid bacteria

HealthFerm has received funding from the European Union under the Horizon Europe grant agreement No. 101060247 and from the Swiss State Secretariat for Education, Research and Innovation (SERI) under contract No. 22.00210.

Checklist	**Y (Yes), N (No)**
1) The deliverable has been sent to all relevant partners for review (i.e. in the same task/WP):	**Y**
2) The deliverable has been endorsed by the WP lead and co-lead:	**Y**
3) Input from other WPs has been sought, if applicable: ***Please specify:***	N

[1] **Type**: Use one of the following codes (in consistence with the Description of the Action):
R: Document, report (excluding the periodic and final reports)
DEM: Demonstrator, pilot, prototype, plan designs
DEC: Websites, patents filing, press & media actions, videos, etc.

[2] **Dissemination level**: Use one of the following codes (in consistence with the Description of the Action)
PU: Public, fully open, e.g. web
SEN: Sensitive, limited under conditions of the Grant Agreement

Table of Contents

Publishable Summary

As part of the HealthFerm project, which aims to explore microbial diversity for improving plant-based food fermentations, a total of 102 high-quality microbial genomes were successfully sequenced, annotated, and submitted to the European Nucleotide Archive (ENA) of the European Bioinformatics Institute (EBI; Hinxton, United Kingdom). These microbial genomes include 81 bacterial and 21 yeast strains isolated from diverse fermented foods and countries in Europe. Of the bacterial genomes, 67 are fully circularised, indicating complete assemblies that include chromosomes and putative plasmids, while 14 remain non-circular but are of high quality. The yeast genomes were assembled at the contig level, in line with the complexity of their eukaryotic genome structures.

All strains sequenced were selected based on their (assumed) functional relevance to key activities such as the production of short-chain fatty acids (SCFA), exopolysaccharides (EPS), and vitamin B12, as well as the degradation of fibres, proteins and antinutritional factors. This genomic resource forms a robust foundation for understanding the genetic potential of microorganisms relevant to food fermentation at the genetic level. Furthermore, these genomes will support comparative genomics, genome mining, and genome-to-phenotype association studies in subsequent project stages, integrating these predictive genomic insights with laboratory validation to develop healthier and more sustainable fermented foods.

1. Introduction

Work Package 1 (WP1) aims to identify, characterise and functionally analyse microbial resources for the design of innovative plant-based fermented foods. Within this context, Task 1.3 focuses on generating high-quality genome sequences of 100 microbial strains. The microbial strains originate from internal culture collections of project partners and the Community Science campaign (Task 1.1). Strains available in culture collections of project partners were selected based on their potential functional relevance to key fermentation activities, such as the production of SCFA, EPS, and vitamin B12, as well as the degradation of fibres, proteins and/or antinutritional factors. The screening for these functionalities was done through direct phenotypic assays or inferred by enzymatic activity assays, with or without confirmatory biochemical testing. Strains coming from the Community Science campaign (Task 1.1) were selected based on their abundance in the citizens' samples.

Task 1.3 aims to provide complete, high-quality, publicly accessible genome sequences of strains relevant for food fermentation processes. The genomic data overcomes the current limitation of existing microbial genome databases, namely that such strains are underrepresented, and serves as a foundational resource for comparative genomics, phenotype-genotype associations, identifying genetic markers linked to key microbial functionalities, mining for metabolic pathways, and predicting and optimising the strains' performance in target substrates. By linking genomic features to phenomics data from Task 1.2, this work establishes a predictive base for rational starter cultures design that enhances nutritional profiles, reduces anti-nutrients, and maximises sustainability in cereal/pulse fermentations. These efforts align with the broader HealthFerm goal of developing tailored microbial inocula for novel food substrates.

2. Description of Activities

The complete microbial genomes were obtained via long-read sequencing [Oxford Nanopore Technologies (ONT), Oxford, United Kingdom] of microbial DNA, followed by the use of bioinformatics tools to achieve high-quality assembled genomes that were submitted to the ENA database. The detailed description of the process is as follows:

2.1 Microbial isolates

Microbial strains (N =102) originating from internal culture collections of project partners and the Community Science campaign were received as washed pellets with a minimum cell density of 4.50 x 10^9 CFU/ml for bacterial isolates, and 1.50 x 10^9 CFU/ml for yeast isolates. Washed pellets were used directly for DNA extraction.

2.2 DNA extraction and quality check

High-molecular-weight DNA was extracted and purified using the Genomic-tip 20/G kit according to the manufacturer's instructions (Qiagen, Hilden, Germany). The recovered genomic DNA was washed with 1 ml of cold 70% ethanol, resuspended in 100 µL of TE buffer (10 mM Tris·HCl, 1 mM EDTA, pH 8.0) and dissolved by incubating in a water bath at 55°C for 2 hours. DNA purity was evaluated using a NanoDrop ND-2000 spectrophotometer (Thermo Fisher Scientific, Waltham, Massachusetts, USA), and DNA concentration was determined using a Qubit fluorometer (Thermo Fisher Scientific).

2.3 Library preparation and data processing

Library preparation was carried out using the ligation sequencing gDNA - Native Barcoding Kit 24 V14 (SQK-NBD114.24) for the R10.4.1 flow cell (ONT). Sequencing was performed using either a MinION or a PromethION device (ONT). The datasets, in POD5 file format, were transferred for processing to a Linux server in VUB's private cloud.

Raw nanopore data was base-called using Dorado v5.3.0 (ONT) in super accurate mode, loading the model "dna_r10.4.1_e8.2_400bps_sup" and specifying the kit number. Classified reads were separated per-barcode using the "demux" Dorado function. Read statistics were acquired via NanoPlot v1.20.0, and quality was checked via NanoQC v0.9.4. Subsequently, reads were filtered and trimmed using NanoFilt v2.8.0 with a minimum quality score of 10, a headcrop of 20, and a tailcrop of 100.

2.4 Assembly, refining and first quality check

Filtered reads were assembled using Flye v2.9.3. (Kolmogorov *et al.*, 2019) in default mode. After assembly, one refining iteration on Medaka v1.11.3 (ONT) was done, using the model r1041_e82_400bps_sup_v4.3.0. The refined assemblies obtained after Medaka refining were evaluated for their quality using Benchmarking Universal Single-Copy Orthologs BUSCO v5.4.3 (Manni *et al.*, 2021) to assess genome completeness and duplication frequencies. Analyses were conducted using genome mode (--mode genome) with the lineage-specific datasets bacteria_odb10 for bacterial assemblies and saccharomycetales_odb10 for yeast assemblies.

2.5 Taxonomy check

For bacterial assemblies, a contigs database was created using Anvi'o (Eren *et al.*, 2015) with a minimum contig length of 1,000 bp. Prodigal (Hyatt *et al.*, 2010) was employed for gene prediction to identify open reading frames (ORFs) and reduce false positives. Annotation was carried out using Hidden Markov Model (HMM) profiles with the Bacteria_71 set of single-copy core genes (SCG) (Lee, 2019). Taxonomy was confirmed by comparing the genome single-copy core genes against the Genome Taxonomy Database (GTDB) based on standardised genome phylogeny (Parks *et al.*, 2022) using DIAMOND (Buchfink *et al.*, 2015). For genomes that could only be classified at the genus level, the Average Nucleotide Identity (ANI) was computed using PyANI v0.2.10 (Pritchard *et al.*, 2016), comparing the genomes with the closest reference found via SCG.

For yeast assemblies, refined genomes were cleaned and sorted using Minimap2 v.2.28. A soft masking step was performed using Tantan v51 (Frith, 2011) to minimise false-positive gene predictions from non-coding regions. Gene prediction was performed using AUGUSTUS v3.5.0 (Nachtweide & Stanke, 2019). The AUGUSTUS gene calls were used to create a contigs database using Anvi'o with the function external-gene-calls, only considering contigs with a minimum length of 1,000 bp. Prodigal was employed for gene prediction, *i.e.*, identifying open reading frames (ORFs), and reducing false

positives. To confirm the taxonomy, the Average Nucleotide Identity (ANI) was computed using PyANI (Pritchard *et al.*, 2016), comparing the sequenced genomes with their corresponding reference genomes.

2.6 Genome annotation

Functional annotation for bacterial genomes was performed using Prokka v1.14.6 (Seemann, 2014), on refined assemblies specifying the option for bacterial genomes with a minimum contig length of 200 bp. For yeast, gene prediction accuracy was maximised via the application of GeneMark-ES v4 (Lomsadze *et al.*, 2005) and GlimmerHMM v3.0.4 (Majoros *et al.*, 2004) on previous AUGUSTUS gene calls. The final consensus gene models were integrated using EvidenceModeler v1.1.1 (Haas *et al.*, 2008) and functionally annotated via InterProScan5 (Jones *et al.*, 2014) and Eggnog-Mapper v2.1.12 (Cantalapiedra *et al.*, 2021).

2.7 Submission to public database

Assembled genomes and their corresponding gene annotations were submitted to ENA. The submission process was conducted using the command-line tool Webin-CLI v8.2.0 that included a validation step before actual submission. A tailored pipeline was created, which involved the creation of the required submission files, such as the EMBL flat file (.embl) using genome data (assembly and annotation), a chromosome list, and a manifest file specifying metadata. In all cases, the correct Taxonomy ID (taxid) was provided to ensure accurate classification within ENA. Submitted, processed, and validated genomes were assigned a unique accession number by ENA for public access and permanent reference.

3. Results

A total of 102 high-quality complete microbial genomes were generated and submitted to the European Nucleotide Archive (ENA) under the **Project ID PRJEB89506,** fulfilling the objective of Deliverable D1.3. The genomes encompass a broad taxonomic and functional diversity, including:

- 81 bacterial genomes, of which 67 are fully circularised, indicating closed, high-quality assemblies that include chromosome and putative plasmids. The remaining 14 bacterial genomes are non-circular but contain a principal linear contig of the expected chromosome size.
- 21 yeast genomes, assembled at the contig-level due to the complexity of their eukaryotic genome structures, with the expected genome size.

Assembled genomes were related to 17 different isolation sources, among which three principal ones: sourdough with 76 strains (from distinct types of flour), borş with 6 strains, and water kefir with 4 strains. 68.6% of the strains were lactic acid bacteria (LAB), 20.6% were yeasts, 6.9% were acetic acid bacteria (AAB), 2.9% were propionic acid bacteria (PAB), and 1% were others, as detailed in Table 1.

Table 1: Detail of the number of isolates per type, genus and species.

Type	Genus	Species	Isolates	
LAB	*Lactiplantibacillus*	2	13	**68.6%**
	Furfurilactobacillus	2	11	
	Levilactobacillus	2	11	
	Fructilactobacillus	1	7	
	Pediococcus	4	6	
	Leuconostoc	3	6	
	Lacticaseibacillus	2	5	
	Companilactobacillus	2	4	
	Limosilactobacillus	2	3	
	Lactococcus	1	1	
	Latilactobacillus	1	1	
	Lentilactobacillus	1	1	
	Weissella	1	1	
Yeast	*Naumovozyma*	1	5	**20.6%**
	Saccharomyces	2	5	
	Maudiozyma	2	4	
	Monosporozyma	1	4	
	Schizosaccharomyces	1	1	
	Torulaspora	1	1	
	Zygosaccharomyces	1	1	
AAB	*Acetobacter*	4	6	**6.9%**
	Gluconobacter	1	1	
PAB	*Propionibacterium*	1	3	**2.9%**
Other	*Hafnia*	1	1	**1.0%**
	Total	**40**	**102**	

For each strain, the following metadata were compiled: ENA public accession number, strain number, species name, isolation source, project partner code, and country of origin, as detailed in Table 2 and Table 3.

Table 2: Detailed metadata of bacterial genomes submitted to the ENA database.

#	Accession number	Strain	Species	Chromosome size (Mbp)	Number of plasmids	Protein encoding genes	Isolation source	Country of origin	Project partner
1	GCA_965286205	VUB-B007 A11	*Acetobacter cerevisiae*	3.2	1	2988	Sourdough	BE	VUB
2	GCA_965286365	VUB-H039 E5	*Acetobacter cerevisiae*	3.3	1	2936	Sourdough	BE	VUB
3	GCA_965638085	VUB-H056 E3	*Acetobacter orientalis*	3.0	1	2691	Sourdough	BE	VUB
4	GCA_965636875	VUB-B009 C2	*Acetobacter oryzifermentans*	2.9	3	2958	Sourdough	BE	VUB
5	GCA_965286175	VUB-H014 D2	*Acetobacter oryzifermentans*	2.8	2	2894	Sourdough	BE	VUB
6	GCA_965286085	VUB-B018 D7	*Acetobacter pasteurianus*	2.8	4	2900	Sourdough	BE	VUB
7	GCA_965286325	VUB-H036 M5	*Companilactobacillus nantensis*	Contig level	-	2669	Sourdough	BE	VUB
8	GCA_965286075	PUR-C17-01	*Companilactobacillus paralimentarius*	2.6	3	2568	Sourdough	IT	PUR
9	GCA_965286185	UH015_11	*Companilactobacillus paralimentarius*	Contig level	-	2645	Sourdough	FI	UH
10	GCA_965286235	VUB-H002 M9	*Companilactobacillus paralimentarius*	Contig level	-	2653	Sourdough	BE	VUB
11	GCA_965286195	PUR-C54-01	*Fructilactobacillus sanfranciscensis*	1.3	1	1385	Sourdough	FR	PUR
12	GCA_965638135	VUB-B008 M14	*Fructilactobacillus sanfranciscensis*	1.3	1	1395	Sourdough	BE	VUB
13	GCA_965286165	VUB-H036 M3	*Fructilactobacillus sanfranciscensis*	1.3	1	1316	Sourdough	BE	VUB
14	GCA_965286215	VUB-H051 M3	*Fructilactobacillus sanfranciscensis*	1.3	1	1419	Sourdough	BE	VUB
15	GCA_965803135	UH002_13	*Fructilactobacillus sanfranciscensis*	1.3	1	1359	Sourdough	FI	UH
16	GCA_965815465	UH009_14	*Fructilactobacillus sanfranciscensis*	1.3	-	1304	Sourdough	FI	UH
17	GCA_965286155	UH015_15	*Fructilactobacillus sanfranciscensis*	1.3	1	1352	Sourdough	FI	UH
18	GCA_965614345	ETH-STR969	*Furfurilactobacillus milii*	2.6	3	2601	Sourdough	CH	ETHZ
19	GCA_965614395	ETH-STR978	*Furfurilactobacillus milii*	2.6	1	2456	Sourdough	CH	ETHZ
20	GCA_965614355	ETH-STR1086	*Furfurilactobacillus milii*	2.7	1	2586	Sourdough	CH	ETHZ
21	GCA_965614405	ETH-STR1087	*Furfurilactobacillus milii*	2.7	0	2540	Sourdough	CH	ETHZ
22	GCA_965614415	ETH-STR965	*Furfurilactobacillus milii*	2.7	0	2574	Sourdough	USA	ETHZ
23	GCA_965286245	UH008_13	*Furfurilactobacillus rossiae*	2.6	3	2507	Sourdough	FI	UH
24	GCA_965806495	UH004_10	*Furfurilactobacillus rossiae*	2.6	4	2636	Sourdough	FI	UH
25	GCA_965286125	UH033_2	*Furfurilactobacillus rossiae*	Contig level	-	2732	Sourdough	FI	UH
26	GCA_965812195	UH035_10	*Furfurilactobacillus rossiae*	2.5	2	2479	Sourdough	FI	UH

Co-funded by the European Union

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

27	GCA_965638095	VUB-B018 M6	*Furfurilactobacillus rossiae*	2.5	3	2583	Sourdough	BE	VUB
28	GCA_965815565	VUB-81D1	*Furfurilactobacillus rossiae*	2.5	1	2387	Sourdough	BE	VUB
29	GCA_965286395	VUB-H016 D1	*Gluconobacter potus*	3.1	4	3178	Sourdough	BE	VUB
30	GCA_965614565	VUB-B012 M6	*Hafnia alvei*	4.6	1	4223	Sourdough	BE	VUB
31	GCA_965638145	KFAL1	*Lacticaseibacillus paracasei*	3.0	4	3042	Kefir water	IT	UBZ
32	GCA_965286145	HF-IBB-214	*Lacticaseibacillus paracasei*	2.8	3	2995	Borş	RO	IBB
33	GCA_965638105	E-97949	*Lacticaseibacillus paracasei*	Contig level	-	3184	Human biopsy	FI	VTT
34	GCA_965286355	HF-IBB-287	*Lacticaseibacillus paracasei*	2.9	3	3001	Borş	RO	IBB
35	GCA_965614495	E-052741	*Lacticaseibacillus rhamnosus*	3.0	0	2820	Infant feces	FI	VTT
36	GCA_965286285	HF-IBB-019	*Lactiplantibacillus argentoratensis*	3.1	3	2924	Borş	RO	IBB
37	GCA_965286295	UH016_15	*Lactiplantibacillus argentoratensis*	3.0	1	2784	Sourdough	FI	UH
38	GCA_965286025	PUR-C47-01	*Lactiplantibacillus plantarum*	3.1	2	3087	Sourdough	FR	PUR
39	GCA_965286625	IBB-BR9	*Lactiplantibacillus plantarum*	3.1	9	3452	Braga	RO	IBB
40	GCA_965286275	E-78076	*Lactiplantibacillus plantarum*	3.1	2	2917	Beer	UK	VTT
41	GCA_965286415	2MM1	*Lactiplantibacillus plantarum*	3.0	4	3003	Hemp	IT	UH
42	GCA_965286015	POM1	*Lactiplantibacillus plantarum*	Contig level	-	3094	Tomato	IT	UH
43	GCA_965286555	10MM0	*Lactiplantibacillus plantarum*	Contig level	-	3007	Hemp	IT	UH
44	GCA_965286335	VUB-B043 E5M	*Lactiplantibacillus plantarum*	Contig level	-	3063	Sourdough	BE	VUB
45	GCA_965286095	VUB-H044 M1	*Lactiplantibacillus plantarum*	Contig level	-	3153	Sourdough	BE	VUB
46	GCA_965638115	VUB-H046 M4	*Lactiplantibacillus plantarum*	Contig level	-	3196	Sourdough	BE	VUB
47	GCA_965286635	UH033_10	*Lactiplantibacillus plantarum*	3.1	5	3245	Sourdough	FI	UH
48	GCA_965286495	HF-IBB-082	*Lactiplantibacillus plantarum*	3.0	0	3112	Borş	RO	IBB
49	GCA_965286575	E-991172	*Lactococcus cremoris*	2.5	3	2801	Beer	FI	VTT
50	GCA_965638125	E-153429	*Latilactobacillus sakei*	1.9	3	2033	Chicken	FI	VTT
51	GCA_965284755	VUB-H032 M23	*Lentilactobacillus diolivorans*	3.1	1	2922	Sourdough	BE	VUB
52	GCA_965286305	IBB-P109	*Leuconostoc citreum*	1.7	3	1741	Bell pepper	RO	IBB
53	GCA_965614535	S7d10	*Leuconostoc citreum*	Contig level	-	1910	Sauerkraut	IT	UBZ
54	GCA_965814075	VUB-19M7	*Leuconostoc citreum*	1.7	-	1779	Sourdough	BE	VUB
55	GCA_965286525	GSL1	*Leuconostoc mesenteroides*	1.9	1	2034	Apple	IT	UBZ
56	GCA_965286515	IBB-P124	*Leuconostoc mesenteroides*	2.0	2	2115	Beans	RO	IBB

57	GCA_965286035	HF-IBB-469	*Leuconostoc pseudomesenteroides*	2.2	1	2230	Kefir water	RO	IBB
58	GCA_965286385	PUR-C03-01	*Levilactobacillus brevis*	2.4	2	2483	Sourdough	IT	PUR
59	GCA_965286315	PUR-C42-01	*Levilactobacillus brevis*	2.5	4	2641	Sourdough	HU	PUR
60	GCA_965286535	VUB-B006 M9	*Levilactobacillus brevis*	Contig level	-	2405	Sourdough	BE	VUB
61	GCA_965286105	VUB-H025 M21	*Levilactobacillus brevis*	2.3	3	2366	Sourdough	BE	VUB
62	GCA_965286605	VUB-H042 M19	*Levilactobacillus brevis*	2.3	2	2416	Sourdough	BE	VUB
63	GCA_965286465	UH013_6	*Levilactobacillus brevis*	2.3	2	2324	Sourdough	FI	UH
64	GCA_965286455	UH033_1	*Levilactobacillus brevis*	2.3	4	2644	Sourdough	FI	UH
65	GCA_965286615	VUB-H037 M3	*Levilactobacillus brevis*	2.3	3	2598	Sourdough	BE	VUB
66	GCA_965814455	VUB-6DM1	*Levilactobacillus brevis*	2.3	5	2707	Sourdough	BE	VUB
67	GCA_965814695	PUR-C17-03	*Levilactobacillus namurensis*	2.5	3	2386	Sourdough	IT	PUR
68	GCA_965286135	VUB-H028 M10	*Levilactobacillus cerevisiae*	3.2	4	3076	Sourdough	BE	VUB
69	GCA_965286115	HF-IBB-062	*Limosilactobacillus fermentum*	2.0	0	1988	Borş	RO	IBB
70	GCA_965286585	HF-IBB-205	*Limosilactobacillus fermentum*	2.0	1	2053	Borş	RO	IBB
71	GCA_965286255	HF-IBB-392	*Limosilactobacillus panis*	2.1	1	2036	Borş	RO	IBB
72	GCA_965286485	15M9	*Pediococcus acidilactici*	1.9	0	1902	Hemp	IT	UH
73	GCA_965638075	9MM1	*Pediococcus acidilactici*	1.9	1	2010	Hemp	IT	UH
74	GCA_965286065	VUB-H028 M3	*Pediococcus inopinatus*	Contig level	-	1983	Sourdough	BE	VUB
75	GCA_965286505	VUB-H015 M7	*Pediococcus parvulus*	2.0	3	2008	Sourdough	BE	VUB
76	GCA_965286055	UH017_6	*Pediococcus parvulus*	2.0	3	2143	Sourdough	FI	UH
77	GCA_965286425	HF-IBB-737	*Pediococcus pentosaceus*	1.8	2	1947	Sourdough	RO	IBB
78	GCA_965286565	AS5	*Propionibacterium freudenreichii*	2.6	0	2344	Cheese	FI	UH
79	GCA_965286045	AS9	*Propionibacterium freudenreichii*	2.6	0	2343	Cheese	FI	UH
80	GCA_965286265	J117	*Propionibacterium freudenreichii*	2.6	1	2275	Cheese	FI	UH
81	GCA_965286445	BRL2	*Weissella confusa*	Contig level	-	2358	Beetroot	IT	UBZ

Table 3: Detailed metadata of yeast genomes submitted to the ENA database.

#	Accession number	Strain	Species	Genome size (Mbp)	Number of contigs	Protein encoding genes	Isolation source	Country of origin	Project partner
82	GCA_965637955	JB6Y2	*Zygosaccharomyces bisporus*	11.1	38	5032	Kombucha	IT	UBZ
83	GCA_965637945	KFAY2	*Saccharomyces cerevisiae*	12.1	46	5466	Kefir water	IT	UBZ
84	GCA_965637975	VUB-H027 Y12	*Saccharomyces cerevisiae*	12.0	65	5457	Sourdough	BE	VUB
85	GCA_965638015	VUB-H042 Y18	*Saccharomyces cerevisiae*	12.0	43	5461	Sourdough	BE	VUB
86	GCA_965830275	UH002_3	*Saccharomyces cerevisiae*	12.0	112	5498	Sourdough	FI	UH
87	GCA_965609025	PUR-C42-02	*Saccharomyces uvarum*	12.0	36	5410	Sourdough	BE	PUR
88	GCA_965637995	JB15IG1	*Schizosaccharomyces pombe*	13.0	36	4712	Kombucha	IT	UBZ
89	GCA_965608555	PUR-C40-03	*Torulaspora delbrueckii*	9.3	18	4854	Sourdough	BE	PUR
90	GCA_965637235	VUB-B008 Y5	*Maudiozyma bulderi*	18.3	691	6583	Sourdough	BE	VUB
91	GCA_965637255	VUB-H025 Y8	*Maudiozyma bulderi*	17.3	681	6168	Sourdough	BE	VUB
92	GCA_965637265	VUB-H040 Y10	*Maudiozyma bulderi*	17.2	663	6102	Sourdough	BE	VUB
93	GCA_965637245	VUB-H001 C1	*Maudiozyma exigua*	25.6	46	9758	Sourdough	BE	VUB
94	GCA_965637915	KFBY1	*Monosporozyma unispora*	12.8	39	5229	Kefir water	IT	UBZ
95	GCA_965637925	VUB-H002 C7	*Monosporozyma unispora*	12.5	17	5189	Sourdough	BE	VUB
96	GCA_965637905	VUB-H033 C9	*Monosporozyma unispora*	12.5	26	5169	Sourdough	BE	VUB
97	GCA_965637875	VUB-H037 C10	*Monosporozyma unispora*	12.5	22	5209	Sourdough	BE	VUB
98	GCA_965637895	ETH-STR981	*Naumovozyma castellii*	11.5	18	5443	Sourdough	CH	ETHZ
99	GCA_965638005	ETH-STR985	*Naumovozyma castellii*	11.5	21	5433	Sourdough	CH	ETHZ
100	GCA_965637935	VUB-B026 Y4	*Naumovozyma castellii*	11.5	18	5418	Sourdough	BE	VUB
101	GCA_965637885	VUB-H022 Y10	*Naumovozyma castellii*	11.5	15	5448	Sourdough	BE	VUB
102	GCA_965637965	VUB-H027 Y8	*Naumovozyma castellii*	11.5	36	5495	Sourdough	BE	VUB

4. Conclusion

The generation of complete, well-annotated microbial genomes via long-read sequencing allowed to identify genes related to the functional targets defined in HealthFerm's WP1 - Task 1.2 and in WP2. The genomics data, together with the phenomics data obtained in Task 1.2, will provide a comprehensive data set for each of the strains assessed, given their potential use as starter cultures, providing technological and health benefits to foodstuffs via food fermentation.

The genomics data will lay the basis of follow-up research, including the detection of DNA polymorphisms, such as SNPs and structural variations, advanced genotype-phenotype association studies, and genome-scale metabolic modelling (GSMM) by linking genome content and polymorphism data to strain-level phenotypic profiles to validate the predictions at the genotypic and phenotypic level.

5. References

Buchfink, B., Xie, C., & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, *12*(1), 59–60. https://doi.org/10.1038/nmeth.3176

Cantalapiedra, C. P., 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. *Molecular Biology and Evolution*, *38*(12), 5825–5829. https://doi.org/10.1093/molbev/msab293

Eren, A. M., Esen, Ö. C., Quince, C., Vineis, J. H., Morrison, H. G., Sogin, M. L., & Delmont, T. O. (2015). Anvi'o: an advanced analysis and visualisation platform for 'omics data. *PeerJ*, *3*, e1319. https://doi.org/10.7717/peerj.1319

Frith, M. C. (2011). A new repeat-masking method enables specific detection of homologous sequences. *Nucleic Acids Research*, *39*(4), e23–e23. https://doi.org/10.1093/nar/gkq1212

Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., White, O., Buell, C. R., & Wortman, J. R. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biology*, *9*(1), R7. https://doi.org/10.1186/gb-2008-9-1-r7

Hyatt, D., Chen, G. L., LoCascio, P. F., Land, M. L., Larimer, F. W., & Hauser, L. J. (2010). Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, *11*. https://doi.org/10.1186/1471-2105-11-119

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., Pesseat, S., Quinn, A. F., Sangrador-Vegas, A., Scheremetjew, M., Yong, S.-Y., Lopez, R., & Hunter, S. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics*, *30*(9), 1236–1240. https://doi.org/10.1093/bioinformatics/btu031

Kolmogorov, M., Yuan, J., Lin, Y., & Pevzner, P. A. (2019). Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology*, *37*(5), 540–546. https://doi.org/10.1038/s41587-019-0072-8

Lee, M. D. (2019). GToTree: a user-friendly workflow for phylogenomics. *Bioinformatics*, *35*(20), 4162–4164. https://doi.org/10.1093/bioinformatics/btz188

Lomsadze, A., Ter-Hovhannisyan, V., Chernoff, Y. O., & Borodovsky, M. (2005). Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Research*, *33*(20), 6494–6506. https://doi.org/10.1093/nar/gki937

Majoros, W. H., Pertea, M., & Salzberg, S. L. (2004). TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics*, *20*(16), 2878–2879. https://doi.org/10.1093/bioinformatics/bth315

Manni, M., Berkeley, M. R., Seppey, M., & Zdobnov, E. M. (2021). BUSCO: Assessing Genomic Data Quality and Beyond. *Current Protocols*, *1*(12), e323. https://doi.org/https://doi.org/10.1002/cpz1.323

Nachtweide, S., & Stanke, M. (2019). Multi-Genome Annotation with AUGUSTUS. In M. Kollmar (Ed.), *Gene Prediction: Methods and Protocols* (pp. 139–160). Springer New York. https://doi.org/10.1007/978-1-4939-9173-0_8

Parks, D. H., Chuvochina, M., Rinke, C., Mussig, A. J., Chaumeil, P.-A., & Hugenholtz, P. (2022). GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalised and complete genome-based taxonomy. *Nucleic Acids Research*, *50*(D1), D785–D794. https://doi.org/10.1093/nar/gkab776

Pritchard, L., Glover, R. H., Humphris, S., Elphinstone, J. G., & Toth, I. K. (2016). Genomics and taxonomy in diagnostics for food security: soft-rotting enterobacterial plant pathogens. *Analytical Methods*, *8*(1), 12–24. https://doi.org/10.1039/C5AY02550H

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, *30*(14), 2068–2069. https://doi.org/10.1093/bioinformatics/btu153