CORA Cork Open Research Archive
Cartlann Taighde Oscailte Chorcaí

Title	Shotgun metagenomics-based exploration of kefir microbiomes
Authors	Walsh, Liam
Publication date	2023
Original Citation	Walsh, L. 2023. Shotgun metagenomics-based exploration of kefir microbiomes. PhD Thesis, University College Cork.
Type of publication	Doctoral thesis
Rights	© 2023, Liam Walsh. - https://creativecommons.org/publicdomain/zero/1.0/
Download date	2026-11-16 11:36:35
Item downloaded from	https://hdl.handle.net/10468/15912



UCC
University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

Shotgun metagenomics-based exploration of kefir microbiomes

A thesis presented to the National University of Ireland for the degree of Doctor of Philosophy

By

Liam Walsh BSc

Teagasc Food Research Centre, Moorepark, Fermoy, Co. Cork, Ireland

APC Microbiome Ireland, University College Cork, Cork, Ireland

Department of Microbiology, University College Cork, Cork, Ireland

October 2023

Research supervisors: Professor Paul Cotter, Professor Paul O'Toole

Table of Contents

Declaration vii
Thesis Abstract viii
Publications x
List of Abbreviations xi
List of Figures xiv
List of Tables xxvi
Chapter 1 1
Bioinformatic approaches for studying the microbiome of fermented food 1
Section 1.0: Abstract 2
Section 1.1: Introduction- Microbiome research and its relevance to food 3
Section 1.2: Shotgun metagenomics 4
1.2.1 Overview of bioinformatic approaches for metagenomic data 5
1.2.2 Short-read approaches 16
1.2.3 Metagenome assembly-based approaches 17
1.2.4 Advantages and disadvantages of the short read and assembly-based approaches 21
1.2.5 Specialised bioinformatics tools 24
1.2.6 Limitations of current comparative performance methods for bioinformatic tools 29
Section 1.3: Applications of bioinformatic approaches to fermented food microbiomes. 30
1.3.1 Research into the fermentation process 30
1.3.2 Community structure and dynamics 31
1.3.3 Flavour 35
1.3.4 Food safety and quality 40
1.3.5 Health benefits 43
Section 1.4: Conclusion 46
Section 1.5: References 49
Section 1.6: Supplementary information 58
Supplementary Figures 58
Chapter 2 59
Analysis of the milk kefir pan-metagenome reveals four community types, core species and associated metabolic pathways 59
Section 2.0: Abstract 60
Section 2.1: Introduction 61
Section 2.2: Methods 62
2.2.1 Sample collection, preparation and pre-processing 62
2.2.2 Bioinformatic analysis 63

2.2.3 Statistical analysis 64
2.2.4 Data visualisation 65
Section 2.3: Results 65
2.3.1 Taxonomic profiling reveals distinct community types in the kefir microbiome 65
2.3.2 Phylogenetic analysis of assembled metagenomes identifies three putatively novel species in the kefir microbiome. 69
2.3.3 Strain level phylogenetic analysis of the kefir microbiome uncovers strain subclusters and long term strain-type retention 70
2.3.4 The metabolic potential of kefir metagenomes separates by community type 76
2.3.5 Functional comparisons of the representatives of each community- type reveals distinguishing metabolic and functional features 78
2.3.6 Phylogenetic placement of representatives of each community- type provides taxonomic classification and supports the elevation of representatives of *Lla. lactis* subcluster - 2 to a distinct species 82
2.3.7 Virus identification highlights the prevalence of phages in kefir fermentations 83
Section 2.4: Discussion 84
Section 2.5: References 88
Section 2.6: Supplementary information 91
2.6.1 Supplementary method 91
2.6.2 Supplementary Results 92
2.6.3 Supplementary Discussion 94
2.6.4 Supplementary Figures 94
Chapter 3 123
Comparison of the relative impacts of acute consumption of an inulin-enriched diet, milk kefir or a commercial probiotic product on the human gut microbiome and metabolome 123
Section 3.0: Abstract 124
Section 3.1: Introduction 125
Section 3.2: Methods 126
3.2.1 Participant enrolment, intervention design, urine and stool collection 126
3.2.2 Clinical measurements 128
3.2.3 Microbial DNA extraction from faeces 128
3.2.4 High-throughput DNA sequencing 128
3.2.5 Bioinformatics 128
3.2.6 ^{1}H NMR Metabolic profiling analysis 129
3.2.8 Short chain fatty acids (SCFAs) analysis 130
3.2.9 Statistics 130

Section 3.3: Results 131
3.3.1 Clinical outcomes 131
3.3.2 Compositional analysis reveals an increased abundance of *Lactococcus raffinolactis* after kefir consumption 131
3.3.3 Comparative genomics reveals a number of host adaptability features within the pan-genome of *Lla. raffinolactis* 132
3.3.4 Kefir consumption results in modest changes to urinal metabolites but not functional potential 134
Section 3.4: Discussion 136
Section 3.5: References 138
Section 3.6: Supplementary information 142
3.6.1 Supplementary figures 142
3.6.2 Supplementary table 143
Chapter 4 198
Kefir4All, a Citizen Science initiative to raise awareness of the roles microbes play in food and beverage fermentation 198
Section 4.0: Abstract 199
Section 4.1: Introduction 200
Section 4.2: Methods 201
4.2.1 Overview of the study 201
4.2.2 Participant recruitment 202
4.2.3 Fermentation equipment 202
4.2.4 Surveys and evaluation measures 203
4.2.5 Outreach activities 203
4.2.6 Bioinformatic analysis 204
4.2.7 Statistics and data visualisation 205
Section 4.3: Results 206
4.3.1 The 'Workshop Survey' highlights the importance of promoting microbiology to the general population. 206
4.3.2 Tracking participation within the Kefir4All cohort provides insights into the advantages and disadvantages of citizen scientist recruitment based on a 'Good faith approach' 212
4.3.3 Continued participation played an important role in the success of Kefir4All project 213
4.3.4 Survey data reveals a relationship between fermentation parameters and the occurrences of fermentation characteristics. 216
4.3.5 The Kefir4All project provided important new research data, and resulted in an enhanced interest and scientific knowledge with respect to fermented foods and microbiology among participants. 218

Section 4.4: Discussion 224
Section 4.5: References 229
Section 4.6: Supplementary information 231
4.6.1 Supplementary method 231
4.6.2 Supplementary result 251
4.6.2 Supplementary Figures 280
Chapter 5 286
Adaptations and community changes in milk and water kefir microbiomes in response to environmental parameters (Kefir4All-Citizen Science Project) 286
Section 5.0: Abstract 287
Section 5.1 Introduction 288
Section 5.2. Materials and methods 289
5.2.1 Participant recruitment and equipment: 289
5.2.2 Study design: 289
5.2.3 Collection of metadata and samples: 290
5.2.4 DNA extraction, preparation of libraries, shotgun metagenomic sequencing and data pre-processing 290
5.2.5 Food metagenomics datasets 291
5.2.6 Nanopore library preparation and sequencing 291
5.2.7 Bioinformatic analysis 292
5.2.8 Statistics 294
5.2.9 Data visualisation 294
Section 5.3 Results 295
5.3.1 Compositional analysis reveals a diverse range of prevalent microorganisms, including several not detected in initial kefir fermentations. 295
5.3.2 Compositional analysis of kefir metagenomes suggests the majority of compositional change within the milk and water kefir grain microbiota occurs in the initial periods of the project. 298
5.3.3 Functional analysis of kefir metagenomes suggests the majority of functional change within the milk and water kefir grain microbiota occurs in the initial periods of the project. 300
5.3.4 Fermentation of kefir grains with a common single community type can lead to the development of multiple distinct community types 303
5.3.5 Fermentation conditions may influence the compositional profile of both milk and water kefir through the introduction of contaminating microbial species 305
5.3.6 Fermentation parameters, and abiotic factors are modest predictors of milk and water kefir compositional profiles 307

5.3.7 Phylogenetic analysis of kefir metagenomes suggests the majority of evolutionary change for most representative species occurs in the initial periods of the project 311
Section 5.4 Discussion 315
Section 5.5 References 320
Section 5.6. Supplementary information 323
5.6.1 Supplementary Figures 323
5.6.2 Supplementary References 326
General Discussion 327
References 330

Declaration

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of PhD, is entirely my own work (except as declared, hereafter), and has not been submitted for another degree, either at University College Cork or elsewhere.

Signed: Liam Walsh

Student Number: 115707141

Date: 14/08/2023

Thesis Abstract

Fermentation is among the world's oldest forms of food processing, having been applied for millennia to preserve or enhance foods and still accounts for a significant component in the human diet. One such fermented food is milk kefir, which is regularly consumed in eastern European countries and is becoming increasingly popular in western society as a functional food, with numerous scientific studies and reviews associating health benefits with its consumption. Water kefir represents another fermented food of considerable interest, which is increasingly being perceived as a non-dairy alternative to milk kefir. In this thesis, we demonstrate that metagenomic analysis is a valuable tool to (i) expand our, and the general public's understanding on the microbiology of kefir fermentations, (ii) determine the potential functionality of specific microbes therein and (iii) investigate the impact of milk kefir on the host gut microbiome. In chapter 1, we critically analyse the tools and pipelines that have been used, or that could be applied, to the analysis of metagenomic and metatranscriptomic data relating to fermented foods. In addition, we critically analyse a number of studies of fermented foods in which these tools have previously been applied to highlight the insights that these approaches can provide. Chapter 2 is focused on the shotgun metagenomics-based analysis of 256 kefir milk samples produced from milk kefir grains sourced from various parts of the world. This chapter provides considerable insight into the heterogeneity of these populations, while also uncovering conserved features such as the presence of *Lactococcus lactis* and *Lactococcus cremoris*, which may help to define the minimal components required for a fermented milk product to be considered a milk kefir. In chapter 3, we show that shotgun metagenomics, when used alongside metabolomics, can provide evidence that milk kefir modulates the gut microbiome. We show that daily consumption of kefir in a healthy cohort has a subtle impact on the urinal metabolome and gut microbiome. The principal change to the gut microbiome was the detection of *Lactococcus raffinolactis* post kefir consumption. The detection of *Lactococcus raffinolactis* is particularly notable given its low relative abundance across kefir metagenomes generated in chapter 2. In chapter 4, we describe Kefir4All, a citizen science project designed to provide the general public with an opportunity to expand their

awareness, knowledge and practical skills relating to microbiology, introduced from the perspective of producing a fermented food, i.e., milk kefir or water kefir. In chapter 5, we highlight how research relating to the milk kefir and water kefir microbiome was greatly extended through the efforts of the Kefir4All citizen scientists through the investigation of compositional, functional and evolutionary change in milk and water kefir microbial communities over 21 weeks of repeat regular fermentation by citizen scientists. Overall, this thesis highlights that bioinformatic analysis of high throughput sequencing datasets can expand our knowledge of microbial communities associated with fermented foods and in the host following consumption, while also highlighting the merits of employing fermented food-related studies to raise awareness, knowledge and interest in microbiology and fermentation.

Publications

Liam H. Walsh, Mairéad Coakley, Aaron M. Walsh, Paul W. O'Toole & Paul D. Cotter (2022). Bioinformatic approaches for studying the microbiome of fermented food, Critical Reviews in Microbiology, DOI: 10.1080/1040841X.2022.2132850

Liam H. Walsh, Mairéad Coakley, Aaron M. Walsh, Paul W. O'Toole & Paul D. Cotter. (2023). Analysis of the Milk Kefir Pan-Metagenome Reveals Four Community Types, Core Species and Associated Metabolic Pathways, iScience, DOI: 10.1016/j.isci.2023.108004

Cotter, Paul & Walsh, Liam & Walsh, Aaron & Perez, Isabel & Crispie, Fiona & Costabile, Adele & Ellis, Richard & Finlayson, Jim & Finneganan, Laura & Claesson, Marcus & Holmes, Elaine. (2023). Comparison of the relative impacts of acute consumption of an inulin-enriched diet, milk kefir or a commercial probiotic product on the human gut microbiome and metabolome, *npj Science of Food* DOI: 10.21203/rs.3.rs-2542702/v1.

Liam H. Walsh, Samuel Breselge, José Guilherme Prado Martin, Mairéad Coakley, Eimear Ferguson, Aimee Stapleton, Fiona Crispie, Paul W. O'Toole and Paul D. Cotter. Kefir4All, a Citizen Science initiative to raise awareness of the roles that microbes play in food fermentation, Journal of Microbiology & Biology Education, DOI: 10.1128/jmbe.00155-23

List of Abbreviations

°C	Degrees Celcius
AMR	Anti-Microbial Resistance
ANI	Average Nucleotide Identity
ANOVA	Analysis of variance
ANOSIM	analysis of similarities
ARG	Antibiotic Resistance Genes
ARG-ANNOT	Antibiotic Resistance Gene-ANNOTation
AST	Antimicrobial susceptibility testing
BC	Bray- Curtis index
CAMI	Critical assessment of metagenome interpretation
CARD	Comprehensive Antibiotic Resistance Database
CDs	Coding sequences
CLARK	CLAssifier based on Reduced K-mers
CONCOCT	Clustering cONtigs on COverage and ComposiTion
Contigs	Single contiguous sequences
CPM	Counts Per Million
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
DIAMOND	Double index alignment of next-generation sequencing data
DNA	Deoxyribonucleic Acid
D-PLA	D-phenyllactic acid
DPPH	1,1-diphenyl-2-picrylhydrazyl
EC	Enzyme Commission
ENA	European Nucleotide Archive
EPS	Exopolysaccharide
FDA	Food and Drug Administration
FDR	False discovery rate
FSIS	The Food Safety and Inspection Service
g	Gram
GABA	Gamma-aminobutyric Acid
GB	Gigabyte
GC content	Guanine-cytosine
GC-MS	Gas Chromatography-Mass Spectrometry
GEMs	Genome-scale metabolic models
GOLD	Genomes Online Database
GUI	graphical user interface
HGT	Horizontal gene transfer
HMM	Hidden Markov Models
HPLC-DAD	High-Performance Liquid ChromatographyDiode Array Detection
HS	Headspace
HTS	High Throughput Sequencing
HUMAnN	HMP (Human Microbiome Project) Unified Metabolic anlaysis Network
IBD	Inflammatory Bowel disease
IBS	Irritable Bowel Syndrome

IDBA	Iterative de Bruijn graph assembler
IDBA-UD	Iterative De Bruijn graph de novo Assembler for short reads
ISAPP	International Scientific Association for Probiotics and Prebiotics
ITS	Internal Transcribed Spacers
kb	Kilobase
*k*La	Oxygen transfer coefficient
KO	KEGG Orthology
J	Jaccard distance
LAB	Lactic Acid Bacteria
LDA	Linear Discrimination Analyses
LEfSe	LDA effect size
MAG	Metagenomic Assembled Genome
MAGO	Metagenome-assembled genomes orchestra
MB	Megabyte
MDS	Multidimensional scaling
MetaBAT	Metagenome Binning with Abundance and Tetra-nucleotide frequencies
MetaPhlAn	Metagenomic phylogenetic analysis
MG-RAST	The Metagenomics RAST (Rapid Annotation using SubsystemTechnology) server
MRD	Maximum Recovery Diluent
NGS	Next Generation Sequencing
NMDS	Non-metric Multidimensional Scaling
OTU	Operational taxonomic unit
PanPhlAn	Pangenome-based phylogenomic analysis
PBS	Phosphate buffered saline
PCoA	Principal Coordinates Analysis
PCR	Polymerase chain reaction
PD	Phylogenetic diversity
PERMANOVA	Permutational analysis of variance
PHAGCs	Potentially Health Associated Gene Clusters
PLS-DA	Partial Least Squares Discriminant Analyses
Prodigal	Prokaryotic Dynamic Programming Genefinding Algorithm
RCR	Relative citation ratio
rpm	Revolutions Per Minute
rRNA	Ribosomal Ribonucleic Acid
SCOBY	Symbiotic Colony Of Bacteria and Yeast
SNVs	Single-nucleotide variants
SPME	Solid phase microextraction
SSMS	Shallow shotgun metagenome sequencing
SUPER-FOCUS	SUbsystems Profile by databasE Reduction using FOCUS (Find Organisms by Composition USage)
TNF	tetra-nucleotide frequency
Viromes	Viral metagenomes
VOCs	Volatile organic compounds
w/v	weight/volume

WMGS	Whole metagenome shotgun sequencing
FMP	Fermented milk product

Acetobacter	*A.*
Acinetobacter	*Ac.*
Bacillus	*B.*
Bifidobacterium	*Bi.*
Brettanomyces	*Br.*
Brochothrix thermosphacta	*Bro.*
Clostridium	*C.*
Escherichia	*E.*
Enterococcus	*En.*
Gluconobacter	*Glu.*
Hafnia	*H.*
Klebsiella	*K.*
Komagataeibacter	*Kom.*
Lactobacillus	*L.*
Lacticaseibacillus	*Lc.*
Lentilactobacillus	*Len.*
Leuconostoc	*Leu.*
Levilactobacillus	*Lev.*
Limosilactobacillus	*Lim.*
Lactococcus	*Lla.*
Lactiplantibacillus	*Lpb.*
Methylobacterium	*M.*
Oenococcus	*O.*
Pseudomonas	*P.*

Pantoea	*Pa.*
Propionibacterium	*Pr.*
Rahnella	*R.*
Streptococcus	*S.*
Saccharomyces	*Sac.*
Zymomonas	*Z.*
Zygotorilaspora	*Zg.*

List of Figures

Fig n	description	Page no
Figure 1	**Figure 1:** Relative citation ratio scores for selected peer-reviewed research papers, introducing established metagenomic tools (2 years and older) with the highest influence to the research community. Peer-reviewed research papers were selected through a literature mining methodology and assessed using a citation analysis as described above. Peer-reviewed research papers, discussed Bioinformatic tools with applications in ■ Short read taxonomic classification, ■ Short read functional classification, ■ Metagenomic assembly, ■ Metagenomic binning and ■ Analysis of recovered genomes.	7
Figure 2	**Figure 2:** Annual growth and availability in reference data available for the interpretation of metagenomic data including reference genomes in the NCBI database (https://www.ncbi.nlm.nih.gov/). (A) Annual growth (from 2005 to 2021) in reviewed research papers introducing novel and improved tools designed for the interpretation of metagenomic data, metagenomes available in the NCBI GenBank database and reference genomes in the NCBI GenBank and NCBI RefSeq (Tatusova et al., 2014; Clark et al., 2016); (B) Available Metagenome, NCBI Genbank and NCBI Refseq reference data separated by assembly level. (C) Available Metagenome, NCBI Genbank and NCBI Refseq reference data separated by domain. (D) Cladogram of bacterial and fungal species found in milk kefir, species information obtained from (Bourrie, Willing and Cotter, 2016). Tip labels represent the species detected, e.g., *Lentilactobacillus parabuchneri*, tip colour represents the phylum, and shape represents the kingdom of the species. Presence/absence heatmap displays the availability of reference genomes for each of the listed species in both NCBI RefSeq and NCBI GenBank databases. Metagenome, Refseq and Genbank derived data was generated and analysed in R using the Biomartr package (Drost and Paszkowski, 2017). Visualisation was performed using the ggplot2 package (Wickham, 2016) and ggtree package (Yu, 2020).	21
Figure 3	**Figure 3**: Conceptual overview of the primary steps, outputs and applications associated with the application of short read and assembly based approaches to fermented food metagenomic data.	31
Figure S1	**Figure S1:** Cladogram of bacterial and fungal species found in water kefir, species information obtained from (Bourrie, Willing and Cotter, 2016). Tip labels	58

	represent the species detected, e.g., *Lacticaseibacillus paracasei*, tip colour represents the phylum, and shape represents the kingdom of the species. Presence/absence heatmap displays the availability of reference genomes for each of the listed species in both NCBI RefSeq and NCBI GenBank databases.	
Chapter 2		
Figure 1	**Figure 1:** Species and their relative abundance across milk kefir metagenomes. Cladogram presenting a hierarchical overview of species detected in the kefir microbiome using the custom kraken2 pipeline. Tips represent the species detected, e.g., *Lla. lactis*, and the associated colour of the tips describes the order of the species ■ *Bacillales*, ■ *Enterobacterales*, ■ *Lactobacillales*, ■ *Propionibacteriales*, ■ *Pseudomonadales*, ■ *Rhodospirillales*, ■ *Saccharomycetales,* ■ *Thermales* and ■ Unclassified bacteriophage. The outer rings represent a circular heatmap displaying the relative abundance of each species across 256 metagenomes.	66
Figure 2	**Figure 2:** Variance in the compositional data among kefir metagenomes explained by the dominating microbial species. **(A.)** Principal Coordinate analysis (PCoA) of beta diversity measured by Bray-Curtis dissimilarity of 256 kefir metagenomes, calculated for species-level composition. The PCoA plot displays the microbiome differences between kefir metagenomes dominated by different microbial species. Metagenomes are coloured by the dominant species ■ *A. orientalis,* ■ *Ac. sp.* TTHO-4, ■ *L. helveticus*, ■ *L. kefiranofaciens,* ■ *Lla. lactis,* ■ *Leuc. mesenteroides* and ■ *P. helleri*. Metagenomes are shaped according to time point (▲ 8 hr or ● 24 hr). The 95% confidence ellipse is shown for metagenomes dominated by ■ *Lla. lactis,* ■ *L. helveticus* and ■ *L. kefiranofaciens*.**(B.)** PCoA of beta diversity separated by the dominant species *L. helveticus*, *L. kefiranofaciens* and *Lla. lactis.* Metagenomes are coloured by the relative abundance of the dominant species.	68
Figure 3	**Figure 3:** Putative new species of the genus *Acetobacter* and *Lactobacillus* within the kefir microbiome.Phylogenetic tree of high quality MAGs of the genus (**A.** and **B.**) Phylogenetic tree of high-quality MAGs of the genus **(A.)** Acetobacter and **(B.)** Lactobacillus, including putative new species. Outer band coloured green, blue and red represent % contamination, completion and strain heterogeneity, respectively, as assessed using checkM. The coloured tip of the phylogram corresponds to the genus classification of the MAG. Genus classifications include ■ *Acetobacter* and ■ *Lactobacillus*. The tip shape of the phylogram indicates which MAGs are known species (●) or potentially novel species (▲).	70
Figure 4	**Figure 4:** Strain level phylograms of *Lla. lactis* and corresponding multiple sequence alignment (MSA) files generated by StrainPhlAn 3. Colored tips of the phylogram corresponds to the time point of the kefir metagenome from which the strain was recovered. The MSAs contains completely specified nucleotides (A, C, G, T), ambiguous characters (N) and alignment gaps (–) inserted between the nucleotides of some of the sequences Other representative species phylograms are available in Figures S5–S20.	71
Figure 5	**Figure 5:** Analysis of representative strains, with joint identification by StrainPhlAn 3 and PanPhlAn 3. **(A.)** Collective intra-sample strain diversity across metagenomes for each of the outlined 15 bacterial species (X-axis). Each coloured boxplot displays the range of polymorphic sites expressed as percentages (Y-axis); along the MLSA of strains. **(B.)** Results of correlation analysis performed using the genetic distance between StrainPhlAn 3 (Y-axis) and PanPhlAn 3 (X-axis) predicted strains calculated using the Kendall correlation method. **(C.)** Graphical presentation of PanPhlAn 3 predicted gene families (Y-axis) that compose the pangenome of the representative species (X-axis). Details include number of gene families detected per pangenome and the number of gene families classified as ■	72

	accessory, ■ core, ■ specific to cluster 1, ■ specific to cluster 2, ■ specific to cluster 3 and ■ strain specific. **(D.)** Number of strains detected (Y-axis) for each representative species (X-axis) by the ■ StrainPhlAn 3 or ■ PanPhlAn 3 approach. Only representative species, with joint identification by StrainPhlAn 3 and PanPhlAn 3 were considered in this graph.	
Figure 6	**Figure 6:** Pangenome of *Lla. lactis* strains detected within the milk kefir microbiome. Representative heatmap representing the pangenome of *Lla. lactis* strains; rows represent PanPhlAn 3 detected strains and columns represent the UniRef.90 gene families identified per strain. Row side annotations display the grouping of each gene family, e.g., core gene or specific to cluster 1. Column side annotations display the assigned cluster of the detected strain according to the PanPhlAn 3 hierarchical clustering methodology (see Methods) and the time point that the strain was recovered from.	74
Figure 7	**Figure 7:** Variance in the metabolic potential among kefir metagenomes explained by the abundances of species/strains types. Principal Coordinate analysis (PCoA) of beta diversity measure by Bray-Curtis dissimilarity for 256 kefir samples, calculated for metabolic composition as predicted using HUMAnN 3.0. Samples are shaped according the dominant species/strain subcluster (● *Lla. lactis*, ● *Lla. lactis* subcluster - 1, ▲ *Lla. lactis* subcluster - 2, ■ *L. helveticus* or ◆ *L. kefiranofaciens*). ● *Lla. lactis* represent kefir metagenomes in which complimentary strain profiling data was absent. Samples are coloured by the relative abundance of the dominant species/strain subcluster.	78
Figure 8	**Figure 8:** Diversity measures of community types within the kefir microbiome and characteristic metabolic function of *Lla. lactis* subcluster - 1, *Lla. lactis* subcluster - 2, *L. helveticus* and *L. kefiranofaciens* representative MAGs. **(A.)** Alpha diversity index observed for each community type when considering compositional data obtained using Kraken2. **(B.)** Alpha diversity index observed for each community type when considering functional data obtained using HUMAnN 3.0. The colouring of the boxplots represent samples dominated by ■ *L. helveticus*, ■ *L. kefiranofaciens,* ■ *Lla. lactis* subcluster - 1 and ■ *Lla. lactis* subcluster - 2. Pairwise tests were carried out between the dominant microbial species/strain using the Dunn's test for nonparametric data. **(C.)** Presence/absence of annotations indicative of characteristic metabolic functions involved in catalyzing carbohydrate degradation (CAZymes), methanogenesis and methanotropy, other reductases and short chain fatty acids and alcohol conversions (Y-axis) across representative MAGs identified using dRep and derived from the community types *Lla. lactis* subcluster - 1, *Lla. lactis* subcluster - 2, *L. helveticus* and *L.* kefiranofaciens (X-axis). The colour of each cell of the heat map represents the presence ■ or absence ■ of annotations.	81
Figure 9	**Figure 9:** Separation of *Lla. lactis* subclusters based on Average nucleotide identity (ANI) and relative abundance across kefir metagenomes **(A.)** Average nucleotide identity (ANI) gaps exist near ~91% ANI, between representatives of *Lla. lactis* subcluster - 1 and *Lla. lactis* subcluster - 2. Each dot represents the ANI (X-axis) and genome alignment percentage values (Y-axis) resulting from pairwise comparison within and between subclusters. Colours represent the ANI values. **(B.)** Heatmap displaying the relative abundance across 256 metagenomes of the species *Lla. lactis*, predicted using Kraken2, and *Lla. lactis* subsp. *lactis* and *Lla. cremoris* subsp. *cremoris* both predicted using metacache.	83
Figure S1	**Figure S1:** Kefir grain sampling initiative undertaken in this study. World map showing the countries from which the kefir grains were sourced. The different colour circles represent the country in which the kefir grains originated. The diameter of the circle displays the number of grains sourced from that country.	94

Figure S2	**Figure S2:** Graphical representation of the experimental design implemented in this study. Steps involved in the experimental design included kefir fermentations, sample collection, DNA extraction, library preparation, sequencing and bioinformatics analysis.	95
Figure S3	**Figure S3**: The microbial community patterns of the milk kefirs with time. **(A.)** Species that were detected at ≥10% relative abundance in at least one 8 hr sample. **(A.)** Species that were detected at ≥10% relative abundance in at least one 8 hr-replicate sample. **(C.)** Species that were detected at ≥10% relative abundance in at least one 24 hr sample and **(D.)** Species that were detected at ≥10% relative abundance in at least one 24 hr-replicate sample.	96
Figure S4	**Figure S4:** High quality metagenome assembled genomes assembled in this study. Phylogenetic tree of the 613 high quality MAGs assembled from kefir metagenomics datasets, obtained in this study. Outer rings coloured green, blue and red represent % contamination, completion and strain heterogeneity respectively, as assessed using checkM. The coloured tip of the phylogram corresponding to the genus classification of the MAG.	97
Figure S5	**Figure S5:** Strains of *A. fabarum* detected across kefir metagenomes. Strain level phylograms of *A. fabarum* strains and corresponding multiple sequence alignment (MSA) files generated by StrainPhlAn 3. Coloured tips of the phylogram corresponds to the time point of the kefir metagenome from which the strain was recovered. Colouring of labels represents the strain subcluster by which the strain was grouped.	98
Figure S6	**Figure S6:** Strains of *A. orientalis* detected across kefir metagenomes. Strain level phylograms of *A. orientalis* strains and corresponding multiple sequence alignment (MSA) files generated by StrainPhlAn 3. Coloured tips of the phylogram corresponds to the time point of the kefir metagenome from which the strain was recovered. Colouring of labels represents the strain subcluster by which the strain was grouped.	99
Figure S7	**Figure S7:** Strains of *Ac. guillouiae* detected across kefir metagenomes. Strain level phylograms of *Ac. guillouiae* strains and corresponding multiple sequence alignment (MSA) files generated by StrainPhlAn 3. Coloured tips of the phylogram corresponds to the time point of the kefir metagenome from which the strain was recovered. Colouring of labels represents the strain subcluster by which the strain was grouped.	100
Figure S8	**Figure S8:** Strains of *L. helveticus* detected across kefir metagenomes. Strain level phylograms of *L. helveticus* strains and corresponding multiple sequence alignment (MSA) files generated by StrainPhlAn 3. Coloured tips of the phylogram corresponds to the time point of the kefir metagenome from which the strain was recovered. Colouring of labels represents the strain subcluster by which the strain was grouped.	101
Figure S9	**Figure S9:** Strains of *L. kefiranofaciens* detected across kefir metagenomes. Strain level phylograms of *L. kefiranofaciens* strains and corresponding multiple sequence alignment (MSA) files generated by StrainPhlAn 3. Coloured tips of the phylogram corresponds to the time point of the kefir metagenome from which the strain was recovered. Colouring of labels represents the strain subcluster by which the strain was grouped.	102
Figure S10	**Figure S10:** Strains of *Lla. chungangensis* detected across kefir metagenomes**.** Strain level phylograms of *Lla. chungangensis* strains and corresponding multiple sequence alignment (MSA) files generated by StrainPhlAn 3. Coloured tips of the phylogram corresponds to the time point of the kefir metagenome from which the strain was recovered. Colouring of labels represents the strain subcluster by which the strain was grouped.	103
Figure S11	**Figure S11:** Isolates of *Lla. phage bIL285* detected across kefir metagenomesStrain level phylograms of *Lla. phage bIL285* strains and corresponding multiple sequence alignment (MSA) files generated by StrainPhlAn 3. Coloured tips of the phylogram corresponds to the time point of the kefir metagenome from which	104

	the strain was recovered. Colouring of labels represents the strain subcluster by which the strain was grouped.	
Figure S12	**Figure S12:** Isolates of *Lla. phage bIL309* detected across kefir metagenomes. Strain level phylograms of *Lla. phage bIL309* strains and corresponding multiple sequence alignment (MSA) files generated by StrainPhlAn 3. Coloured tips of the phylogram corresponds to the time point of the kefir metagenome from which the strain was recovered. Colouring of labels represents the strain subcluster by which the strain was grouped.	105
Figure S13	**Figure S13:** Isolates of *Lla.* phage bIL310 detected across kefir metagenomes Strain level phylograms of *Lla.* phage bIL310 strains and corresponding multiple sequence alignment (MSA) files generated by StrainPhlAn 3. Coloured tips of the phylogram corresponds to the time point of the kefir metagenome from which the strain was recovered. Colouring of labels represents the strain subcluster by which the strain was grouped.	106
Figure S14	**Figure S14:** Strains of *Leuc. mesenteroides* detected across kefir metagenomes. Strain level phylograms of *Leuc. mesenteroides* strains and corresponding multiple sequence alignment (MSA) files generated by StrainPhlAn 3. Coloured tips of the phylogram corresponds to the time point of the kefir metagenome from which the strain was recovered. Colouring of labels represents the strain subcluster by which the strain was grouped.	107
Figure S15	**Figure S15:** Strains of *Leuc. pseudomesenteroides* detected across kefir metagenomes. Strain level phylograms of *Leuc. pseudomesenteroides* strains and corresponding multiple sequence alignment (MSA) files generated by StrainPhlAn 3. Coloured tips of the phylogram corresponds to the time point of the kefir metagenome from which the strain was recovered. Colouring of labels represents the strain subcluster by which the strain was grouped.	108
Figure S16	**Figure S16:** Strains of *Leuc. suionicum* detected across kefir metagenomes. Strain level phylograms of *Leuc. suionicum* strains and corresponding multiple sequence alignment (MSA) files generated by StrainPhlAn 3. Coloured tips of the phylogram corresponds to the time point of the kefir metagenome from which the strain was recovered. Colouring of labels represents the strain subcluster by which the strain was grouped.	109
Figure S17	**Figure S17:** Strain level phylograms of *Pr. freudenreichii* and corresponding multiple sequence alignment (MSA) files generated by StrainPhlAn 3. Coloured tips of the phylogram corresponds to the time point of the kefir metagenome from which the strain was recovered. Colouring of labels represents the strain subcluster by which the strain was grouped.	110
Figure S18	**Figure S18:** Strain level phylograms of *P. helleri* and corresponding multiple sequence alignment (MSA) files generated by StrainPhlAn 3. Coloured tips of the phylogram corresponds to the time point of the kefir metagenome from which the strain was recovered. Colouring of labels represents the strain subcluster by which the strain was grouped.	111
Figure S19	**Figure S19:** Strain level phylograms of *P. lundensis* and corresponding multiple sequence alignment (MSA) files generated by StrainPhlAn 3. Coloured tips of the phylogram corresponds to the time point of the kefir metagenome from which the strain was recovered. Colouring of labels represents the strain subcluster by which the strain was grouped.	112
Figure S20	**Figure S20:** Strain level phylograms of *S. thermophilus* and corresponding multiple sequence alignment (MSA) files generated by StrainPhlAn 3. Coloured tips of the phylogram corresponds to the time point of the kefir metagenome from which the strain was recovered. Colouring of labels represents the strain subcluster by which the strain was grouped.	113
Figure S21	**Figure S21:** Pangenome of *A. fabarum* strains detected within the milk kefir microbiome. Representative heatmap representing the pangenome of *A. fabarum* strains; rows represent PanPhlAn 3 detected strains; columns represent the gene families identified per strain. Row side annotations display the grouping of each	114

	gene family, e.g., core gene or specific to cluster 1. Column side annotations display the assigned cluster of the detected strain according to the StrainPhlAn 3 and PanPhlAn 3 clustering methodology (see methods) and the time point that the strain was recovered from.	
Figure S22	**Figure S22:** Pangenome of *A. orientalis* strains detected within the milk kefir microbiome. Representative heatmap representing the pangenome of *A. orientalis* strains; rows represent PanPhlAn 3 detected strains; columns represent the UniRef.90 gene families identified per strain. Row side annotations display the grouping of each gene family, e.g., core gene or specific to cluster 1. Column side annotations display the assigned cluster of the detected strain according to the StrainPhlAn 3 and PanPhlAn 3 clustering methodology (see methods) and the time point that the strain was recovered from.	115
Figure S23	**Figure S23:** Pangenome of *L. helveticus* strains detected within the milk kefir microbiome. Representative heatmap representing the pangenome of *L. helveticus* strains*;* rows represent PanPhlAn 3 detected strains; columns represent the UniRef.90 gene families identified per strain. Row side annotations display the grouping of each gene family, e.g., core gene or specific to cluster 1. Column side annotations display the assigned cluster of the detected strain according to the StrainPhlAn 3 and PanPhlAn 3 clustering methodology (see methods) and the time point that the strain was recovered from.	116
Figure S24	**Figure S24:** Pangenome of *L. kefiranofaciens* strains detected within the milk kefir microbiome. Representative heatmap representing the pangenome of *L. kefiranofaciens* strains; rows represent PanPhlAn 3 detected strains; columns represent the UniRef.90 gene families identified per strain. Row side annotations display the grouping of each gene family, e.g., core gene or specific to cluster 1. Column side annotations display the assigned cluster of the detected strain according to the StrainPhlAn 3 and PanPhlAn 3 clustering methodology (see methods) and the time point that the strain was recovered from.	117
Figure S25	**Figure S25:** Pangenome of *Leuc. mesenteroides* strains detected within the milk kefir microbiome. Representative heatmap representing the pangenome of *Leuc. mesenteroides* strains*;* rows represent PanPhlAn 3 detected strains; columns represent the UniRef.90 gene families identified per strain. Row side annotations display the grouping of each gene family, e.g., core gene or specific to cluster 1. Column side annotations display the assigned cluster of the detected strain according to the StrainPhlAn 3 and PanPhlAn 3 clustering methodology (see methods) and the time point that the strain was recovered from.	118
Figure S26	**Figure S26:** Pangenome of *Leuc. pseudomesenteroides* strains detected within the milk kefir microbiome. Representative heatmap representing the pangenome of *Leuc. pseudomesenteroides* strains; rows represent PanPhlAn 3 detected strains; columns represent the UniRef.90 gene families identified per strain. Row side annotations display the grouping of each gene family, e.g., core gene or specific to cluster 1. Column side annotations display the assigned cluster of the detected strain according to the StrainPhlAn 3 and PanPhlAn 3 clustering methodology (see methods) and the time point that the strain was recovered from.	119
Figure S27	**Figure S27:** Pangenome of *P. helleri* strains detected within the milk kefir microbiome. Representative heatmap representing the pangenome of *P. helleri* strains; rows represent PanPhlAn 3 detected strains; columns represent the UniRef.90 gene families identified per strain. Row side annotations display the grouping of each gene family, e.g., core gene or specific to cluster 1. Column side annotations display the assigned cluster of the detected strain according to the StrainPhlAn 3 and PanPhlAn 3 clustering methodology (see methods) and the time point that the strain was recovered from.	120
Figure S28	**Figure S28:** Pangenome of *S. thermophilus* strains detected within the milk kefir microbiome. Representative heatmap representing the pangenome of *S. thermophilus* strains; rows represent PanPhlAn 3 detected strains; columns represent the UniRef90 gene families identified per strain. Row side annotations	121

	display the grouping of each gene family, e.g., core gene or specific to cluster 1. Column side annotations display the assigned cluster of the detected strain according to the StrainPhlAn 3 and PanPhlAn 3 clustering methodology (see methods) and the time point that the strain was recovered from.	
Figure S29	**Figure S29:** Core metabolic pathway and the associated species in the kefir microbiome. Stacked bar chart of species contributions to the relative abundance of each core metabolic pathway in the kefir microbiome across samples. The contributing species are coloured individually; as *L. helveticus*, *L. kefiranofaciens*, *Lla. lactis and* other.	122
Chapter 3		
Figure 1	**Figure 1:** Changes in microbiome diversity following FMP, Inulin and Kefir consumption. **(A.)**: Principal Coordinate analysis (PCoA) of beta diversity by intervention, measured by Bray-Curtis dissimilarity and calculated for species-level composition **(B.)**: Changes in alpha diversity post interventions, values calculated using the Shannon and Simpson diversity metrics. P- values represent the results of separate Wilcoxon signed-rank tests. **(C.)**: Changes in relative abundance of *Lactococcus raffinolactis* post kefir consumption. The bounds, whiskers and percentile of each box plot represented maximum, 75 percentile, median, 25 percentile and minimum from the top to the bottom respectively.	132
Figure 2	**Figure 2:** Prevalence of *Lla. raffinolactis* strains within the milk kefir microbiome and host adaptability genes within the pan- genome of *Lla. raffinolactis* **(A.)** Prevalence of *Lla. raffinolactis* across 256 kefir metagenomes generated by Walsh *et al* and assessed using Kraken2. **(B.)** Prevalence of genes with potential applications in host adaptability (Y -axis) across references of species *Lla. raffinolactis* (X -axis). The color of each cell of the heatmap represents the presence (red) or absence (blue) of host adaptability genes. Row side annotations include gene classifications separated into "Colonisation", and "Survival", summary descriptions of the pan- genomes of *Lla. raffinolactis*, separated into "Core", "Accessory" and "Strain specific" based on the prevalence of genes across reference genomes. Column side annotations represent the source data and sub cluster of the (Meta) genome.	134
Figure 3	**Figure 3:** The presence of *Lla. raffinolactis* coincides with statistically significant shifts in the metabolome of the kefir cohort. **(A.)**: Correlations between the concentrations of metabolites and the relative abundance of *Lla. raffinolactis.* **(B.)**: Changes in the quantities of the positively correlated metabolites Acetic, Betaine, N,N-Dimethylglycine and Succinic acid, post kefir treatment, separated according to the undetectable (gray) or present (red) status of *Lla. raffinolactis.* P- values reflect the output of separate Wilcoxon signed ranked tests assessing the change in metabolites between participants of the kefir cohort. The bounds, whiskers and percentile of each box plot represented maximum, 75 percentile, median, 25 percentile and minimum from the top to the bottom respectively.	135
Figure S1	**Figure S1:** Changes in metabolic pathway diversity in the gut microbiome following FMP, Inulin and Kefir consumption. Principal Coordinate analysis (PCoA) of beta diversity by intervention measured by Bray-Curtis dissimilarity and calculated from the HUMANn3 metabolic pathway abundance profiles.	142
Chapter 4		
Figure 1	**Figure 1:** Experimental design of Kefir4All (including details of the surveys undertaken). Stages included 'Project Initiation', 'Project Monitoring' and 'Project Completion'. 'Project Initiation' included participant recruitment, training and distribution of the fermentation equipment. A 'Workshop Survey' was used to capture feedback from participants at this stage of the project. In the 'Project Monitoring' stage, participants performed repeat fermentations in private households for up to 21 weeks, and recorded details of milk kefir or water kefir fermentations using the Survey Monkey platform [24]. During this period the 'Fermentation Survey' was used to assess the participation of the citizen	207

	scientists. Further insights into the level of participation was gained by assessing the number of kefir grains and liquids returned at week (wk) 1, wk5, wk9, wk13, wk17 and wk21. T0 samples represent control milk kefir and water kefir grain and liquid samples fermented under laboratory/controlled conditions and produced by professional scientists. At 'Project Completion', participants received a report in which individualised results of their fermenting activities were detailed. Participants were further invited to provide feedback through the 'Project Completion Survey'.	
Figure 2	**Figure 2:** Overview of the survey results collected for the 'Workshop Survey' with significant associations as determined using the Fisher's exact test. Plots display the relationship between categorical variables collected in the 'Workshop Survey', conducted prior to commencing the Kefir4All project. Plots are detailed as the following **(A.)** The relationship between variables **'**Rate Interest in fermented foods' and 'Are you participating in the Citizen Science project?', variables include Yes and No ; **(B.)** The relationship between variables What is your current interest in Microbiology?' and 'Are you participating in the Citizen Science project?'; **(C.)** The relationship between variables 'Has this talk improved your interest and general knowledge of microbiology and fermented foods?' and 'What is your current interest in Microbiology?', variables include Interested, Neutral, Disinterested, Very disinterested and I don't know what it means; **(D.)** The relationship between variables 'Are you continuing Biology and or Home-Economics' and 'What is your current interest in Microbiology?'; **(E.)** The relationship between variables 'Rate your background knowledge about fermentation' and **'**Rate Interest in fermented foods', variables include Extremely interested, Interested, Not at all interested, Not so interested and Somewhat interested; **(F.)** The relationship between variables 'Rate your background knowledge about fermentation' and **'**What is your current interest in Microbiology? **(G.)** A world cloud generated for individualised comments made by responders to the question entitled 'Please tell us what you thought of our workshop'.	211
Figure 3	**Figure 3:** Overview of the contributions of the citizen scientists during Kefir4All. **(A.)** Number of citizen scientists assigned to the fermentation categories; 'Regular fermenters', 'Non-fermenters', 'Rare fermenters', and 'Occasional fermenters'. **(B.)** Breakdown of the number of fermentation categories according to the groups (Schools and general members of the public) (X-axis). **(C.)** Number of survey answers, according to weeks of the study (X-axis); the colours on the bar charts represents the number of surveys completed by participants of each of the fermentation categories during each timeframe of the study (X-axis). **(D.)** Number of samples (milk kefir and water kefir grain and liquid samples) provided at each week of the study (X-axis); the colouring of the bar charts represents the number of samples provided by participants of the fermentation categories at each week of the study.	214
Figure 4	**Figure 4: (A.)** The first of the leaderboard plots displaying the collective survey responses per group recorded within a week of fermentation, and ranked according to the number of surveys completed, with the highest group represented with the icons 'First place' and a trophy, line plots are coloured by the school/group. **(B.)** Collective number of survey's (Y-axis) answered each day (X-axis) since the beginning of the week bar charts are coloured according to the number of surveys contributed by the group/school during each day. **(C.)** Collective number of survey's (Y-axis) answered by each group per day (X-axis) since the beginning of the project, line plots are coloured by the school/group.	215

Figure 5	**Figure 5:** Example figures included in the personalised report of a representative citizen scientist. **(A.)** Samples provided by a representative citizen scientist during the Kefir4All study, separated by weeks (wk), a red colour indicates that samples (kefir grain and liquid) (Y-axis) were provided at the outlined time point (X- axis) a white colour indicates that samples were not provided at the outlined time point (X-axis). **(B.)** Changes in the weight of the citizen scientist's kefir milk grain (in grams) over the course of the project, separated by the weeks (wk) of the study. **(C.)** Changes over the weeks (wk) of the study in the detection and relative abundance of species (X-axis) detected at >1% relative abundance (RA) in at least one of the samples provided by the citizen scientists, separated by the grain and liquid samples.	216
Figure 6	**Figure 6:** Association of fermentation parameters in the milk and water kefir fermentations and the occurrence of fermentation characteristics, variables include ■ It is bubbling, ■ It looks unchanged, ■ The milk has separated (whey & curds), ■ The milk has thickened and ■ There is a biofilm on top of the liquid. **(A.)** Breakdown of the number of 'Fermentation categories' (X-axis) according to the fermentation characteristics **(B.)** Bubble plot displaying the Chi- square residuals values between the fermentation parameters in the milk and water kefir fermentations. Sub plots show the associations between 'Fermentation categories', 'Grain Wash', Grain weight' ,'Liquid volume', 'pH' and 'Room temperature' and the fermentation characteristics 'Bubbling', 'Coagulation', 'Unchanged' and 'Whey & curds' in the milk kefir fermentations and associations between the fermentation parameters 'Fermentation categories', 'Grain Wash', Grain weight' ,'Liquid volume', 'pH', 'Room temperature', 'Sugar weight', and Water type' and the fermentation characteristics 'Biofilm', 'Bubbling' and 'Unchanged' in the water kefir fermentations.	218
Figure 7	**Figure 7: (A.)** Number of metagenomes (X-axis) derived from kefir grains sourced from different countries (Y-axis). **(B.)** Dendrogram presenting a hierarchical overview of species profiles detected in the milk kefir and water kefir microbiomes using the metachache pipeline. Tips of the dendrogram are coloured according to the kefir type; ■ Grain or ■ Liquid. The first outer ring displays the results of hierarchical clustering; colour indicates the grouping of each metagenome into one of the eight clusters. Clusters are coloured as follows ■ 1, ■ 2, ■ 3, ■ 4, ■ 5, ■ 6, ■ 7 and ■ 8. The second outer ring represents a circular heatmap displaying the Jaccard coefficients of each of the eight clusters detected using hierarchical clustering. The third outer ring represents the data source the metagenomes were generated in and is coloured as follows; ■ 'Breselge *et al* - Water kefir', ■ 'Kefir4All - Water kefir', ■ 'Walsh *et al* - Milk kefir' and ■ 'Kefir4All - Milk kefir'. **(C.)** Number of grain or liquid metagenomes, separated according to the kefir type - milk kefir or water kefir grain and liquid, and coloured according to the data source.	221
Figure 8	**Figure 8:** Survey answers from the citizen scientist (n=18) and the supervisory teachers (n=6) who coordinated the students in the school environment. **(A.)** 'Citizen scientists - How would you rate your current interest in the project?'; **(B.)** 'Teachers - How would you rate your current interest in the project?'; **(C.)**. 'Citizen scientists - Has this project improved your interest and general knowledge of microbiology and fermented foods?'; **(D.)** 'Teachers - Has this project improved your interest and general knowledge of microbiology and fermented foods?'; **(E.)** 'Citizen scientists - Has this project improved your scientific vocabulary?'; **(F.)** 'Teachers - Has this project improved the scientific vocabulary of participating citizen scientists in your school'; **(G).** Number of yes/no survey answers to the question entitled, 'Citizen scientists - Are you more interested in science as a result of this project?'; **(H.)** Number of yes/no survey answers to the question	222

	entitled, 'Teachers - Are your students more interested in science as a result of this project?'; **(I.)** Number of survey answers to the question entitled, 'Citizen scientists - How would you rate your overall satisfaction with the help you received from representatives of our project?'; **(J.)** Number of survey answers to the question entitled, 'Teachers - How would you rate your overall satisfaction with the help you received from representatives of our project?'; **(K.)** Number of yes/no survey answers to the question entitled, 'Citizen scientists - Did the arranged activities/merchandise' help to maintain your interest in the project?'; **(L.)** Number of yes/no survey answers to the question entitled, 'Teachers - Did the arranged activities/merchandise' help to maintain your interest in the project?'; **(M.)** 'Citizen scientists - Which activities/merchandise helped to maintain your interest?'; **(N.)** 'Teachers - Which activities/merchandise helped to maintain your interest?'; **(O.)** 'Citizen scientists - Did this project spark your interest in participating in other Citizen Science projects?'; **(P.)**'Teachers - Did engaging in the Kefir4All project interest you in getting your school involved in future Citizen Science projects?'.	
Figure S1	**Figure S1:** Overview of the survey results collected for the 'Workshop Survey', conducted prior to commencing the Kefir4All project. Responses are detailed for the following questions; **(A.)** 'Are you participating in the Citizen Science project?'; **(B.)** 'What is your current interest in Microbiology?'; **(C.)** 'Rate your background knowledge about Microbiology.' **(D.)** 'Rate you Interest in fermented foods'; **(E.)** 'Rate your background knowledge about fermentation'; **(F.)** ' How often do you eat fermented foods'; **(G.)** 'Have you ever fermented food before?'; **(H.)** 'Are you continuing Biology and or Home-Economics' **(I.)** 'Where do you currently get information about science?'; **(J.)** 'Has this talk improved your interest and general knowledge of microbiology and fermented foods?'.	280
Figure S2	**Figure S2:** World cloud of comments and replies from citizen scientists to the question entitled 'Why did you proceed with or stop the project?'.	281
Figure S3	**Figure S3:** World cloud generated for individualised comments made by citizen scientists to the question entitled 'What did you dislike about the project?'.	281
Figure S4	**Figure S4:** World cloud generated for individualised comments made by teachers responders to the question entitled 'What did you dislike about the project?'.	282
Figure S5	**Figure S5:** World cloud generated for individualised comments made by citizen scientists to the question entitled 'What did you like about the project?'.	282
Figure S6	**Figure S6:** World cloud generated for individualised comments made by teachers responders to the question entitled 'What did you like about the project?'	283
Figure S7	**Figure S7:** World cloud generated for individualised comments made by citizen scientists to the question entitled 'Do you have any suggestions for other interactions/activities that you feel would have benefited this project?'	283
Figure S8	**Figure S8:** World cloud generated for individualised comments made by teachers responders to the question entitled 'Do you have any suggestions for other interactions/activities that you feel would have benefited this project?'	284
Figure S9	**Figure S9:** World cloud generated for individualised comments made by citizen scientists to the question entitled 'Do you have any suggestions for future Citizen Science projects?'	284
Figure S10	**Figure S10:** World cloud generated for individualised comments made by teachers responders to the question entitled 'Do you have any suggestions for future Citizen Science projects?'	285
Chapter 5		
Figure 1	**Figure 1:** Graphical representation of the experimental design implemented in this citizen science initiative termed Kefir4All. Details include the total number of citizen scientists, further divided into milk and water kefir arms, the fermentation categories that the citizen scientists could choose from, and the steps involved in Kefir4All from the participants perspective. The steps involved in Kefir4All from the participants perspective were divided into (1) 'Regular fermentation', in which	295

	citizen scientists were asked to perform regular fermentations in private households for up to 21 weeks, (2) 'Collect metadata', in which citizen scientists recorded details of milk kefir or water kefir fermentations using the Survey Monkey platform [24] and (3) 'Collect samples' during which citizen scientists were asked to provide grains and liquids samples at week (wk) 01, wk05, wk09, wk13, wk17 and wk21. T0 samples represent control milk kefir and water kefir grain and liquid samples fermented under laboratory/controlled conditions and produced by professional scientists.	
Figure 2	**Figure 2: (A.)** Bubble plot displaying the species and relative abundance (RA) of detected microbial species at a quality threshold of ≥ 0.1% relative abundance in both the T0 milk and water kefir grain metagenomes. T0 grain metagenomes, representing the initial microbiome composition of the kefir grains prior to distribution to citizen scientists. **(B.)** Heatmap displaying the prevalent microbial species detected in milk and water kefir metagenomes (liquids and grains), derived from samples collected from citizen scientists over the course of the Kefir4All project (wk01-wk21) and T0 samples. T0 samples represent the initial microbiome composition of the kefir grains and the resulting liquids produced through fermentation of these grains in a controlled laboratory setting.	297
Figure 3	**Figure 3: (A.)** Beta-diversity boxplots displaying the Bray Curtis values across milk kefir grain metagenomes representing different timeframes of the project ('wk01', 'wk05, 'wk09', 'wk13' , 'wk17' and 'wk21') compared to the T0 milk kefir grain metagenome. Kruskal-Wallis and Dunn test was used to assess differences in beta diversity, since the data was nonparametric. Red dot represents the mean Bray Curtis between milk kefir grain metagenomes representing different timeframes of the project compared to the T0 milk kefir grain. **(B.)** Beta-diversity boxplots displaying the Bray Curtis values across water kefir grain metagenomes representing different timeframes of the project ('wk01', 'wk05, 'wk09', 'wk13', 'wk17' and 'wk21') compared to the T0 water kefir grain metagenome. Kruskal-Wallis and Dunn test was used to assess differences in beta diversity, since the data was nonparametric. Red dot represents the mean Bray Curtis between water kefir grain metagenomes representing different timeframes of the project compared to the T0 milk kefir grain. **(C.)** Heatmap displaying the Bray Curtis between kefir grain metagenomes from the same participant, representing different timeframes of the project.	300
Figure 4	**Figure 4: (A.)** Beta-diversity boxplots displaying the Bray Curtis values as determined based on functional composition, across milk grain metagenomes representing different timeframes of the project ('wk01', 'wk05, 'wk09', 'wk13', 'wk17' and 'wk21') compared to the T0 milk kefir grain metagenome. **(B.)** Beta-diversity boxplots displaying the Bray Curtis values as determined based on functional composition, across milk grain metagenomes representing different timeframes of the project ('wk01', 'wk05, 'wk09', 'wk13', 'wk17' and 'wk21') compared to the T0 milk kefir grain metagenome. Kruskal-Wallis and Dunn test was used to assess differences in beta diversity, since the data was nonparametric. Red dot represents the mean Bray Curtis between milk kefir grain metagenomes representing different timeframes of the project compared to the T0 milk kefir grain. **(C.)** Statistically significant strong spearman correlations (-.7 to -.1 or .7 to 1) between the relative abundance of level 3 subsystems and functional change as determined using the Bray Curtis.	302
Figure 5	**Figure 5: (A.)** Principal Coordinate analysis (PCoA) plot of beta diversity measure by Bray-Curtis dissimilarity of the total metagenomes considered in this study, as calculated for species composition obtained using metacache. Metagenomes are coloured according to the datasource, ■ 'Breselge *et al* - Water kefir', ■ 'Kefir4All - Milk kefir', ■ 'Kefir4All - Water kefir' and ■ 'Walsh et al - Milk kefir'. Larger circles ● represent centroid points between each data source. Distribution, skewness, and kurtosis of metagenomes are visualised by the Kernel density	305

	estimate displayed on the X and Y-axis (B.) Subplots of the PCoA plot of beta diversity separated by dominant species (most abundance), *Lac. paracasei, L. helveticus, L. kefiranofaciens*, *Lla. cremoris*, *Lla. lactis, Len. hilgardii* and *Z. mobilis.* Metagenomes are coloured according to the relative abundance of the metagenome with values ranging from 0% relative abundance (RA) (Blue) to 60% RA (Red). Metagenomes are shaped according the kefir type, ● Grain and▲ Liquid.	
Figure 6	**Figure 6: (A.)** Alpha-diversity boxplots displaying the Shannon diversity values across kefir types ('Milk- Grains', 'Milk -liquid', 'Water -grains' and 'Water liquid'). Kruskal-Wallis and Dunn test was used to assess differences in alpha diversity between kefir metagenome types, since the data was nonparametric. **(B.)** Alpha-diversity boxplots displaying the Shannon diversity values across fermentation conditions (Laboratory controlled or household conditions). Kruskal-Wallis and Dunn test was used to assess differences in alpha diversity between fermentation conditions, since the data was nonparametric. **(C.)** Persistence of putative environmental microbial species (Y-axis) in either milk or water kefir grain metagenomes. Persistence was determined through detections at sequential timeframes (X-axis) in metagenomes derived from the same citizen scientists; corresponding liquid samples are also displayed. Size and colour of the bubbles represents the relative abundance of each of the putative environmental microbial species (Y-axis) at the various timeframes (wk01, wk05, wk09, wk13, wk17 and wk21).	307
Figure 7	**Figure 7: (A.)** Heatmap displaying the Spearman rank correlation coefficient between the fermentation parameters 'Total fermentation response frequency', 'Room temperature', 'Grain weight', Liquid volume and 'pH' and the prevalent microbial species (X-axis) detected in milk and water kefir metagenomes (liquids and grains), derived from samples collected from citizen scientists over the course of the Kefir4All project (wk01-wk21) and T0 samples. **(B.)** Lefse covariate analysis reveals the effect size (LDA score) (X-axis) of the fermentation parameters 'Fermentation category', 'Other fermented food', 'Pets' and 'Water type' on the relative abundance of the outlined species (Y-axis), separated according to the kefir type - milk kefir or water kefir, and coloured according to the various options available to each fermentation parameter. For example If the citizen scientists selected milk kefir, they had the option to ferment in Cow milk (whole), Cow milk (low fat), Cow milk (skim) and Goat milk	311
Figure 8	**Figure 8: (A.)** Bar chart displaying the number (X-axis), of strains detected (Y-axis), separated according to the kefir type - milk kefir or water kefir, and coloured according to the data source; ■ 'Milk kefir - Kefir4All ', ■ 'Water kefir - Kefir4All ', ■'Water kefir - Breselge *et al*.' and ■ 'Milk kefir - Walsh *et al*.'. **(B.)** Heatmap displaying the mean rate of nucleotide change for prevalent microbial species detected in milk and water kefir metagenomes (liquids and grains), derived from samples collected from citizen scientists over the course of the Kefir4All project (wk01-wk21) compared to T0 samples. T0 samples represent the initial microbiome composition of the kefir grains and the resulting liquids produced through fermentation of these grains in a controlled laboratory setting. **(C.)** Bar chart displaying the percentages of metagenome assembled genomes (MAGs) (X-axis) for each of the prevalent species (Y- axis) deemed to be the 'same strains' vs 'different strain'.	314
Figure S1	**Figure S1: (A.)** Beta-diversity boxplots displaying the Bray Curtis values between milk kefir grain metagenomes from the same participant, representing different timeframes of the project ('wk01 vs wk05', 'wk05 vs wk09', 'wk13 vs wk17' and 'wk17 vs wk21'). **(B.)** Beta-diversity boxplots displaying the Bray Curtis values between water kefir grain metagenomes from the same participant, representing different timeframes of the project ('wk01 vs wk05', 'wk05 vs wk09', 'wk13 vs wk17' and 'wk17 vs wk21'). **(C.)** Beta-diversity boxplots displaying the the Bray	323

	Curtis between milk kefir liquid metagenomes from the same participant, representing different timeframes of the project ('wk01 vs wk05', 'wk05 vs wk09', 'wk13 vs wk17' and 'wk17 vs wk21'). **(D.)** Beta-diversity boxplots displaying the Bray Curtis values between water kefir liquid metagenomes from the same participant, representing different timeframes of the project ('wk01 vs wk05', 'wk05 vs wk09', 'wk13 vs wk17' and 'wk17 vs wk21'). Kruskal-Wallis and Dunn test was used to assess differences in beta diversity, since the data was nonparametric. Red dot represents the mean Bray Curtis between metagenomes representing different timeframes of the project. **(E.)** Gower-diversity boxplots displaying the Gower distance between milk kefir fermentation surveys completed by the same participant, representing different timeframes of the project ('wk01 vs wk05', 'wk05 vs wk09', 'wk13 vs wk17' and 'wk17 vs wk21'). **(F.)** Gower-diversity boxplots displaying the Gower distance between water kefir fermentation surveys completed by the same participant, representing different timeframes of the project ('wk01 vs wk05', 'wk05 vs wk09', 'wk13 vs wk17' and 'wk17 vs wk21').	
Figure S2	**Figure S2:** Collective intra-sample strain diversity across metagenomes for each of the outlined bacterial species (X-axis). Each coloured boxplot displays the range of polymorphic sites expressed as percentages (Y-axis); along the MLSA of strains.	324
Figure S3	**Figure S3:** Phylogenetic-diversity boxplots displaying the phylogenetic distance values between Strain comparisons 'Intra-participant comparisons', 'Inter-participant comparisons' and 'Unrelated comparisons', of the outlined species (see title).	325
Figure S4	**Figure S4:** Correlations between species relative abundance and strain rates of nucleotide change.	326

List of Tables

Fig n	Description	Page no
Chapter 1		
Table 1	**Table 1**: Source codes, synopsis, tutorial and basic characteristics of the reviewed bioinformatic tools. Basic characteristics include RCR citation (a measure of the influence that publications introducing tools have on the research community), advantages and disadvantages of the tool e.g., extent of documentation.	8-15
Table 2	**Table 2:** Description of short-read based and metagenome assembly-based approaches available for the interpretation of metagenomic data. Details include the limitations and advantages of both approaches.	23
Chapter 3		
Table S1	**Table S1:** Changes in clinical parameters of participants following FMP, Inulin and Kefir consumption.	143- 197

Chapter 1

Bioinformatic approaches for studying the microbiome of fermented food

Published in *Critical reviews in microbiology*

(doi: 10.1080/1040841X.2022.2132850)

Authors: Liam H Walsh, Mairéad Coakley, Paul W O'Toole, and Paul D. Cotter

Contributions:

- **Candidate** wrote the review, with guidance from **MC**, **PWT**, and **PDC**

Section 1.0: Abstract

High-throughput DNA sequencing-based approaches continue to revolutionize our understanding of microbial ecosystems, including those associated with fermented foods. Metagenomic and metatranscriptomic approaches are state of the art biological profiling methods and are employed to investigate a wide variety of characteristics of microbial communities, such as taxonomic membership, gene content and the range and level at which these genes are expressed. Individual groups and consortia of researchers are utilising these approaches to produce increasingly large and complex datasets, representing vast populations of microorganisms. There is a corresponding requirement for the development and application of appropriate bioinformatic tools and pipelines to interpret this data. This review critically analyses the tools and pipelines that have been used or that could be applied to the analysis of metagenomic and metatranscriptomic data from fermented foods. In addition, we critically analyse a number of studies of fermented foods in which these tools have previously been applied, to highlight the insights that these approaches can provide.

Section 1.1: Introduction- Microbiome research and its relevance to food

The growth and activities of microbial communities associated with food influence its biological state, for example by promoting preservation through fermentation or degradation by spoilage. Since ancient times, fermentation-associated microorganisms and human nutrition have been linked, with the earliest records dating back to 7,000 BC. Fermentation of food is still a routinely practiced form of food production across the globe, producing culturally significant foods that are consumed daily by millions of people [1]. Fermented foods and beverages are classified by the International Scientific Association for Probiotics and Prebiotics (ISAPP) as "foods made through desired microbial growth and enzymatic conversions of food components" [2]. These enzymatic processes cause significant changes in the properties of foods, including dairy, meat, fish, vegetable, fruit and cereal substrates [3], providing a stabilizing effect, while adding flavour, aroma and distinctive features to the foods [4].

Fermented foods have recently attracted renewed interest among Western consumers, particularly because of an enhanced appreciation of their associated health benefits. These health benefits are attributed to the food's nutritional content, the fermenting microorganisms themselves and the by-products (postbiotics) of their metabolic activities. From a nutritional perspective, fermentation may remove some anti-nutrients such as allergens, and typically improves the micronutrient content of most foods by increasing the bioavailability of minerals and vitamins. Additionally, macronutrients such as carbohydrates and proteins may be more digestible following fermentation [5, 6]. Some studies support the hypothesis that fermented foods host microbial species/strains with health-promoting functionality, using food as a transport matrix and thus improving health outcomes [4, 7, 8].

By-products of fermentation have a significant effect on the sensory and nutritional properties of food [5, 9]. Some of these by-products, such as bioactive peptides and exopolysaccharides, have been associated with reduced occurrence of conditions such as diabetes, obesity, cardiovascular disease and hypertension [10, 11]. Despite evidence of the health benefits provided by some fermented foods, there is a lack of appropriately designed human trials to determine the effects of specific fermented foods on human health

[12]. Additionally, most health claims lack a sufficient mechanistic understanding, highlighting the need for further research to understand the methods by which fermented food microorganisms and/or their by-products might contribute to health, as well as the consumption level of the fermented food required for these health benefits to be conferred [5]. Despite the need for more clinical studies, our understanding of the microbial components of fermented foods has progressed significantly with recent advances in culture-independent high-throughput sequencing approaches. The expanded use of shotgun metagenomics and metatranscriptomics to study a considerable variety of microbial environments has provided a deep insight into the microbial composition, functional potential and gene expression in foods [13].

Section 1.2: Shotgun metagenomics

Whole metagenome shotgun sequencing (WMGS) provides an untargeted sequencing-based approach to assess metagenomic DNA from biological samples. Historically, culture-dependent approaches were used to characterise microbial communities associated with fermented foods. However, these methods are limited to culturable microorganisms, failing to provide insights for the yet to be cultured or difficult-to-culture microbes [14]. Such microbes can play significant roles in fermentation and prior to the introduction of high-throughput sequencing (HTS), few experimental applications existed to examine their physiology [15]. For example, in Pu-erh tea, a fermented beverage typically containing an array of species corresponding to multiple genera, only a select number of microorganisms have been successfully cultured, namely *Aspergillus* sp. and *Blastobotrys* sp., due to difficulties in replicating the fermentation environment in laboratory settings [16, 17]. Most studies analysing the compositional structure of microbial communities to date have used targeted sequencing approaches such as 16S rDNA gene sequencing, as opposed to WMGS, with the routine utilisation of WMGS initially being impeded by the associated costs and computational challenges [13, 18]. Such issues continue to be a major consideration with respect to the application of WMGS for large-scale longitudinal studies [19]. However, this is offset by the fact that WMGS sequencing offers several advantages.

Whole metagenome shotgun sequencing can provide both a functional and species level taxonomic profile

of bacterial, viral, archaeal and eukaryotic taxa, and recover fragmented draft genomes and genes present in the microbial ecosystem. In comparison, amplicon sequencing typically achieves a genus-level compositional insight and provides little functional insight for bacterial, archaeal or eukaryotic taxa (depending on the target marker selected), and fails to detect viral members due to an absence of appropriate phylogenetic single marker sequences [20-22]. The strength of WMGS is highlighted when considering strain level phylogenetic reconstructions at the population level. Previously this was only possible through time-consuming sequencing of isolates, but several papers have reported that the adoption of WMGS can produce comparable results [23]. Other key factors influencing the uptake of WMGS include further reductions in the cost of sequencing and the development of shallow shotgun metagenome sequencing (SSMS). Shallow shotgun metagenome sequencing provides an economic approach, by sequencing at a limited depth and has been demonstrated to produce comparable species and functional outputs to shotgun sequencing for a number of well characterised and simulated human metagenomes [19].

1.2.1 Overview of bioinformatic approaches for metagenomic data

An ever-increasing array of bioinformatic tools, pipelines and databases are available for the interpretation of metagenomic datasets (Figures 1, 2 & Table 1). The selection of these tools prior to experimentation is typically driven by the scientific question, sample type, properties of the tools and availability of reference databases (Figure 2). Before reviewing the bioinformatic tools available, a preliminary PubMed search was carried out using the query string “"Algorithms*[MESH Terms] OR Software* [MESH Terms] AND High-Throughput Nucleotide Sequencing* [MESH Terms] OR Metagenomics*[MESH Terms] OR Metagenomics / methods* [MESH Terms]" to acquire peer-reviewed research papers introducing novel and improved tools designed for the interpretation of metagenomic data. Records were assessed initially using title/abstract analysis, followed by full-text screening. Publications deemed appropriate were selected for citation analysis using the relative citation ratio (RCR) metric [24, 25] as calculated by iCite (https://icite.od.nih.gov/).

Through this analysis, a measure of the influence that established tools (2 years and older) had on the research community was obtained. 2-3 highly cited tools were selected per bioinformatic category, to demonstrate how underlying principles of bioinformatic tools (see below) are applied (Figure 1 & Table 1). These bioinformatic tools facilitate the analysis of millions of sequences in parallel from a diverse range of metagenomes or can be designed specifically for a niche application, such as the pathogen profiling of metagenomes obtained from archaeological sites [26]. Most tools employ heuristic approaches, due to the general complexity of metagenomic datasets. This data complexity arises due to the considerable volume of information that can be generated and the frequent need to integrate results obtained from multiple tools, which are often used in a non-standardised fashion [27]. Most bioinformatic tools available for the interpretation of metagenomic data are tailored for short-read sequencing data, reflecting the widespread use of Illumina platforms within the research community (Figure 1 & Table 1). Whole metagenome shotgun sequencing tools can be broadly assigned to two distinct approaches, referred to as ‘read-based' and 'assembly-based' approaches, which can be used separately or in parallel, depending on the study design. Many of the ‘read-based' and 'assembly-based' tools (Figure 1 & Table 1), while tailored to the complexity of metagenomics data can serve dual purposes and be applied to Metatranscriptional data. Caution is advised when considering the dual application of such tools, as many have not been sufficiently tested for their efficacy in classifying Metatranscriptional data [28]. For a detailed review of bioinformatic tools applied to Metatranscriptional data please see [28].

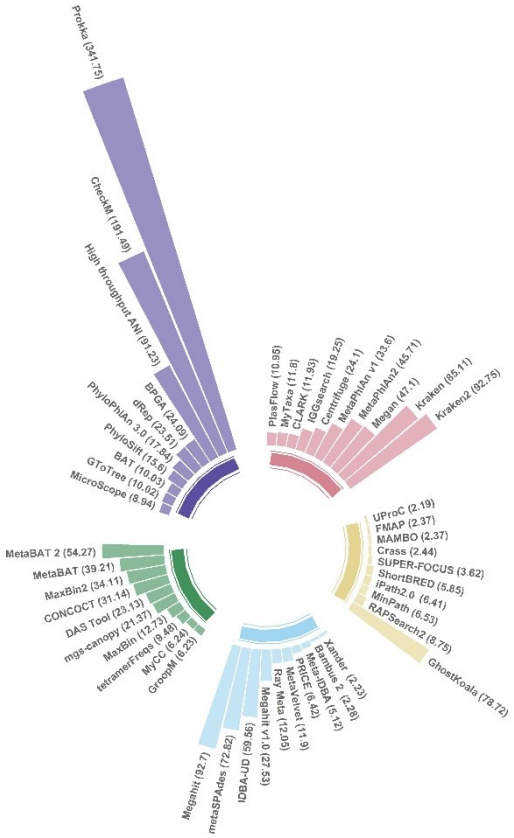


Figure 1: Relative citation ratio scores for selected peer-reviewed research papers, introducing established metagenomic tools (2 years and older) with the highest influence to the research community. Peer-reviewed research papers were selected through a literature mining methodology and assessed using a citation analysis as described above. Peer-reviewed research papers, discussed Bioinformatic tools with applications in ■ Short read taxonomic classification, ■ Short read functional classification, ■ Metagenomic assembly, ■ Metagenomic binning and ■ Analysis of recovered genomes.

Table 1: Source codes, synopsis, tutorial and basic characteristics of the reviewed bioinformatic tools. Basic characteristics include RCR citation (a measure of the influence that publications introducing tools have on the research community), advantages and disadvantages of the tool e.g., extent of documentation.

Tool type	Source code	Synopsis	Reference	Advantages	Disadvantages	Tutorial
Short-read approaches:						
Short read taxonomic classification						
MetaPhlAn3	https://github.com/biobakery/MetaPhlAn/tree/3.0	Short-read compositional tool that functions by aligning reads to an internal database of clade-specific and quasi marker genes through BLASTn (Altschul *et al.*, 1990).	Beghini *et al.*, 2021	• Improves viral detection • Bypasses the need for modifications and error checking of the initial metagenomic dataset • Compatible with many other bioinformatics tools e.g. HUMAnN3 and StrainPhlAn • Low memory requirements • Can be applied to metatranscriptomic data	• Poor fungal detection • Marker genes make up only a small proportion of a microbial genome	https://github.com/biobakery/biobakery/wiki/metaphlan3
Kraken2	https://github.com/DerrickWood/kraken2	Kraken2 is a taxonomic classification tool that uses k-mers to profile metagenomic reads or their translated protein sequences. The minimizer content of both k-mers contained in query sequences and Kraken's internal database are compared.	Wood, Lu and Langmead, 2019	• Highly customisable • Reports distinct minimiser count • Can be applied to metatranscriptomic data • Short run time	• Prone to false positives from closely related genomes • High memory requirements • Additional steps required to acquire a community composition estimate	https://github.com/DerrickWood/kraken2/wiki/Manual
Short read functional classification						
HUMAnN3	http://huttenhower.sph.harvard.edu/humann	HUMAnN3 queries metagenomic and metatranscriptomic datasets using a multi-search phase methodology. Initially, reads are aligned against a customized pangenome database of functionally annotated genes, built from the taxonomic identifications of	Beghini *et al.*, 2021	• Ability to assign functional features to contributing species • Up-to-date sequence database	• Long run-time • UniRef gene families often lack annotations	https://github.com/biobakery/biobakery/wiki/humann3

		MetaPhlAn3. Unmapped reads are translated and queried against a selectable protein database through DIAMOND by default.		• Bypasses the need for modifications and error checking of the initial metagenomic dataset • Highly customisable		
SUPER-FOCUS	https://edwards.sdsu.edu/SUPERFOCUS	Prior to functional analysis users may use FOCUS to acquire a taxonomic profile which directs the construction of a greatly reduced customized seed database, containing functionally related protein families found in the identified taxa. Functional profiling is achieved by querying metagenomics reads against the custom built database and one of the four clustered non-redundant SUPER-FOCUS databases.	Silva *et al.*, 2016	• Simply and easy to use command line functions • Ability to assign functional features to contributing species • Tiered classification system	• Limited documentation	https://github.com/metageni/SUPER-FOCUS
GhostKOALA	www.kegg.jp/ghostkoala/	KEGG Web service, annotations (K numbers) are assigned to genes within a metagenome through alignment to the non-redundant GENES database.	Kanehisa, Sato and Morishima, 2016	• Does not require command line interfaces to function • Provides high level overviews in visual format	• Limited data upload hinders its application in large metagenomics applications	https://www.kegg.jp/blastkoala/help_ghostkoala.html
Metagenome assembly-based approaches:						
Metagenomic assembly						
IDBA-UD	https://github.com/loneknightpy/idba	IDBA-UD uses increasing k-mer values to build a de Bruijn graph and accounts for errors through a flexible cut off values and contig alignment. Low depth repeat patterns are resolved using the dual ends of paired reads.	Peng *et al.*, 2012	• Designed to assemble low-depth regions and resolve repeat structures using progressive depth on contigs	• Only works with paired-end reads • Limited documentation	https://github.com/loneknightpy/idba https://denbi-metagenomics-workshop.readthedocs.io/en/lates

				•		t/assembly/idba_ud.html
MetaSPAdes	http://cab.spbu.ru/software/meta-spades/	MetaSPAdes adopts the iterative multiple paired k-mer approach of SPAdes to form a de Brujin graph. MetaSPAdes attempts to reconstruct species level consensus sequences using coverage information.	Nurk *et al.*, 2017	• Produces high quality assemblies	• Time-consuming • Large memory requirement	https://github.com/ngs-docs/2017-cicese-metagenomics/blob/master/assemble-metaspades.md
MEGAHIT	https://hku-bal.github.io/megabox	MEGAHIT shares the same multiple k-mer strategy as the IDBA assemblers to build a succinct de Bruijn graph, a memory efficient version of a de Bruijn graph. To further improve the memory requirements of MEGAHIT, low abundance k-mers are excluded from the graph. Despite the removal of low abundance k-mers, METAHIT implements a mercy-k-mer strategy.	Li *et al.*, 2016	• Produces large assemblies	• Prone to misassemblies	https://github.com/voutcn/megahit https://www.hadriengourle.com/tutorials/meta_assembly/
<u>Metagenomic binning</u>						
MetaBAT	https://bitbucket.org/berkeleylab/metabat/src/master/	MetaBAT utilises a clustering algorithm based on empirically derived statistics specific for each contig pair. Statistics include the TNF distance probability (TDP) and abundance distance probability (ADP), which are used to form a distance matrix. A specialise k-medoid clustering algorithm performs binning based on the distance matrix.	Kang *et al.*, 2019	• Optimal performance compared to MaxBin2 and CONCOCT (Uritskiy, DiRuggiero and Taylor, 2018)	• Poor performance when assembling similar genomes	https://www.hadriengourle.com/tutorials/meta_assembly/
CONCOCT	https://github.com/BinPro/CONCOCT	CONCOCT is a binning tool that groups contigs based on their k-mer frequency and abundances. Clustering of contigs is performed using a mixture of Gaussian distribution supplemented with a variational Bayesian approximation that identifies the optimal OTU number.	Alneberg *et al.*, 2014		• Produces lower number of assemblies compared to other tools such as MetaBAT 2 and MaxBin2	https://metagenomics-workshop.readthedocs.io/en/stable/binning/concoct.html

					• Poor performance when assembling similar genomes	
MaxBin2	https://sourceforge.net/projects/maxbin2/	MaxBin2 uses TNFs and abundance profiles to bin assembled contigs into draft genomes. Clustering is performed by an expectation-maximization algorithm, which assigns scaffolds to their respective bins based on median values of single-copy marker genes and probabilities scores.	Wu, Simmons and Singer, 2016	• Incorporates single-copy marker genes to distinguish similar genomes	• Difficulties in reporting the number of genomes estimated in high complexity datasets due to the application of single-copy marker genes.	https://hpc.ilri.cgiar.org/maxbin2-software
Downstream analysis of MAGs						
Prokka	https://github.com/tseemann/prokka#invoking-prokka	Prokka relies on established prediction tools and databases to perform genome annotation. Prediction based tools such as Prodigal compare genome queries against reference genomes at the nucleotide level. Gene annotation is performed by a translated search against multiple databases in a hierarchical fashion.	Seemann, 2014	• Suitable for use on a standard laptop computer • Fast run time • Customisable	• Additional steps required for data synthesis and visualization.	https://github.com/tseemann/prokka
CheckM	http://ecogenomics.github.io/CheckM/	Estimates quality of bacterial (meta)genomic assemblies using a hidden Markov model. Completion and contamination are determined based on frequency of marker genes and duplicate marker genes, respectively. Strain heterogeneity is informed by the differences in amino acid content between the identified marker genes.	Parks *et al.*, 2015	• Low error rate due to the usage of lineage-specific marker genes	• Limited application in eukaryotes due to the absence of eukaryotic sequences in the CheckM reference database	https://github.com/Ecogenomics/CheckM/wiki https://onestopdataanalysis.com/checkm-completeness-contamination/
IMG/M	https://img.jgi.doe.gov/cgi-bin/m/main.cgi	IMG/M uses multiple prediction tools such as Prodigal (Hyatt *et al.*, 2010) to detect gene CDSs within MAGs and other unique features. Predicted CDS genes are concatenated and, depending on the depth of coverage information,	Chen *et al.*, 2020	• Does not require command line interfaces to function • Identification of novel biosynthetic	• Exponential growth of the gene number	https://img.jgi.doe.gov/submit/doc/IMGSubmissionUserGuide.pdf

		an estimated measure of gene copies in scaffolds/contigs can be obtained. During functional annotation, the CDSs are subject to separate assignment steps.		gene clusters (BGCs)		
Genome scale metabolic models						
CarveMe	https://github.com/cdanielmachado/carveme	CarveMe introduces an automated top down approach to genome scale metabolic modelling. A high quality reference model is used as a template and tailored towards the organism of interest. Genome annotations are used to inform the removal of features from the template models, e.g., biological pathways predicted to be absent in the input organism.	Machado *et al.*, 2018	• Identification of minimal communities • Identification of key species • Predictions can be influenced by providing experimental data during the data input stage • Simple and easy to use command line functions		https://carveme.readthedocs.io/en/latest/usage.html
Metage2Metabo	https://github.com/AuReMe/metage2metabo	Input annotations of genomes/metagenomes are processed by Pathway Tools (Karp *et al.*, 2016) to construct a GEM (Genome Scale Metabolic Model). The GEM is then examined using a network expansion algorithm, which iteratively tracks the relationship between successive reactions, and metabolites including those produced through reactions, ceasing when the metabolites present can no longer fulfil the requirements for all potential pathways.	Belcour *et al.*, 2020		•	https://metage2metabo.readthedocs.io/en/latest/
Specialised bioinformatics tool:						
Viral Detection						

VirSorter 2	https://github.com/simroux/VirSorter	Input reads are initially screened to detect reads of a circular nature. Quality inspected sequences with 2 or more predicted ORFs are aligned to reference sequences in the PFAM and RefSeqABVir or Viromes databases, respectively. All annotated sequences are added into 5 separate groupings and their gene content is assessed using a sliding window plot. Each grouping provides a quantitative measure for each sequence of one of the following features, presence of viral hallmark genes, enrichment of viral-like or non-*Caudovirales* genes, uncategorised and short-reads, as well as depletions in PFAM associated genes and strand switching. These preselected features allow for the identification of both viral-specific, viral-like and non-viral features in a sequence.	Guo *et al.*, 2021	• Identification of rare viruses • Output is compatable with the (meta)genome annotation tool-DRAMv	• Prone to false positive results • High CPU usage • Additional tools required for downstream metagenomic analyses	https://www.protocols.io/view/getting-started-with-virsorter2-bhdij24e?version_warning=no&step=5
VirFinder	https://github.com/jessieren/VirFinder	Virfinder employs a k-mer-based machine learning approach to identify viral sequences in metagenomic data. Virfinder identifies the number of k-tuple word counts per sequence. A logistic regression model with lasso regularization predicts a confidence value between 0 and 1 based on the k-tuple word frequencies, where 0 indicates the sequence is not likely of viral origin and 1 suggests the query is likely of viral origin.	Ren *et al.*, 2017	• Improved prediction accuracy compared to alignment-based techniques and is more likely to characterise shorter sequences and those lacking reference genomes • Low CPU usage • Quick run-time	• Additional tools required for downstream metagenomic analysis • Prone to false positive results due to an absence of Eukaryotic sequences in VirFinder's training datasets. •	https://jessieren.github.io/VirFinder/

Strain level analysis						
PanPhlAn	https://github.com/SegataLab/panphlan/wiki/Home_3_0	PanPhlAn was designed to identify strains, and their gene content. Initially metagenomic reads are aligned against a reference species level pangenome. Gene family coverage is used to compute species unique gene presence/absence profiles. Abundant gene families of similar coverage are grouped together and are assumed to be from the same genome.	Beghini *et al.*, 2021	• Functional insight	• Reduced strain detection compared to StrainPhlAn	https://github.com/SegataLab/panphlan/wiki/Tutorial-3_0
StrainPhlAn	https://github.com/biobakery/MetaPhlAn/wiki/StrainPhlAn-3.0	StrainPhlAn examines the nucleotide differences of consensus marker genes obtained using MetaPhlAn3.	Beghini *et al.*, 2021	• Improved strain detection • Provides a polymorphic output file to report likelihood of the presence of multiple strains in the same sample	• No functional insight	https://github.com/biobakery/biobakery/wiki/StrainPhlAn3
<u>Antimicrobial Resistance:</u>						
ARG-ANNOT	http://backup.mediterranee-infection.com/article.php?laref=282&titre=arg-annot	Employs a local alignment-methodology to predict and identify AR features in bacteria. Features include existing, emerging, and putative ARGS, as well as chromosomal point mutations associated with an AR phenotype.	Gupta *et al.*, 2014	• Detection is limited to known ARG sequences • Prone to false negative results	• low false positive rate	https://www.mediterranee-infection.com/wp-content/uploads/2019/03/arg-annot-tutorials_doc.pdf
DeepARG	https://bitbucket.org/gusphdproj/deeparg-ss/src/master/	deepARG-LS and deepARG-SS are deep learning prediction models used to detect AR features. Both models map inputs to ARG categories based on a series of transfer and softMax activator functions which propagate through	Arango-Argoty *et al.*, 2018	• Increased sensitivity • Allows for the detection of novel ARGs and those with little	• Influenced by the quality of reference databases •	https://bench.cs.vt.edu/deeparg

		neurons (CPUs) in hidden layers of the models, to calculate the associated probability of relevant features.		sequence similarity to known ARGs		

1.2.2 Short-read approaches

Short-read tools profile microbial communities at the compositional and functional level (Figure 1). Short-read taxonomic classification tools function by aligning or mapping sequences to informative subsections of meta(genomic) data, e.g., taxon specific genes and k-mer features (short strings of DNA sequences). The short read taxonomic tool MetaPhlAn3 (metagenomic phylogenetic analysis 3) [29] functions by aligning reads to an internal database of clade-specific and quasi marker genes through BLASTn [30]. Quasi markers expand the scope of analysis enabling the profiling of viral and eukaryotic reads. The distinct nucleotide composition of marker genes reduces the occurrence of false positives and negatives, allowing a simplified workflow without needing to modify or error check the initial metagenomic dataset [29]. Kraken2 [31] is another popular taxonomic classification tool that uses k-mers to profile metagenomic reads or their translated protein sequences. The minimizer content of both k-mers contained in query sequences and Kraken's internal database are compared. Minimizers are nucleotide/protein sequences that are shared by multiple k-mers [31, 32].

Functional short read approaches often incorporate taxonomic identifications to construct a customised database of gene or protein features, known to be attributed to the identified taxa. Reads or k-mers are aligned or matched respectively to this customised database [33] and unmapped reads are further examined against larger protein databases containing annotated sequences to identify homologs [14]. These tools provide a summary-level characterisation of a metagenome, estimating feature abundance profiles [34, 35] and assigning reads to the most likely microbial lineage [33, 36]. The selective methodology reduces the time and computational requirements compared to traditional mapping approaches that attempt to label every read sequence from a metagenome [37]. HUMAnN3 (HMP Unified Metabolic Analysis Network) [29] is a short read functional profiling tool, querying metagenomic and metatranscriptomic datasets using a multi-search phase methodology. Initially, reads are aligned against a customised pangenome database of functionally annotated genes, built from the taxonomic identifications of MetaPhlAn3. By incorporating MetaPhlAn3, functional units can be linked to their taxonomic source.

Unmapped reads are translated and queried against a selectable protein database through DIAMOND by default. SuperFocus [33] also applies a similar methodology and can build a reduced seed database informed by the taxonomic identifications of FOCUS [38] for the initial alignment phase [33].

1.2.3 Metagenome assembly-based approaches

1.2.3.1 Metagenomic assembly

Assembly methods merge the consecutive k-mer content of metagenomic reads, based on their overlapping sequence similarity, into single contiguous sequences (contigs) of the shortest possible length. Assembly is performed to simplify data analysis processes, such as homology-based searches, as a single read contains limited information. A commonly used metagenomic assembly technique is global assembly, which attempts to construct all the genomes present within a sample [39]. Most popular metagenomic assemblers construct De Bruijn graphs [40] and are tailored to assemble short-reads (Figure 1 & Table 1), failing to reproduce the same results with longer reads, e.g., all IDBA (iterative de Bruijn graph assembler) methods available incorporate the stepwise use of increasing k-mer values to build a de Bruijn graph (Peng et al., 2012).

Metagenomic assembly tools employ different algorithmic approaches to address the caveats of assembling microbial communities. Such caveats include uneven sequencing depth and coverage, sequencing errors, repeat structures and strain mixtures, all of which contribute to the development of error prone and fragmented contigs [39, 41]. The extension of the IDBA algorithms IDBA-UD [42] is used to assemble *de novo* paired-end reads of uneven coverage, and accounts for errors in metagenomic data through multiple approaches. Approaches include a flexible cut-off value for removing erroneous contigs that is determined by the sequence depth of adjacent contigs. Errors are further removed via contig alignment. Low depth repeat patterns are resolved using the dual ends of paired reads. Paired-end reads that are unaligned on one end but are uniquely aligned to contigs of high confidence are identified and grouped. Local assembly is then performed on the unaligned end to resolve repeat structures (Peng et al., 2012). MetaSPAdes [41] attempts to reconstruct species level consensus sequences using coverage information, extending the

sequence based on the highest coverage values of approaching edges (k-mers). Furthermore, coverage ratios of all adjacent edges are examined. Those of a low coverage ratio are detached from their source vertex (k-1-mers), but the information is retained to enable the construction of strain contigs. MetaSPAdes then uses these strain contigs to influence its prediction of hypothetical reads, using them to resolve repeat structures [41].

1.2.3.2 Metagenomic binning

The assembly methods discussed above typically achieve contig-level resolution only, because of multiple limiting factors. These factors include repetitive genomic sequences, strain level variation, sequencing errors and low coverage of sequences due to technological limitations, all of which contribute to fragmentation after the assembly process. Fragmented contigs that fall short of chromosome level resolution are often insufficient representations of microbial populations, and can hamper insights into the physiology of microbial communities [43]. Thus, additional bioinformatic tools such as binners (Figure 1 & Table 1) are required to further profile the informational content of reads [44]. Metagenomic binning involves classifying reads and/or contigs of interest into separate groups referred to as OTU's (operational taxonomic units), e.g., genera or species-level clusters. Binning methods are typically used to recover Metagenomic-Assembled Genomes (MAGs), which are fragmented draft genomes that can be used for a diverse range of analyses, including taxonomic assessment that extends to unknown species, and functional comparisons/associations. Binning algorithms are separated into two stages, i.e., read extraction and assignment. Reads/contigs are typically extracted based on similarity in sequence composition/features, e.g., oligonucleotide frequency and abundance profiles, all of which display taxon specific patterns [45]. A typical method to assess oligonucleotide frequency is to determine the tetra-nucleotide frequency (TNF) content of reads, which is employed in metagenomics assemblers such as MetaBAT (Metagenome Binning with Abundance and Tetra-nucleotide frequencies) [46] and MetaBAT2 [47]. Tetra-nucleotide frequency is the frequency at which a set of unique k-mers, four base pairs in length, e.g., AGTC, appear in the reads to be binned. Tetra-nucleotide frequencies provide an insight into phylogeny as closely related genomes can often contain a similar TNF content [48].

CONCOCT [44] is a binning tool that groups contigs based on their abundances and k-mer frequency. Fragmented contigs form two vectors, referred to as the coverage vector and the composition vector. The coverage vector is the amount of reads that are available for a given base in the genomic sequence, while the composition vector is a concatenation of k-mer frequencies of any selected length and their respective complement [44]. As described previously, the extraction strategies used by binning tools rely on multiple parameters. While most tools share similar extraction strategies, the algorithms employed at the assignment stage distinguish the binning tools [45]. Supervised and unsupervised binning methods are examples of widely used assignment techniques in metagenomics. Supervised methods compare reads or contigs to reference databases and their accuracy is influenced by the completeness of the respective database. Unsupervised methods use machine learning algorithms to bin read or contigs based on abundance profiles or sequence composition [49, 50]. MaxBin 2 [50] performs binning using an expectation-maximization algorithm, which assigns scaffolds to their respective bins based on median values of single-copy marker genes and probabilities scores [50]. Genome binning can be performed before or after assembly and, regardless of the stage, it improves downstream analysis by providing more information from sequencing reads, which can be summed to calculate a taxonomic profile.

1.2.3.3 Downstream analysis of MAGs

The application of genome binning methods often results in the construction of MAGs with varying levels of completion and contamination. These metrics are often not provided by the binning tool and other downstream analysis tools such as CheckM [51] must be applied. CheckM examines the single-copy genes including duplicates of MAGs to determine their number. Completion is reported based on the number of single-copy genes compared to the expected number and contamination is determined by the percentage of duplicates [51]. Good quality MAGs represent composite genomes containing the majority of the total gene content of a representative species [46, 52]. GOLD (Genomes Online Database) contains 18,945 MAGs (April 2022), but this is only a fraction of the total of >200,000 MAGs that are publicly available [53]. High quality MAGs are close approximations of individual genomes and provide a comprehensive data source for applications such as comparative genomics [54] and reference data [55]. Analysis of MAGs is the final step

in assembly-based workflows and the beginning of biology-based analysis. Phylogenetic placement involves surveying the gene content of MAGs to detect taxonomically informative marker genes [53]. GTDB-Tk [56] initially classifies the domain of MAGs by comparing the gene content of MAGs predicted using Prodigal (Prokaryotic Dynamic Programming Genefinding Algorithm) [57] to a reference set of bacterial and 122 archaeal marker genes using HMMER (Eddy, 2011). The domain with the highest number of marker genes is chosen and the reference marker gene sets are aligned again using HMMER. The output of this pipeline is a concatenated MSA used for phylogenetic placement by pplacer [56, 58].

Annotation of MAGs typically involves identifying all coding genes of a MAG and deducing their possible biological function through alignment with functionally annotated reference sequences and or translated homology searches. Multiple webservers and standalone tools exist specifically for the analysis of MAGs [59]. Prokka [60] relies on established prediction tools and databases to perform genome annotation. Prediction-based tools such as Prodigal compare genome queries against reference genomes at the nucleotide level [57]. Through this search, Prokka can acquire positional information about contig bound features, i.e., coding sequences (CDs), and acquire their protein output. Gene annotation is performed by a translated search against multiple databases in a hierarchical fashion. The following databases are queried sequentially, a user selectable database of annotated proteins (optional), UniProt [61], RefSeq, Pfam [62] and other hidden Markov model databases. Genomic features in question are assigned an identity by the first accurate match between the query sequence and one of the mentioned databases, if a match occurs further searches are not performed [60].

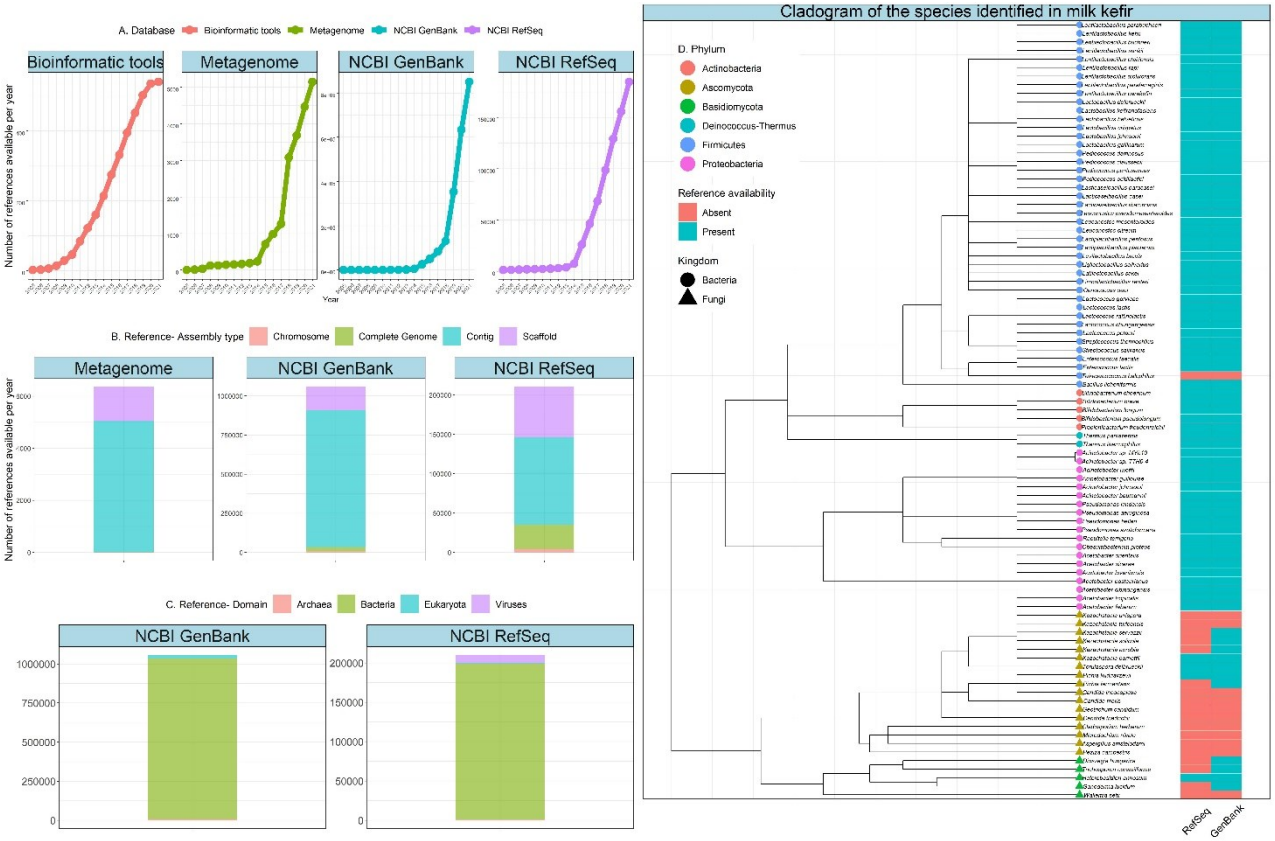


Figure 2: Annual growth and availability in reference data available for the interpretation of metagenomic data including reference genomes in the NCBI database (https://www.ncbi.nlm.nih.gov/). **(A.)** Annual growth (from 2005 to 2021) in reviewed research papers introducing novel and improved tools designed for the interpretation of metagenomic data, metagenomes available in the NCBI GenBank database and reference genomes in the NCBI GenBank and NCBI RefSeq [63, 64]; **(B.)** Available Metagenome, NCBI Genbank and NCBI Refseq reference data separated by assembly level. **(C.)** Available Metagenome, NCBI Genbank and NCBI Refseq reference data separated by domain. **(D.)** Cladogram of bacterial and fungal species found in milk kefir, species information obtained from [65]. Tip labels represent the species detected, e.g., *Lentilactobacillus parabuchneri*, tip colour represents the phylum, and shape represents the kingdom of the species. Presence/absence heatmap displays the availability of reference genomes for each of the listed species in both NCBI RefSeq and NCBI GenBank databases. Metagenome, Refseq and Genbank derived data was generated and analysed in R using the Biomartr package [66]. Visualisation was performed using the ggplot2 package [67] and ggtree package [68].

1.2.4 Advantages and disadvantages of the short read and assembly-based approaches

Current bioinformatic approaches fail to achieve 100% accuracy and have several limitations and flaws that are difficult to address. Given the general short-read nature of WMGS, short read-based approaches commonly utilise sequence similarity and alignment length criteria separately or in parallel, which can determine if reads are evenly distributed across a reference genome/marker gene or are recruited to uninformative subsections of the genome, and represent a false positive result. Such requirements ensure

accuracy and limit the quantity of false positives but restrict the sensitivity of the tool and as such only features of high confidence are identified. Given this, short read -based approaches excel at identifying well documented and highly conserved features such as ARGs but are limited when profiling novel features sharing little resemblance to those contained in reference databases [69, 70]. For example, it was noted that the short-read marker gene-based classification tool MetaPhlAn2 failed to accurately profile metagenomic reads belonging to the *Brettanomyces* genus due to a complete absence of *Brettanomyces* reference marker genes in MetaPhlAn2. Reference availability remains an issue in the updated versions of the tool (April 2022) [71]. Furthermore, marker genes make up only a small proportion of a microbial genome and as such have limited profiling capabilities if a species is present at low abundance compared to host and environmental DNA [26]. Similar issues are also reported for short read k-mer based approaches, which can only compare query sequences to known k-mers of a reference database. K-mer based approaches such as Kraken2 are prone to false positive results generated by closely related genomes with similar k-mer content to true identifications [31]

Assembly-based approaches achieve more satisfactory results when profiling novel environments, allowing for the construction of unclassified composite genomes. Assembly-based approaches are useful for the analysis of viral metagenomes (viromes), in particular viral prophage, which are dominated by poorly annotated sequences with few reference genomes available (54,352 in the Genomes - NCBI Datasets, April 2022 & Figure 2) [72]. Nonetheless, assembly-based approaches produce incomplete draft genomes containing contaminated sequences that hinder downstream binning applications. Constructed assemblies are further limited by sample type, strain-level diversity, population heterogeneity, sequence repeats and a requirement for an average 100-fold coverage difference, compared to short-read approaches to detect low abundance species [73, 74]. These limitations cause inconsistencies in results, which vary according to the tool used and restrict comparisons between studies. Machine learning allows for moderate constraints in cut off values, increasing the sensitivity of such tools. However, the implementation of machine learning classifiers is hampered by the availability of suitable features, e.g., most reference AR data is derived from clinical bacterial isolates and represent a poor database for training machine learning methods [75].

Assembly-based approaches can often only produce complete genomes for the most dominant species, producing low resolution fragmented assemblies for lower abundant microbial taxa and fail to capture all the functional diversity of a microbiota; which would require the assembly and functional annotation of hundreds or thousands of microorganisms in complex communities such as the ocean. While capturing a complete functional and compositional profile of a complex environment can be a computationally challenging process, certain studies may wish to focus on certain aspects of an environment, such as the reconstruction of cholesterol-lowering genes, in such cases a targeted assembly based approach may be employed guided by a reference database [76].

Table 2: Description of short-read based and metagenome assembly-based approaches available for the interpretation of metagenomic data. Details include the limitations and advantages of both approaches.

Approach	Description	Limitations	Advantages
Short read-based	• Short-read taxonomic classification tools function by aligning or mapping sequences to informative subsections of meta(genomic) data. Tools evaluate coverage breadth and depth to determine if reads are evenly distributed across a reference genome/marker gene or are recruited to uninformative subsections of the genome, and represent a false positive result. Such classification methods rely on fast lookup algorithms to handle the enormous data sets generated by next-generation sequencing.	• Requires reference data for classification • Limited applications when profiling novel features • Limited application in viral metagenomics due to an absence of a universal barcode gene	• Enables the profiling a wide variety of bacteria, eukaryotes and some viruses • Robust analysis of the abundant microbial features in a microbiome • Compositional and functional profiles of the microbiome, where each feature is described in terms of relative abundance • Lower resource requirements
Metagenome assembly-based	• Assembly methods merge the consecutive k-mer content of metagenomic reads, based on their overlapping sequence similarity, into single contiguous sequences (contigs) of the shortest possible length. Metagenomic assembly methods can reconstruct large sections of the genomes of some species in a microbial community, if the sequencing depth is sufficient.	• Requires high read count • Computationally challenging • Produces fragmented draft genomes • Incomplete picture of the compositional and functional diversity of a microbiome • Software is only available for Linux systems	• Construction of unclassified composite genomes • Provides reference material including viruses to genomic databases • In-depth functional genome analyses, including the construction of genome scale metabolic models

1.2.5 Specialised bioinformatics tools

Most bioinformatic tools available report on a broad range of taxonomic and or functional features detected in a metagenome (Figure 1 & Table 1), which often meet the needs of the researcher. However, in some cases, the research question is too specific and specialised bioinformatic tools are required to interpret the data. The proliferation of genomic and molecular information (Figure 2) enables the profiling of more specific compositional and functional features, which have been restricted in the past due to a lack of appropriate databases that can unify curated data, sequencing coverage requirements and appropriate algorithms. For example, the profiling of functional features such as carbohydrate-active enzymes, requires databases that reflect the bio-curation efforts of researchers [77]. Such specialised bioinformatic tools are becoming available and a number of assembly-based and short read approaches can be used to predict specific compositional elements and/or functional traits such as predominate strains, viral and or pathogenic taxa and antibiotic resistance genes (ARGs). Other specialised tools exist to profile functional features, such as bacteriocin and *CRISPR-Cas* systems, which were not included in this review, but are often included in pipelines designed for the downstream analysis of MAGs [78].

1.2.5.1 Strain level analysis

Strain profiling represents the pinnacle of metagenomics analysis. A comprehensive understanding of strain variants is needed to address ambiguities arising from substantial genetic variation that strains of the same species can possess [79]. For example different strains of *Escherichia coli* have vastly different impacts on host health, with some strains displaying beneficial probiotic properties and others being classified as pathogenic [80]. Ambiguities with respect to species level resolution impact insights at the functional level and can prevent association with different host/microbial phenotypes, limiting the application of metagenomics in clinical and environmental settings. Several tools are available for profiling microbial communities at higher genomic resolution and represent a mixture of short read and assembly-based approaches. Available tools attempt to introduce methods that can overcome the challenges imposed by

genetic variation within a population, e.g., data complexities arising from the presence of multiple polymorphic sites at varying frequencies in genes/genomes. Tools often distinguish between strains using compositional features for whole genomes or marker genes, such as single-nucleotide variants (SNVs), overall similarity to strain-reference genomes or shared gene composition [81]. StrainPhlAn [29] examines the nucleotide differences of consensus marker genes obtained using MetaPhlAn3 [37], while PanPhlAn (pangenome-based phylogenomic analysis) [29] utilises gene family coverage information to compute species unique gene presence/absence profiles. Abundant gene families of similar coverage are grouped together and are assumed to be from the same genome. Comparisons of gene presence/absence between reference genomes and reference strains allows for the identification of novel or existing strains, and their genetic repertoire, which is used to infer functionality [29, 82].

1.2.5.2 Viral detection

Viruses that use microorganisms as hosts for viral replication are universally present in biomes, including fermented foods [83, 84]. In such environments, viruses can interact with residential microbes, contributing to lytic infections and horizontal gene transfer, influencing their composition and genetic makeup, respectively [85]. Changes driven by viruses can pose a distinct threat to the production of certain fermented foods, which rely on the viability and functionality of the microorganisms involved. Viruses can influence the overall functionality of the fermenting ecosystem, affecting the sensory and safety properties of the final product [86, 87]. Despite the recognized importance of bacteriophages (phage, i.e., viruses that target bacteria) in microbial ecosystems, further research into phage biology, diversity, and interactions is needed to mitigate the negative impact that phage may have in these environments. Viral discovery has increased rapidly in recent years due to technological advancements in genomics and metagenomics (Figure 2). Metagenomic datasets capture viral sequences contained in actively infected cells and their genomes, as well as those located outside of the cellular environment, and as such represent a unique opportunity to investigate phage biology, diversity, and interactions with host microorganisms [88, 89]. However, despite the widespread acceptance of the utility of viral metagenomics, robust bioinformatic tools are still needed to address the complexity associated with viral profiling that arises due to the intra-species genetic diversity,

short size, and absence of marker sequences in viral features [90]. Virfinder [89] employs a k-mer-based machine learning approach to identify viral sequences in metagenomic data. Viruses often contain similarities in their k-mer profile to other viral sequences and its host, compared to non-host species. By employing a k-mer-based approach, the tool benefits from improved prediction accuracy, compared to alignment-based techniques and is more likely to characterise shorter sequences and those lacking reference genomes. VirSorter [91] is designed to detect viral elements in high-throughput sequencing datasets using an iterative process. VirSorter examines input sequences using a sliding window plot providing a quantitative measure for each sequence of one of the following features; presence of viral hallmark genes, enrichment of viral-like or non-*Caudovirales* genes, uncategorised and short-reads, as well as depletions in PFAM-associated genes and strand switching. These preselected features allow for the identification of both viral-specific, viral-like and non-viral features in metagenomics data that are leveraged in downstream probability calculations [92].

1.2.5.3 Antimicrobial resistance

Antimicrobial resistance (AMR) refers to genetic adaptations in microorganisms that confer resilience to an antimicrobial agent. Adaptations commonly occur by horizontal gene transfer or chromosomal mutation, and their incidence rate is heavily influenced by antibiotic exposure [93, 94]. The widespread distribution of ARGs , existing AMR mechanisms to all known treatments and limited availability of novel antibiotics poses a threat to public health and medical practices, for example by impacting the safety of surgical procedures [93, 95]. Many factors have contributed to the spread of AMR, such as excessive antibiotic usage, limited antibiotic discovery pipelines and ineffective detection methods which prevent appropriate treatments [96, 97]. Current detection methods such as antimicrobial susceptibility testing (AST) and PCR, are laborious, time-consuming and fail to capture the complete resistome (the entire AMR gene collection of a metagenome) [98, 99]. Metagenomics represents a potential culture-free surveillance and monitoring application, allowing for the *in silico* detection and profiling of AMR mechanisms present in a resistome. Additionally, metagenomics enables comparisons of resistomes within and between different target populations, revealing patterns of transmission [100]. However, it should be noted that the presence of a

putative AMR gene does not always correspond with resistance and the corollary is also true as resistance may be conferred by a gene not already assigned as a resistance determinant.

Most bioinformatic approaches examining AMR elements in environmental samples use alignment-based search tools such as BLAST and Bowtie2 to compare raw reads or predicted open reading frames (ORFs) from assemblies against existing ARG databases, which are subsequently assigned an annotation from the "best hits" obtained. Several reference databases such as CARD (Comprehensive Antibiotic Resistance Database) [101], ARDB [102] and ResFinder [103] are commonly used for such alignments. ARG-ANNOT (Antibiotic Resistance Gene-ANNOTation) [104] is an example of a bioinformatic tool that employs a local alignment-methodology to predict and identify AR features in bacteria [104]. Machine-learning algorithms provide an alternative approach to detect AR elements by training a prediction model on a complex dataset. The dataset provides a representative sample of similarity distributions found in nature of features in ARG reference databases. DeepARGdeepARG-LS and deepARG-SS [105] are deep learning prediction models used to detect AR features in raw reads and predicted genes, respectively, from metagenomic data. Both models were created using a machine learning methodology. Models were trained and evaluated using a distance distribution. The distribution represents sequence similarity values between hypothetical ARGs provided by UNIPROT and known ARGs taken from the ARGminer, CARD, ARDB, UniProt and SARG databases [105].

1.2.5.4 Genome scale metabolic models

Functional profiling tools have been discussed, which predict metabolic genes/pathways present in a MAG (Figure 1 & Table 1). Such tools provide a description of metabolic features, but in isolation cannot provide a complete understanding of the phenotypic properties of the studied organism. Such genome assessments can however inform the creation of genome-scale metabolic models (GEMs). These GEMs are mathematical constructs that predict, for the organism studied, the relationship between metabolic genes, enzymes, metabolites and metabolic reactions, storing the metabolic features, set of reactions between them and the association biomass equations in separate metrics [106]. GEMs provide a representation of the

metabolic capabilities that the organism may be capable of based on their repertoire of metabolic genes and/or experimentally obtained information. Such knowledge can advance culturing approaches by providing insights into the supplemented nutrients required by an organism to sustain growth in different substrates/mediums. GEMs can further highlight knowledge gaps in the metabolism of the species in question [107]. Organism-specific GEMs can be combined with other GEMs recovered from the same environment and expanded to community level models to provide insights into cross feeding and/or nutrient competition between members of the same community [108, 109]. The construction of high quality GEMs can be a laborious process requiring extensive manual curation, including reviewing each metabolic gene, metabolite and pathway association, which is further complicated by the diverse functional roles of enzymes that can take part in multiple reactions, or alternatively the requirement of multiple enzymes in certain reactions and missing reactions [110]. Missing reactions are often addressed by gap filling, which involves inferring missing reactions based on the metabolites present in the metabolic network [111]. The manual curation step has in the past been reported to take up to two years to complete [112]. A number of steps are available that can help to automate the construction of GEMs, such as literature mining and comparisons to reference templates housed in repositories such as BioModels [113] and BiGG [114]. A number of bioinformatic tools are available for the automated construction of GEMs that can be scaled to metagenomic applications [115], such as CarveMe [116] and Metage2Metabo [117]. CarveMe introduces an automated top-down approach to genome scale metabolic modelling. A high quality reference model is used as a template and tailored towards the organism of interest through genome annotations, which are used to inform the removal of features from the template models [116].

1.2.6 Limitations of current comparative performance methods for bioinformatic tools

As described above, a range of tools can be used and combined to profile a metagenome, all of which employ different reference sources and classification algorithms, contributing to performance differences. Walsh et al., 2018 reported that short read bioinformatic approaches can have a statistically significant impact on the results of taxonomic and functional profiling, highlighting the importance of questioning and validating when possible the choice of bioinformatic tools [21].

While bioinformatic tools and algorithms are rapidly improving, it is often difficult to identify which tools are optimal for specific applications due to the fast pace development of these tools, the absence of a uniform evaluation strategy and an appropriate comparison model, such as a real metagenomic dataset derived from multiple sample types [49, 59, 118]. Most published studies introducing novel or improved bioinformatic tools implement comparative performance methods for the assessment of the proposed tool against other established bioinformatic tools. Such studies employ a diverse range of evaluation techniques, performance metrics and sample datasets preventing the generation of uniform results that can be compared across studies [119].

In 2014, the critical assessment of metagenome interpretation (CAMI) was proposed that, through collaboration with the metagenomic community, was intended to develop uniform and reproducible datasets, standards, and strategies for the evaluation of bioinformatic applications [120]. A number of bioinformatic tools have been developed through the CAMI initiative to evaluate the performance of short-read [121], assembly [122] and binning [123] tools applied to metagenomic data. These tools rely on simulated shotgun metagenomic datasets generated using the metagenomic simulator CAMISIM [124]. Through CAMISIM, users can create custom made synthetic microbial communities built from input genomes or taxonomic profiles or recreate existing benchmark datasets for the evaluation of different bioinformatic tools. However, the lack of a detailed user manual hinders the applicability of the former approach. Users can apply these stimulated datasets to the tools, reviewed extensively in Meyer *et al*., 2021 to identify tools with optimal performance for the benchmarked dataset. An alternative approach is to

sequence DNA mixtures of known origin. Such an approach allows for technical variation and biases introduced during data generation and is particularly suitable for the simpler microbial communities found in fermented food [118]. Recent publications [119, 121, 123-125] utilised the CAMI benchmarking platform to evaluate the performance of bioinformatic tools used in metagenomic analysis. Unfortunately, these publications failed to account for either parameter settings or variability in sample type, and assessed only a select number of established tools [126]. Whole metagenome shotgun sequencing based profiling is a relatively novel field with many improvements still needed, such as the need for advancements in tools supporting long read and hybrid sequencing. Currently the key problem faced by bioinformatic analysis is the availability of reference data (Figure 2), the selection and integration of multiple tools and parameter settings for optimal classification accuracy and recall. The utilisation of multiple tools of the same category, or the development of tools that use integrated results from different bioinformatic applications as input data. e.g., BinSanity [127] MetaMeta [128], MetaWRAP [129] and bioBakery workflows [130] may help improve the accuracy of analysis [21].

Section 1.3: Applications of bioinformatic approaches to fermented food microbiomes.

1.3.1 Research into the fermentation process

Many fermentation processes are uncharacterised or not fully understood and, as a result, the primary application of bioinformatics tools is often to enhance knowledge about these systems. Studies employing these tools will typically report on population composition at the genus or species level including putative new species and, occasionally, will reach strain level resolution (Figure 3). To date, compositional profiles of many fermented foods have been acquired through bioinformatic analysis of kefir, kimchi, wine, beer, meat, and cocoa metagenomic datasets [7, 71, 131-134]. Through functional analysis, it is possible to detect and profile potential functional and metabolic features present in the fermentation process and assign them to a contributing species [22]. Many studies are applying metagenomics tools beyond their feature profiles to better understand diverse aspects of fermented food microbiomes such as uncovering the functional

features driving co-occurrence and succession patterns. Here we refer to a subset of studies as examples of the novel insights that bioinformatic tools and pipelines can provide into fermented food microbiomes (Figure 3).

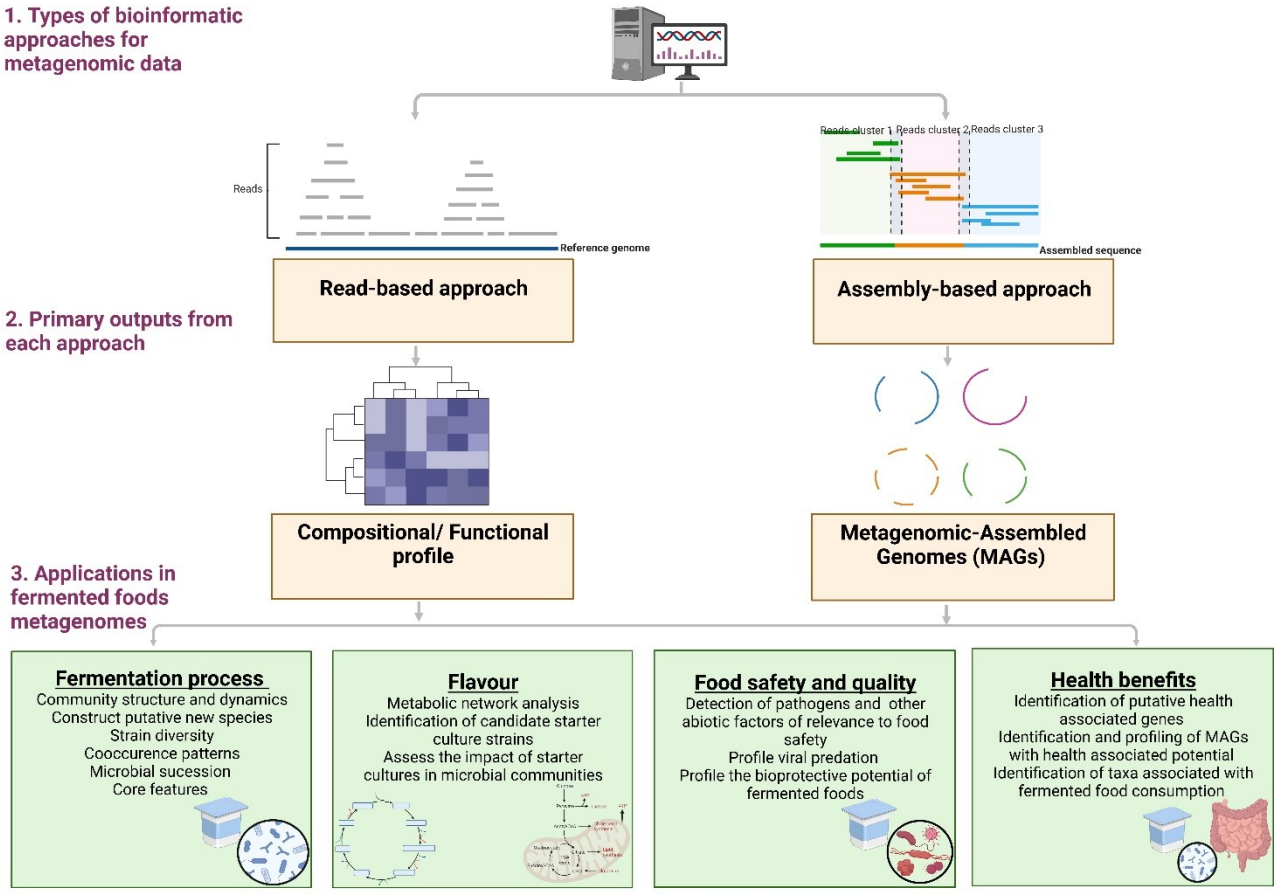


Figure 3: Conceptual overview of the primary steps, outputs and applications associated with the application of short read and assembly based approaches to fermented food metagenomic data.

1.3.2 Community structure and dynamics

Firstly, we discuss a representative case study that highlights the results that can be expected from applying both short-read and assembly-based approaches to examine a fermentation ecosystem. Through a combination of short-read alignment and assembly based approaches Verce, De Vuyst and Weckx, 2019, examined the microbiota of water kefir at 24 and 72 hours of fermentation. Multiple short-read

compositional tools and databases were used in combination to produce a taxonomic profile. The tools Kraken, BLASTn, BLASTx, Kaiju [135] and metagenomic recruitment plotting were used to perform compositional classification using a k-mer or alignment strategy, while MetaPhlAn3, rRNA selector [136] and ITSx [137] facilitated classification using a marker genes strategy. Each tool varied slightly in taxa assignment, providing differing species level results but agreeing at the genus level detecting *Lactobacillus*, *Oenococcus*, *Bifidobacterium*, *Saccharomyces* and *Brettanomyces* as the dominant bacteria and yeast genera. Metagenomic recruitment plotting was the preferred compositional approach to further profile the microbial community to the species level, resulting in the identification of a novel *Oenococcus* species, *Candidatus Oenococcus aquikefiri*. In addition, assembly-based methods were employed to acquire water kefir-based MAGs, including the novel species *Candidatus O. aquikefiri*. The workflow involved contig assembly using MEGAHIT [138], followed by annotation using Prokka, which informed further annotation by HMMER using various databases specific for carbohydrate-active enzymes. Binning was carried out by CONCOCT, and the resulting contigs were mapped using BWA-MEM [139] to reference genomes, selected by the results of short-read taxonomic profiling. The genes identified were linked to roles in amino acid, vitamin, and cofactor biosynthesis and were further assigned to their taxonomic source. For example, genes for glycerol biosynthesis were located on assembled contigs of the species *Saccharomyces cerevisiae*. Identified genes were also linked to metabolites with roles in carbohydrate, pyruvate, citrate, and malate metabolism. For example, the protein coding genes known to encode for glycerol kinase and glycerol-3-phosphate dehydrogenase were located on contigs of the species *Lactobacillus*, indicating their potential to use glycerol as part of their metabolism and suggesting a possible interaction with *S. cerevisiae* [71]. Overall, the interpretations made from the bioinformatic methods applied in this study advance knowledge into the microbial taxa that can be found in water kefir, their functional contributions to the fermentation ecosystem and suggested the potential of interspecies interactions during the fermentation process.

1.3.2.1 Microbial succession

It is difficult to characterise the microbial ecosystems associated with a fermentation process using a single

fermented sample due to the occurrence of temporal changes. Shifts in the abundance, diversity and functionality of microbial communities often occur along the fermentation process. Bioinformatic tools in combination with a longitudinal study design involving the profiling of samples representing different time points of a fermentation, can provide insights into how microbial communities change and adapt over the fermentation period. Such shifts in community structure are referred to as microbial succession [140]. Bioinformatic tools have been used in this manner in a number of publications focusing on the fermented foods kefir, soy sauce and cheese rinds to name a few [7, 71, 140, 141]. Walsh *et al*., 2016 identified a clear pattern of microbial succession in kefir throughout the 24-hour fermentation period using the taxonomic profiling tools, Kraken and MetaPhlAn2. At 8 hour, *Lactobacillus kefiranofaciens* was the dominant microbial resident in kefir, but as the fermentation progressed its proportions decreased and the relative abundance of *Leuconostoc mesenteroides* increased considerably becoming the dominant bacterial resident at 24 hour. Further investigation using the functional profiling tool HUMAnN2, revealed the presence of aromatic amino acid biosynthesis genes in *Leuc. mesenteroides* and their absence in *L. kefiranofaciens,* suggesting a potential mechanism driving microbial succession [7]. Insights into microbial succession are particularly important to assess the distribution of inoculated species and strains throughout fermentation. Bertuzzi *et al.*, 2018 profiled microbial communities associated with surface ripened cheese and their corresponding smear cultures using the taxonomic profiling tool, Kaiju and strain profiling tool PanPhlAn and performed functional analysis via Super-Focus. Through this bioinformatic workflow they identified patterns of microbial succession in two distinct surface ripened cheeses at the species level, which were further profiled to provide insights into the strain level distributions. Such analysis revealed the present/absence of inoculated smear cultures strains during the ripening process [142].

1.3.2.2 Core feature detection

Bioinformatic analysis of metagenomics datasets derived from a fermented sample at different time points as opposed to a single representative sample can reveal core functions present throughout the fermentation process. Sulaiman *et al*., 2014 used metagenome assembly-based approaches, specifically CLC

Genomic Workbench to assemble contigs from metagenomic reads, derived from different stages of fermentation from zero to 6 months. Prodigal and AUGUSTUS [143] were used to predict ORFS from prokaryotic and eukaryotic contigs respectively. Contigs were used as references to map unassembled reads, to determine gene abundance values. Predicted genes were then functionally classified using the KEGG classification workflow provided by MEGAN [22]. Through this bioinformatic approach, they observed core functional features, in soy sauce fermentation that were consistently recovered after 6 months of fermentation [141].

3.2.3 Evaluate the effectiveness of laboratory-based techniques specific for fermented foods

Bioinformatic tools can be used to assess and compare certain laboratory-based techniques. Dugat-Bony *et al.*, 2020 evaluated the effectiveness of four viral extraction methods designed specifically for cheese environments. Viral DNA extracted from treated samples of Epoisses cheese, and vesicle DNA were selected for WMGS-based analysis. Sequence reads were processed by a global assembly-based approach, using SPAdes [144] to produce contigs from paired-end reads. VirSorter compared quality assessed contigs against the Viromes database and a catalogue of non-redundant genes predicted from viral metagenomes, identifying contigs likely to be of viral origin. A subset of the contigs were selected based on sequence coverage, viral origin, and circular characteristics, with contigs matching one or more of the selection criteria included in the finished contig dataset. The final contig dataset represents the first Epoisses cheese surface virome acquired for a cheese environment. Contigs were characterised using PHASTER [145] and an alignment-based approach involving Blast searches which queried contig sequences against all nucleotide and viral specific information contained in the NCBI database. The characterisation tools separated the contigs into four distinct categories, namely putative phage encoding contigs, unclassified contigs, putative and plasmid-derived contigs. An abundance table was acquired by mapping sequencing reads from each of the extraction methods to the putatively viral contigs through Bowtie2. The abundance table then informed statistical testing using Spearman correlations, to evaluate the impact of the different extraction protocols on the composition of the virome. Reads from each extraction method were further examined by

SortMeRNa v2.0 (Kopylova et al., 2012) to detect ribosomal DNAs, which served as indicators of microbial contamination. The computational workflow that examined the viral profile and microbial contamination levels identified the best performing extraction method, but further wet lab techniques are required to validate its effectiveness [146].

1.3.3 Flavour

The metabolism of microbial communities in fermented foods and beverages play an essential role in the development of flavour, a combination of aroma and taste sensations. An in-depth knowledge of the metabolic activity and network of microbial communities during fermentation is necessary to understand, at a systems level, how microbes contribute to the organoleptic properties of food.

Bioinformatic approaches can provide a systems level understanding and be employed to provide a theoretical basis for the improvement of fermentation processes without the use of detailed genome scale metabolic models [147], which is of particular importance for industrial applications with increasing quality requirements. Potential avenues of improvement include the selection of starter cultures, appropriate nutrients, environmental conditions and cross-contamination events with production environments, all of which can contribute to the sensory characteristics of fermented foods. The identification of such improvements depends on the ability to characterise the microbial community in question and is thus applicable to bioinformatic analysis where careful consideration into the functional characteristics of microbial strains is possible.

1.3.3.1 Metabolic network analysis

Short-read and assembly profiling tools can be used to perform metabolic network analysis as demonstrated by Wu *et al*., 2017. The microbial metabolic network of vinegar pei, a starter culture mixture used to produce Zhenjiang vinegar, was reconstructed using multiple bioinformatic tools, detailed below. The short-read tool MetaCV [148] profiled the vinegar samples at the taxonomic and functional level, using the KEGG database for gene assignment. Metavelvet assembled contigs [149] were analysed using FragGeneScan [150] to

detect contig bound ORFs, which were subsequently annotated by BLASTx using the Genbank, NR and KEGG databases. All genes assigned a KO term were mapped to their respective KEGG pathways by the tool KEGGMapper and assigned an EC number [151]. This assembly-based analysis revealed the metabolic pathways and enzymes that are potentially involved in flavour production in cereal vinegar, which were traced to contributing microbial members at the genus level by MetaCV [152].

1.3.3.2 Identification of candidate starter culture strains

Variability during fermentation prevents the production of uniform products with consistent flavour. As such, the production of many fermented foods would benefit from the identification of an appropriate starter culture to minimize variability. Starter cultures are inoculums of selected microbial strains, which are deliberately introduced into a food matrix to ensure a more controlled fermentation process, ideally producing an homogeneous product with enhanced nutritional and sensory qualities [153, 154]. Starter cultures routinely used to produce fermented foods, such as cheese, and can influence cheese-based microbiomes and, in turn, flavour and appearance [142]. Liang *et al*., 2018 identified potential sources of starter cultures from an assembly-based approach that examined industrially produced paocai. Paocai is a dish composed of an assortment of vegetables including cabbage, radish, long beans and peppers, all of which are fermented together in a jar with a paocai brine solution [155]. Metagenomic reads were assembled using IDBA-UD, and ORFs were predicted and translated from the assemblies using Prodigal, followed by clustering based on sequence identity and coverage values using CD-HIT [156]. Clustering was used to construct a non-redundant gene catalog from input sequences obtained in the clustering phase. The gene catalog subsequently informed searches against the NCBI non-redundant (NR) protein sequences and KEGG databases using diamond and GhostKOALA [157], respectively. From this analysis the authors were able to identify the species *Lactobacillus paralimentarius* and *Lactobacillus alimentarius* in Paocai that, on the basis of their abundance and functional roles, were predicted to substantially contribute to a successful fermentation process and could potentially serve as reservoirs for selecting candidate starter culture strains [154].

Bioinformatic tools can compare functional subsystems found in metagenomes or metagenomic contigs to isolated starter cultures or potential candidates to investigate the potential of a particular isolate to be used as a starter. Illeghems, Weckx and De Vuyst, 2015 examined the cocoa bean fermentation process through WMGS of a single representative sample after a 30-hour fermentation. Using the data generated from this analysis, they performed a sequence-based comparison of candidate starter culture strains and Genovo assembled contigs, using EDGAR [158]. This analysis revealed the proportion of CDs shared between the candidate starter cultures and the metagenomic contigs, as well as a description of their functional categories. Based on shared gene content, the potential contribution of candidate starter cultures to the overall fermentation process was revealed, e.g., *Limosilactobacillus fermentum* 222 was 82.7% identical in gene content to the collective gene content of the metagenomic contigs, highlighting its usefulness as a starter culture [159]. Further research into the metabolic activities of computationally identified or inspected strains, and their association with flavour compounds are needed to validate their potential as starter cultures [154].

1.3.3.3 Assess the impact of starter cultures and fermentation parameters in microbial communities

Taxonomic and functional profiling tools can be used to study changes in taxonomic membership and gene content due to a trait/response of interest [35]. Zepeda-Mendoza *et al.*, 2018 employed a short read based, taxonomic and functional pipeline to profile multiple microbial communities in control and strain inoculated samples of Cabernet Sauvignon wine. Profiling was performed to examine how inoculated strains of *Oenococcus oeni* and *Brettanomyces bruxellensis* interact with microbial communities and thus effect the flavour of Cabernet Sauvignon wine. Taxonomic profiling was performed using MG mapper [160]. The functional annotation workflow included contig assembly using IDBA-UD, gene prediction by Prodigal, and KO (KEGG Orthology) assignment through a translated homology search against the KEGG database using BLASTx. Univariate differential abundance testing using pairwise fisher exact tests, highlighted taxa, KO terms and their associated KEGG pathways inferred to be differentially abundant in the strain inoculated samples compared to spontaneously fermented control samples [161].

Bioinformatic approaches can also be employed to detect how additional factors such as raw ingredients, e.g., herbs and spices and manufacturing conditions, influence the microbial communities and its metabolic activities during fermentation and in turn, the organoleptic properties of the fermented product. Bertuzzi *et al.*, 2018 profiled microbial communities associated with surface ripened cheese and their corresponding smear cultures using a previously described pipeline. Through this bioinformatic analysis they observed a correlation between salting the surface of ripening cheese and the increased relative abundance of osmotic stress resistance and the osmoprotectant-related gene families [142]. Link the volatilome profile to contributing taxa and metabolic pathways

Despite the capacity to determine differentially abundant metabolic pathways in fermented foods, it is often unclear which metabolic pathways generate the appropriate volatile organic compounds in the necessary quantity for a successful fermentation process. Bioinformatic profiling coupled with metabolomics can be used to reveal links between meta pathways and fermentation-based volatilomes to detect formation pathways that may be preferentially used and their microbial contributors [159]. Li *et al.*, 2018 employed a combination of metagenomic and metabolomic-based approaches to examine interactions between microbial communities and flavour production in Pu-erh tea. The metabolomic workflow involved volatile organic compounds (VOCs) identification by headspace (HS) solid phase microextraction (SPME) and subsequent analysis by gas chromatography mass spectrometry (GC-MS). VOCs represent a group of evaporated carbon based chemicals emitted from products or processes. Examples of VOCs include aldehydes, ketones, acids, benzene derivatives and hydrocarbons. The metabolomic analysis detected five dominant flavours in the fermentation process of Pu-erh tea, namely methoxy phenolic compounds, theabrownin, alcohol and caravone. The Pu-erh tea microbiome was further examined using metagenomics. An assembly-based approach was taken to characterise the microbial community. Sequence reads were assembled using Soapdenovo [162] and the resulting assemblies were subjected to gene prediction and annotation using MetaGeneMark and DIAMOND, respectively. DIAMOND performed multiple translated searches against the NR, KEGG, eggNOG and CAZy databases. This multi-omic approach allowed for a detailed understanding of the factors associated with the biosynthesis of methoxy-phenolic

compounds; these included genes, enzymes (methyltransferase) and their taxonomic affiliations inferred from detected protein families. In addition to identification, this workflow also monitored the contribution of the identified microorganisms at the genus level to community production of methyltransferase during fermentation by recalculating their relative abundance using the total sum abundance of the enzyme at two distinct timepoints. The biological interactions involved in the production of the 4 other dominant compounds were also examined in this manner, which revealed similar results, extending the knowledge into metabolic pathways that contribute to the dominant flavours of Pu-erh tea [163].

This method can be further applied to strain level analysis. Ferrocino *et al.*, 2018 computationally profiled the microbial communities in both spontaneously fermented and strain inoculated samples of Salame Felino, a type of Italian cured pork sausage. Computational profiling was achieved through taxonomic classification by MetaPhlAn2 and a multistage assembly approach. Reads were assembled using Velvet, followed by gene prediction via MetaGeneMark. Subsequently genes were clustered through Usearch [164] based on sequence identity and alignment length criteria, creating a gene catalogue. The gene catalogue was annotated by querying its content against the NCBI-NR database using mBLASTX. Lastly, meta pathway analysis was performed using MEGAN, which assigned the identified genes to KEGG pathways. VOCs generated during the fermentation process were examined using metabolomics, which involved VOC extraction using HS SPME and analysis by GC/MS. Similarly to Zepeda-Mendoza *et al.*, 2018, the bioinformatic pipeline used by Ferrocino *et al.*, 2018, coupled with statistical analysis (Spearman's rank-order correlation), revealed differences in gene content and metapathways between the Salame samples examined, which, in combination with metabolomics data, could be further linked to expressed volatile compounds. For example, bioinformatic profiling reported a higher abundance of KEGG genes encoding acetate kinase and butanediol dehydrogenase in the strain inoculated samples that correlated with elevated levels of acetic acid and reduced product acceptance. A higher concentration of acetic acid may have contributed to their inferior sensory properties compared to control samples reported during a liking test. Similarly, other sources have reported less fragrant products when employing starter cultures in the fermentation process [141]. Such results highlight the application of this bioinformatic approach in assessing

the potential impact of strains on fermentation mechanisms, which contribute to flavour production [165].

1.3.4 Food safety and quality

Foodborne illness can result in sickness and death, the full impact of which cannot be quantified due to limitations in surveillance methods [166]. There is an associated need for swift and accurate techniques for the detection of foodborne pathogens and spoilage microorganisms. Metagenomic tools can identify pathogenic or spoilage-causing microorganisms, which can be further examined computationally to reveal insights of relevance to food safety. For example, such tools can monitor changes in the diversity or proportion of undesirable microbes in food to anticipate food borne diseases and or microbial spoilage [167, 168]. In terms of exploratory potential, several short read and assembly based bioinformatic tools can detect and trace foodborne pathogens to the strain level [23, 82, 169].

1.3.4.1 Detection of pathogens other abiotic factors of relevance to food safety

Several short read bioinformatic tools were employed in a proof-of-concept study to demonstrate that metagenomic shotgun sequencing can detect pathogenic strains in the dairy product nunu. Initially, metagenomic data was taxonomically profiled using MethPhlAn2, revealing that the majority of samples were dominated by the potentially pathogenic species *Streptococcus infantarius.* It was with some concern that the presence of the species *Escherichia coli* and *Klebsiella pneumoniae* was also detected at varying abundances across the samples. Three short-read bioinformatic tools, MetaMLST [170], PanPhlAn, and StrainPhlAn were used to detect strains of *E. coli* and *K. pneumoniae* present in 10 nunu samples. StrainPhlAn detected several *E. coli* and *K. pneumoniae* marker genes while MetaMLST found an *E. coli* strain in one sample and *K. pneumoniae* strains in seven samples. PanPhlAn identified *E. coli* strains in two nunu samples and several *K. pneumoniae* strains [171]. *E. coli* and *K. pneumoniae* specific marker genes detected by StrainPhlAn were aligned against *E. coli* and *K. pneumoniae* reference genomes, respectively, and the outputs were visualised as phylogenetic trees created using GraPhlAn [172]. Phylogenetic analysis reveal

that an *E. coli* strain detected in one of the nunu samples was closely related to the outbreak associated strain *E. coli* O139:H28 E24377A and shared the same ShET2 enterotoxin-encoding gene. Two *K. pneumoniae* strains showed antibiotic resistance properties and were closely related to *K. pneumoniae* KpQ3, another outbreak associated strain [171].

Other abiotic factors of relevance to food safety include food production practices and food processing environments, which potentially introduce opportunistic microorganisms to food and can contribute to the development of AR [173]. Alexa *et al.*, 2020 employed functional metagenomics to inspect a cheese production chain using a representative recombinant library to identify potential sources of pathogens, ARGs and virulence factors. The library consisted of cloned DNA fragments obtained from cheese production facilities, as well as raw cheese and milk samples. Recombinant clones (9,216) were randomly selected from each sample category and sequenced using a Nextseq 500. Filtered paired-end reads were taxonomically profiled by Kraken2 revealing phylum and species level profiles, interestingly the production environment displayed the highest microbial diversity of all tested samples, containing multiple species of low abundance. Functional annotation involved two alignment methods used specifically to examine the AR potential of the sample. Bowtie2 was used to align reads against the MEGARes database [174] to detect ARGs . BLASTx was employ to detect any gene with quorum quenching activity by performing a translated search against a custom built protein database, detailing all enzymes recorded by LaSarre and Federle, 2013 to inhibit quorum sensing. Detected ARGs were characterised into nine categories, and separated by sample type, which upon comparison displayed a similar resistome profile between samples. A limited number of ARGs could be assigned to their species of origin (11%), but highlighted the potential of *E. coli* and *Lactococcus lactis* as reservoirs for multi-drug resistant genes. Notably raw cheese samples seem to contain a high fraction of *Lla. lactis* with AR potential compared to other samples. The BLASTx based annotation identified homologues to genes with predicted quorum quenching activity for example *QSDH* was reported to have the highest content of homologous genes [175]. Taxonomic profiling assigned most QQ genes to their host species or strains (72.7%) identifying members such as *Lla. lactis*, and *Stenotrophomonas sp*. LM091, which based on their gene content may help regulate the spread of AR in the dairy food chain [175].

1.3.4.2 Profile viral predation

As noted above, viruses are known contributors to fermentation failure despite this, for many fermented foods interactions between viral and host communities are poorly described. Many well-studied fermented foods such as cheese, milk -kefir and water kefir (Figure 2 & Figure S1) lack detailed insights into viral based interactions despite extensive research efforts over the last decade [146]. Colombo *et al*., 2018 performed a shotgun metagenomics analysis on airborne virus-like particles (VLP) in two cheese production plants to assess their aptitude for microbial contamination, which is of relevance to cheese production given the prolonged exposure of certain cheese surfaces to air during ripening [176]. An assembly-based approach was taken involving contig assembly via MIRA and gene prediction of the assemblies using Prodigal. Predicted ORFs were subjected to a BLAST+ search against a custom-built viral database. Further searches, using BLASTn, BLASTp and BLASTx were employed to compare viral and microbial data, using the NCBI NR nucleotide and protein databases, as well as CARD. Through this analysis detected ORFs were linked to taxonomic members of viral and bacterial origin, for example ORFs mapped to *Legionella pneumophila* and the human Papillomavirus, raising some safety concerns. ORFs were further assigned to AR categories based on gene function and antibiotic drug targets. The frequency of ARGs genes was calculated by dividing the representative ARG-related ORFs by the total ORFs content detected. Viral taxonomy was assessed to determine potential interactions with bacteria at the genus level, highlighting that the virome can target both fermenting and non-fermenting bacteria. Interactions were quantified using the same ORF based method employed to determine the frequency of antibiotic resistant genes [177].

1.3.4.3 Profile the bioprotective potential of fermented foods

Bioinformatic tools can reveal functionalities in fermented foods, some of which have a protective effect against spoilage and disease. Features include the occurrence of CRISPRs elements, production of bacteriocins and other antimicrobials, highlighting the bioprotective potential of some foods. [159] detected such functions in a spontaneous cocoa bean fermentation process using the following computational workflow. Metagenomic data representing the fermentation process was assembled using Newbler, CABOG

[178], Genovo [179] and CAMERA. Notably Genova had the best performance statistics in terms of contig length, abundance and predicted functional genes as determined through the gene finding tools Glimmer [180], GeneMarks and Augustus. Genovo assembled contigs were selected for further annotation through GenDB. GenDB detected five bacteriocin encoding genes, six CRISPR elements and reported minimal evidence for the presence of antibiotic resistant genes [159].

1.3.5 Health benefits

Foods with added benefits beyond their nutritional value that possibly contribute to overall health and wellbeing are referred to as functional foods, with many fermented foods falling under this category. The market value of fermented functional foods continues to increase annually, achieving its most rapid growth in recent years [181, 182]. Despite this growth, the further development and marketing of fermented functional foods may be limited by a lack of sufficient research to provide the evidence base required for robust health or functional claims.

To date, numerous studies have examined the health promoting properties of fermented food, with yogurt being examined most extensively [183]. Such studies offer interesting insights into the effects of fermented foods, but often fail to elucidate the mechanisms driving such effects. Metagenomics can be used to profile the gene and metabolic content of microbial communities to provide further evidence of the features driving any observed changes attributed to fermented foods.

1.3.5.1 Identification of putative health associated genes

Bioinformatic applications can be used to provide insights into the effects of fermented foods on human health. For example, fermenting microorganisms can be computationally assessed for the presence or absence of putative health promoting genes [184]. Leech *et al.*, 2020 investigated the microbial component of 58 diverse fermented foods and non-fermented milk samples using shotgun metagenomics. BAGEL [185] compared Prodigal predicted genes to the BAGEL4 bacteriocin database to identify potential bacteriocin producing genes. Only four fermented food samples completely lacked gene clusters associated with

bacteriocin production, with 55 putative bacteriocin-encoding gene clusters detected. Additional putative health-promoting genes were detected from UniRef gene clusters obtained through HUMAnN2 using a list of search terms. Health associated genes were detected and grouped into three categories; host colonisation, intestinal survival and host modulation with a notable difference in the number of detected colonisation genes in the fermented samples as opposed to the unfermented samples [184]. Such insights can help to predict if fermenting microorganisms contribute to the health promoting properties of fermented foods.

1.3.5.2 Identify and profile MAGs with health associated potential

Pasolli *et al.*, 2020 compared MAGS derived from fermented milk products (n=303) to those of human (n=9,445) and reference genomes, to examine the colonisation potential of lactic acid bacteria (LAB) found in dairy based fermented foods. Metagenomes were taxonomically profiled using MetaPhlAn2, and their compositions were compared to identify LAB species found in both sets of samples. While their viabilty in the gut is uncertain *Streptococcus thermophilus* and *Lla. lactis* were the most prevalent species in the gut, and an additional 37-foodborne species were also detected at varying prevalence. Most overlapping LAB species shared a relative abundance of <2% with a small number of exceptions. Metadata associated with the human samples were used to identify correlations between lifestyle, age and geography and the prevalence of food associated LAB. Strong correlations were detected between the three categories of metadata and the LAB species. MAGs were then reconstructed by reassembling sequencing reads using a combination of IDBA-UD and metaSPAdes and binning by MetaBat2. Reconstructed MAGs and GenBank reference genomes were integrated together and clustered into species level bins based on sequence similarity values determined using MASH. Species-level bins showed that 52.4% of fermented food MAGs clustered with species detected in the human gut. Species-level bins were further visualised via phylogenetic trees and multidimensional scaling plots. PhyloPhlAn3 and GraPhlAn built phylogenies, while FastANI was employed for the multidimensional scaling plots. The phylogenies enabled a genomic comparison of the species detected in different samples. Notably, the phylogenetic analysis of *S. thermophilus* strains revealed

no unique subclades between the sample types, providing evidence that gut ecosystems may acquire *S. thermophilus* through the diet. Other strains such as *Lactobacillus delbrueckii* subsp*. bulgaricus* and *Lactobacillus delbrueckii* subsp*. lactis* also displayed a similar pattern when phylogenetically assessed [186].

1.3.5.3 Identify taxa associated with fermented food consumption

Metagenomic analysis can assess the impact of fermented food consumption in shaping the composition of host microbiomes. A recent intervention study examined the impact of fermented food consumption on the composition of the gut microbiome [183]. Participants of the longitudinal study also took part in the American Gut Project (https://microsetta.ucsd.edu/), and were selected based on their answer to a fermented plant question from the American gut survey. Participants were separated into consumers of fermented foods and non-consumers based on their answer to the aforementioned survey [187]. A further fermented food questionnaire was given to the participants, which captured the types of food consumed over the 4-week intervention study. Foods include fermented dairy products such as kefir and yogurt, fermented grains such as miso and fermented vegetables such as sauerkraut and kimchi. Stool samples were collected from 115 participants each week for a period of four weeks. All stool samples collected were subjected to 16S rDNA sequencing and metabolomic analysis with a subset of samples derived from the same single time point chosen for shotgun sequencing. SHOGUN [188] was used in combination with bowtie2 for taxonomic annotations and quality filtering, NCBI RefSeq [189] was used as a reference database. Species level taxonomic identifications with an abundance >0.01 in at least one sample were reported in a biological Observation Matrix (BIOM) format taxa table. Firstly, a literature research was conducted to identify taxa that were commonly found in fermentation environments and could also be found in the metagenomic samples. The species *Lactobacillus acidophilus, Levilactobacillus brevis, Limosilactobacillus fermentum*, *Lla. lactis*, *Leuc. mesenteroides, Lacticaseibacillus paracasei*, *Lactiplantibacillus plantarum* and *Lacticaseibacillus rhamnosus* were identified as a result of the literature search and were classified as set 1. The taxa of set 1 were compared to 40 taxa that were found across all samples, referred to as set 2. The log ratio of both sets of taxa was computed for the consumer and non-

consumer group using churro [190]. Notably the consumer group had a higher log ratio of both sets of taxa compared to the non-consumer group, suggesting that the consumption of fermented foods is associated with the taxa contained in set one. To assess overall differences between the groups, differential abundance testing was performed using the statistical tool songbird [191] which ranks taxa that have changed the most between groups. The log ratio of 40 of the highest (set 5) and lowest ranked species (set 6) associated with fermented food consumption were calculated using churro, with a higher log ratio of the set 5 to set 6 displayed by the consumer group. Such analysis identified compositional differences between consumers and non-consumers of fermented foods and identifies taxa that could be potential contributing to such differences. Furthermore the metagenomics data was integrated with metabolomics data to identify occurrence patterns between the taxa associated with the consumption of fermented foods and a isomer of the putative health promoting fatty acid conjugated linoleic acid (CLA) [183].

Section 1.4: Conclusion

Distinct bioinformatic tools applied in combination can be used to profile fermentation ecosystems, allowing for a detailed understanding of complex fermentation processes and the potential roles of individual species in the fermentation environment [22, 159]. The assessment of microbial communities in this manner has significantly advanced our understanding of mixed populations, and the microorganism's individual contributions to ecosystems. Metagenomic-associated methods, when applied to fermented foods, have resulted in the documentation of fermentation processes and production methods associated with many traditional foods. Such studies have shared and preserved knowledge concerning production practices, and in some cases have assessed/applied such insights to novel applications. Metabolic network analysis (see section 3) may in the future provide insights into mechanisms needed to customise pathways, ensuring the generation of amino acids that are associated with a desirable taste while avoiding those that lead to off flavours, allowing for increased product acceptance. For example, the production of glutamic acid in a sufficient quantity to promote or enhance the umami taste of a food. Despite, the insights provided by WMGS and metabolomics into flavour production, experimental validations such as, sensory analysis will

be key to confirming any hypotheses generated by WMGS [152].

Organisations such as the Food Safety and Inspection Service (FSIS), an agency of the United States Department of Agriculture, have recognised that sequencing-based approaches have the potential to become future microbial pathogen and spoilage detection tools [192]. Metagenomics is gathering considerable interest in such applications, and It is technologically feasible that in the future real time *in situ* metagenomics analysis will be possible for field research [193]. Current limitations of metagenomics in assessing the safety of food and food processing environments include its inability to provide sufficient gene expression level data and to definitively validate all taxonomic and functional outputs as associated with pathogenic microorganisms. Metagenomics is restricted to the detection of features with homology to those previously described and included in the available databases. Such homology-based inferences are hampered by the limited number of reference pathogens in current databases, which cannot yet account for all taxa of interest [175].

Despite this, given appropriate sequencing depth and coverage, shotgun metagenomics can provide a valuable overview of potential pathogens, virulence factors, and antimicrobial resistance genes and their source (Figure 3). Shotgun metagenomics provides a mechanistic understanding of the microbes in question furthering knowledge relating to associated pathogenic and/or spoilage processes and any additional factors that contribute to their occurrence, which can for now be used to complement culture dependent techniques [194], until validated and certified shotgun approaches for industry are established.

Few studies exist that have applied WMGS to assess the health-promoting aspects of fermented foods, and to our knowledge no metagenomic specific tool currently available, is designed to link taxonomic and functional outputs with biomarkers of health. The absence of health-related tools and publications is reflective of the limited mechanistic understanding of how fermented foods can promote health. A recent milk kefir review reported the findings of various animal models and *in vitro* studies, specifically investigating the effects of kefir consumption on cholesterol metabolism, cancer, wound healing, the immune system and the gut microbiome [65]. While the publications discussed in the review showed promising results, few were able to provide mechanistic insights, which could be integrated with

metagenomics tools and studies [195].

Metagenomics analysis will continue to benefit from the development and improvement of tools and sequencing platforms, as the field continues to embrace the ideas and principles of computer science. One significant improvement is database expansion, with an increasing number of genomes and metagenomes sequenced each year (Figure 2), allowing for improved annotations that can scale to bigger datasets. Tools such as InStrain [55] allow for the flexible construction of customised reference databases, enabling the incorporation of genomes, with publically available and or environmental specific MAGs. Such tools would be particularly suitable to fermented foods such as milk and water kefir, where there is a notable absence in the availability of Eukaryotic reference genomes (Figure 2 & Figure S1).

Despite this positive trend, the forecast advances in computational tools will largely be accessible through command line interfaces at institutes with sufficient computational infrastructures. Future tool developers should incorporate user-friendly functionalities or preferably a graphical user interface (GUI) in their tools to facilitate the participation of laboratory oriented microbiologists [196]. Concerning tool development, no single tool is available to date that encompasses all the steps required to interpret a metagenome, and gold standard methods capable of full-fledged metagenomics analyses have not yet been developed. Therefore performing bioinformatic analysis to meet a publishable scientific standard requires the researcher to identify, install and configure bioinformatic tools that will fulfill the research requirements and construct personal scripts to ensure the outputs of the various tools align [129].

Given the interest in this field as well as the developments in sequencing and bioinformatics, it is reasonable to assume that gold standard methods are on the horizon and will ensure accurate comparability between studies. In fact, the first steps have been taken towards this, with the development of MAGO (metagenome-assembled genomes orchestra) [126], MetaWRAP [129] and bioBakery workflows [29], all of which provides a standardised automated approach for metagenomic analysis, incorporating many separate softwares into an integrated workflow. Other tools have also been developed such as Squeezemeta that, similar to MAGO, provides an integrated standardised metagenomic and metatranscriptomic workflow, which is amenable to a standard local computer [27].

In conclusion, further research is required to identify and develop gold standard methods for metagenomics analysis that will allow for accurate standardised workflows for foods, fermented or otherwise and other environments. Such workflows should incorporate all steps required in a short-read and or assembly-based approach; ensuring comparability between studies by controlling for intrinsic differences in the selection of bioinformatics tools [119, 197-199]. Further research should begin by benchmarking the performance of tools across different biological ecosystems to assess their aptitude on the selected environment and develop new parameters to optimize performance based on use case scenarios. Such research would allow for the cataloguing of bioinformatic tools, enabling researchers to make informed decisions on their application.

Section 1.5: References

1. Selhub, E.M., A.C. Logan, and A.C. Bested, *Fermented foods, microbiota, and mental health: ancient practice meets nutritional psychiatry.* Journal of Physiological Anthropology, 2014. **33**(1): p. 2.
2. Marco, M.L., et al., *The International Scientific Association for Probiotics and Prebiotics (ISAPP) consensus statement on fermented foods.* Nature Reviews Gastroenterology & Hepatology, 2021. **18**(3): p. 196-208.
3. Kabak, B. and A.D. Dobson, *An introduction to the traditional fermented foods and beverages of Turkey.* Crit Rev Food Sci Nutr, 2011. **51**(3): p. 248-60.
4. Marco, M.L., et al., *Health benefits of fermented foods: microbiota and beyond.* Current Opinion in Biotechnology, 2017. **44**: p. 94-102.
5. Şanlier, N., B.B. Gökcen, and A.C. Sezgin, *Health benefits of fermented foods.* Critical Reviews in Food Science and Nutrition, 2019. **59**(3): p. 506-527.
6. Nout, M.J.R., *Food Technologies: Fermentation*, in *Encyclopedia of Food Safety*, Y. Motarjemi, Editor. 2014, Academic Press: Waltham. p. 168-177.
7. Walsh, A.M., et al., *Microbial Succession and Flavor Production in the Fermented Dairy Beverage Kefir.* mSystems, 2016. **1**(5): p. e00052-16.
8. Bove, P., et al., *Lactobacillus plantarum passage through an oro-gastro-intestinal tract simulator: carrier matrix effect and transcriptional analysis of genes associated to stress and probiosis.* Microbiol Res, 2013. **168**(6): p. 351-359.
9. Chaves-López, C., et al., *Traditional Fermented Foods and Beverages from a Microbiological and Nutritional Perspective: The Colombian Heritage.* Comprehensive Reviews in Food Science and Food Safety, 2014. **13**(5): p. 1031-1048.
10. Martinez-Villaluenga, C., E. Peñas, and J. Frias, *Chapter 2 - Bioactive Peptides in Fermented Foods: Production and Evidence for Health Effects*, in *Fermented Foods in Health and Disease Prevention*, J. Frias, C. Martinez-Villaluenga, and E. Peñas, Editors. 2017, Academic Press: Boston. p. 23-47.
11. Nampoothiri, K.M., et al., *Chapter 3 - Health Benefits of Exopolysaccharides in Fermented Foods*, in *Fermented Foods in Health and Disease Prevention*, J. Frias, C. Martinez-Villaluenga, and E. Peñas, Editors. 2017, Academic Press: Boston. p. 49-62.

12. Gille, D., et al., *Fermented Food and Non-Communicable Chronic Diseases: A Review.* Nutrients, 2018. **10**(4).
13. Cao, Y., et al., *A Review on the Applications of Next Generation Sequencing Technologies as Applied to Food-Related Microbiome Studies.* Frontiers in microbiology, 2017. **8**: p. 1829-1829.
14. Zepeda Mendoza, M.L., T. Sicheritz-Pontén, and M.T. Gilbert, *Environmental genes and genomes: understanding the differences and challenges in the approaches and software for their analyses.* Brief Bioinform, 2015. **16**(5): p. 745-58.
15. Solden, L., K. Lloyd, and K. Wrighton, *The bright side of microbial dark matter: lessons learned from the uncultivated majority.* Curr Opin Microbiol, 2016. **31**: p. 217-226.
16. Abe, M., et al., *Characteristic fungi observed in the fermentation process for Puer tea.* Int J Food Microbiol, 2008. **124**(2): p. 199-203.
17. Tian, J., et al., *Bacterial and fungal communities in Pu'er tea samples of different ages.* J Food Sci, 2013. **78**(8): p. M1249-56.
18. De Filippis, F., E. Parente, and D. Ercolini, *Metagenomics insights into food fermentations.* Microbial biotechnology, 2017. **10**(1): p. 91-102.
19. Hillmann, B., et al., *Evaluating the Information Content of Shallow Shotgun Metagenomics.* mSystems, 2018. **3**(6): p. e00069-18.
20. Quince, C., et al., *Shotgun metagenomics, from sampling to analysis.* Nat Biotechnol, 2017. **35**(9): p. 833-844.
21. Walsh, A.M., et al., *Species classifier choice is a key consideration when analysing low-complexity food microbiome data.* Microbiome, 2018. **6**(1): p. 50-50.
22. Beier, S., R. Tappu, and D.H. Huson, *Functional Analysis in Metagenomics Using MEGAN 6*, in *Functional Metagenomics: Tools and Applications*, T.C. Charles, M.R. Liles, and A. Sessitsch, Editors. 2017, Springer International Publishing: Cham. p. 65-74.
23. Truong, D.T., et al., *Microbial strain-level population structure and genetic diversity from metagenomes.* Genome research, 2017. **27**(4): p. 626-638.
24. Hutchins, B.I., et al., *Relative Citation Ratio (RCR): A New Metric That Uses Citation Rates to Measure Influence at the Article Level.* PLOS Biology, 2016. **14**(9): p. e1002541.
25. Hutchins, B.I., et al., *The NIH Open Citation Collection: A public access, broad coverage resource.* PLOS Biology, 2019. **17**(10): p. e3000385.
26. Hübler, R., et al., *HOPS: automated detection and authentication of pathogen DNA in archaeological remains.* Genome Biology, 2019. **20**(1): p. 280.
27. Tamames, J. and F. Puente-Sánchez, *SqueezeMeta, A Highly Portable, Fully Automatic Metagenomic Analysis Pipeline.* Frontiers in Microbiology, 2019. **9**(3349).
28. Shakya, M., C.-C. Lo, and P.S.G. Chain, *Advances and Challenges in Metatranscriptomic Analysis.* Frontiers in Genetics, 2019. **10**.
29. Beghini, F., et al., *Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3.* eLife, 2021. **10**: p. e65088.
30. Altschul, S.F., et al., *Basic local alignment search tool.* J Mol Biol, 1990. **215**(3): p. 403-10.
31. Wood, D.E., J. Lu, and B. Langmead, *Improved metagenomic analysis with Kraken 2.* Genome Biology, 2019. **20**(1): p. 257.
32. Wood, D.E. and S.L. Salzberg, *Kraken: ultrafast metagenomic sequence classification using exact alignments.* Genome Biology, 2014. **15**(3): p. R46.
33. Silva, G.G., et al., *SUPER-FOCUS: a tool for agile functional analysis of shotgun metagenomic data.* Bioinformatics, 2016. **32**(3): p. 354-61.
34. Gloor, G.B., et al., *Microbiome Datasets Are Compositional: And This Is Not Optional.* Frontiers in Microbiology, 2017. **8**(2224).
35. Calle, M.L., *Statistical Analysis of Metagenomics Data.* Genomics & informatics, 2019. **17**(1): p. e6-e6.
36. Segata, N., et al., *Metagenomic microbial community profiling using unique clade-specific marker genes.* Nature Methods, 2012. **9**(8): p. 811-814.

37. Truong, D.T., et al., *MetaPhlAn2 for enhanced metagenomic taxonomic profiling.* Nat Methods, 2015. **12**(10): p. 902-3.
38. Silva, G.G.Z., et al., *FOCUS: an alignment-free model to identify organisms in metagenomes using non-negative least squares.* PeerJ, 2014. **2**: p. e425.
39. Ayling, M., M.D. Clark, and R.M. Leggett, *New approaches for metagenome assembly with short reads.* Brief Bioinform, 2020. **21**(2): p. 584-594.
40. Pevzner, P.A., H. Tang, and M.S. Waterman, *An Eulerian path approach to DNA fragment assembly.* Proceedings of the National Academy of Sciences, 2001. **98**(17): p. 9748-9753.
41. Nurk, S., et al., *metaSPAdes: a new versatile metagenomic assembler.* Genome research, 2017. **27**(5): p. 824-834.
42. Peng, Y., et al., *IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth.* Bioinformatics, 2012. **28**(11): p. 1420-8.
43. Chen, L.X., et al., *Accurate and complete genomes from metagenomes.* Genome Res, 2020. **30**(3): p. 315-333.
44. Alneberg, J., et al., *Binning metagenomic contigs by coverage and composition.* Nat Methods, 2014. **11**(11): p. 1144-6.
45. Sedlar, K., K. Kupkova, and I. Provaznik, *Bioinformatics strategies for taxonomy independent binning and visualization of sequences in shotgun metagenomics.* Computational and Structural Biotechnology Journal, 2017. **15**: p. 48-55.
46. Kang, D.D., et al., *MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities.* PeerJ, 2015. **3**: p. e1165.
47. Kang, D.D., et al., *MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies.* PeerJ, 2019. **7**: p. e7359.
48. Pride, D.T., et al., *Evolutionary implications of microbial genome tetranucleotide frequency biases.* Genome Res, 2003. **13**(2): p. 145-58.
49. Wang, Z., et al., *Assessment of metagenomic assemblers based on hybrid reads of real and simulated metagenomic sequences.* Briefings in Bioinformatics, 2019.
50. Wu, Y.W., B.A. Simmons, and S.W. Singer, *MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets.* Bioinformatics, 2016. **32**(4): p. 605-7.
51. Parks, D.H., et al., *CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes.* Genome Res, 2015. **25**(7): p. 1043-55.
52. Imelfort, M., et al., *GroopM: an automated tool for the recovery of population genomes from related metagenomes.* PeerJ, 2014. **2**: p. e603.
53. Asnicar, F., et al., *Precise phylogenetic analysis of microbial isolates and genomes from metagenomes using PhyloPhlAn 3.0.* Nature Communications, 2020. **11**(1): p. 2500.
54. Mukherjee, S., et al., *Genomes OnLine database (GOLD) v.7: updates and new features.* Nucleic acids research, 2019. **47**(D1): p. D649-D659.
55. Olm, M.R., et al., *inStrain profiles population microdiversity from metagenomic data and sensitively detects shared microbial strains.* Nature Biotechnology, 2021. **39**(6): p. 727-736.
56. Chaumeil, P.-A., et al., *GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database.* Bioinformatics, 2019. **36**(6): p. 1925-1927.
57. Hyatt, D., et al., *Prodigal: prokaryotic gene recognition and translation initiation site identification.* BMC bioinformatics, 2010. **11**: p. 119-119.
58. Matsen, F.A., R.B. Kodner, and E.V. Armbrust, *pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree.* BMC Bioinformatics, 2010. **11**(1): p. 538.
59. Dong, X. and M. Strous, *An Integrated Pipeline for Annotation and Visualization of Metagenomic Contigs.* Frontiers in genetics, 2019. **10**: p. 999-999.
60. Seemann, T., *Prokka: rapid prokaryotic genome annotation.* Bioinformatics, 2014. **30**(14): p. 2068-2069.
61. Apweiler, R., et al., *UniProt: the Universal Protein knowledgebase.* Nucleic Acids Research, 2004. **32**(suppl_1): p. D115-D119.

62. Punta, M., et al., *The Pfam protein families database.* Nucleic Acids Res, 2012. **40**(Database issue): p. D290-301.
63. Tatusova, T., et al., *RefSeq microbial genomes database: new representation and annotation strategy.* Nucleic Acids Res, 2014. **42**(Database issue): p. D553-9.
64. Clark, K., et al., *GenBank.* Nucleic acids research, 2016. **44**(D1): p. D67-D72.
65. Bourrie, B.C.T., B.P. Willing, and P.D. Cotter, *The Microbiota and Health Promoting Characteristics of the Fermented Beverage Kefir.* Frontiers in Microbiology, 2016. **7**(647).
66. Drost, H.-G. and J. Paszkowski, *Biomartr: genomic data retrieval with R.* Bioinformatics, 2017. **33**(8): p. 1216-1217.
67. Wickham, H., *ggplot2: elegant graphics for data analysis*. 2016: springer.
68. Yu, G., *Using ggtree to Visualize Data on Tree-Like Structures.* Current Protocols in Bioinformatics, 2020. **69**(1): p. e96.
69. Sunagawa, S., et al., *Structure and function of the global ocean microbiome.* Science, 2015. **348**(6237): p. 1261359.
70. Menzel, P., et al., *Comparative Metagenomics of Eight Geographically Remote Terrestrial Hot Springs.* Microb Ecol, 2015. **70**(2): p. 411-24.
71. Verce, M., L. De Vuyst, and S. Weckx, *Shotgun Metagenomics of a Water Kefir Fermentation Ecosystem Reveals a Novel Oenococcus Species.* Frontiers in Microbiology, 2019. **10**(479).
72. Aggarwala, V., G. Liang, and F.D. Bushman, *Viral communities of the human gut: metagenomic analysis of composition and dynamics.* Mobile DNA, 2017. **8**(1): p. 12.
73. Prosser, J.I., *Dispersing misconceptions and identifying opportunities for the use of 'omics' in soil microbial ecology.* Nature Reviews Microbiology, 2015. **13**(7): p. 439-446.
74. Boyd, J.A., B.J. Woodcroft, and G.W. Tyson, *GraftM: a tool for scalable, phylogenetically informed classification of genes within metagenomes.* Nucleic Acids Research, 2018. **46**(10): p. e59-e59.
75. Wallace, J.C., et al., *FARME DB: a functional antibiotic resistance element database.* Database (Oxford), 2017. **2017**.
76. Guo, J., et al., *Review, Evaluation, and Directions for Gene-Targeted Assembly for Ecological Analyses of Metagenomes.* Frontiers in Genetics, 2019. **10**(957).
77. Zhang, H., et al., *dbCAN2: a meta server for automated carbohydrate-active enzyme annotation.* Nucleic Acids Research, 2018. **46**(W1): p. W95-W101.
78. Almeida, O.G.G. and E.C.P. De Martinis, *Bioinformatics tools to assess metagenomic data for applied microbiology.* Applied Microbiology and Biotechnology, 2019. **103**(1): p. 69-82.
79. Segata, N., *On the Road to Strain-Resolved Comparative Metagenomics.* mSystems, 2018. **3**(2): p. e00190-17.
80. Leimbach, A., J. Hacker, and U. Dobrindt, *E. coli as an all-rounder: the thin line between commensalism and pathogenicity.* Curr Top Microbiol Immunol, 2013. **358**: p. 3-32.
81. Van Rossum, T., et al., *Diversity within species: interpreting strains in microbiomes.* Nature Reviews Microbiology, 2020. **18**(9): p. 491-506.
82. Scholz, M., et al., *Strain-level microbial epidemiology and population genomics from shotgun metagenomics.* Nature Methods, 2016. **13**(5): p. 435-438.
83. Hendrix, R.W., *Bacteriophages: evolution of the majority.* Theoretical population biology, 2002. **61**(4): p. 471-480.
84. Zahn, J.A. and M.C. Halter. *Surveillance and Elimination of Bacteriophage Contamination in an Industrial Fermentation Process*. 2018.
85. Weinbauer, M.G. and F. Rassoulzadegan, *Are viruses driving microbial diversification and diversity?* Environ Microbiol, 2004. **6**(1): p. 1-11.
86. Fernández, L., et al., *Bacteriophages in the Dairy Environment: From Enemies to Allies.* Antibiotics (Basel), 2017. **6**(4).
87. Garneau, J.E. and S. Moineau, *Bacteriophages of lactic acid bacteria and their impact on milk fermentations.* Microb Cell Fact, 2011. **10 Suppl 1**(Suppl 1): p. S20.
88. Roux, S., *A Viral Ecogenomics Framework To Uncover the Secrets of Nature's "Microbe Whisperers".* mSystems, 2019. **4**(3).

89. Ren, J., et al., *VirFinder: a novel k -mer based tool for identifying viral sequences from assembled metagenomic data.* Microbiome, 2017. **5**(1): p. 69.
90. Roux, S., et al., *Benchmarking viromics: an in silico evaluation of metagenome-enabled estimates of viral community composition and diversity.* PeerJ, 2017. **5**: p. e3817.
91. Guo, J., et al., *VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses.* Microbiome, 2021. **9**(1): p. 37.
92. Roux, S., et al., *VirSorter: mining viral signal from microbial genomic data.* PeerJ, 2015. **3**: p. e985.
93. Blair, J.M.A., et al., *Molecular mechanisms of antibiotic resistance.* Nature Reviews Microbiology, 2015. **13**(1): p. 42-51.
94. Gillings, M.R. and H.W. Stokes, *Are humans increasing bacterial evolvability?* Trends Ecol Evol, 2012. **27**(6): p. 346-52.
95. Macesic, N., F. Polubriaginof, and N.P. Tatonetti, *Machine learning: novel bioinformatics approaches for combating antimicrobial resistance.* Curr Opin Infect Dis, 2017. **30**(6): p. 511-517.
96. Brown, E.D. and G.D. Wright, *Antibacterial drug discovery in the resistance era.* Nature, 2016. **529**(7586): p. 336-43.
97. Ventola, C.L., *The antibiotic resistance crisis: part 1: causes and threats.* P t, 2015. **40**(4): p. 277-83.
98. Guitor, A.K., et al., *Capturing the Resistome: a Targeted Capture Method To Reveal Antibiotic Resistance Determinants in Metagenomes.* Antimicrobial Agents and Chemotherapy, 2019. **64**(1): p. e01324-19.
99. Satlin, M.J., et al., *Bacteremia due to carbapenem-resistant Enterobacteriaceae in neutropenic patients with hematologic malignancies.* J Infect, 2016. **73**(4): p. 336-45.
100. Berglund, F., et al., *Identification and reconstruction of novel antibiotic resistance genes from metagenomes.* Microbiome, 2019. **7**(1): p. 52.
101. Alcock, B.P., et al., *CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database.* Nucleic Acids Research, 2019. **48**(D1): p. D517-D525.
102. Liu, B. and M. Pop, *ARDB—Antibiotic Resistance Genes Database.* Nucleic Acids Research, 2008. **37**(suppl_1): p. D443-D447.
103. Zankari, E., et al., *Identification of acquired antimicrobial resistance genes.* Journal of Antimicrobial Chemotherapy, 2012. **67**(11): p. 2640-2644.
104. Gupta, S.K., et al., *ARG-ANNOT, a new bioinformatic tool to discover antibiotic resistance genes in bacterial genomes.* Antimicrob Agents Chemother, 2014. **58**(1): p. 212-20.
105. Arango-Argoty, G., et al., *DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data.* Microbiome, 2018. **6**(1): p. 23.
106. Zhang, C. and Q. Hua, *Applications of Genome-Scale Metabolic Models in Biotechnology and Systems Medicine.* Frontiers in Physiology, 2016. **6**(413).
107. Lieven, C., et al., *MEMOTE for standardized genome-scale metabolic model testing.* Nature Biotechnology, 2020. **38**(3): p. 272-276.
108. Zomorrodi, A.R. and C.D. Maranas, *OptCom: a multi-level optimization framework for the metabolic modeling and analysis of microbial communities.* PLoS Comput Biol, 2012. **8**(2): p. e1002363.
109. Zelezniak, A., et al., *Metabolic dependencies drive species co-occurrence in diverse microbial communities.* Proc Natl Acad Sci U S A, 2015. **112**(20): p. 6449-54.
110. Cuevas, D.A., et al., *From DNA to FBA: How to Build Your Own Genome-Scale Metabolic Model.* Frontiers in Microbiology, 2016. **7**(907).
111. Karlsen, E., C. Schulz, and E. Almaas, *Automated generation of genome-scale metabolic draft reconstructions based on KEGG.* BMC Bioinformatics, 2018. **19**(1): p. 467.
112. Thiele, I. and B. Palsson, *A protocol for generating a high-quality genome-scale metabolic reconstruction.* Nat Protoc, 2010. **5**(1): p. 93-121.
113. Li, C., et al., *BioModels Database: An enhanced, curated and annotated resource for published quantitative kinetic models.* BMC Systems Biology, 2010. **4**(1): p. 92.
114. Norsigian, C.J., et al., *BiGG Models 2020: multi-strain genome-scale models and expansion across the phylogenetic tree.* Nucleic Acids Research, 2019. **48**(D1): p. D402-D406.

115. Mendoza, S.N., et al., *A systematic assessment of current genome-scale metabolic reconstruction tools.* Genome Biology, 2019. **20**(1): p. 158.
116. Machado, D., et al., *Fast automated reconstruction of genome-scale metabolic models for microbial species and communities.* Nucleic Acids Research, 2018. **46**(15): p. 7542-7553.
117. Belcour, A., et al., *Metage2Metabo, microbiota-scale metabolic complementarity for the identification of key species.* eLife, 2020. **9**: p. e61968.
118. Meyer, F., et al., *Tutorial: assessing metagenomics software with the CAMI benchmarking toolkit.* Nature Protocols, 2021.
119. Sczyrba, A., et al., *Critical Assessment of Metagenome Interpretation—a benchmark of metagenomics software.* Nature Methods, 2017. **14**(11): p. 1063-1071.
120. McHardy, A., A. Sczyrba, and T. Rattei. *The Critical Assessment of Metagenome Interpretation (CAMI) competition*. 2014.
121. Meyer, F., et al., *Assessing taxonomic metagenome profilers with OPAL.* Genome Biol, 2019. **20**(1): p. 51.
122. Mikheenko, A., V. Saveliev, and A. Gurevich, *MetaQUAST: evaluation of metagenome assemblies.* Bioinformatics, 2016. **32**(7): p. 1088-90.
123. Meyer, F., et al., *AMBER: Assessment of Metagenome BinnERs.* GigaScience, 2018. **7**(6): p. giy069.
124. Fritz, A., et al., *CAMISIM: simulating metagenomes and microbial communities.* Microbiome, 2019. **7**(1): p. 17.
125. Ye, S.H., et al., *Benchmarking Metagenomics Tools for Taxonomic Classification.* Cell, 2019. **178**(4): p. 779-794.
126. Murovec, B., L. Deutsch, and B. Stres, *Computational Framework for High-Quality Production and Large-Scale Evolutionary Analysis of Metagenome Assembled Genomes.* Molecular Biology and Evolution, 2019. **37**(2): p. 593-598.
127. Graham, E.D., J.F. Heidelberg, and B.J. Tully, *BinSanity: unsupervised clustering of environmental microbial assemblies using coverage and affinity propagation.* PeerJ, 2017. **5**: p. e3035.
128. Piro, V.C., M. Matschkowski, and B.Y. Renard, *MetaMeta: integrating metagenome analysis tools to improve taxonomic profiling.* Microbiome, 2017. **5**(1): p. 101.
129. Uritskiy, G.V., J. DiRuggiero, and J. Taylor, *MetaWRAP-a flexible pipeline for genome-resolved metagenomic data analysis.* Microbiome, 2018. **6**(1): p. 158.
130. McIver, L.J., et al., *bioBakery: a meta'omic analysis environment.* Bioinformatics, 2017. **34**(7): p. 1235-1237.
131. Jung, J.Y., et al., *Metagenomic Analysis of Kimchi, a Traditional Korean Fermented Food.* Applied and Environmental Microbiology, 2011. **77**(7): p. 2264.
132. Bellon, J.R., et al., *Designing and creating Saccharomyces interspecific hybrids for improved, industry relevant, phenotypes.* Appl Microbiol Biotechnol, 2015. **99**(20): p. 8597-609.
133. Meersman, E., et al., *Breeding Strategy To Generate Robust Yeast Starter Cultures for Cocoa Pulp Fermentations.* Appl Environ Microbiol, 2015. **81**(18): p. 6166-76.
134. Flores, M., et al., *Yeast strains as potential aroma enhancers in dry fermented sausages.* Int J Food Microbiol, 2015. **212**: p. 16-24.
135. Menzel, P., K.L. Ng, and A. Krogh, *Fast and sensitive taxonomic classification for metagenomics with Kaiju.* Nature Communications, 2016. **7**(1): p. 11257.
136. Lee, J.-H., H. Yi, and J. Chun, *rRNASelector: a computer program for selecting ribosomal RNA encoding sequences from metagenomic and metatranscriptomic shotgun libraries.* Journal of microbiology (Seoul, Korea), 2011. **49**(4): p. 689-691.
137. Bengtsson-Palme, J., et al., *Improved software detection and extraction of ITS1 and ITS2 from ribosomal ITS sequences of fungi and other eukaryotes for analysis of environmental sequencing data.* Methods in Ecology and Evolution, 2013. **4**(10): p. 914-919.
138. Li, D., et al., *MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices.* Methods, 2016. **102**: p. 3-11.
139. Li, H., *Aligning sequence reads, clone sequences and assembly contigs with*

BWA-MEM. 2013: arXiv [Preprint].
140. Wolfe, B.E., et al., *Cheese rind communities provide tractable systems for in situ and in vitro studies of microbial diversity.* Cell, 2014. **158**(2): p. 422-433.
141. Sulaiman, J., et al., *Microbial succession and the functional potential during the fermentation of Chinese soy sauce brine.* Front Microbiol, 2014. **5**: p. 556.
142. Bertuzzi, A.S., et al., *Omics-Based Insights into Flavor Development and Microbial Succession within Surface-Ripened Cheese.* mSystems, 2018. **3**(1): p. e00211-17.
143. Stanke, M. and B. Morgenstern, *AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints.* Nucleic Acids Res, 2005. **33**(Web Server issue): p. W465-7.
144. Bankevich, A., et al., *SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing.* J Comput Biol, 2012. **19**(5): p. 455-77.
145. Arndt, D., et al., *PHASTER: a better, faster version of the PHAST phage search tool.* Nucleic Acids Res, 2016. **44**(W1): p. W16-21.
146. Dugat-Bony, E., et al., *Viral metagenomic analysis of the cheese surface: A comparative study of rapid procedures for extracting viral particles.* Food Microbiol, 2020. **85**: p. 103278.
147. Melkonian, C., et al., *Finding Functional Differences Between Species in a Microbial Community: Case Studies in Wine Fermentation and Kefir Culture.* Frontiers in Microbiology, 2019. **10**(1347).
148. Liu, J., et al., *Composition-based classification of short metagenomic sequences elucidates the landscapes of taxonomic and functional enrichment of microorganisms.* Nucleic Acids Research, 2012. **41**(1): p. e3-e3.
149. Afiahayati, K. Sato, and Y. Sakakibara, *MetaVelvet-SL: an extension of the Velvet assembler to a de novo metagenomic assembler utilizing supervised learning.* DNA Res, 2015. **22**(1): p. 69-77.
150. Rho, M., H. Tang, and Y. Ye, *FragGeneScan: predicting genes in short and error-prone reads.* Nucleic Acids Res, 2010. **38**(20): p. e191.
151. Kanehisa, M., et al., *Data, information, knowledge and principle: back to metabolism in KEGG.* Nucleic Acids Research, 2013. **42**(D1): p. D199-D205.
152. Wu, L.-H., et al., *Metagenomics reveals flavour metabolic network of cereal vinegar microbiota.* Food Microbiology, 2017. **62**: p. 23-31.
153. Laranjo, M., M.E. Potes, and M. Elias, *Role of Starter Cultures on the Safety of Fermented Meat Products.* Frontiers in Microbiology, 2019. **10**(853).
154. Liang, H., et al., *Dynamic and Functional Characteristics of Predominant Species in Industrial Paocai as Revealed by Combined DGGE and Metagenomic Sequencing.* Frontiers in microbiology, 2018. **9**: p. 2416-2416.
155. Chen, W. and A. Narbad, *Lactic acid bacteria in foodborne hazards reduction*. 2018: Springer.
156. Huang, Y., et al., *CD-HIT Suite: a web server for clustering and comparing biological sequences.* Bioinformatics, 2010. **26**(5): p. 680-2.
157. Kanehisa, M., Y. Sato, and K. Morishima, *BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences.* J Mol Biol, 2016. **428**(4): p. 726-731.
158. Edgar, R.C., *MUSCLE: multiple sequence alignment with high accuracy and high throughput.* Nucleic Acids Res, 2004. **32**(5): p. 1792-7.
159. Illeghems, K., S. Weckx, and L. De Vuyst, *Applying meta-pathway analyses through metagenomics to identify the functional properties of the major bacterial communities of a single spontaneous cocoa bean fermentation process sample.* Food Microbiology, 2015. **50**: p. 54-63.
160. Petersen, T.N., et al., *MGmapper: Reference based mapping and taxonomy annotation of metagenomics sequence reads.* PLOS ONE, 2017. **12**(5): p. e0176469.
161. Zepeda-Mendoza, M.L., et al., *Influence of Oenococcus oeni and Brettanomyces bruxellensis on Wine Microbial Taxonomic and Functional Potential Profiles.* American Journal of Enology and Viticulture, 2018. **69**(4): p. 321-333.
162. Luo, R., et al., *SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler.* Gigascience, 2012. **1**(1): p. 18.

163. Li, Z., et al., *Revealing the influence of microbiota on the quality of Pu-erh tea during fermentation process by shotgun metagenomic and metabolomic analysis.* Food Microbiology, 2018. **76**: p. 405-415.
164. Edgar, R.C., *Search and clustering orders of magnitude faster than BLAST.* Bioinformatics, 2010. **26**(19): p. 2460-2461.
165. Ferrocino, I., et al., *Shotgun Metagenomics and Volatilome Profile of the Microbiota of Fermented Sausages.* Applied and environmental microbiology, 2018. **84**(3): p. e02120-17.
166. World Health Organization, *The burden of foodborne diseases in the WHO European Region (2017).* 2017.
167. Kable, M.E., et al., *The core and seasonal microbiota of raw bovine milk in tanker trucks and the impact of transfer to a milk processing facility.* MBio, 2016. **7**(4): p. e00836-16.
168. Ercolini, D., et al., *Monitoring of microbial metabolites and bacterial diversity in beef stored under different packaging conditions.* Applied and environmental microbiology, 2011. **77**(20): p. 7372-7381.
169. Stasiewicz, M.J., H.C. den Bakker, and M. Wiedmann, *Genomics tools in microbial food safety.* Current Opinion in Food Science, 2015. **4**: p. 105-110.
170. Zolfo, M., et al., *MetaMLST: multi-locus strain-level bacterial typing from metagenomic samples.* Nucleic Acids Res, 2017. **45**(2): p. e7.
171. Walsh, A.M., et al., *Strain-Level Metagenomic Analysis of the Fermented Dairy Beverage Nunu Highlights Potential Food Safety Risks.* Appl Environ Microbiol, 2017. **83**(16).
172. Asnicar, F., et al., *Compact graphical representation of phylogenetic data and metadata with GraPhlAn.* PeerJ, 2015. **3**: p. e1029.
173. Oniciuc, E.-A., et al., *Food processing as a risk factor for antimicrobial resistance spread along the food chain.* Current Opinion in Food Science, 2019. **30**: p. 21-26.
174. Lakin, S.M., et al., *MEGARes: an antimicrobial resistance database for high throughput sequencing.* Nucleic Acids Research, 2016. **45**(D1): p. D574-D580.
175. Alexa , E.A., et al., *Dairy Products and Dairy-Processing Environments as a Reservoir of Antibiotic Resistance and Quorum-Quenching Determinants as Revealed through Functional Metagenomics.* mSystems, 2020. **5**(1): p. e00723-19.
176. Salazar, J.K., et al., *Metagenomics of pasteurized and unpasteurized gouda cheese using targeted 16S rDNA sequencing.* BMC Microbiology, 2018. **18**(1): p. 189.
177. Colombo, S., et al., *Characterization of airborne viromes in cheese production plants.* J Appl Microbiol, 2018. **125**(5): p. 1444-1454.
178. Miller, J.R., et al., *Aggressive assembly of pyrosequencing reads with mates.* Bioinformatics, 2008. **24**(24): p. 2818-24.
179. Laserson, J., V. Jojic, and D. Koller, *Genovo: de novo assembly for metagenomes.* J Comput Biol, 2011. **18**(3): p. 429-43.
180. Delcher, A.L., et al., *Improved microbial gene identification with GLIMMER.* Nucleic Acids Res, 1999. **27**(23): p. 4636-41.
181. Shahbandeh, M., *Global functional food market revenue 2019-2025.* 2019.
182. Behera, S., et al., *Advances in Microbial Fermentation and Fermented Food for Health*. 2018. p. 53-69.
183. Taylor, B.C., et al., *Consumption of Fermented Foods Is Associated with Systematic Differences in the Gut Microbiome and Metabolome.* mSystems, 2020. **5**(2): p. e00901-19.
184. Leech, J., et al., *Fermented-Food Metagenomics Reveals Substrate-Associated Differences in Taxonomy and Health-Associated and Antibiotic Resistance Determinants.* mSystems, 2020. **5**(6): p. e00522-20.
185. de Jong, A., et al., *BAGEL: a web-based bacteriocin genome mining tool.* Nucleic Acids Res, 2006. **34**(Web Server issue): p. W273-9.
186. Pasolli, E., et al., *Large-scale genome-wide analysis links lactic acid bacteria from food with the gut microbiome.* Nature Communications, 2020. **11**(1): p. 2610.

187. McDonald, D., et al., *American Gut: an Open Platform for Citizen Science Microbiome Research.* mSystems, 2018. **3**(3): p. e00031-18.
188. Hillmann, B., et al., *SHOGUN: a modular, accurate and scalable framework for microbiome quantification.* Bioinformatics, 2020. **36**(13): p. 4088-4090.
189. O'Leary, N.A., et al., *Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation.* Nucleic Acids Res, 2016. **44**(D1): p. D733-45.
190. Fedarko, M.W., et al., *Visualizing 'omic feature rankings and log-ratios using Qurro.* NAR Genomics and Bioinformatics, 2020. **2**(2).
191. Morton, J.T., et al., *Establishing microbial composition measurement standards with reference frames.* Nature Communications, 2019. **10**(1): p. 2719.
192. Bauer, N., et al. *White paper: current and future development and use of molecular subtyping by USDA-FSIS. Food Safety and Inspection Service, U.S. Department of Agriculture, Washington, DC:*. 2014; Available from: https://www.fsis.usda.gov/wps/wcm/connect/6c7f71fd-2c0c-4ff0-b2bc-4977c7947516/Molecular-Subtyping-White-Paper.pdf?MOD=AJPERES.
193. Albrecht, B., C. Bağcı, and D.H. Huson, *MAIRA- real-time taxonomic and functional analysis of long reads on a laptop.* BMC Bioinformatics, 2020. **21**(13): p. 390.
194. Hazards, E.Panel o.B., et al., *Whole genome sequencing and metagenomics for outbreak investigation, source attribution and risk assessment of food-borne microorganisms.* EFSA Journal, 2019. **17**(12): p. e05898.
195. Slattery, C., P.D. Cotter, and P.W. O'Toole, *Analysis of Health Benefits Conferred by Lactobacillus Species from Kefir.* Nutrients, 2019. **11**(6): p. 1252.
196. Liu, Y.-X., et al., *A practical guide to amplicon and metagenomic analysis of microbiome data.* Protein & Cell, 2021. **12**(5): p. 315-330.
197. Couto, N., et al. *Critical steps in clinical shotgun metagenomics for the concomitant detection and typing of microbial pathogens*. Scientific reports, 2018. **8**, 13767 DOI: 10.1038/s41598-018-31873-w.
198. Angers-Loustau, A., et al., *The challenges of designing a benchmark strategy for bioinformatics pipelines in the identification of antimicrobial resistance determinants using next generation sequencing technologies.* F1000Res, 2018. **7**.
199. Gruening, B., et al., *Recommendations for the packaging and containerizing of bioinformatics software.* F1000Res, 2018. **7**.

Section 1.6: Supplementary information

Supplementary Figures

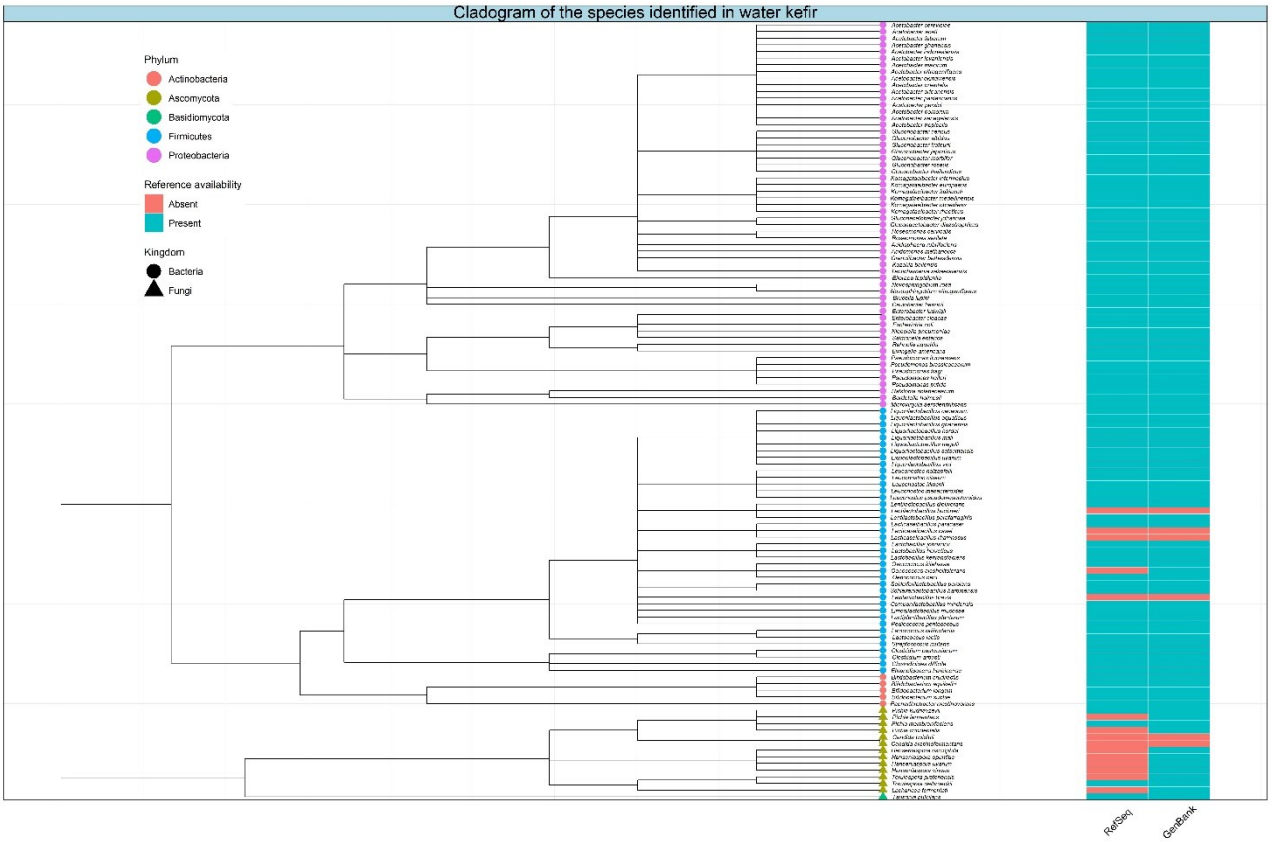


Figure S1: Cladogram of bacterial and fungal species found in water kefir, species information obtained from (Bourrie, Willing and Cotter, 2016). Tip labels represent the species detected, e.g., Lacticaseibacillus paracasei, tip colour represents the phylum, and shape represents the kingdom of the species. Presence/absence heatmap displays the availability of reference genomes for each of the listed species in both NCBI RefSeq and NCBI GenBank databases.

Chapter 2

Analysis of the milk kefir pan-metagenome reveals four community types, core species and associated metabolic pathways

Published in iScience - Cell Press

(https://doi.org/10.1016/j.isci.2023.108004)

Authors: Liam H Walsh, Mairéad Coakley, Fiona Crispie, Paul W O'Toole, and Paul D. Cotter

Contributions:

- **Candidate** performed fermentations, sample collection, DNA extractions, sequencing library preparations, and bioinformatic and statistical analysis
- **Candidate** wrote the manuscript with contributions from M.C., A.M.W, P.W.O.T. and P.D.C
- **F.C.** conducted sequencing
- **PWOT**, and **PDC** supervised the study

Section 2.0: Abstract

Milk kefir, a fermented dairy beverage, has a long history of safe consumption. Particular milk kefir beverages, and the strains therein, have also been associated with specific health benefits. A comprehensive metagenomics-based understanding of the microorganisms present within milk kefir communities from across the globe can help to define the milk kefir pan-metagenome, which includes details such as how the core milk kefir microbiome and the extent non-core components can vary. Such insights are of considerable scientific and commercial value and can contribute to future efforts to create tailored kefir-based microbial communities for different applications. Indeed, the identification of a core microbiome can assist regulators and kefir producers to ensure that products marketed as kefir contain a microbial composition that reflects the artisanal beverage. To facilitate this we undertook a large-scale metagenomic analysis of milk kefir samples, generated by inoculating full fat, pasteurised cow's milk with 64 separate kefir grains sourced from 25 different countries, to study patterns of strain, species, and functional pathway variation. The detection of core, highly prevalent and sample specific, compositional, strain and functional features help to define the milk kefir pan-metagenome. This analysis identified core features in the milk kefir microbiome, revealing a consistent pattern of domination by one microbial strain from the species *Lactobacillus helveticus* or the sub-species *Lactobacillus kefiranofaciens* subsp*. kefiranofaciens*, *Lactococcus lactis* subsp*. lactis* and *Lla. cremoris* subsp*. cremoris* in each kefir. Notably, even in kefirs where *Lla. lactis* subsp*. lactis* and *Lla. cremoris* subsp*. cremoris* did not dominate, these strains and 51 associated metabolic pathways were identified across all metagenomes. Strain level phylogenetic/functional analysis provided an unprecedented opportunity to assess the genomic differences of related bacteria within the milk kefir microbiome, revealing long-term strain-type retention and the presence of multiple strain subclusters among the 12 prevalent microbial species found in kefir. The detection of core and highly prevalent compositional, strain and functional features can help shape further regulatory information concerning kefir, and highlights how insights provided by metagenomic analysis has the potential to contribute to valuable insights of relevance to regulatory practices relating to fermented foods.

Section 2.1: Introduction

Milk kefir is a fermented milk product traditionally produced by inoculating milk with a kefir grain, which is a protein and exopolysaccharide matrix containing a variety of microorganisms, including those responsible for the fermentation process [1]. The live cultures represent a mixture of microbial species, including, for example, representatives from the genera *Lactobacillus*, *Lactiplantibacillus*, *Lentilactobacillus*, *Lactococcus*, *Leuconostoc*, *Streptococcus*, *Kluyveromyces* and *Saccharomyces* [2]. Previous research efforts have verified the presence and dominance of kefiran producing bacteria, such as *Lactobacillus kefiranofaciens* and *Lentillactobacillus kefiri* within the milk kefir grain microbiome [3]. This pattern of dominance has been verified through the microbial analysis of kefir grains from different geographic regions, with heterogeneity typically arising from the low-abundance bacterial and eukaryotic communities [4, 5]. In contrast, the resulting fermented milk known as 'milk kefir' has been reported to have a more varied compositional profile compared to the grain [2] and can undergo substantial compositional changes within a few hours [5, 6]. The organoleptic and health-promoting properties of milk kefir are attributed to the food's nutritional content, the fermenting microorganisms and the by-products (including postbiotics) produced from their metabolic activities [7]. Numerous studies examining the effects of milk kefir on host health have reported positive results [2, 8]. Indeed, milk kefir consumption has been reported to reduce the symptoms of lactose intolerance in humans [9], while testing in rodent models has indicated that specific milk kefirs and/or components thereof can contribute to wound healing, reduce cholesterol, impact the gut-brain axis and demonstrate anti-carcinogenic, anti-obesity, anti-inflammatory or anti-pathogenic properties. However, in many instances, the mechanisms for these purported effects have not been elucidated [10-16]. Kefir-associated microbes can also be employed in fermented food production to improve sensory characteristics and/or enhance microbiological safety [17-19]. To further optimise milk kefir, and the combination of microbes responsible for its production, it is necessary to first understand and harness the heterogeneous microbial communities present in different beverages, which can extend to different batches of the same product [20]. This is necessary as it is also clear that at least some health benefits are not universal, but are associated with specific grains and, thus specific microbes [14], with the latter providing the justification for

isolating specific strains and creating simple 'pitched' communities that retain the key traits of the beverage, while facilitating greater quality control [21]. Another challenge relates to regulations, with inconsistencies across different jurisdictions regarding what constitutes a milk kefir and, in other cases, a lack of regulation [22], which can lead to products being marketed as kefir without having a microbial composition that reflects the artisanal beverage [23]. Metagenomics-based identification and characterisation of the microorganisms present within kefir communities of different origin has the potential to define the milk kefir pan-metagenome. According to Imchen *et al,* a pan-metagenome is "the collective study of all or several metagenomes from all possible units belonging to a particular type of ecosystem or host". [24]. Exploration of the pan-metagenome of milk kefir can provide considerable insights into the heterogeneity of milk kefir microbial populations, while also uncovering key core or conserved features. This information might also help to identify genes, and ultimately mechanisms, which contribute to health benefits. In milk kefir, sampling of both the milk kefir grain and liquid microbiome can provide insights into the pan-metagenome. However, the fermented beverage often represents a more varied compositional profile, which can include the residual milk microbes that persist in the fermentation medium, and as such represents the more informative kefir microbial community to study microbial diversity [5]. Here we describe the metagenomics-based characterisation of milk kefir generated by 64 kefir grains sourced from 25 different countries, providing a framework upon which the minimal microbial and functional criteria to define a milk product as a milk kefir can be established through the identification of a core pan-metagenome in milk kefir fermentations.

Section 2.2: Methods

2.2.1 Sample collection, preparation and pre-processing

64 milk kefir grains and two milk kefir samples were obtained from 25 different countries (Figure S1). Samples were donated or purchased from private households that consented to their use in the study. Technical details used for kefir fermentation, sample collection, DNA extraction, preparation of libraries, shotgun metagenomic sequencing and data pre-processing are described in Supplementary Methods.

2.2.2 Bioinformatic analysis

Taxonomic analysis of the kefir metagenomic data was performed by Kraken 2 [25] against a custom Kraken 2 database (see Supplementary Information: Result S2 for additional information). The relative abundance of each species was calculated using Bracken [26]. The predicted function of the metagenomic sequence data was generated using HUMAnN 3 [27]. Melonnpan v. 0.99 [28] was used to predict the metabolites produced from each kefir microbiome. The top 10 most abundant predicted metabolites detected in ≥10% of samples were then selected for downstream analysis. Metagenomic assembly was performed using MEGAHIT [29]. MetaWRAP [30] was used for genome binning, with default settings. CheckM [31] was implemented to check the quality of metagenome-assembled genomes (MAGs). Lower-quality MAGs, i.e., <80% completeness and/or >5% contamination, were removed from downstream analysis. GTDB-Tk [32] was used to assign taxonomy to the MAGs and to identify potentially new species. dRep [33] was used to cluster MAGs representing putative new species into primary and secondary clusters on the basis of their relative similarities. MAGs considered in this study were annotated with DRAM (Distilled and Refined Annotation of Metabolism) [34] using default parameters.

Strain-level assessment was performed using both PanPhlAn 3 [35] and StrainPhlAn 3 [36]. Phylograms of the StrainPhlAn 3 predicted strains were constructed from multiple sequence alignment (MSA) files of reconstructed markers, as documented in Asnicar *et al*., 2020. Metacache was used to determine the prevalence and relative abundance of sub-species [37].

VirSorter2 [38] was used to detect prophage in MAGs and contigs. A cut-off score of 0.5 (default) was selected for maximal sensitivity, and prophage less than 1,500 bp in length were discarded. VirSorter2 detections were then quality inspected using checkV to remove potential host genes bordering predicted viral sequences [39]. VirSorter2 was again applied on trimmed viral sequences that passed the quality inspection of checkV. Viral sequences obtained through the combined outputs of VirSorter2 and checkV were annotated with the viral mode of DRAM [34]. For all tools applied (see above), default parameters were used unless specified otherwise. The downstream analysis of data generated by each tool applied was completed in R-4.0.2 [40].

2.2.3 Statistical analysis

Statistical analysis of compositional and metabolic data was carried out with vegan v2.6.2 [41], to compute alpha and beta diversity values. The stats v4.1.2 kruskal.test function was used to perform the Kruskal-Wallis rank sum test, to identify significant differences in alpha diversity values. The adonis function in the vegan package [41], was used for permutational analysis of variance (PERMANOVA). PERMDISP [42] was used to test the null hypothesis of no difference between groups dispersion using vegan v2.6.2. The linear discriminant analysis (LDA) effect size (LEfSe) [43] was used to identify differentially abundant species and metabolic pathways. Genetic distances between strains were calculated through the dist.DNA function in the package ape v 5.6.2 [44], using the p-distance, Kimura's 2-parameter evolutionary models and the Jaccard coefficient for PanPhlAn 3 predicted strains. The genetic distances of strains were initially assessed by the Hopkins statistic using the get_clust_tendency function in the package factoextra v1.0.7 [45]. The fviz_nbclust function included in factoextra v1.0.7 was used to predict the optimal number of clusters using the average silhouette method. Cluster validation was performed using the Silhouette coefficient through the fviz_silhouette function in factoextra v1.0.7. Correlation between genetic distances of PanPhlAn 3 and StrainPhlAn 3 predicted strains was calculated using the Kendall correlation method [46] through the ggscatter function included in ggpubr v0.4.0. Clustering analysis of StrainPhlAn 3 predicted strains was performed using the PAM function of the cluster package v2.1.3 and hierarchical clustering was performed on PanPhlAn 3 predicted strains using the Ward criterion (ward.D) through the stats v4.1.2 hclust function. The species *Propionibacterium freudenreichii* was excluded for clustering due to low sample size. To evaluate if subclusters differed statistically in their gene content, the Mann-Whitney-Wilcoxon test or the Kruskal-Wallis H test was performed. The Mann-Whitney-Wilcoxon tests were performed using the stats v4.1.2 wilcox.test function, the Kruskal-Wallis H tests was performed using the stats v4.1.2 kruskal.test function. The resulting p-values from both functions were corrected for multiple comparisons, using the Bonferroni adjustment method through the stats v4.1.2 p.adjust function.

2.2.4 Data visualisation

The mapBubbles function in the package rworldmap v1.3.6 [47], was used to produce a worldmap of the kefir grains analysed (Figure S1).The experimental design figure was produced using BioRender [48] (Figure S2). Ordination plots, correlation scatter plots, bar charts and Venn diagrams were produced using ggplot2 v3.3.6 [49]. Heat maps were generated using both ggplot2 and pheatmap v1.0.12 [50]. ggtree v3.2.1 [51] was used to visualise phylogenetic trees from the StrainPhlAn 3 and GTDB-Tk results. The taxize v0.9.100 package [52] was used to construct a cladogram from the custom Kraken 2 outputs. The resulting cladogram was visualised using the ggtree v3.2.1 package.

Section 2.3: Results

2.3.1 Taxonomic profiling reveals distinct community types in the kefir microbiome

64 individual milk kefir grain samples sourced from 25 different countries (Figure S1) were used to carry out fermentations, in duplicate, to produce milk kefir from full fat pasteurised cow's milk, with samples being collected after 8 and 24 hr of fermentation for shotgun metagenomic analysis (Figure S2); yielding 108.7 GB of data (see Supplementary information: Result S1, Figure S2). The short read taxonomic profiler, Kraken2 [25], was employed to determine the microbiome composition using a custom built Kraken2 database (see Supplementary Information: Result S2 for additional information). This identified 46 species of viral, eukaryotic and bacterial origin at different levels of relative abundance above a 1% threshold in the kefir metagenomics dataset (Figure 1).

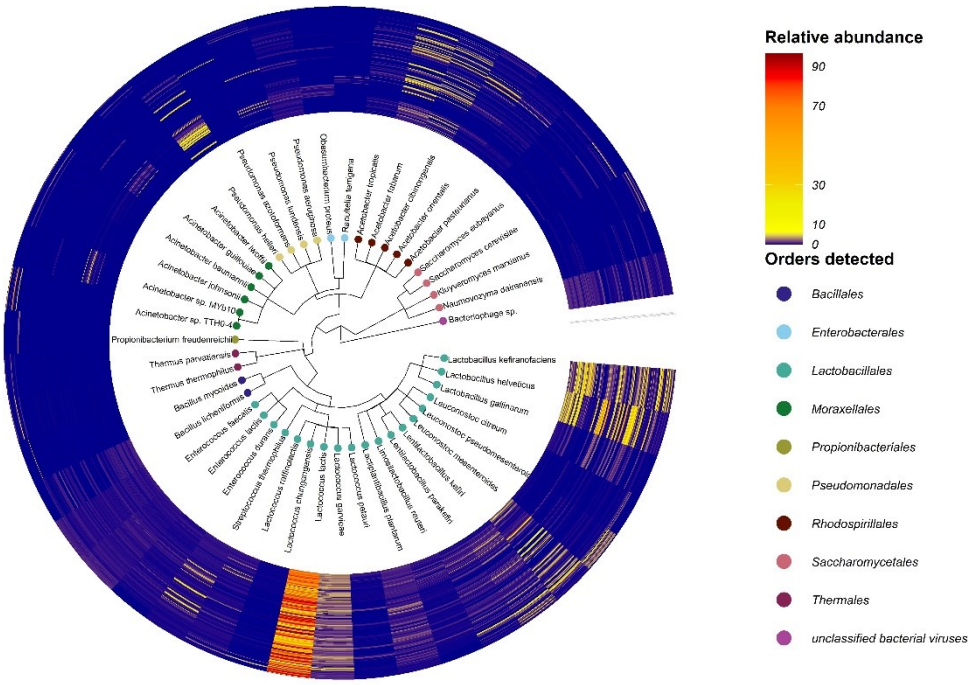


Figure 1: Species and their relative abundance across milk kefir metagenomes. Cladogram presenting a hierarchical overview of species detected in the kefir microbiome using the custom kraken2 pipeline. Tips represent the species detected, e.g., *Lla. lactis*, and the associated colour of the tips describes the order of the species *Bacillales*, *Enterobacterales*, *Lactobacillales*, *Propionibacteriales*, *Pseudomonadales*, *Rhodospirillales*, *Saccharomycetales,* *Thermales* and Unclassified bacteriophage. The outer rings represent a circular heatmap displaying the relative abundance of each species across 256 metagenomes.

Of the 46 species, *Thermus thermophilus* and *Pseudomonas lundensis* were consistently detected in the control, unfermented milk samples, and thus not regarded as kefir-associated. The yeast species *Saccharomyces eubayanus, Kluyveromyces marxianus* and *Saccharomyces cerevisiae*, and combined bacteriophage were detected at low relative abundance (<2%) (Figure 1). *Acetobacter cibinongensis, Acetobacter fabarum, Lactiplantibacillus plantarum, Lactobacillus gallinarum, Lactobacillus helveticus, Lactobacillus kefiranofaciens, Lactococcus garvieae, Lactococcus lactis, Lentilactobacillus kefiri, Lentilactobacillus parakefiri, Leuconostoc mesenteroides* and *Pseudomonas helleri* were detected at ≥1% relative abundance in ≥10% of samples (the defined threshold of prevalence for this study) (Figure 1). *Lla. lactis* was detected in 253 of the 256 samples at varying levels of relative abundance (2-96%). The three

samples in which *Lla. lactis* was not detected at ≥1% relative abundance had been collected after 8 hr of fermentation. Even in these instances, *Lla. lactis* accounted for sizeable proportions of the relative abundance of the corresponding 24 hr samples (16-78%) (Figure 1).

Overall, the kefir metagenomes were consistently dominated by one bacterial species (Figure 1 and Figure 2). This species was *Lla. lactis* in 215 samples (ranging in abundance values from 24.8-96.5%), *L. helveticus* in 19 samples (30.9-77%), *L. kefiranofaciens* in 14 samples (28-57.6%), *Acetobacter orientalis* in 3 samples (41.5-65.6%), *Acinetobacter* sp. TTH0-4 in 2 samples (39.3-58.9%) and *Leuc. mesenteroides* in one sample (84.6%) (Figure 2A). There were no major changes in patterns of dominance between the 8 and 24 hr time points for specific kefir fermentations in the majority of cases (i.e., n=104 fermentations corresponding to 52 different starter grains), but notable changes were observed in 24 metagenomes corresponding to 12 grains. LEfSe analysis revealed that *L. kefiranofaciens, Len. kefiri* and *L. helveticus* was more abundant at 8 hr whereas *Lla. lactis* was more abundant at 24 hr (See Supplementary Information: Result S3, Figure S3).

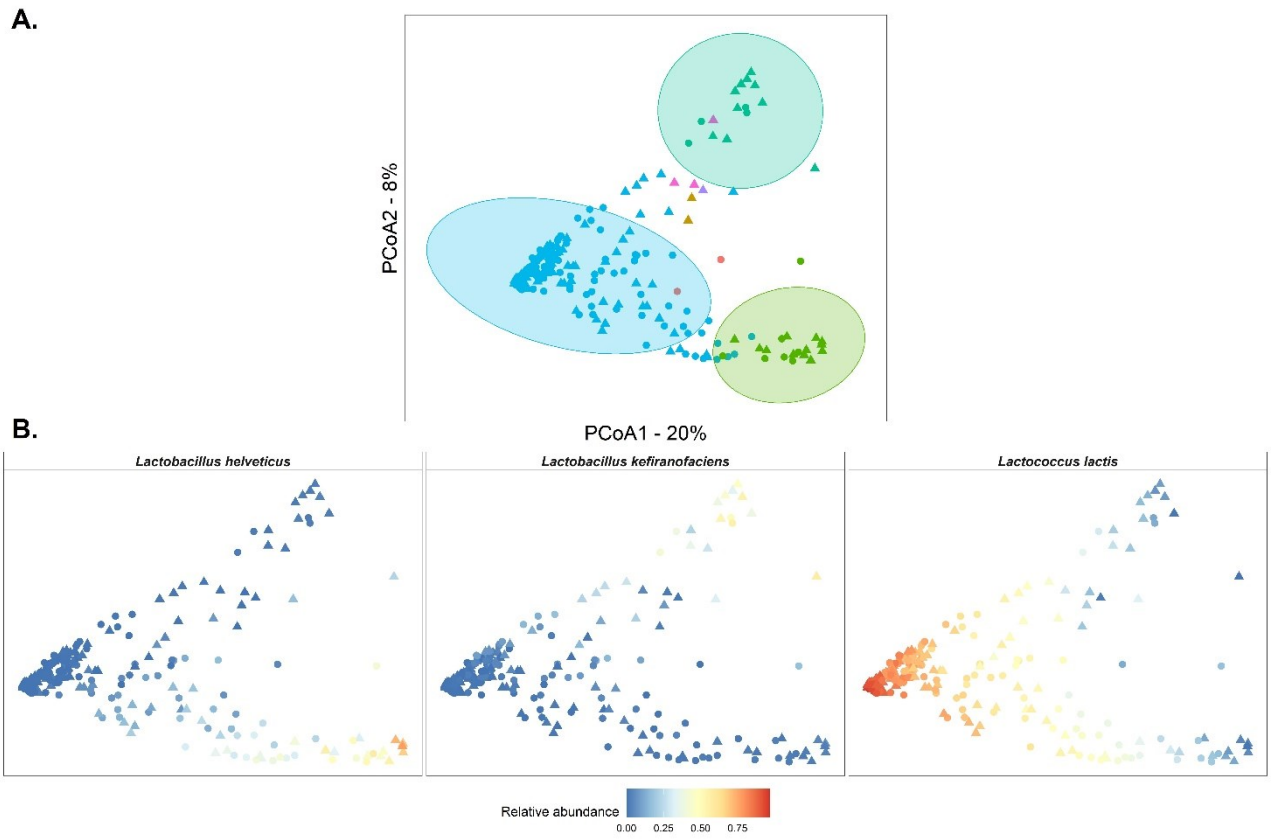


Figure 2: Variance in the compositional data among kefir metagenomes explained by the dominating microbial species. **(A.)** Principal Coordinate analysis (PCoA) of beta diversity measured by Bray-Curtis dissimilarity of 256 kefir metagenomes, calculated for species-level composition. The PCoA plot displays the microbiome differences between kefir metagenomes dominated by different microbial species. Metagenomes are coloured by the dominant species ■ *A. orientalis,* ■ *Ac. sp.* TTHO-4, ■ *L. helveticus*, ■ *L. kefiranofaciens,* ■ *Lla. lactis,* ■ *Leuc. mesenteroides* and ■ *P. helleri*. Metagenomes are shaped according to time point (▲ 8 hr or ● 24 hr). The 95% confidence ellipse is shown for metagenomes dominated by ■ *Lla. lactis,* ■ *L. helveticus* and ■ *L. kefiranofaciens*. **(B.)** PCoA of beta diversity separated by the dominant species *L. helveticus*, *L. kefiranofaciens* and *Lla. lactis.* Metagenomes are coloured by the relative abundance of the dominant species.

The permutational analysis of variance (PERMANOVA) statistical test revealed significant dissimilarity between kefir metagenomes, with the dominant microbial species accounting for the highest fraction of total variance (PERMANOVA: R2=0.36, p= <0.001). Pairwise PERMANOVA further confirmed significant differences between kefir metagenomes and revealed specific groupings that were significantly dissimilar, which included metagenomes dominated by *Lla. lactis*, *L. kefiranofaciens* and *L. helveticus* (Figure 2A). There were not enough metagenomes dominated by other prevalent microorganisms to be statistically regarded as additional subgroups. No dissimilarities were apparent when the influence of fermentation time point (Figure 2) or country of origin were considered.

2.3.2 Phylogenetic analysis of assembled metagenomes identifies three putatively novel species in the kefir microbiome.

In total, 718 MAGs with an estimated completion of ≥50% and ≤10% contamination were generated, and assigned to 18 different species (Figure S4). Of these, 613 were high quality MAGs (over 80% complete, with less than 5% contamination) and corresponded to the species *A. fabarum*, *A. orientalis*, *Ac. guillouiae*, *Ac. albensis, L. helveticus*, *L. kefiranofaciens*, *Len. kefiri*, *Lla. lactis*, *Lla. raffinolactis*, *Leuc. mesenteroides*, *Leuc. pseudomesenteroides*, *P. helleri*, *P. azotoformans* and *S. thermophilus*. 41 MAGs had less than 95% average nucleotide identity (ANI) to the closest known NCBI prokaryote reference genome and thus may represent novel species. 39 of these MAGs are predicted to be representatives of the genus *Acetobacter*, with two belonging to the genus *Lactobacillus* (Figure 3). De-replication of the putatively novel species revealed that the MAGs assigned to the genus *Acetobacter* grouped together into the same species level bin, while the MAGs assigned to the genus *Lactobacillus* exhibited greater differences at the nucleotide level and grouped into separate clusters (Figure 3).

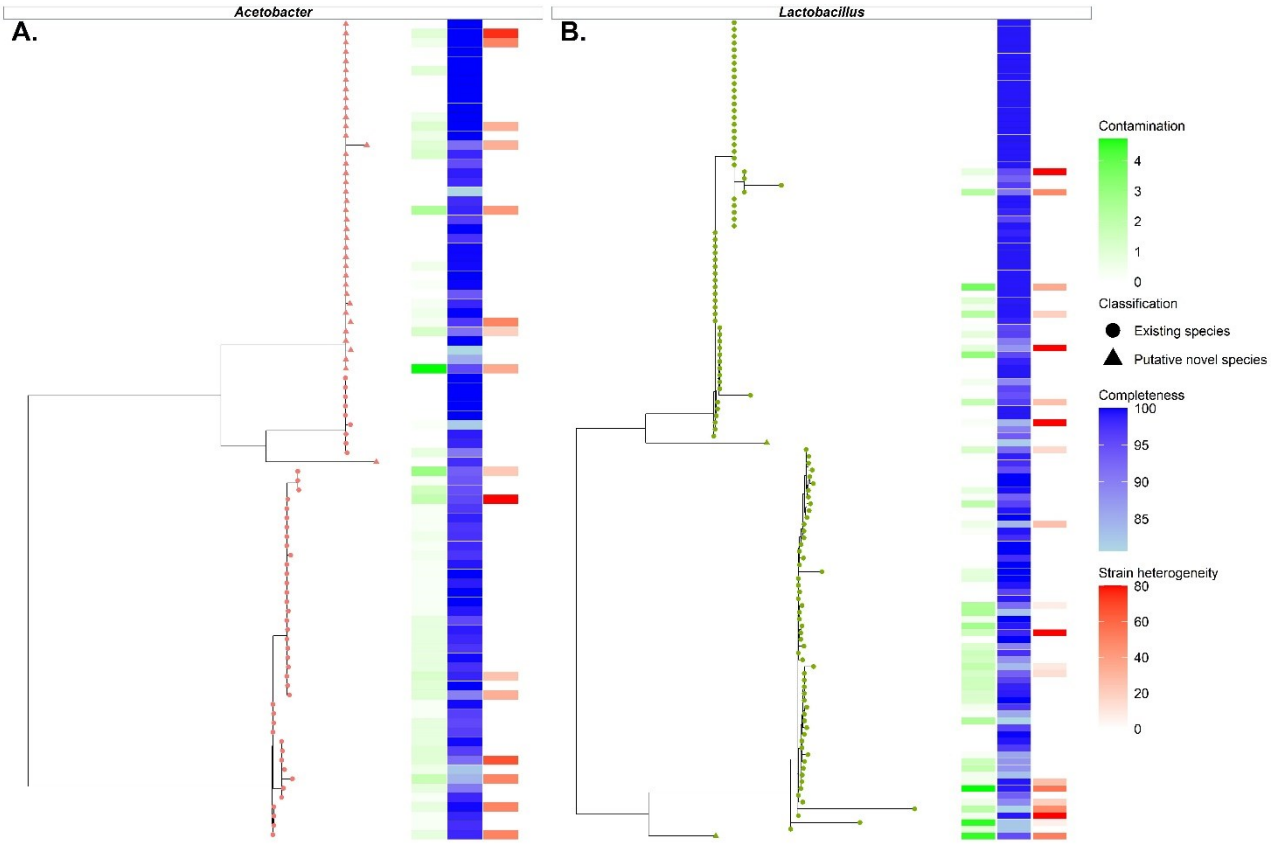


Figure 3: Putative new species of the genus *Acetobacter* and *Lactobacillus* within the kefir microbiome.Phylogenetic tree of high quality MAGs of the genus (**A.** and **B.**) Phylogenetic tree of high-quality MAGs of the genus **(A.)** *Acetobacter* and **(B.)** *Lactobacillus*, including putative new species. Outer band coloured green, blue and red represent % contamination, completion and strain heterogeneity, respectively, as assessed using checkM. The coloured tip of the phylogram corresponds to the genus classification of the MAG. Genus classifications include ■ *Acetobacter* and ■ *Lactobacillus*. The tip shape of the phylogram indicates which MAGs are known species (●) or potentially novel species (▲).

2.3.3 Strain level phylogenetic analysis of the kefir microbiome uncovers strain subclusters and long term strain-type retention

StrainPhlAn 3 was chosen as the primary strain profiling tool in this study due to its sensitivity. StrainPhlAn 3 evaluates the alignment of marker genes to reference genomes to capture a sufficient portion of the reference genomes for variant calling [53, 54]. StrainPhlAn additionally reports the number of polymorphic sites across alignments. Sites are deemed polymorphic if they display a certain degree of variability in the base call that the dominant allele accounts for less than 80% of reads [53]. StrainPhlAn 3 profiling of the 256-kefir metagenomes detected a total of 861 strains that represented 15 different species of bacterial origin and 3 different species of lactococcal phage (Figure 4 and Figure S5-20).

Intrasample strain diversity was evaluated across metagenomes, which established that >0.1 of nucleotides on detected marker genes for the examined species were polymorphic in 95% of metagenomes [36, 53], consistent with the presence of multiple strains for the considered species across metagenomes. Examination of the strains of the species *A. fabarum*, *A. orientalis*, *Ac. guillouiae*, *E. faecalis*, *L. helveticus*, *L. kefiranofaciens*, *Lla. chungangensis*, *Leuc. mesenteroides*, *Leuc. pseudomesenteroides*, *Leuc. suionicum*, *Pr. freudenreichii*, *P. helleri* and *S. thermophilus* revealed a percentage of polymorphic sites <2% suggesting the dominance of a single strain within each individual kefir metagenome. While strains of *Lla. lactis* (Figure 4) and *P. lundensis* (Figure S17) had a median percentage of polymorphic sites <2% in many cases, some specific metagenomes contained strains with a greater number of polymorphic rates, suggesting the possibility of multiple strains without single strain dominance in these metagenomes (Figure 5A). Despite these exceptions, in general, the majority of the strain level diversity of a kefir metagenome is captured when one profiles the dominant strain present from each species.

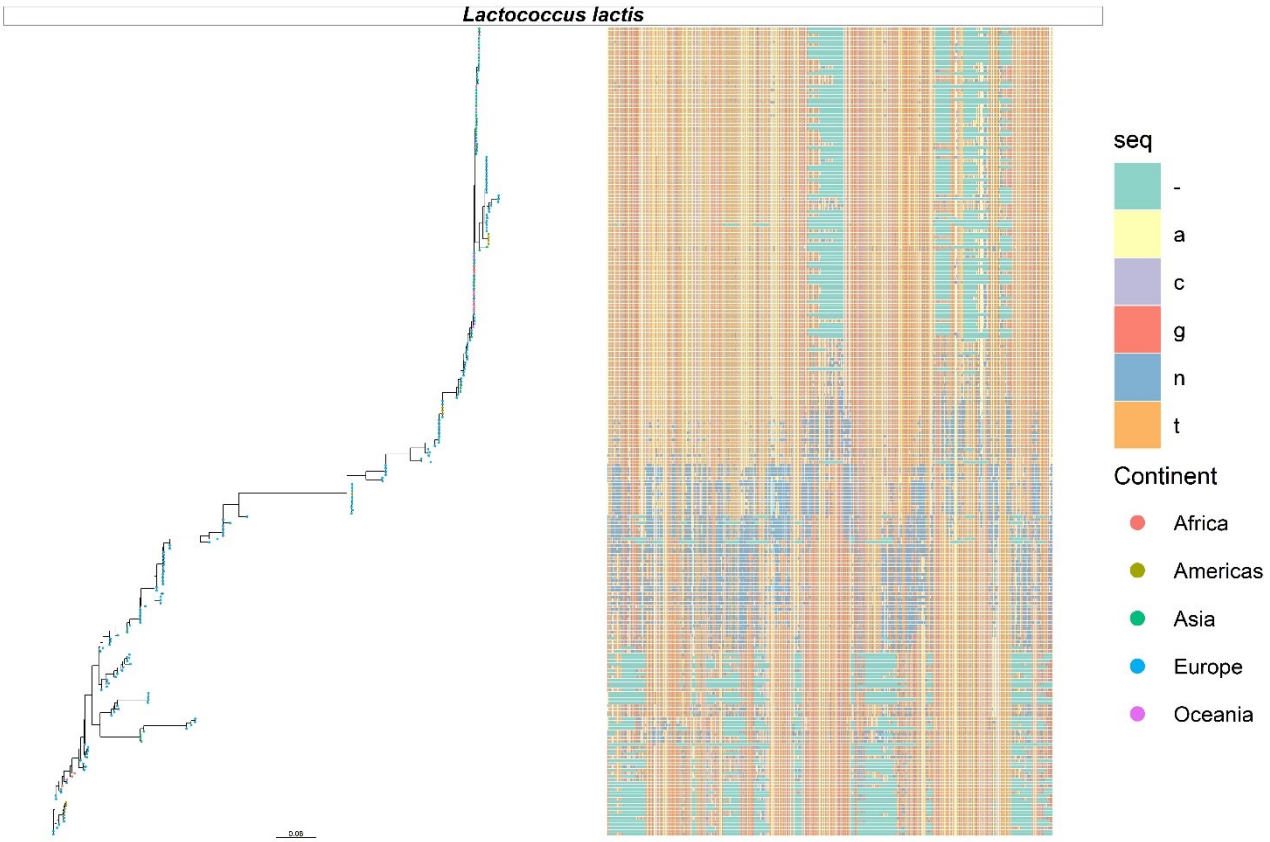


Figure 4: Strain level phylograms of *Lla. lactis* and corresponding multiple sequence alignment (MSA) files generated by StrainPhlAn 3. Colored tips of the phylogram corresponds to the time point of the kefir metagenome from which the strain was recovered. The MSAs contains completely specified nucleotides (A, C, G, T), ambiguous characters (N) and alignment gaps (–) inserted between the nucleotides of some of the sequences Other representative species

phylograms are available in Figures S5–S20.

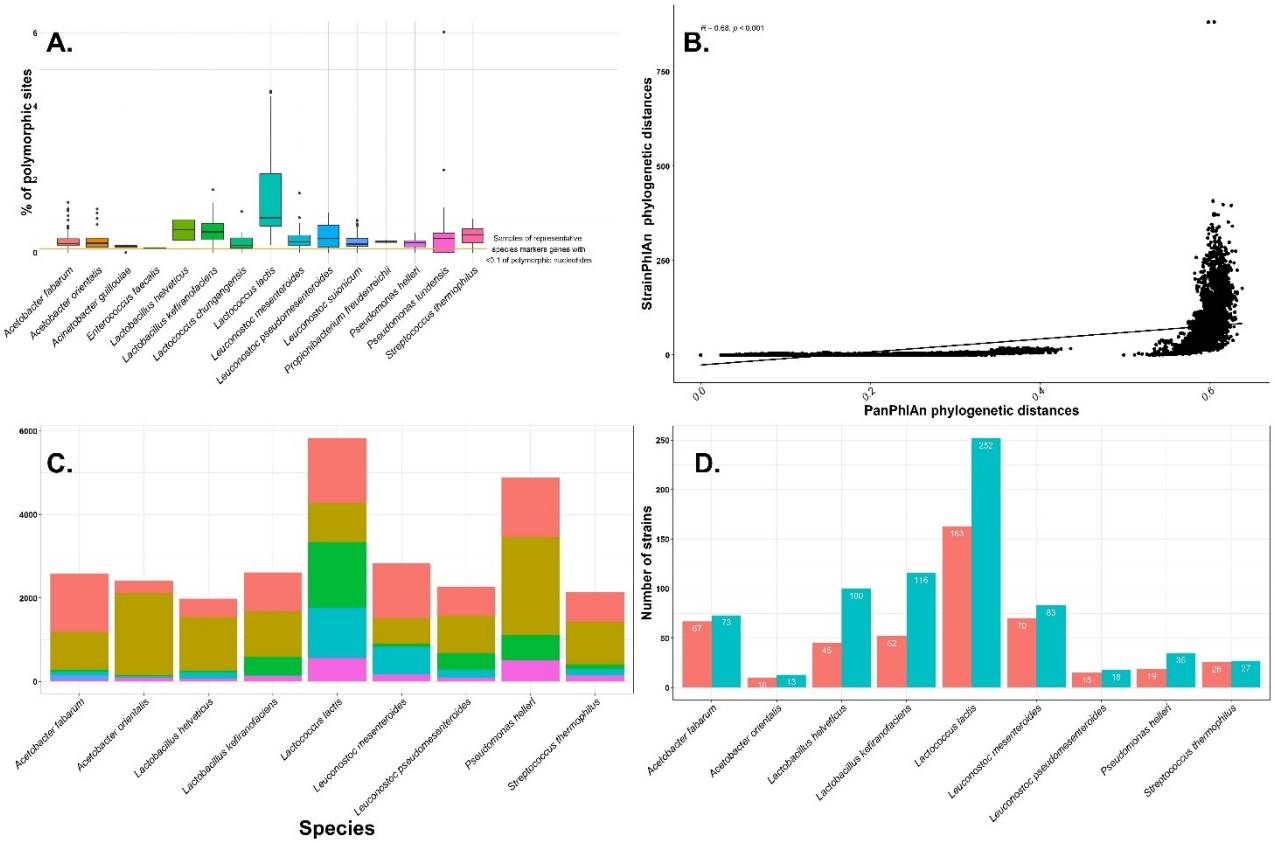


Figure 5: Analysis of representative strains, with joint identification by StrainPhlAn 3 and PanPhlAn 3. **(A.)** Collective intra-sample strain diversity across metagenomes for each of the outlined 15 bacterial species (X-axis). Each coloured boxplot displays the range of polymorphic sites expressed as percentages (Y-axis); along the MLSA of strains. **(B.)** Results of correlation analysis performed using the genetic distance between StrainPhlAn 3 (Y-axis) and PanPhlAn 3 (X-axis) predicted strains calculated using the Kendall correlation method. (**C.)** Graphical presentation of PanPhlAn 3 predicted gene families (Y-axis) that compose the pangenome of the representative species (X-axis). Details include number of gene families detected per pangenome and the number of gene families classified as accessory, core, specific to cluster 1, specific to cluster 2, specific to cluster 3 and strain specific. (**D.)** Number of strains detected (Y-axis) for each representative species (X-axis) by the StrainPhlAn 3 or PanPhlAn 3 approach. Only representative species, with joint identification by StrainPhlAn 3 and PanPhlAn 3 were considered in this graph.

As the metagenomics data collected was temporally ordered, an assessment of the succession patterns of strains was possible. Strains recovered from metagenomes produced from the same grain were compared using the p-distance evolutionary model to assess if the same strains were retained, or if lower abundance strains became dominant as the fermentation process progressed. Strains of *Pr. freudenreichii* derived from the same kefir grain varied substantially in nucleotide composition, differing at 40-60% of sites, indicating a strain succession event (Figure S17) may have occurred in the initial fermentations. No nucleotide

differences were observed between strains from other species, including those derived from metagenomes with a higher rate of polymorphic sites, indicating that the strain of that species that dominated at 8 hr continued to do so at 24 hr.

In addition to comparing strains derived from the same kefir grain to uncover patterns of short term strain succession/retention, we also compared the phylogenetic distance between strains of the same species across all metagenomes using the Kimura 2-parameter distance evolutionary model. Firstly when considering variability, we observed large genetic distances between a number of strains of the species *Lla. lactis* (Figure 4), which were not observed for strains of the other, less abundant species examined. The large genetic distance between the *Lla. lactis* strains was further assessed using the p-distance evolutionary DNA model, which reported that a number of strains differed by ≥70% of sites, hindering downstream clustering analysis at the nucleotide level (Figure 4). For other species, PAM clustering analysis was used to identify the presence/absence of strain subclusters (Figures S5–S20). 2-4 distinct subclusters were predicted for the bacterial species *A. fabarum* (Figure S5), *A. orientalis* (Figure S6), *L. helveticus* (Figure S8), *L. kefiranofaciens* (Figure S9), *Leuc. mesenteroides* (Figure S14), *Leuc. pseudomesenteroides* (Figure S15), *Leuc. suionicum* (Figure S16), *P. helleri* (Figure S18) and *S. thermophilus* (Figure S20). Subclusters of these species contained strains with identical nucleotide composition along shared sites in the multiple sequence alignment and therefore were inferred to be of the same strain-type. Specifically, subcluster - 1 and - 2 of *A. fabarum* (Figure S5) and *Leuc. suionicum* (Figure S16), 1, 2 and 4 of *L. helveticus* (Figure S8), subcluster - 2 of *A. orientalis* (Figure S6) and *S. thermophilus* (Figure S20), subcluster - 1 of *L. kefiranofaciens* (Figure S9) and *Leuc. mesenteroides* (Figure S14) contained the same strain-types, despite the strains being derived from kefir grains sourced from different geographical locations.

To further investigate the variability between strains, we assessed the functional potential of strains through the application of a meta-pangenomic approach using PanPhlAn 3. PanPhlAn 3 evaluates the alignment of metagenomics reads to reference species pangenomes to capture genes and coverage information used to determine gene family presence/absence profiles per strain. PanPhlAn 3 was used to confirm and expand on information relating to subclusters predicted using the more sensitive StrainPhlAn 3 approach [53]. This

methodology facilitated the construction of pangenomes for the representative bacterial species *A. fabarum*, *A. orientalis*, *L. helveticus*, *L. kefiranofaciens*, *Lla. lactis*, *Leuc. mesenteroides*, *Leuc. pseudomesenteroides*, *P. helleri* and *S. thermophilus* (Figure 5B, 6 and Figures S21-28).

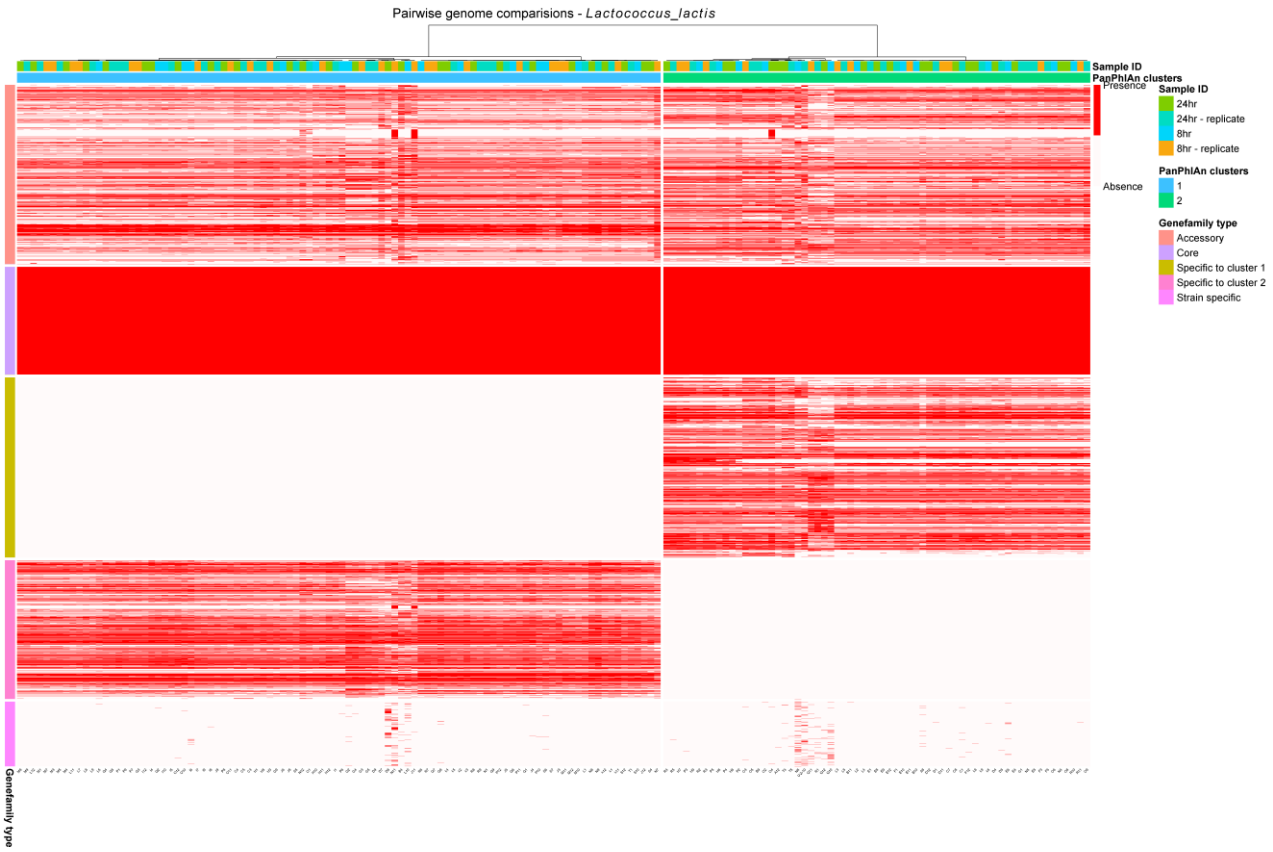


Figure 6: Pangenome of *Lla. lactis* strains detected within the milk kefir microbiome: Representative heatmap representing the pangenome of *Lla. lactis* strains*;* rows represent PanPhlAn 3 detected strains and columns represent the UniRef.90 gene families identified per strain. Row side annotations display the grouping of each gene family, e.g., core gene or specific to cluster 1. Column side annotations display the assigned cluster of the detected strain according to the PanPhlAn 3 hierarchical clustering methodology (see Methods) and the time point that the strain was recovered from.

Gene families within pangenomes could be grouped into three categories; 'core' if they were consistently recovered across strains, 'accessory' if they were identified in more than one strain and 'strain specific' if they were identified solely in one strain (Figure 5B, Figure 6, Figure S21-28). Correlation analysis was performed using the genetic distance between StrainPhlAn 3 and PanPhlAn 3 strains. Positive linear correlations of medium to very high strength were consistently observed for strains for each of the nine representative species (Figure 5C) detected using both StrainPhlAn 3 and PanPhlAn 3. PanPhlAn 3 predicted

strains were clustered using hierarchical clustering (see Methods). Functional differences and hierarchical clustering between strains of *Lla. lactis* suggested the presence of two distinct subclusters. Hierarchical clustering typically agreed with the results of the PAM clustering methodology; however exact matching was not possible due to the detection of a different number of strains by both approaches (Figure 5D). As expected StrainPhlAn 3 consistently detected a higher number of strains (Figure 5D) than PanPhlAn 3, which in a small number of cases such as with respect to *L. helveticus* (45 vs 100), prevented the confirmation and further functional analysis of subclusters 3 and 4. Similar results were observed for subclusters of *L. kefiranofaciens* and *A. orientalis* strains. As clustering results predicted using the PAM clustering methodology were mostly replicated using hierarchical clustering, we could further identify gene families unique to each subcluster and determine the proportions of the pan genome that these subclusters accounted for. Statistically significant functional differences between clusters was observed for the species *A. fabarum, L. helveticus, Lla. lactis, Leuc. mesenteroides, Leuc. pseudomesenteroides*, *P.* helleri and S. *thermophilus*, as determined using either the Mann-Whitney U test to compare two clusters, or the Kruskal-Wallis test to compare three clusters. Further assessment of the gene families unique to each subcluster revealed slight functional differences between subclusters of the species *A. fabarum* (Figure S21), *A. orientalis* (Figure S23), *L. helveticus* (Figure S23) and *S. thermophilus* (Figure S28) with gene families unique to each cluster accounting for less than 10% of the species pangenome (Figure 5B). For all other species, sizeable proportions of the pangenome were attributed to clusters (Figure 5B, Figure S24-S27). The *Lla. lactis* pangenome displayed the most variability, with 27% and 21% of the pangenome being assigned to cluster one and two, respectively (Figure 5B, Figure 6), an observation that was consistent with the results of the StrainPhlAn 3 analysis. Gene family differences between clusters of *Leuc. mesenteroides* and *P. helleri* were consistent with the absence of gene families from one of the two clusters (Figure 5B, Figure S25 and Figure S27).

2.3.4 The metabolic potential of kefir metagenomes separates by community type

As well as being used to investigate differences in the functional potential across strain-types, we further assessed functional potential across kefir metagenomes using HUMAnN 3.0, which led to the identification of 402 diverse metabolic pathways in the kefir microbiome. Pathways were grouped into 36 higher level classifications based MetaCyc (Metabolic Pathway Database) [55]. Many pathways were involved in biosynthesis, specifically for glycerol biosynthesis, glucose, ADP-L-glycero-beta-D-manno-heptose, chitin, starch, UDP-N-acetyl-D-galactosamine, CMP-3-deoxy-D-manno-octulosonate, colanic acid, mannose and UDP-N-acetyl-D-glucosamine. Fewer pathways were attributed to degradation. The carbohydrate degradation pathways detected suggest that the carbohydrates chitin, fructose, fucose, galactose, glucose, lactose, stachyose, starch, sucrose, trehalose and xylose can be used by some kefir-associated microorganisms. Pathways involved in amino acid metabolism indicate that the amino acids L-glutamate, L-arginine, L-histidine, L-leucine, L-phenylalanine, L-tryptophan, L-valine, putrescine, 4-aminobutanoate and L-ornithine may be utilised during fermentation. The number of metabolic pathways detected per sample ranged from 79 (M11) to 282 (F8) with a median value of 163 pathways. Pathways were further grouped based on their presence across samples into the following 6 categories 'core' (n=51) if they were present in 100% of metagenomes, 'high' (n=105) if they were present in 70% but less than 100% of metagenomes, 'moderate' (n=39) if they were present in 50% but less than 70%, of metagenomes, 'low' (n=185) if they were present in multiple but less than 50% of metagenomes and 'sample specific' (n=21). 51 pathways were deemed core pathways, as they were present in all kefir metagenomes. Notably, all core pathways detected could be attributed to the presence of *Lla. lactis*, as estimated using the per species metabolic pathway abundance output of HUMAnN 3.0 [53] (Figure S29). Additionally the relative abundance attributed to *Lla. lactis* typically accounted for the majority of the observed relative abundance per core pathway among metagenomes. Core pathways principally involved generic housekeeping functions such as carbohydrate metabolism and biosynthesis of amino acids (Table S6). Melonnpan [28] was also used to predict the metabolites potentially produced by each kefir metagenome using the UniRef90 genefamiles identified by

HUMAnN 3.0. The output was restricted to the top 10 most abundant metabolites in ≥10% of samples to ensure high accuracy. The top 10 metabolites predicted ordered according to their maximum relative abundance included the secondary bile acid, deoxycholic acid, glutamate, chenodeoxycholate and deoxycholate, cholate, xanthine, lithocholate, butyrate and isobutyrate, nicotinic acid, cholestenone and lithocholic acid.

Insights provided by strain subclustering together with species level taxonomic profiles were combined and utilised to provide insights into patterns of dissimilarity within the metabolic potential data among kefir metagenomes. Consistent with the dissimilar patterns identified in the composition data, we observed significant dissimilarity between the metabolic potential of kefir metagenomes, which can be explained when considering the dominant microbial species/strain type. Ordination clearly displayed the grouping of metagenomes by dominating microbial species/strain subclusters and relative abundance, in a pattern suggesting four distinct community types, extending the result obtained when considering the compositional data (see Result section: 2.3.1). PERMANOVA (R2=0.5, p= 0.001) and Pairwise PERMANOVA further confirmed the significant differences between kefir metagenomes when considering the dominant microbial species/strain type (Figure 7).

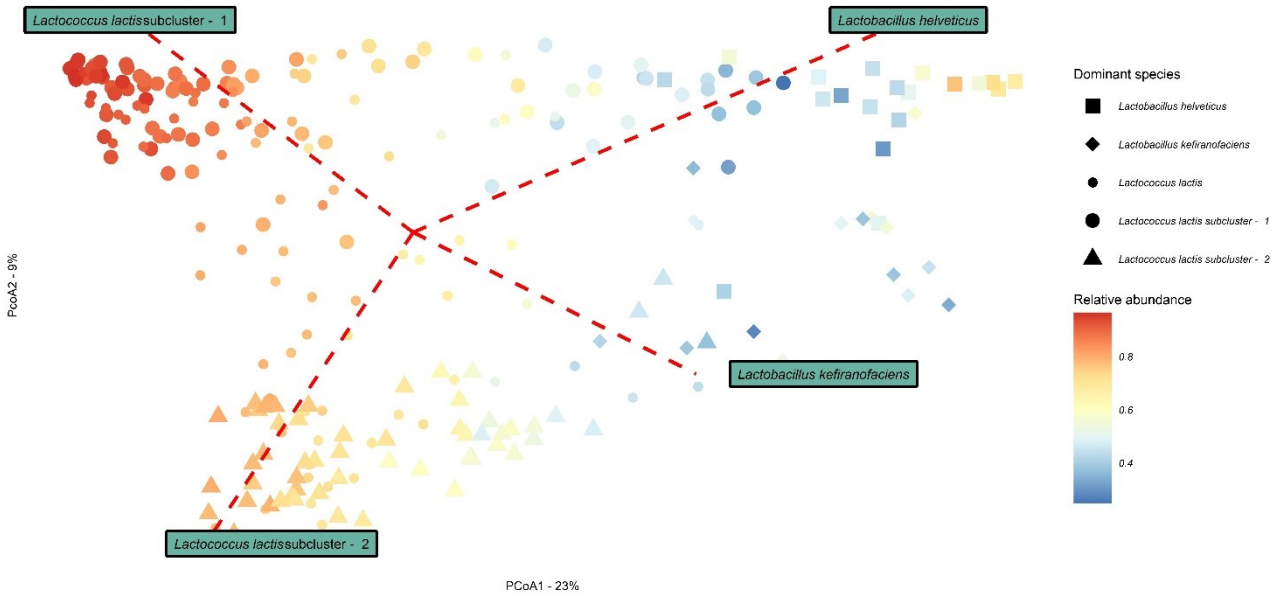


Figure 7: Variance in the metabolic potential among kefir metagenomes explained by the abundances of species/strains types. Principal Coordinate analysis (PCoA) of beta diversity measure by Bray-Curtis dissimilarity for 256 kefir samples, calculated for metabolic composition as predicted using HUMAnN 3.0. Samples are shaped according the dominant species/strain subcluster (● *Lla. lactis*, ● *Lla. lactis* subcluster - 1, ▲ *Lla. lactis* subcluster - 2, ■ *L. helveticus* or ◆ *L. kefiranofaciens*). ● *Lla. lactis* represent kefir metagenomes in which compliementary strain profiling data was absent. Samples are coloured by the relative abundance of the dominant species/strain subcluster.

2.3.5 Functional comparisons of the representatives of each community- type reveals distinguishing metabolic and functional features

Further evidence of separation in metabolic potential between community types was provided by the identification of unique and differentially abundant metabolic pathways. LEfSe detected differentially abundant pathways between the community types, *Lla. lactis* subcluster - 1 (n= 24), *Lla. lactis* subcluster - 2 (n= 40), *L. helveticus* (n= 75) and *L. kefiranofaciens* (n= 80). When considering the presence/absence of metabolic pathways across community types, all community types shared a collective of 222 metabolic pathways, while unique metabolic pathways were also observed in the community types *Lla. lactis* subcluster - 1 (n= 35), *Lla. lactis* subcluster - 2 (n= 16) and *L. helveticus* (n= 3). Differentially abundant pathways detected in the *Lla. lactis* subcluster - 1 community type included lactose and galactose degradation I, numerous vitamin biosynthesis pathways such as menaquinol (vitamin K-2),

tetrahydrofolate (vitamin B9). Notably of the 24 differentially abundant pathways 5 of these produce tetrahydrofolate as the final product. Additionally a number of amino acid biosynthesis pathways were enriched in this community type such as the superpathway of aromatic amino acid biosynthesis and other pathways involved in the production of the amino acids L-methionine, L-isoleucine and seleno amino acids. Similarly the *Lla. lactis* subcluster - 2 community type included multiple differentially abundant biosynthesis pathways for the amino acids L-arginine, L-alanine , L-proline, L-isoleucine, L-valine, L-serine, L-threonine, L-homoserine, L-histidine and L-ornithine. Of note multiple pathways for glycolysis and L-arginine were detected for this community type. The *L. helveticus* community type contained differentially abundant pathways for *inosline,* sugar and galactose degradation, and pathways for L-rhamnose, inosine purines, L-lysine, L-glutamine, L-phenylalanine, L-tyrosine and L-tryptophan biosynthesis. Lastly *the L. kefiranofaciens* community type, contained differentially abundant biosynthesis pathways for *ubiquinol, fat*ty acids, thiamine and the amino acids L- lysine, L-citrulline, L-arginine, L-threonine and L-methionine, L-ornithine.

Within-sample diversity at the species and metabolic level were compared across samples representing the four distinct community types using alpha diversity measures (Figure 8A-8B). Samples corresponding to *Lla. lactis* subcluster - 1 exhibit significantly lower species alpha diversity compared to the other community types (p= <0.001) (Kruskal-Wallis multiple comparison). *L. helveticus* and *Lla. lactis* subcluster - 2 dominated samples exhibit significantly higher alpha diversity than the *L. kefiranofaciens*-dominated samples (p= <0.001) and *Lla. lactis* subcluster - 1 dominated samples (p= <0.001) (Kruskal-Wallis multiple comparison) (Figure 8A). When considering metabolic diversity, samples corresponding to *Lla. lactis* subcluster - 1 exhibit significantly lower alpha diversity compared to the *L. kefiranofaciens*-dominated samples (p= <0.001) (Kruskal-Wallis multiple comparison) (Figure 9B). As many of the differentially abundant pathways identified in each community type related to metabolisms we selected high quality MAGs, representing the dominant microbial species in each community type, to compare functional annotations encoding metabolic traits. Selection was based on superior quality scores and included >80% completion, <5% contamination and 0% strain heterogeneity as assessed using checkm. This revealed the potential capacity of MAGs representing the dominant microbial species in each community types to produce alcohol and produce the short chain

fatty acid (SCFA) acetate. L-lactate production was prevalent across MAGs with only one exception occurring in the *Lla. lactis* subclusters - 1. *Lla. lactis* subclusters were further able to metabolise xylans, mixed-linkage glucans, cellulose, and arabinan (Figure 8C). *Lla. lactis* subclusters also contained annotations supporting the complete biosynthetic pathways of the amino acids cysteine, valine, isoleucine, leucine, arginine, proline, 2-oxoisocaproate and histidine. The *L. helveticus* and *L. kefiranofaciens* representatives, contained no biosynthetic pathways and few CAZY annotations, with annotations supporting arabinose cleavage in both sets of MAGs and chitin degradation in *L. kefiranofaciens* (Figure 8C). A number of genes were highly prevalent and solely identified in MAGs of *Lla. lactis* subclusters - 1, which encoded for components of the pathways Puu, reported for the production of the neurotransmitter γ-aminobutyric acid (GABA) [56]. Specifically, the identified genes included the *puuD* gene encoding for a γ-Glu-GABA hydrolase, which plays a role in the Puu pathway by converting γ-Glu-GABA into GABA [57] and the genes *gadB* and *gadC* encoding for glutamate decarboxylases (GadB) and Glu/GABA antiporter, respectively (Figure 4). The glutamic acid decarboxylase (GAD) system provides a full mechanism by which *Lla. lactis* subclusters - 1 can produce GABA. Additional genes encoding for the arginine deiminase system were further detected. Genes supporting the exopolysaccharide production protein ExoZ and ExoY, were commonly detected in MAGs of *L. helveticus* and *L. kefiranofaciens*, and undetected across MAGs of *Lla. lactis* subclusters, except for the detection of genes supporting ExoY in one representative MAG of *Lla. lactis* subclusters - 2.

Finally, we have harnessed previous research by Walsh *et al* 2016, whereby paired metagenomics and metabolomics data was used to reveal strong associations between the production of specific metabolites with particular kefir-associated species. More specifically, here we examined our data to identify metabolic features that may explain these associations. For example the metabolites phenylethyl alcohol (R= 0.6 and p= 0.04) and ethyl acetate (R= 0.52 and p= 0.05) associated with the presence of *L. kefiranofaciens*, with supportive annotations including the presence of genes encoding for the enzymes with aryl-alcohol dehydrogenase and acetyl transferases activity, respectively, across *L. kefiranofaciens* MAGs [6].

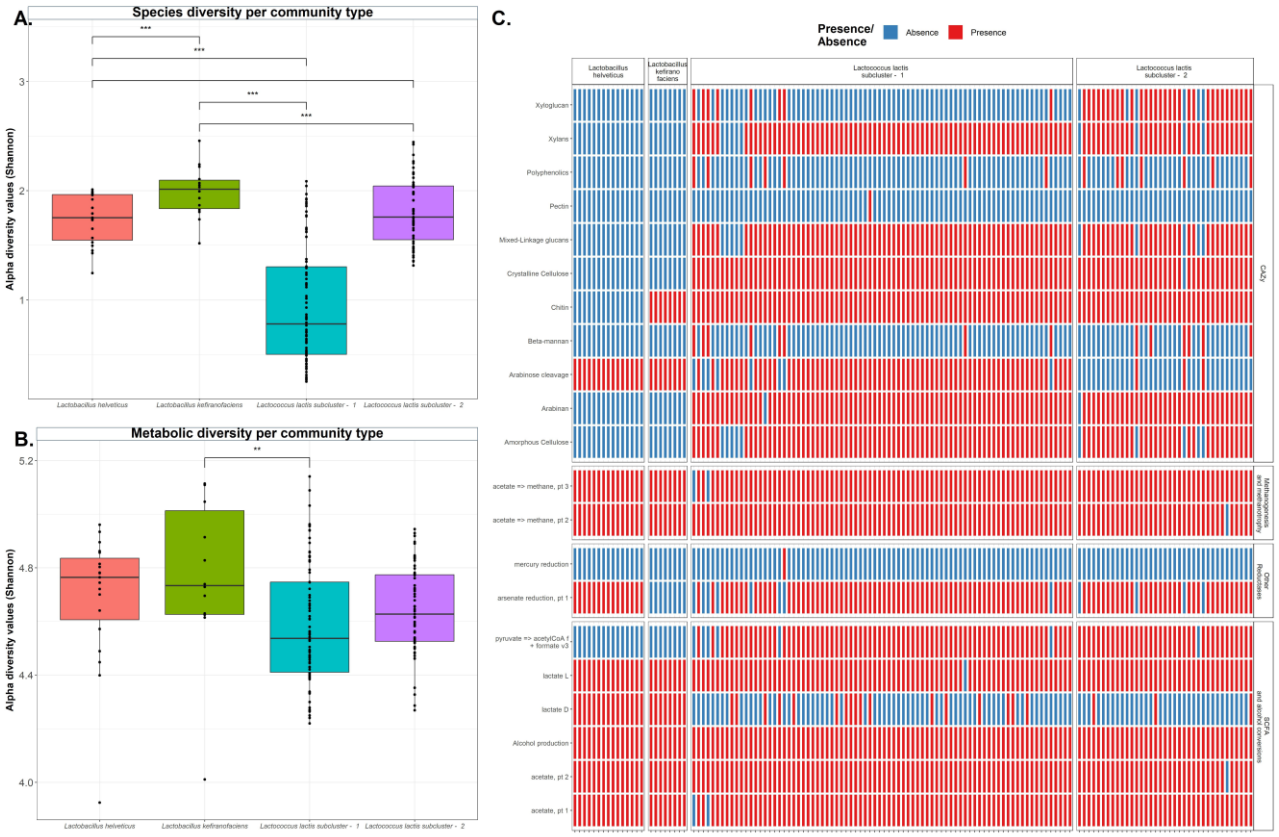


Figure 8: Diversity measures of community types within the kefir microbiome and characteristic metabolic function of *Lla. lactis* subcluster - 1, *Lla. lactis* subcluster - 2, *L. helveticus* and *L.* kefiranofaciens representative MAGs. **(A.)** Alpha diversity index observed for each community type when considering compositional data obtained using Kraken2. **(B.)** Alpha diversity index observed for each community type when considering functional data obtained using HUMAnN 3.0. The colouring of the boxplots represent samples dominated by *L. helveticus*, *L. kefiranofaciens,* *Lla. lactis* subcluster - 1 and *Lla. lactis* subcluster - 2. Pairwise tests were carried out between the dominant microbial species/strain using the Dunn's test for nonparametric data. Corrected p-values indicated as $p < 0.05$, ∗; $p < 0.01$, ∗∗; $p < 0.001$, ∗∗∗; $p < 0.0001$. ∗∗∗∗. **(C.)** Presence/absence of annotations indicative of characteristic metabolic functions involved in catalyzing carbohydrate degradation (CAZymes), methanogenesis and methanotropy, other reductases and short chain fatty acids and alcohol conversions (Y-axis) across representative MAGs identified using dRep and derived from the community types *Lla. lactis* subcluster - 1, *Lla. lactis* subcluster - 2, *L. helveticus* and *L.* kefiranofaciens (X-axis). The colour of each cell of the heat map represents the presence or absence of annotations.

2.3.6 Phylogenetic placement of representatives of each community- type provides taxonomic classification and supports the elevation of representatives of *Lla. lactis* subcluster - 2 to a distinct species

High quality MAGs, representing the dominant microbial species in each community type, were further analysed to identify closely related reference genomes. The *L. helveticus* representatives were mostly related to the reference genome GCF_000160855.1 with ANI values within the 97% range. The *L. kefiranofaciens* representatives were most related to the reference genome GCF_900103655.1, which represents the sub-species *L. kefiranofaciens subsp. kefiranofaciens* with ANI values within the 99% range and as such were deemed closely related genomes. The *Lla. lactis* subcluster - 1 MAGs were mostly related to the reference genome GCF_900099625.1, which is a representative of the sub-species *Lla. lactis* subsp*. lactis*, with ANI values distributed within the 98% range. The *Lla. lactis* subcluster - 2 MAGs were mostly related to the reference genome GCF_002078765.2, which is a representative of the sub-species *Lla. lactis* subsp*. cremoris* with ANI values distributed within the 97-99% range. As a recent study has proposed the elevation of *Lla. lactis* subsp*. cremoris* to the species level as *Lla. cremoris* [58], we performed pairwise genome comparisons using dRep between MAGs of the *Lla. lactis* subclusters. ANI values, between MAGs of *Lla. lactis* subcluster - 1 and *Lla. lactis* subcluster - 2 were lower than 95% ANI (Figure 9A), which generally represents a threshold for species delimitation [59]. As *Lla. cremoris* has yet to be validated under the rules of the International Code of Nomenclature of Prokaryotes [60] many bioinformatic tools, including those employed in this study, will not have adjusted their databases accordingly. Therefore, we re-evaluated the prevalence and relative abundance of *Lla. lactis* predicted using Kraken 2, which may have provided overestimations solely for the identification of the species *Lla. lactis*. Initially, when using Kraken 2, *Lla. lactis* dominated in 215 samples (ranging in abundance values from 24.8-96.5%), now through the application of metacache, we determined *Lla. lactis*, specifically *Lla. lactis* subsp*. lactis*, dominated in 125 samples (ranging in abundance values from 16-63%) and *Lla. cremoris*, specifically *Lla. cremoris* subsp*. cremoris*, dominated in 90 samples (ranging in abundance values from 23-60%). Furthermore, *Lla. lactis* subsp*. lactis* and *Lla. cremoris* subsp*. cremoris* were detected across all kefir metagenomes. We further confirmed the

dominance of sub-species *L. kefiranofaciens* subsp. *kefiranofaciens* across metagenomes compared to the *L. kefiranofaciens* subsp. *kefirgranum* sub-species, the latter accounting for just 0-0.001 % relative abundance across samples.

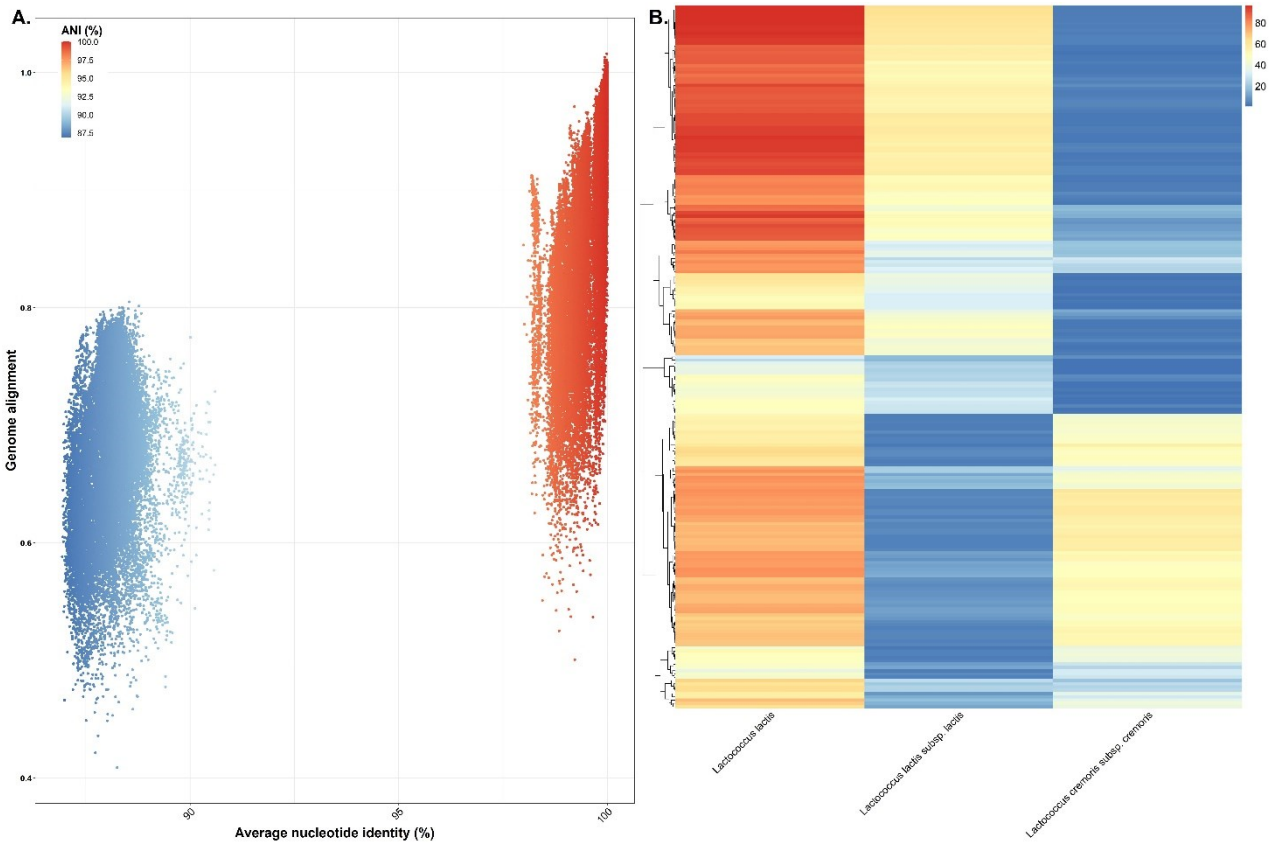


Figure 9: Separation of *Lla. lactis* subclusters based on Average nucleotide identity (ANI) and relative abundance across kefir metagenomes. **(A.)** Average nucleotide identity (ANI) gaps exist near ~91% ANI, between representatives of *Lla. lactis* subcluster - 1 and *Lla. lactis* subcluster - 2. Each dot represents the ANI (X-axis) and genome alignment percentage values (Y-axis) resulting from pairwise comparison within and between subclusters. Colours represent the ANI values. **(B.)** Heatmap displaying the relative abundance across 256 metagenomes of the species *Lla. lactis*, predicted using Kraken2, and *Lla. lactis* subsp. *lactis* and *Lla. cremoris* subsp. *cremoris* both predicted using metacache.

2.3.7 Virus identification highlights the prevalence of phages in kefir fermentations

Short read taxonomic profiling consistently detected the presence of bacteriophage at low relative abundance across samples, but was unable to assign the reads to specific species (Figure 1). Subsequent strain level profiling detected *Lla.* phage bIL285 (Figure S11), *Lla.* phage bIL309 (Figure S12) and *Lla.* phage bIL310 (Figure S13), but was unable to detect additional bacteriophage or provide associated functional annotations for the detected strains. Putative lysogenic viral contigs were consistently recovered in all high

quality MAGs constructed in this study. The level of phage detected was consistent with phage infection being common across all predominant species detected in the kefir metagenomes considered in this study, where one to a maximum of two prophage were identified per MAG. Additional viral sequences were also obtained by inspection of assemblies generated for viral features using the empirical screening criteria. Viral contigs may be representatives of prophage and/or free phage. Viral contigs were assessed for the presence of virally-encoded auxiliary metabolic genes (AMGs) using DRAM-v [34]. AMGs represent a unique category of viral genes that share homology with host metabolic genes but are adapted in viral genomes and function to reconfigure the metabolism of the host often to benefit viral synthesis [61]. The DRAM-v analysis detected 65 potential AMGs such as *pstS* and *phoU* across the kefir metagenomics dataset, 26 and 39 of which were deemed high (auxiliary scores 1-2) and medium-ranked (auxiliary scores 3) confidence AMGs, respectively. Notably, no AMGs were recovered from milk samples. Annotations associated with these genes suggests that kefir-associated bacteriophage encode AMGs that could play roles in host energy generation, carbohydrate utilization and organic nitrogen transformation, transport and other miscellaneous functions. Among the genes detected was *pstS*, which encodes a phosphate transport system substrate-binding protein which may function to redirect phosphate acquisition for phage genome replication. Furthermore, the presence of *phoU* may enable rapid uptake of free phosphate for use in DNA synthesis [62] to increase viral production.

Section 2.4: Discussion

In this study we applied metagenomics-based identification and characterisation of the microorganisms present in milk kefir produced by milk kefir grains from different sources to investigate the pan-metagenome in milk kefir fermentations. A detailed knowledge of the pan- metagenome within milk kefir, provides considerable insights into the heterogeneity of these populations, while also uncovering highly prevalent and key core or conserved features, which is of commercial and scientific value. In line with Ibacache-Quiroga *et al.*, 2022, this study found that, for the fermentation substrate and conditions employed, Eukarya represented less than 2% of the kefir microbiome, with the majority of the microbiota

attributed to bacterial taxa. We further build on previous publications, which determined the continued occurrence and dominance of *L. kefiranofaciens* and *Len. kefiri* within the less diverse milk kefir grain microbiome [5]. Specifically species, coupled with strain, level analysis revealed that one microbial strain, most frequently strains of the sub-species *Lla. lactis* subsp*. lactis* or *Lla. lactis* subsp*. cremoris* (named under the current International Code of Nomenclature of Prokaryotes), tended to dominate in the individual kefir (Figure 9). Other strains that dominated were representative of the sub-species *L. kefiranofaciens* subsp*. kefiranofaciens* and the species *L. helveticus* (Figure 1, Figure 2 and Figure 9). The combination of short read profiling to acquire relative abundance information and metagenomics assembly for phylogenetic placement could not further assign strains of *L. helveticus* to the sub-species level as determined using a combination of short read profiling to acquire relative abundance information and metagenomics assembly for phylogenetic placement. We note that in the majority of the fermentations analysed (n=104 of the 128), the pattern of dominance remained stable between the microbiome of 8 and 24 hr fermentations, suggesting stability in the kefir ecosystem as the fermentation process progresses (Figure S3, See Supplementary Information: Discussion for additional information).

The observed pattern of dominance was the principal factor driving the separation of kefir metagenomes into four distinct community types, labelled as *Lla. lactis* - subcluster - 1 (*Lla. lactis* subsp*. lactis*)*, Lla. lactis* - subcluster - 2 (*Lla. lactis subsp. cremoris*), *L. helveticus* and *L. kefiranofaciens* (*L. kefiranofaciens* subsp*. kefiranofaciens*) (Figure 7). We provide further evidence to support the reclassification of sub-species *Lla. lactis* subsp*. cremoris* to the distinct species, *Lla. cremoris* [58] through the pairwise genome comparisons of high quality representative MAGs. It has been shown that a number of phenotypic features differentiate *Lla. lactis* from *Lla. cremoris*, such as, maltose utilisation, growth temperature, salt tolerance, production of ammonia from arginine via the arginine deiminase system and GABA production [63]. Here we further confirm some of the aforementioned features at the genotypic level, including the presence of the entire complement of gene annotations supporting GABA production and the production of ammonia, uniquely detected in the *Lla. lactis* - subcluster - 1 representatives. Even though not always dominant, *Lla. lactis* and *Lla. cremoris* were detected in all samples at varying levels of relative abundance (Figure 9B), providing

evidence that they constitute part of the core microbiome in kefir fermentations. The identification of a core microbiome, community patterns and highly prevalent microbial species in kefir fermentations helps to define the core pan-metagenome and determine the minimal compositional framework for a fermented milk product to be considered as a kefir.

The genetic relationship between representatives of *Lla. lactis* subsp*. lactis* and *Lla. lactis* subsp*. cremoris* (~91% ANI) (Figure 9A) is in line with Li *et al* [58] and further supports the elevation of *Lla. lactis subsp. cremoris* to the species level as *Lla. cremoris* [58]. Such insights are particularly noteworthy as most bioinformatic tools available, including those implemented here, failed to delineate between *Lla. lactis* and the proposed *Lla. cremoris* at the species and strain level. As a result of both species being consistently detected in the kefir metagenome and often accounting for a high relative abundance, previous studies concerning kefir may have overrepresented the abundance of *Lla. lactis*.

Species sub-structures were observed for 12 microbial species (Figure S5, Figure S6, Figures S8, Figures S9, Figures S12-16, Figure S18, and Figure S20). Short-term strain retention was apparent in repeat fermentations with only one apparent instance of a different strain from the same species becoming dominant as the fermentation progressed. We also provide evidence of long-term strain-type retention through evolutionary comparisons between strains of the same secondary subcluster using the Kimura 2-parameter model. Such evolutionary comparisons identified the presence of the same strain-type, defined using the stringent threshold of identical nucleotide compositions in reconstructed marker genes, in kefirs derived from grains sourced from different geographical locations. Strain level identification was complemented by an in-depth assessment of functional differences between strains derived from the kefir ecosystem (Figure 5B*,* Figure 6, Figures S21-S28). Assessment at the functional level further confirmed a number of species sub-structures (Figure 5, Figure 7) and revealed few functional intra-species differences between strains of *A. fabarum* (Figure S21), *A. orientalis* (Figure S22), *L*. *helveticus* (Figure S23) and *S. thermophilu*s (Figure S28). The detection of few functional differences suggests the presence of strains derived from the same sub-species. However, for other representative species such as *Leuc. mesenteroides* (Figure S25)*,* we identified notable differences across the recovered strains that could be attributed to the

presence/absence of cluster specific gene families, and identified shared and unique gene families between strains of *Lla. lactis* and *Lla. cremoris*. The identification of a species pangenome and the grouping of gene families according to their presence/absence in strains provides a novel insight into the gene catalogue of kefir-associated strains and can inform future efforts to assess sub-species diversity.

The 222 metabolic pathways shared by the different community types may help to explain why some kefir fermentations that are dominated by different species/strains can have similar characteristics such as flavour and smell [64]. Furthermore the detection of 51 metabolic pathways across all kefir metagenomes considered in this study and 105 highly prevalent pathways provides a means to go beyond taxonomy to define the minimal functional framework for a fermented milk product to be considered as a kefir Melonnpan predicted a number of metabolites to be abundant and prevalent across kefir samples, including the metabolites deoxycholic acid and GABA, of interest as an neurotransmitter [65], respectively. We further identify potentially supportive genome annotations for a number of kefir derived species and associated metabolic compounds made in previous publications. Annotations consistent with the ability to produce of ethyl acetate in *L. kefiranofaciens* are particularly noteworthy as GC-MS analysis revealed a 59.15% increase in this metabolite in a kefir milk, as a result of a spike in with the strain *L. kefiranofaciens* NCFB 2797 [6]. For further details concerning the volatile compounds produced by milk kefir see [5, 6, 66]. Our analysis also provided insights into the identification and functional annotation of viral sequences, resulting in the metabolic categorization of putative AMGs, such as *pstS* and *phoU*, found across kefir metagenomes as thus representative of the accessory pan-metagenome of milk kefir. The presence of these phage-associated genes adds to the paradigm that phosphate scavenging is critical to phosphate-intensive viral reproduction [67, 68]. Such analysis highlights the presence of certain viral species and strains and provides insights into the functional potential of the viral component of the kefir microbiome.

Ultimately, this study provides the most comprehensive insight into the microbiomes of kefir to date through the establishment of a pan-metagenome, yielding novel information into the species, including putative new species and those residual of the unfermented milk, strain variability and potential functionality that can occur in kefir microbiomes derived from kefir grains sourced from across the globe.

Such insights extend public knowledge of the microbial species present in the artisanal beverage, which can inform further guidelines on kefir production and help to raise awareness of the microbial species and metabolic pathways commonly found in the milk kefir microbiome. The insights and workflows outlined in this manuscript could also be applied to other fermented foods to help shape future regulatory guidelines relating to the microbial content of specific foods.

Section 2.5: References

1. Chin-Wen, L., C. Hsiao-Ling, and J.-R. Liu, *Identification and characterisation of lactic acid bacteria and yeasts isolated from kefir grains in Taiwan.* Australian Journal of Dairy Technology, 1999. **54**(1): p. 14.
2. Bourrie, B.C.T., B.P. Willing, and P.D. Cotter, *The Microbiota and Health Promoting Characteristics of the Fermented Beverage Kefir.* Frontiers in Microbiology, 2016. **7**(647).
3. Georgalaki, M., et al., *Lactobacillus kefiranofaciens: From Isolation and Taxonomy to Probiotic Properties and Applications.* Microorganisms, 2021. **9**(10).
4. Farnworth, E.R. and I. Mainville, *Kefir: a fermented milk product.* Handbook of fermented functional foods, 2003. **2**: p. 89-127.
5. Blasche, S., et al., *Metabolic cooperation and spatiotemporal niche partitioning in a kefir microbial community.* Nat Microbiol, 2021. **6**(2): p. 196-208.
6. Walsh, A.M., et al., *Microbial Succession and Flavor Production in the Fermented Dairy Beverage Kefir.* mSystems, 2016. **1**(5).
7. Leeuwendaal, N.K., et al., *Fermented Foods, Health and the Gut Microbiome.* Nutrients, 2022. **14**(7): p. 1527.
8. Slattery, C., P.D. Cotter, and P.W. O'Toole, *Analysis of Health Benefits Conferred by Lactobacillus Species from Kefir.* Nutrients, 2019. **11**(6): p. 1252.
9. Hertzler, S.R. and S.M. Clancy, *Kefir improves lactose digestion and tolerance in adults with lactose maldigestion.* J Am Diet Assoc, 2003. **103**(5): p. 582-7.
10. Rodrigues, K.L., et al., *Antimicrobial and healing activity of kefir and kefiran extract.* Int J Antimicrob Agents, 2005. **25**(5): p. 404-8.
11. Lee, M.Y., et al., *Anti-inflammatory and anti-allergic effects of kefir in a mouse asthma model.* Immunobiology, 2007. **212**(8): p. 647-54.
12. de Moreno de Leblanc, A., et al., *Study of immune cells involved in the antitumor effect of kefir in a murine breast cancer model.* J Dairy Sci, 2007. **90**(4): p. 1920-8.
13. Liu, J.R., et al., *Hypocholesterolaemic effects of milk-kefir and soyamilk-kefir in cholesterol-fed hamsters.* Br J Nutr, 2006. **95**(5): p. 939-46.
14. Bourrie, B.C.T., P.D. Cotter, and B.P. Willing, *Traditional kefir reduces weight gain and improves plasma and liver lipid profiles more successfully than a commercial equivalent in a mouse model of obesity.* Journal of Functional Foods, 2018. **46**: p. 29-37.
15. Bourrie, B.C.T., et al., *Kefir microbial composition is a deciding factor in the physiological impact of kefir in a mouse model of obesity.* Br J Nutr, 2021. **125**(2): p. 129-138.
16. van de Wouw, M., et al., *Distinct actions of the fermented beverage kefir on host behaviour, immunity and microbiome gut-brain modules in the mouse.* Microbiome, 2020. **8**(1): p. 67.
17. Duboc, P. and B. Mollet, *Applications of exopolysaccharides in the dairy industry.* International Dairy Journal, 2001. **11**(9): p. 759-768.
18. Dimitrellou, D., et al., *Effect of freeze–dried kefir culture on proteolysis in feta-type and whey-cheeses.* Food Chemistry, 2010. **119**(2): p. 795-800.

19. Mantzourani, I., et al., *Study of kefir grains application in sourdough bread regarding rope spoilage caused by Bacillus spp.* Food Chem, 2014. **143**: p. 17-21.
20. Dimidi, E., et al., *Fermented Foods: Definitions and Characteristics, Impact on the Gut Microbiota and Effects on Gastrointestinal Health and Disease.* Nutrients, 2019. **11**(8): p. 1806.
21. Bourrie, B.C.T., C. Richard, and B.P. Willing, *Kefir in the Prevention and Treatment of Obesity and Metabolic Disorders.* Current Nutrition Reports, 2020. **9**(3): p. 184-192.
22. Mukherjee, A., et al., *Global Regulatory Frameworks for Fermented Foods: A Review.* Frontiers in Nutrition, 2022. **9**.
23. Nejati, F., et al., *Traditional Grain-Based vs. Commercial Milk Kefirs, How Different Are They?* Applied Sciences, 2022. **12**(8): p. 3838.
24. Imchen, M., R.K. Vennapu, and R. Kumavath, *Chapter 17 - Pan-metagenomics: An overview of the human microbiome*, in *Pan-genomics: Applications, Challenges, and Future Prospects*, D. Barh, et al., Editors. 2020, Academic Press. p. 335-342.
25. Wood, D.E., J. Lu, and B. Langmead, *Improved metagenomic analysis with Kraken 2.* Genome Biology, 2019. **20**(1): p. 257.
26. Lu, J., et al., *Bracken: estimating species abundance in metagenomics data.* PeerJ Computer Science, 2017. **3**: p. e104.
27. Beghini, F., et al., *Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3.* bioRxiv, 2020: p. 2020.11.19.388223.
28. Mallick, H., et al., *Predictive metabolomic profiling of microbial communities using amplicon or metagenomic sequences.* Nature Communications, 2019. **10**(1): p. 3136.
29. Li, D., et al., *MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph.* Bioinformatics, 2015. **31**(10): p. 1674-6.
30. Uritskiy, G.V., J. DiRuggiero, and J. Taylor, *MetaWRAP-a flexible pipeline for genome-resolved metagenomic data analysis.* Microbiome, 2018. **6**(1): p. 158.
31. Parks, D.H., et al., *CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes.* Genome Res, 2015. **25**(7): p. 1043-55.
32. Chaumeil, P.-A., et al., *GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database.* Bioinformatics, 2019. **36**(6): p. 1925-1927.
33. Olm, M.R., et al., *dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication.* The ISME Journal, 2017. **11**(12): p. 2864-2868.
34. Shaffer, M., et al., *DRAM for distilling microbial metabolism to automate the curation of microbiome function.* Nucleic Acids Res, 2020. **48**(16): p. 8883-8900.
35. Asnicar, F., et al., *Precise phylogenetic analysis of microbial isolates and genomes from metagenomes using PhyloPhlAn 3.0.* Nature Communications, 2020. **11**(1): p. 2500.
36. Truong, D.T., et al., *Microbial strain-level population structure and genetic diversity from metagenomes.* Genome research, 2017. **27**(4): p. 626-638.
37. Müller, A., et al., *MetaCache: context-aware classification of metagenomic reads using minhashing.* Bioinformatics, 2017. **33**(23): p. 3740-3748.
38. Guo, J., et al., *VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses.* Microbiome, 2021. **9**(1): p. 37.
39. Nayfach, S., et al., *CheckV assesses the quality and completeness of metagenome-assembled viral genomes.* Nature Biotechnology, 2021. **39**(5): p. 578-585.
40. R Core Team, *R: A language and environment for statistical computing.* 2020.
41. Oksanen, J., et al., *vegan: Community Ecology Package*. 2020.
42. Anderson, M.J., *Distance-based tests for homogeneity of multivariate dispersions.* Biometrics, 2006. **62**(1): p. 245-53.
43. Segata, N., et al., *Metagenomic biomarker discovery and explanation.* Genome biology, 2011. **12**(6): p. R60-R60.
44. Paradis, E. and K. Schliep, *ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R.* Bioinformatics, 2018. **35**(3): p. 526-528.

45. Kassambara, A. and F. Mundt, *factoextra: Extract and Visualize the Results of Multivariate Data Analyses*. 2020.
46. Kassambara, A., *ggpubr:"ggplot2" based publication ready plots (R package version 0.4. 0)[Computer software]*. 2020, ggpubr:'ggplot2'based publication ready plots (R Package Version 0.4. 0 ….
47. South, A. and M.A. South, *Package 'rworldmap'.* Mapping Global Data, 2016: p. 1.3-6.
48. BioRender, *BioRender.* 2021.
49. Wickham, H., *ggplot2: elegant graphics for data analysis*. 2016: springer.
50. Kolde, R., *pheatmap: Pretty Heatmaps. R package version 1.0. 12.* CRAN. R-project. org/package= pheatmap, 2019.
51. Yu, G., *Using ggtree to Visualize Data on Tree-Like Structures.* Current Protocols in Bioinformatics, 2020. **69**(1): p. e96.
52. Chamberlain, S.A. and E. Szöcs, *taxize: taxonomic search and retrieval in R.* F1000Research, 2013. **2**: p. 191-191.
53. Beghini, F., et al., *Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3.* eLife, 2021. **10**: p. e65088.
54. Zolfo, M., et al., *Detecting contamination in viromes using ViromeQC.* Nat Biotechnol, 2019. **37**(12): p. 1408-1412.
55. Caspi, R., et al., *The MetaCyc database of metabolic pathways and enzymes - a 2019 update.* Nucleic Acids Research, 2019. **48**(D1): p. D445-D453.
56. Wu, Q., et al., *Common Distribution of gad Operon in Lactobacillus brevis and its GadA Contributes to Efficient GABA Synthesis toward Cytosolic Near-Neutral pH.* Frontiers in Microbiology, 2017. **8**.
57. Diez-Gutiérrez, L., et al., *Gamma-aminobutyric acid and probiotics: Multiple health benefits and their future in the global functional food and nutraceuticals market.* Journal of Functional Foods, 2020. **64**: p. 103669.
58. Li, T.T., W.L. Tian, and C.T. Gu, *Elevation of Lactococcus lactis subsp. cremoris to the species level as Lactococcus cremoris sp. nov. and transfer of Lactococcus lactis subsp. tructae to Lactococcus cremoris as Lactococcus cremoris subsp. tructae comb. nov.* International Journal of Systematic and Evolutionary Microbiology, 2021. **71**(3).
59. Richter, M. and R. Rosselló-Móra, *Shifting the genomic gold standard for the prokaryotic species definition.* Proceedings of the National Academy of Sciences, 2009. **106**(45): p. 19126-19131.
60. *International Code of Nomenclature of Prokaryotes.* International Journal of Systematic and Evolutionary Microbiology, 2019. **69**(1A): p. S1-S111.
61. Warwick-Dugdale, J., et al., *Host-hijacking and planktonic piracy: how phages command the microbial high seas.* Virology Journal, 2019. **16**(1): p. 15.
62. Muda, M., N.N. Rao, and A. Torriani, *Role of PhoU in phosphate transport and alkaline phosphatase regulation.* J Bacteriol, 1992. **174**(24): p. 8057-64.
63. Wels, M., et al., *Comparative genome analysis of Lactococcus lactis indicates niche adaptation and resolves genotype/phenotype disparity.* Frontiers in microbiology, 2019. **10**: p. 4.
64. Farag, M.A., et al., *The Many Faces of Kefir Fermented Dairy Products: Quality Characteristics, Flavour Chemistry, Nutritional Value, Health Benefits, and Safety.* Nutrients, 2020. **12**(2).
65. Cryan, J.F. and O.F. O'Leary, *A glutamate pathway to faster-acting antidepressants?* Science, 2010. **329**(5994): p. 913-914.
66. Guangsen, T., L. Xiang, and G. Jiahu, *Microbial diversity and volatile metabolites of kefir prepared by different milk types.* CyTA - Journal of Food, 2021. **19**(1): p. 399-407.
67. Bratbak, G., J.K. Egge, and M. Heldal, *Viral mortality of the marine alga Emiliania huxleyi (Haptophyceae) and termination of algal blooms.* Marine Ecology Progress Series, 1993. **93**(1/2): p. 39-48.
68. CLASEN, J.L. and J.J. ELSER, *The effect of host Chlorella NC64A carbon : phosphorus ratio on the production of Paramecium bursaria Chlorella Virus-1.* Freshwater Biology, 2007. **52**(1): p. 112-122.
69. Walsh, L.H., et al., *Bioinformatic approaches for studying the microbiome of fermented food.* Critical Reviews in Microbiology, 2022: p. 1-33.

Section 2.6: Supplementary information

2.6.1 Supplementary method

Method S1: Kefir fermentation, sample collection and DNA extraction

Samples were kindly donated or purchased from private households that consented to their use in the study. Upon arrival, the grains were weighed and divided in two. Half of the divided grain was frozen immediately at -20^{o}C and the other half was cultured continuously for one to two weeks until the grains reached 4 grams in weight. After this growth period, each kefir grain was cultured twice by inoculating the grains into 200 ml of full fat cow's milk (Avonmore, Ireland; 2 % wt/vol) in a laminar flow hood and placing in an orbital shaker incubator at 24^{o}C. 20 ml fermented milk was collected after 8 hours (hr) and 24 hr of the fermentation process for each of the successive fermentations. In total, four samples were collected for each fermentation initiated by each of the 64 kefir grains sourced in this study, resulting in a total of 256 kefir samples. Following collection, the kefir milk samples were stored at −20°C prior to DNA extraction, where 15 ml milk was used. Non-fermented full fat cow's milk samples were used as controls.

Prior to DNA extraction, 15 ml of kefir milk was centrifuged at 5,444 × g for 30 minutes (min) at 4°C. Samples were washed with 10 ml of phosphate buffered saline (PBS) and centrifuged again at 5,444 × g for 15 min, to remove protein and fat residues. Two additional rounds of washing and centrifugation were performed before pelleting the microbial cells. The cell pellet was resuspended in 800 µl of Solution CD1 from the DNeasy PowerSoil Pro Kits. Total DNA from the resuspended pellets was extracted and purified according to the standard DNeasy PowerSoil Pro Kits kit protocol.

Method S2: Preparation of Libraries, Shotgun Metagenomic Sequencing and Data Pre-processing

Total DNA was initially quantified and qualified by gel electrophoresis and using the NanoDrop 1000 (BioSciences, Dublin, Ireland) prior to more accurate quantification with the Qubit High Sensitivity DNA kit (BioSciences, Dublin, Ireland). Whole-metagenome shotgun libraries were prepared in accordance with the Nextera XT DNA Library Preparation Guide from Illumina (Clooney *et al.*, 2016) and stored at 20°C until further processing. The libraries were assessed using the Agilent 2100 Bioanalyzer system, quantified with

the Qubit High Sensitivity DNA kit, and sequenced on the Illumina NextSeq sequencing platform using a v2 NextSeq 500/550 high-output reagent kit (300 cycles) in accordance with standard Illumina sequencing protocols. Sequencing was performed at the Teagasc Sequencing Centre (Moorepark, Cork, Ireland). All metagenomic processing was run using the Teagasc high performance computing cluster. Raw paired-end FASTQ files containing the metagenomic shotgun sequences were trimmed using Trimgalore v0.6.1 (Krueger, 2015) to remove adapter content and low quality reads. Reads were queried against a reference bovine genome using Bowtie2 v2.4.4 (Langmead and Salzberg, 2012) to identify contaminated DNA .

2.6.2 Supplementary Results

Result S1: Overview of metagenomics data

The 256 quality-trimmed kefir specific data sets had a combined size of approximately 102.9 Gb and nine quality-trimmed milk specific data sets with a combined size of approximately 0.4 GB. Both datasets contained reads with median lengths of 150. The kefir specific data set had an average of 0.45 Gb (with a maximum of 0.68 Gb) per sample. The milk specific data set had an average of 0.041 Gb (with a maximum of 0.35 Gb) per sample.

Result S2: Construction of a Kraken 2 database

Taxonomic analysis of the kefir metagenomic data was initially performed by Kraken 2 against the standard Kraken 2 database (Wood, Lu and Langmead, 2019), MetaPhlAn 3 (Beghini *et al.*, 2021) and Kaiju using the refseq Kaiju's database index (Menzel, Ng and Krogh, 2016). Applying these well-known taxonomic methods (see methods) to assess the microbial composition among kefir metagenomes resulted in differing compositional profiles and classification rates. None of the tools in isolation successfully detected taxa of bacterial, viral and fungal origin. In an effort to improve the number of detections The short read taxonomic profiler, Kraken2 (Wood, Lu and Langmead, 2019) was employed to determine the microbial compositions of 256 kefir metagenomes using a custom build Kraken2 database (see methods).

Result S3: Microbial succession

Microbial succession patterns or changes in community structure were analysed. Initially, taxonomic profiles across samples derived from the same kefir grains were compared. Such comparisons revealed little compositional change between the 8 and 24 hr time points for the majority of samples (n=208 samples corresponding to 52 different starter grains), however succession patterns were observed for the remaining samples (n=48 samples corresponding to 12 grains), where the dominant microbial species changed over the course of 8 hr to 24 hr fermentations to another of the dominant microbial species. Such patterns often represented a decrease in the relative abundance of species such as *L. kefiranofaciens* (n=7), *L. helveticus* (n=6) or *P. helleri* (n=2), and a subsequent increase in the relative abundance of *Lla. lactis*. No distinct pattern was observed between time point and or community types, however community types dominated by *P. helleri, Acinetobacter* sp. TTHO-4 and *Leuc*. *mesenteroides* did not persist as the fermentation process progressed to 24 hr. LEfSe was used to identify differentially abundant taxa between the timepoints. *Pseudomonas* species and *T. thermophilus* tended to more abundant and were consistently most prevalent at 0 hr (i.e., representing the microbiome of the milk substrate), *L. kefiranofaciens, Lentilactobacillus kefiri and L. helveticus*, tended to more abundant at 8 hr whereas *Lla. lactis* tended to be more abundant at 24 hr.

Result S4: Metabolic potential of the kefir microbiome and core metabolic pathways

Separation was observed between the milk kefir samples at 8 hr and 24 hr and unfermented milk controls in terms of the prevalence of metabolic pathways, with 140 metabolic pathways detected among the unfermented milk controls in contrast to 389 and 382 pathways detected across the 8 hr and 24 hr milk kefir samples, respectively, with the majority of these kefir-associated pathways (n=369) being shared. Further evidence of separation between time points was provided by the statistical tool LEfSe; of the total of 402 pathways, LEfSe detected 1, 59, and 89 differentially abundant pathways between 0, 8 and 24 hr samples, respectively. Notably, we observed that pathways involved in carbohydrate metabolism and biosynthesis and unsaturated fatty acid biosynthesis were most abundant at 8 hr, whereas those involved in amino acid

biosynthesis cofactor, carrier, and vitamin biosynthesis and metabolism were most abundant at 24 hr.

2.6.3 Supplementary Discussion

Little functional and compositional difference were observed between the 8 and 24 hr fermentations, suggesting relative stability in the kefir ecosystem as the fermentation process progresses. Notably *L. kefiranofaciens* and *Len. kefiri* was differentially abundant at the 8 hr category indicating their frequent contributions as early stage fermenting taxa, while *Lla. lactis* was most prevalent at 24 hr, indicating its role as a late stage fermenting microorganism. Microbial succession events have also been reported in other publications assessing the kefir microbiome. Walsh *et al.*, 2016 noted changes in the relative abundance of *L. kefiranofaciens* and *Leuc. mesenteroides,* which decreased and increased as the fermentation process progressed (Walsh *et al.*, 2016).

2.6.4 Supplementary Figures

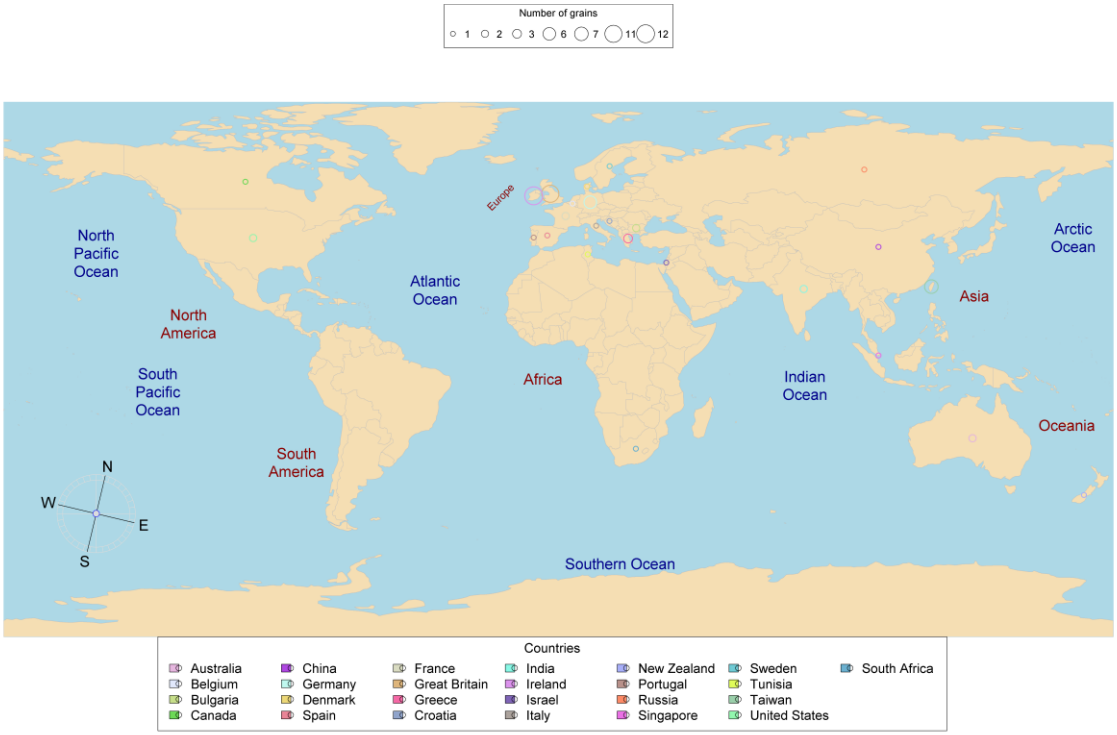


Figure S1: Kefir grain sampling initiative undertaken in this study. World map showing the countries from which the kefir grains were sourced. The different colour circles represent the country in which the kefir grains originated. The diameter of the circle displays the number of grains sourced from that country.

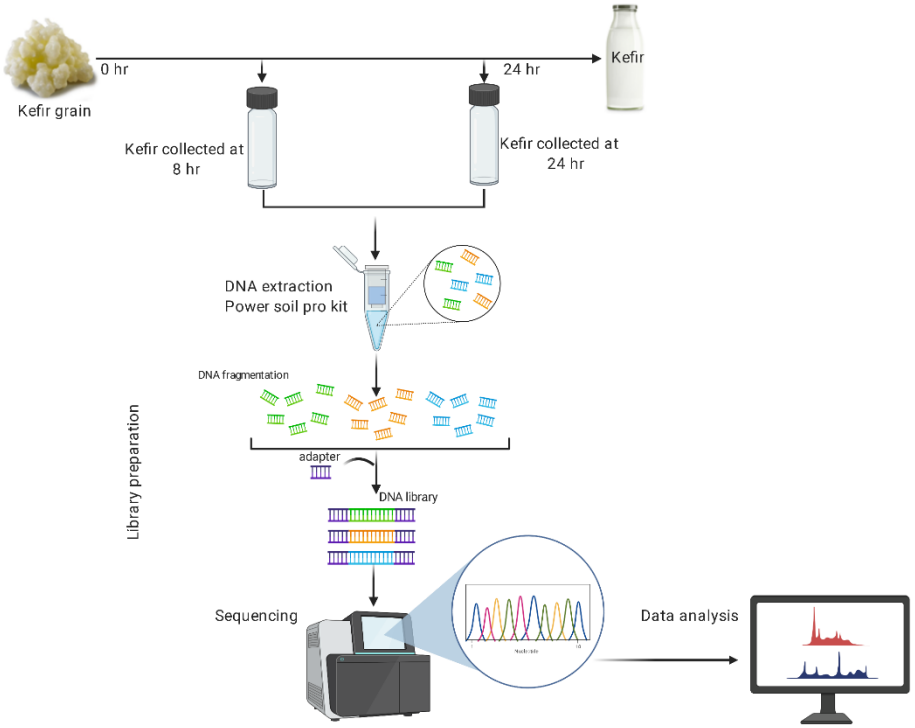


Figure S2: Graphical representation of the experimental design implemented in this study. Steps involved in the experimental design included kefir fermentations, sample collection, DNA extraction, library preparation, sequencing and bioinformatics analysis.

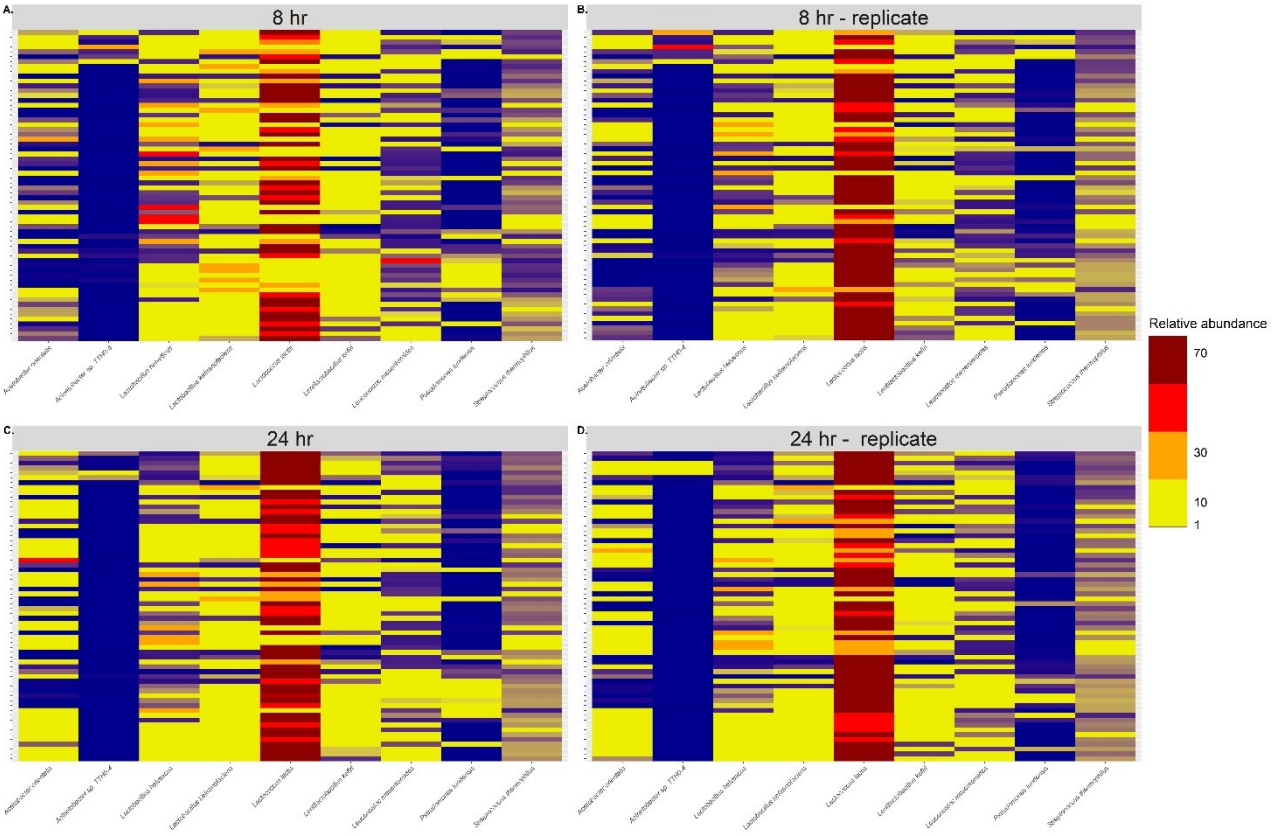


Figure S3: The microbial community patterns of the milk kefirs with time. **(A.)** Species that were detected at ≥10% relative abundance in at least one 8 hr sample. **(B.)** Species that were detected at ≥10% relative abundance in at least one 8 hr-replicate sample. **(C.)** Species that were detected at ≥10% relative abundance in at least one 24 hr sample and **(D.)** Species that were detected at ≥10% relative abundance in at least one 24 hr-replicate sample.

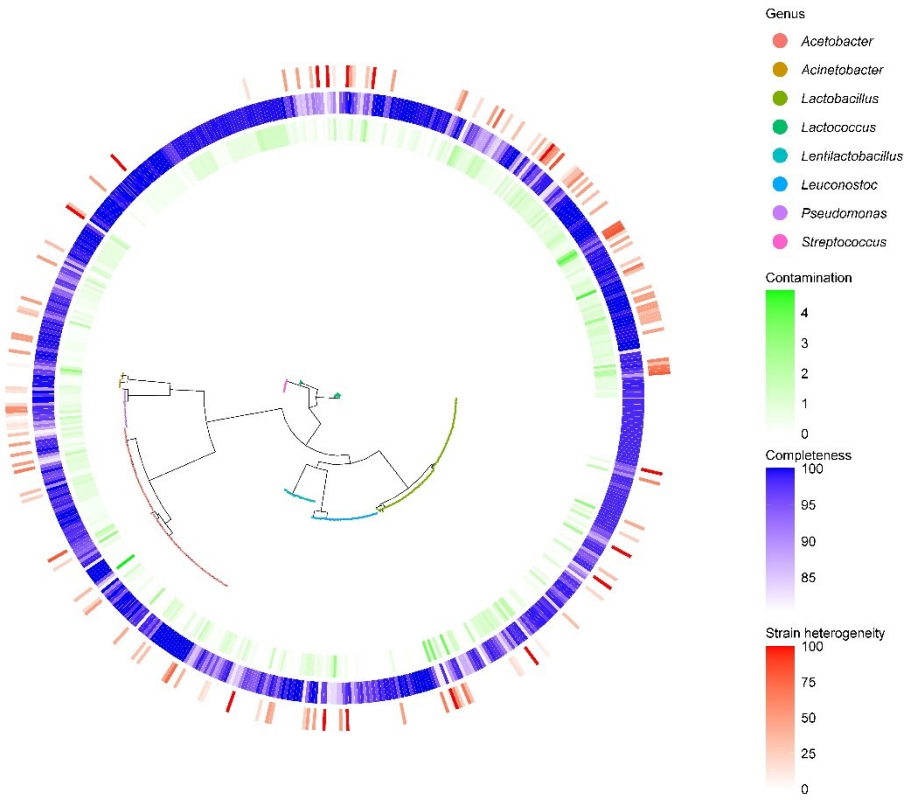


Figure S4: High quality metagenome assembled genomes assembled in this study Phylogenetic tree of the 613 high quality MAGs assembled from kefir metagenomics datasets, obtained in this study. Outer rings coloured green, blue and red represent % contamination, completion and strain heterogeneity respectively, as assessed using checkM. The coloured tip of the phylogram corresponding to the genus classification of the MAG.

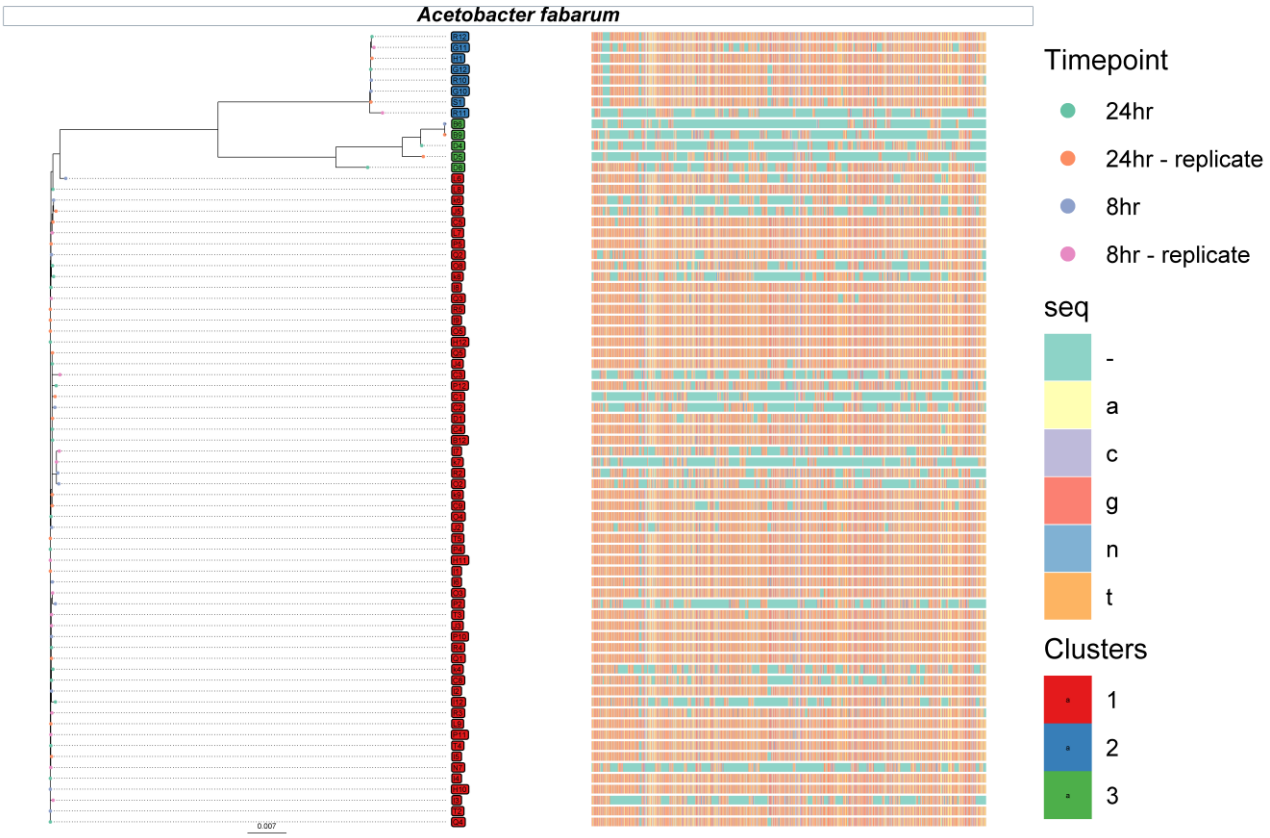


Figure S5: Strains of *A. fabarum* detected across kefir metagenomes. Strain level phylograms of *A. fabarum* strains and corresponding multiple sequence alignment (MSA) files generated by StrainPhlAn 3. Coloured tips of the phylogram corresponds to the time point of the kefir metagenome from which the strain was recovered. Colouring of labels represents the strain subcluster by which the strain was grouped.

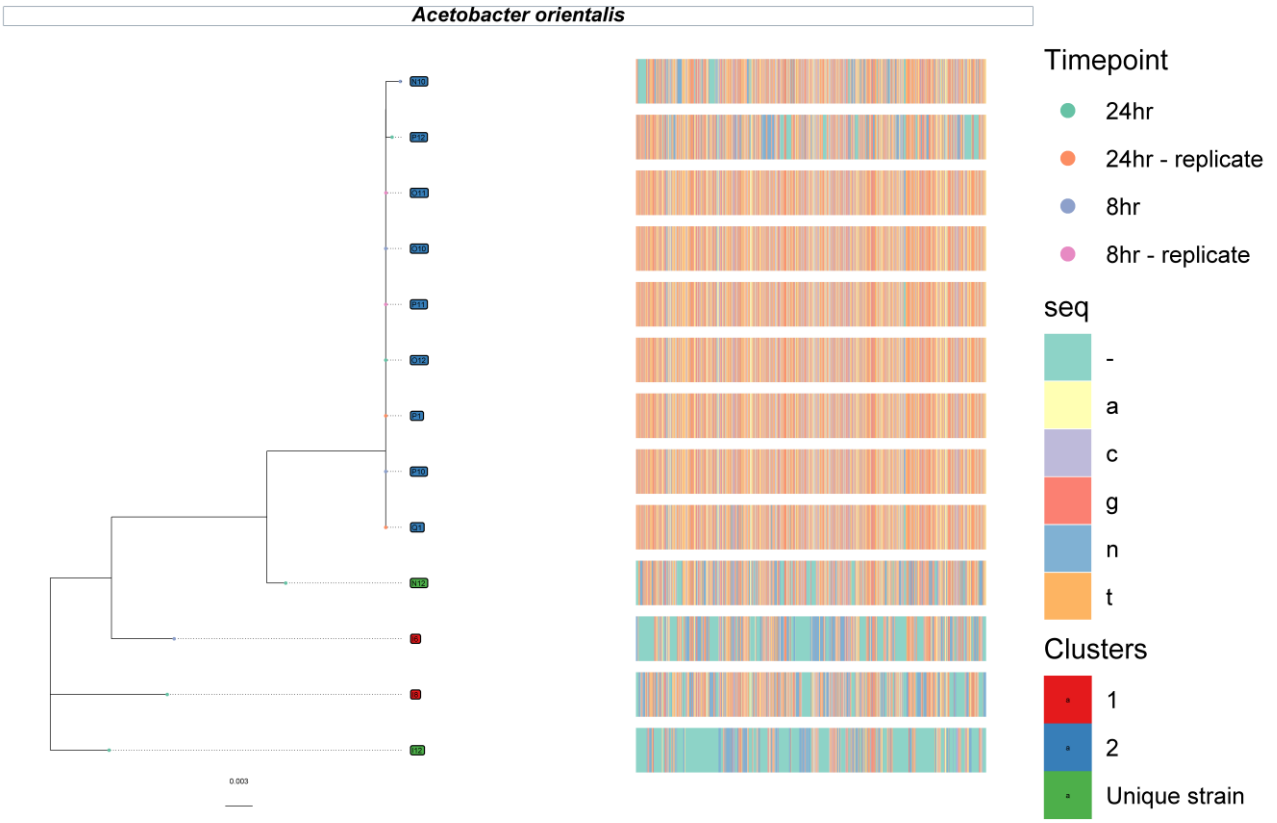


Figure S6: Strains of *A. orientalis* detected across kefir metagenomes. Strain level phylograms of *A. orientalis* strains and corresponding multiple sequence alignment (MSA) files generated by StrainPhlAn 3. Coloured tips of the phylogram corresponds to the time point of the kefir metagenome from which the strain was recovered. Colouring of labels represents the strain subcluster by which the strain was grouped.

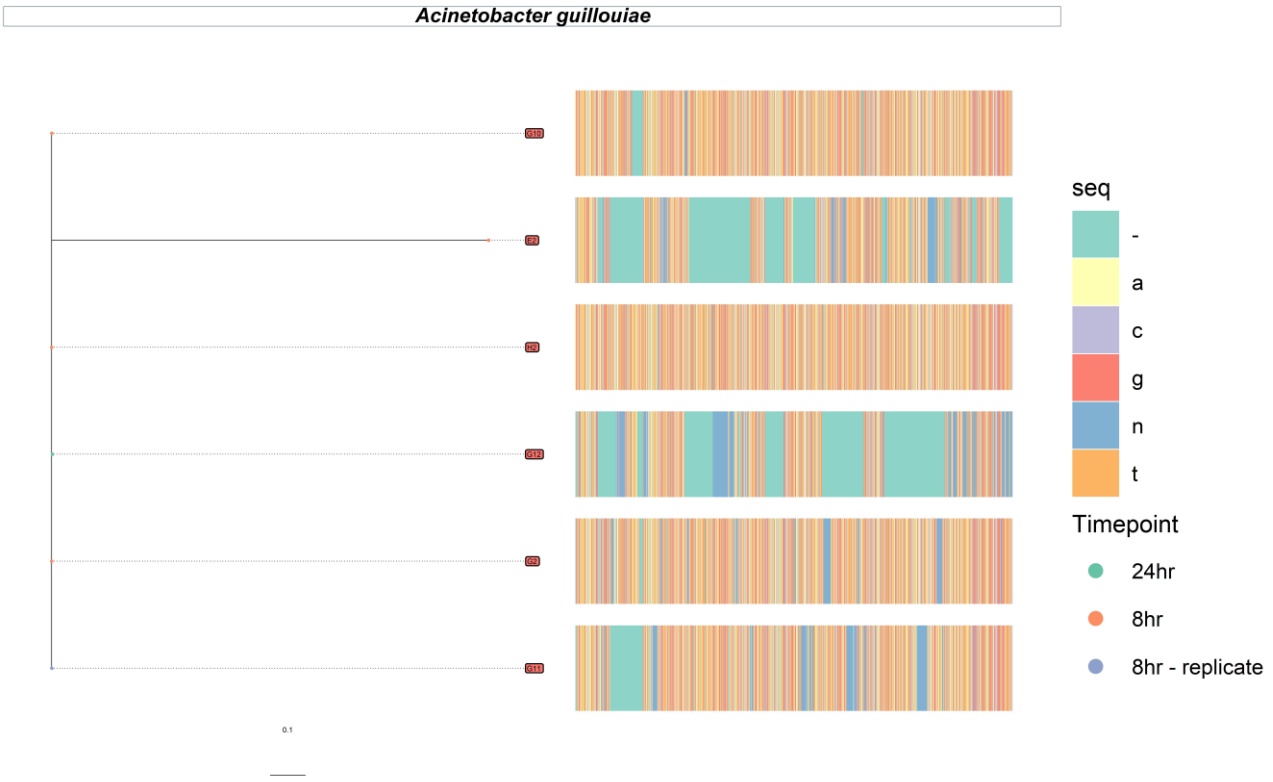


Figure S7: Strains of *Ac. guillouiae* detected across kefir metagenomes**.** Strain level phylograms of *Ac. guillouiae* strains and corresponding multiple sequence alignment (MSA) files generated by StrainPhlAn 3. Coloured tips of the phylogram corresponds to the time point of the kefir metagenome from which the strain was recovered. Colouring of labels represents the strain subcluster by which the strain was grouped.

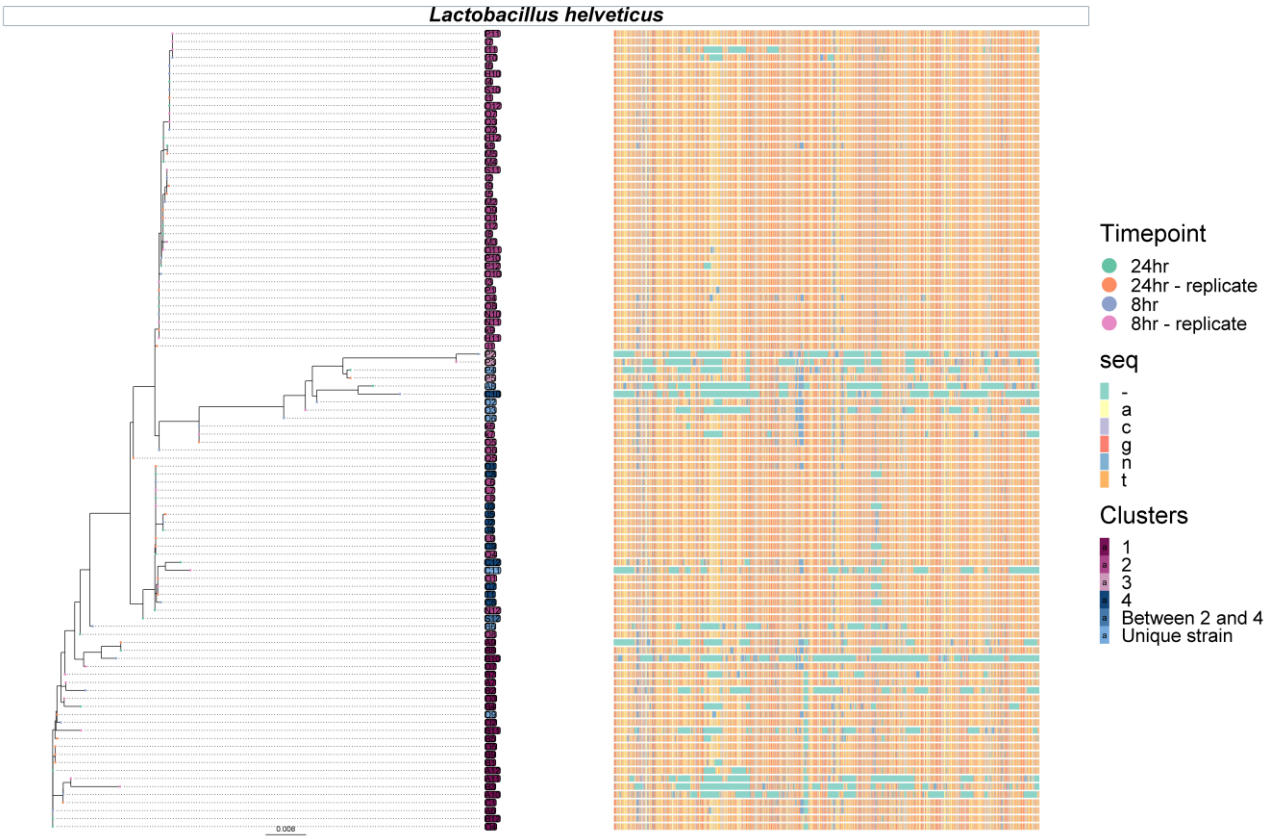


Figure S8: Strains of *L. helveticus* detected across kefir metagenomes. Strain level phylograms of *L. helveticus* strains and corresponding multiple sequence alignment (MSA) files generated by StrainPhlAn 3. Coloured tips of the phylogram corresponds to the time point of the kefir metagenome from which the strain was recovered. Colouring of labels represents the strain subcluster by which the strain was grouped.

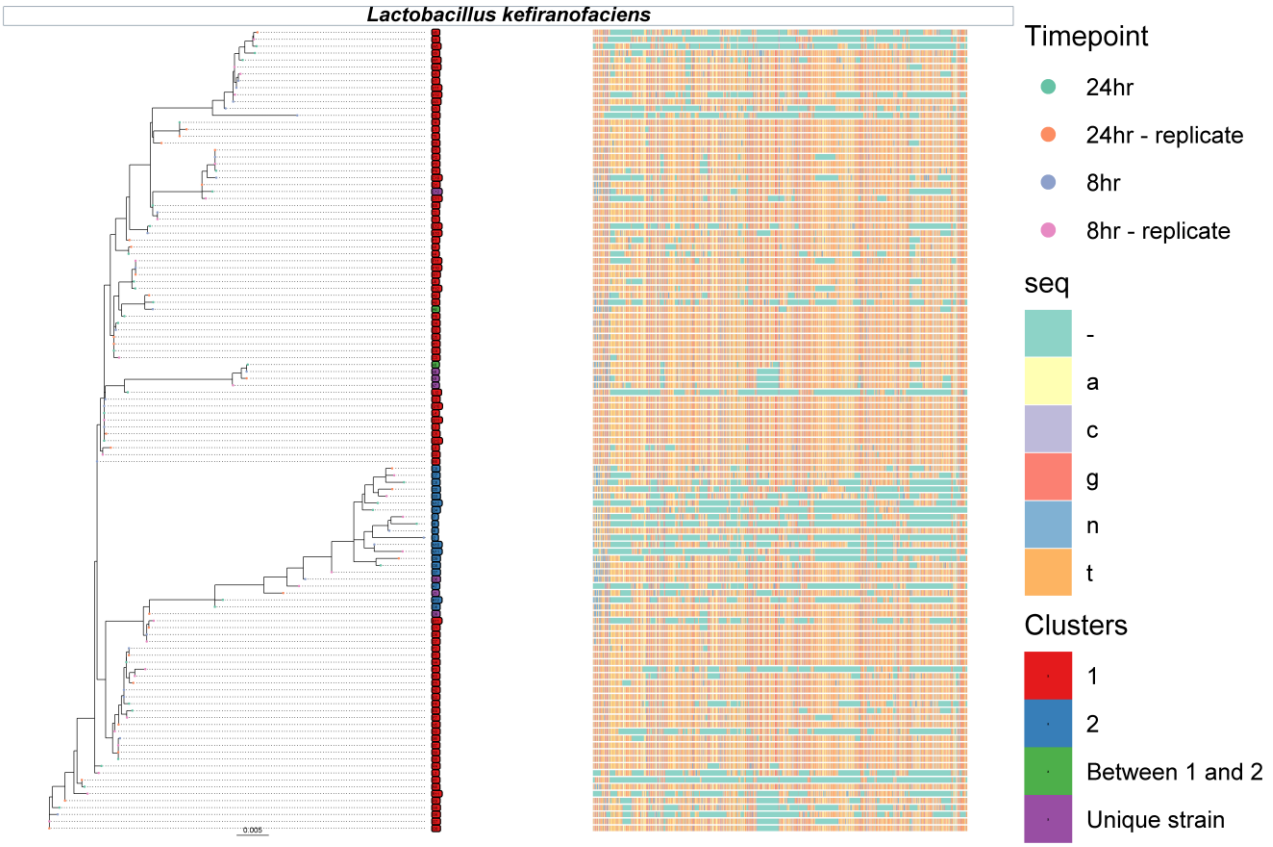


Figure S9: Strains of *L. kefiranofaciens* detected across kefir metagenomes.Strain level phylograms of *L. kefiranofaciens* strains and corresponding multiple sequence alignment (MSA) files generated by StrainPhlAn 3. Coloured tips of the phylogram corresponds to the time point of the kefir metagenome from which the strain was recovered. Colouring of labels represents the strain subcluster by which the strain was grouped.

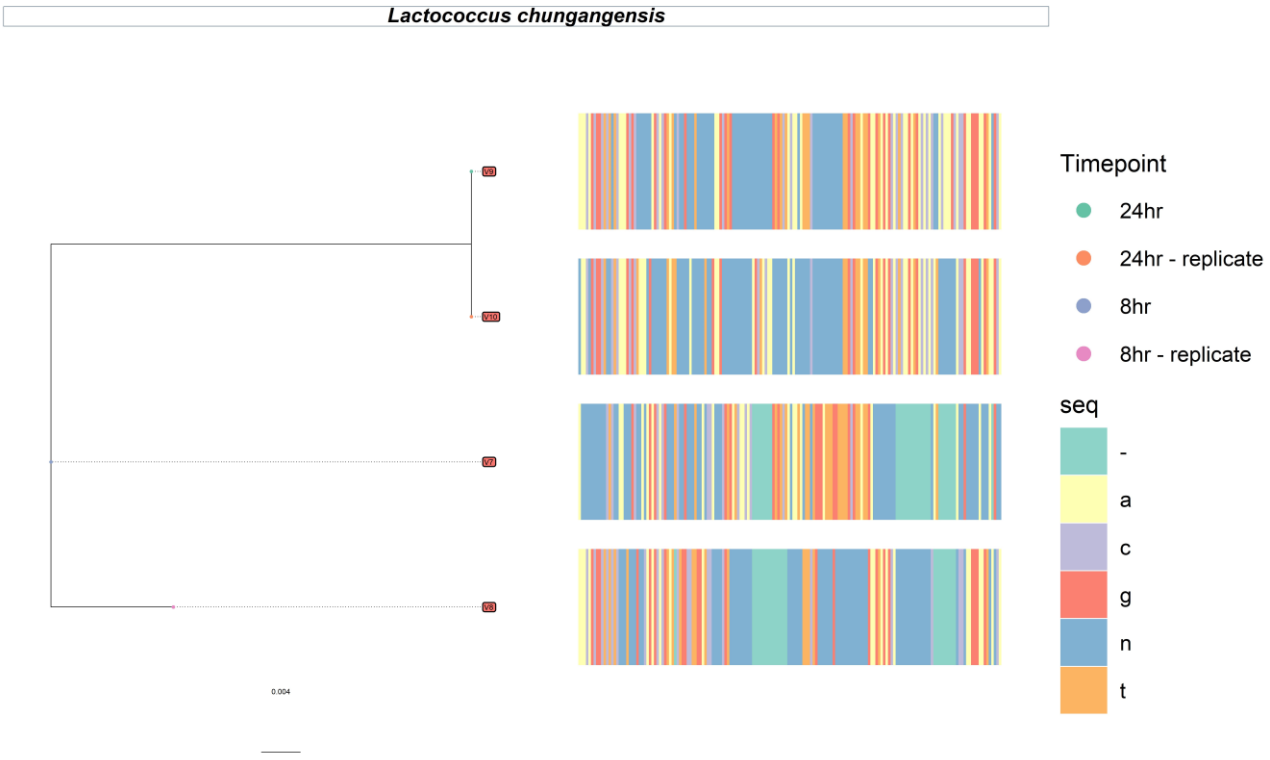


Figure S10: Strains of *Lla. chungangensis* detected across kefir metagenomes. Strain level phylograms of *Lla. chungangensis* strains and corresponding multiple sequence alignment (MSA) files generated by StrainPhlAn 3. Coloured tips of the phylogram corresponds to the time point of the kefir metagenome from which the strain was recovered. Colouring of labels represents the strain subcluster by which the strain was grouped.

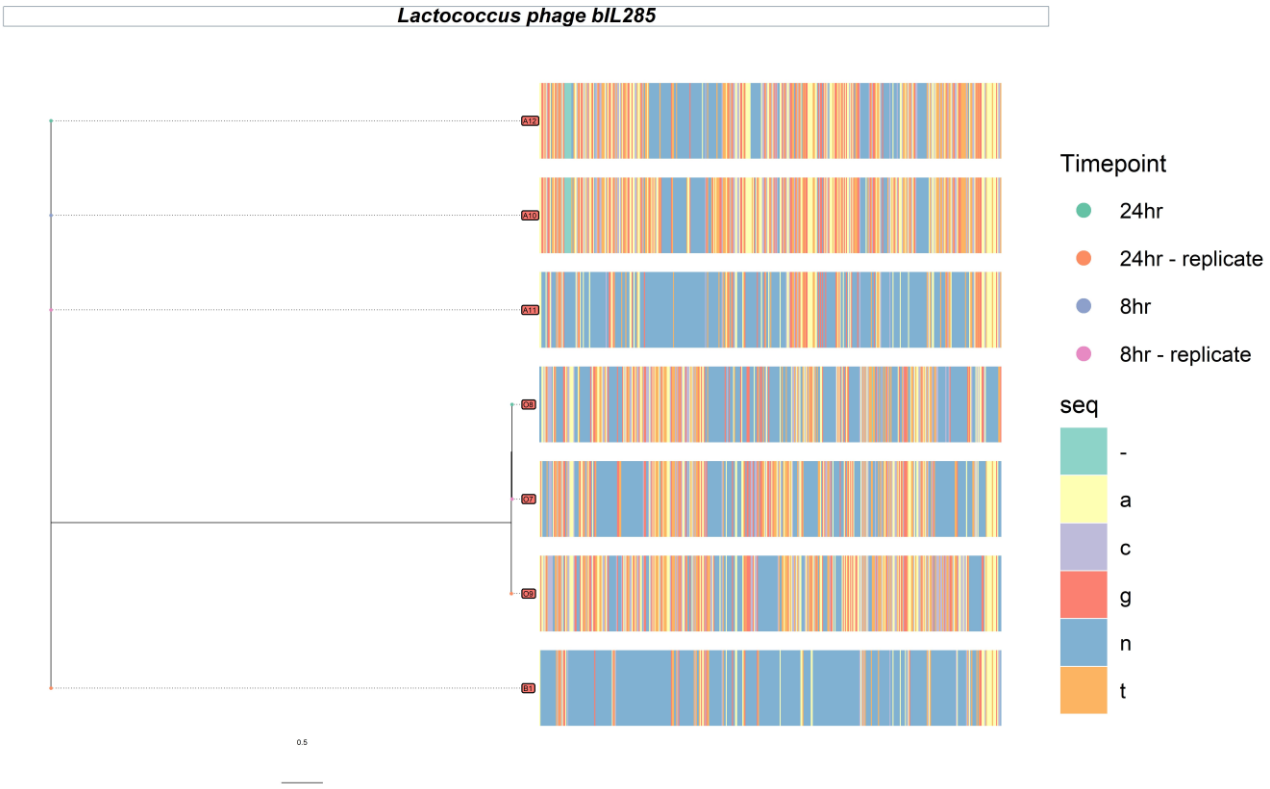


Figure S11: isolates of *Lla. phage bIL285* detected across kefir metagenomes. Strain level phylograms of *Lla. phage bIL285* strains and corresponding multiple sequence alignment (MSA) files generated by StrainPhlAn 3. Coloured tips of the phylogram corresponds to the time point of the kefir metagenome from which the strain was recovered. Colouring of labels represents the strain subcluster by which the strain was grouped.

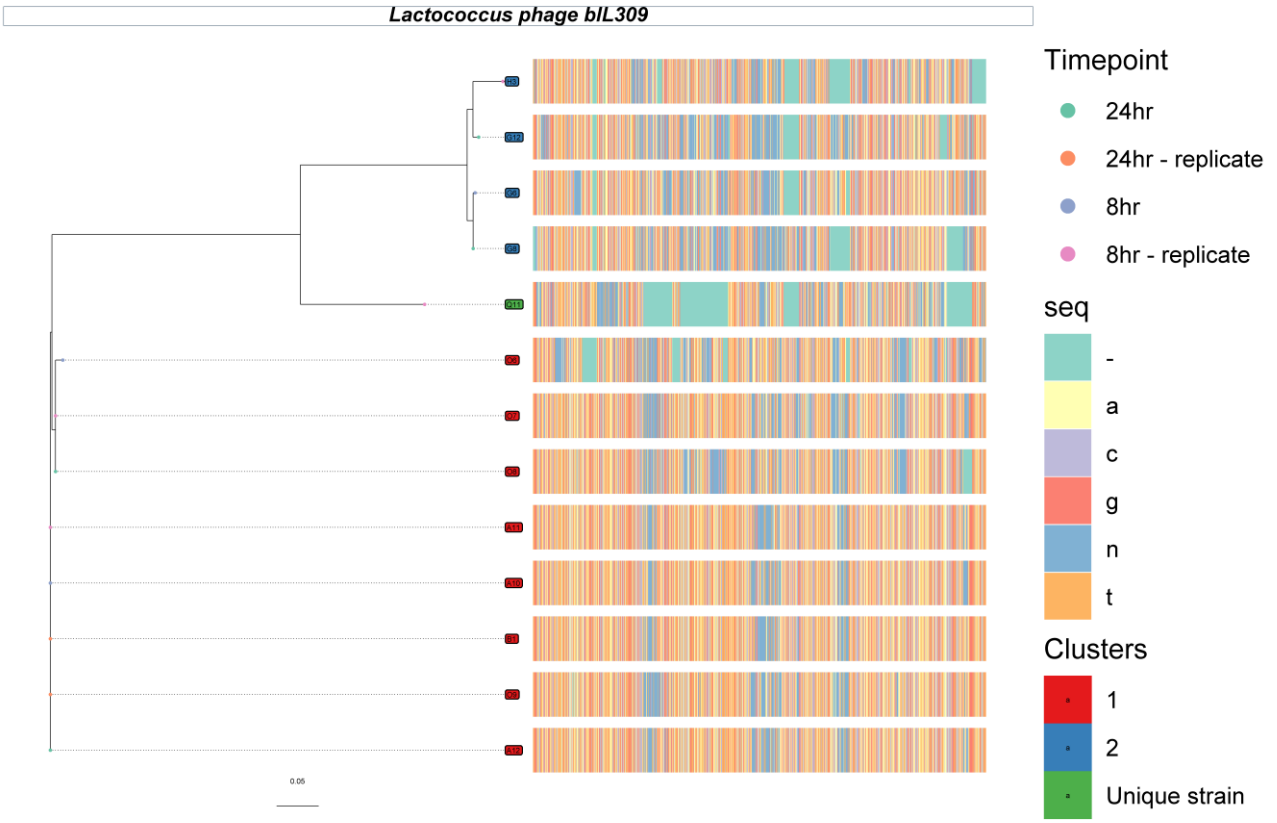


Figure S12: Isolates of *Lc. phage bIL309* detected across kefir metagenomes. Strain level phylograms of *Lc. phage bIL309* strains and corresponding multiple sequence alignment (MSA) files generated by StrainPhlAn 3. Coloured tips of the phylogram corresponds to the time point of the kefir metagenome from which the strain was recovered. Colouring of labels represents the strain subcluster by which the strain was grouped.

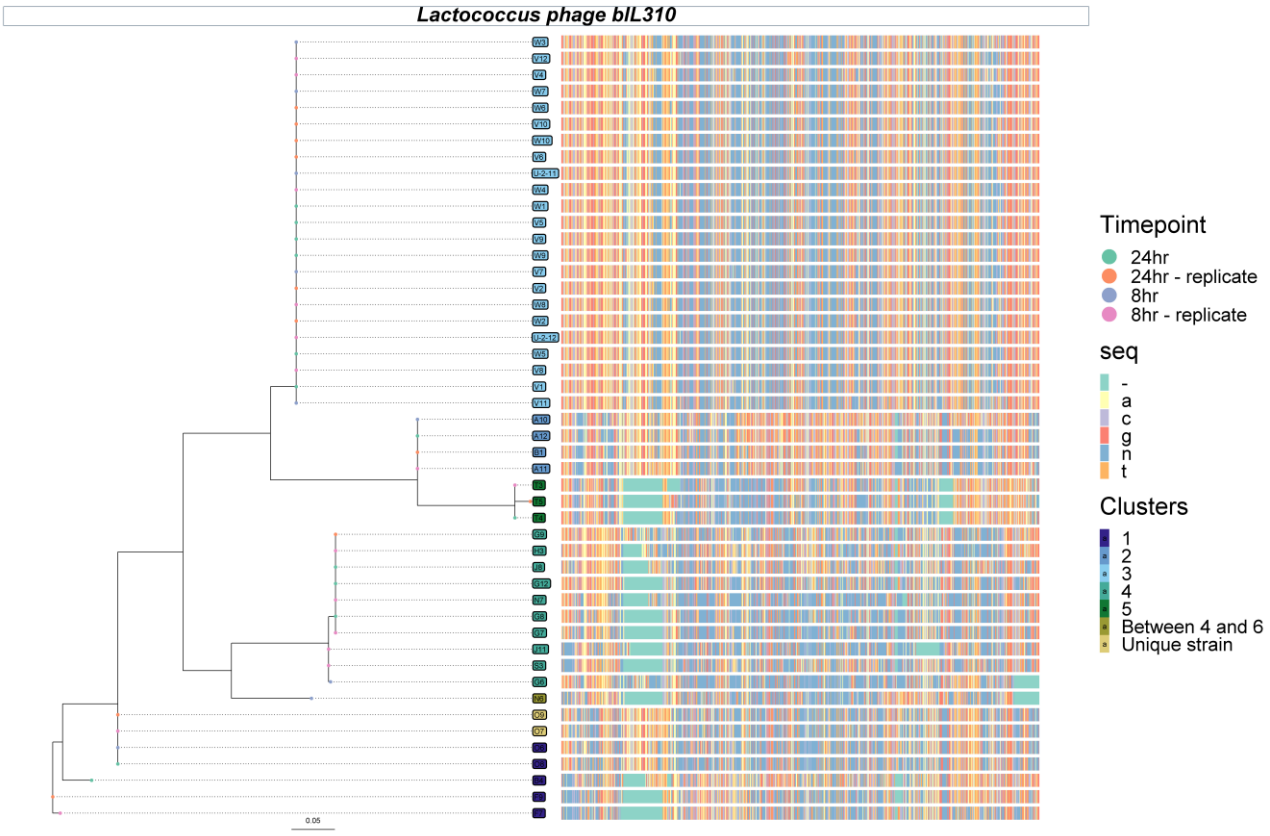


Figure S13: Isolates of *Lla. phage bIL310* strains detected across kefir metagenomes. Strain level phylograms of *Lla. phage bIL310* strains and corresponding multiple sequence alignment (MSA) files generated by StrainPhlAn 3. Coloured tips of the phylogram corresponds to the time point of the kefir metagenome from which the strain was recovered. Colouring of labels represents the strain subcluster by which the strain was grouped.

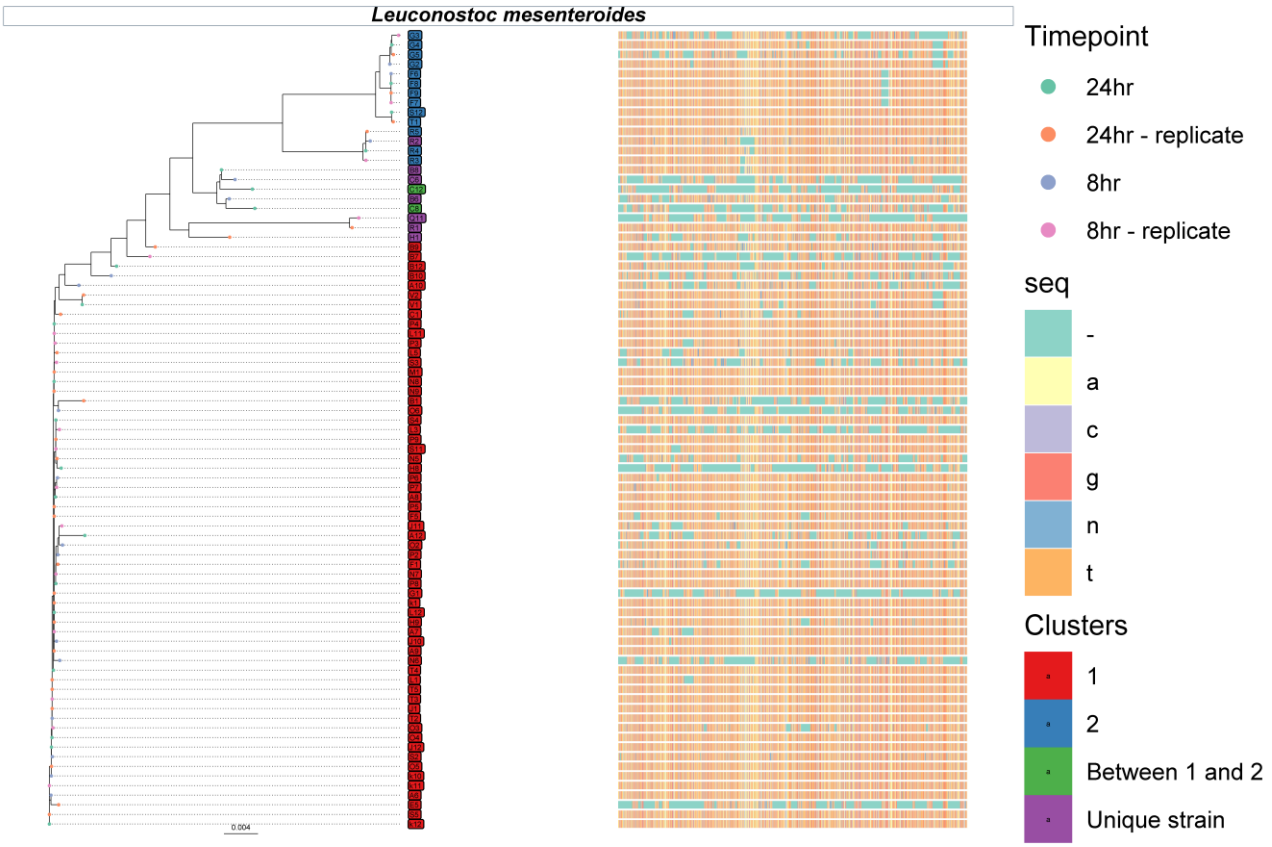


Figure S14: Strains of *Leuc. mesenteroides* detected across kefir metagenomes. Strain level phylograms of *Leuc. mesenteroides* strains and corresponding multiple sequence alignment (MSA) files generated by StrainPhlAn 3. Coloured tips of the phylogram corresponds to the time point of the kefir metagenome from which the strain was recovered. Colouring of labels represents the strain subcluster by which the strain was grouped.

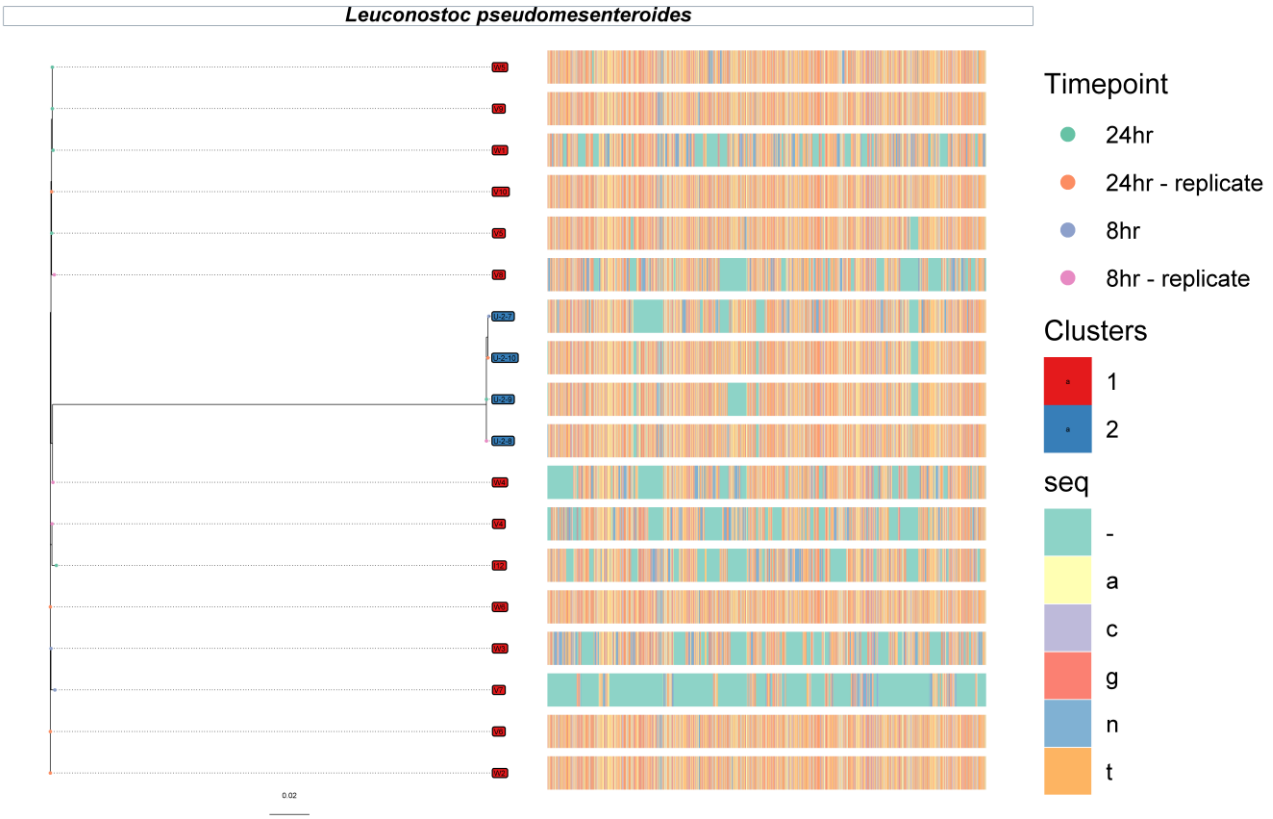


Figure S15: Strains of *Leuc. pseudomesenteroides* detected across kefir metagenomes. Strain level phylograms of *Leuc. pseudomesenteroides* strains and corresponding multiple sequence alignment (MSA) files generated by StrainPhlAn 3. Coloured tips of the phylogram corresponds to the time point of the kefir metagenome from which the strain was recovered. Colouring of labels represents the strain subcluster by which the strain was grouped.

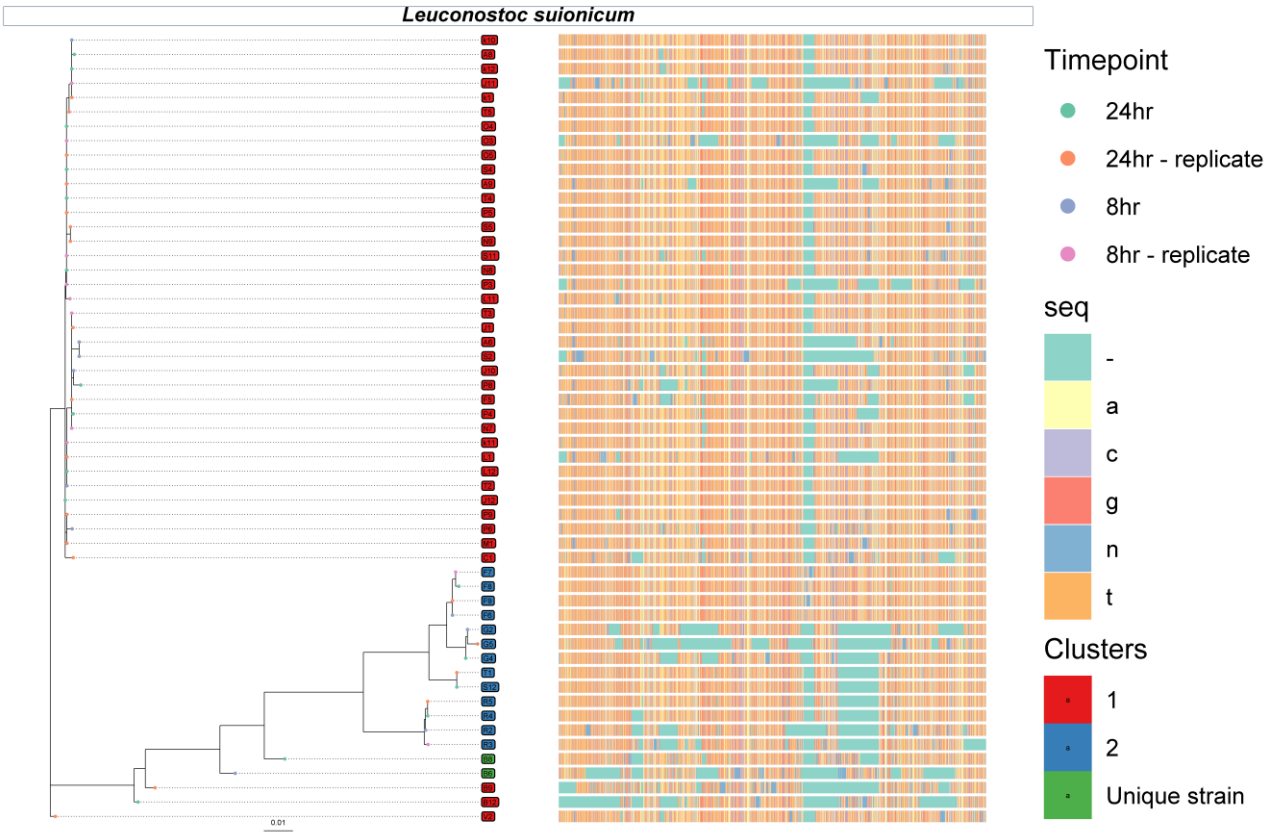


Figure S16: Strains of *Leuc. suionicum* detected across kefir metagenomes. Strain level phylograms of *Leuc. suionicum* strains and corresponding multiple sequence alignment (MSA) files generated by StrainPhlAn 3. Coloured tips of the phylogram corresponds to the time point of the kefir metagenome from which the strain was recovered. Colouring of labels represents the strain subcluster by which the strain was grouped.

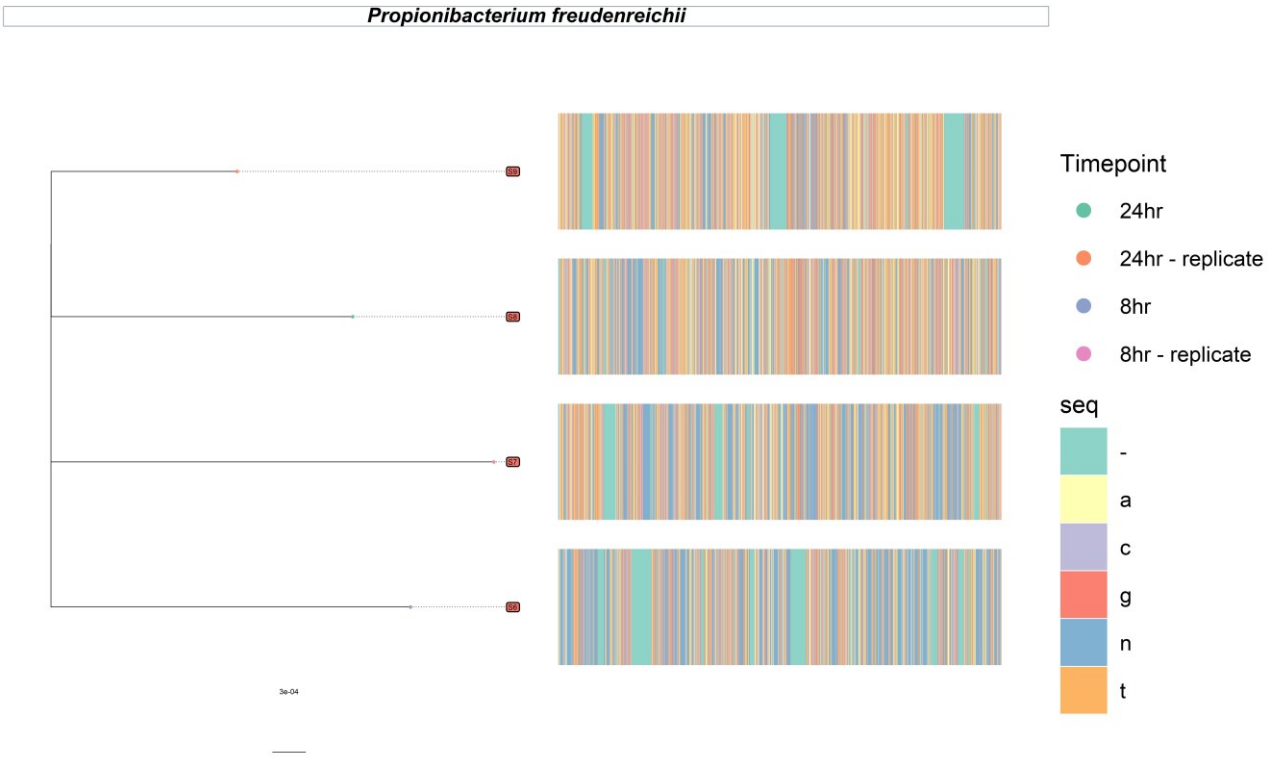


Figure S17: Strains of *Pr. freudenreichii* detected across kefir metagenomes. Strain level phylograms of *Pr. freudenreichii* strains and corresponding multiple sequence alignment (MSA) files generated by StrainPhlAn 3. Coloured tips of the phylogram corresponds to the time point of the kefir metagenome from which the strain was recovered. Colouring of labels represents the strain subcluster by which the strain was grouped.

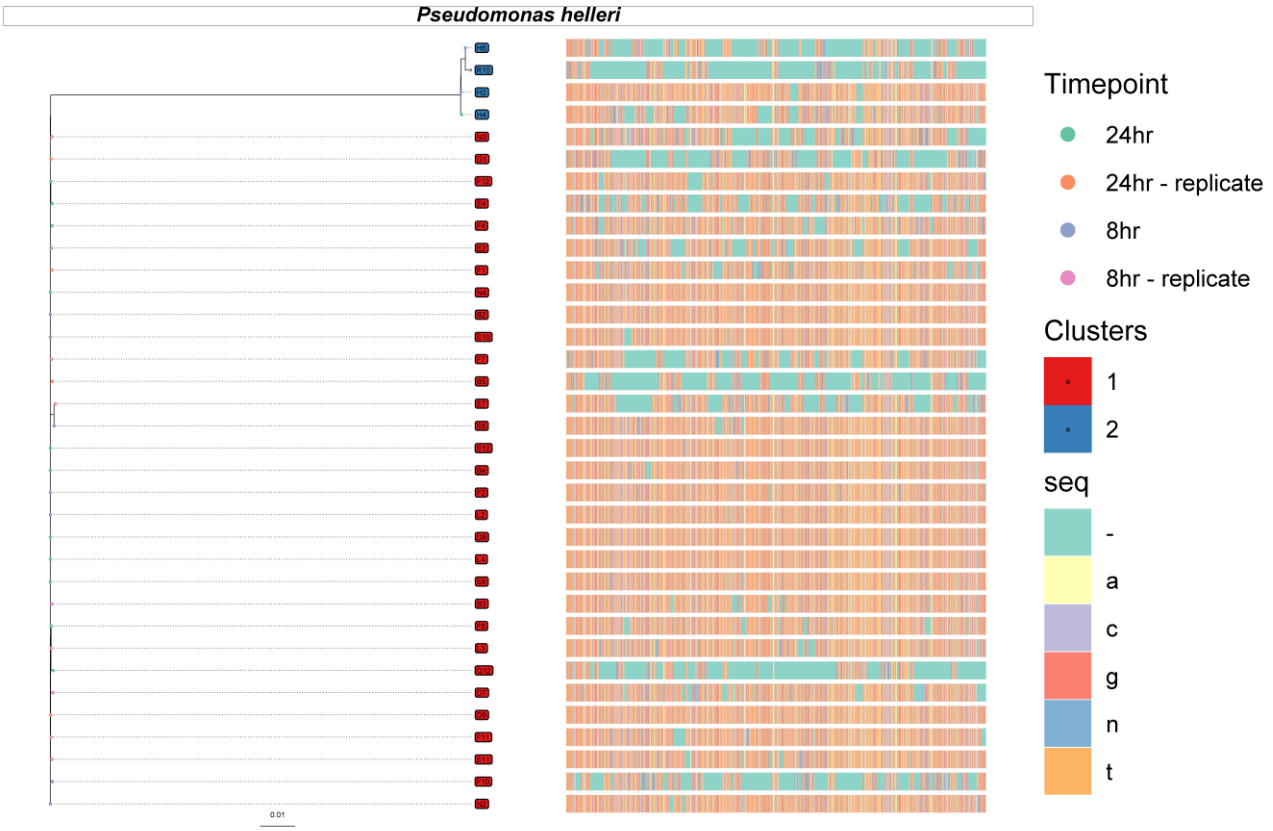


Figure S18: Strains of *P. helleri* detected across kefir metagenomes. Strain level phylograms of *P. helleri* strains and corresponding multiple sequence alignment (MSA) files generated by StrainPhlAn 3. Coloured tips of the phylogram corresponds to the time point of the kefir metagenome from which the strain was recovered. Colouring of labels represents the strain subcluster by which the strain was grouped.

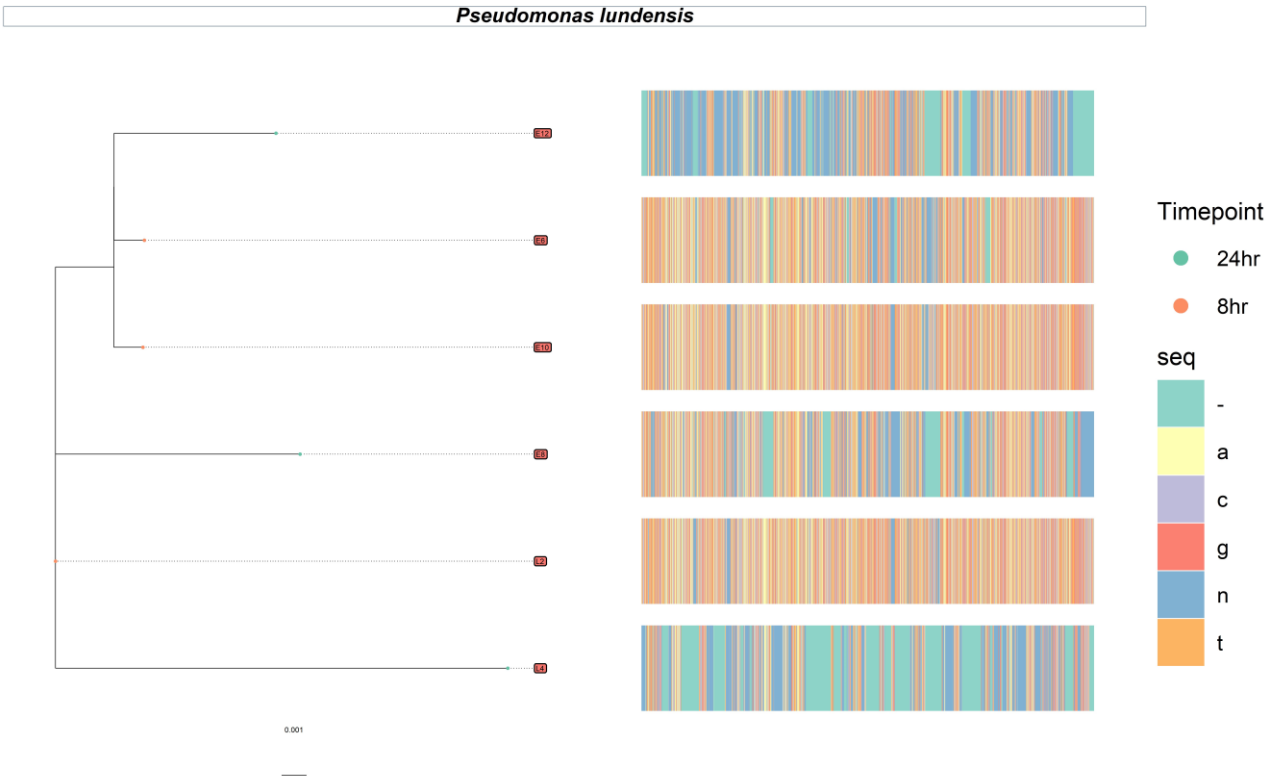


Figure S19: Strains of *P. lundensis* detected across kefir metagenomes**.** Strain level phylograms of *P. lundensis* strains and corresponding multiple sequence alignment (MSA) files generated by StrainPhlAn 3. Coloured tips of the phylogram corresponds to the time point of the kefir metagenome from which the strain was recovered. Colouring of labels represents the strain subcluster by which the strain was grouped.

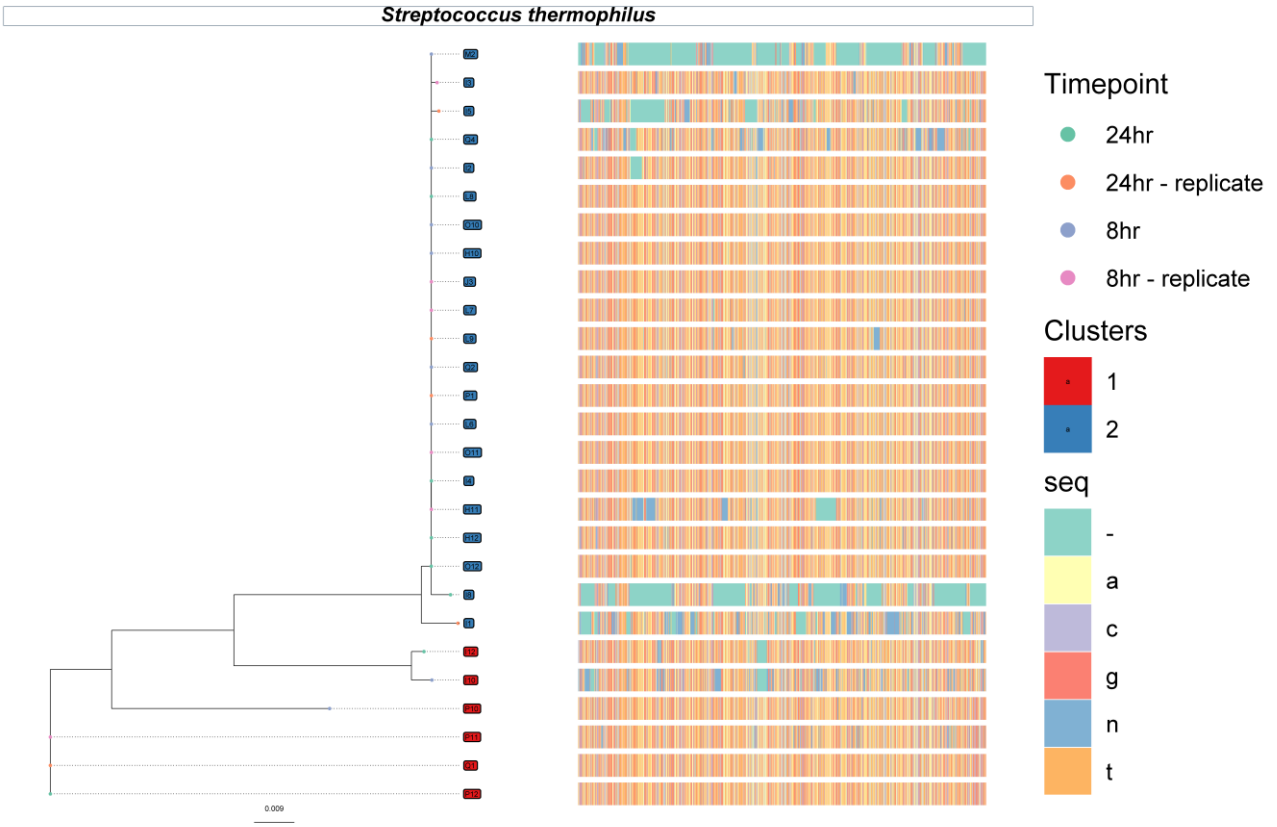


Figure S20: Strains of *S. thermophilus* detected across kefir metagenomes. Strain level phylograms of *S. thermophilus* strains and corresponding multiple sequence alignment (MSA) files generated by StrainPhlAn 3. Coloured tips of the phylogram corresponds to the time point of the kefir metagenome from which the strain was recovered. Colouring of labels represents the strain subcluster by which the strain was grouped.



Figure S21: Pangenome of *A. fabarum* strains detected within the milk kefir microbiome. Representative heatmap representing the pangenome of *A. fabarum* strains*;* rows represent PanPhlAn 3 detected strains; columns represent the gene families identified per strain. Row side annotations display the grouping of each gene family, e.g., core gene or specific to cluster 1. Column side annotations display the assigned cluster of the detected strain according to the StrainPhlAn 3 and PanPhlAn 3 clustering methodology (see methods) and the time point that the strain was recovered from.

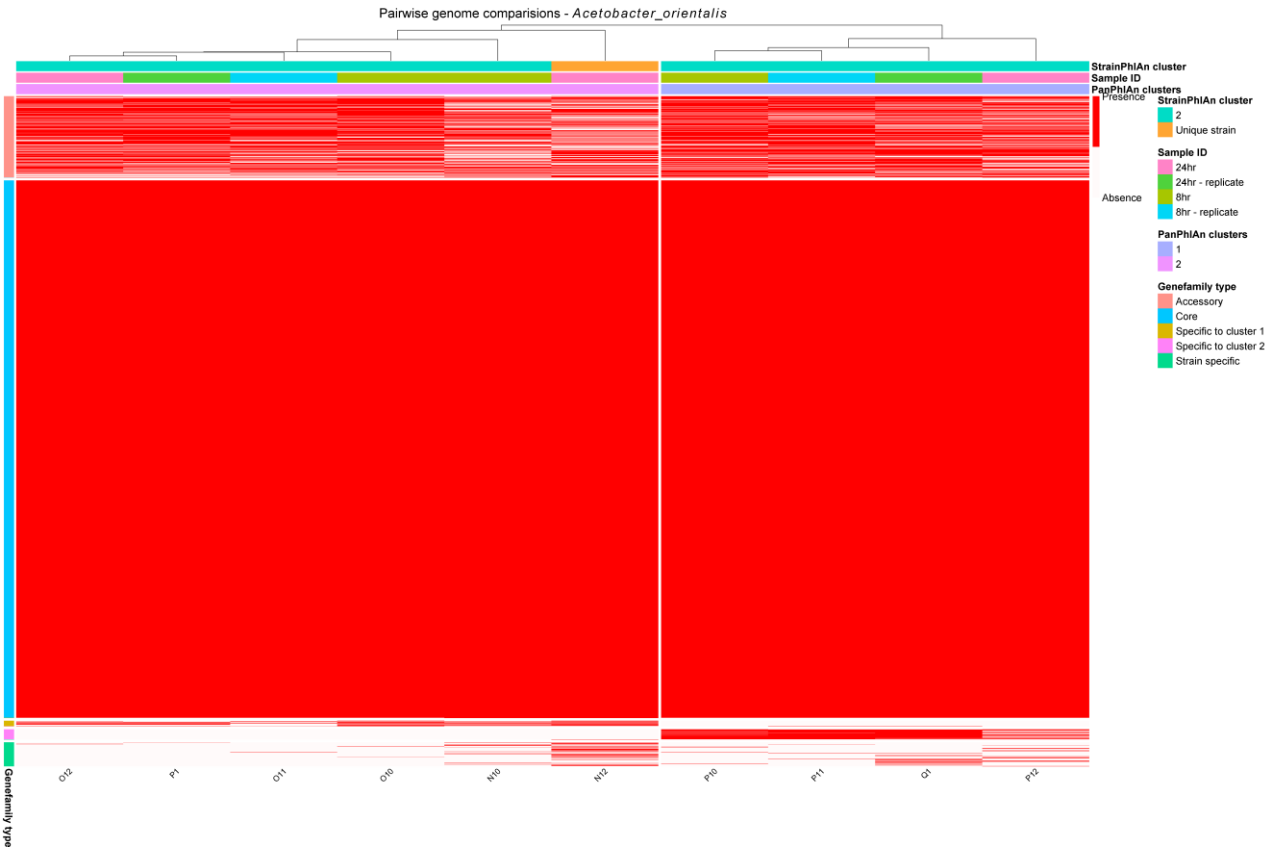


Figure S22: Pangenome of *A. orientalis* strains detected within the milk kefir microbiome. Representative heatmap representing the pangenome of *A. orientalis* strains*;* rows represent PanPhlAn 3 detected strains; columns represent the UniRef.90 gene families identified per strain. Row side annotations display the grouping of each gene family, e.g., core gene or specific to cluster 1. Column side annotations display the assigned cluster of the detected strain according to the StrainPhlAn 3 and PanPhlAn 3 clustering methodology (see methods) and the time point that the strain was recovered from.

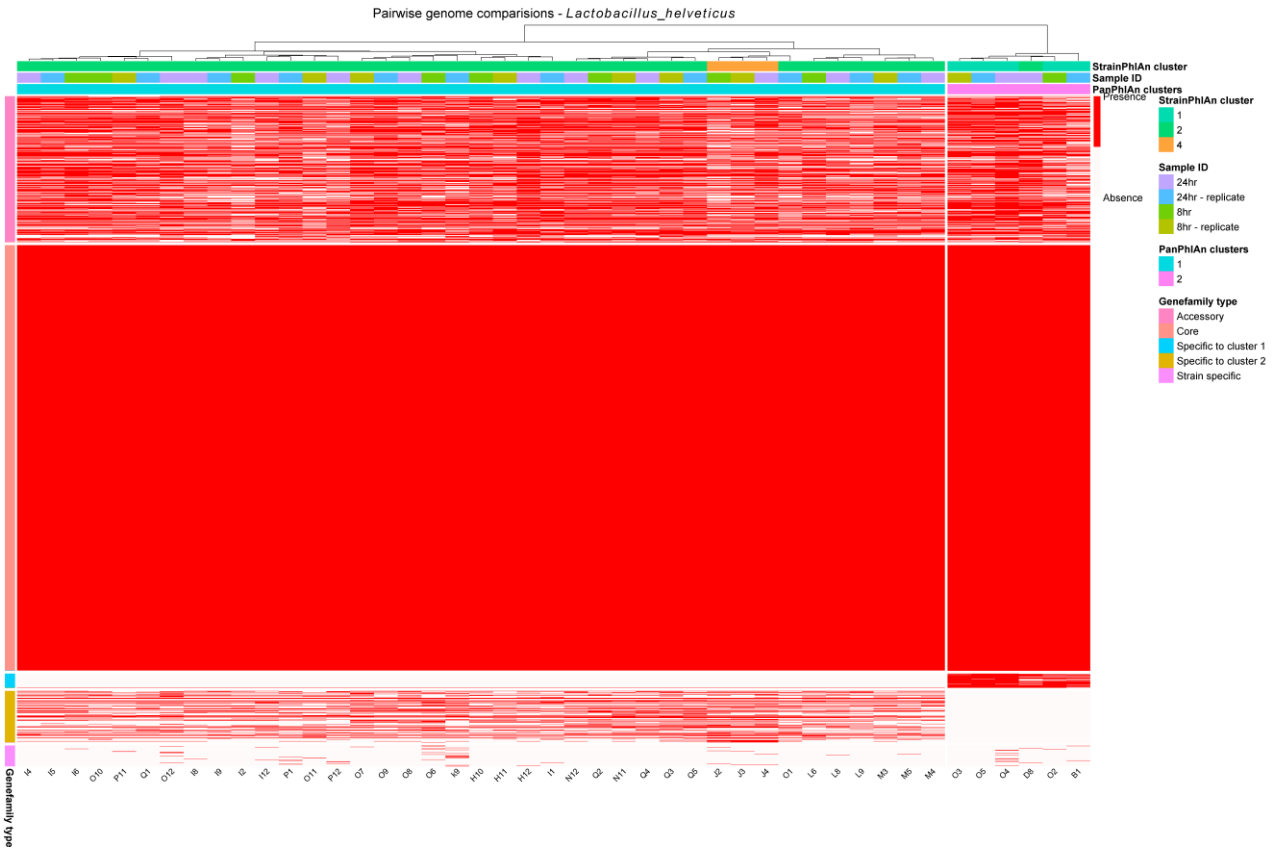


Figure S23: Pangenome of *L. helveticus* strains detected within the milk kefir microbiome. Representative heatmap representing the pangenome of *L. helveticus* strains*;* rows represent PanPhlAn 3 detected strains; columns represent the UniRef.90 gene families identified per strain. Row side annotations display the grouping of each gene family, e.g., core gene or specific to cluster 1. Column side annotations display the assigned cluster of the detected strain according to the StrainPhlAn 3 and PanPhlAn 3 clustering methodology (see methods) and the time point that the strain was recovered from.

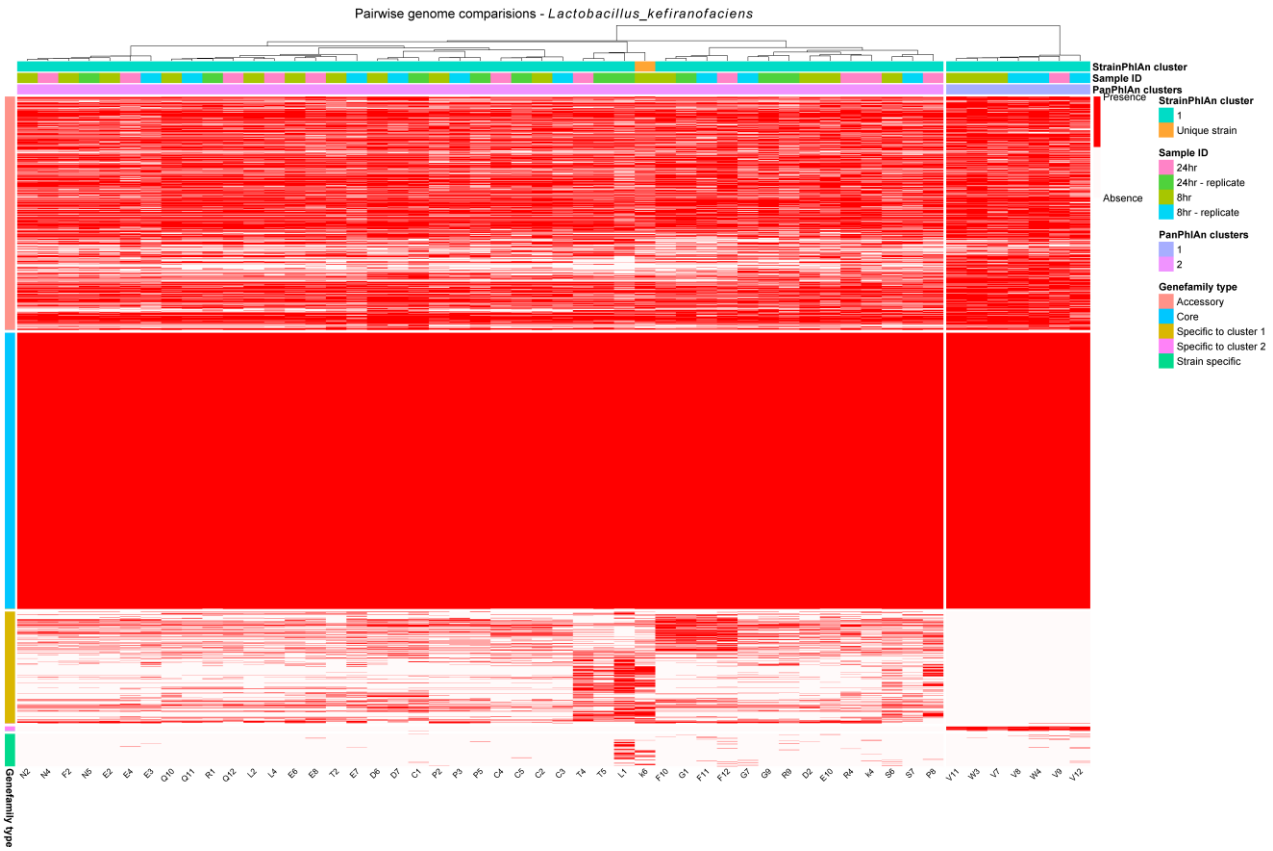


Figure S24: Pangenome of *L. kefiranofaciens* strains detected within the milk kefir microbiome. Representative heatmap representing the pangenome of *L. kefiranofaciens* strains*;* rows represent PanPhlAn 3 detected strains; columns represent the UniRef.90 gene families identified per strain. Row side annotations display the grouping of each gene family, e.g., core gene or specific to cluster 1. Column side annotations display the assigned cluster of the detected strain according to the StrainPhlAn 3 and PanPhlAn 3 clustering methodology (see methods) and the time point that the strain was recovered from.

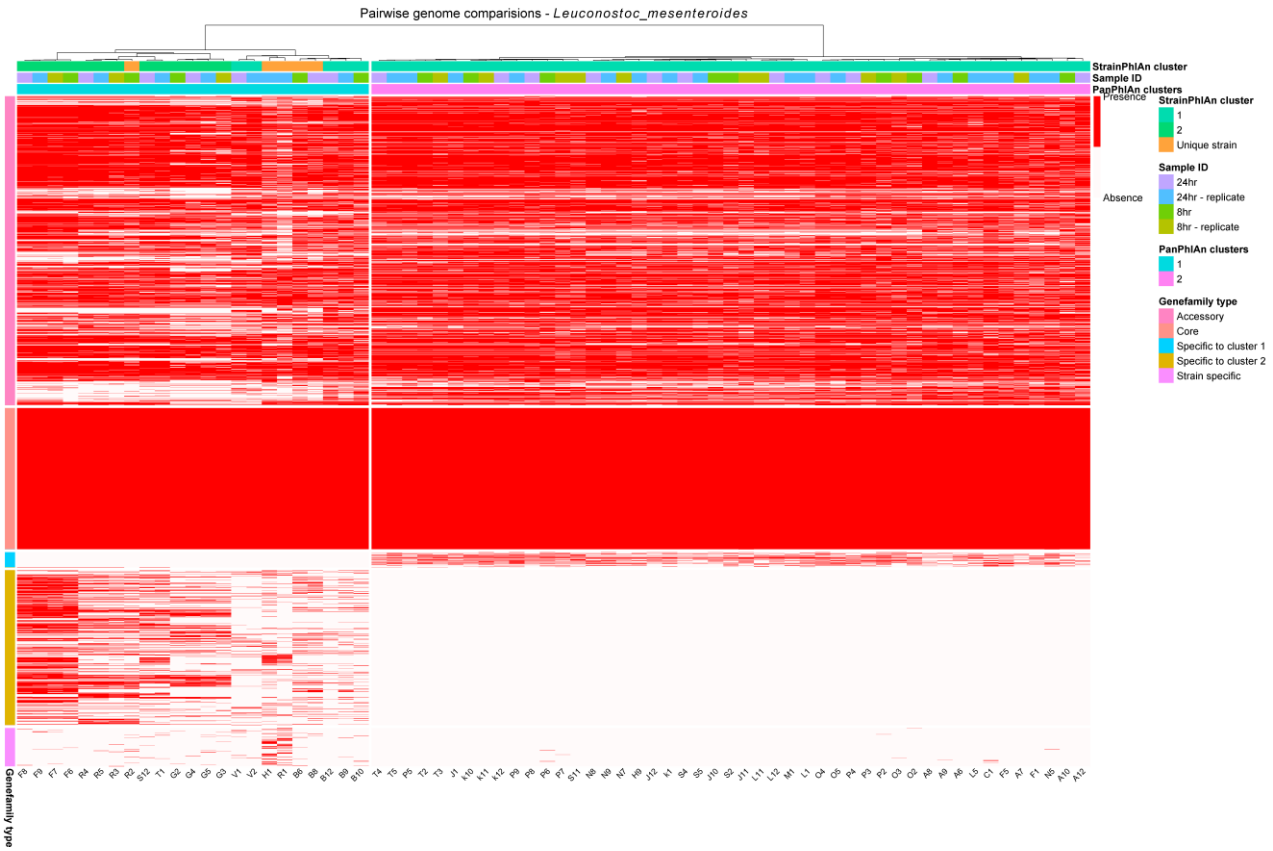


Figure S25: Pangenome of *Leuc. mesenteroides* strains detected within the milk kefir microbiome. Representative heatmap representing the pangenome of *Leuc. mesenteroides* strains*;* rows represent PanPhlAn 3 detected strains; columns represent the UniRef.90 gene families identified per strain. Row side annotations display the grouping of each gene family, e.g., core gene or specific to cluster 1. Column side annotations display the assigned cluster of the detected strain according to the StrainPhlAn 3 and PanPhlAn 3 clustering methodology (see methods) and the time point that the strain was recovered from.

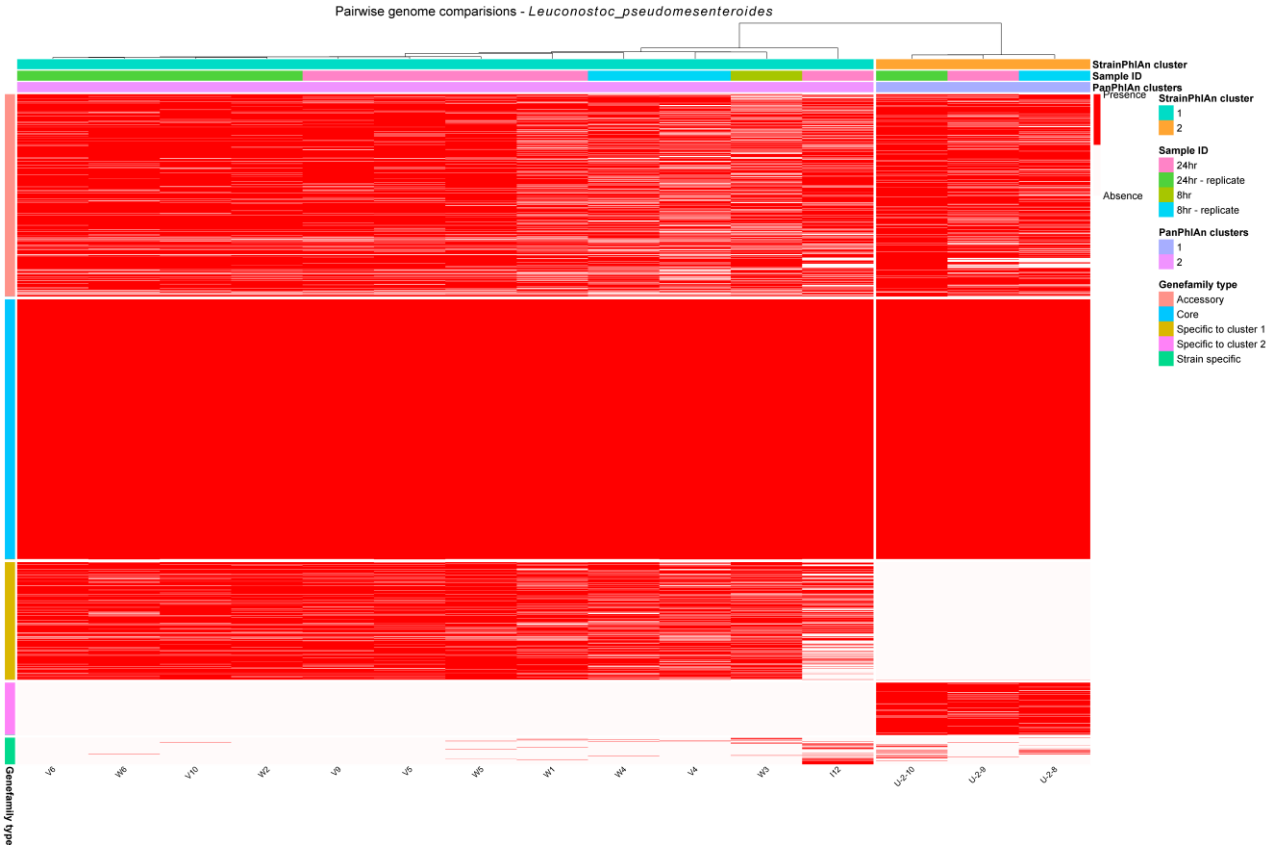


Figure S26: Pangenome of *Leuc. pseudomesenteroides* strains detected within the milk kefir microbiome. Representative heatmap representing the pangenome of *Leuc. pseudomesenteroides* strains*;* rows represent PanPhlAn 3 detected strains; columns represent the UniRef.90 gene families identified per strain. Row side annotations display the grouping of each gene family, e.g., core gene or specific to cluster 1. Column side annotations display the assigned cluster of the detected strain according to the StrainPhlAn 3 and PanPhlAn 3 clustering methodology (see methods) and the time point that the strain was recovered from.

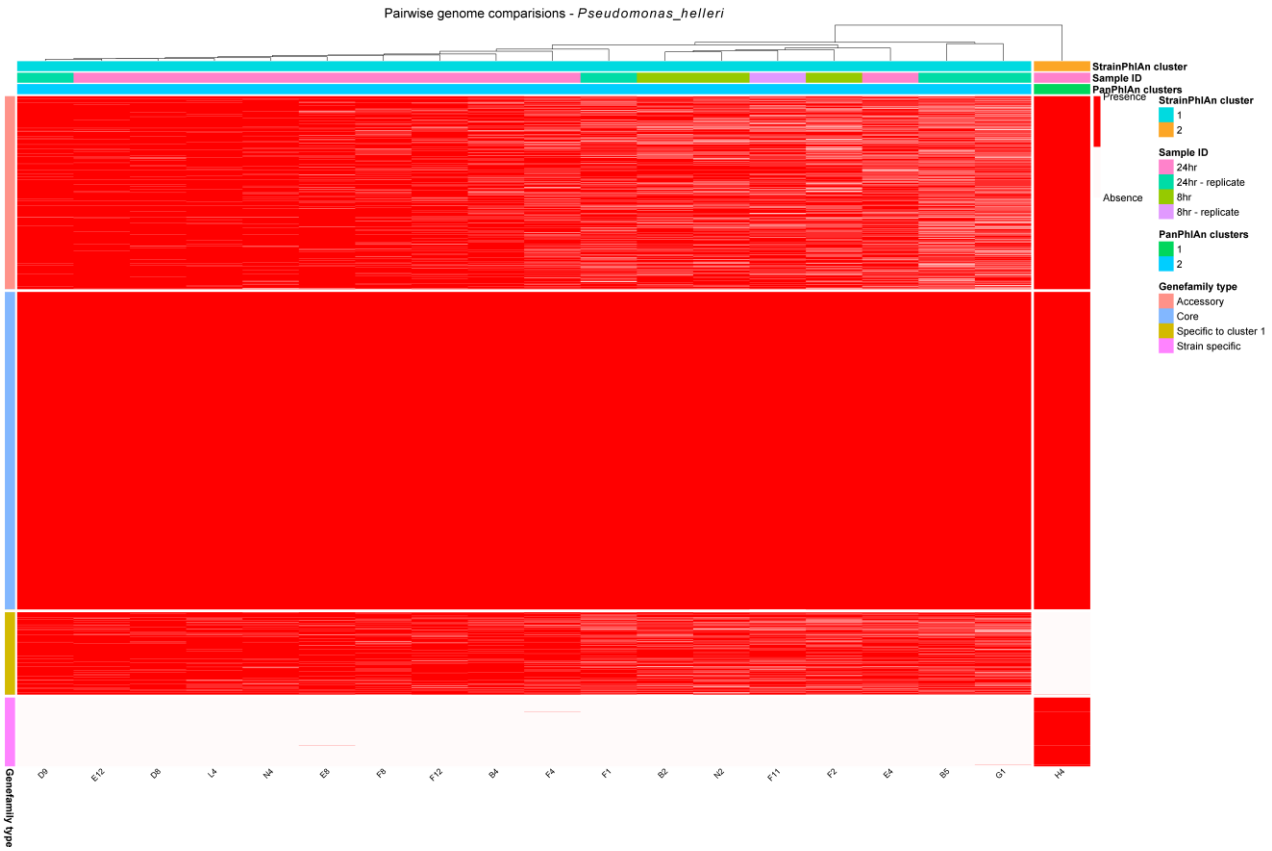


Figure S27: Pangenome of *P. helleri* strains detected within the milk kefir microbiome. Representative heatmap representing the pangenome of *P. helleri* strains; rows represent PanPhlAn 3 detected strains; columns represent the UniRef.90 gene families identified per strain. Row side annotations display the grouping of each gene family, e.g., core gene or specific to cluster 1. Column side annotations display the assigned cluster of the detected strain according to the StrainPhlAn 3 and PanPhlAn 3 clustering methodology (see methods) and the time point that the strain was recovered from.

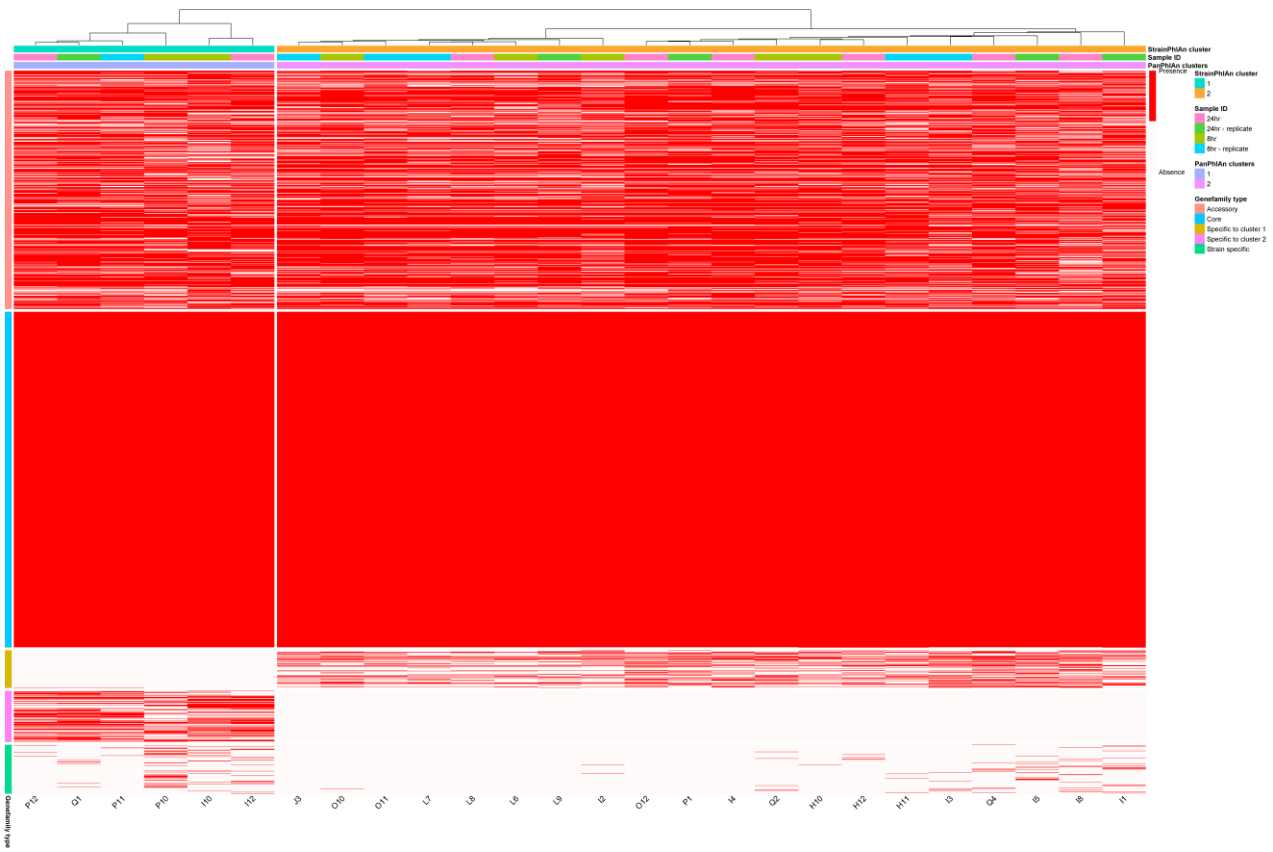


Figure S28: Pangenome of *S. thermophilus* strains strains detected within the milk kefir microbiome. Representative heatmap representing the pangenome of *S. thermophilus* strains; rows represent PanPhlAn 3 detected strains; columns represent the UniRef.90 gene families identified per strain. Row side annotations display the grouping of each gene family, e.g., core gene or specific to cluster 1. Column side annotations display the assigned cluster of the detected strain according to the StrainPhlAn 3 and PanPhlAn 3 clustering methodology (see methods) and the time point that the strain was recovered from.

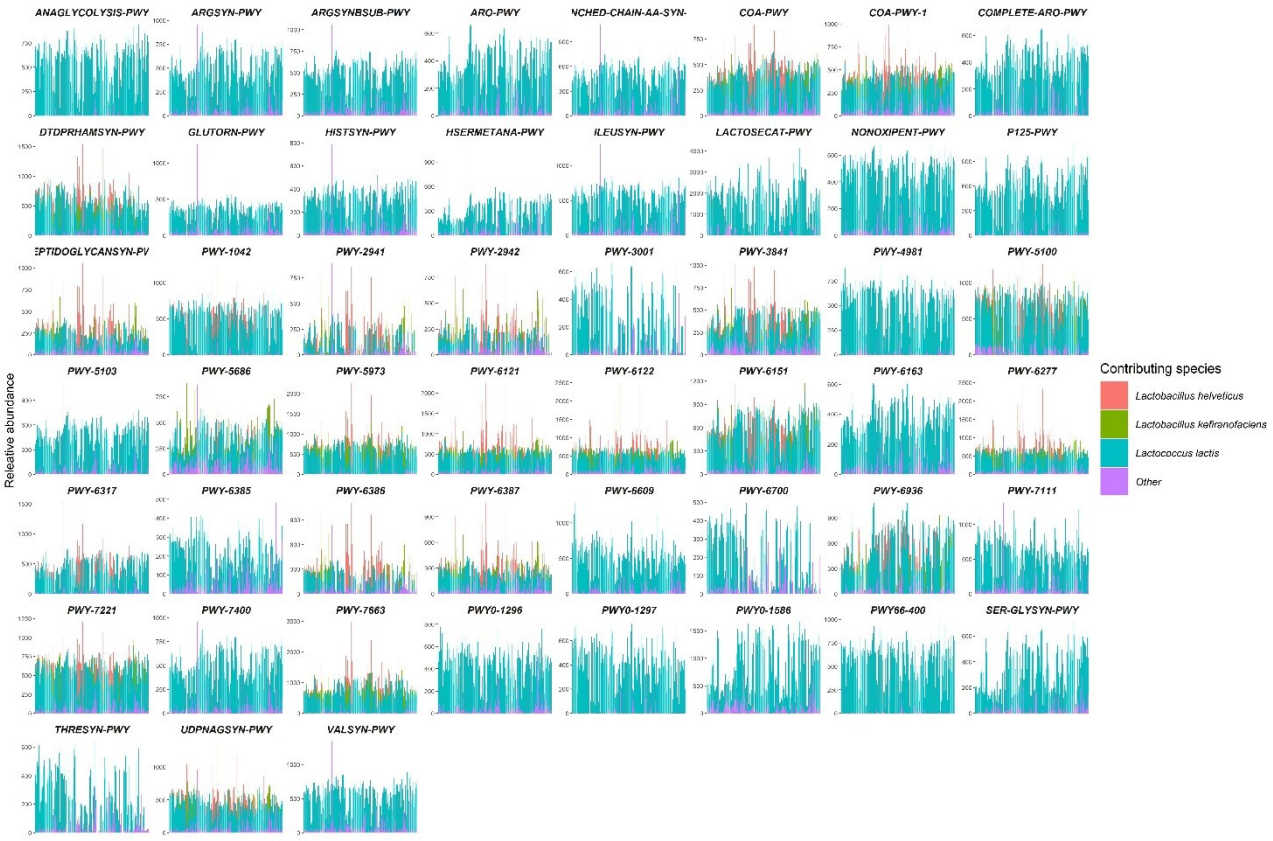


Figure S29: Core metabolic pathway and the associated species in the kefir microbiome. Stacked bar chart of species contributions to the relative abundance of each core metabolic pathway in the kefir microbiome across samples. The contributing species are coloured individually; as *L. helveticus*, *L. kefiranofaciens*, *Lla. lactis and* o

Chapter 3

Comparison of the relative impacts of acute consumption of an inulin-enriched diet, milk kefir or a commercial probiotic product on the human gut microbiome and metabolome

Published in *npj Science of Food*

(https://dx.doi.org/10.2139/ssrn.4423993)

Authors: Liam H. Walsh, Aaron M. Walsh, Isabel Garcia-Perez, Fiona Crispie, Adele Costabile, Richard Ellis, Jim Finlayson, Laura A. Finnegan, Marcus J. Claesson, Elaine Holmes, Paul D. Cotter.

Contributions

- **A. C**, **R.E**, **J.F**, **E.H**, and **P.D.C**. study conception and design
- **I.G.P** performed SCFAs and H NMR Metabolic profiling analysis.
- **F.C** and **L.A.F** performed DNA extraction, DNA library preparations and Illumina sequencing.
- **A.M.W** helped **Candidate**: with bioinformatics, statistical analysis, and data visualisation.
- **A.M.W** helped **Candidate** to draft and edit the manuscript.
- **P.D.C.** supervised **Candidate** throughout.

Section 3.0: Abstract

It has been established that the human gut microbiota is central to health, and, consequently, there has been a growing desire to positively modulate its composition and/or function through, for example, the use of fermented foods, prebiotics or probiotics. Here, we compare the relative impact of the daily consumption of an inulin-enriched diet (n=10), a commercial probiotic-containing fermented milk product (FMP) (n=10), or a traditional kefir FMP (n=9), over a 28-day period on the gut microbiome and urine metabolome of healthy human adults. None of the treatments resulted in significant changes to clinical parameters or biomarkers tested. However, shotgun metagenomic analysis revealed that kefir consumption resulted in a significant change in taxonomy, in the form of an increased abundance of the sub-dominant FMP-associated species *Lactococcus raffinolactis*, which further corresponded to shifts in the urine metabolome. Overall, our results indicated that daily consumption of a single portion of kefir alone resulted in detectable changes to the gut microbiota and metabolome of consumers.

Section 3.1: Introduction

It is becoming increasingly apparent that the microorganisms in the human gut are pivotal to many aspects of our health. Disturbance of the gut microbiota has been linked to a variety of diseases, including colon cancer, diabetes, inflammatory bowel disease, and obesity [1]. Consequently, the gut microbiota has emerged as a potential target for the prevention or treatment of such diseases. Targeted manipulation of the gut microbiota can be achieved by dietary intervention, including through the use of fermented foods, probiotics or prebiotics [2].

A probiotic is a live microorganism that exerts health benefits upon the host when it is ingested at sufficient quantities [3]. Many probiotic strains are consumed as constituents of yoghurt/fermented milk-type drinks or as supplements in the form of capsules. To access the lower gastrointestinal tract, probiotics must first survive transit through the acidic upper gastrointestinal tract. Some probiotic strains can subsequently colonise the gut by adhering to the intestinal epithelial cells [4]. A number of potential mechanisms can contribute to the health benefits of specific strains. These include pathogen inhibition via bacteriocin production or mucosal competitive exclusion, and immunomodulation [5]. Human studies have demonstrated that probiotic strains, such as *Bifidobacterium infantis* 35624, *Lacticaseibacillus casei* DN-114001, and *Lacticaseibacillus casei* Shirota, confer health benefits, including alleviation of the symptoms of irritable bowel syndrome, prevention of antibiotic-associated diarrhoea, or protection against infection [6-8].

Health benefits have also been attributed to several traditional fermented foods and the microbes therein [9]. Particular examples of kefir, a traditional fermented milk product (FMP), have been linked to health benefits, including anti-cholesterolemic, anti-inflammatory, and anti-pathogenic effects [10-12], and several investigations indicate that specific microorganisms in kefir contribute to these effects [13-15]. Despite this promise, it should be noted there has been a relative lack of research into the effects of kefir on human health, with most investigations relying on *in vitro* or animal models [16].

Prebiotics can also beneficially influence the gut microbiota. Most prebiotics are non-digestible oligosaccharides that stimulate the growth and or metabolic activity of health-promoting commensal

microorganisms in the gut [17]. The most frequently studied prebiotics are fructooligosaccharides, galactooligosaccharides, and inulin [18]. Prebiotics naturally occur in various foods, including artichokes, chicory and wheat, but can also be provided in the form of food supplements [19]. Human studies have demonstrated that prebiotics can alter the gut microbiome, with many studies reporting an increase in *Bifidobacterium* and or members of the former genus *Lactobacillus* [20], as well as confering various health benefits, including improved satiety, lowered insulin concentrations, and reduced infection [21-23]. Metagenomic and metabolomic sequencing efforts are beginning to improve our understanding of the effects of probiotics, prebiotics and fermented foods on the human gut microbiota and metabolome [24, 25]. However, the majority of studies have been directed at preventing or treating the symptoms of specific diseases [26] and/or evaluating the impact of consumption of high quantities of a probiotic(s), prebiotic(s) and/or fermented food(s) [27]. Few studies to date have analysed the relative impact of supplementing diets with moderate portions of one of these products on the microbiota and metabolome of healthy participants [28].

In the present study, we use metagenomic sequencing and metabolomics to identify changes in the human gut microbiota and metabolome of healthy participants following daily consumption, over 28 days, of either a single portion of a commercial FMP containing the probiotic *Lc. casei*, a traditional fermented milk beverage kefir or a diet enriched with the prebiotic inulin.

Section 3.2: Methods

3.2.1 Participant enrolment, intervention design, urine and stool collection

We recruited 29 healthy volunteers, aged from 18 to 65, for this study. Volunteers were members of the public and were enrolled from the Centre for Health Science Inverness (https://www.nes.scot.nhs.uk/contact-us/centre-for-health-science-inverness/). Volunteers provided written informed consent to take part in the study. Volunteers had not consumed probiotic-containing products, including commercial probiotic drinks or probiotic supplements, within the 6 weeks preceding the study. They had not taken antibiotics, antacids or proton pump inhibitors within the 2 weeks

preceding the study. Volunteers did not have any of the following medical conditions: inflammatory bowel disease (IBD), irritable bowel syndrome (IBS), celiac disease, food allergies, gastroenteritis (within the 4 weeks preceding the study), heart valve abnormalities, prior rheumatic fever, diabetes or any other immune disorder. Volunteers did not follow any other dietary or lifestyle recommendations during this intervention. After a 7 day run-in period (Day -7 to Day 0), volunteers consumed one of the following treatments daily over a 28 day intervention period (Day 0 to Day 28): 7 g of inulin (n=10), 247ml of traditional kefir (n=9), and 65 ml of a commercial probiotic (*Lc. casei* Shirota)-containing dairy beverage (n=10). The inulin group was asked to consume additional portions of foods such as vegetables and fruits that constituted part of their normal diet and contained fibre, which amounted to 7 grams of natural inulin. 7 grams of natural inulin were selected based on the GI tolerance [52], as short- and long-term consumptions of inulin, given at a daily dose containing at least 5 g of inulin, is reported to be well tolerated by healthy subjects [53]. The kefir group were asked to consume a kefir milk product, produced by the inoculation of milk with a kefir grain. Kefir samples were kindly provided by Nourish Kefir (https://www.nourishkefir.co.uk/buy-kefir-here/). Lastly, the probiotic group where asked to consume a commercially available probiotic yoghurt style drink, containing live cultures of *Lc. casei* Shirota, Each commercial probiotic contained a minimum of 2×10^{10} CFUs, according to the manufacturers (https://www.hcp.yakult.co.uk/). Subjects were controlled weekly via phone call for possible side effects and compliance in achieving the dietary requirements. Stool and urine samples were collected on Day 0 and Day 28. NHS Ethical Approval was not required as this study involved healthy volunteers only, who provided authorisation for the use of their bodily matter for research purposes in accordance with the NHS research governance systems (https://www.nhsresearchscotland.org.uk/services/research-governance#:~:text=Research%20Governance%20concerns%20setting%20standards,preventing%20poor%20performance%20and%20misconduct). R&D Management Approval was obtained from the NHS Highland R&D Department. The study was conducted according to the Declaration of Helsinki following Good Clinical Practice (GCP).

3.2.2 Clinical measurements

Several clinical parameters were measured before and after treatment, including height, weight, and body fat. Anxiety was assessed using the GAD-7 questionnaire [54], while quality of life was assessed using the EQ-5D questionnaire [55]. Abdominal symptoms regarding bloating, flatulence and bowel habit were also recorded. Stool consistency was classified using the Bristol Stool Scale (BSS) [56].

3.2.3 Microbial DNA extraction from faeces

Microbial DNA was extracted from 250 mg of faecal samples, using the QIAamp DNA Stool Mini Kit (QIAGEN, UK) according to the protocol described in [57].

3.2.4 High-throughput DNA sequencing

Whole-metagenome shotgun libraries were prepared using the Nextera XT kit in accordance with the Nextera XT DNA Library Preparation Guide from Illumina, with the exception that tagmentation time was increased to 7 min. Samples were sequenced on the Illumina NextSeq 500 in the Teagasc sequencing facility, with a NextSeq 500/550 High Output Reagent Kit v2 (300 cycles), in accordance with standard Illumina sequencing protocols.

3.2.5 Bioinformatics

Raw whole-metagenome shotgun sequencing reads derived from faecal samples collected in this study and kefir samples generated by Walsh *et al* [29], were quality filtered and trimmed using a combination of Picard Tools v2.18.23 (https://github.com/broadinstitute/picard) and SAMtools v1.10 [58]. Kraken2 v 2.1.1 [59] was used to determine the species-level microbial composition of the gut and kefir microbiome. Relative abundance of each species among samples was calculated using Bracken v2.2 [60]. To distinguish between false positive and present taxa, the Kraken2 script was performed using the parameter -report-minimizer-data [59]. Identifications displaying a high distinct minimiser count but accounting for <1 % relative abundance were confirmed as true positive/negative results using the bioinformatics tool raspir v101 following the methodology outlined in [31]. Microbial pathway analysis was performed using

HUMAnN3 v3.1.1. The pan- genome of *Lactococcus raffinolactis* (*Lla. raffinolactis*) was constructed by Roary v3.13 [61], using complete reference genomes available in NCBI [62]. Kefir-derived metagenome assembled genomes of the species *Lla. raffinolactis* were acquired from [29]. Meta(genomes) were clustered using dREP v3.2.0 [63] (Figure 2). Meta(genome) annotation was performed using DRAM v1.2 [64] and Prokka v1.14 [65]. A list of genes associated with survival/colonisation associated activity were sourced from Mils *et al.* [66] , Leech *et al.* [67] and Kim *et al.* [68] and used to examine the host adaptability potential of meta(genomes).

3.2.6 ^{1}H NMR Metabolic profiling analysis

Urine samples were stored at -80^{o}C prior to analysis. Urine samples were thawed at 4^{o}C and then vortexed and centrifuged at 1600 × g for 10 min to remove precipitated proteins and particulates. An aliquot of each urine sample (540 µL) was mixed with 60 µL of phosphate buffer (pH 7.4, 80% H_2O) containing 1 mM of the internal standard, 3-(trimethylsilyl)-[2,2,3,3,-2H4]-propionic acid (TSP) and 2mM sodium azide (Na^3N), as described previously [69]. Urine samples were analyzed in 96-well plates containing one quality control (QC) sample every ten samples. QC samples were prepared by pooling 20µl volumes of each urine sample. During the analyses, samples were maintained at 4^{o}C in the autosampler. ^{1}H NMR spectroscopy was performed at 300 K on a Bruker 600 MHz spectrometer (Bruker BioSpin, Karlsruhe, Germany) using the following standard one-dimensional pulse sequence with saturation of the water resonance: RD−gz,1−90°−t−90°−tm−gz,2−90°−ACQ, where RD is the relaxation delay, t is a short delay typically of about 4 µs, 90° represents a 90° radio frequency (RF) pulse, tm is the mixing time (10 ms), gz,1 and gz,2 are magnetic field z-gradients both applied for 1 ms, and ACQ is the data acquisition period (2.7s). Water suppression was achieved through continuous wave irradiation at the water resonance frequency using 25 Hz RF strength during RD and also during tm. The receiver gain was set to 90.5 for all experiments. Each urine spectrum was acquired using 4 dummy scans, 32 scans, 64K time domain points and with a spectral window set of 20 ppm. Prior to Fourier transformation, the free induction decays were multiplied by an exponential function corresponding to a line broadening of 0.3 Hz.

The ^{1}H NMR spectra were digitized over the range of δ10.0 to −0.5 and imported into MATLAB (2014a, Mathworks Inc., USA), and automatically corrected for phase and baseline distortions and referenced to the TSP singlet at δ 0.0 using TopSpin 3.1 software. Spectra were then digitized into 20k data points at a resolution of 0.0005ppm using an in-house MATLAB R2014a (Mathworks) script. Subsequently, spectral regions corresponding to the internal standard (δ -0.5 to 0.5) and water (δ 4.6 to 5) peaks were removed. In addition, urea (δ 5.4 to 6.3) was removed from the urinary spectra. Spectra were normalized using median fold change normalization to the median spectrum [70]. The resulting ^{1}H-NMR spectrum were imported into MATLAB to conduct multivariate statistical analysis. Data was centered and scaled to account for the repeated-measures design and then modelled using Partial Least Squares - Discriminant Analysis (PLS - DA) with Monte-Carlo cross-validation (MCCV) [71]. The fit and predictability of the models obtained was determined by the R-squared and Q-squared values, respectively.

3.2.8 Short chain fatty acids (SCFAs) analysis

The GC-MS targeted SCFA analysis was conducted on an Agilent 7000C Triple Quadrupole GC/MS System according to a previously published method (24). To interrogate the SCFAs results the non-parametric Wilcoxon's signed rank test was used to assess pairwise differences.

3.2.9 Statistics

Statistical analysis was performed in R-3.2.2 [72]. The Wilcoxon -signed rank test was used to identify significant changes in the gut microbiota of faecal samples collected on Day 0 (i.e., before consumption) versus Day 28 (i.e., after consumption) (Figure 1b). Resulting p-values were corrected for multiple comparisons using the Benjamini-Hochberg (BH) method. A result was considered statistically significant if the adjusted p-value was less than 0.1. The significance threshold was relaxed to account for the high number of comparisons. The meta.mds function in the vegan package [73] was used for multidimensional scaling (MDS) analysis, while the adonis function, also in the vegan package, was used for PERMANOVA analysis (Figure 1a & Supplementary figure 1). Data visualisation of summary statistics generated from the above tests was performed using the ggplot2 package [74].

Section 3.3: Results

3.3.1 Clinical outcomes

For this study, we investigated clinical and anthropomorphic parameters, the composition and functional potential of the gut microbiota and metabolome profile of healthy participants following daily consumption (28 days) of either a commercial probiotic-containing FMP, a traditional kefir FMP or a diet enriched with the prebiotic inulin. Several parameters were measured before and after treatment and none of the treatments resulted in significant changes to the clinical parameters or biomarkers tested, i.e., percentage body fat, anxiety, and abdominal symptoms (Table S1).

3.3.2 Compositional analysis reveals an increased abundance of *Lactococcus raffinolactis* after kefir consumption

Shotgun metagenomics was used to compare the gut microbial diversity and taxonomic abundance for all participants before versus after treatment, using the taxonomic profiling tool Kraken 2. When considering microbial diversity, no significant changes were observed, as determined using statistical analysis and MDS with no perceivable shift in the composition of the gut following inulin (PERMANOVA: p=0.94, R^2=2.2%), kefir (PERMANOVA: p=0.99, R^2=1.6%), or commercial FMP (PERMANOVA: p=0.99, R^2=1.5%) (Figure 1a & b), Shotgun metagenomics was used to determine which of the species detected in the gut microbiome of participants were differentially abundant after consumption of either inulin, kefir or a commercial FMP. Species were selected based on the reported microbial compositions of the probiotic (https://www.hcp.yakult.co.uk/) and kefir [29] and through a literature search for inulin, for example, inulin has previously been shown to modulate the genera *Anaerostipes*, *Bifidobacterium*, and *Bilophila* [30]. No differentially abundant species were observed in the commercial FMP or inulin group. However, in the kefir group we observed the increased relative abundance of *Lactococcus raffinolactis* in 4 of the 10 participants, which was below detectable levels prior to kefir consumption (Figure 1c). The presence of *Lla. raffinolactis* in these samples was confirmed using the alignment based approach raspir (rare species identifier) [31]. *Lla. raffinolactis* is a prevalent microorganism within the milk kefir microbiome, taxonomic

profiling of 256 kefir milk metagenomes revealed its presence in 169 of those metagenomes at a relative abundance of 0.1 to 9%. (Figure 2a)

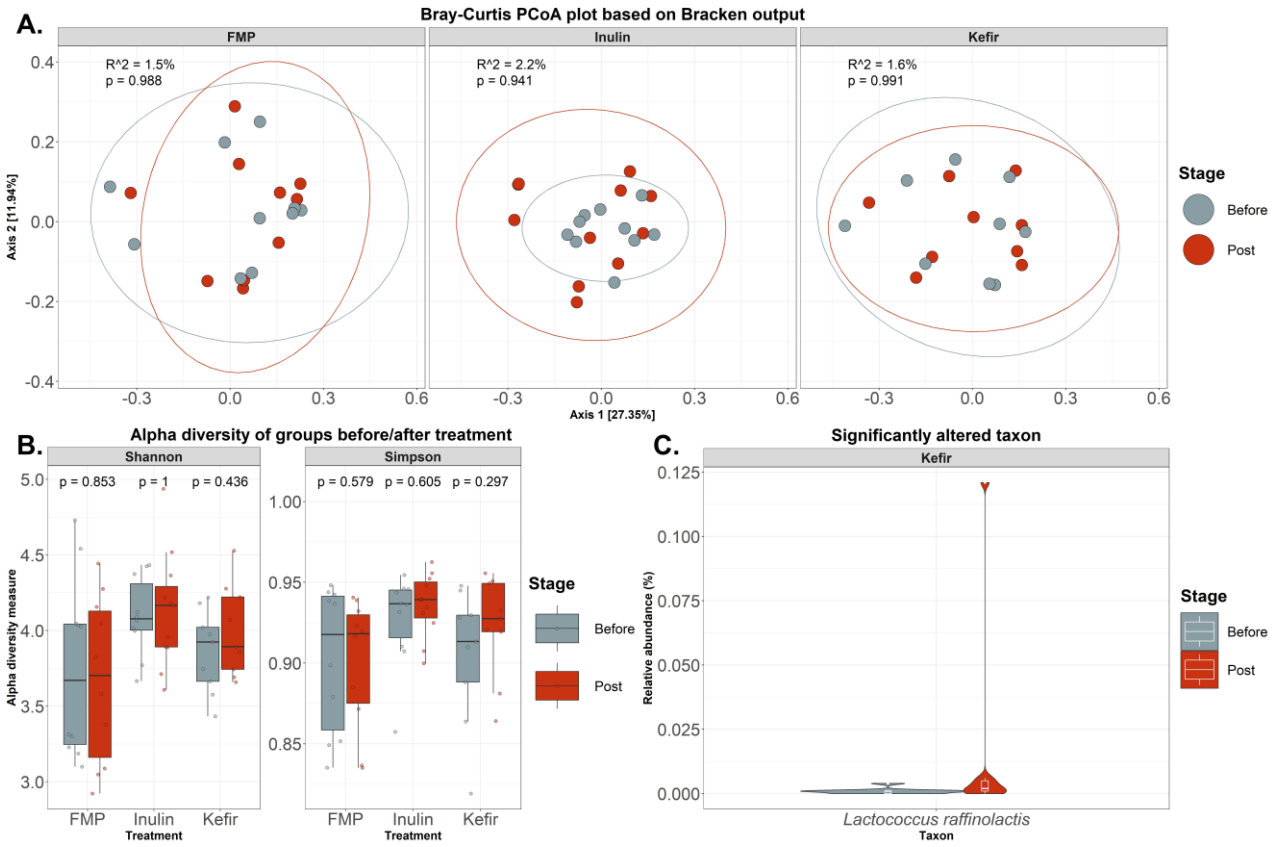


Figure 1: Changes in microbiome diversity following FMP, Inulin and Kefir consumption. **(A)**: Principal Coordinate analysis (PCoA) of beta diversity by intervention, measured by Bray-Curtis dissimilarity and calculated for species-level composition **(B.)**: Changes in alpha diversity post interventions, values calculated using the Shannon and Simpson diversity metrics. P- values represent the results of separate Wilcoxon signed-rank tests. **(C.)**: Changes in relative abundance of *Lactococcus raffinolactis* post kefir consumption. The bounds, whiskers and percentile of each box plot represented maximum, 75 percentile, median, 25 percentile and minimum from the top to the bottom respectively.

3.3.3 Comparative genomics reveals a number of host adaptability features within the pan-genome of *Lla. raffinolactis*

Reads mapped to the *Lla. raffinolactis* genome during the rapsir methodology were further inspected for annotated genes whereby the survival genes *dltA*, *clpE*, *copB_1* and *tuf* were identified from gut metagenomics data post kefir consumption. Given the recovery of survival genes from gut metagenomics data post kefir consumption, a comparative genomics approach was employed to uncover further, host adaptability features across the *Lla. raffinolactis* species (Figure 2b) [32]. The pan- genome of *Lla.*

raffinolactis contained 22 survival genes and suggested resistance to multiple stress conditions. The pan-genome included genes encoding stress-related proteins such as the molecular chaperones DnaK, GroES and GroEL, heat shock protein such as GrpE and oxidative stress resistant genes such as thiol peroxidase and thioredoxin reductase [33]. Furthermore, transport system associated genes/annotations such as *gbuA* were identified. *gbuA* encodes a component of the choline ABC transport system with applications in osmoregulation [34, 35] (Figure 2b). Three genes associated with adhesion-related functions were identified in the pan- genome of *Lla. raffinolactis* including cell-surface related proteins, such as elongation factor Tu (*tuf*), and sortase A-dependent enzymes (StrA) (Figure 2b), which is known to affect the mucin binding ability of microorganisms [36]. 5 Glycoside hydrolase family (GH) enzymes (GH36 - EC 3.2.1.22, GH42 - EC 3.2.1.23, GH109 - EC 3.2.1.49, GH29 - EC 3.2.1.51 and GH20 - EC 3.2.1.52) were recovered with potential applications in mucin cleavage. Additionally the GH29 family enzymes (alpha-L-fucosidase - EC 3.2.1.51, alpha-1,3/1,4-L-fucosidase - EC 3.2.1.111) suggested strains could remove Fuc moieties contained in intestinal mucin glycans [37-39]. The short-chain fatty acids (SCFAs), lactate and acetate and a number of genes linked to GABA production were widely found across *the Lla. raffinolactis* references. Specifically, the identified genes included the *puuD* gene encoding for a γ-Glu-GABA hydrolase, which plays a role in the Puu pathway by converting γ-Glu-GABA into GABA [40] and the genes *gadB* and *gadC* encoding for glutamate decarboxylases (GadB) and Glu/GABA antiporter, respectively. The glutamic acid decarboxylase (GAD) system provides a full mechanism by which *Lla. raffinolactis* can produce GABA [41]. All genomes encoded for β-galactosidase activity (EC 3.2.1.23), which could aid the digestion of lactose [42].

Within the genus *Lactococcus, Lla. lactis* strains are generally recognized as safe (GRAS) by the Food and Drug Administration (FDA) [43]. *Lla. lactis* together with *Lla. raffinolactis* are listed among the inventory of microbial food cultures (MFC) of fermented food products as species with demonstrated safety [44].Comparative genomic analysis of *Lla. raffinolactis* revealed no functional genes encoding enterotoxins, transferable antibiotic resistance, and an absence of D-lactate dehydrogenase (D-LDH) (EC 1.1.1.28) across strains which harboured only L-LDH (EC 1.1.1.27). D-LDH represents a safety concern as D-

lactate produced by commensals may induce D-lactate acidosis [45]. All *Lla. raffinolactis* genomes also contained the cold shock gene *cspL*, which may support survival of cold stress conditions imposed during cold temperature storage [46].

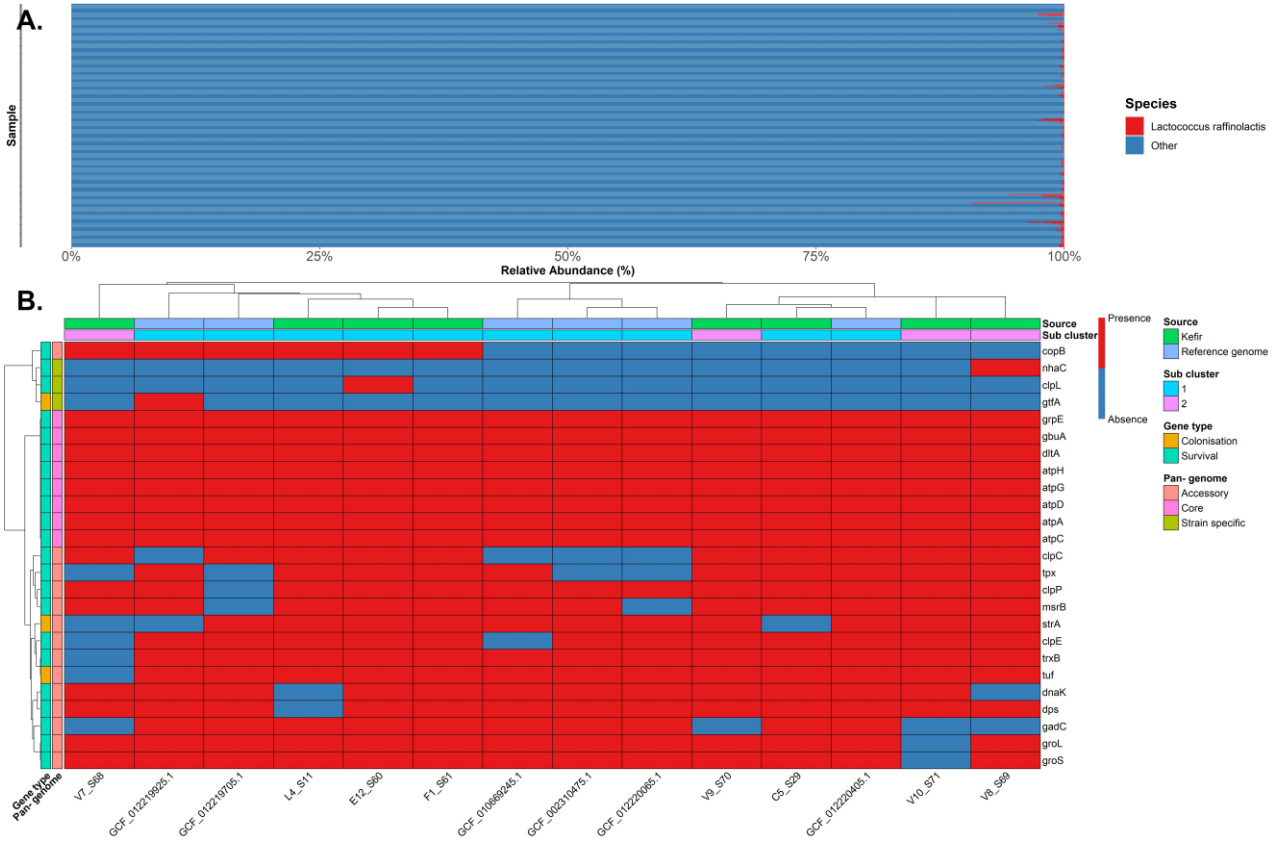


Figure 2: Prevalence of *Lla. raffinolactis* strains within the milk kefir microbiome and host adaptability genes within the pan- genome of *Lla. raffinolactis* **(A.)** Prevalence of *Lla. raffinolactis* across 256 kefir metagenomes generated by Walsh *et al* and assessed using Kraken2. **(B.)** Prevalence of genes with potential applications in host adaptability (Y -axis) across references of species *Lla. raffinolactis* (X -axis). The color of each cell of the heatmap represents the presence (red) or absence (blue) of host adaptability genes. Row side annotations include gene classifications separated into "Colonisation", and "Survival", summary descriptions of the pan- genomes of *Lla. raffinolactis*, separated into "Core", "Accessory" and "Strain specific" based on the prevalence of genes across reference genomes. Column side annotations represent the source data and sub cluster of the (Meta) genome.

3.3.4 Kefir consumption results in modest changes to urinal metabolites but not functional potential

HUMAnN3 was used to measure the abundance of pathways encoded within the faecal microbiome. MDS coupled with statistical analysis indicated that, overall, there was no perceivable shift in the functional potential of the microbiome following the consumption of inulin (PERMANOVA: p=0.95, R^2= 0.2%), kefir

(PERMANOVA: p=0.75, R^2=0.8%), or commercial FMP (PERMANOVA: p=0.573, R^2=2.1%) (Supplementary figure 1).

Three MCCV-PLS-DA models were built to investigate the impact of kefir, yogurt and inulin interventions on urinary metabolites. No significant differences were found following consumption of commercial inulin (R^2Y=0.98, Q^2Y=-0.28), FMP (R^2Y=0.96, Q^2Y=-0.8) or kefir (R^2Y=0.99, Q^2Y=0.57).

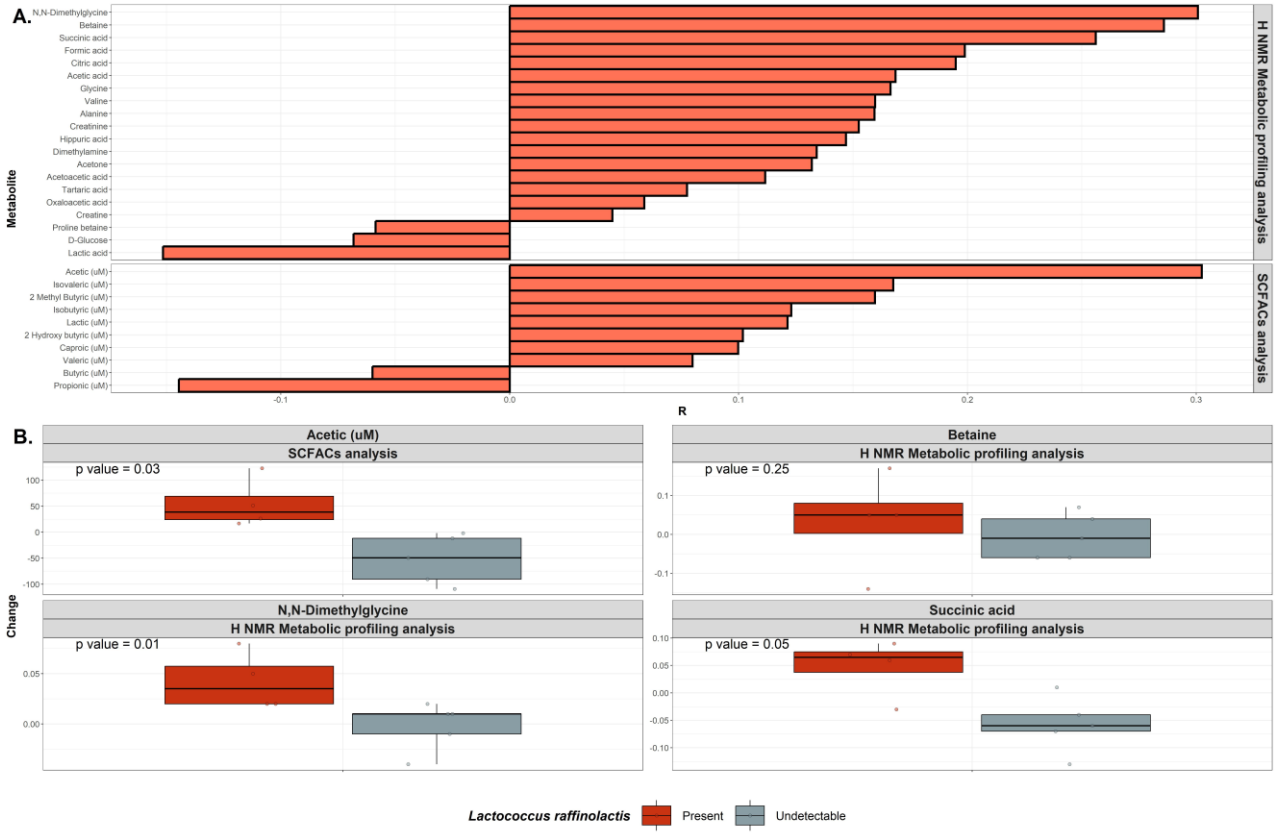


Figure 3: The presence of *Lla. raffinolactis* coincides with statistically significant shifts in the metabolome of the kefir cohort. **(A.)**: Correlations between the concentrations of metabolites and the relative abundance of *Lla. raffinolactis*. **(B.)**: Changes in the quantities of the positively correlated metabolites Acetic, Betaine, N,N-Dimethylglycine and Succinic acid, post kefir treatment, separated according to the undetectable (gray) or present (red) status of *Lla. raffinolactis.* P- values reflect the output of separate Wilcoxon signed ranked tests assessing the change in metabolites between participants of the kefir cohort. The bounds, whiskers and percentile of each box plot represented maximum, 75 percentile, median, 25 percentile and minimum from the top to the bottom respectively.

The relative abundance of *Lla. raffinolactis* was positively correlated with the concentrations of acetic, N,N-dimethylglycine, betaine and succinic acid (Figure 3a). However, none of these correlations were significant following p-value adjustment. While none of the metabolites were significantly altered when considering the kefir cohort in its entirety, we noted the metabolites acetic (Wilcoxon: p= 0.03), N,N-dimethylglycine (Wilcoxon: p= 0.01), succinic acid (Wilcoxon: p= 0.05) were statistically increased post

kefir consumption in participants with increased relative abundance of *Lla. raffinolactis*, compared with the remaining participants of the kefir cohort (Figure 3b). Furthermore, the examined *Lla. raffinolactis* genomes contained complete annotations for acetic acid production and supporting annotations for succinic acid production through the TCA cycle and N,N-dimethylglycine through the glycine betaine/proline transport system. Furthermore, *Lla. raffinolactis* genomes have supportive metabolic pathway to produce GABA, which can be metabolised to succinic acid, through the metabolic pathway GABA shunt, which was increased post kefir consumption.

Section 3.4: Discussion

In the present study, we combined whole metagenome shotgun sequencing and metabolomics to investigate the respective effects of a daily portion of an inulin, kefir, and a commercial probiotic drink containing *Lacticaseibacillus casei* on the human gut microbiota. All participants across the three treatment groups were deemed healthy and displayed individualised microbiomes (Figure 1a). As expected, our findings indicated that none of the treatments caused major changes in the overall human gut microbiota structure. It is important to note that these findings result from a relatively small number of participants. As the effect size is minimal, greater numbers of samples would be required to detect more subtle changes to the faecal microbiome arising as a result of consumption of these treatments. Despite limitations in sample number, whole metagenome shotgun sequencing revealed changes in the relative abundance of *Lla. raffinolactis* in 4 out of 10 participants after kefir consumption (Figure 1). The detection of *Lla. raffinolactis* is of particular interest given its “non - dominant”, status compared to other lactococci in dairy foods [47] and this study is to our knowledge the first to link the detection of *Lla. raffinolactis* in the gut microbiome to fermented food consumption. As the study did not include a wash-out period, it is not clear if the *Lla. raffinolactis* detected represent transient or mucosa-adherent strains, though the detection of the species in only four participants was notable. A number of publications have reported short-term resilience of gut microbiomes to dietary changes including the incorporation of fermented foods, which may explain why *Lla. raffinolactis* was absent in 5 of the 9 participants in the kefir

treatment group [27, 48]. Alternatively other studies have demonstrated that the persistence of potentially colonising microbes is dependent on the initial composition of the microbiota [49], in which the participants of this study displayed individuality (Figure 1).

Comparative analysis was employed to explore the genomic potential of strains of *Lla. raffinolactis* as a potentially colonising species. A total of 25 genes associated with host adaptability were identified to be contained in at least one of the *Lla. raffinolactis* strains. Specifically 22 survival genes and 3 colonisation genes were recovered across the strains (Figure 2). The recovery of such genes primarily provides an insight into the possible defence mechanisms employed by *Lla. raffinolactis* to counteract intracellular damage or to enhance the robustness of the cell to withstand the challenging conditions of the gastrointestinal tract and other environmental conditions. CAZymes and peptidases, were identified that further supported a colonisation effect and suggested a potential interaction between strains and the colonic mucus, for example the enzyme families GH109 and GH42 associated with mucin cleavage functionality [69-71] and the peptidase family Sortase A- C60A associated with mucin binding ability [36]. It should be noted that while surveying and reporting of genes associated with host adaptability highlights the potential of *Lla. raffinolactis* as a candidate coloniser, experimental verification is lacking.

Metabolic profiling analysis was used to assess the respective impacts of each treatment on the urine metabolome. No significant changes were observed following either inulin or commercial FMP consumption, but there were significant changes following kefir consumption in a number of these four participants. More specifically, statistically significant changes were observed at the metabolic level in response to consumption of kefir and the presence of *Lla. raffinolactis* with a statistically significant increase in the metabolites acetic acid, N, N - dimethylglycine and succinic acid (Figure 3b). However, correlation analysis of increased relative abundance of *Lla. raffinolactis* did not display statistically significant linear relationships with the concentrations of any of these metabolites following p-value adjustment. Thus, there was no evidence to suggest that *Lla. raffinolactis* had direct effects on the urinal metabolome.

Our findings suggest that kefir consumption in a healthy cohort can have a subtle impact on the urinal

metabolome and gut microbiome in a subset of participants. The principal change to the gut microbiome was the detection of *Lla. raffinolactis*. It should be noted that this, and the other interventions, could have more considerably impacted the microbiome of the upper GI tract, like the cecum or the ileum, which would not be detectable in stool samples to the same degree [50]. Regardless, it is unclear if the observed changes are sufficient to confer noticeable health benefits in this already healthy population, especially as we did not find any significant differences in the clinical parameters measured here. It is also important to note that this study used a small sample size of healthy participants, and *Lla. raffinolactis* was present in low relative abundance. Due to these limiting factors the sample size was insufficient to detect functional/metabolic features associated with its presence. Ultimately, this study highlights the need to employ deep metagenomics sequencing in larger cohorts to examine the potentially colonising microbial members of "live foods", which will likely account for low abundance in the gut [51], and/or metatranscriptomics to investigate their influence on microbial gene expression in the gut, or confirm their viability.

Section 3.5: References

1. Marchesi, J.R., et al., *The gut microbiota and host health: a new clinical frontier.* Gut, 2016. **65**(2): p. 330-339.
2. Sheth, R.U., et al., *Manipulating bacterial communities by in situ microbiome engineering.* Trends in Genetics, 2016. **32**(4): p. 189-200.
3. Hill, C., et al., *Expert consensus document: The International Scientific Association for Probiotics and Prebiotics consensus statement on the scope and appropriate use of the term probiotic.* Nature reviews Gastroenterology & hepatology, 2014. **11**(8): p. 506-514.
4. Bienenstock, J., et al., *New insights into probiotic mechanisms: a harvest from functional and metagenomic studies.* Gut microbes, 2013. **4**(2): p. 94-100.
5. Bermudez-Brito, M., et al., *Probiotic mechanisms of action.* Annals of Nutrition and Metabolism, 2012. **61**(2): p. 160-174.
6. Whorwell, P.J., et al., *Efficacy of an encapsulated probiotic Bifidobacterium infantis 35624 in women with irritable bowel syndrome.* The American journal of gastroenterology, 2006. **101**(7): p. 1581.
7. Dietrich, C.G., T. Kottmann, and M. Alavi, *Commercially available probiotic drinks containing Lactobacillus casei DN-114001 reduce antibiotic-associated diarrhea.* World Journal of Gastroenterology: WJG, 2014. **20**(42): p. 15837.
8. Gleeson, M., et al., *Daily probiotic's (Lactobacillus casei Shirota) reduction of infection incidence in athletes.* International journal of sport nutrition and exercise metabolism, 2011. **21**(1): p. 55-64.
9. Marco, M.L., et al., *Health benefits of fermented foods: microbiota and beyond.* Current opinion in biotechnology, 2017. **44**: p. 94-102.

10. Liu, J.-R., et al., *Hypocholesterolaemic effects of milk-kefir and soyamilk-kefir in cholesterol-fed hamsters.* British journal of nutrition, 2006. **95**(05): p. 939-946.
11. Lee, M.-Y., et al., *Anti-inflammatory and anti-allergic effects of kefir in a mouse asthma model.* Immunobiology, 2007. **212**(8): p. 647-654.
12. Rodrigues, K.L., et al., *Antimicrobial and healing activity of kefir and kefiran extract.* International journal of antimicrobial agents, 2005. **25**(5): p. 404-408.
13. Huang, Y., et al., *Characterization of Lactobacillus plantarum Lp27 isolated from Tibetan kefir grains: a potential probiotic bacterium with cholesterol-lowering effects.* Journal of dairy science, 2013. **96**(5): p. 2816-2825.
14. Carasi, P., et al., *Impact of kefir derived Lactobacillus kefiri on the mucosal immune response and gut microbiota.* Journal of immunology research, 2015. **2015**.
15. Zanirati, D.F., et al., *Selection of lactic acid bacteria from Brazilian kefir grains for potential use as starter or probiotic cultures.* Anaerobe, 2015. **32**: p. 70-76.
16. Bourrie, B.C.T., B.P. Willing, and P.D. Cotter, *The Microbiota and Health Promoting Characteristics of the Fermented Beverage Kefir.* Frontiers in Microbiology, 2016. **7**(647).
17. Gibson, G.R., et al., *The International Scientific Association for Probiotics and Prebiotics (ISAPP) consensus statement on the definition and scope of prebiotics.* 2017.
18. Al-Sheraji, S.H., et al., *Prebiotics as functional foods: A review.* Journal of Functional Foods, 2013. **5**(4): p. 1542-1553.
19. Slavin, J., *Fiber and prebiotics: mechanisms and health benefits.* Nutrients, 2013. **5**(4): p. 1417-1435.
20. Le Bastard, Q., et al., *The effects of inulin on gut microbial composition: a systematic review of evidence from human studies.* Eur J Clin Microbiol Infect Dis, 2020. **39**(3): p. 403-413.
21. Kellow, N.J., M.T. Coughlan, and C.M. Reid, *Metabolic benefits of dietary prebiotics in human subjects: a systematic review of randomised controlled trials.* British Journal of Nutrition, 2014. **111**(7): p. 1147-1161.
22. Beserra, B.T., et al., *A systematic review and meta-analysis of the prebiotics and synbiotics effects on glycaemia, insulin concentrations and lipid parameters in adult patients with overweight or obesity.* Clinical Nutrition, 2015. **34**(5): p. 845-858.
23. Lohner, S., et al., *Prebiotics in healthy infants and children for prevention of acute infectious diseases: a systematic review and meta-analysis.* Nutrition reviews, 2014. **72**(8): p. 523-531.
24. Sánchez, B., et al., *Probiotics, gut microbiota, and their influence on host health and disease.* Molecular nutrition & food research, 2017. **61**(1).
25. Hutkins, R.W., et al., *Prebiotics: why definitions matter.* Current opinion in biotechnology, 2016. **37**: p. 1-7.
26. Washburn, R.L., D. Sandberg, and M.A. Gazdik Stofer, *Supplementation of a single species probiotic does not affect diversity and composition of the healthy adult gastrointestinal microbiome.* Human Nutrition & Metabolism, 2022. **28**: p. 200148.
27. Wastyk, H.C., et al., *Gut-microbiota-targeted diets modulate human immune status.* Cell, 2021. **184**(16): p. 4137-4153.e14.
28. Eckburg, P.B., et al., *Diversity of the Human Intestinal Microbial Flora.* Science, 2005. **308**(5728): p. 1635-1638.
29. Cotter, P., et al., *Analysis of the Milk Kefir Pan-Metagenome Reveals Four Community Types, Core Species and Associated Metabolic Pathways*. 2023.
30. Vandeputte, D., et al., *Prebiotic inulin-type fructans induce specific changes in the human gut microbiota.* Gut, 2017.
31. Pust, M.-M. and B. Tümmler, *Identification of core and rare species in metagenome samples based on shotgun metagenomic sequencing, Fourier transforms and spectral comparisons.* ISME Communications, 2021. **1**(1): p. 2.
32. Lebeer, S., J. Vanderleyden, and S.C.J.D. Keersmaecker, *Genes and Molecules of Lactobacilli Supporting Probiotic Action.* Microbiology and Molecular Biology Reviews, 2008. **72**(4): p. 728-764.

33. Zhang, W., et al., *Complete Genome Sequencing and Comparative Genome Characterization of Lactobacillus johnsonii ZLJ010, a Potential Probiotic With Health-Promoting Properties.* Front Genet, 2019. **10**: p. 812.
34. Annamalai, T. and K. Venkitanarayanan, *Role of proP and proU in betaine uptake by Yersinia enterocolitica under cold and osmotic stress conditions.* Appl Environ Microbiol, 2009. **75**(6): p. 1471-7.
35. Bouvier, J., et al., *Characterization of OpuA, a glycine-betaine uptake system of Lactococcus lactis.* Journal of molecular microbiology and biotechnology, 2000. **2**(2): p. 199-205.
36. Banla, L.I., et al., *Sortase-Dependent Proteins Promote Gastrointestinal Colonization by Enterococci.* Infect Immun, 2019. **87**(5).
37. Katayama, T., et al., *Molecular cloning and characterization of Bifidobacterium bifidum 1,2-alpha-L-fucosidase (AfcA), a novel inverting glycosidase (glycoside hydrolase family 95).* J Bacteriol, 2004. **186**(15): p. 4885-93.
38. Nagae, M., et al., *Structural basis of the catalytic reaction mechanism of novel 1,2-alpha-L-fucosidase from Bifidobacterium bifidum.* J Biol Chem, 2007. **282**(25): p. 18497-18509.
39. Ashida, H., et al., *Two distinct alpha-L-fucosidases from Bifidobacterium bifidum are essential for the utilization of fucosylated milk oligosaccharides and glycoconjugates.* Glycobiology, 2009. **19**(9): p. 1010-7.
40. Diez-Gutiérrez, L., et al., *Gamma-aminobutyric acid and probiotics: Multiple health benefits and their future in the global functional food and nutraceuticals market.* Journal of Functional Foods, 2020. **64**: p. 103669.
41. Wu, Q., et al., *Common Distribution of gad Operon in Lactobacillus brevis and its GadA Contributes to Efficient GABA Synthesis toward Cytosolic Near-Neutral pH.* Frontiers in Microbiology, 2017. **8**.
42. Papadimitriou, K., et al., *Discovering probiotic microorganisms: in vitro, in vivo, genetic and omics approaches.* Front Microbiol, 2015. **6**: p. 58.
43. Wessels, S., et al., *The lactic acid bacteria, the food chain, and their regulation.* Trends in Food Science & Technology, 2004. **15**(10): p. 498-505.
44. Bourdichon, F., et al., *Food fermentations: microorganisms with technological beneficial use.* Int J Food Microbiol, 2012. **154**(3): p. 87-97.
45. Kowlgi, N.G. and L. Chhabra, *D-lactic acidosis: an underrecognized complication of short bowel syndrome.* Gastroenterol Res Pract, 2015. **2015**: p. 476215.
46. Mbye, M., et al., *Updates on understanding of probiotic lactic acid bacteria responses to environmental stresses and highlights on proteomic analyses.* Comprehensive Reviews in Food Science and Food Safety, 2020. **19**(3): p. 1110-1124.
47. Meslier, V., V. Loux, and P. Renault, *Genome sequence of Lactococcus raffinolactis strain 4877, isolated from natural dairy starter culture.* J Bacteriol, 2012. **194**(22): p. 6364.
48. Fragiadakis, G.K., et al., *Long-term dietary intervention reveals resilience of the gut microbiota despite changes in diet and weight.* Am J Clin Nutr, 2020. **111**(6): p. 1127-1136.
49. Zhang, C., et al., *Ecological robustness of the gut microbiota in response to ingestion of transient food-borne microbes.* Isme j, 2016. **10**(9): p. 2235-45.
50. Derrien, M. and J.E. van Hylckama Vlieg, *Fate, activity, and impact of ingested bacteria within the human gut microbiota.* Trends in microbiology, 2015. **23**(6): p. 354-366.
51. Pasolli, E., et al., *Large-scale genome-wide analysis links lactic acid bacteria from food with the gut microbiome.* Nature Communications, 2020. **11**(1): p. 2610.
52. Bonnema, A.L., et al., *Gastrointestinal tolerance of chicory inulin products.* J Am Diet Assoc, 2010. **110**(6): p. 865-8.
53. Ripoll, C., et al., *Gastrointestinal tolerance to an inulin-rich soluble roasted chicory extract after consumption in healthy subjects.* Nutrition, 2010. **26**(7-8): p. 799-803.
54. Spitzer, R.L., et al., *A brief measure for assessing generalized anxiety disorder: the GAD-7.* Arch Intern Med, 2006. **166**(10): p. 1092-7.
55. Rabin, R. and F. de Charro, *EQ-5D: a measure of health status from the EuroQol Group.* Ann Med, 2001. **33**(5): p. 337-43.

56. Lewis, S.J. and K.W. Heaton, *Stool form scale as a useful guide to intestinal transit time.* Scand J Gastroenterol, 1997. **32**(9): p. 920-4.
57. Costabile, A., et al., *Wholegrain fermentation affects gut microbiota composition, phenolic acid metabolism and pancreatic beta cell function in a rodent model of type 2 diabetes.* Front Microbiol, 2022. **13**: p. 1004679.
58. Li, H., et al., *The sequence alignment/map format and SAMtools.* Bioinformatics, 2009. **25**(16): p. 2078-2079.
59. Wood, D.E., J. Lu, and B. Langmead, *Improved metagenomic analysis with Kraken 2.* Genome Biology, 2019. **20**(1): p. 257.
60. Lu, J., et al., *Bracken: estimating species abundance in metagenomics data.* PeerJ Computer Science, 2017. **3**: p. e104.
61. Page, A.J., et al., *Roary: rapid large-scale prokaryote pan genome analysis.* Bioinformatics, 2015. **31**(22): p. 3691-3693.
62. O'Leary, N.A., et al., *Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation.* Nucleic Acids Res, 2016. **44**(D1): p. D733-45.
63. Olm, M.R., et al., *dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication.* Isme j, 2017. **11**(12): p. 2864-2868.
64. Shaffer, M., et al., *DRAM for distilling microbial metabolism to automate the curation of microbiome function.* Nucleic Acids Res, 2020. **48**(16): p. 8883-8900.
65. Seemann, T., *Prokka: rapid prokaryotic genome annotation.* Bioinformatics, 2014. **30**(14): p. 2068-2069.
66. Mills, S., et al., *Enhancing the stress responses of probiotics for a lifestyle from gut to product and back again.* Microbial Cell Factories, 2011. **10**(1): p. S19.
67. Leech, J., et al., *Fermented-Food Metagenomics Reveals Substrate-Associated Differences in Taxonomy and Health-Associated and Antibiotic Resistance Determinants.* mSystems, 2020. **5**(6).
68. Kim, E., et al., *Complete Genome Sequencing and Comparative Genomics of Three Potential Probiotic Strains, Lacticaseibacillus casei FBL6, Lacticaseibacillus chiayiensis FBL7, and Lacticaseibacillus zeae FBL8.* Frontiers in Microbiology, 2022. **12**.
69. Dona, A.C., et al., *Precision high-throughput proton NMR spectroscopy of human urine, serum, and plasma for large-scale metabolic phenotyping.* Anal Chem, 2014. **86**(19): p. 9887-94.
70. Veselkov, K.A., et al., *Optimized preprocessing of ultra-performance liquid chromatography/mass spectrometry urinary metabolic profiles for improved information recovery.* Anal Chem, 2011. **83**(15): p. 5864-72.
71. Garcia-Perez, I., et al., *Objective assessment of dietary patterns by use of metabolic phenotyping: a randomised, controlled, crossover trial.* Lancet Diabetes Endocrinol, 2017. **5**(3): p. 184-195.
72. Team, R.C., *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2013*. 2014, ISBN 3-900051-07-0.
73. Oksanen, J., et al., *The vegan package.* Community ecology package, 2007. **10**: p. 631-637.
74. Wickham, H., *ggplot2: elegant graphics for data analysis*. 2016: springer.

Section 3.6: Supplementary information

3.6.1 Supplementary figures

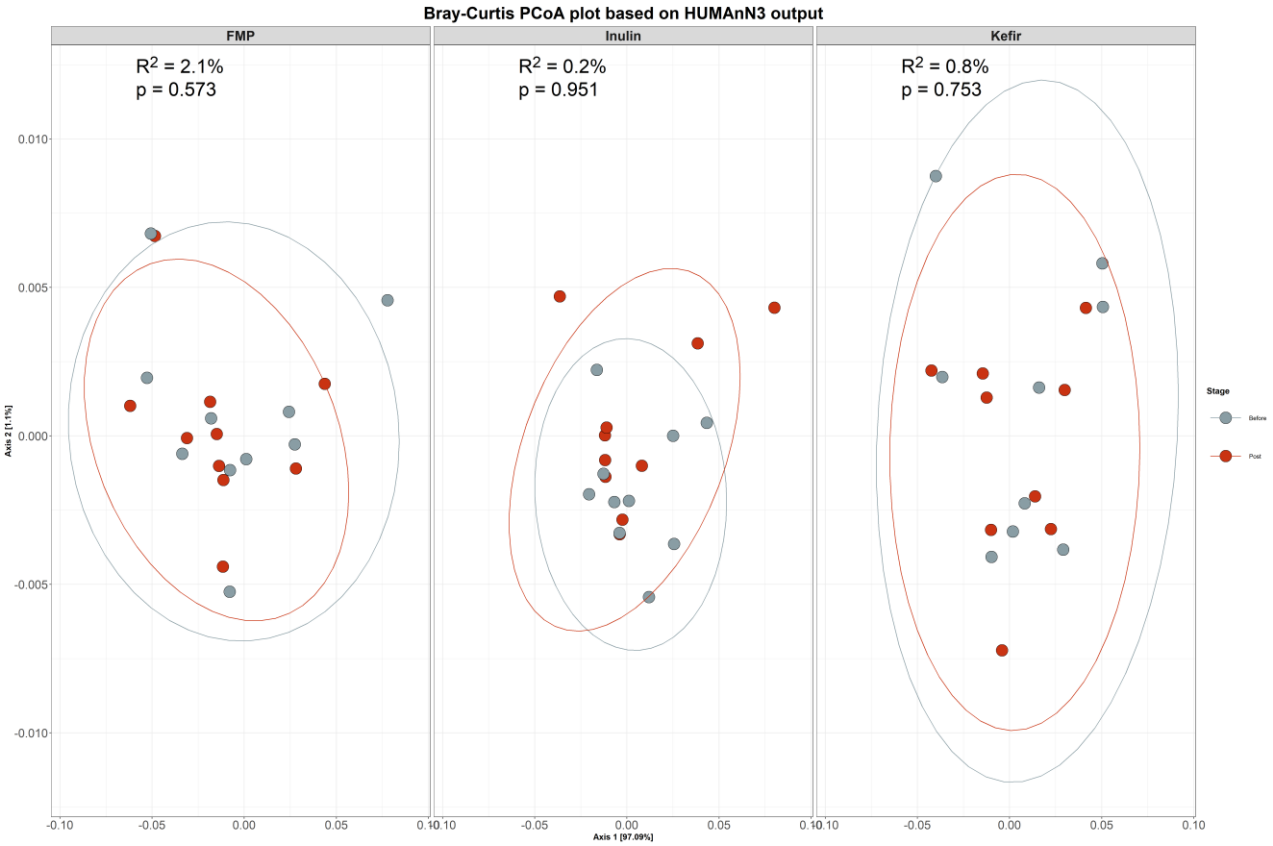


Figure S1: Changes in metabolic pathway diversity in the gut microbiome following FMP, Inulin and Kefir consumption. Principal Coordinate analysis (PCoA) of beta diversity by intervention measured by Bray-Curtis dissimilarity and calculated from the HUMANn3 metabolic pathway abundance profiles.

3.6.2 Supplementary table

Table S1: Changes in clinical parameters of participants following FMP, Inulin and Kefir consumption.

Observations	BMI(printout)	21.8	31	24
	Height (cm)	171	168	168
	Heart Rate	62	65	63
	Dia BP	68	94	71
	Sys BP	125	144	110
	Weight(kg)	63.8	87.6	67.8
	Fat %	12	41.8	30.9
	Bristol Stool	4	4	3
	Date	Tuesday 17 May 2016	Tuesday 17 May 2016	Tuesday 17 May 2016
	Sex	M	F	F
	Age	44	49	53
Patients metadata	Patients number	1	4	8
	Intervention group	Inulin Patients	Inulin Patients	Inulin Patients

23.5	29.9	26.3	26.9	23.6	31.3	25.5
174	167	161	171	167	163	173
66	96	62	61	56	84	86
78	93	69	71	71	92	66
110	124	111	116	115	138	113
71.2	83.3	68.1	78.8	65.9	83.2	76.2
23.9	44.4	31.5	38.8	28.7	45.2	34.5
4	5	4	3	Not answered	4	3
Tuesday 17 May 2016	Tuesday 17 May 2016	Tuesday 17 May 2016	Tuesday 17 May 2016	Tuesday 17 May 2016	Tuesday 17 May 2016	Tuesday 17 May 2016
F	F	F	F	F	F	F
52	60	46	56	45	57	37
10	15	18	20	22	25	30
Inulin Patients	Inulin Patients	Inulin Patients	Inulin Patients	Inulin Patients	Inulin Patients	Inulin Patients

23.2	25.3	28.3	32.5	21.5	25.7	24
184	163	164	179	179	191	166
56	67	71	72	86	57	67
58	61	83	94	102	69	71
113	109	142	146	152	122	107
78.7	67.2	76	104.1	68.9	93.8	66.2
15	30.1	39.3	46.9	16.2	19.7	27
4	3	4	4	4	4	3
Tuesday 17 May 2016	Tuesday 17 May 2016	Wednesday 18 May 2016	Tuesday 17 May 2016	Tuesday 17 May 2016	Tuesday 17 May 2016	Tuesday 17 May 2016
M	F	F	F	M	M	F
20	35	51	42	57	40	50
3	6	7	12	13	16	19
Kefir Patients	Kefir Patients	Kefir Patients	Kefir Patients	Kefir Patients	Kefir Patients	Kefir Patients

22.5	23	16.7	24.1	23.3	30.8	32
191	160	178	170	177	191	186
60	80	82	70	66	78	75
77	67	74	76	83	89	89
131	114	126	120	138	130	143
82.2	58.8	52.8	69.6	73.1	112.3	110.8
10.9	24.1	13.6	16.9	15.2	32.6	34
3	Not answered	3	3	4	4	4
Tuesday 17 May 2016	Tuesday 17 May 2016	Tuesday 17 May 2016	Wednesday 18 May 2016	Tuesday 17 May 2016	Tuesday 17 May 2016	Tuesday 17 May 2016
M	F	F	M	M	M	M
31	25	31	28	40	42	51
23	27	28	2	5	9	11
Kefir Patients	Kefir Patients	Kefir Patients	Yakult patients	Yakult patients	Yakult patients	Yakult patients

24.1	23.4	22.6	34.7	21.6	25.2	
168	155	160	156	153	163	
79	74	75	69	58	63	
88	70	65	98	74	73	
139	108	125	135	119	130	
67.9	56.2	57.9	84.5	50.5	66.9	
34.7	25.4	28.2	39.9	21.5	31.7	
3	4	4	2	4	5	EQ-5D-5L
Tuesday 17 May 2016	Tuesday 17 May 2016	Tuesday 17 May 2016	Tuesday 17 May 2016	Tuesday 17 May 2016	Tuesday 17 May 2016	
F	F	F	F	F	F	
57	29	60	51	58	36	
14	17	21	24	26	29	
Yakult patients	Yakult patients	Yakult patients	Yakult patients	Yakult patients	Yakult patients	

Health(0-100)	80	90	80	70	80	85
Anx/Dep(1-5)	1	1	1	1	1	1
Pain(1-5)	1	2	1	2	1	1
Usual(1-5)	1	1	1	1	1	1
Self-Care(1-5)	1	1	1	1	1	1
Mobility(1-5)	1	1	1	1	1	1

90	90	80	90	90	75	85
1	1	2	1	1	1	1
1	1	2	1	1	1	1
1	1	1	1	1	1	1
1	1	1	1	1	1	1
1	1	1	1	1	1	1

90	85	85	90	85	100	95
1	1	1	1	1	1	1
1	1	1	1	1	1	1
1	1	1	1	1	1	1
1	1	1	1	1	1	1
1	1	1	1	1	1	1

100	95	60	90	100	81	91
1	1	1	1	1	1	2
1	1	1	1	1	1	1
1	1	1	1	1	1	1
1	1	1	1	1	1	1
1	1	1	1	1	1	1

				Nausea(1-7)	2	1
				Empty(1-7)	1	1
				Bloated(1-7)	1	2
				Abdo rumbling(1-7)	2	2
75	90	90		**Regurg(1-7)**	1	1
1	1	2	**Abdo symptoms**	**Heartburn(1-7)**	1	1
3	1	1		**Doesn't decline(1-7)**	1	1
1	1	1		**Fasting(1-7)**	1	1
1	1	1		**After eat(1-7)**	1	1
1	1	1		Continuous(1-7)	1	1

2	1	1	1	1	1	1
1	1	1	2	2	2	1
1	1	1	2	1	1	1
2	1	1	2	2	2	1
1	1	1	1	1	1	1
2	1	2	1	1	1	3
1	1	1	Not answered	1	1	1
1	1	1	1	1	1	1
1	1	1	2	1	1	1
1	1	1	1	1	1	1

2	1	1	1	1	1	1
1	1	1	2	1	2	1
1	1	4	2	1	1	1
2	1	2	2	1	2	3
1	1	1	1	1	1	1
2	1	1	1	1	1	1
1	1	3	1	Not answered	1	Not answered
1	1	1	1	Not answered	1	1
1	1	3	1	Not answered	1	1
1	1	1	1	1	1	1

1	1	1	1	1	1	1
1	1	1	2	1	2	1
1	1	1	1	1	1	1
1	2	1	2	4	2	1
1	1	1	1	1	1	1
1	1	1	1	1	1	1
1	1	1	1	1	1	1
1	1	1	1	1	1	1
1	1	1	1	1	1	1
1	1	1	1	1	1	1

1	1	1	1	1	1	1
1	1	2	2	3	1	Not answered
1	1	1	1	3	1	2
2	1	2	2	3	1	2
1	1	1	1	2	1	1
2	1	1	1	2	1	1
1	Not answered	1	1	1	Not answered	1
1	1	1	1	1	Not answered	1
1	1	1	1	1	1	1
1	1	2	1	1	1	1

	Bloody stools(1-7)	1	Not answered	1	1	1
	Black stools(1-7)	1	Not answered	1	1	1
	Diff swall solid(1-7)	1	1	1	1	1
	Diff swall liq(1-7)	1	1	1	1	1
	Vomit blood(1-7)	1	1	1	1	1
Abdo symptoms	**Flatulence(1-7)**	3	2	2	2	1
	Belching(1-7)	2	2	1	1	2
	Fullness(1-7)	2	1	2	1	1
	Loss Appetite(1-7)	2	1	1	1	1
	Vomiting(1-7)	1	1	1	1	1

1	1	1	1	1	1	1
1	1	1	1	1	1	1
1	1	1	1	1	1	1
1	1	1	1	1	1	1
1	1	1	1	1	1	1
2	2	2	2	3	2	4
1	1	2	2	3	1	1
2	2	2	1	4	3	4
1	1	1	1	1	1	1
1	1	1	Not answered	1	1	1

1	Not answered	1	1	Not answered	1	1
1	1	1	1	Not answered	3	1
1	Not answered	1	1	1	1	Not answered
1	1	1	1	1	1	Not answered
1	1	1	1	1	1	1
2	1	2	3	2	2	1
1	1	2	1	2	1	1
2	Not answered	1	1	2	2	1
1	1	1	1	1	1	1
1	1	1	1	1	1	1

1	1	1	1	1	1	1
1	1	2	1	1	1	1
1	1	1	1	1	1	1
1	1	1	1	1	1	1
1	1	1	1	1	1	1
2	4	2	1	2	2	3
2	1	2	1	2	2	1
1	2	3	1	2	2	2
1	1	1	1	1	1	1
1	1	1	1	1	1	1

1	1	Not answered	1			
1	1	Not answered	1		**Pain (0-12.3cm)**	0.65
1	1	Not answered	1		**Incomplete(1-7)**	1
1	1	1	1		**Urgency(1-7)**	1
1	1	1	1		**freq with pain(1-7)**	1
2	4	1	2	**Abdo symptoms**	**constipation(1-7)**	1
1	2	1	3		**alt solid/loose(1-7)**	2
2	1	2	3		**Diarrhoea(1-7)**	2
1	1	1	1		**Hard stools(1-7)**	1
1	1	1	1		with Mucus(1-7)	1

2.75	0.75	0.65	0.00	0.80	0.30	0.35
Not answered	1	1	1	1	1	1
Not answered	1	3	1	1	1	2
Not answered	1	1	1	1	1	1
Not answered	2	1	1	1	1	1
1	1	1	2	1	1	1
Not answered	1	1	1	1	1	1
Not answered	2	1	1	1	1	1
Not answered	1	1	1	1	1	1

0.30	1.00	0.00	3.20	0.35	0.00	0.30
2	1	1	1	1	Not answered	2
2	1	1	1	1	Not answered	1
1	1	1	2	1	Not answered	1
1	1	1	2	1	Not answered	1
1	4	2	2	1	Not answered	1
1	1	1	1	1	Not answered	1
1	1	2	3	2	Not answered	1
1	1	1	1	1	Not answered	1

0.00	0.00	0.50	0.00	0.00	0.00	0.00
1	Not answered	2	1	1	1	1
1	Not answered	1	1	1	1	1
1	Not answered	1	1	1	1	1
1	Not answered	1	1	1	1	1
1	1	1	1	1	1	2
1	Not answered	1	1	1	1	2
1	Not answered	1	1	1	1	3
1	Not answered	1	1	1	1	2

0.00	0.30	0.00	4.65	0.90	3.80	0.00
1	1	1	1	1	3	Not answered
1	1	1	1	1	1	Not answered
1	1	1	1	1	3	Not answered
1	1	1	1	1	4	Not answered
1	1	1	2	2	2	Not answered
1	2	1	1	2	1	Not answered
1	1	1	2	1	4	Not answered
1	1	1	1	1	1	Not answered

0.40						
1						
3	GAD-7	**Feeling afraid...(0-3)**	0	0	0	0
1		**Becoming easily..(0-3)**	1	1	1	0
1		**Being so...(0-3)**	0	0	0	0
1		**Trouble...(0-3)**	1	0	0	0
1		**Worrying...(0-3)**	1	0	0	0
1		**Not being...(0-3)**	0	0	0	0
1		Feel nervous...(0-3)	1	0	0	0

0	0	0	0	0	0	0
1	1	0	0	0	1	0
0	0	0	0	0	0	0
0	1	0	0	1	0	0
0	0	0	0	1	0	0
0	0	0	0	0	0	0
0	0	0	0	1	0	0

0	0	1	0	0	0	1
0	1	0	1	0	0	0
0	0	0	0	0	0	0
1	1	0	0	0	0	1
1	1	0	1	0	0	1
0	0	0	0	0	0	0
1	1	0	0	0	0	0

0	0	0	0	0	0	0
0	2	0	0	0	0	1
0	0	0	0	0	0	0
0	1	0	0	0	0	1
0	1	0	0	0	0	Not answered
0	0	0	0	0	0	0
0	1	0	0	0	0	0

					Observations	**BMI(printout)**
						Height (cm)
1	1	1	0	1		**Heart Rate**
1	1	1	0	2		**Dia BP**
0	0	0	0	1		**Sys BP**
0	1	1	0	3		**Weight(kg)**
0	1	0	0	2		**Fat %**
0	0	0	0	2		**Bristol Stool**
0	0	1	0	2		Date

21.3	30.2	23.5	23.3	30.0	26.6	27.0
171	168	168	174	167	161	171
57	62	76	61	91	67	57
66	93	73	76	96	69	85
122	139	116	110	143	113	137
62.3	85.3	66.4	70.6	83.6	68.9	79.0
12.1	42.1	31.6	27.4	43.4	31.1	40.3
4	4	3	4	5	4	3
########## #####	########## #####	########## #####	########## #####	########## #####	########## #####	########## #####

23.6	31.3	Not answered	23.3	24.8	28.5	32.6
167	163	173	184	163	164	179
61	99	75	61	62	84	61
64	78	63	62	56	78	87
103	130	109	128	100	125	140
65.8	83.1	78.0	78.8	66.0	76.7	104.4
29.3	45.4		15.1	27.0	39.0	45.3
4	4	4	5	3	4	4
########## #####	########## #####	########## #####	########## #####	########## #####	########## #####	########## #####

21.9	25.8	23.8	22.7		16.9	23.8
179	191	166	191		178	170
78	54	70	66		80	63
89	71	65	71		83	75
158	129	113	122		123	131
70.1	94.3	65.5	82.8		53.6	68.9
16.6	21.9	29.1	12.6		14.6	16.2
4	4	3	3		3	4
########## #####	########## #####	########## #####	########## #####		########## #####	########## #####

23.5	31.7	32.1	24.3	23.5	22.7	34.7
177	191	186	168	155	160	156
58	70	64	81	94	74	70
81	84	85	84	60	69	96
131	140	157	132	109	115	149
73.5	115.8	110.9	68.7	56.4	58.1	84.5
15.2	32.0	33.3	35.8	26.2	27.6	40.4
4	4	4	4	2 to 3	5	1
########## #####	########## #####	########## #####	########## #####	########## #####	########## #####	########## #####

43.0	25.4					
153	163					
64	60					
72	69	EQ-5D-5L	**Health(0-100)**	75	90	70
113	124		**Anx/Dep(1-5)**	1	1	2
50.3	67.4		**Pain(1-5)**	1	2	1
24.0	41.6		**Usual(1-5)**	1	1	1
4	5		**Self-Care(1-5)**	1	1	1
##########	##########		Mobility(1-5)	1	1	1

75	80	80	95	85	85	80
1	1	1	1	1	1	1
1	1	1	1	2	1	1
1	1	1	1	1	1	1
1	1	1	1	1	1	1
1	1	1	1	2	1	1

50	80	95	90	95	79	100
Not answered	1	1	1	1	1	1
Not answered	1	1	1	1	1	1
Not answered	1	1	1	1	1	1
Not answered	1	1	1	1	1	1
Not answered	1	1	1	1	1	1

79		95	98	95	80	90
1		1	1	1	1	1
2		1	1	2	1	2
1		1	1	1	1	1
1		1	1	1	1	1
1		1	1	1	1	1

						Nausea(1-7)
						Empty(1-7)
						Bloated(1-7)
						Abdo rumbling(1-7)
100	90	98	80	90	90	**Regurg(1-7)**
1	1	1	1	1	1	**Heartburn(1-7)**
1	1	1	2	1	1	**Doesn't decline(1-7)**
1	1	1	1	1	1	**Fasting(1-7)**
1	1	1	1	1	1	**After eat(1-7)**
1	1	1	1	1	1	Continuous(1-7)

1	1	2	1	1	3	1
1	1	1	2	1	2	2
1	4	2	1	1	4	3
1	2	1	1	2	2	3
1	1	1	1	1	1	1
1	1	1	1	2	1	1
1	Not answered	1	1	Not answered	1	1
1	1	1	1	1	3	2
Not answered	2	1	1	1	2	2
1	1	1	1	1	1	2

1	2	3	1	1	1	1
1	1	1	4	1	1	1
1	2	1	1	4	2	2
2	1	1	4	1	2	1
1	1	1	1	1	1	1
1	1	4	3	1	1	1
1	Not answered	1	1	1	Not answered	1
1	1	1	1	2	1	1
1	2	3	2	2	1	1
1	1	1	1	1	1	1

1	1	1	2		1	1
1	1	1	1		1	1
1	1	1	2		1	1
2	1	1	1		2	3
1	1	1	1		2	1
1	1	1	2		2	1
1	1	1	1		1	1
1	1	1	2		2	1
1	1	1	2		1	1
1	1	1	1		1	1

1	1	1	2	2	1	1
2	1	1	1	3	1	2
1	1	1	2	1	1	3
2	1	2	2	2	1	2
1	1	1	1	2	1	1
1	1	2	1	2	1	2
1	1	1	1	3	1	1
1	1	1	1	Not answered	1	1
1	1	1	1	Not answered	1	1
1	1	1	1	3	1	1

1	2					
1	1					
1	5					
1	4	**Diff swall liq(1-7)**	1	1	1	1
1	1	**Vomit blood(1-7)**	1	1	1	1
1	2	**Flatulence(1-7)**	3	5	2	1
Not answered	4	**Belching(1-7)**	1	1	1	1
1	1	**Fullness(1-7)**	Not answered	1	2	1
1	4	**Loss Appetite(1-7)**	1	1	1	1
1	1	Vomiting(1-7)	1	1	1	1

1	1	1	1	1	1	1
1	1	1	1	1	1	1
2	4	4	2	1	2	3
1	3	1	1	1	2	3
1	1	4	1	2	3	1
1	1	1	1	1	2	1
1	1	1	1	1	1	1

1	1	1	1	1	1	1
1	1	1	1	1	1	1
2	2	1	3	1	1	2
1	1	1	2	1	1	2
3	1	2	2	1	1	2
1	1	1	1	1	1	1
1	1	1	1	1	1	1

	1	1	1	1	1	1
	1	1	1	1	1	1
	2	3	2	1	2	4
	2	1	2	1	2	4
	2	2	3	1	1	2
	1	1	1	2	1	1
	1	1	1	1	1	1

					Pain (0-12.3cm)	0.35
					Incomplete(1-7)	1
					Urgency(1-7)	1
					freq with pain(1-7)	1
					constipation(1-7)	1
1	1	1	1	1	**alt solid/loose(1-7)**	1
1	1	1	1	1	**Diarrhoea(1-7)**	1
3	2	3	1	6	**Hard stools(1-7)**	1
1	1	3	1	2	**with Mucus(1-7)**	1
1	2	3	1	5	**Bloody stools(1-7)**	1
1	1	1	1	1	**Black stools(1-7)**	1
1	1	1	1	2	Diff swall solid(1-7)	1

8.00	0.35	0.00	0.30	5.75	1.80	0.80
1	1	1	1	1	3	1
2	1	2	1	2	1	2
1	1	1	1	1	1	1
1	2	1	1	1	1	1
3	3	1	1	1	1	1
2	1	1	1	1	1	2
1	3	1	1	1	1	1
1	1	1	1	1	1	1
1	1	1	1	1	1	1
1	1	1	1	1	1	1
Not answered	1	1	1	1	1	1

1.20	2.80	6.00	8.90	0.00	0.00	0.30
1	Not answered	Not answered	1	1	1	2
1	Not answered	1	1	1	1	1
1	Not answered	Not answered	Not answered	1	1	1
1	Not answered	Not answered	3	1	1	1
1	1	Not answered	1	1	1	1
1	Not answered	Not answered	1	1	1	1
1	Not answered	Not answered	3	1	1	1
1	Not answered	Not answered	1	1	1	1
1	Not answered	Not answered	1	1	1	1
1	Not answered	Not answered	1	1	1	1
Not answered	1	1	1	1	1	1

0.00	0.20	3.00		0.45	0.00	0.20
1	1	2		1	1	1
1	1	1		1	1	1
1	1	1		1	1	1
1	1	2		1	1	1
1	1	1		1	1	2
1	1	1		1	1	1
1	1	2		1	1	2
1	1	1		1	1	2
1	1	1		1	1	1
1	1	Not answered		1	1	1
1	1	1		1	1	1

0.00	0.70	0.25	4.35	0.35	2.70	0.00
1	1	1	Not answered	1	3	Not answered
1	2	1	3	1	1	Not answered
1	1	1	1	1	3	Not answered
1	1	1	1	1	4	Not answered
1	1	1	2	2	2	Not answered
1	2	1	1	1	1	Not answered
1	1	1	2	1	4	Not answered
1	1	1	2	1	1	Not answered
1	1	1	1	1	1	Not answered
1	1	1	1	1	1	Not answered
1	1	1	1	1	1	1

7.70						
6						
6						
4						
5						
6	GAD-7	**Feeling afraid...(0-3)**	0	0	0	0
4		**Becoming easily..(0-3)**	1	0	1	0
4		**Being so...(0-3)**	0	0	0	0
1		**Trouble...(0-3)**	1	0	0	0
2		**Worrying...(0-3)**	1	1	0	0
1		**Not being...(0-3)**	0	0	0	0
1		Feel nervous...(0-3)	1	0	1	0

0	0	0	1	0	0	0
1	1	0	1	0	1	2
0	0	0	1	0	0	1
1	0	0	1	0	0	1
1	1	2	1	0	0	0
0	0	0	1	0	0	0
0	1	0	1	1	1	0

0	0	1	0	0	0	0
1	1	0	0	0	0	0
0	0	0	0	0	0	0
1	0	1	0	0	0	0
0	0	1	0	0	0	1
0	0	0	0	0	0	0
0	0	0	0	0	0	0

	0	0	0	0	0	0
	2	0	0	1	0	1
	1	0	0	0	0	0
	2	0	0	0	0	1
	2	0	0	0	0	0
	1	0	0	0	0	0
	2	0	0	0	0	1

0	0	0	0	0
1	0	1	0	2
0	0	0	0	0
0	0	1	0	2
0	0	1	0	1
0	0	0	0	1
0	1	0	0	1

Chapter 4

Kefir4All, a Citizen Science initiative to raise awareness of the roles microbes play in food and beverage fermentation

Accepted - The Journal of Microbiology & Biology Education (JMBE)

Authors: Liam H. Walsh, Samuel Breselge, José Guilherme Prado Martin, Mairéad Coakley, Eimear Ferguson, Aimee Stapleton, Fiona Crispie, Paul W. O'Toole and Paul D. Cotter

Contributions:

- **Candidate**, **S.B.** and **P.D.C.** conceived and/or designed the study
- **Candidate** and **S.B.** constructed the surveys
- **Candidate**, **S.B.**, **M.C., E.F.** and **A.S.** designed and conducted the outreach activities
- **Candidate**, **S.B.** and **J.G.P.M.** collected samples, extracted DNA. **S.B.** and **J.G.P.M** performed library preparation. **F.C.** conducted sequencing
- **Candidate** conducted analysis of survey and high throughput sequencing data
- **Candidate** wrote the manuscript with contributions from **M.C.**, **S.B.**, **P.W.O.T.** and **P.D.C**
- **M.C.**, **S.B.**, **P.W.O.T.** and **P.D.C** supervised the project

Section 4.0: Abstract

Microorganisms are ubiquitous in nature and are central to human, animal, environmental and planetary health. They play a particularly important role in the food chain, and the production of high quality, safe and health-promoting foods, especially fermented foods. This important role is not always apparent to members of the public. Here we describe Kefir4All, a citizen science project designed to provide the general public with an opportunity to expand their awareness, knowledge and practical skills relating to microbiology, introduced through the medium of producing a fermented food, i.e., milk kefir or water kefir. During the course of Kefir4All, 123 citizen scientists, from second level school and non-school settings, participated in a study to track changes in the microbial composition of kefirs, by performing and recording details of milk kefir or water kefir fermentations they performed in their homes or schools over the 21 week project. At the start of the study, the citizen scientists were provided with milk or water kefir grains to initiate the fermentations. Both types of kefir grain are semi-solid, gelatinous like substances, composed of expolysaccharides and proteins, containing a symbiotic community of bacteria and yeast. The experimental component of the project was complemented by a number of education and outreach events, including career talks and a site visit to our research centre (Kefir Day). At the end of the study, a report was provided to each citizen scientist, in which individualised results of their fermenting activities were detailed. A number of approaches were taken to obtain feedback and other insights from the citizen scientists. Evaluations took place before and after the Kefir4All project to gauge the citizen scientist's self-reported awareness, knowledge and interest in microbiology and fermented foods. Further insights into the level of citizen science participation were gained through assessing the number of samples returned for analysis and the level of participation of the citizen scientists throughout the project. Notably, the survey results revealed a self-reported, increased interest in, and general knowledge of, science among the Kefir4All citizen scientists after undertaking the project, and a willingness to take part in further citizen science projects. Ultimately, Kefir4All represents an example of the successful integration of citizen science into existing education and research systems.

Section 4.1: Introduction

Although the term ‘citizen science’ can be flexibly applied, due to an absence of an agreed international definition [1], citizen science, typically involves collaborative research between scientists and non-scientists, i.e., the ‘citizen scientists’ [2]. Citizen science has been used to date to carry out studies across a range of disciplines [3], e.g., for wildlife surveys, such as birdwatching [4], medicine in projects like ‘PatientsLikeMe’ [5], in food fermentation projects such as ‘Ferme Pekes’ [6] and the ‘Global Sourdough Project’ [7]. Interest in citizen science has increased, as reflected by the number of associated studies and publications [8]. It is worth noting the growing number of guidelines aiming to establish the best practice principles for citizen science, including the European Citizen Science Association's ‘Ten Principles of Citizen Science’ [9]. However at present there are few recommendations derived from quantitative and qualitative assessments, coupled with personalised accounts from citizen scientists actively involved in a project, with a consistent level of participation. One clear advantage of citizen science relates to the synergistic relationship between participants and researchers, encouraging scientific openness [1]. A number of additional benefits can be availed of by researchers by involving participants, who can provide a greater quantity and/or quality of samples and/or data than would be possible for a small number of professional scientists in a research setting [10]. Citizen science also provides unique opportunities to communicate with the general public, aiding in the identification of gaps in public knowledge and key research questions. For example in the Global Sourdough Project, the community-scientist network of bakers provided 500 sourdough starters from four continents and personalised accounts from over 1,000 people concerning the origin, culinary practices and distinct flavours of their own sourdoughs. The results of the sourdough study further informed consumer/producers of the factors influencing the microbial diversity of sourdough, dispelling widespread assumptions that the geographical location of the sourdough was a primary influencer of diversity. While the Global Sourdough Project revealed similarities and differences in the microbial composition of different examples of a specific fermented food, an alternative question that can be addressed through citizen science relates to the extent to which a single common-source fermentation-associated community can change when disseminated to participants.

Notably in this regard, milk kefir and water kefir are examples of fermented foods that are accessible to the general population due to their ease of production, low-cost production and history of safe consumption [11] and are suited to microbiology-related citizen science projects. In particular, the starting point of the fermentation, the symbiotic communities of bacteria and yeasts (SCOBY) contained within an exopolysaccharide matrix, referred to as kefir 'grains' [11], produce a wide variety of fermented beverages with different characteristics [12-14] and can be easily shared with numerous participants.

Here we describe Kefir4All, a kefir focused citizen science project during which citizen scientists took part in scientific learning and discussion, while performing fermentations and recording observations in a household environment. We use survey data collected throughout the course of the project to measure the initial interest in the project, and to track that interest as the project progressed. We also examine the impact of outreach activities, for example in maintaining continuous participation, a recognised limiting factor for projects of this nature [15], and most importantly we examine the educational impact of taking part in a citizen science project from the participant's perspective [15].

Section 4.2: Methods

4.2.1 Overview of the study

102 citizen scientists were recruited from across nine school settings and were refered to as school groups, while 21 citizen scientists were recruited outside of school settings, and were refered to as 'non school group - milk kefir' or 'non school group - water kefir', based on the kefir type they were fermenting during the project. At the start of the project, the 123 citizen scientists received a milk kefir or water kefir grain (each originating from a common source), the required fermentation equipment (details below) and fermentation guidelines. Milk or water kefir was then produced by the citizen scientists in private households for up to 21 weeks. Production of the milk or water kefir required the 'transfer' of the kefir grain to fresh medium every 1-3 days. The milk kefir medium represented different types of milk, while the water kefir medium were different combinations of sugary water, with or without additional fruits. Kefir grains and liquids were collected at week (wk) 1, wk5, wk9, wk13, wk17 and wk21, and metadata

(such as information about storage conditions, type of milk or sugar water used, frequency with which the grains were 'transferred') were reported after each fermentation by citizen scientists using the Survey Monkey platform [16].

4.2.2 Participant recruitment

Citizen scientists were recruited during a series of fermentation workshops, given in-person in a school for the school group, or online for the non-school group. The workshop was divided into (1) a 40-minute talk (Introduction to microbiology, Introduction to DNA sequencing, and Introduction to fermented foods), (2) a 20-minute workshop on kefir fermentation, followed by an opportunity for questions from the potential citizen scientists and (3) distribution of the kefir fermentation kits. The workshop was used to provide background scientific knowledge about microbiology, fermentation and bioinformatics, which aided in the explanation of the research project and the role of the citizen scientists, with an emphasis on how their contributions would aid the research project and its objectives. Citizen scientists further received recordings of the fermented food workshop and steps involved in sample collection to use as supporting materials throughout the project.

4.2.3 Fermentation equipment

Citizen scientists received fermentation kits that included a kefir grain, fermentation vessels, thermometers, utensils, pH strips, collection tubes and fermentation guidelines. Fermentation equipment was purchased from Kefirako (*Borgia*, Slovenia)[1], including the fermentation jars, pH test strips and thermometers. Collection tubes were purchased from Thermo Scientific™, UK. The milk kefir grains were sourced from *Every good thing*, UK[2]. The water kefir grains were sourced from *Live Ferments*, Ireland[3]. Fermentation guidelines generated by researchers conducting the study, included detailed written

[1] Borgla d.o.o.,Valvasorjeva ulica 40, 2000 Maribor, Slovenija
[2] https://everygoodthing.co.uk/products/milk-kefir-grains
[3] https://kefirgrains.ie/

instructions on the steps involved in the fermentation process and in sample collection, recordings of the workshop[4] and links to surveys used to collect metadata (see Supplementary Materials).

4.2.4 Surveys and evaluation measures

The Kefir4All project and the associated engagement activities were evaluated through a series of surveys. Our study included four surveys: 'Workshop Survey', 'Fermentation Survey', 'Project Completion Survey - Citizen Scientists' and 'Project Completion Survey - Teachers' (see Supplementary Materials), all constructed using Survey Monkey [16]. The Citizen Scientists were asked to complete the 'Workshop Survey', at the end of our recruitment workshop; which included opinions provided by members of the public who did not subsequently take part in the Kefir4All study. The 'Workshop Survey' examined the self reported educational impact on attendees as a result of attending a 1 hour workshop. The 'Workshop Survey' also assessed the pre-existing microbiology and fermentation-related knowledge of the participants. The 'Fermentation Survey' was used to collect metadata about each fermentation process completed by the citizen scientists. The 'Project Completion Survey - Citizen Scientists' and the 'Project Completion Survey - Teachers' primarily examined the self reported educational impact on the citizen scientists, as a result of taking part in the project, and the personal accounts of the citizen scientists and teachers respectively, relating to strengths of the project and /or suggested improvements. Citizen scientists were asked to complete the 'Fermentation Survey' throughout the project, and the 'Project Completion Survey - Citizen Scientists' at the end of the project. School-based coordinators (teachers) of the Kefir4All project were asked to complete the 'Project Completion Survey - Teachers' at the end of the project. All survey responses were analysed qualitatively.

4.2.5 Outreach activities

[4] https://youtube.com/playlist?list=PLlgXUwWwJdfm5KcWpFpa0f2pO0NxohG3T

To facilitate continuous participation with, and maintain interest from the citizen scientists, a number of activities and challenges were devised. Six Careers Events (three speakers per event) were arranged, which the citizen scientists could attend virtually and hear the accounts and experiences of researchers at different stages of their careers. To encourage continued participation of the citizen scientists in the project, a leader board was circulated every week, highlighting the citizen scientists with most activity. The leader board also identified the 'citizen scientist of the week', the one who completed most fermentations that week (based on number of surveys they completed). The citizen scientists received goodie bags containing beanie hats, cups, and stationary upon providing their final set of samples. Citizen scientists were able to continuously engage with and avail of the expertise of the researchers throughout the project, *via* email or in person during sample collection and return school visits. The school based citizen scientists were invited to the Teagasc Food Research Centre, Moorepark for a 'Kefir Day' in May 2022. During this visit, the citizen scientists took part in tours of the research facilities, short laboratory based workshops, and visited scientific stands covering a number of topics such as bacteriophage, food fermentation, the gut microbiome and next generation sequencing. Completion certificates and preliminary data reports, including the microbial composition of the participant's kefir grains and liquid samples were made available to each contributing citizen scientist at the end of the project. Preliminary data was presented in a PDF document, together with a recap of the conceptions discussed during the project.

4.2.6 Bioinformatic analysis

Milk kefir and water kefir shotgun metagenomic sequences generated in Kefir4All, or acquired from Walsh *et al*. [17]and Breselge *et al.* (in preparation) were processed using the Teagasc high performance computing cluster. Sequences from Walsh *et al*. [17] and Breselge *et al*. (in preparation) were included to compare compositional profiles between metagenomes acquired using a more traditional laboratory based approach and a citizen science approach. The laboratory based approach included the fermentation and sampling of milk and water kefir under standardized conditions in a laboratory setting. Initially, raw

paired-end FASTQ files containing the metagenomic shotgun sequences were trimmed using Trimgalore (v.0.6.1) [19], to remove adapter sequences and low quality reads (average quality score <Q20), fragmented (<75 bp) and with more than two ambiguous nucleotides. Contaminating reads were removed, through an initial alignment of metagenomes against a reference database, followed by read removal of any aligned reads. Specifically milk kefir metagenomes were searched against a reference database containing a bovine, phiX 174, and human genome using Bowtie2 (v.2.3.4) [20] with the parameter 'sensitive-local'. Water kefir metagenomes were searched against a reference database containing a phiX 174 and human genome (GRCh38) using Bowtie2 (v2.3.4) with the parameter 'sensitive-local'. Reference genomes were downloaded from iGenomes[5] and represented the most up to date versions available in NCBI (November 2022). Remaining high-quality reads were sorted and split to create forward, reverse and unpaired reads output FASTQ formatted files for each metagenome. Compositional analysis was performed with MetaCache (v 2.3) [21].

4.2.7 Statistics and data visualisation

Analysis of survey data from each of the outlined survey's was completed in R-4.0.2 [22]. Correlation analysis was calculated using the Kendall correlation method [23], through the ggscatter function included in ggpubr (v0.4.0). Frequency counts for all survey responses collected in this study were described in a contingency table using the stats package [3]. The Fisher's Exact Test was used to determine significant associations between frequency counts of categorical survey responses in the "Workshop survey". The Chi-square test of independence was used to determine significant associations between frequency counts of the categorical variables 'Fermentation characteristics' and 'Fermentation categories' , 'Grain wash', 'Room temperature', 'pH', 'Grain weight', 'Water type', 'Liquid volume' and 'Sugar weight', all of which were collected in the "Fermentation survey". P values obtained using the Fisher's Exact Test and the Chi-square test were adjusted using the Benjimani- Hochberg correction to account for multiple comparisons. Beta diversity values were calculated using the vegdist function of the vegan (v2.6.2)

[5] https://support.illumina.com/sequencing/sequencing_software/igenome.html

package [24]. The fviz_nbclust function included in factoextra (v1.0.7) [25] was used to predict the optimal number of clusters using the average silhouette method. Cluster validation was performed using the Dunn index test [26] through the dunn function in clValid (v 0.7). Hierarchical clustering was performed using the hclust function using the ward.D2 agglomeration method in stats (v4.1.2) [22]. Bootstrapping was applied to assess cluster stability using the clusterboot function of the fpc (v2.2.10) package [27]. Jaccard coefficient values obtained using bootstrapping were used to evaluate cluster stability, values <0.6 were considered unstable and values > .85 were considered highly stable. Bar charts were produced from survey data using the geom_segment function of the ggplot2 (v3.3.6) package [28]. Word clouds were produced using the geom_text_wordcloud_area function in the ggwordcloud package (v0.5.0) [29]. The ordination plot was produced from Bray-Curtis dissimilarity beta diversity values between milk and water kefir metagenomes using the geom_point function of the ggplot2 (v3.3.6) package.

Section 4.3: Results

4.3.1 The 'Workshop Survey' highlights the importance of promoting microbiology to the general population.

The Kefir4All project involved a series of steps (Figure 1) that began with the organisation and advertisement of workshops (see methods), including letters being sent to local secondary schools to encourage their participation and posts on social media. The citizen scientists were recruited at the end of a 1 hour workshop (see methods), primarily given in schools settings. At the beginning of this workshop, the potential citizen scientists were asked to complete Questions 2-15 of the 'Workshop Survey', to assess their self-reported background knowledge and interest relating to the topics of microbiology and fermentation. The remaining questions (Questions 1 & 16-19) of the 'Workshop Survey' were completed at the end of the workshop to assess if attendance at the workshop, raised awareness and interest in potential participants, of the topics microbiology and fermentation (Figure 1 and S1).

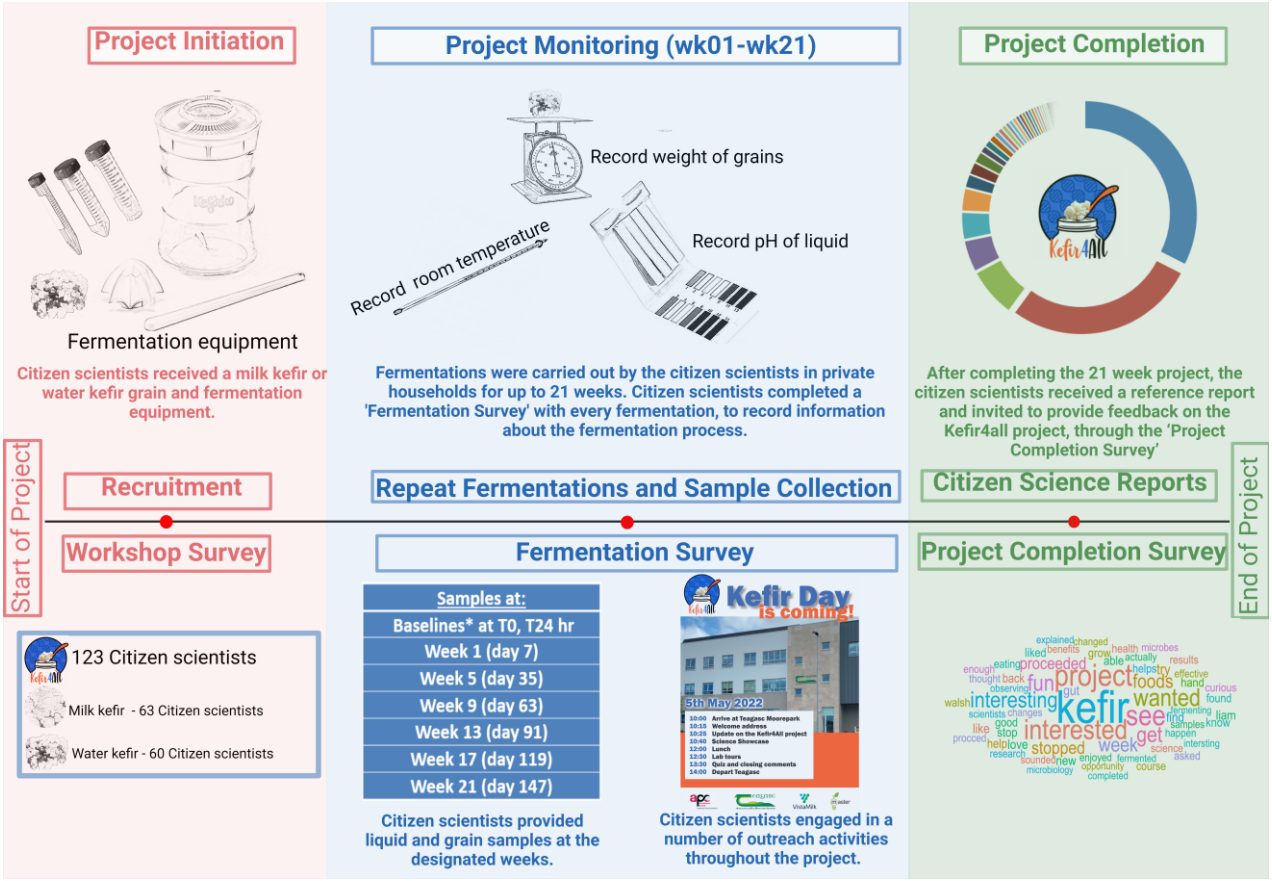


Figure 1: Experimental design of Kefir4All (including details of the surveys undertaken). Stages included 'Project Initiation', 'Project Monitoring' and 'Project Completion'. 'Project Initiation' included participant recruitment, training and distribution of the fermentation equipment. A 'Workshop Survey' was used to capture feedback from participants at this stage of the project. In the 'Project Monitoring' stage, participants performed repeat fermentations in private households for up to 21 weeks, and recorded details of milk kefir or water kefir fermentations using the Survey Monkey platform [24]. During this period the 'Fermentation Survey' was used to assess the participation of the citizen scientists. Further insights into the level of participation was gained by assessing the number of kefir grains and liquids returned at week (wk) 1, wk5, wk9, wk13, wk17 and wk21. T0 samples represent control milk kefir and water kefir grain and liquid samples fermented under laboratory/controlled conditions and produced by professional scientists. At 'Project Completion', participants received a report in which individualised results of their fermenting activities were detailed. Participants were further invited to provide feedback through the 'Project Completion Survey'.

The 'Workshop Survey' was completed by 87 individuals, 58 (67%) of whom volunteered for the Kefir4All project (Figure S1A, Figure 2A and Figure 2B). The survey responses collected from responders who took part in the project generally followed the same patterns as the survey responses from responders who did not take part in the project, with two notable exceptions, i.e., participants were more likely to take part in the project if they had a greater interest in fermented foods (Figure 2A) and/or microbiology (Figure 2B)(Fisher's Exact Test, adjusted p-values <0.01).

Of the total responders, 25 (29%) had an interest in microbiology, few were disinterested in microbiology (n=5, 5%), and approximately half of the responders (n=43, 50%) had a neutral interest (Figure 2B). A neutral interest was the most selected option in responders who agreed to take part in the project and those who did not. However, more of the non - participating responders selected the option 'I don't know what it means' (30% of the total non - participating responders) compared to participating responders (7% of the total participating responders)(Figure 2B). Most responders felt they lacked background knowledge, approximately half of the responders (n=41, 47%) considered that they had a little background knowledge, and a further 28 (32%) responders felt they had no background knowledge, while only one (1%) and 11 of the 87 responders (13%) regarded themselves as knowledgeable and very knowledgeable, respectively (Figure S1C). 64 responders (74%) expressed an interest in fermented foods (Figure S1D), despite many responders 'never' consuming fermented foods (Figure S1F) and the majority having no previous experience of fermentation (n=72, 84%) (Figure S1G). Knowledge gaps were also evidenced by the fact that, prior to the workshop, 37 responders (45%) were not aware that foods they previously consumed were fermented. Additionally, a small minority of responders (n=8, 14%) referred to unfermented foods such as milk (n=4) or alternatively stated that they were not sure (n=3) when asked the question what fermented foods do you eat?

While many of the non-participating responders expressed an interest in fermented food (45%), the remaining non-participating responders selected 'Not so interested' (37%) or 'Not at all interested' (19%), which was in stark contrast to the participating responders with few selecting the 'Not so interested' (9%) or Not at all interested' (2%) options (Figure 2A). We examined the association between interest in fermented foods and consumption of fermented foods. When asked how often they consumed fermented foods, 32 (37%) of the responders selected 'Never', 21 (25%) selected 'Occasionally', 22 (26%) selected 'Regular' and nine (9%) selected 'Daily'. Of those who declared that they had previously consumed fermented food, only 33% of responders (n=18) enjoyed the sour taste of the food. The responders who 'Never' consumed fermented food (Figure S1F) also tended to select 'Somewhat interested' (n=13, 34%), 'Not so interested' (n=10, 67%), or Not at all interested' (n=5, 83%) in fermented

foods. In contrast, most 'Regular' and 'Daily' consumers of fermented food expressed an interest in fermentation with only one (17%) daily consumer and two (13%) regular consumer expressing a disinterest in fermented foods. Similarly, only one (11%) daily consumer and seven (32%) regular consumers were unaware that some of the foods they had consumed were fermented foods prior to the workshop compared to 18 (62%) of those who reported that they had not previously consumed fermented foods. However, known prior consumption of fermented food did not associate with taking part in the project (Fisher's Exact Test, p value >0.05), with 19 of the total 32 responders (52%) agreeing to take part in the project despite never, to their knowledge, consuming fermented foods.

We then examined the association between interest in fermented foods/microbiology and subject choices among students in second level education of relevance to the citizen science project, specifically if the responder chose to study the topics of Biology (n= 36, 41%), Home-economics (n=8, 9%), both (n=36, 30%) or neither subject (n=16, 19%). It was established that subject choice was not associated with interest in fermented foods (Fisher's Exact Test, p value >0.05), with a higher portion of responders (n=5, 31%) who did not take either subject expressing an interest in fermented foods. Subject choice did associate with interest in microbiology (Fisher's Exact Test, p value >0.05), however many of the responders who selected the option "I don't know what it means" when asked about their interest in microbiology, had selected biology (n= 10, 83%) as part of their studies. Furthermore, subject choice did not associate with participation rates in the project (Fisher's Exact Test, p value >0.05), with similar subject choice profiles between those that participated in the project and those that did not. Indeed, 13 responders (81%) who did not select Biology and or Home-economics as part of their studies took part in the study, accounting for 22% (n=13) of the recruited responders. When considering where the responders source their scientific information, the majority of the responders used YouTube (n=34, 39%), followed closely by information received through schools (n=30, 34%), with few responders using websites (n=9, 10%) and/or books (n=3, 3%) (Figure S1I). However, responders who utilised books and websites tended to engage with the information provided with only one responder (11%) who selected 'websites' choosing to never engage with those sources of information. Six responders (20%) engaged with scientific information in

schools every day and a further 12 (40%) responders engaged on a regular basis with scientific information provided in schools. Those who identified schools as the primary source of their scientific knowledge also had higher levels of daily and regular engagement compared to YouTube, which while having lower proportions of no engagement had higher portions of occasional engagement compared to schools.

71 of the 87 (82%) responders found the workshop to be interesting/very interesting, with 77 responders (89%) reporting an improved interest and general knowledge of microbiology and/or fermented foods as a result of the workshop. Specifically, 54 responders (62%) reported an improved knowledge relating to both microbiology and fermented foods, 18 reported an improvement solely in microbiology and five in fermented foods, only 9 of the 87 (10%) responders felt no benefit as a result of attending the workshop (Figure S1J). The majority of the responders who did not feel they benefited from the workshop (Figure S1J) expressed an absence of interest in fermentation and or an absence of background knowledge in microbiology. Of these nine responders, four (50%) were “not at all interested in fermentation” and five (62%) selected “I don’t know what it means” when asked about background knowledge of microbiology. Despite this, of the 12 responders who did not know what microbiology was prior to the workshop, seven (58%) reported an improvement of both their interest and general knowledge of microbiology by the end of the workshop. Greater learning outcomes tended to be achieved by responders who expressed an interest in the topic Microbiology (Fisher's Exact Test, adjusted p-value <0.05) (Figure 2C), which in turn had a relationship with subject choice (Fisher's Exact Test, adjusted p-value <0.05) (Figure 2D). Similarly, responders who expressed a higher interest in microbiology and/or fermentation were more likely to take part in the citizen science project (Fisher's Exact Test, p value >0.05) (Figure 2A and 2B). For example 83% of responders (n=5) who selected ‘Not at all Interested’ when asked to rate their interest in fermented foods did not taking part in the project, while in contrast 81 responders (n=17) who selected ‘Interested’ and 100% of responders (n=5) who selected ‘Extremely Interested’ took part in the project (Figure 2A and 2B). Furthermore, 75% of responder who expressed that they did not benefit from the workshop did not take part in the project, while only 23% of responders who expressed an increased interest and general

knowledge of microbiology and fermented foods after attending the workshop did not take part in the project.


A.
Please rate you interest in fermented foods.
Somewhat Interested
Interested
Not so Interested
Not at all Interested
Extremely Interested
Are you participating in the Citizen Science project?
Yes
No
Fisher's exact test
p < 0.05
Number of responses
B.
What is your current interest in Microbiology?
Neutral
Interested
I don't know what it means
Very disinterested
Disinterested
Are you participating in the Citizen Science project?
Yes
No
Fisher's exact test
p < 0.05
Number of responses
C.
Has this talk improved your interest and general knowledge?
Yes both microbiology and fermented foods
Yes but just microbiology
No
Yes but just fermented foods
What is your current interest in Microbiology?
Interested
Neutral
Disinterested
Very disinterested
I don't know what it means
Fisher's exact test
p < 0.05
Number of responses
D.
Are you continuing Biology and or Home-Economics
Yes but just biology
Yes
No
Yes but just home-economics
What is your current interest in Microbiology?
Interested
Neutral
Disinterested
Very disinterested
I don't know what it means
Fisher's exact test
p < 0.05
Number of responses
E.
Please rate your background knowledge about fermentation.
A little background knowledge
No background knowledge
Some background knowledge
Good background knowledge
Awesome background knowledge
Please rate you interest in fermented foods
Extremely Interested
Interested
Not at all Interested
Not so Interested
Somewhat Interested
Fisher's exact test
p < 0.05
Number of responses
F.
Please rate your background knowledge about fermentation.
A little background knowledge
No background knowledge
Some background knowledge
Good background knowledge
Awesome background knowledge
What is your current interest in Microbiology?
Interested
Neutral
Disinterested
Very disinterested
I don't know what it means
Fisher's exact test
p < 0.05
Number of responses
G.
Please tell us what you thought of our workshop
interesting
good
lot
information
people
science
discussing
parts
biology
different
loved
enjoyed
microbiology
just
kept
presented
get
project
add
step
talking
experiment
class
topic
really
improve
something
muchthink
know
alot
practical
health
foods
idea
thought
learned
learn
fun
presentation
engaging
made
still
started
fermented
nefits
liked
selected
food
informative
interested
new
like
also
kefir
dont
didn't
thing
ferment
dna
microbiology
time
knew
laptop
perfect
babout
find
opportunity
viruses
overload
learnt
knowledge
entertained
missed
making
comfort
sounded
unresponsive
spoke
yet
bconfused
stop
encouraging
enjoy
future
happening
important
subject
without
research
tter
types
highly
help
plus
look
example
points
incredibly
educated
bored
love
activities
anything
serious
informed
involved
seats
listeningnice
fermenting
discussed
covered
understand
questions
speaker
concise
room
material
haven't
videos
don't
take
fore
neficial
explain
ansr
interest
pictures
dragged
fermentations
drink
citizens
nevercleary
voice
speakers
effort
sound
old
m
big
rar
try
ts
want
way
felt
motivated
participate
healthy
process
watch
workshop
asking
fermentation
aspect
constantly
youtu
asking
encouraging
stop
bconfused
yet
unresponsive
sounded
entertained
missed
making
comfort
plus
laptop
perfect
babout
find
opportunity
viruses
overload
learnt
knowledge
points
incredibly
educated
example
time
look
alot
practical
health
foods
idea
thought

4.3.2 Tracking participation within the Kefir4All cohort provides insights into the advantages and disadvantages of citizen scientist recruitment based on a 'Good faith approach'

Kefir4All consisted of 123 participants, separated into nine local school groups (with an average of 11 participants per school group) (n=102) and two non-school groups (n=21) separated into non-school group - milk kefir and non-school group - water kefir (Figure 1). The participants were selected based on their interest in taking part in the project. After quality filtering (which included the removal of blank/incomplete surveys or those missing essential details such as the participants ID number), the 'Fermentation Survey' consisted of 1,823 responses. These were provided over the course of 21 weeks (wk1-wk21) (Figure 1 and 3B). The 'Fermentation Survey' functioned as a fermented food log, in which citizen scientists could record information and observations about each fermentation they performed, e.g., grain growth, fermentation characteristics (bubbles, biofilm formation, etc.) and room temperature. The completion rate of these surveys provided a measure of the number of citizen scientists who contributed to the project. The number of samples provided also reflected the citizen scientists participation, with 516 kefir samples provided by citizen scientists across six time points over 21 weeks (Figure 3D). As recruitment was performed in 'good faith', with no prior knowledge of the citizen scientist's interest in, and or likelihood to complete the project, we initially used the 'Fermentation Survey' to evaluate the contributions of the citizen scientists. This was judged based on the number of surveys completed by each participant. If a citizen scientist completed $\geq$20 'Fermentation Surveys', they were considered 'Regular fermenters' (n=44). Participants were deemed 'Occasional fermenters' (n=10) if they completed greater than 10, but less than 20 surveys, 'Rare fermenters' (n=34) if they completed <10 'Fermentation Surveys' or 'Non-fermenters' (n=35), if no surveys were complete during the project (Figure 3A). Across all of the 11 groups recruited, there was an average of four 'Regular fermenters', one 'Occasional fermenters', four 'Rare fermenters' and three 'Non-fermenters'. Citizen scientists were asked to provide six kefir liquid samples and six kefir grain samples during the project (Figure 1). On average, the citizen scientists provided four kefir liquid samples and four kefir grain samples. Correlation analysis using the Spearman method indicated that there was a positive relationship between the group size and the

frequency of ‘Non-fermenters’ (Spearman's correlation adjusted p-value=0.02) and ‘Rare fermenters’ (Spearman's correlation adjusted, p-value= 0.03). Of particular interest, we noted a higher proportion of ‘Regular fermenters’ in school groups where teachers classified their contribution as ‘A lot’ or ‘A moderate amount’, compared to teachers who did not complete the evaluation survey, or who selected ‘None at all’.

4.3.3 Continued participation played an important role in the success of Kefir4All project

We monitored the number of ‘Fermentation Surveys’ completed during the Kefir4All study, separated according to the weeks of the study (Figure 3C). During the first five weeks, participation was high with 214 surveys completed in the seven-day period of wk1 and 91 active participants providing 109 samples, and 502 surveys completed within the 28-day period of wk5, with 75 active participants providing 99 samples in wk5. However, the survey completion rate, the active number of participants and the number of samples provided (Figure 3C and 3D) all gradually decreased following wk5, with the lowest values recorded at wk21 with 182 surveys completed, 39 active participants and 65 samples provided.

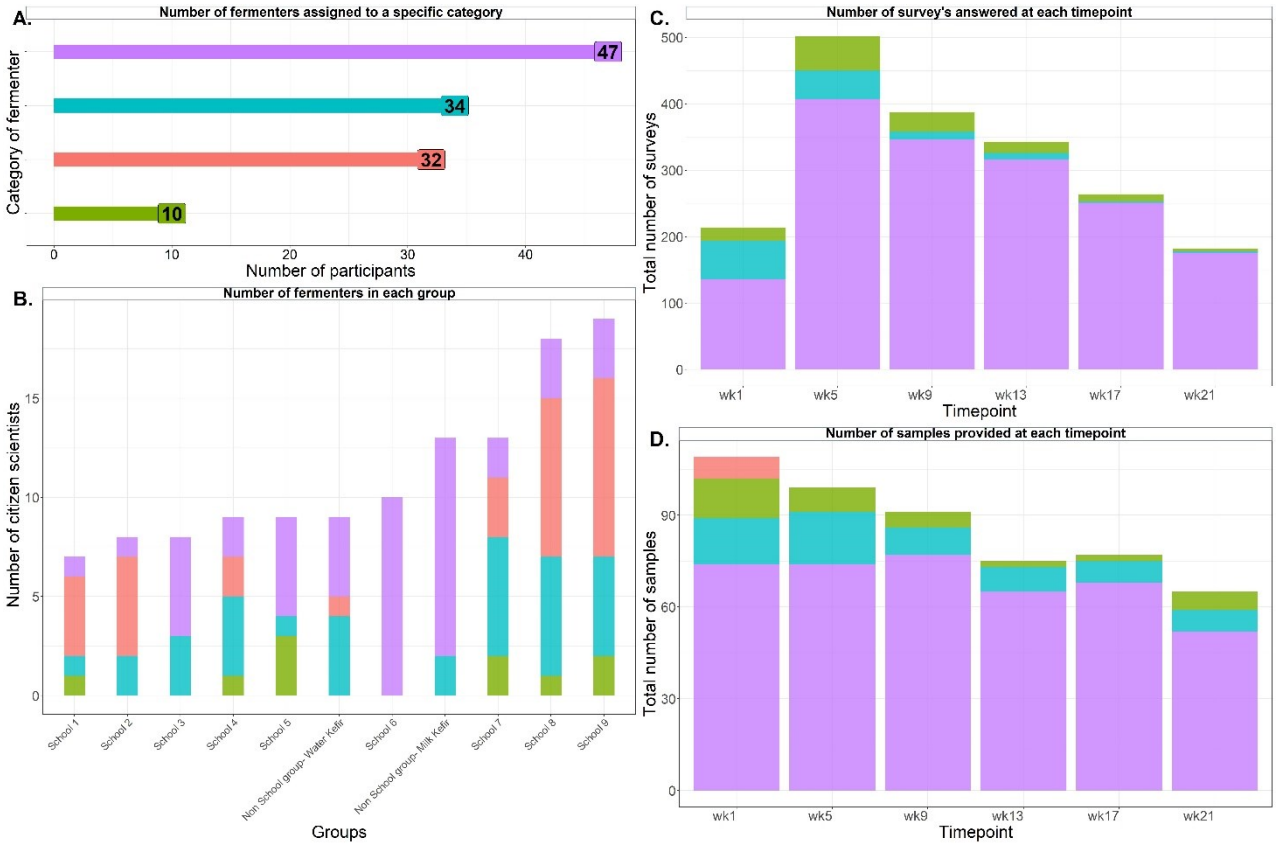


Figure 3: Overview of the contributions of the citizen scientists during Kefir4All. **(A.)** Number of citizen scientists assigned to the fermentation categories; ‘Regular fermenters’, ‘Non-fermenters’, ‘Rare fermenters’, and ‘Occasional fermenters’. **(B.)** Breakdown of the number of fermentation categories according to the groups (Schools and general members of the public) (X-axis). **(C.)** Number of survey answers, according to weeks of the study (X-axis); the colours on the bar charts represents the number of surveys completed by participants of each of the fermentation categories during each timeframe of the study (X-axis). **(D.)** Number of samples (milk kefir and water kefir grain and liquid samples) provided at each week of the study (X-axis); the colouring of the bar charts represents the number of samples provided by participants of the fermentation categories at each week of the study.

When this pattern of decreased participation over time became apparent, the school coordinators requested engagement activities after wk5. To address this, a leaderboard (Figure 4) was developed and circulated weekly to the citizen scientists, graphically displaying the collective survey responses per group that week (Figure 4A), the collective number of surveys answered each day that week, separated according to group (Figure 4B), and the collective number of surveys answered since the beginning of the project (Figure 4C). In addition, career events were organised, during which career opportunities available in multiple sectors of science, such as chemistry, microbiology and agricultural science, were discussed, and the career paths of numerous researchers, at different stages of their careers were discussed. Citizen

scientists were also invited to a 'Kefir Day', for which they were introduced to the research and laboratory facilities within the Teagasc Food Research Centre, Moorepark. Finally, completion certificates, Kefir4All merchandise and a personalised report detailing and graphically displaying the microorganisms in the participant's kefir (Figure 5) were highlighted as materials that would be provided at the end of the project.

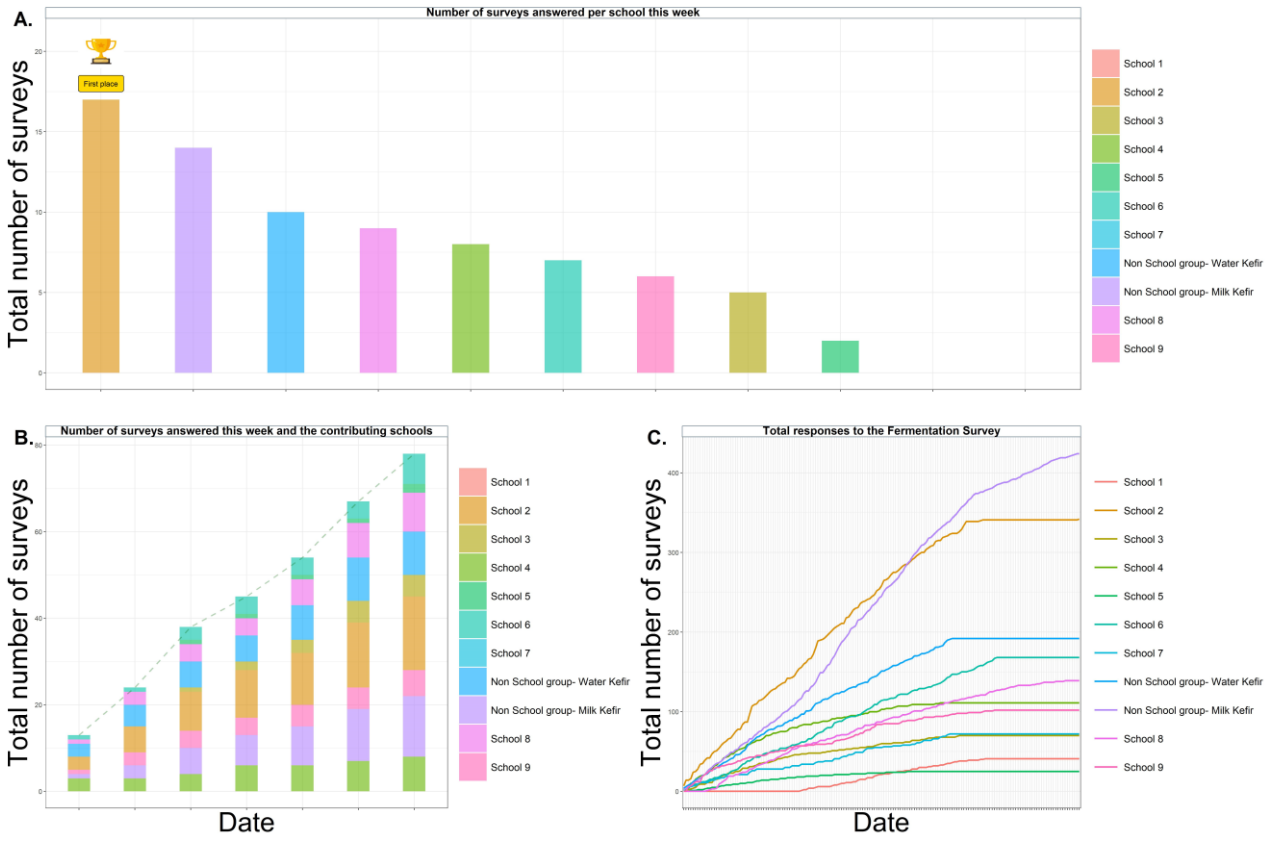


Figure 4: (A.) The first of the leaderboard plots displaying the collective survey responses per group recorded within a week of fermentation, and ranked according to the number of surveys completed, with the highest group represented with the icons 'First place' and a trophy, line plots are coloured by the school/group. **(B.)** Collective number of survey's (Y-axis) answered each day (X-axis) since the beginning of the week bar charts are coloured according to the number of surveys contributed by the group/school during each day. **(C.)** Collective number of survey's (Y-axis) answered by each group per day (X-axis) since the beginning of the project, line plots are coloured by the school/group.

We noted the consistent number of samples provided by 'Regular Fermenters', across all stages of the project and a subtle increase in the number of samples provided by 'Regular Fermenters' at wk09 and wk17, which corresponded to the introduction of both the engagement activities described (Figure 3D). Furthermore, according to the 'Evaluation Survey', the majority of citizen scientists and school

coordinators reported that the engagement activities contributed to continued participation (see below for further details).

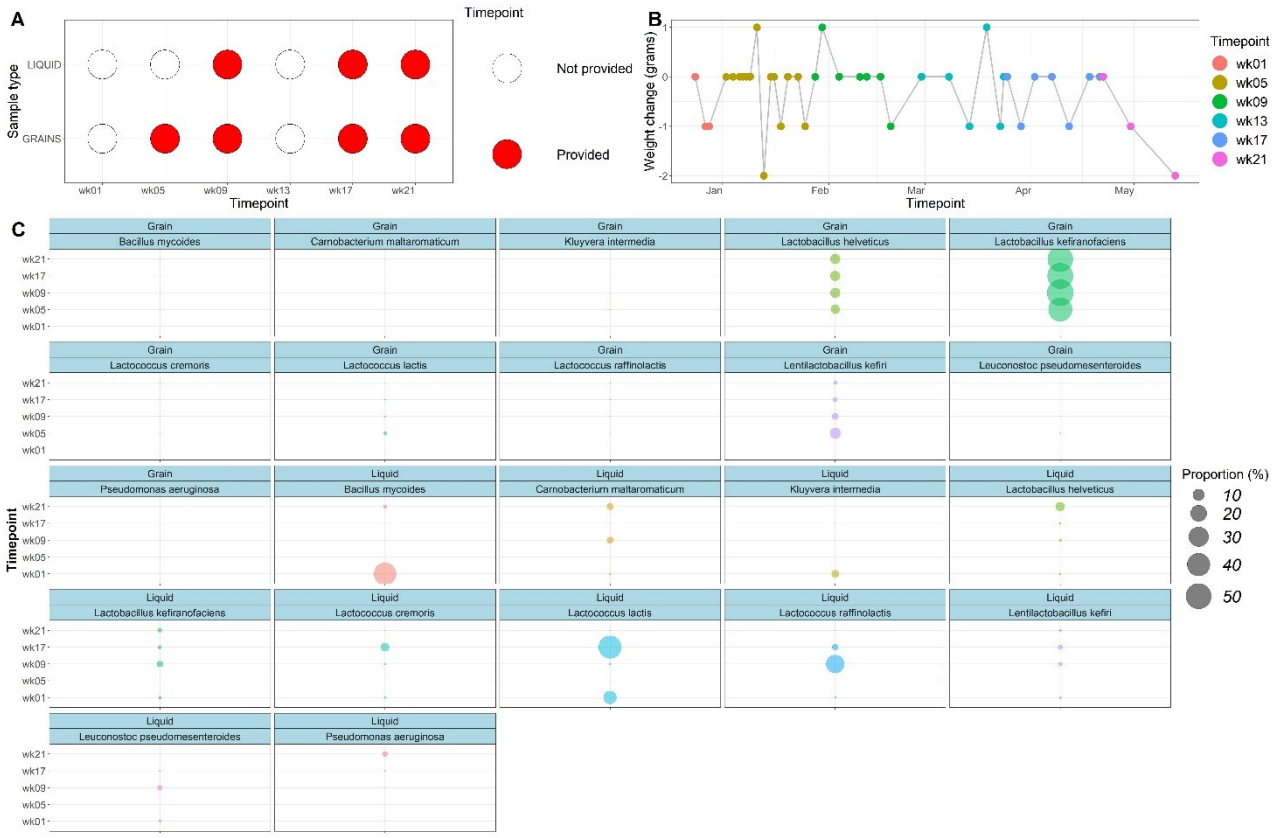


Figure 5: Example figures included in the personalised report of a representative citizen scientist. **(A.)** Samples provided by a representative citizen scientist during the Kefir4All study, separated by weeks (wk), a red colour indicates that samples (kefir grain and liquid) (Y-axis) were provided at the outlined time point (X- axis) a white colour indicates that samples were not provided at the outlined time point (X-axis). **(B.)** Changes in the weight of the citizen scientist's kefir milk grain (in grams) over the course of the project, separated by the weeks (wk) of the study. **(C.)** Changes over the weeks (wk) of the study in the detection and relative abundance of species (X-axis) detected at >1% relative abundance (RA) in at least one of the samples provided by the citizen scientists, separated by the grain and liquid samples.

4.3.4 Survey data reveals a relationship between fermentation parameters and the occurrences of fermentation characteristics.

The chi-square test of independence was employed to analyse the relationship between the categorical variables in the 'Fermentation survey' and fermentation characteristics such as grain growth, pH values and the frequency of the observed 'Fermentation characteristics' (Figure 6). Categorical variables included 'Fermentation categories', which, in the case of milk kefir included '300 ml of full fat milk', '300 ml of low fat milk', '300 ml of goat milk' and '300 ml of skimmed milk'. 'Fermentation characteristics' in milk kefir

included ‘Bubbling’, ‘Coagulation’, ‘Whey and curds’ and ‘Unchanged’. For water kefir, categorical variables included ‘Fermentation categories‘ such as ‘12 gram of white sugar and 1 dried fig’, ‘12 gram of white sugar, 1 dried fig, 1 slice of fresh lemon’, ‘12 gram of white sugar and 1 dried apricot’ and ‘12 gram of brown sugar’. The ‘Fermentation characteristics’ in water kefir included ‘Biofilm’, ‘Bubbling’, and ‘Unchanged’ (Figure 6A).

Following p value adjustment for multiple comparisons, the categorical variables ‘Fermentation categories‘, ‘Grain wash’, ‘Room temperature’, ‘pH’, ‘Grain weight’ and ‘Liquid volume were all shown to be significantly associated with the observed fermentation characteristics (see methods) in milk kefir (Pearson's Chi-squared test, adjusted p-values <0.01). Similarly, the categorical variables ‘Fermentation categories’, ‘Grain wash’, ‘Room temperature’, ‘pH’, ‘Grain weight’, ‘Water type’, ‘Liquid volume’ and ‘Sugar weight’ were all shown to be significantly associated with the observed fermentation characteristics in water kefir (Pearson's Chi-squared test, adjusted p-values <0.01). The strongest positive correlations were primarily observed between the ‘Fermentation categories’ and ‘pH’ values for both milk and water kefir when considering ‘Fermentation characteristics’. ‘Liquid volume’ and ‘Grain weight’ demonstrated a number of strong correlations in the milk kefir cohort (Figure 6B). Notably, for water kefir, the fermentation parameters ‘Liquid volume’ and ‘Sugar weight’ contained a limited range of values as participants largely adhered to the recommended set ratios of 300ml of water to 12 grams of water kefir grains during the study (Figure 1). However, the 300ml of milk to 12 grams of milk kefir grains was not as readily adopted due to difficulties in grain growth throughout the project.

Within the closing comments of the fermentation survey, grain was the word that most frequently appeared (16 participants in the water kefir cohort and 11 participants in the milk kefir cohort). Closer inspection highlighted that six participants in the milk kefir cohort and 11 participants in the water kefir cohort expressed concerns over grain growth, for example “I'm quite concerned at the fact that they didn't grow”. Another common theme in relation to the grain specific comments from citizen scientists was a change in the appearance of the kefir grains with five water kefir participants and two milk kefir participants reporting physical deviations in their grains.

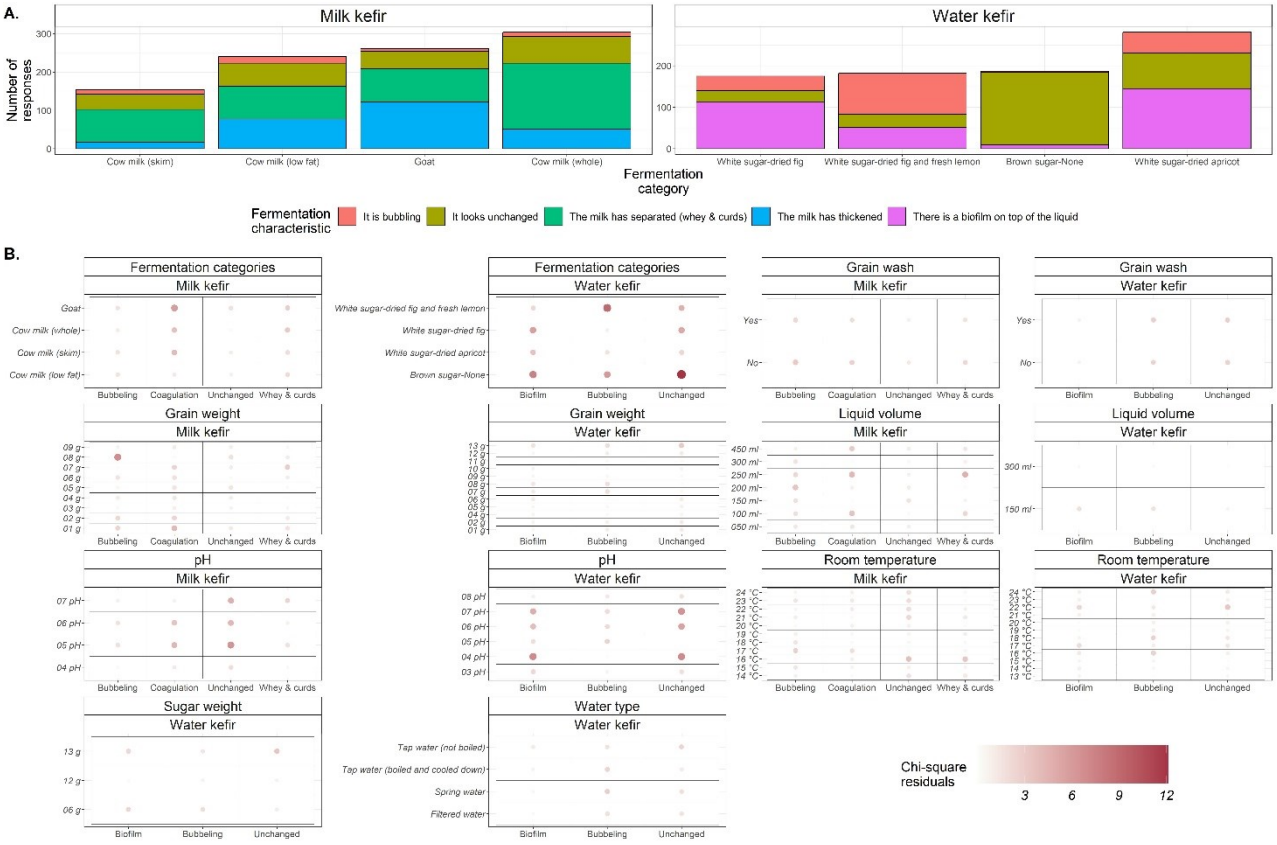


Figure 6: Association of fermentation parameters in the milk and water kefir fermentations and the occurrence of fermentation characteristics, variables include It is bubbling, It looks unchanged, The milk has separated (whey & curds), The milk has thickened and There is a biofilm on top of the liquid. **(A.)** Breakdown of the number of 'Fermentation categories' (X-axis) according to the fermentation characteristics **(B.)** Bubble plot displaying the Chi-square residuals values between the fermentation parameters in the milk and water kefir fermentations. Sub plots show the associations between 'Fermentation categories', 'Grain Wash', Grain weight' ,'Liquid volume', 'pH' and 'Room temperature' and the fermentation characteristics 'Bubbling', 'Coagulation', 'Unchanged' and 'Whey & curds' in the milk kefir fermentations and associations between the fermentation parameters 'Fermentation categories', 'Grain Wash', Grain weight' ,'Liquid volume', 'pH', 'Room temperature', 'Sugar weight', and Water type' and the fermentation characteristics 'Biofilm', 'Bubbling' and 'Unchanged' in the water kefir fermentations.

4.3.5 The Kefir4All project provided important new research data, and resulted in an enhanced interest and scientific knowledge with respect to fermented foods and microbiology among participants.

As a consequence of the recruitment and engagement initiatives undertaken during the Kefir4All project, Kefir4All collected 516 milk and water kefir grains and liquid samples from 72 participants (Figure 3D, 7C). These 516 samples were further divided into 223 milk kefir samples from 29 participants, and 293 water kefir samples from 43 water kefir participants. Thanks to the contributions of citizen scientists, Kefir4All provided a unique opportunity to study natural changes in the milk and water kefir microbiome within household settings and in response to different fermentation parameters, while also facilitating a

comparison of the species level data with that from other kefir studies employing a more traditional method of sample generation and collection (Figure 7). After DNA extraction and high throughput shotgun metagenomic sequencing, the microbiome composition of milk and water kefir samples collected in Kefir4All was determined and compared with that of the 256 milk [17] [18]and 375 water kefir microbial profiles (publication in preparation) (Figure 7B and 7C), produced under controlled laboratory settings. The milk and water kefirs produced under controlled laboratory settings were initiated using 64 distinct milk kefir grains, and 69 distinct water kefir grains, respectively. These grains were originally sourced from >20 different countries, and repeat 24-48 hour fermentations for each kefir grain sourced are represented in Figure 7A (Y-axis). Milk and water kefir profiles generated by the Kefir4All citizen scientists initiated using single source microbial communities mirrored the diversity observed in both milk and water kefir profiles initiated using different grains (Figure 7A and 7B) with the added advantage of associated metadata in the form of 1,823 fermentation surveys collected over 21 weeks (Figure 3C) (see methods). Specifically, hierarchical clustering analysis identified 8 community type clusters across the study groups, with 'Breselge et al. - Water kefir' , 'Kefir4All - Water kefir' , 'Kefir4All - Milk kefir' and 'Walsh et al. - Milk kefir' metagenomes representing the data sources considered in this study (Figure 7B). A value of 0.1 Dunn's index score was calculated for this clustering analysis, Jaccard coefficient values obtained using bootstrapping were used to evaluate cluster stability, values indicated the frequency in which metagenomes of each cluster grouped together. Values ranged from 0.46 to 0.91, values <0.60 were considered unstable, values > 0.85 were considered highly stable. Of the eight clusters identified (Figure 7B), Clusters 1, 2 and 3 were considered unstable. Cluster 1, 3, 4 and 5 contained metagenomes from 'Breselge et al. - Water kefir' and metagenomes from 'Kefir4All - Water kefir'. Cluster 2 contained metagenomes from all data sources but most frequently from water kefir specific data sources, but also contained the lowest Jaccard coefficient of 0.46. Cluster 6, 7 and 8 contained metagenomes from 'Walsh et al. - Milk kefir' and 'Kefir4All - Milk kefir' (Figure 7B). As expected, more representative metagenomes for each cluster were identified in 'Walsh et al. - Milk' and 'Breselge et al. - Water kefir', with notable exceptions such as cluster 4, containing 15 metagenomes from 'Breselge et al. - Water kefir' and 69

metagenomes from ‘Kefir4All - Water kefir’ and cluster 7 containing 23 metagenomes from ‘Walsh et al. - Milk kefir’ and 109 metagenomes from ‘Kefir4All - Milk kefir’.

As noted, the citizen scientists were informed that at the end of the study they would receive a report detailing the microbial composition of the kefirs they produced, including in the microbial profile over the course of the project (Figure 5C, Supplementary Materials: report 1 and 2). Within this report, appropriate reading material was provided, to ensure participants had the necessary scientific knowledge to interpret the data. The appropriate reading material conveyed background information concerning microbiology, fermented foods, the microbial communities contained within milk kefir and water kefir (Figure 5C), and the different measures available for the description of milk and water kefir microbial communities, including taxonomy, relative abundance, alpha and beta diversity. Participants from which no samples were received were provided with a reference report to ensure they were informed about the overall outcomes of the study (Supplementary Report 1 and 2)

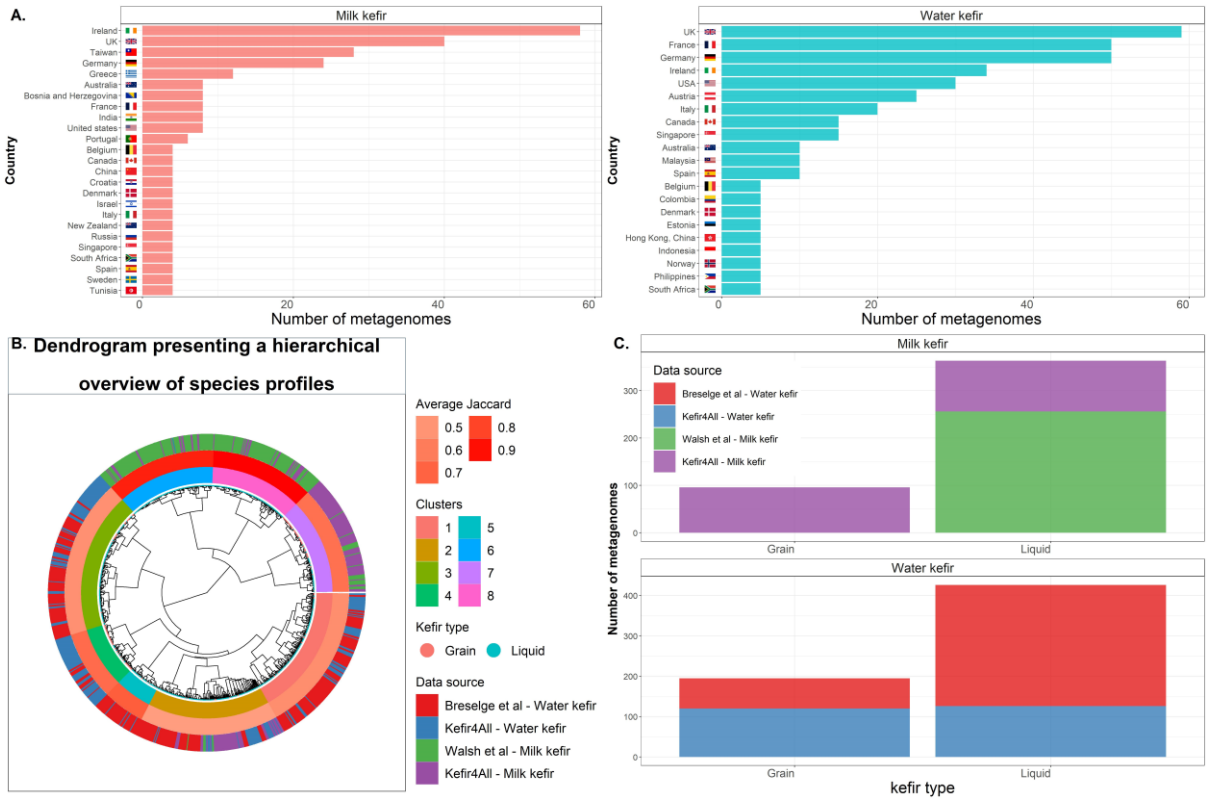


Figure 7: (A.) Number of metagenomes (X-axis) derived from kefir grains sourced from different countries (Y-axis). **(B.)** Dendrogram presenting a hierarchical overview of species profiles detected in the milk kefir and water kefir microbiomes using the metachache pipeline. Tips of the dendrogram are coloured according to the kefir type; Grain or Liquid. The first outer ring displays the results of hierarchical clustering; colour indicates the grouping of each metagenome into one of the eight clusters. Clusters are coloured as follows 1, 2, 3, 4, 5, 6, 7 and 8. The second outer ring represents a circular heatmap displaying the Jaccard coefficients of each of the eight clusters detected using hierarchical clustering. The third outer ring represents the data source the metagenomes were generated in and is coloured as follows; 'Breselge *et al* - Water kefir', 'Kefir4All - Water kefir', 'Walsh *et al* - Milk kefir' and 'Kefir4All - Milk kefir'. **(C.)** Number of grain or liquid metagenomes, separated according to the kefir type - milk kefir or water kefir grain and liquid, and coloured according to the data source.

The 'Evaluation Survey' (Figure 8), completed by both citizen scientists (n=18) and supervisory teachers (n=6) revealed that all responding citizen scientists and teachers felt that their interest in science (8A-D, 8G, 8H, 8O and 8P) and understanding of scientific vocabulary relating to fermented foods or microbiology (Figure 8E and 8F) was increased as a result of taking part in the project. Specifically, 10 (42%) of the responders, including both supervisory teachers and citizen scientists, felt that they or their students had moderate improvements, while seven responders reported a large improvement, in their understanding of scientific vocabulary (Figure 8E and 8F). Additionally, both teachers and citizen scientists felt that their general knowledge of microbiology and/or fermented foods, or those of their students were improved

after taking part in the project. Specifically, none of the 24 responders (0%) reported that they did not benefit in terms of knowledge and interest of microbiology and fermented foods at the end of the project, and 17 responding citizen scientists (94%) and four teachers (67%) reported improvements in knowledge and interest in both microbiology and fermented foods (Figure 8C and 8D). Lastly, 16 of the 18 responding citizen scientists (89%) and all six teachers (100%) expressing an interest in participating in future citizen science studies (Figure 8O and 8P).

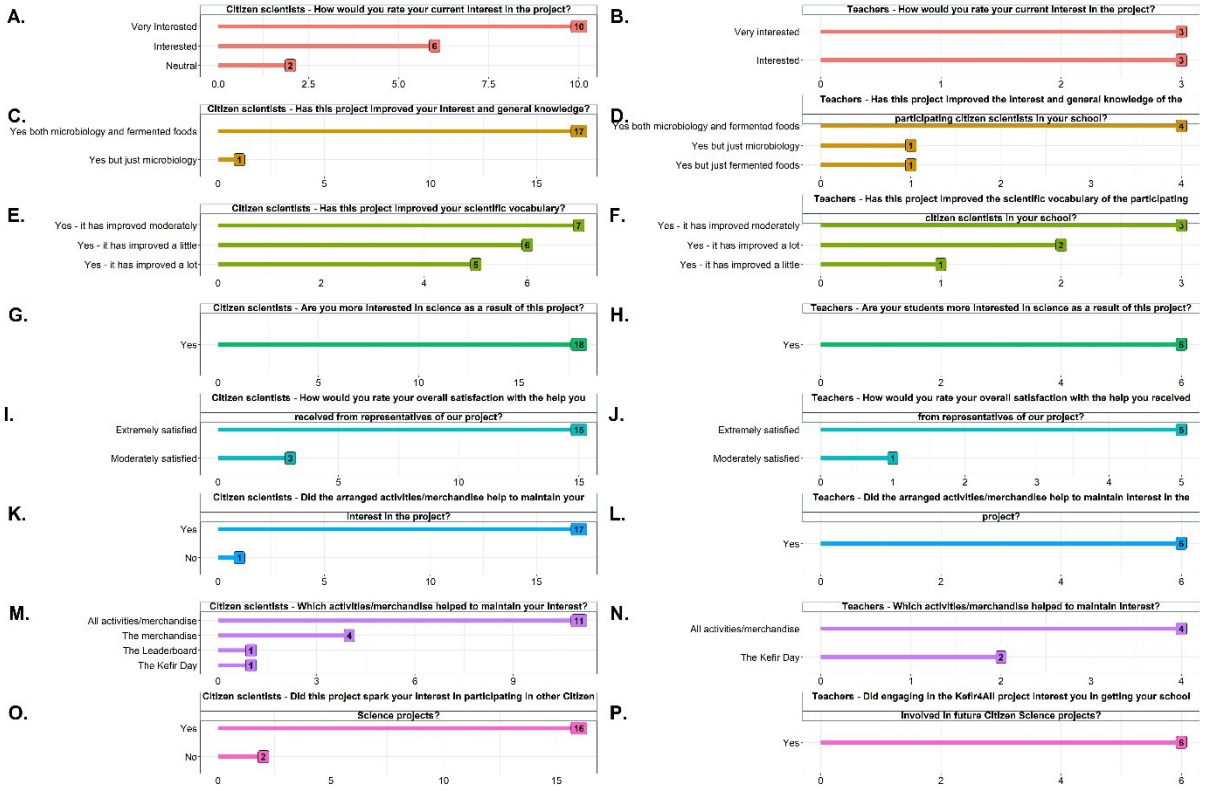


Figure 8: Survey answers from the citizen scientist (n=18) and the supervisory teachers (n=6) who coordinated the students in the school environment. **(A.)** 'Citizen scientists - How would you rate your current interest in the project?'; **(B.)** 'Teachers - How would you rate your current interest in the project?'; **(C.)**. 'Citizen scientists - Has this project improved your interest and general knowledge of microbiology and fermented foods?'; **(D.)** 'Teachers - Has this project improved your interest and general knowledge of microbiology and fermented foods?'; **(E.)** ' Citizen scientists - Has this project improved your scientific vocabulary?'; **(F.)** ' Teachers - Has this project improved the scientific vocabulary of participating citizen scientists in your school'; **(G).** Number of yes/no survey answers to the question entitled, 'Citizen scientists - Are you more interested in science as a result of this project?'; **(H.)** Number of yes/no survey answers to the question entitled, 'Teachers - Are your students more interested in science as a result of this project?'; **(I.)** Number of survey answers to the question entitled, 'Citizen scientists - How would you rate your overall satisfaction with the help you received from representatives of our project?'; **(J.)** Number of survey answers to the question entitled, 'Teachers - How would you rate your overall satisfaction with the help you received from representatives of our project?'; **(K.)** Number of yes/no survey answers to the question entitled, 'Citizen scientists -

Did the arranged activities/merchandise' help to maintain your interest in the project?'; **(L.)** Number of yes/no survey answers to the question entitled, 'Teachers - Did the arranged activities/merchandise' help to maintain your interest in the project?'; **(M.)** ' Citizen scientists - Which activities/merchandise helped to maintain your interest?'; **(N.)** 'Teachers - Which activities/merchandise helped to maintain your interest?'; **(O.)** 'Citizen scientists - Did this project spark your interest in participating in other Citizen Science projects?'; **(P.)**'Teachers - Did engaging in the Kefir4All project interest you in getting your school involved in future Citizen Science projects?'.

When analysing the individualised responses to the questions 'Why did you proceed with or stop the project?', the most frequently occurring word was 'kefir', which was referenced in six of the 18 responses (33%), and further referred to as 'it' or 'fermented food' in four additional comments (22%) (Figure S2). Full comments included 'I was interested in observing changes in the kefir from week to week', 'I was very curious to see how the microbes in my kefir changed over the course of the project' and 'I enjoyed doing it and it was fun to see what would happen'. 'Interested' and 'interesting' were the next most frequent words, which were referenced in five (28%) and three (16%) of the 18 responses, respectively, and concerned the areas of microbiology and food fermentation (Figure S2). We observed identical trends when analysing the individualised comments to the question 'What did you like about the project?', 'kefir' was again the most frequent word accounting for eight of the total 24 responses (33%), followed closely by liked (n=7, 29%), interesting (n=5, 21%), different (n=5, 21%) and seeing (n=4) (Figure S3 and S4). Full comments included 'I learned more about kefir and good conditions for growing it', 'I liked seeing different results after each fermenting' and 'That students could complete it at home and they were expected to be responsible'. Engagement activities helped to maintain interest in the responding participants, with just one participant recording that the engagement activities did not maintain their interest. The majority of responders, including both teachers (n=4, 67%) and citizen scientists (n=11, 61%), felt that all of the engagement activities combined assisted in maintaining interest (Figures 7K and 7L), while others, including both teachers and citizen scientists, highlighted the importance of the Kefir Day (n=3, 12%) or the provision of Kefir4All merchandise (n=3, 12%) (Figures 7M and 7N). 'Nothing' was the most frequent word included in the personalised responses to the question 'What did you dislike about the project?' (Figure S5 and S6), accounting for seven of the total 24 responses (29%), with no other general trend of interest occurring in terms of word frequency (Figure S7-S10).

Section 4.4: Discussion

The benefits of citizen science to education systems and wider society are being actively investigated by international citizen science associations [30]. However, previous citizen science projects applied in school settings have already established that a high level of participation is found to have greater impacts on learning outcomes [31]. In Kefir4All, we conducted a citizen science project involving 123 citizen scientists in which scientific learnings and practices were applied in formal and informal learning environments [32]. The project provided science education and learning in formal environments, by providing workshops facilitating the integration of the project into the scientific curriculum of the schools involved, and in informal environments, by supporting the citizen scientist to carry out repeat fermentations in private households for up to 21 weeks (Figure 1).

In Kefir4All, we circulated the 'Workshop Survey' at the end of the introductory workshop to determine the public's background knowledge and perceptions of, as well as interest in microbiology and fermented foods (Figure S1). The survey results revealed a general lack of knowledge (Figures S1C, S1E and Figure 2E and 2F) relating to both topics. Often, individuals who expressed a neutral interest in the project provided personalised comments to the question 'What did you like and dislike about Microbiology?', such as 'I'm curious about it and feel like this project will be great to learn more, I just don't know enough about it at the moment' and 'I guess its not really talked about so I don't have an opinion on it, but it'd be nice to learn something new'. The 'Workshop Survey' also revealed the approaches that were most commonly used by members of the public to access scientific information, with 'YouTube' and 'School' being considerably more popular than 'News channels' and 'Books' (Figure S1I). Interestingly, from the personalised comments provided in relation to the workshop, one responder noted, 'Well for one it was engaging and it sounded just like the YouTube videos I watch so it was fun and enjoyable'. These combined results, including the personalised comments of citizen scientists suggest that an absence of knowledge is one of the primary limiting factors impacting interest in scientific disciplines, and highlights the importance of YouTube and school settings as key sources of scientific information' (Figures S1I). The 'Evaluation Survey' was then used to determine self-reported improvements in interest and general

knowledge concerning microbiology and fermentation, among the participating citizen scientists as a result of taking part in the project (Figure 8) [33]. Here we show that taking part in the project and, to a lesser extent, attending a workshop can greatly influence public perception, interest and general scientific knowledge (Figure S1J,Figure 2C, Figures 8A-8H, 8O and 8P). This pattern was consistent with the positive feedback received in other citizen science projects such as Symbiosis in the soil [33]. Indeed, we received numerous comments from citizen scientists such as 'The information that the microbiologists had talked to us about has created an interest in the subject for me'. Furthermore, many citizen scientists and teachers reported benefits in terms of improved scientific knowledge, vocabulary and interest as well as a desire to be part of further citizen science projects (Figures 8A-8H, 8O and 8P), even among those with relatively less interest in the topic.

Unfortunately, perhaps as a result of recruiting on the basis of good faith alone, a number of the citizen scientists that were originally recruited did not contribute significantly to the project, with 35 'Non-fermenters' and 34 'Rare fermenters' (Figure 3A), representing 56% of the total participants. It is worth noting that evaluation surveys were only completed by citizen scientists who completed the study and regularly engaged with the project, and thus it was not possible to get further insights into the low rate of participation by some individuals. However, from the personalised comments about the project, one teacher noted, 'I'd like to pitch it to only certain students as many of ours didn't engage well unfortunately'. Such comments, when taken together with the low rate of participation by some individuals, highlight the potential benefits of working within research networks and/or having a community structure with successful citizen scientists and their schools/groups to ensure recruitment of citizen scientists that are likely to complete the project, and is particularly applicable as many participants and teachers in Kefir4All expressed an interest in partaking in future citizen science studies (Figures 8O and 8P). A number of such research networks are currently available, including the US-based Citizen Science Association (CSA), the European Citizen Science Association, and the Australian Citizen Science Association [8, 34], which may present a better alternative for participant recruitment compared to the good faith approach applied in Kefir4All [4]. One potential mitigating factor, in terms of recruiting in good

faith for future projects, may be to recruit smaller groups, as recruitment as part of larger groups positively correlated with a greater frequency of 'Non-fermenters' and 'Rare fermenters'. From the personalised comments about the project, one teacher noted when asked about what they disliked about the project, 'Trying to keep students motivated - should not have selected so many'. Smaller groups may facilitate direct participation and the requirement for less supervision time for the teacher involved, while still ensuring the development of new connections between participants and citizen science. Alternatively, a practical learning activity could be introduced during the workshop to assess the citizen scientist's interest for the project, indeed such an activity was one of the areas of potential improvement proposed in the "Workshop survey". A learning activity could include dividing the workshop attendees into groups, who would then be asked to produce different fermented foods, explain their food, and discuss the similarities and differences between the different fermented foods. [35]

We noted a high degree of initial participation, which is a likely reflection of the initial enthusiasm of participants in taking part in a citizen science project following interactions with the Kefir4All team (Figure 3C and 3D, Figure 8A and 8B). However, after wk5 the participants began to show signs of diminished interest (Figure 3C). Indeed one of the main points of dislike raised by some participants during the "Evaluation Survey" was the project length and level of participation required. We therefore recommend that, where possible, citizen science projects attempt to design their project within a 6-week range to ensure maximum interest, or the incorporation of engagement activities to ensure a longer lasting project. When asked what citizen scientists liked about the project, one of the main points related to the satisfaction derived from contributing to scientific research. This was reflected in comments such as 'Knowing my involvement was being used for research - It was interesting', 'Opportunity to show family members some scientific research' and 'being part of a large citizen science project - a first for me'. Another point of success was the requirement to make observations and to use fermentation equipment. During the study many citizen scientists reported an appeal in seeing different fermentation characteristics over time, with comments such as 'The equipment was all supplied and it was great standard and new! I liked seeing how the kefir was different each time', 'I enjoyed observing any changes.

I liked using the equipment we received such as the thermometer and pH strips' and 'I liked the repetitiveness of doing the fermentations as the time it took up was little and often so it wasn't stressful. I also liked gathering the information each time and seeing if it was different or the same as usual'. Indeed in their closing comments, many participants noted that they had tried and/or had read about a number of different approaches to improve grain growth such as "I left the milk longer to try for more grains" and "I read up on some ways to have a more stable environment for kefir growth (like adding some lemon drops)".

The incorporation of engagement activities in the project contributed to continued participation, while also providing further opportunities for further education of the participants through career talks and data interpretation skills (through the graphical displays presented in the weekly leader board) (Figure 4), and individualised citizen science reports (Figure 5). A number of responders highlighted that prior to the expanded array of engagement activities, the project had begun to become repetitive by virtue of the fermentation process having become routine. One teacher questioned at wk5 'Is there any further information from the project that we can show the students involved to keep them motivated?'. While the outreach activities helped to maintain continuous participation in the project, they resulted in minimal changes to the number of fermentations recorded and samples provided (Figure 3C and 3D). Some teachers communicated that the sharing of results generated from their personalised kefir metagenomic datasets during the project would have had a very positive impact with respect to ensuring continual participation, and a citizen scientist noted that the project would have been improved if 'Results were provided during the project, to help keep me motivated'. Therefore, future citizen science projects of this type should design the project structure to enable the sharing of data (Figure 5 and Supplementary Report 1 and 2) during the earlier stages of the project and not at the end, as was done in Kefir4All (as time was needed for DNA extraction, sequencing and data analysis). If data sharing is not possible within the timeframe of the project, derivatives of the 'Which Variables Matter?' lesson module developed and utilised in the Wild Sourdough project [35] could be explored. The Wild Sourdough project aimed to study patterns of change within sourdough by asking citizen scientists to produce sourdough under varying

conditions and the lesson module represents a graphing exercise where citizen scientists can explore which conditions influence their sourdough. Recommendations discussed in this manuscript that are derived from the challenges encountered, and/or from feedback from participants, can help shape and improve guidelines for future projects and highlight the importance of communication between researchers and participants.

One clear advantage of citizen science from a researchers perspective is the capacity for such a project to produce an extensive database, containing information such as observations, measurements, samples and/or assessments collected or produced by citizen scientists [36]. A number of previous citizen science projects have reported conflicting results in terms of sample quality [37] with, in some instances, concerns over the reliability and quality of data [10] while, in other cases, in the results provided by the citizen scientists were comparable with those provided by professionals [38][39, 40]. Here, while reliability differed between participants, a comparison, using hierarchical clustering, between the newly generated data and that from previous kefir studies performed by professional researchers highlighted a similar level of diversity between data sources at the compositional level (Figure 7B). Furthermore, the data sources in Kefir4All had clear advantages in terms of associated metadata and a time series structure, which would have been difficult to reproduce at the same scale through a traditional laboratory-based approach. [39, 40]This citizen science project, with its time series design, enabled the unique opportunity to assess the milk kefir and water kefir microbiome in terms of long-term stability, potential strain level evolution, and adaption to the different environments (different households and different fermentation practices), including resilience of the grains to the introduction of environmental microbes. Another benefit to the researchers is the sense of having contributing to education and public education and receiving feedback such as 'Thank you for having me as a citizen scientist, it was so fun - definitely solidifying the area of chemistry / biochemistry for my future work area'.

Overall, our results further highlight the value of schools as hubs for citizen science projects. From a researcher's perspective, schools can act as central hubs in which to train, re-visit, collect samples and communicate with citizen scientists, reducing logistical concerns. As outlined by Roche *et al*., some

citizen science projects do not incorporate suitable flexibility in their experimental design to enable a smooth integration in schools curricula [32]. Here, we show that teachers involved in the project can often incorporate a leadership role maintain continuous participation, and can provide guidance and/or play an active role in shaping the project structure to suit their school and their students [32]. However further developments are still needed for schools to reach their optimal potential in citizen science. Possible further developments that should be considered include grants and or other funding opportunities to establish and maintain a citizen science infrastructure within a school system such as a citizen science club, which will reduce the reliance of recruiting in 'good faith' and promote meaningful participation of schools in scientific research. Ultimately citizen science projects in a school setting, provide an opportunity to improve scientific research and extend the current scientific curriculum, and represent a unique opportunity to communicate scientific principles while sharing and discussing personalised results, providing a participant with an individualised practical insight into how science is conducted in the real world [41].

Section 4.5: References

1. Heigl, F., et al., *Toward an international definition of citizen science.* Proceedings of the National Academy of Sciences, 2019. **116**(17): p. 8089-8092.
2. National Academies of Sciences, E. and Medicine, *Learning through citizen science: Enhancing opportunities by design.* 2018.
3. Yosef, R. and P. Tryjanowski, *Editorial: Citizen Science for Future Generations.* Frontiers in Ecology and Evolution, 2022. **10**.
4. Randler, C., *Leaders Inspiring the Next Generation of Citizen Scientists – An Analysis of the Predictors of Leadership in Birding.* Frontiers in Ecology and Evolution, 2021. **9**.
5. Smith, C.A. and P.J. Wicks, *PatientsLikeMe: Consumer health vocabulary as a folksonomy.* AMIA Annu Symp Proc, 2008. **2008**: p. 682-6.
6. Wuyts, S., et al., *Carrot Juice Fermentations as Man-Made Microbial Ecosystems Dominated by Lactic Acid Bacteria.* Appl Environ Microbiol, 2018. **84**(12).
7. Landis, E.A., et al., *The diversity and function of sourdough starter microbiomes.* eLife, 2021. **10**: p. e61644.
8. Storksdieck, M., et al., *Associations for citizen science: regional knowledge, global collaboration.* Citizen Science: Theory and Practice, 2016. **1**(2).
9. Robinson, L.D., et al., *Ten principles of citizen science*, in *Citizen science: Innovation in open science, society and policy*. 2018, UCL Press. p. 27-40.
10. Lewandowski, E., et al., *Public perceptions of citizen science.* Citizen Science: Theory and Practice, 2017. **2**(1).
11. Lavelle, C. and J.-B. Boulé, *Fermentation: A Short Scientific and Culinary Overview of Kefir.* Handbook of Molecular Gastronomy, 2021.

12. Bourrie, B.C., B.P. Willing, and P.D. Cotter, *The Microbiota and Health Promoting Characteristics of the Fermented Beverage Kefir.* Front Microbiol, 2016. **7**: p. 647.
13. Farag, M.A., et al., *The many faces of kefir fermented dairy products: Quality characteristics, flavour chemistry, nutritional value, health benefits, and safety.* Nutrients, 2020. **12**(2): p. 346.
14. Tu, C., et al., *Quality and metagenomic evaluation of a novel functional beverage produced from soy whey using water kefir grains.* Lwt, 2019. **113**: p. 108258.
15. Asingizwe, D., et al., *Why (not) participate in citizen science? Motivational factors and barriers to participate in a citizen science program for malaria control in Rwanda.* PLoS One, 2020. **15**(8): p. e0237396.
16. SurveyMonkey, *SurveyMonkey: the world's most popular free online survey tool*. 2021.
17. Walsh, L.H., et al., *Analysis of the milk kefir pan-metagenome reveals four community types, core species, and associated metabolic pathways.* iScience, 2023. **26**(10): p. 108004.
18. Walsh, L.H.C., M. Walsh, A. M. Crispie, F. and P.W.C. O'Toole, P. D. , *Analysis of the milk kefir pan-metagenome reveals four community types, core species and associated metabolic pathways*. Submitted- 2022.
19. Andrews, S., et al., *Trim Galore.* Trim Galore, 2015.
20. Langmead, B. and S.L. Salzberg, *Fast gapped-read alignment with Bowtie 2.* Nature methods, 2012. **9**(4): p. 357-359.
21. Müller, A., et al., *MetaCache: context-aware classification of metagenomic reads using minhashing.* Bioinformatics, 2017. **33**(23): p. 3740-3748.
22. R Core Team, *R: A language and environment for statistical computing.* 2020.
23. Kassambara, A., *ggpubr:"ggplot2" based publication ready plots (R package version 0.4. 0)[Computer software]*. 2020, ggpubr:'ggplot2'based publication ready plots (R Package Version 0.4. 0 ….
24. Oksanen, J., et al., *vegan: Community Ecology Package*. 2020.
25. Kassambara, A. and F. Mundt, *Package 'factoextra'.* Extract and visualize the results of multivariate data analyses, 2017. **76**(2).
26. Davis, R.H. and C.E. Economou, *A review of fuzzy clustering methods.* Advances in Engineering Software (1978), 1984. **6**(4): p. 189-191.
27. Hennig, C. and M. Imports, *Package 'fpc'.* Flexible Procedures for Clustering, 2015.
28. Wickham, H. and M.H. Wickham, *The ggplot package.* Google Scholar, 2007.
29. Le Pennec, E. and K. Slowikowski, *ggwordcloud: A Word Cloud Geom for'ggplot2'.* R package version 0.5. 0, 2019.
30. Roche, J., et al., *Citizen Science in Ireland.* Frontiers in Communication, 2021. **6**.
31. Lüsse, M., et al., *Citizen science and its potential for science education.* International Journal of Science Education, 2022. **44**(7): p. 1120-1142.
32. Roche, J., et al., *Citizen Science, Education, and Learning: Challenges and Opportunities.* Frontiers in Sociology, 2020. **5**.
33. McKenney, E., et al., *Symbiosis in the Soil: Citizen Microbiology in Middle and High School Classrooms.* J Microbiol Biol Educ, 2016. **17**(1): p. 60-2.
34. Roche, J. and N. Davis, *Should the science communication community play a role in political activism?* Journal of Science Communication, 2017. **16**(1): p. L01.
35. McKenney, E. and L. Nichols, *Wild Sourdough.* Fermentology, 2021.
36. Poisson, A.C., et al., *Quantifying the contribution of citizen science to broad-scale ecological databases.* Frontiers in Ecology and the Environment, 2020. **18**(1): p. 19-26.
37. Balázs, B., et al., *Data Quality in Citizen Science*, in *The Science of Citizen Science*, K. Vohland, et al., Editors. 2021, Springer International Publishing: Cham. p. 139-157.
38. D'Alessio, M., G. Rushing, and T.L. Gray, *Monitoring water quality through citizen science while teaching STEM undergraduate courses during a global pandemic.* Sci Total Environ, 2021. **779**: p. 146547.
39. Danielsen, F., et al., *A Multicountry Assessment of Tropical Resource Monitoring by Local Communities.* BioScience, 2014. **64**(3): p. 236-251.

40. Delaney, D.G., et al., *Marine invasive species: validation of citizen science and implications for national monitoring networks.* Biological Invasions, 2008. **10**(1): p. 117-128.
41. Armstrong, B., et al., *How Does Citizen Science Compare to Online Survey Panels? A Comparison of Food Knowledge and Perceptions Between the Zooniverse, Prolific and Qualtrics UK Panels.* Frontiers in Sustainable Food Systems, 2021. **4**.

Section 4.6: Supplementary information

4.6.1 Supplementary method

Method S1: Instructions for milk kefir fermentations

Step 1: Open Survey Monkey by opening the link below and fill in your ID and the type of kefir you are fermenting.

https://www.surveymonkey.com/r/KefirFermentation

Step 2: Study your fermentation jar, does the liquid look like any of the below pictures? Record your room temperature with the thermometer provided. Report your observations to Survey Monkey. Ignore this step if it is your first fermentation

Step 3: Fill out questions 1-3 on Survey Monkey.



Its Bubbling



It looks unchanged



The milk and whey have separated

Step 4: Unscrew and remove the top lid and strainer from the fermentation vessel.

Step 5: Dip the pH test strip for two seconds in the liquid and wait 10 seconds, report the pH of the liquid to SurveyMonkey.

Step 6: Reapply the strainer lid to the fermentation vessel and attach the scraper tool provided on the strainer lid, by inserting the scraper into the hole in the middle of the strainer lid, apply light pressure to the scraper to get it to snap into place.

Step 7 (optional): Remove the strainer lid and gently stir the fermented liquid using the stirrer provided, this will help with the straining process.

Step 8: Get a bowl/container to collect the fermentation liquid. Use a 15 ml and 50 ml collection tube on the days for sample collection.

Step 9: Tilt the fermentation vessel in the direction of the collection bowl/container to begin the straining process. Turn the scraper in half-circle motions while straining. After straining, the grains will remain in the jar and the liquid will be transferred to the bowl.

Step 10: Wash the kefir grains using either milk or water, pour the washing liquid not the kefir grains down the sink after washing (optional). Report if the grains were washed or not, on SurveyMonkey.

Step 11: Remove the kefir grains from the vessel and weigh them, report the weight on SurveyMonkey.

Step 12: Wash the fermentation equipment with hot soapy water; this must be performed at least once a week.

Step 13: To do the milk kefir fermentation, fill the vessel to the first line (300 ml) with milk and place up to 6 grams of grains into the jar. If you have 6 grams of grains available, please use them. Attach the strainer lid back to the fermentation vessel followed by the top lid. Report the quantity of milk on SurveyMonkey.

Step 14: Record the temperature of the room using the thermometer provided and report the temperature to SurveyMonkey.

Step 15: Screw the top lid firmly into place if you have been chosen to carry out anaerobic fermentation, and lightly/loosely if you have been chosen to carry out aerobic fermentations.

Step 16 (optional): Place the fermentation clock in the middle of the top lid. The fermentation clock will let you keep track of the time and date the fermentation started.

Step 17: Complete all remaining questions on SurveyMonkey.

Step 18: Leave your milk kefir fermentation for no more than 72 hours and then repeat all required steps outlined in this document.

Thank you!

Survey for each fermentation https://www.surveymonkey.com/r/KefirFermentation

Workshop Survey https://www.surveymonkey.com/r/WM6JF7N

Survey for the start of the project https://www.surveymonkey.com/r/KefirLocation

Note: the survey for the start of the project needs to be filled out only once or again if the household location for the fermentation is changed)

Sample Collection Dates:

Sample Number	Date	Week
1	Thursday 30 Dec	1
2	Thursday 27 Jan	5
3	Thursday 24 Feb	9
4	Thursday 24 Mar	13
5	Thursday 21 Apr	17
6	Thursday 19 May	21

Sample Collection:

1. Start a new fermentation 24h before you want to collect samples
2. Gently mix your kefir
3. Collect 15 ml liquid in 15 ml tube
4. Collect 50 ml liquid in 50 ml tube
5. Discard remaining liquid & rinse the grains
6. Collect about "5 ml" of grains in 15ml tube
7. Freeze all samples in a freezer (-20°C)
8. Bring frozen samples to school (use ice packs to keep samples frozen)
9. Keep samples frozen until collection by us

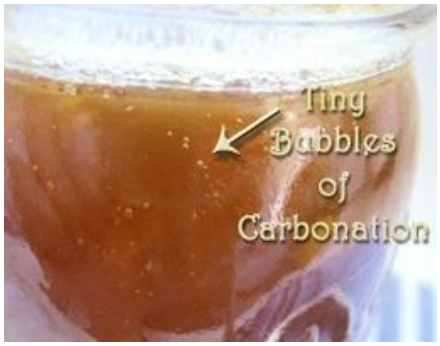
Methods S2: Instructions for water kefir fermentations

The day before: boil water and let it cool down to room temperature (this step can be skipped if you use bottled or purified water)

Step 1: Open Survey Monkey by opening the link below and fill in your ID and type of kefir you are fermenting.

https://www.surveymonkey.com/r/KefirFermentation

Step 2: Study your fermentation jar, does the liquid look like any of the below pictures? Record room temperature with the thermometer provided. Report your observations to Survey Monkey. Ignore this step if it is your first fermentation



Its Bubbling



It looks unchanged



There is a biofilm on top of the liquid

Step 3: Unscrew and remove the top lid and strainer from the fermentation vessel

Step 4: Dip the pH test strip for one second in the liquid and wait 10 seconds, report the pH of the liquid to SurveyMonkey

Step 5: Reapply the strainer lid to the fermentation vessel and attach the scraper tool provided on the strainer lid, by inserting the scraper into the hole in the middle of the strainer lid, apply light pressure to the scraper to get it to snap into place.

Step 6 (optional): Remove the strainer lid and gently stir the fermented liquid using the stirrer provided, this will help with the straining process.

Step 7: Get a bowl/container to collect the fermentation liquid. Use a 15 ml and a 50 ml collection tube on the days for sample collection.

Step 8: Pour off the liquid. After pouring off the liquid, the grains will remain in the jar.

Step 10: Wash the kefir grains with water, pour the water, not the kefir grains, down the sink after washing (optional). Report if the grains were washed or not, on SurveyMonkey

Step 11: Remove the kefir grains from the vessel and weigh them, report the weight on SurveyMonkey

Step 12: Wash the fermentation equipment with hot, soapy water; this should be performed at least once a week.

Step 13: To do a new kefir fermentation, fill the vessel to the first line (300ml) with water (boiled and cooled down tap water, filtered water, or bottled water), add one of the following:

a) 12 gram of white sugar and 1 dried fig
b) 12 gram of white sugar, 1 dried fig, 1 slice of fresh lemon
c) 12 gram of white sugar and 1 dried apricot
d) 12 gram of brown sugar

Please stick with your assigned category (a-d).

Mix well until sugar is dissolved. Add 12 gram of the water kefir grains into the jar, attach the strainer lid back to the fermentation vessel, followed by the top lid. Report the quantity of water, sugar type, and fruit on SurveyMonkey. Please keep the ratios the same.

Step 14: Screw the top lid loosly back on, as you want to carry out an aerobic fermentations.

Step 15 (optional): Place the fermentation clock in the middle of the top lid. The fermentation clock will let you keep track of the time and date the fermentation started.

Step 16: Complete all remaining questions on SurveyMonkey

Step 17: Leave your water kefir fermentation for no more than 72 hours and then repeat all required steps outlined in this document

---

Thank you!

Survey for each fermentation
https://www.surveymonkey.com/r/KefirFermentation

Workshop Survey
https://www.surveymonkey.com/r/WM6JF7N

Survey for the start of the project
https://www.surveymonkey.com/r/KefirLocation

Note: the survey for the start of the project needs to be filled out only once or again if the household location for the fermentation is changed)

Sample Collection Dates:

Sample Number	Date	Week
1	Friday 14 Jan	1
2	Friday 11 Feb	5
3	Friday 11 Mar	9
4	Friday 8 Apr	13
5	Friday 6 May	17
6	Friday 3 Jun	21

Sample Collection:

10. Start a new fermentation 24h before you want to collect samples
11. Gently mix your kefir
12. Collect 15 ml liquid in 15 ml tube (dark blue cap)
13. Collect 50 ml liquid in 50 ml tube (dark blue cap)
14. Discard remaining liquid & gently rinse the grains with cold water
15. Collect about "3 ml" of grains in 15ml tube (light blue cap)
16. Freeze all samples in a freezer (-20°C)
17. Bring frozen samples to school (use ice packs to keep samples frozen)

4.6.1.3 Methods S3: Survey entitled 'Workshop Survey'

1. Are you participating in the Citizen Science project?

- [ ] Yes
- [ ] No

2. What is your current interest in Microbiology?

- [ ] Interested
- [ ] Neutral
- [ ] Disinterested
- [ ] Very disinterested
- [ ] I don't know what it means

3. Please tell us why you selected the above, for example what you like and dislike about Microbiology.

answer

4. Please rate your background knowledge about microbiology.

- [ ] Very knowledgeable
- [ ] Knowledgeable
- [ ] Neutral
- [ ] A little knowledge
- [ ] No knowledge what so ever

5. Where do you currently get information about science?

- [ ] Youtube
- [ ] Facebook
- [ ] Instagram
- [ ] Websites
- [ ] New channels
- [ ] Word of mouth
- [ ] School
- [ ] Books (not school books)
- [ ] Other

6. How regularly do you engage with the above materials?

- ◯ Daily
- ◯ Regularly
- ◯ Occasionally
- ◯ Never

7. Please indicate if you are continuing Biology and or Home-Economics into the leaving cert cycle.

- [ ] Yes
- [ ] Yes but just home-economics
- [ ] Yes but just biology
- [ ] No

8. Please rate you interest in fermented foods.

- [ ] Extremely interested
- [ ] interested
- [ ] Somewhat interested
- [ ] Not so interested
- [ ] Not at all interested

9. Please rate your background knowledge about fermentation.

- ◯ No background knowledge
- ◯ A little background knowledge
- ◯ Some background knowledge
- ◯ Good background knowledge
- ◯ Awesome background knowledge

10. Have you ever fermented food before?

- [ ] Yes
- [ ] No

11. If you answered yes, what have you fermented and how long have you been fermenting?

12. If you answered yes to the previous question, did you enjoy the sour taste of the food?

- ◯ Yes
- ◯ No

13. How often do you eat fermented foods?

- ◯ Daily
- ◯ Regularly
- ◯ Occasionally
- ◯ Never

14. If you selected Daily, Regularly or Occasionally - what types of fermented foods do you eat?

15. Were you aware that those foods were fermented prior to the workshop?

- [ ] Yes
- [ ] No

16. Please tell us what you thought of our workshop today.

- ◯ Very interesting
- ◯ Interesting
- ◯ Neutral
- ◯ Boring
- ◯ Very boring

17. Please tell us why you selected the above answer and how we can improve.

18. Has this talk improved your interest and general knowledge of microbiology and fermented foods?

- ◯ No
- ◯ Yes but just microbiology
- ◯ Yes but just fermented foods
- ◯ Yes both microbiology and fermented foods

19. Can think of a good name for our Citizen science project?

4.6.1.4 Methods S4: Survey entitled ‘Fermentation Survey

This survey will collect some information on how your current fermentation is going and afterwards the survey will guide you through setting up a new fermentation.

Please fill out this survey each time you start a new fermentation.

1. Please enter your ID

2. What kind of kefir have you gotten from us?

- ◯ Water Kefir
- ◯ Milk Kefir

Milk Kefir Fermentation

3. Before you start a new fermentation, lets have a look at the current fermentation. How is the current fermentation going?

(Please skip this question, if this is your very first fermentation with the grains.)

- [ ] It is bubbling
- [ ] Looks unchanged
- [ ] The milk has separated (whey & curds)
- [ ] The milk has thickened
- [ ] Other (please specify)

4. What is currently the temperature in the room you are fermenting? (in °C)

5. Which pH did you detect?

- ◯ 14
- ◯ 13
- ◯ 12
- ◯ 11
- ◯ 10
- ◯ 9
- ◯ 8
- ◯ 7
- ◯ 6
- ◯ 5
- ◯ 4
- ◯ 3
- ◯ 2
- ◯ 1

6. Cool! Lets start a new fermentation. The next questions will guide you through setting up the next fermentation. Please enter the date of todays fermentation

Date

Date

DD/MM/YYYY

7. Did you wash the grains before you started a new fermentation? (optional step)

- ◯ Yes
- ◯ No

8. What do your grains weigh after the last fermentation? (in g)

PS: we want to know if your grains grew.
(Please skip this question, if this is your very first fermentation with the grains.)

9. Which milk volume did you use for your fermentation (in ml)? The first fill line on the jar represents 300 ml.

10. What kind of milk have you used?

- [ ] Cow milk (whole)
- [ ] Cow milk (low fat)
- [ ] Cow milk (skim)
- [ ] Goat

11. What is the weight (in g) of the grains you have added to your milk?
Should be 6g per 300 ml of milk - you want to always start with the same amount if possible.

12. What kind of fermentation did you choose?

Aerobic fermentation: lid slightly opened - allowing gas exchange
Anaerobic fermentation: lid tightly closed - no gas exchange

Please choose either option and continue with it throughout the project.

- [ ] Aerobic fermentation
- [ ] Anaerobic fermentation

13. Would you like to add some other comments? Todays survey will finish after this question.
Please enter "no" if you don't have any other comments.

Water Kefir Fermentation

14. Before you start a new fermentation, lets have a look at the current fermentation. How is the current fermentation going?

(Please skip this question, if this is your very first fermentation with the grains.)

- [ ] It is bubbeling
- [ ] It looks unchanged
- [ ] There is a biofilm on top of the liquid
- [ ] Other (please specify)

15. What is currently the temperature in the room you are fermenting? (in °C)

16. Which pH did you detect?

- ◯ 14
- ◯ 13
- ◯ 12
- ◯ 11
- ◯ 10
- ◯ 9
- ◯ 8
- ◯ 7
- ◯ 6
- ◯ 5
- ◯ 4
- ◯ 3
- ◯ 2
- ◯ 1

17. Cool! Lets start a new fermentation. The next questions will guide you through setting up the next fermentation. Please enter the date of todays fermentation

Date

Date

DD/MM/YYYY

18. Did you wash the grains before you started a new fermentation? (optional step)

- ◯ Yes
- ◯ No

19. What do your grains weigh after the last fermentation? (in g)

PS: we want to know if your grains grew.
(Please skip this question, if this is your very first fermentation with the grains.)

20. What kind of water are you using for your fermentation

- ◯ Tap water (boiled and cooled down)
- ◯ Tap water (not boiled)
- ◯ Filtered water
- ◯ Spring water

21. Which water volume did you use for your fermentation (in ml)? The first fill line on the jar represents 300 ml.

22. What kind of sugar have you used?

- ☐ White sugar
- ☐ Brown sugar

23. How much sugar did you add? (in g)

24. Did you add some kind of fruit to your water kefir medium?

- ◯ Yes, dried fig
- ◯ Yes, dried fig and fresh lemon
- ◯ Yes, dried apricot
- ◯ None

25. What is the weight of the grains you have added to your water kefir medium? (in g)
Should be 12g for 300 ml of media - you want to always start with the same ratio if possible.

26. What kind of fermentation did you choose?

Aerobic fermentation: lid slightly opened - allowing gas exchange
Anaerobic fermentation: lid tightly closed - no gas exchange

Please choose either option and continue with it throughout the project.

- ◯ Aerobic fermentation
- ◯ Anaerobic fermentation

27. Would you like to add some other comments? Todays survey will finish after this question. Please enter "no" if you don't have any other comments.

4.6.1.5 Methods S5: Survey entitled 'Project Completion Survey - Citizen Scientists'

1. What was your unique project ID?
You can skip this question, if you want

2. Did you complete the Kefir4All project?

- Yes
- No

Why did you proceed with or stop the project?

3. How would you rate your current interest in the project?

- Very Interested
- Interested
- Neutral
- Disinterested
- Very disinterested

Please tell us why you selected the above.

4. How would you rate your participation in the Citizen Science project?

- Regularly fermented (Completed a fermentation every 1-3 days during the project)
- Occasionally fermented (Completed 10-20 fermentations during the project)
- Barely fermented (Competed 1-10 fermentations during the project)
- Never fermented during the project

5. What did you like about the project?

6. What did you dislike about the project?

7. How could we have improved this project from your perspective?

8. Has this project improved your interest and general knowledge of microbiology and fermented foods?

- ◯ Yes but just microbiology
- ◯ Yes but just fermented foods
- ◯ Yes both microbiology and fermented foods
- ◯ No

9. Has this project improved your scientific vocabulary

- ◯ Yes - it has improved a lot
- ◯ Yes - it has improved moderately
- ◯ Yes - it has improved a little
- ◯ No

10. Are you more interested in science as a result of this project?

- ◯ Yes
- ◯ No

11. How would you rate your overall satisfaction with the help you received from representatives of our project?

- ◯ Extremely satisfied
- ◯ Moderately satisfied
- ◯ Slightly satisfied
- ◯ Neither satisfied nor dissatisfied
- ◯ Slightly dissatisfied
- ◯ Moderately dissatisfied
- ◯ Extremely dissatisfied

12. Did the arranged activities/merchandise (Kefir Day, Leaderboard, Career Talks, hoodies, goodie bags and certificate for project completion) help to maintain your interest in the project?

- ◯ Yes
- ◯ No

13. If you answered yes to the above, please select the activities/merchandise that helped to maintain your interest?

- ◯ All activities/merchandise
- ◯ The Kefir Day
- ◯ The Leaderboard
- ◯ The Career Talks
- ◯ The merchandise

14. Do you have any suggestions for other interactions/activities that you feel would have benefited this project?

15. Did this project spark your interest in participating in other Citizen Science projects

- ◯ Yes
- ◯ No

16. Do you have any suggestions for future Citizen Science projects?

Thank you for taking part in the Kefir4All project and for taking the time to complete this survey

4.6.1.6 Methods S6: Survey entitled 'Project Completion Survey - Teachers'

1. What school do you represent?

2. How would you rate your current interest in the project?

- ◯ Very interested
- ◯ Interested
- ◯ Neutral
- ◯ Disinterested
- ◯ Very disinterested

Please tell us why you selected the above.

3. How would you rate your own contribution to the project?

- ◯ A great deal
- ◯ A lot
- ◯ A moderate amount
- ◯ A little
- ◯ None at all

4. What did you like about the project?

5. What did you dislike about the project?

6. How could we have improved this project from your perspective?

7. In your opinion, has this project improved the interest and general knowledge in microbiology and fermented foods of the participating citizen scientists in your school?

- ◯ Yes but just microbiology
- ◯ Yes but just fermented foods
- ◯ Yes both microbiology and fermented foods
- ◯ No

8. Has this project improved the scientific vocabulary of the participating citizen scientists in your school?

- ◯ Yes - it has improved a lot
- ◯ Yes - it has improved moderately
- ◯ Yes - it has improved a little
- ◯ No

9. Are your students more interested in science as a result of this project?

- ◯ Yes
- ◯ No

10. How would you rate your overall satisfaction with the help you received from representatives of our project?

- ◯ Extremely satisfied
- ◯ Moderately satisfied
- ◯ Slightly satisfied
- ◯ Neither satisfied nor dissatisfied
- ◯ Slightly dissatisfied
- ◯ Moderately dissatisfied
- ◯ Extremely dissatisfied

11. Did the arranged activities/merchandise (Kefir Day, Leaderboard, Career Talks, hoodies, goodie bags and certificate for project completion) help to maintain your interest in the project?

- ◯ Yes
- ◯ No

12. If you answered yes to the above, please select the activities/merchandise that helped to maintain interest?

- ◯ All activities/merchandise
- ◯ The Kefir Day
- ◯ The Leaderboard
- ◯ The Career Talks
- ◯ The merchandise

13. Do you have any suggestions for other interactions/activities that you feel would have benefited this project?

14. Did engaging in the Kefir4All project interest you in getting your school involved in future Citizen Science projects?'

- ◯ Yes
- ◯ No

15. Do you have any suggestions for future Citizen Science projects?

Thank you for taking part in the Kefir4All project and for taking the time to complete this survey

4.6.2 Supplementary result

Result S1: Supplementary Report 1 - Milk kefir. Personalised report provided to the citizen scientists ID46 at the end of the study, detailing and graphically displaying the microorganisms in the participant's kefir samples (kefir grain and liquid).

Metagenomic sequencing of kefir samples from ID49

Liam Walsh[1,2,3], Eimear Ferguson[4], Dr Samuel Mortensen[1,3], Dr Mairead Coakley[1], Dr Fiona Crispie[1], Prof Guilherme Martin[1] and Prof Paul Cotter[1,2,3,4]

[1]Food Biosciences Dept, Teagasc Food Research Centre, Moorepark, Fermoy, Co. Cork, Ireland.
[2]School of Microbiology, University College Cork, Ireland.
[3]APC Microbiome Ireland SFI Research Centre, University College Cork, Ireland.
[4]VistaMilk SFI Research Centre, Teagasc, Moorepark, Fermoy, Co. Cork, Ireland.

March 2023

Please contact paul.cotter@teagasc.ie for any follow-up on the report.

Introduction

In this report, we are excited to reveal the microorganisms that lived in your kefirs, and how they changed as a result of your hard efforts over the 5 months of the Kefir4All project. Firstly, we will revise concepts about microbiology and fermented foods, which you may find useful when looking at your results, which are described in Section 2.

Introduction to Microbiology

Aim

The aim of this introduction section, is to reintroduce you to the exciting world of microbes. Here we will learn about bacteria, viruses and fungi, and their importance and existence in all aspects of our lives.

Background

Microbiology is the study of microscopic organisms that are too small to be visible with the naked eye. These microscopic organisms are what we refer to as microorganisms, or microbes for short. Microbes and their activities are important to virtually all processes on Earth. Microbes can be used as a resource in many ways, e.g., making life-saving drugs, biofuels, foods and drinks, and even cleaning up the environment. Microorganisms also matter because they affect every aspect of our lives, they are in us, on us and around us, and can be both beneficial and harmful. There are three main groups of microbes. Viruses are the smallest of the microbes and cannot grow by themselves. They require a 'host' cell in which to live and reproduce. Once inside the host cell, they often rapidly multiply, destroying the cell in the process! Fungi range in size from being microscopic to very large, and include moulds, mushrooms and mildew. Bacteria are single-celled organisms that, under the optimal conditions, produce offspring, as fast as every 10 minutes. Bacteria can be divided into groups by their shape, e.g., circular (cocci), rod-shaped, spiral, or corkscrew. In cases of illness, doctors find these shapes useful to help identify which infection a patient may have. Microbiology has been, and continues to be, central to meeting many of the current global goals and challenges, such as maintaining food, water and energy supplies. Microbiology research will also help to answer questions such as 'how diverse is life on earth?'

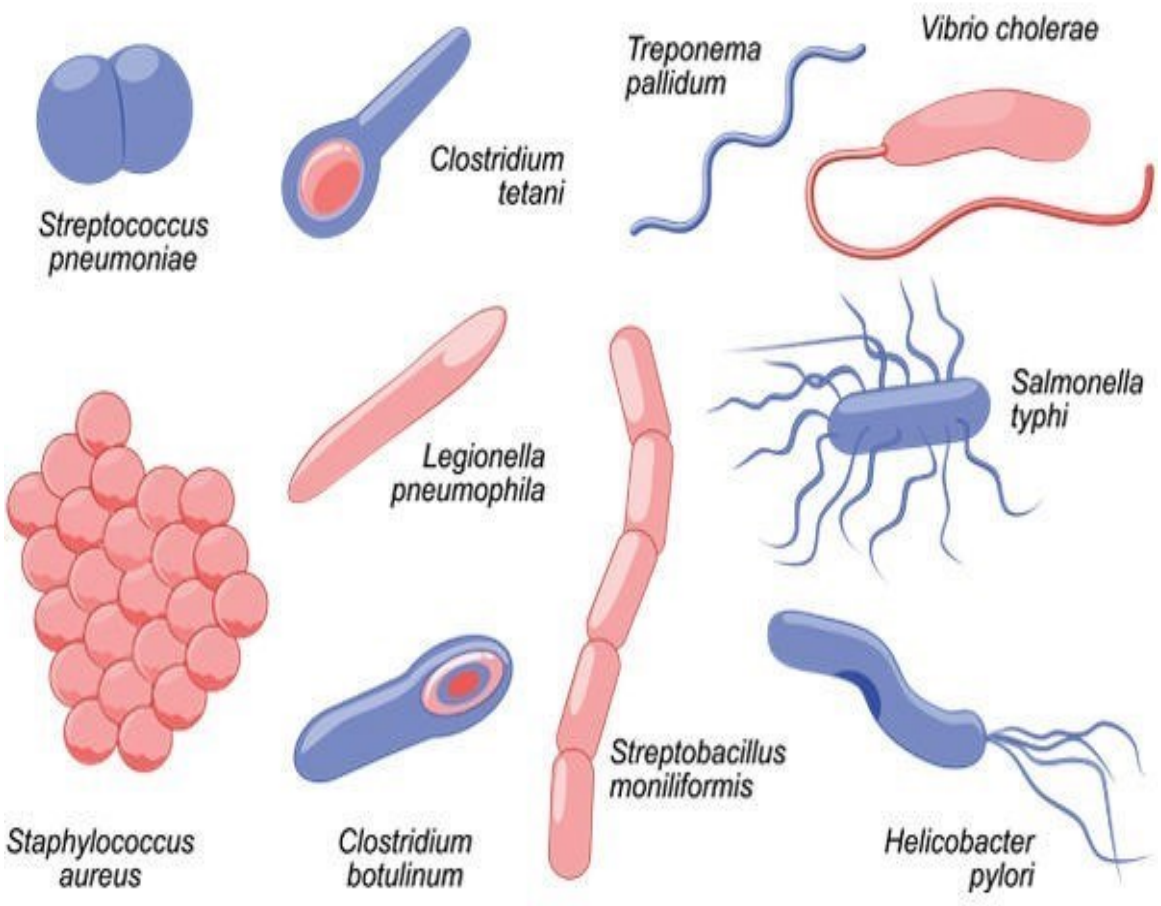


Figure 1: Bacteria are classified into five groups according to their basic shapes: spherical (cocci), rod (bacilli), spiral (spirilla), comma (vibrios) or corkscrew (spirochaetes). They can exist as single cells, in pairs, chains or clusters.

Questions (True or false?)

1. **There are two main types of microbes: bacteria and fungi?**
2. **Bacteria have three main shapes, cocci (balls), bacilli (rods) and spirals.**
3. **Microbes are only in the food we eat.**
4. **Microbes can be useful, harmful or both.**

Introduction to fermented foods

The microbes associated with our foods tend to have a bad name, being linked sometimes to food poisoning. While some microbes make us ill and others can be a nuisance by spoiling our food. Without the activities of certain microbes that are involved in fermentation, there would be no bread, cheese, beer or chocolate. Friend or foe - food microbes are always on the menu. Food fermentation is a distinct form of food processing where foods are produced through the actions of microorganisms, including bacteria and fungi. These microorganisms act as miniature factories, enhancing the nutritional content of the foods and manufacturing new tastes, odours, and health-promoting properties. Fermentation can occur naturally when there is sugars in foods, certain microorganisms are present and the food is stored in a suitable storage condition.

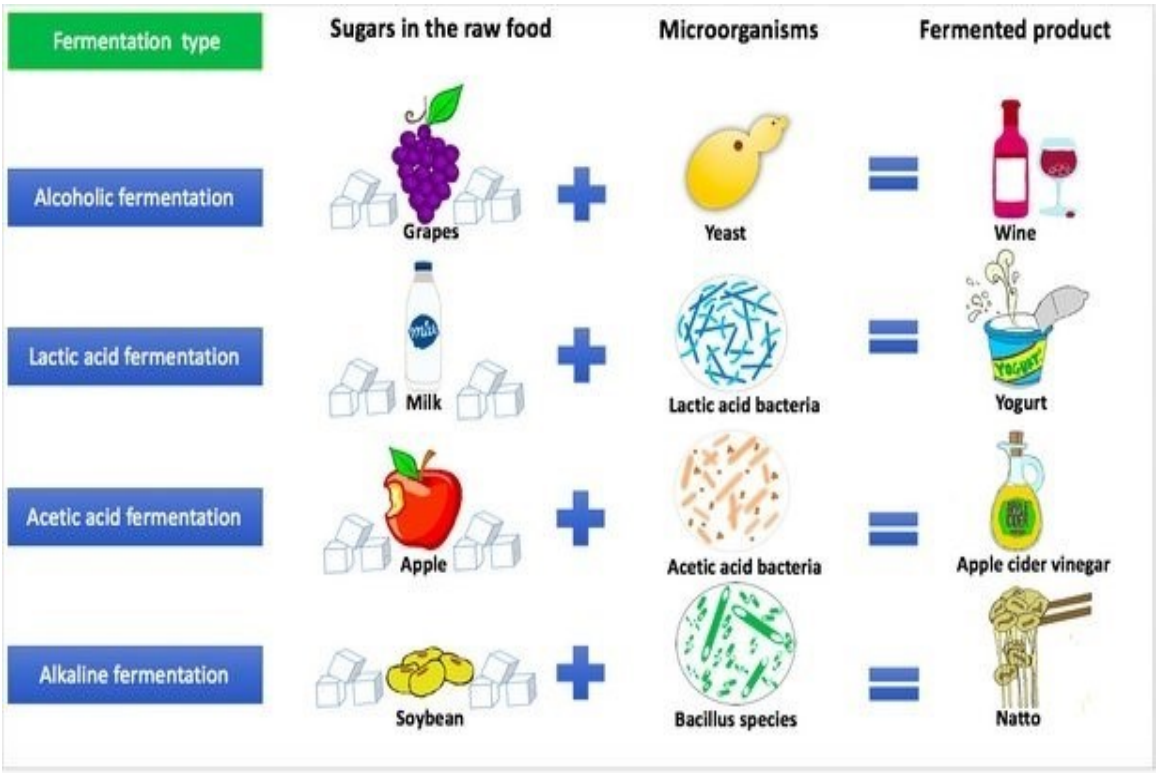


Figure 2: Basics of fermentation. Fermentation is the chemical breakdown of substances by microorganisms; there are four main types of fermentation (highlighted above), and they follow this chemical breakdown process, using different microorganism and producing different fermented foods/drinks, depending on the products the fermenting microorganisms produce.

Food manufacturers continue to use microorganisms today to make a wide range of food products. Fermentation not only gives food a good taste, texture and smell, but it causes changes that reduce the growth of unwanted food microbes. This improves the food's storage time and food safety. Nowadays, fermentation is used to make a wide range of food and drink. To get a clearer understanding of the microbes in fermented foods, both in general and with respect to specific foods, like your kefirs, we need to learn more about the microbes involved. Luckily, DNA sequencing is revolutionising our understanding of the microbes present in fermented foods (as well as in other environments, including the gut). DNA is made up of a series of four nucleotides (think of them as letters - A, T, G and C), and the way these letters are arranged determines the genetic information they encode. In humans, this genetic information might determine what colour eyes or hair we have, while for a microbe it might determine its ability to make a tangy sourdough or a particularly flavorsome cheese. The process of DNA sequencing reads these genetic letters and generates a massive amount of information, which is then interpreted using computing methods.

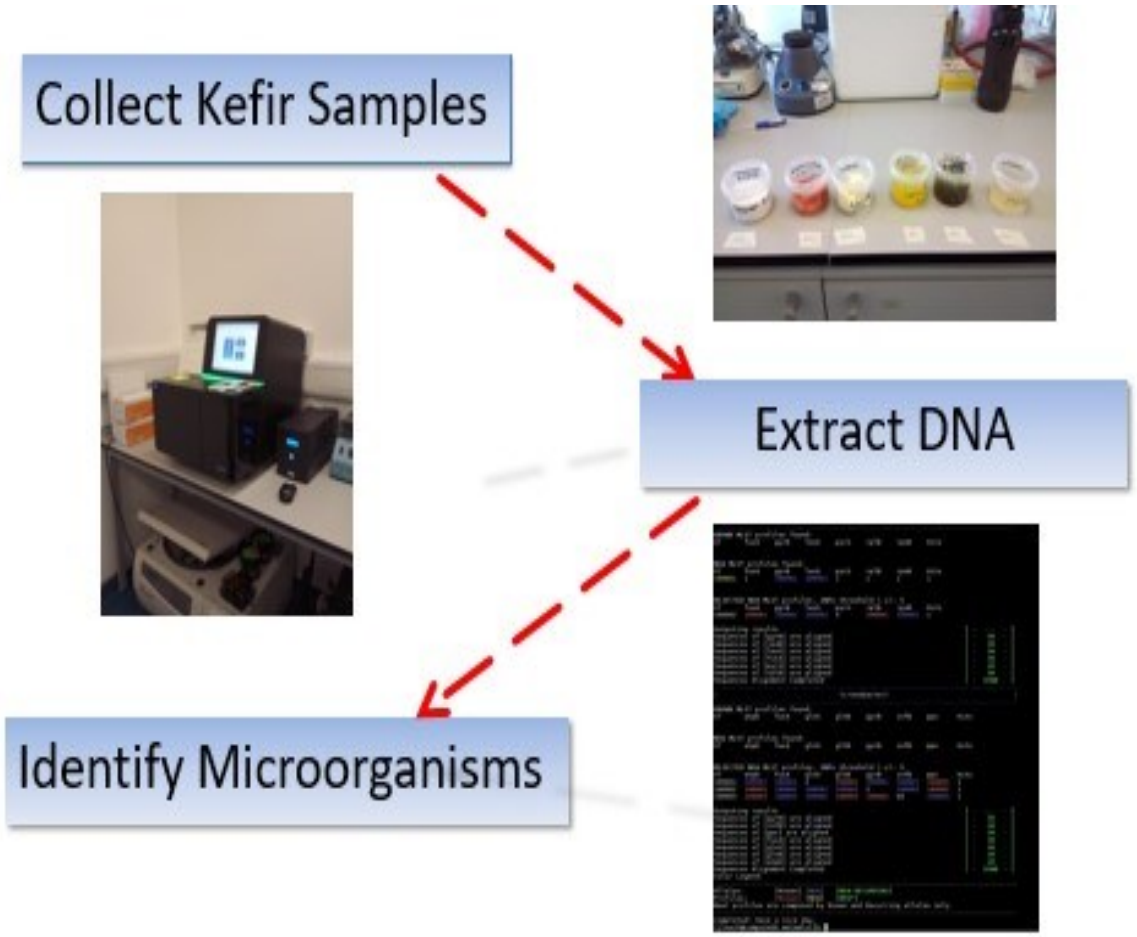


Figure 3: Basics steps involved in Shotgun Metagenomic Sequencing, the approach we applied to your samples. Firstly, we acquired your kefir samples, then we obtained the DNA from the samples through a process called DNA extraction, and then using this DNA, a sequencer, and computer algorithms, we discovered what microorganisms (and the proportion of them) were in your kefir samples .

Questions

1. Can you identify some of the simple sugars and microorganisms involved kefir fermentations?

2. We provided information on all fermented foods, can you expand on our information and provide specific examples of why kefir is so interesting to study?

3. Besides DNA sequencing how else can we study microorganisms?

Your report

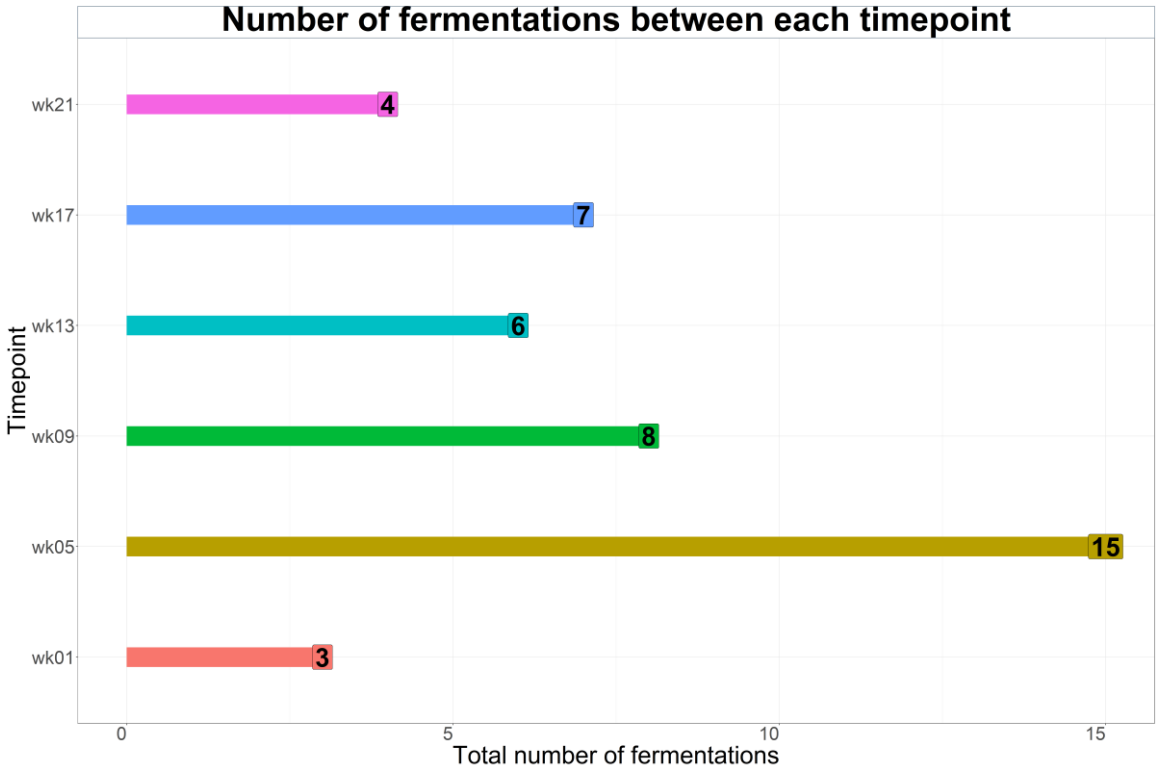


Figure 4: Number of fermentations recorded by you during the Kefir4All study, separated by weeks (wk). Please find an in depth explanation of this figure in the written text below.

During this project you completed 43 fermentations, you can see a breakdown of the numbers of fermentations that you performed over the course of the project in Figure 4. For example, you completed the most surveys from the period of the wk01 to the start of wk05. Thank you for completing those fermentations during the project. On average our Kefir4All citizen scientists completed 36 during this period. Now let's see how your grains grew as a result of all these fermentations!

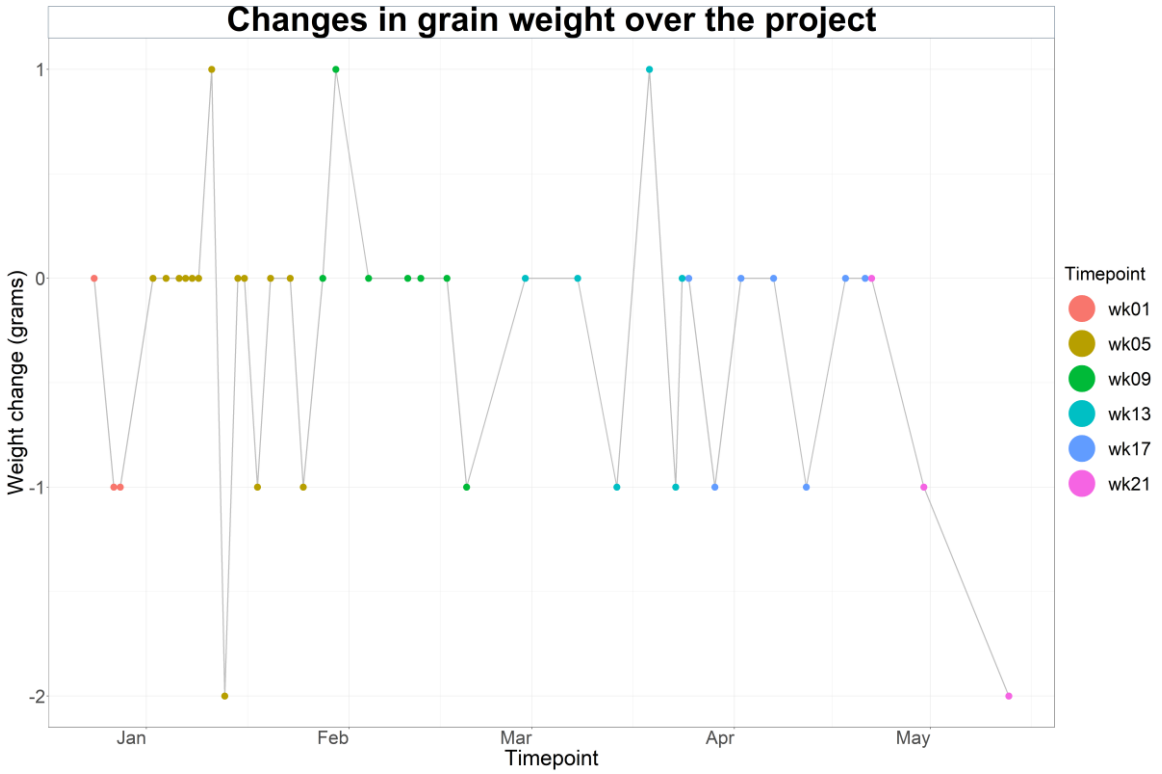


Figure 5: Changes in the weight of your kefir (in grams) over the course of the project, separated by the weeks (wk) of the study.

Figure 5, shows the change in weight of your kefir grains over the project (X-axis). This figure shows, through the different coloured dots, the weight change based on the weight of the grains that you recorded on survey monkey after each fermentation, which is colour coded according to the week of the project. Firstly don't worry about the dips, typically we will see dips in Figure 5 as a result of you providing us with a kefir grain sample. When reading Figure 5, keep in mind that kefir grains are living things. If there are peaks in Figure 5, the kefir grains have responded positively to your care, and in some of your cases will have doubled in size. Why do you think they grew as they did? Did you feed them what they needed? Did you feed them frequently? Did you agitate the grains, by giving the jar a gentle shake every now and again, to spread the microbes throughout the liquid? It is important to feed your microbes regularly in order for your microbes to grow, as they need constant access to nutrients, such as the sugars in the fermentation medium. Microbes in milk kefir fermentations consume sugars in the milk such as lactose, while those in the water kefir fermentations consume the sugars that were added to the water, such as sucrose. It is important that the microbes are feed regularly. If you do not feed the microbes, the landscape of your kefir will resemble that of a desert. In a desert, there is not much food and water available and because of this you will not find many people. Don't worry if you didn't see much grain growth, you may have been fermenting in a nutrient depleted category (e.g. Brown sugar or skimmed milk), or your unique fermentation technique may have favored the growth of other microorganisms that do not produce the substance(s) that makes up the grain. Okay, now that we know how many times you fed your microbes and how they responded to that feeding, let's look at the number of samples that you gave us and that we analysed in the laboratory.

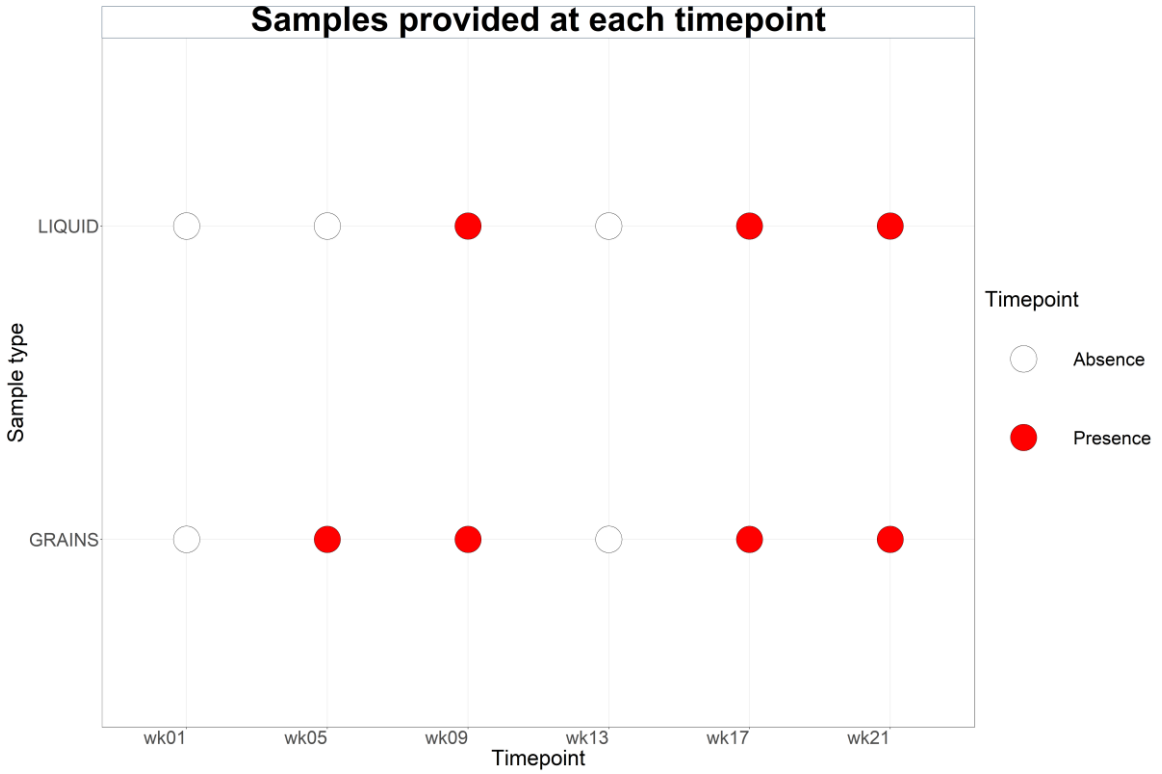


Figure 6: Kefir samples provided over the course of the Kefir4All project, separated by the weeks (wk) of the study.

We were delighted to receive the samples you provided, as shown with the red coloured circles in Figure 6 and we processed them in the lab. We applied a cutting-edge DNA sequencing based approach to study the complete set of microbial genetic information contained in your kefir (grains and liquids). Let look at what we uncovered.

The microbes in your kefir



Figure 7: Changes in the proportions of your kefir samples over the course of the Kefir4ll project, separated by the grain and liquid samples.

Wow, we found 11 microbial species in your kefirs, outlined in Figure 7. These included lactic acid bacteria and spoilage causing bacteria.
Firstly, let's take the time to appreciate that kefir, can only be produced from the actions of microorganisms, which act as miniature factories and change the properties of the initial liquid making a highly acidic product. Microorganisms produce compounds like acetate, diacetyl, ethanol, and acetaldehyde, which shape the unique flavor of kefir. The nutritional benefits of kefir are related to nutrients such as carbohydrates, proteins, minerals, and vitamins. Let's look at some of these microbes in your kefir, what conditions suit them best and how they primarily contribute to fermentations, through the amazing products they have the potential to produce.

We found 8 species of Lactic acid bacteria. These were *Lactococcus lactis, Lactococcus cremoris, Lactococcus raffinolactis, Lactobacillus helveticus, Carnobacterium maltaromaticum, Lentilactobacillus kefiri, Leuconostoc pseudomesenteroides and Lactobacillus kefiranofaciens*. Let's, look at them in more detail below.

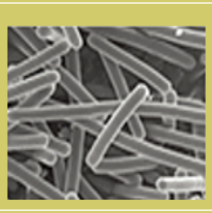
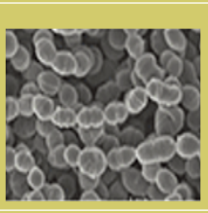
		Lactic acid bacteria
Cell morphology		 Cocci or rod shaped They are often found in pairs or chains 
Sensitivity to oxygen		Can survive with or without oxygen (facultative anaerobic)
Description		Lactic acid bacteria, get almost all of their energy from carbohydates such as lactose and glucose. These microorganisms can for example ferment the sugar lactose, releasing only one by-product, lactic acid. Or convert glucose to lactate and or several other products such as acetic acid, ethanol, and CO2 are produced. The basic fermentative process carried out by these microbes has been used in food production for at least 4000 years. Furthermore, they can produce products with enhanced nutritional properties, as they are able to remove antinutritional factors and, increase the digestibility and the biological value of the proteins
Main metabolite		Lactic acid

Figure 8: Properties of lactic acid bacteria

We found 1 species of spoilage causing bacteria, this was *Pseudomonas aeruginosa*. Let's, look at it in more detail below.

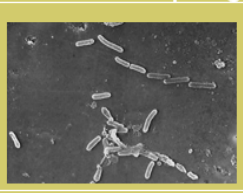
		Spoilage causing bacteria (*Pseudomonas* spp.)
Cell morphology		 Rod-shaped cells that can occur singly, in pairs or chains.
Sensitivity to oxygen		Can survive with or without oxygen (facultative anaerobic)
Description		Milk spoilage is caused by the presence of proteolytic enzymes produced by Pseudomonas spp. These enzymes can cause the liquid to spoil over time. If you have low accounts of this microorganism you know your fermentation was successful.
Main metabolite		Extracellular enzymes (e.g., peptidases and lipases).

Figure 9: Properties of spoilage causing bacteria

That's all very interesting, now, I wonder how the microorganisms changed based on what you were doing at home, when fermenting your kefir. We as scientists look at what Figure 7 refers to as proportions, but its official term is called relative abundance to see how microorganisms changed. Relative abundance is a biodiversity value and shows the commonness or rarity of a species compared to the other species in your kefir. For example we reported *Lactobacillus kefiranofaciens* was one of the microorganisms in your grain with the highest change in proportions over the course of the study. In your first sample *Lactobacillus kefiranofaciens* accounted for 43% of the grain, but in your last kefir grain sample it accounted for 50% of the microorganisms in the grain. Okay we can clearly see what microbes in the grain favoured your unique fermentation technique, I wonder what microorganisms thrived in your liquid samples?

Typically there was a higher proportion of the microorganism *Kluyvera intermedia, Leuconostoc pseudomesenteroides, Carnobacterium maltaromaticum, Bacillus mycoides, Lactococcus raffinolactis, Lactococcus cremoris, Lactococcus lactis* and *Pseudomonas aeruginosa* in the kefir liquid samples compared to the kefir grain samples.

Similarly the microorganism *Lactobacillus kefiranofaciens, Kluyvera intermedia, Leuconostoc pseudomesenteroides, Lentilactobacillus kefiri, Carnobacterium maltaromaticum, Lactobacillus helveticus, Bacillus mycoides, Lactococcus raffinolactis, Lactococcus cremoris, Lactococcus lactis* and *Pseudomonas aeruginosa* all seem to thrive in your kefir liquid samples as the project progressed. Compare the proportions of each of the outlined species in Figure 7, specially compare each week of the study to your first set of samples provided at wk01.

Finally, let's ask a key research question, do you have any microorganisms in your kefir samples (liquids or grains), that weren't detected at the start of the project and may be potentially new? This is a key research question and allows us to study how the microbes in your household influence the fermentation process of the kefir. We found 4 such microorganisms that were not present in the initial grain we provided you with. These were, *Carnobacterium maltaromaticum, Kluyvera intermedia, Leuconostoc*

pseudomesenteroides and *Pseudomonas aeruginosa.* Your microorganisms and their proportions certainly changed during their stay in your household. We have outlined many interesting changes in the microbial landscape of your kefirs (grain and liquid) and in some participants identified the presence of new microbes, not contained within the grain we provided. Now that we have outlined these changes, we at Teagasc need to look at yours and other survey response data to see if there is a link between particular fermentation techniques and changes in any of the microorganisms that we found. See if can figure out where these microbes may have come from?

How diverse was your kefirs over the course of the study?

Definitions used

Relative abundance is a measure of how common a species is relative to the other species in a defined location.
Species richness refers to the number of species in each of your kefirs

For this scientific analysis we used a term called alpha diversity (a-diversity) to describe the 'within-sample' diversity. It is a measure of how diverse a single environment is. Environments with more diversity are considered richer - they are likely to be more ecologically complex and potentially may be more important from environmental and ecosystem functionality perspectives. As explained above, alpha diversity usually takes into account the number of different species observed (species richness). Alpha diversity metrics are also often weighted by the amount in percentages at which the individual microbes are observed.

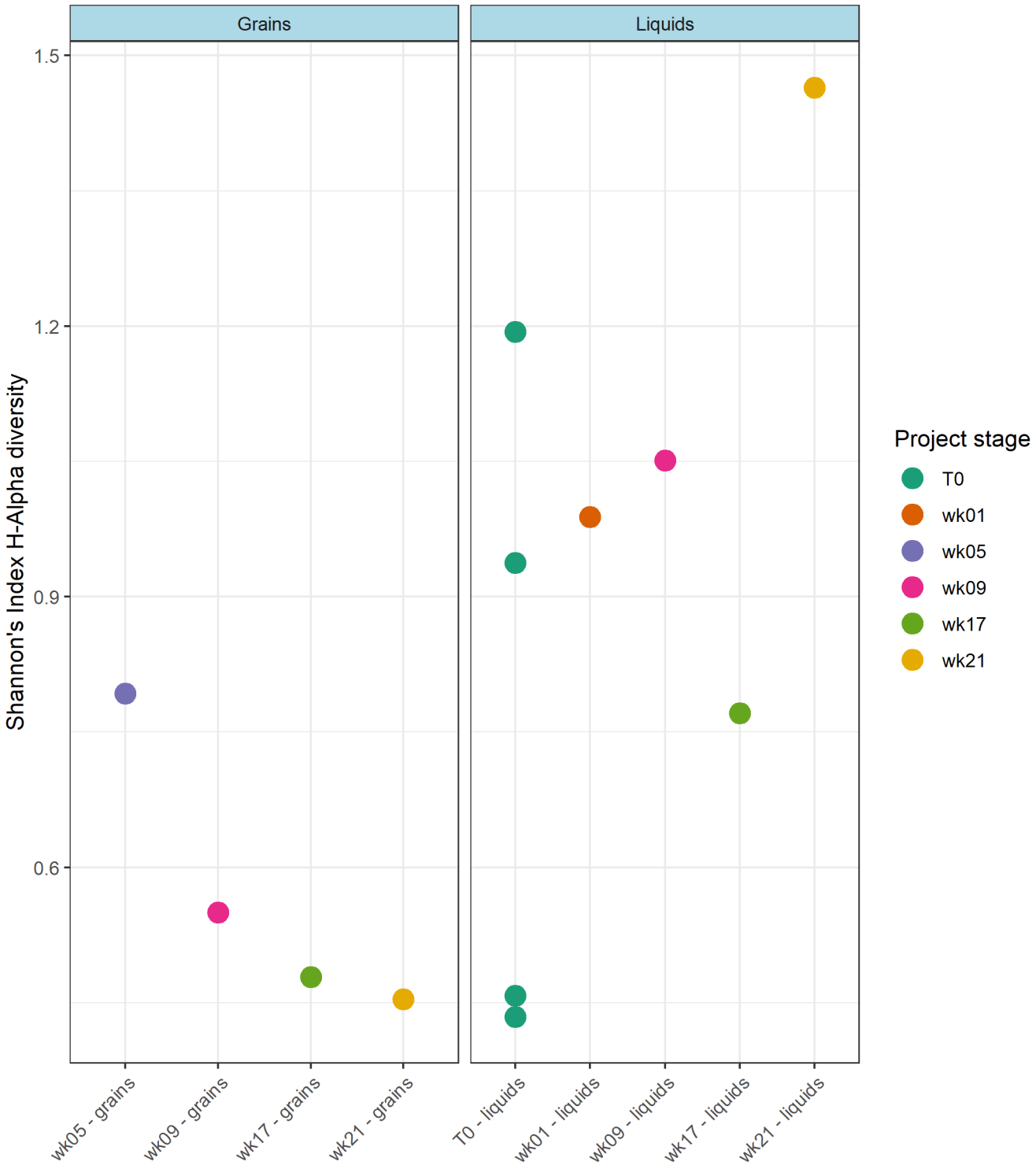


Figure 10: Alpha diversity values for your kefir samples (liquid and grains) over the course of the project, separated by the grain and liquid samples.

Your kefir grain samples typically had less species than your kefir liquid samples. Can you guess why this is? It's likely due to your household, where microorganisms on your surfaces, let's say your kitchen counter, have found brand new homes! Look at Figure 7 and see if you can identify what microorganisms are in your kefir liquid samples but not your kefir grain samples. Such species may be environmental microbes, or they may be in your grain at undetectable levels. Your survey responses will help us to

identify the likely source. Your liquid sample at wk21 was the most diverse, meaning it had the highest number of species. How did this sample, and indeed your other samples, compare to the samples in dark green (T0), which were prepared by Teagasc researchers?

Conclusion

As you can see from this report, through your hard efforts and those of other citizen scientists, we were given the unique opportunity to study how kefir communities are organized by identifying, describing and explaining general patterns that form the structure of communities. As you can see in Section 2.1, we were able to identify many interesting changes such as potential environmental microbes, and to see if any of these microbes became integrated into the kefir grain. In such cases the kefir grain would serve as transport for these microbes into the next fermentation. Your efforts have greatly extended research into the fermented food kefir, and we wholeheartedly thank you for taking the time to contribute to the Kefir4all project. We hope, you as Citizen Scientists enjoyed the project as much as we did. It would be great if you could provide us with some feedback on the project, through the Evaluation Survey circulated with this report.

All the best,
The Kefir4all team

Resources

Below are some of the amazing resources we have generated, including the recordings of the fermented food workshop, which we are very eager to share with you.
Guide to fermentation -
https://www.youtube.com/watch?v=qPxvH5wtvUQ
Fermentation workshop-
https://youtube.com/playlist?list=PLlgXUwWwJdfm5KcWpFpa0f2pO0NxohG3T
Published articles-
https://www.rte.ie/brainstorm/2020/0805/1157452-fermented-foods-kimchi-kombucha-kefir/
https://www.rte.ie/brainstorm/2021/1111/1259091-if-monkeys-are-fermenting-why-arent-you/

Result S2: Supplementary Report 2 - Water kefir. Personalised report provided to the citizen scientists ID113 at the end of the study, detailing and graphically displaying the microorganisms in the participant's kefir samples (kefir grain and liquid).

Metagenomic sequencing of kefir samples from ID113

Liam Walsh[1,2,3], Eimear Ferguson[4], Dr Samuel Mortensen[1,3], Dr Mairead Coakley[1], Dr Fiona Crispie[1], Prof Guilherme Martin[1] and Prof Paul Cotter[1,2,3,4]

[1]Food Biosciences Dept, Teagasc Food Research Centre, Moorepark, Fermoy, Co. Cork, Ireland.
[2]School of Microbiology, University College Cork, Ireland.
[3]APC Microbiome Ireland SFI Research Centre, University College Cork, Ireland.
[4]VistaMilk SFI Research Centre, Teagasc, Moorepark, Fermoy, Co. Cork, Ireland.

March 2023

Please contact paul.cotter@teagasc.ie for any follow-up on the report.

Introduction

In this report, we are excited to reveal the microorganisms that lived in your kefirs, and how they changed as a result of your hard efforts over the 5 months of the Kefir4All project. Firstly, we will revise concepts about microbiology and fermented foods, which you may find useful when looking at your results, which are described in Section 2.

Introduction to Microbiology

Aim

The aim of this introduction section, is to reintroduce you to the exciting world of microbes. Here we will learn about bacteria, viruses and fungi, and their importance and existence in all aspects of our lives.

Background

Microbiology is the study of microscopic organisms that are too small to be visible with the naked eye. These microscopic organisms are what we refer to as microorganisms, or microbes for short. Microbes and their activities are important to virtually all processes on Earth. Microbes can be used as a resource in many ways, e.g., making life-saving drugs, biofuels, foods and drinks, and even cleaning up the environment. Microorganisms also matter because they affect every aspect of our lives, they are in us, on us and around us, and can be both beneficial and harmful. There are three main groups of microbes. Viruses are the smallest of the microbes and cannot grow by themselves. They require a 'host' cell in which to live and reproduce. Once inside the host cell, they often rapidly multiply, destroying the cell in the process! Fungi range in size from being microscopic to very large, and include moulds, mushrooms and mildew. Bacteria are single-celled organisms that, under the optimal conditions, produce offspring, as fast as every 10 minutes. Bacteria can be divided into groups by their shape, e.g., circular (cocci), rod-shaped, spiral, or corkscrew. In cases of illness, doctors find these shapes useful to help identify which infection a patient may have. Microbiology has been, and continues to be, central to meeting many of the current global goals and challenges, such as maintaining food, water and energy supplies. Microbiology research will also help to answer questions such as 'how diverse is life on earth?'

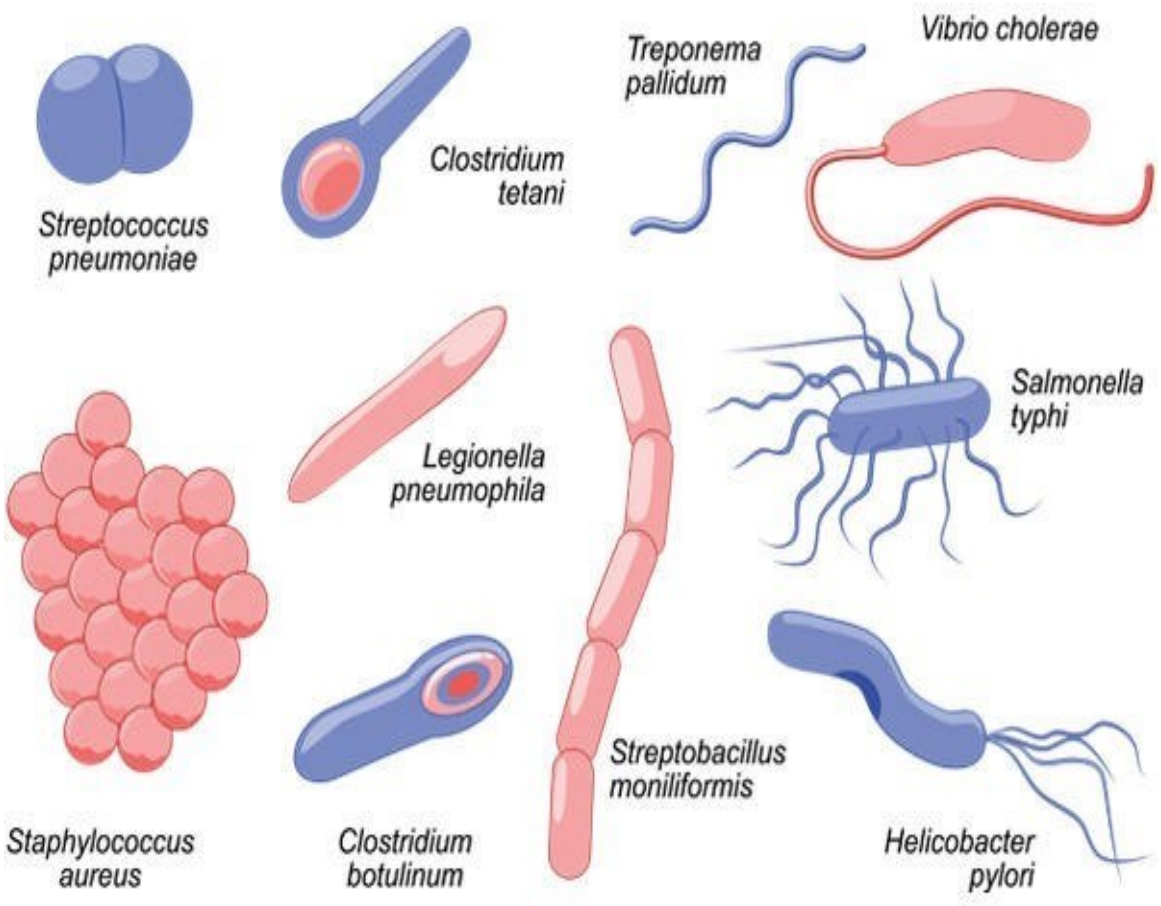


Figure 1: Bacteria are classified into five groups according to their basic shapes: spherical (cocci), rod (bacilli), spiral (spirilla), comma (vibrios) or corkscrew (spirochaetes). They can exist as single cells, in pairs, chains or clusters.

Questions (True or false?)

1. **There are two main types of microbes: bacteria and fungi?**
2. **Bacteria have three main shapes, cocci (balls), bacilli (rods) and spirals.**
3. **Microbes are only in the food we eat.**
4. **Microbes can be useful, harmful or both.**

Introduction to fermented foods

The microbes associated with our foods tend to have a bad name, being linked sometimes to food poisoning. While some microbes make us ill and others can be a nuisance by spoiling our food. Without the activities of certain microbes that are involved in fermentation, there would be no bread, cheese, beer or chocolate. Friend or foe - food microbes are always on the menu. Food fermentation is a distinct form of food processing where foods are produced through the actions of microorganisms, including bacteria and fungi. These microorganisms act as miniature factories, enhancing the nutritional content of the foods and manufacturing new tastes, odours, and health-promoting properties. Fermentation can occur naturally when there is sugars in foods, certain microorganisms are present and the food is stored in a suitable storage condition.

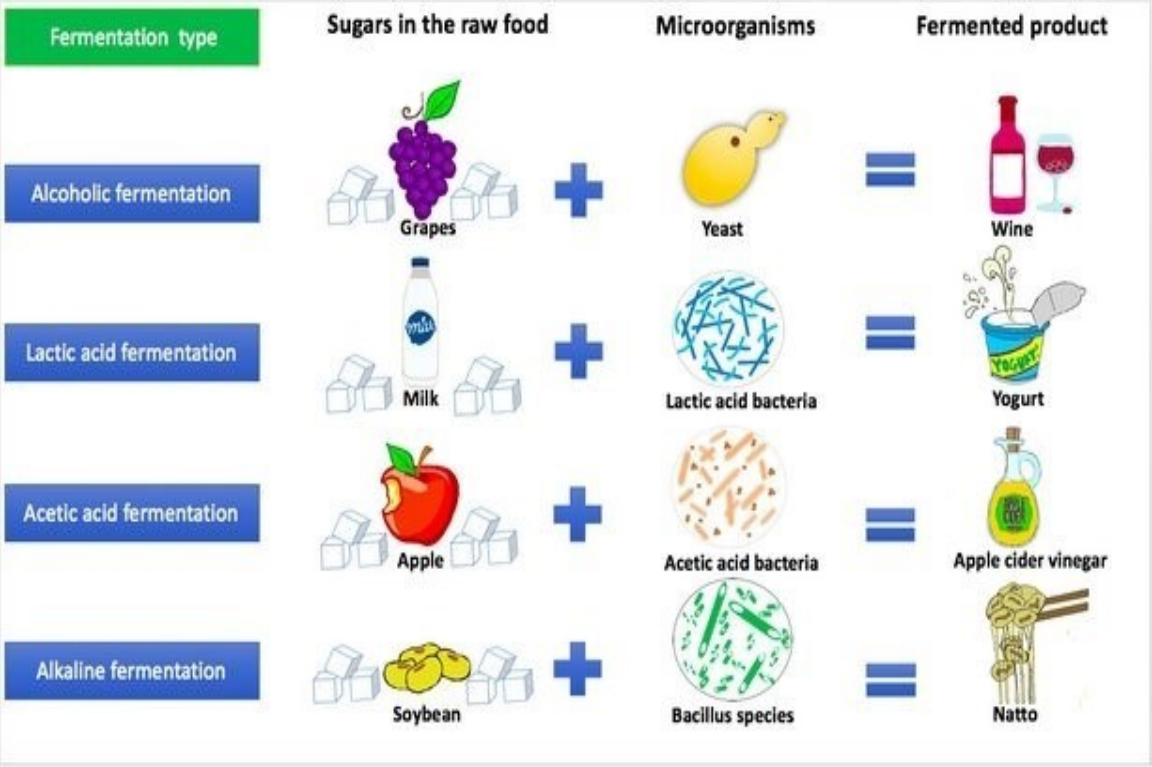


Figure 2: Basics of fermentation. Fermentation is the chemical breakdown of substances by microorganisms; there are four main types of fermentation (highlighted above), and they follow this chemical breakdown process, using different microorganism and producing different fermented foods/drinks, depending on the products the fermenting microorganisms produce.

Food manufacturers continue to use microorganisms today to make a wide range of food products. Fermentation not only gives food a good taste, texture and smell, but it causes changes that reduce the growth of unwanted food microbes. This improves the food's storage time and food safety. Nowadays, fermentation is used to make a wide range of food and drink. To get a clearer understanding of the microbes in fermented foods, both in general and with respect to specific foods, like your kefirs, we need to learn more about the microbes involved. Luckily, DNA sequencing is revolutionising our understanding of the microbes present in fermented foods (as well as in other environments, including the gut). DNA is made up of a series of four nucleotides (think of them as letters - A, T, G and C), and the way these letters are arranged determines the genetic information they encode. In humans, this genetic information might determine what colour eyes or hair we have, while for a microbe it might determine its ability to make a tangy sourdough or a particularly flavorsome cheese. The process of DNA sequencing reads these genetic letters and generates a massive amount of information, which is then interpreted using computing methods.

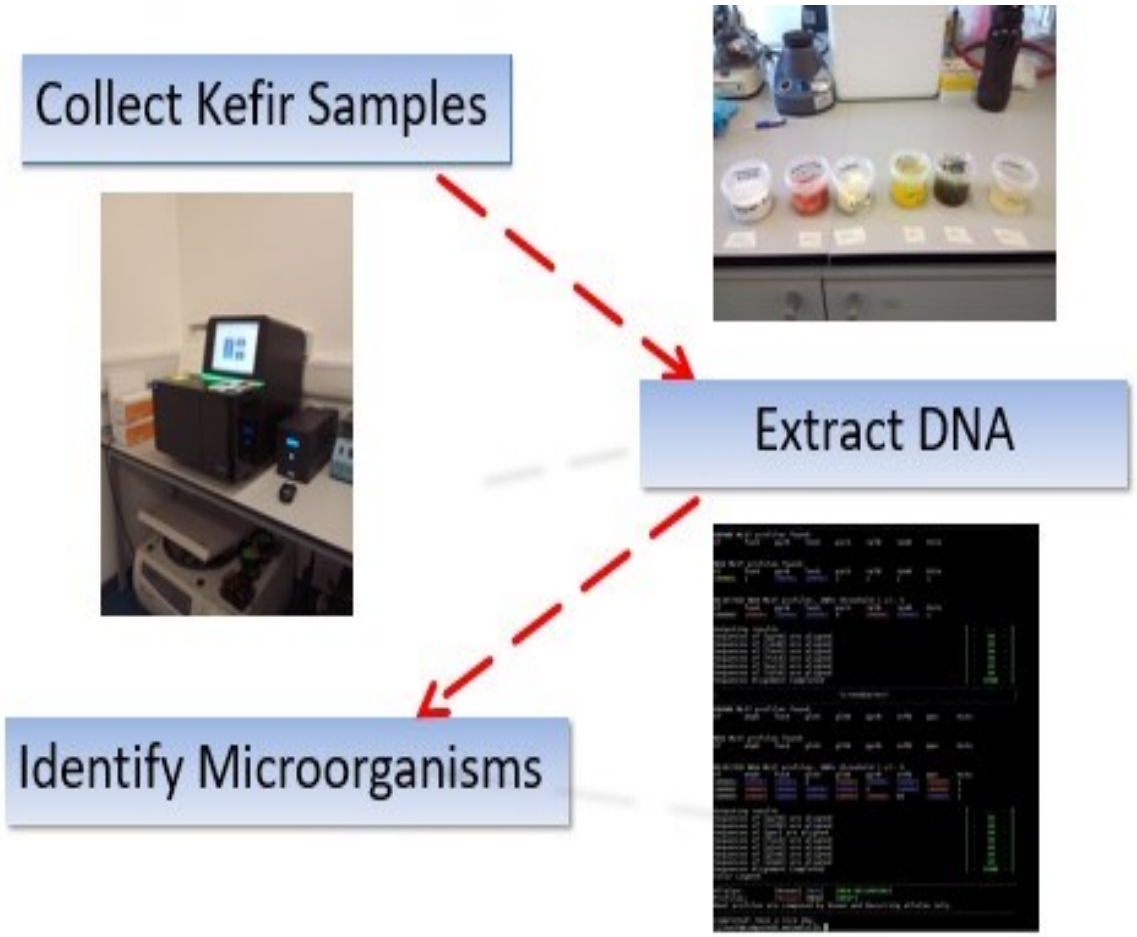


Figure 3: Basics steps involved in Shotgun Metagenomic Sequencing, the approach we applied to your samples. Firstly, we acquired your kefir samples, then we obtained the DNA from the samples through a process called DNA extraction, and then using this DNA, a sequencer, and computer algorithms, we discovered what microorganisms (and the proportion of them) were in your kefir samples.

Questions

1. Can you identify some of the simple sugars and microorganisms involved kefir fermentations?

2. We provided information on all fermented foods, can you expand on our information and provide specific examples of why kefir is so interesting to study?

3. Besides DNA sequencing how else can we study microorganisms?

Your report

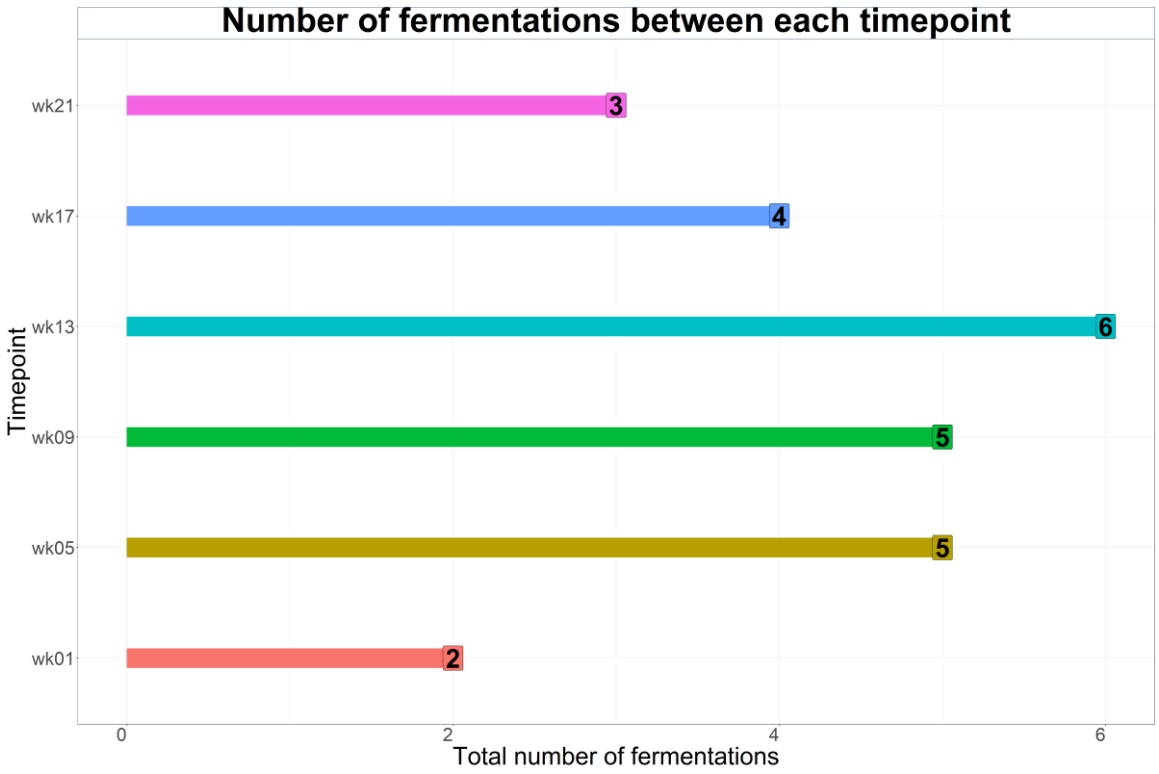


Figure 4: Number of fermentations recorded by you during the Kefir4All study, separated by weeks (wk). Please find an in depth explanation of this figure in the written text below.

During this project you completed 25 fermentations, you can see a breakdown of the numbers of fermentations that you performed over the course of the project in Figure 4. For example, you completed the most surveys from the period of the wk09 to the start of wk13. Thank you for completing those fermentations during the project. On average our Kefir4All citizen scientists completed 36 during this period. Now let's see how your grains grew as a result of all these fermentations!

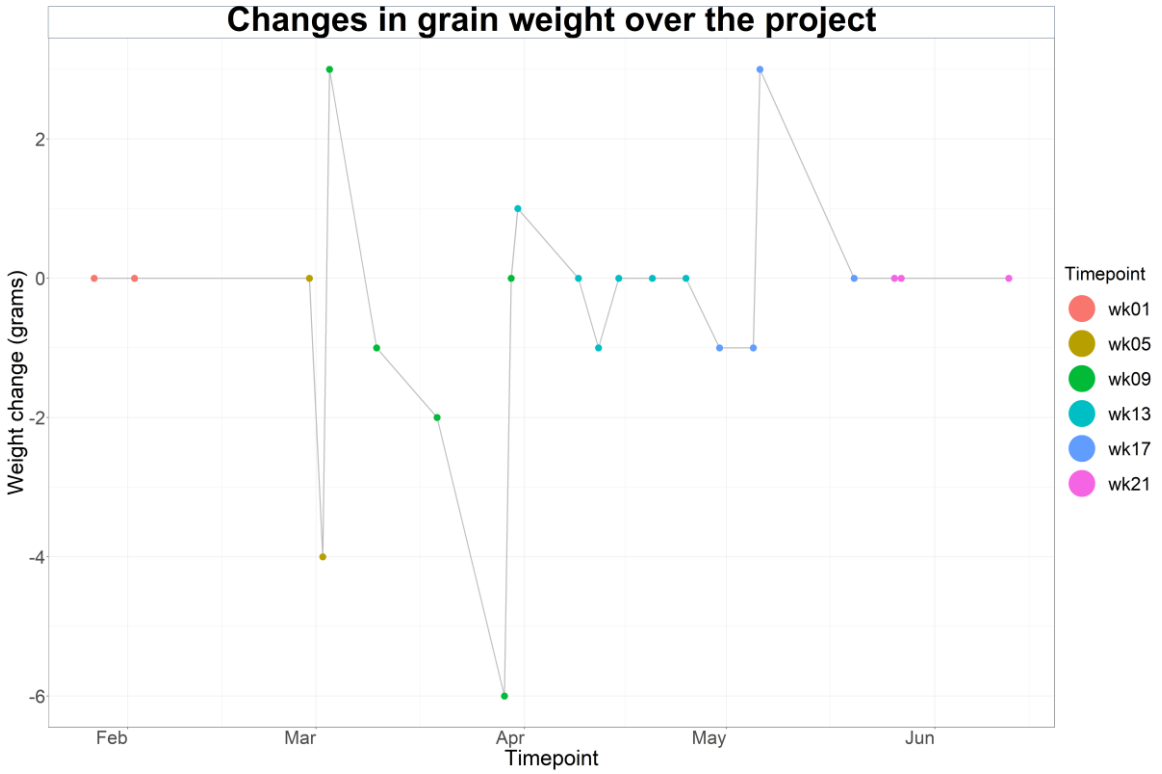


Figure 5: Changes in the weight of your kefir (in grams) over the course of the project, separated by the weeks (wk) of the study.

Figure 5, shows the change in weight of your kefir grains over the project (X-axis). This figure shows, through the different coloured dots, the weight change based on the weight of the grains that you recorded on survey monkey after each fermentation, which is colour coded according to the week of the project. Firstly don't worry about the dips, typically we will see dips in Figure 5 as a result of you providing us with a kefir grain sample. When reading Figure 5, keep in mind that kefir grains are living things. If there are peaks in Figure 5, the kefir grains have responded positively to your care, and in some of your cases will have doubled in size. Why do you think they grew as they did? Did you feed them what they needed? Did you feed them frequently? Did you agitate the grains, by giving the jar a gentle shake every now and again, to spread the microbes throughout the liquid? It is important to feed your microbes regularly in order for your microbes to grow, as they need constant access to nutrients, such as the sugars in the fermentation medium. Microbes in milk kefir fermentations consume sugars in the milk such as lactose, while those in the water kefir fermentations consume the sugars that were added to the water, such as sucrose. It is important that the microbes are feed regularly. If you do not feed the microbes, the landscape of your kefir will resemble that of a desert. In a desert, there is not much food and water available and because of this you will not find many people. Don't worry if you didn't see much grain growth, you may have been fermenting in a nutrient depleted category (e.g. Brown sugar or skimmed milk), or your unique fermentation technique may have favored the growth of other microorganisms that do not produce the substance(s) that makes up the grain. Okay, now that we know how many times you fed your microbes and how they responded to that feeding, let's look at the number of samples that you gave us and that we analysed in the laboratory.

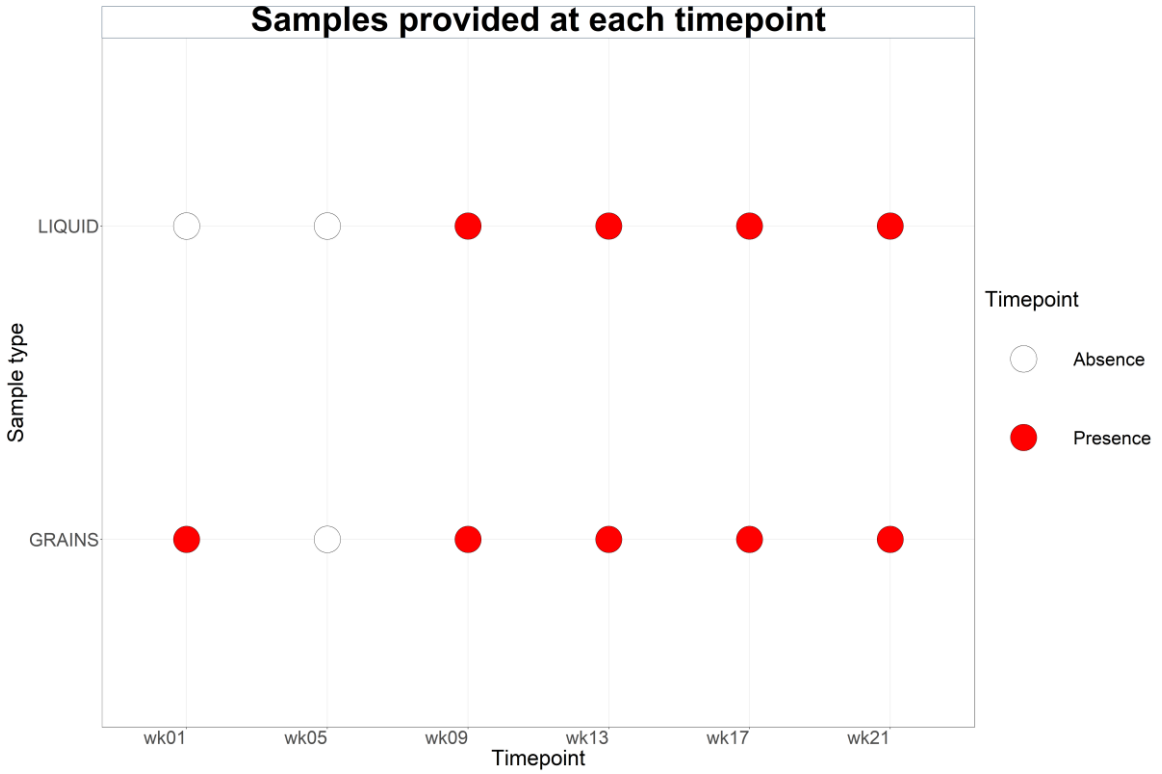


Figure 6: Kefir samples provided over the course of the Kefir4All project, separated by the weeks (wk) of the study.

We were delighted to receive the samples you provided, as shown with the red coloured circles in Figure 6 and we processed them in the lab. We applied a cutting-edge DNA sequencing based approach to study the complete set of microbial genetic information contained in your kefir (grains and liquids). Let look at what we uncovered.

The microbes in your kefir

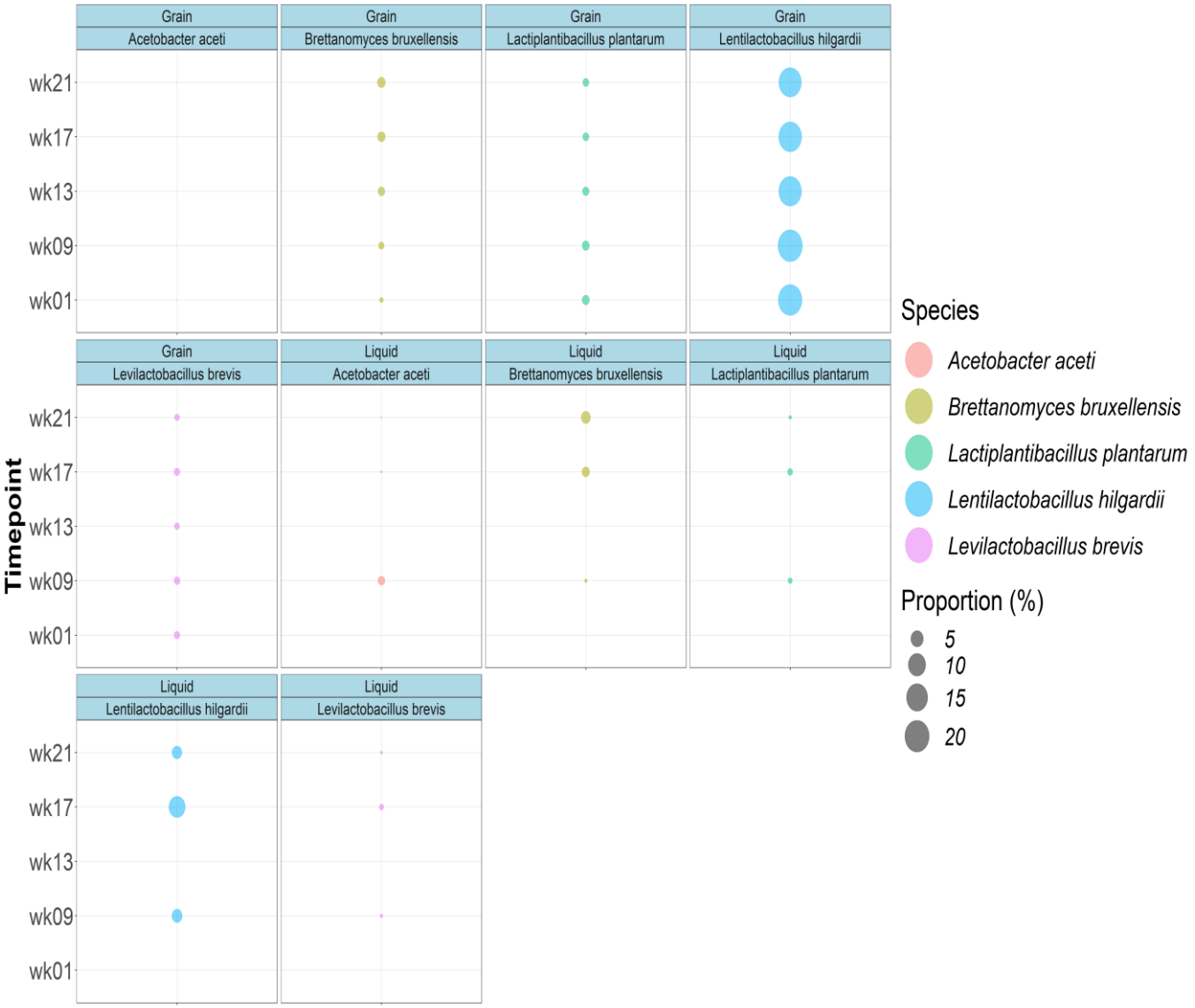


Figure 7: Changes in the proportions of your kefir samples over the course of the Kefir4ll project, separated by the grain and liquid samples.

Wow, we found 5 microbial species in your kefirs, outlined in Figure 7. These included acetic acid bacteria and lactic acid bacteria.
Firstly, let's take the time to appreciate that kefir, can only be produced from the actions of microorganisms, which act as miniature factories and change the properties of the initial liquid making a highly acidic product. Microorganisms produce compounds like acetate, diacetyl, ethanol, and acetaldehyde, which shape the unique flavor of kefir. The nutritional benefits of kefir are related to nutrients such as carbohydrates, proteins, minerals, and vitamins. Let's look at some of these microbes in your kefir, what conditions suit them best and how they primarily contribute to fermentations, through the amazing products they have the potential to produce.

We found 1 species of acetic acid bacteria, this was *Acetobacter aceti*. Let's, look at it in more detail below.

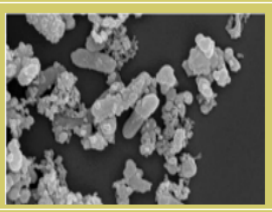
		Acetic acid bacteria
Cell morphology		 Ellipsoidal to rod-shaped cells that can occur singly, in pairs or chains.
Sensitivity to oxygen		Can only survive with oxygen, which can limit their growth during some fermentations
Description		The species of Acetic acid bacteria are well known to transform products such as alcohols, aldehydes, sugars or sugar alcohols in the presence of oxygen to products such as carboxylic acids and acetic acid. Many other bacterial species are also able to transform ethanol under aerobic conditions, but they are unable to do this under high acidic conditions.
Main metabolite		Acetic acid

Figure 8: Properties of acetic acid bacteria

We found 3 species of Lactic acid bacteria. These were *Levilactobacillus brevis, Lentilactobacillus hilgardii and Lactiplantibacillus plantarum*. Let's, look at them in more detail below.

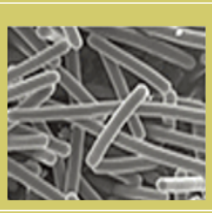
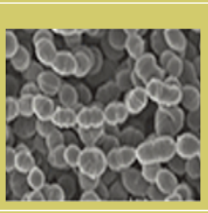
		Lactic acid bacteria
Cell morphology		 Cocci or rod shaped They are often found in pairs or chains 
Sensitivity to oxygen		Can survive with or without oxygen (facultative anaerobic)
Description		Lactic acid bacteria, get almost all of their energy from carbohydates such as lactose and glucose. These microorganisms can for example ferment the sugar lactose, releasing only one by-product, lactic acid. Or convert glucose to lactate and or several other products such as acetic acid, ethanol, and CO2 are produced. The basic fermentative process carried out by these microbes has been used in food production for at least 4000 years. Furthermore, they can produce products with enhanced nutritional properties, as they are able to remove antinutritional factors and, increase the digestibility and the biological value of the proteins
Main metabolite		Lactic acid

Figure 9: Properties of lactic acid bacteria

That's all very interesting, now, I wonder how the microorganisms changed based on what you were doing at home, when fermenting your kefir. We as scientists look at what Figure 7 refers to as proportions, but its official term is called relative abundance to see how microorganisms changed. Relative abundance is a biodiversity value and shows the commonness or rarity of a species compared to the other species in your kefir. For example we reported *Brettanomyces bruxellensis* was one of the microorganisms in your grain with the highest change in proportions over the course of the study. In your first sample *Brettanomyces bruxellensis* accounted for 0% of the grain, but in your last kefir grain sample it accounted for 2% of the microorganisms in the grain. Okay we can clearly see what microbes in the grain favoured your unique fermentation technique, I wonder what microorganisms thrived in your liquid samples?

Typically there was a higher proportion of the microorganism *Brettanomyces bruxellensis* and *Acetobacter aceti* in the kefir liquid samples compared to the kefir grain samples.

Similarly the microorganism *Brettanomyces bruxellensis, Lactiplantibacillus plantarum, Lentilactobacillus hilgardii, Levilactobacillus brevis* and *Acetobacter aceti* all seem to thrive in your kefir liquid samples as the project progressed. Compare the proportions of each of the outlined species in Figure 7, specially compare each week of the study to your first set of samples provided at wk01.

Finally, let's ask a key research question, do you have any microorganisms in your kefir samples (liquids or grains), that weren't detected at the start of the project and may be potentially new? This is a key research question and allows us to study how the microbes in your household influence the fermentation process of the kefir. We found 2 such microorganisms that were not present in the initial grain we provided you with. These were, *Acetobacter aceti* and *Brettanomyces bruxellensis.* Your microorganisms and their proportions certainly changed during their stay in your household. We have outlined many interesting changes in the microbial landscape of your kefirs (grain and liquid) and in some participants identified the presence of new microbes, not contained within the grain we provided. Now that we have outlined these changes, we at Teagasc need to look at yours and other survey response data to see if

there is a link between particular fermentation techniques and changes in any of the microorganisms that we found. See if can figure out where these microbes may have come from?

How diverse was your kefirs over the course of the study?

Definitions used

Relative abundance is a measure of how common a species is relative to the other species in a defined location.
Species richness refers to the number of species in each of your kefirs

For this scientific analysis we used a term called alpha diversity (a-diversity) to describe the 'within-sample' diversity. It is a measure of how diverse a single environment is. Environments with more diversity are considered richer - they are likely to be more ecologically complex and potentially may be more important from environmental and ecosystem functionality perspectives. As explained above, alpha diversity usually takes into account the number of different species observed (species richness). Alpha diversity metrics are also often weighted by the amount in percentages at which the individual microbes are observed.

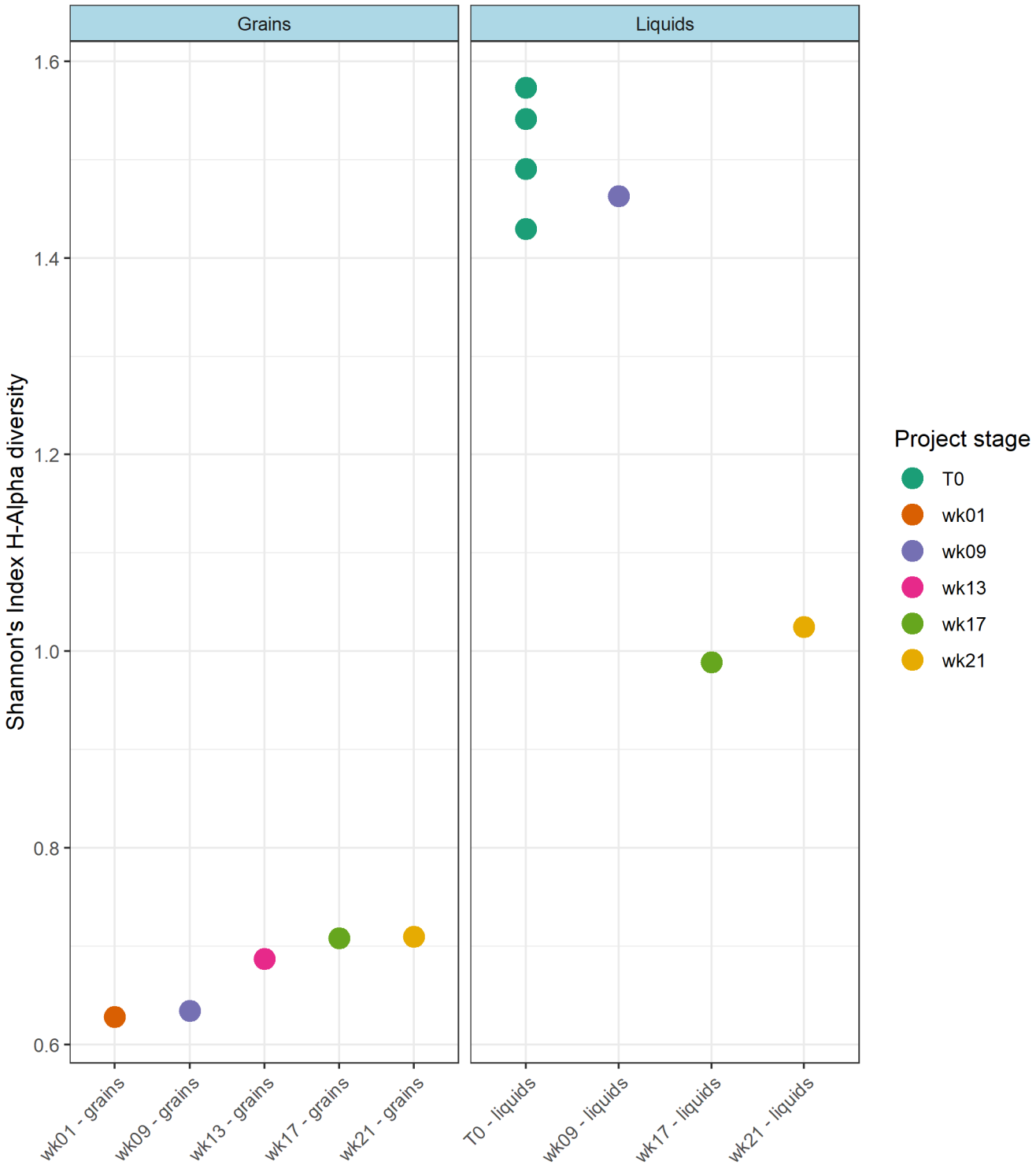


Figure 10: Alpha diversity values for your kefir samples (liquid and grains) over the course of the project, separated by the grain and liquid samples.

Your kefir grain samples typically had less species than your kefir liquid samples. Can you guess why this is? It's likely due to your household, where microorganisms on your surfaces, let's say your kitchen counter, have found brand new homes! Look at Figure 7 and see if you can identify what microorganisms are in your kefir liquid samples but not your kefir grain samples. Such species may be environmental microbes, or they may be in your grain at undetectable levels. Your survey responses will help us to

identify the likely source. Your liquid sample at wk09 was the most diverse, meaning it had the highest number of species. How did this sample, and indeed your other samples, compare to the samples in dark green (T0), which were prepared by Teagasc researchers?

Conclusion

As you can see from this report, through your hard efforts and those of other citizen scientists, we were given the unique opportunity to study how kefir communities are organized by identifying, describing and explaining general patterns that form the structure of communities. As you can see in Section 2.1, we were able to identify many interesting changes such as potential environmental microbes, and to see if any of these microbes became integrated into the kefir grain. In such cases the kefir grain would serve as transport for these microbes into the next fermentation. Your efforts have greatly extended research into the fermented food kefir, and we wholeheartedly thank you for taking the time to contribute to the Kefir4all project. We hope, you as Citizen Scientists enjoyed the project as much as we did. It would be great if you could provide us with some feedback on the project, through the Evaluation Survey circulated with this report.

All the best,
The Kefir4all team

Resources

Below are some of the amazing resources we have generated, including the recordings of the fermented food workshop, which we are very eager to share with you.
Guide to fermentation -
https://www.youtube.com/watch?v=qPxvH5wtvUQ
Fermentation workshop-
https://youtube.com/playlist?list=PLlgXUwWwJdfm5KcWpFpa0f2pO0NxohG3T
Published articles-
https://www.rte.ie/brainstorm/2020/0805/1157452-fermented-foods-kimchi-kombucha-kefir/
https://www.rte.ie/brainstorm/2021/1111/1259091-if-monkeys-are-fermenting-why-arent-you/

4.6.2 Supplementary Figures

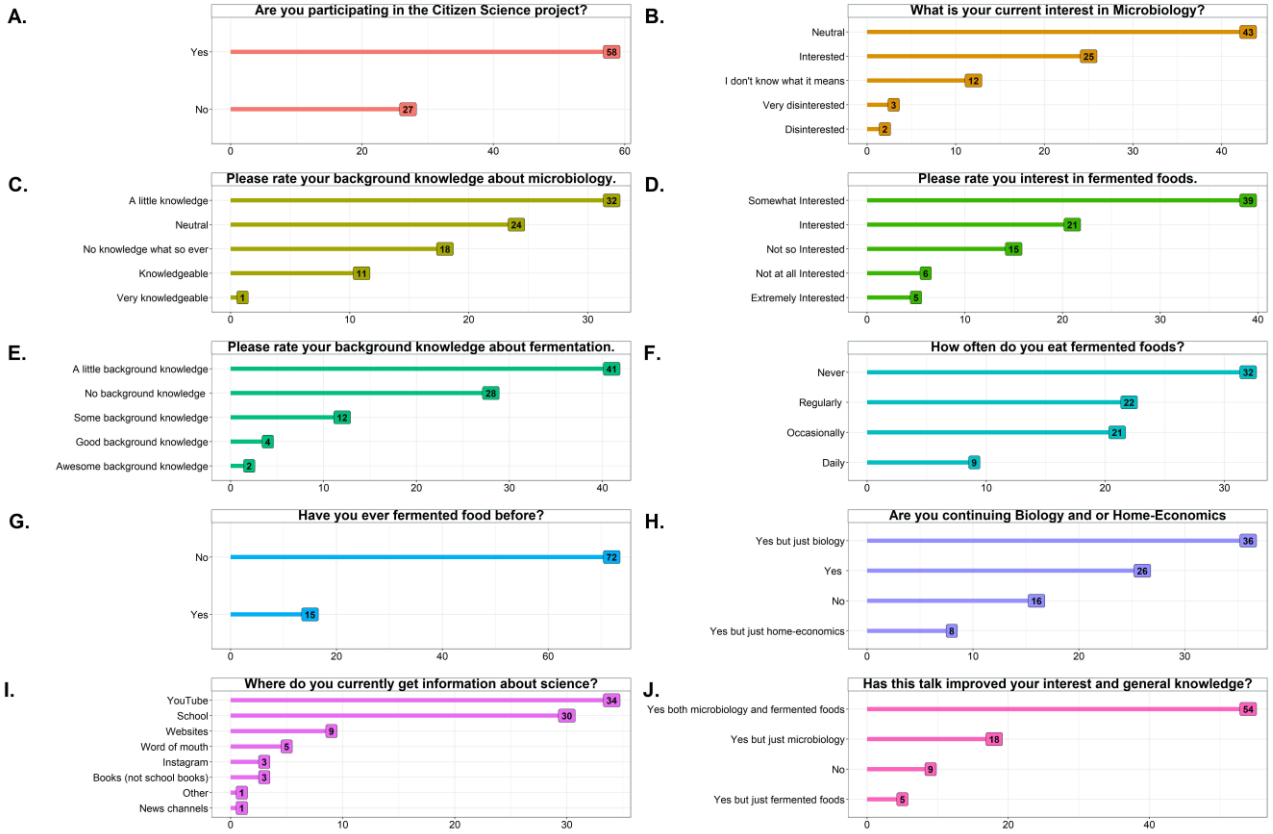


Figure S1: Overview of the survey results collected for the 'Workshop Survey', conducted prior to commencing the Kefir4All project. Responses are detailed for the following questions; **(A.)** 'Are you participating in the Citizen Science project?'; **(B.)** 'What is your current interest in Microbiology?'; **(C.)** 'Rate your background knowledge about Microbiology.' **(D.)** 'Rate you Interest in fermented foods'; **(E.)** 'Rate your background knowledge about fermentation'; **(F.)** ' How often do you eat fermented foods'; **(G.)** 'Have you ever fermented food before?'; **(H.)** 'Are you continuing Biology and or Home-Economics' **(I.)** 'Where do you currently get information about science?'; **(J.)** 'Has this talk improved your interest and general knowledge of microbiology and fermented foods?'.



Figure S2: World cloud of comments and replies from citizen scientists to the question entitled 'Why did you proceed with or stop the project?'.

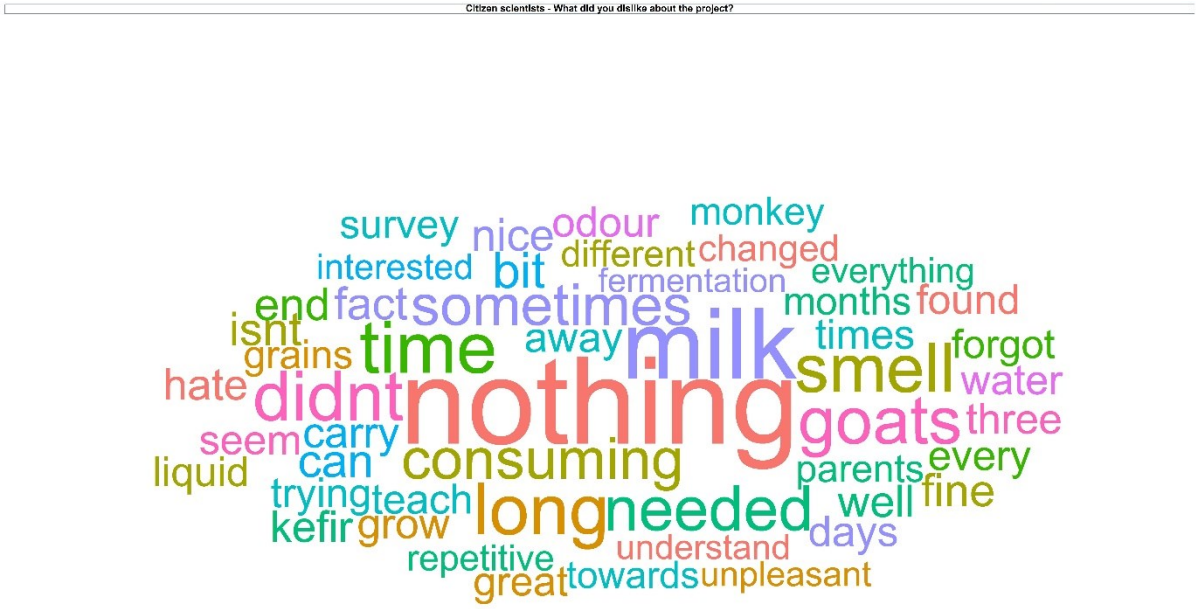


Figure S3: World cloud generated for individualised comments made by citizen scientists to the question entitled 'What did you dislike about the project?'.



Figure S4: World cloud generated for individualised comments made by teachers responders to the question entitled 'What did you dislike about the project?'.



Figure S5: World cloud generated for individualised comments made by citizen scientists to the question entitled 'What did you like about the project?'.

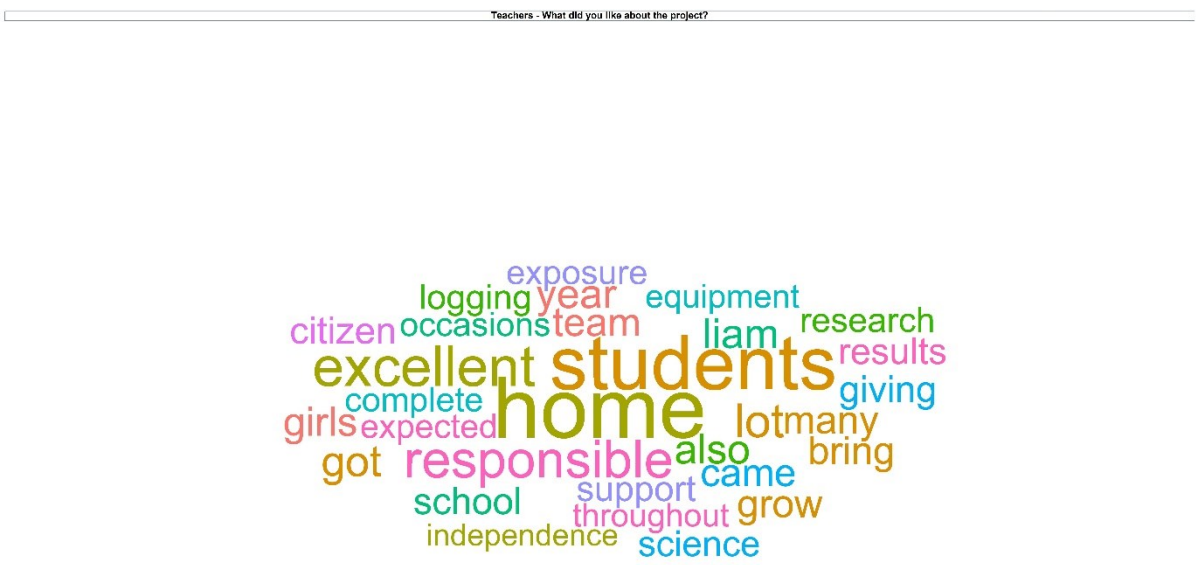


Figure S6: World cloud generated for individualised comments made by teachers responders to the question entitled 'What did you like about the project?'



Figure S7: World cloud generated for individualised comments made by citizen scientists to the question entitled 'Do you have any suggestions for other interactions/activities that you feel would have benefited this project?'



Figure S8: World cloud generated for individualised comments made by teachers responders to the question entitled 'Do you have any suggestions for other interactions/activities that you feel would have benefited this project?'



Figure S9: World cloud generated for individualised comments made by citizen scientists to the question entitled 'Do you have any suggestions for future Citizen Science projects?'



Figure S10: World cloud generated for individualised comments made by teachers responders to the question entitled 'Do you have any suggestions for future Citizen Science projects?'

Chapter 5

Adaptations and community changes in milk and water kefir microbiomes in response to environmental parameters (Kefir4All-Citizen Science Project)

Authors: Liam H. Walsh, Samuel Breselge, José Guilherme Prado Martin, Mairéad Coakley, Fiona Crispie, Paul W. O'Toole and Paul D. Cotter

Contributions:

- **Candidate**, **S.B.** and **P.D.C.** conceived and/or designed the study
- **Candidate** and **S.B.** constructed the surveys
- **Candidate**, **S.B.** and **J.G.P.M.** collected samples, extracted DNA. **S.B.** and **J.G.P.M** performed library preparation. **F.C.** conducted sequencing
- **Candidate** performed bioinformatic and statistical analysis
- **Candidate** wrote the manuscript with contributions from **M.C.**, **S.B.**, **P.W.O.T.** and **P.D.C**
- **M.C.**, **S.B.**, **P.W.O.T.** and **P.D.C** supervised the project

Section 5.0: Abstract

Milk kefir and water kefir are fermented beverages traditionally produced in a household setting through the inoculation of milk or a sugar rich solution, respectively, with a symbiotic microbial ecosystem contained within a milk or water kefir grain. Although traditionally consumed in some regions, the global popularity of, and demand for, both milk and water kefir has increased in recent years. While this has prompted greater research of these beverages and their associated microbiomes, researchers have also studied milk and water kefirs by virtue of their relevance as interesting model microbial communities. Here we apply genome-resolved metagenomic analysis to milk and water kefir metagenomes derived from the same initial microbial population of bacteria, yeast and bacteriophage to study compositional change and microbial evolution at the species and strain level. This was facilitated through the contributions of citizen scientists who used the common source of milk or water kefir grain to carry out fermentations using different substrates and growth conditions for up to 21 weeks. We observed an initial period of rapid compositional change over weeks one (wk01) to nine (wk09) in both the milk and water kefir liquid and grain microbiomes, including in some instances changes in community subtype as represented by the dominant species present, followed by a period of greater stability for many, though not all, metagenomes for the remainder of the study. We document evolutionary change in a number of prevalent species within the milk and water kefir microbiome, which revealed higher rates of nucleotide change on average in the initial periods (wk01-wk09), compared to the latter timeframes of the project (wk13-wk21). Both milk and water kefir metagenomes frequently contained bacteria not typically regarded as being part of milk or water kefir communities and presumably acquired from the environment. Although no potentially pathogenic species persisted in the kefir grains throughout the 21-week project, a number did persist over an extended period. Ultimately, our combined findings provide a new insight into how microbial ecosystems change and evolve over successive fermentations.

Section 5.1 Introduction

Milk and water kefir are fermented beverages containing simple symbiotic microbial ecosystems. Kefir is traditionally produced from the inoculation of kefir grains. Milk and water kefir grains, which are exopolysaccharide matrices [1] containing symbiotic communities of bacteria and yeast, into either milk [2] or sugary water often supplemented with fruits, respectively [3]. Kefir can be produced easily, quickly and cheaply in households, with the grains acting as self-renewing community-scale inocula [4, 5]. These communities have undergone propagation over many generations to co-exist through metabolic cooperation and spatiotemporal niche partitioning [6]. Our understanding of the composition and functionality of kefir microbial communities from different grains has enhanced considerably in recent years due to the application of high throughput DNA sequencing [3, 6-12]. However, considerably less is known about how these communities change over time, especially in response to different environments. Genome-resolved metagenomic analysis provides a means to gain insights into short term evolutionary changes in the composition of microbial environments [13], including those of kefir, as they adapt to evolutionary pressures. Such insights can further our understanding of the influence of evolution on community structure and functionality, and enable targeted manipulation of community dynamics, to, for example, promote the dominance of certain microbial species associated with improved flavour profiles [14] or other desirable features. Furthermore, kefir communities may have applications as simple natural model systems that can be studied in order to predict evolutionary trends in more complex communities. Indeed, kefir has attracted interest from a variety of research groups, and has been utilised as an ecological model for studying community dynamics such as microbial succession [3, 5]. Although few publications have studied the stability of kefir microbiomes over time, it is clear that the community profile of fermenting microorganisms and food characteristics are impacted by the time, conditions, environment, nutrients provided and production technique [4, 15]. Here we used metagenomics based approaches to study changes in kefir communities over 21 weeks, as they adapt to external pressures [14], including different environments, variations in durations between subculturing and nutrients, introduced through the actions of citizen scientists during the citizen science initiative Kefir4All.

Section 5.2. Materials and methods

5.2.1 Participant recruitment and equipment:

A total of 123 local citizen scientists were recruited through a series of fermentation workshops given at public fora. All attendees were also granted access to recordings of the workshop[6]. Prior to commencing the study, citizen scientists received fermentation kits that included a kefir grain (common source of milk kefir or water kefir grain), fermentation vessels, thermometers, utensils, pH strips and fermentation guidelines. Fermentation equipment was purchased from kefirako including fermentation jars, pH test strips and thermometers[7]. Collection tubes were purchased from Thermo Scientific™[8]. The milk kefir grains were sourced from *Every good thing*, UK[9]. The water kefir grains were sourced from *Live Ferments*, Ireland[10]. Fermentation guidelines were also accessible online, and included detailed written instructions on the steps involved in the fermentation process, good hygiene practices and further instructions concerning sample and metadata collection. Furthermore, the instructions provided links to surveys through the Survey Monkey platform, which were used to collect metadata relating to fermentation practices and household environments.

5.2.2 Study design:

Citizen scientists were asked to perform repeat regular fermentations (every 1-3 days) throughout the project (wk01 - wk23) using milk kefir or water kefir grains in one of four fermentation categories. If the citizen scientists selected milk kefir, they had the option to ferment in a) 300 ml of full fat milk, b) 300 ml of low fat milk, c) 300 ml of goat milk or d) 300 ml of skimmed/semi skimmed milk. If the citizen scientists

[6] https://youtube.com/playlist?list=PLlgXUwWwJdfm5KcWpFpa0f2pO0NxohG3T
[7] https://kefirko.com/product/kefir-maker-848-ml/?v=d2cb7bbc0d23
[8] https://www.fishersci.com/shop/products/nunc-15ml-50ml-conical-sterile-polypropylene-centrifuge-tubes/p-4526024
[9] https://everygoodthing.co.uk/products/milk-kefir-grains
[10] https://kefirgrains.ie/

selected water kefir, the substrate options were a) 12 gram of white sugar and 1 dried fig, b) 12 gram of white sugar, 1 dried fig and 1 slice of fresh lemon, c) 12 gram of white sugar and 1 dried apricot or d) 12 gram of brown sugar. Citizen scientists were asked to carry out fermentations for a minimum period of 24 hours (1 day) to a maximum period of 72 hours (3 days). Repeated regular fermentations were carried out by the citizen scientists in private households for up to 21 weeks.

5.2.3 Collection of metadata and samples:

Citizen scientists were asked to complete a survey at the start of the project to record information about their household environment, and then a separate survey relating to each individual fermentation to provide metadata concerning the fermentation process. Metadata collected included the 'Kefir type' (Milk or water kefir and grain or liquid), 'Timeframe' (wk01-wk21), 'Presence/absence of pets', 'Presence/absence of other fermented food', 'Grain wash', 'Room temperature', 'pH', 'Grain weight', 'Water type' (Applicable only in water kefir), 'Liquid volume' and 'Sugar weight' (Applicable only in water kefir), 'Fermentation categories' (see methods section 5.2.2) and 'Fermentation characteristics'. 'Fermentation characteristics' in milk kefir included 'Bubbling', 'Coagulation', 'Whey and curds' and 'Unchanged'. 'Fermentation characteristics' in water kefir included 'Biofilm', 'Bubbling', and 'Unchanged'. The citizen scientists were asked to provide both kefir grain and liquid samples fermented for 24 hours at wk01, wk05, wk09, wk13, wk17 and wk21.

5.2.4 DNA extraction, preparation of libraries, shotgun metagenomic sequencing and data pre-processing

Prior to DNA extraction, 15 ml of kefir liquid was centrifuged at 5,444 × g for 30 minutes (min) at 4°C. Milk kefir liquid samples were washed with 10 ml of phosphate buffered saline (PBS) and centrifuged again at 5,444 × g for 15 min to remove protein and fat residues. Two additional rounds of washing and centrifugation were performed before pelleting the microbial cells. Water kefir liquid samples were washed with 1 ml of phosphate buffered saline (PBS) and centrifuged again at 5,444 × g for 1 min. The cell pellet was resuspended in 800 µl of Solution CD1 from the DNeasy PowerSoil Pro Kits. Total DNA from the

resuspended pellets was extracted and purified according to the standard DNeasy PowerSoil Pro Kits protocol. Total DNA was also extracted from each kefir grain. Fragments of 100 mg were removed from each of the grains and added to separate PowerBead tubes (Cambio, Cambridge, United Kingdom). The grain fragments were homogenized by shaking the PowerBead tube on the TissueLyser II (Qiagen, West Sussex, United Kingdom) at 20 Hz for 10 min. Following homogenization, DNA was purified from the sample by the method outlined above. Total DNA was initially quantified the Qubit High Sensitivity DNA assay (BioSciences, Dublin, Ireland). DNA was then stored at -20°C. Sequencing libraries were prepared using the Illumina DNA Prep (M) Tagmentation kit (Illumina) following the manufacturer's guidelines. Libraries were sequenced for 300 bp paired-end reads using the Illumina NovaSeq6000 platform.

5.2.5 Food metagenomics datasets

In this study, we further analyse two global kefir metagenomics datasets from Walsh *et al*. [11] and Breselge *et al*. (data unpublished), each of which has been described in greater detail and depth in Walsh *et al* and Breselge *et al*, respectively. Briefly, the sampling of Walsh *et al* 2023 comprises 256 milk kefir metagenomes and Breselge *et al* 2023 consisted of 375 water kefir metagenomes separated into 75 water kefir grain metagenomes and 300 water kefir liquid metagenomes.

5.2.6 Nanopore library preparation and sequencing

Approximately 1 µg of DNA extracted by the standard method was used to prepare a sequencing library using the Ligation Sequencing Kit (SQK-LSK109) according to the manufacturer's instructions. For library cleanup, short fragment buffer (SFB) was used to retain DNA fragments of all sizes. The sequencing library for DNA extracted by the DNeasy PowerSoil Pro Kits was prepared using the PCR Sequencing Kit (SQK-PSK004), according to the manufacturer's instructions. Approximately 50 fmol of the prepared library was loaded onto the R10.4.1 (FLO-MIN114.001) flow cell. Sequencing was performed using the ONT GridION sequencing platform.

5.2.7 Bioinformatic analysis

Shotgun metagenomic sequences were processed using the Teagasc high performance computing cluster. Initially raw paired-end FASTQ files containing the metagenomic shotgun sequences were trimmed using Trimgalore (v.0.6.1) [16] to remove adapter content and low quality reads (average quality score <Q20), fragmented (<75 bp) and with more than two ambiguous nucleotides. To identify contaminating DNA (i.e., from the substrate, host or sequencing run controls), milk kefir metagenomes were queried against a reference database containing a bovine, phiX 174 and human genome using Bowtie2 (v.2.3.4) [17] with the parameter '-sensitive-local'. Water kefir metagenomes were queried against a reference database containing a phiX 174 and human genome (GRCh38) using Bowtie2 (v2.3.4) with the parameter '-sensitive-local'. Reference genomes were downloaded from iGenomes (https://support.illumina.com/sequencing/sequencing_software/igenome.html) and represented the most up to date versions available in NCBI (November 2022). Remaining high-quality reads were sorted and split to create forward, reverse and unpaired reads output files for each metagenome.

Compositional analysis was carried out with MetaCache (v 2.3) [18] and functional profiling was performed with SUPER-FOCUS (v 0.34) [19] using the DB_90 database and diamond for alignment, respectively. Strain-level profiling by StrainPhlAn4 [20] was performed using the parameters --mutation_rates --marker_in_n_samples: 1 --sample_with_n_markers: 10 --phylophlan_mode: accurate. The mutation rates represent the number of single nucleotide polymorphisms (SNPs) between strains normalised by alignment length of the multiple sequence alignment (MSA) of all strains of the outlined species [20]. Normalised phylogenetic distance between strains of the same species were calculated using the tree_pairwisedists.py script in PyPhlAn[11]. Phylogenetic distances was calculated as the fraction of mismatching nucleotides between the two strains in the MSA, normalised by the total branch length[20]. Metagenomic assembly was performed using SPAdes (v3.15.3) [21] on individual metagenomes and merged metagenomes to recover individual sample assemblies and co-assemblies respectively. Several SPAdes (v3.15.3) commands were

[11] https://github.com/SegataLab/pyphlan.

performed, specifically using the parameters --meta parameter [21], for general assembly the --metaviral [22] parameter for viral assembly and the --metaplasmid [23] parameter for plasmid assembly. MetaWRAP (v1.3.2) [24] was used for genome binning, with default settings. CheckM 2 (v0.1.3) [25] and BUSCO (v5.1.3) [26] were used to check the quality of the bacterial and eukaryotic metagenome-assembled genomes (MAGs) respectively. Low-quality MAGs, i.e., <50% completeness and/or >5% contamination, were removed from downstream analysis. Taxator-tk (v1.5.0) [27] implemented in the metaWRAP's classify_bins module and GTDB-Tk [28] were used to assign taxonomy to the MAGs. GTDB-Tk was used to further identify putative new species based on ANI values <95%. dRep (v3.2.0) [29] was used to cluster MAGs representing putative new species into primary and secondary clusters on the basis of their relative similarities. Average nucleotide identity between all strains was calculated using dRep (v3.2.0). Same strains were determined based on a protocol outlined in Feehily *et al.*. Briefly, MAGs with nucleotide identity (ANI) ≥ 99.9% and genome coverage ≥ 90 % were deemed to be the same strain. MAGs considered in this study were annotated with DRAM (Distilled and Refined Annotation of Metabolism) [30].

Guppy (v6.3.8) [31] was used to extract the bases from the downloaded long read fast5 data and turn them into standard fastq files. The long read fastq files were processed with porechop v0.2.4 (https://github.com/rrwick/Porechop) to trim off sequencing adapters. Trimmed reads were aligned against the same reference database as the short reads using minimap2 (v0.2.4) with the parameters. As long reads that mapped to the reference database, were considered to be contaminating and set aside. The remaining unmapped reads were considered microbial reads. They were extracted from the sam format file using samtools with the command ‘samtools view -f 4’ [30].

Metagenomic hybrid assembly was carried out with OPERA-MS (v0.8.2) [36]. For all outlined tools applied (see above), default parameters were used unless specified otherwise. Analysis of resulting tables from each tool described was completed in R (v4.1.2) [32].

5.2.8 Statistics

All statistical analysis was performed in R (v4.1.2) implemented through R studio. Statistical analysis of compositional and metabolic data was carried out with vegan (v2.6.2) [33], to compute alpha and beta diversity values. Dissimilarity values between total metadata collected for each sample was computed using the Gower's distance [34] through the StatMatch (v1.4.1) [35] gower.dist function. The stats (v4.1.2) kruskal.test function was used to perform the Kruskal-Wallis rank sum test, to identify significant differences in alpha diversity values. The adonis function in the vegan package [33], was used for permutational analysis of variance (PERMANOVA) [36]. The pairwiseAdonis function in the pairwiseAdonis (v0.4) [37] was used to perform pairwise comparisons of Adonis using Benjimani- Hochberg correction for multiple comparisons. PERMDISP [38] and analysis of variance (ANOVA) [39] was used to assess group dispersion using vegan (v2.6.2) and stats (v4.1.2) respectively. Analysis of Similarity (ANOSIM) [40] was used to assess similarities in compositional profiles between the data sources 'Kefir4All -Milk kefir', 'Kefir4All -Water kefir' metagenomes, 'Breselge *et al* - Water kefir' metagenomes and 'Walsh *et al* - Milk kefir', using the anosim function in vegan (v2.6.2). The linear discriminant analysis (LDA) effect size (LEfSe) [41] was used to perform differentially abundance analysis. The Mantel test [42] was perfomed using the mantel function from the vegan (v2.6.2) package. Envfit from the vegan (v2.6.2) package was used to model 12 different covariates that may influence microbiota beta-diversity of the milk and water kefir microbiome.

5.2.9 Data visualisation

The experimental design figure was produced using BioRender [43]. Ordination plots, correlation scatter plots, heatmaps, density plots, bubble plots, bar charts were produced using ggplot2 (v3.3.6) [44]. Violin and boxplots were produced using ggstatsplot (v0.11.1) [45].

Section 5.3 Results

5.3.1 Compositional analysis reveals a diverse range of prevalent microorganisms, including several not detected in initial kefir fermentations.

During the 21 week citizen science initiative termed Kefir4All, 123 participants were recruited and asked to perform regular repeat fermentations with a common source milk or water kefir grain and provide representative samples for analysis at week 1 (wk01), wk05, wk09, wk13, wk17 and wk21)(Figure 1). In addition, a smaller number of samples referred to as T0 and representing the initial water and milk kefir grain samples and their resulting liquids produced according to the outlined fermentation categories (see methods section 5.2.2) were collected and or produced within laboratory-controlled conditions (Figure 1).

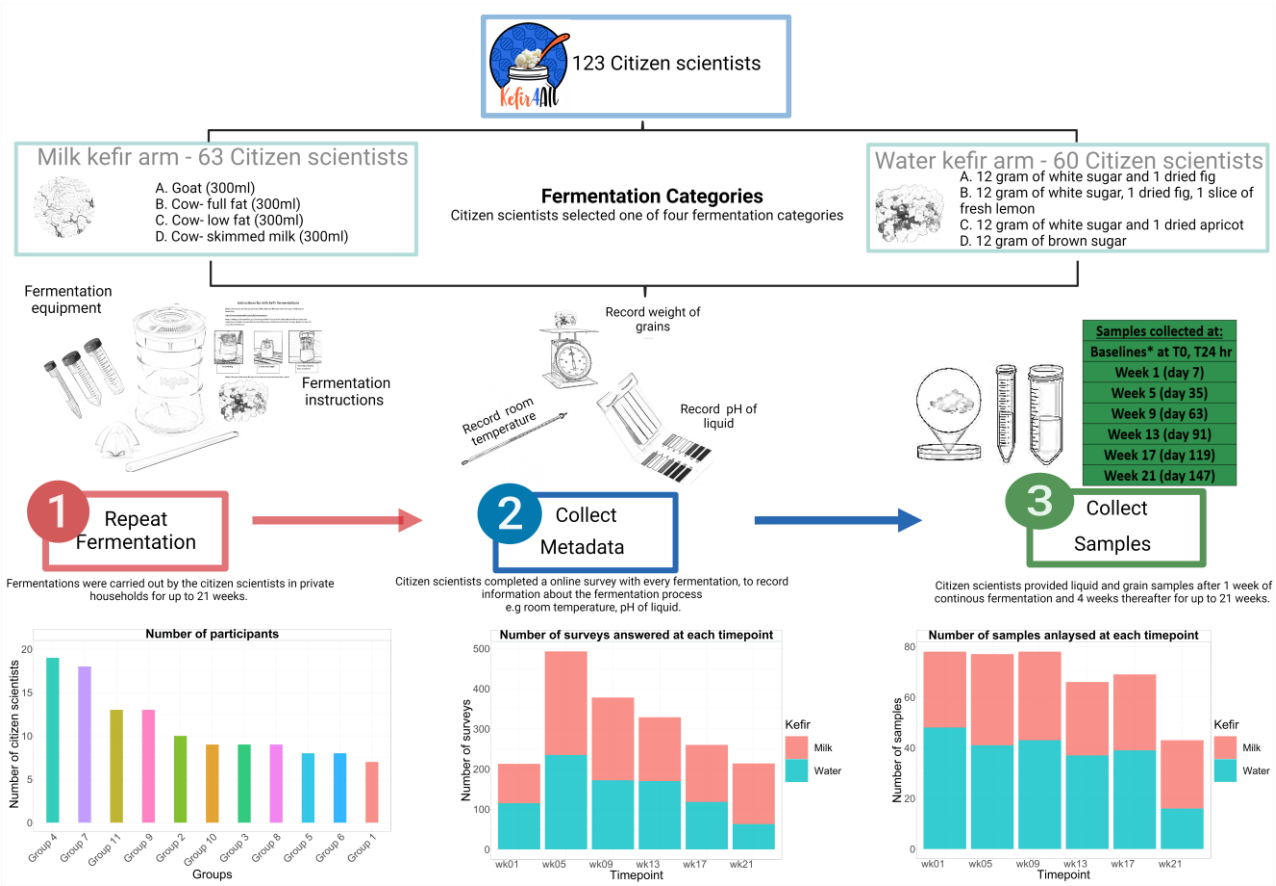


Figure 1: Graphical representation of the experimental design implemented in this citizen science initiative termed Kefir4All. Details include the total number of citizen scientists, further divided into milk and water kefir arms, the fermentation categories that the citizen scientists could choose from, and the steps involved in Kefir4All from the participants perspective. The steps involved in Kefir4All from the participants perspective were divided into (1) 'Regular

fermentation', in which citizen scientists were asked to perform regular fermentations in private households for up to 21 weeks, (2) 'Collect metadata', in which citizen scientists recorded details of milk kefir or water kefir fermentations using the Survey Monkey platform [24] and (3) 'Collect samples' during which citizen scientists were asked to provide grains and liquids samples at week (wk) 01, wk05, wk09, wk13, wk17 and wk21. T0 samples represent control milk kefir and water kefir grain and liquid samples fermented under laboratory/controlled conditions and produced by professional scientists.

Taxonomic profiling of the initial water and milk kefir grain (T0) metagenome prior to circulation to the citizen scientists revealed the presence of seven microbial species in the milk kefir grain and six in the water kefir grain at ≥ 0.1% relative abundance (RA) (Figure 2A). Microbial species detected in the milk kefir grain included *L. kefiranofaciens* (38% RA), *Lactobacillus helveticus* (15% RA), *Lentilactobacillus kefiri* (8% RA), *Lactiplantibacillus plantarum* (0.6% RA), *Lactococcus raffinolactis* (0.2% RA), *L. amylovorus* (0.2% RA) and *L. crispatus* (0.1% RA). Microbial species detected in the water kefir grain included *Len. hilgardii* (27% RA), *Lpb. plantarum* (3% RA), *Levilactobacillus brevis* (2% RA), *Lacticaseibacillus paracasei* (0.7% RA), *Zymomonas mobilis* (0.6% RA), and *Pediococcus damnosus* (0.1% RA) (Figure 2A). Total taxonomic profiling was subsequently performed on 448 kefir samples, including 96 milk kefir grains, 107 milk kefir liquids, 120 water kefir grains and 125 water kefir liquids. Metagenomes were derived from samples (kefir liquid and grains) collected at different timeframes of the project (T0, wk01, wk05, wk09, wk13, wk17 and wk21)(Figure 1).

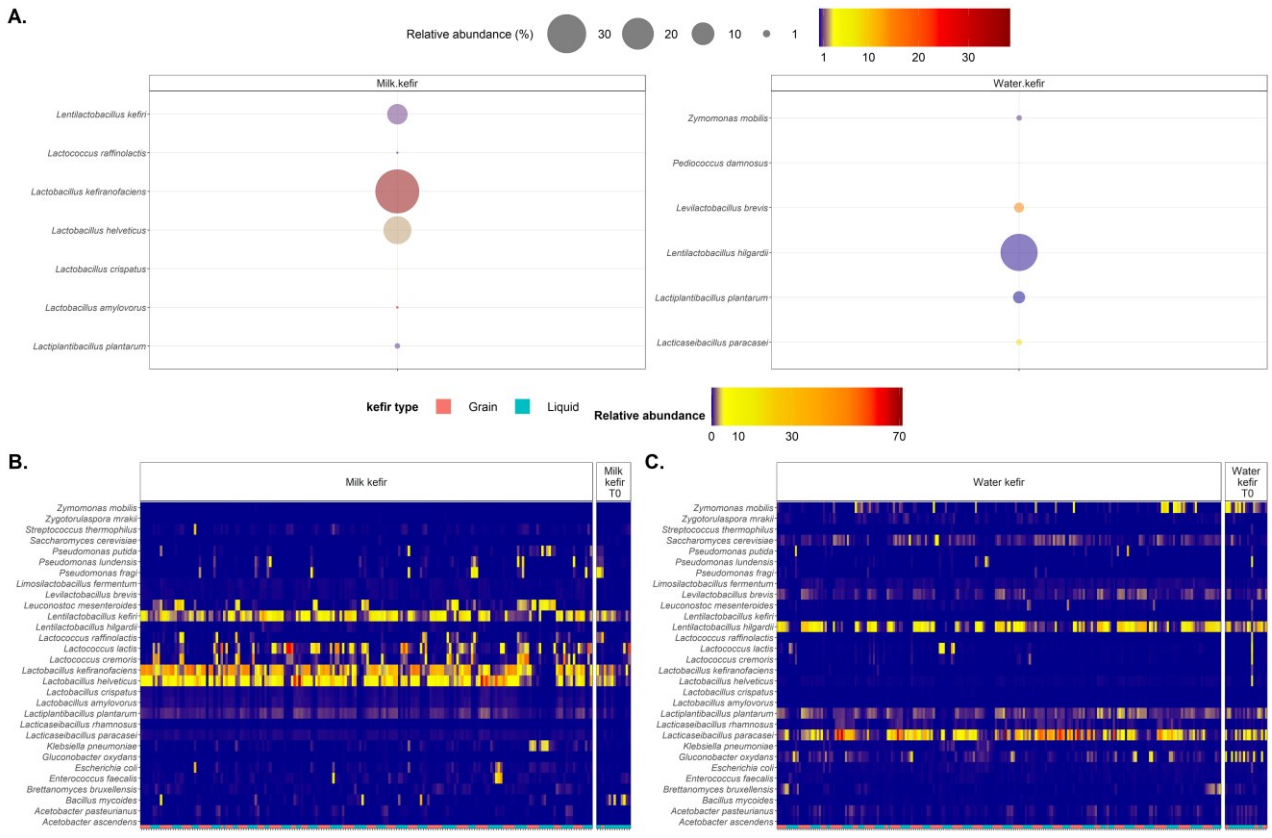


Figure 2: (A.) Bubble plot displaying the species and relative abundance (RA) of detected microbial species at a quality threshold of ≥ 0.1% relative abundance in both the T0 milk and water kefir grain metagenomes. T0 grain metagenomes, representing the initial microbiome composition of the kefir grains prior to distribution to citizen scientists. **(B.)** Heatmap displaying the prevalent microbial species detected in milk and water kefir metagenomes (liquids and grains), derived from samples collected from citizen scientists over the course of the Kefir4All project (wk01-wk21) and T0 samples. T0 samples represent the initial microbiome composition of the kefir grains and the resulting liquids produced through fermentation of these grains in a controlled laboratory setting.

Pairwise PERMANOVA established the absence of significant differences at the compositional level between kefir grain and liquid metagenomes for both milk kefir (Pairwise PERMANOVA, R^2=0.005 and p = 1) and water kefir (Pairwise PERMANOVA, R^2=0.005 and p = 0.3). Taxonomic profiling revealed 21 prevalent bacterial species ≥ 0.1% RA in ≥ 10% of samples (the defined threshold of prevalence in this study) in the milk kefir microbiome. Of these prevalent microbes, 15 could be detected at ≥ 0.1 RA in at least one water kefir metagenome, while, *Bacillus mycoides, L. amylovorus, L. crispatus, Lla. raffinolactis, Len. kefiri* and *Streptococcus thermophilus* were uniquely found in the milk kefir microbiome (Figure 2B). A number of the prevalent microbial species dominated in the milk kefir grain and liquid metagenomes. These included *L. helveticus* (14% of the total milk kefir grains and 24% of the total milk kefir liquids), *L. kefiranofaciens* (76% milk kefir grains and 10% milk kefir liquids), *Len. kefiri* (5% milk kefir grains and 6% milk kefir liquids), *Lla.*

cremoris (0% milk kefir grains and 14% milk kefir liquids), *Lla. lactis* (0% milk kefir grains and 17% milk kefir liquids) and *Leu. mesenteroides* (3% milk kefir grains and 8% milk kefir liquids) (Figure 2A). Of the prevalent species named, these were further assigned at the sub-species level to *L. kefiranofaciens* subsp. *kefiranofaciens* (n=166 of 203), *Lla. lactis* subsp. *lactis* (n=87 of 203), *Lla. cremoris* subsp. *cremoris* (n=41 of 203) and *Leu. mesenteroides* subsp. *mesenteroides* (n=31 of 203) among milk kefir metagenomes.

We identified 11 prevalent bacterial species and 2 prevalent eukaryotic species in the water kefir microbiome. Of these prevalent microbes *Gluconobacter oxydans, Lpb. rhamnosus, Lev. brevis, Zygotorulaspora mrakii* and *Z. mobilis* were uniquely found in the water kefir microbiome (Figure 2C). Similarly, the microbial species *Lc. paracasei* (37% of water kefir grains and 50% of water kefir liquids), *Len. hilgardii* (57% - water kefir grain and 16% water kefir liquids), *Z. mobilis* (3% of water kefir grains and 10% water kefir liquids) were among the dominant species in the water kefir grain and liquid metagenomes (Figure 2C). Of the prevalent species named, these were further assigned at the sub-species level to *Lc. paracasei* subsp. *tolerans* (n=103 of 245), *Z. mobilis* subsp. *mobilis* (n=70 of 245), *Lc. paracasei* subsp. *paracasei* (n=68 of 245), *Lla. lactis* subsp. *lactis* (n=9 of 245), *Lla. cremoris* subsp. *cremoris* (n=3 of 245), *Leu. mesenteroides* subsp. *mesenteroides* (n=3 of 245), *Lpb. plantarum* subsp. *plantarum* (n=2 of 245), *L. kefiranofaciens* subsp. *kefiranofaciens* (n=1 of 245) and *Z. mobilis* subsp. *mobilis* (n=70 of 245) in water kefir metagenomes.

5.3.2 Compositional analysis of kefir metagenomes suggests the majority of compositional change within the milk and water kefir grain microbiota occurs in the initial periods of the project.

Beta diversity, as calculated using the Bray Curtis, was used to measure the dissimilarity between grains and liquid samples provided by citizen scientists to the initial grains and liquids (T0) based on their compositional profile (Figure 3A and 3B) and to determine over what period/timeframe these changes occurred. Dissimilarity values were observed for all kefir grain metagenomes compared to the initial grains (T0), with more stability in the microbiota of milk kefir grains with smaller mean dissimilarity values over time compare to those observed in the microbiota of water kefir grains (Figure 3A and 3B). For the milk kefir grain

metagenomes, this ranged from 0.04 to 0.25 at wk01 and 0.1 to 0.98 at wk05 but did not change considerably thereafter (Figure 3A). Dissimilarity values for the water kefir grain metagenomes compared to the initial grains (T0) followed a similar pattern (Figure 3A), with values ranging from 0.1 to 0.9 at wk01, and 0.2 to 1 at wk05 (Figure 3B) and, again, values did not change considerably thereafter.

We further assessed the stability of both the milk and water kefir grain microbiota over time through pairwise comparisons of dissimilarity values between kefir grain metagenomes derived from the same participant. The collective individual changes within milk liquid and grain metagenomes from the same participant were not significant across the timeframes of the study, however it should be noted that the collective differences in dissimilarity values decreased after the initial periods of the study (wk01 - wk05)(Figure S1A and S1C). We observed a significant decrease in dissimilarity values in both water kefir grain and liquid metagenomes for the periods of wk01 - wk05 compared to wk05 - wk09 (Dunn test, p=0.05 and 0.02), and further to wk09 - wk013 (Dunn test, p=0.03) in the water grain metagenome and wk13-wk17 (Dunn test, p=0.02) in the water liquid metagenome. Similarly to the patterns observed in the milk kefir microbiome, higher mean dissimilarity values were observed in the initial periods of the study (wk01 - wk05) relative to the latter timeframes of the study (wk13 - wk17 and wk17 - wk21), when comparing water kefir metagenomes (Figure S1B and S1D). However, despite this general trend in the milk and water kefir microbiome a subset of kefir grain and liquid metagenomes had high dissimilarity values during the latter timeframe of the project, e.g., metagenomes from milk kefir grain samples provided by ID037 with a dissimilarity value of 0.76 for the period wk17-wk21 and metagenomes from water kefir grain samples provided by ID097 with a value of 0.89 between the grains (Figure 3C). We observed elevated levels of dissimilarity values for milk and water kefir liquid metagenomes across the time points compared to their grain counterparts. For example, we observe mean dissimilarity values of 0.85 in milk kefir and 0.84 in water kefir for the initial periods of wk01-wk05 compared to their grain counterparts, with dissimilarity values of 0.16 in milk kefir and 0.42 in water kefir (Figure S1A-D).

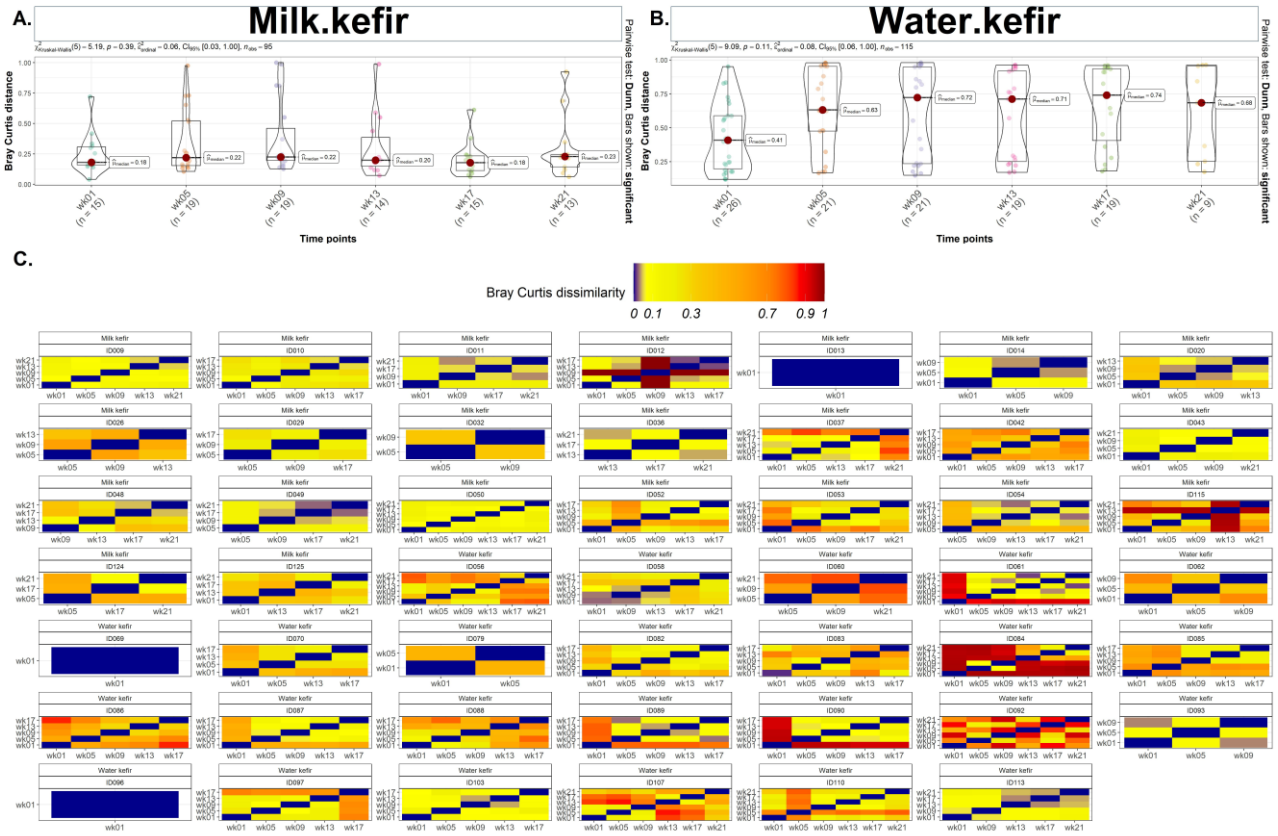


Figure 3: (A.) Beta-diversity boxplots displaying the Bray Curtis values across milk kefir grain metagenomes representing different timeframes of the project ('wk01', 'wk05, 'wk09', 'wk13' , 'wk17' and 'wk21') compared to the T0 milk kefir grain metagenome. Kruskal-Wallis and Dunn test was used to assess differences in beta diversity, since the data was nonparametric. Red dot represents the mean Bray Curtis between milk kefir grain metagenomes representing different timeframes of the project compared to the T0 milk kefir grain. **(B.)** Beta-diversity boxplots displaying the Bray Curtis values across water kefir grain metagenomes representing different timeframes of the project ('wk01', 'wk05, 'wk09', 'wk13', 'wk17' and 'wk21') compared to the T0 water kefir grain metagenome. Kruskal-Wallis and Dunn test was used to assess differences in beta diversity, since the data was nonparametric. Red dot represents the mean Bray Curtis between water kefir grain metagenomes representing different timeframes of the project compared to the T0 milk kefir grain. **(C.)** Heatmap displaying the Bray Curtis between kefir grain metagenomes from the same participant, representing different timeframes of the project.

5.3.3 Functional analysis of kefir metagenomes suggests the majority of functional change within the milk and water kefir grain microbiota occurs in the initial periods of the project.

SUPER-FOCUS [19] was used to provide an overview of the functional potential of the milk and water metagenome. As expected, a significant proportion of the milk and water kefir T0 grain metagenome and their respective liquid metagenomes were assigned to housekeeping functions such as 'Carbohydrate metabolism', 'DNA metabolism', and 'Protein metabolism'. However, SUPER-FOCUS also detected high levels of functions associated with horizontal gene transfer (HGT) and virulence in milk and water kefir. The

level 1 subsystem of 'Phages, prophages, and plasmids' was present at ≥1% average relative abundance in both milk and water kefir grains. Similarly, the level 1 subsystem of 'Virulence' was present at ≥5% average relative abundance in both groups. Dissimilarity values at the functional level were observed for all kefir grain metagenomes compared to the initial grains (T0). Dissimilarity values for the milk kefir grain metagenomes compared to the initial grains (T0) ranged from 0.03 to 0.42. Mean dissimilarity values remained similar for the remaining time points (Figure 4A). Dissimilarity values for the water kefir grain metagenomes compared to the initial grains (T0) followed a similar pattern (Figure 4B), with values ranging from 0.07 to 0.1 at wk01, and 0.12 to 0.69 at wk05 (Figure 4B). Strong positive and or negative correlations were observed between functional change and 39 different level 3 subsystems in the milk kefir grain microbiome. The level 3 subsystems were derived from 23 level 1 subsystems in the milk kefir microbiome, which included 'Cofactors, Vitamins, Prosthetic Groups, Pigments' (n=4), 'RNA Metabolism' (n=3), 'Regulation and Cell signalling (n=3), 'Virulence' (n=2) and 'Virulence, Disease and Defense' (n=2) (Figure 4C). Strong correlations both positive and negative, were observed between functional change and 159 different level 3 subsystems in the water kefir grain microbiome. The level 3 subsystems were derived from 25 level 1 subsystems in the milk kefir microbiome, many of which were assigned to the level 1 subsystem 'Carbohydrates' (n=27), 'Amino Acids and Derivatives' (n=20), and 'Clustering-based subsystems' (n=15)(Figure 4C).

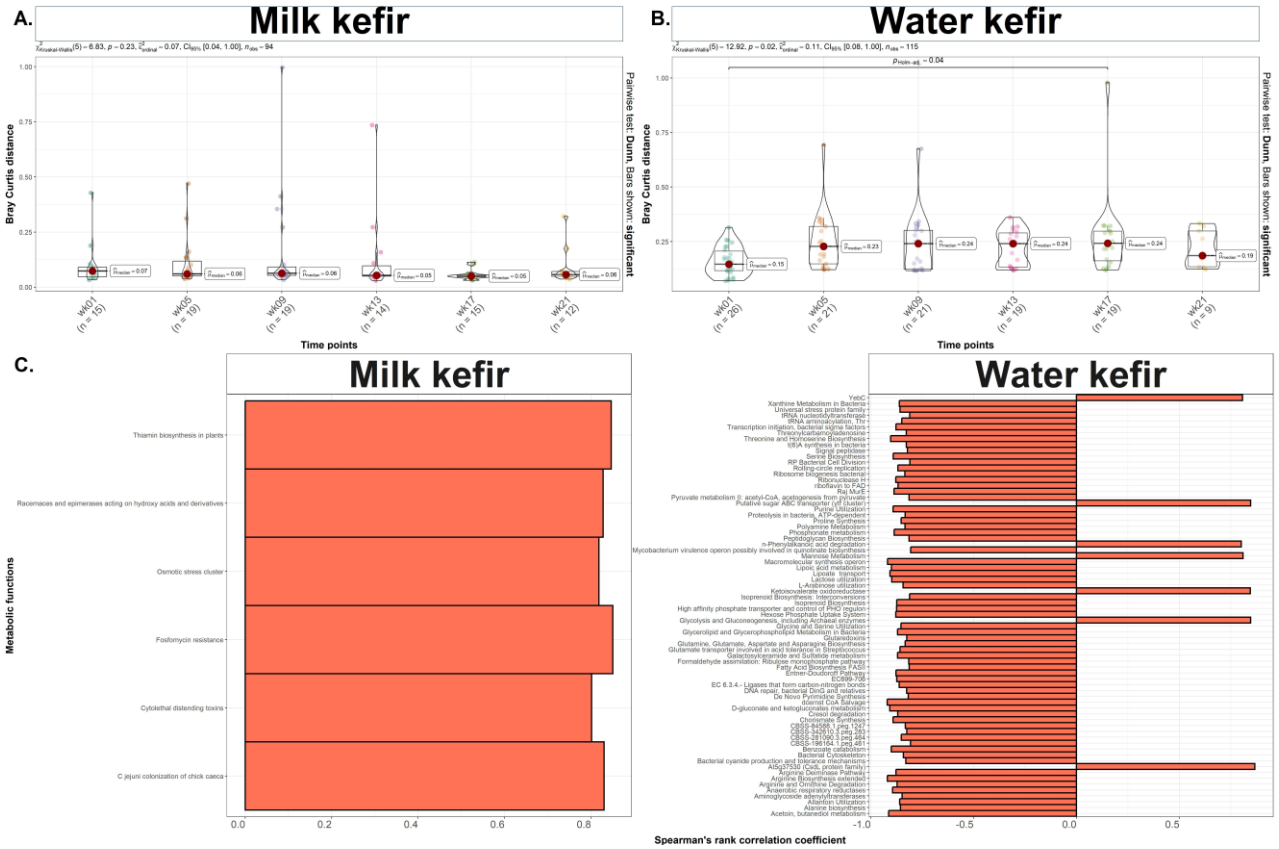


Figure 4: (A.) Beta-diversity boxplots displaying the Bray Curtis values as determined based on metabolic composition, across milk grain metagenomes representing different timeframes of the project ('wk01', 'wk05, 'wk09', 'wk13', 'wk17' and 'wk21') compared to the T0 milk kefir grain metagenome. **(B.)** Beta-diversity boxplots displaying the Bray Curtis values as determined based on functional composition, across milk grain metagenomes representing different timeframes of the project ('wk01', 'wk05, 'wk09', 'wk13', 'wk17' and 'wk21') compared to the T0 milk kefir grain metagenome. Kruskal-Wallis and Dunn test was used to assess differences in beta diversity, since the data was nonparametric. Red dot represents the mean Bray Curtis between milk kefir grain metagenomes representing different timeframes of the project compared to the T0 milk kefir grain. **(C.)** Statistically significant strong spearman correlations (-.7 to -.1 or .7 to 1) between the relative abundance of level 3 subsystems and functional change as determined using the Bray Curtis.

5.3.4 Fermentation of kefir grains with a common single community type can lead to the development of multiple distinct community types

As we identified multiple metagenomes with different dominating microbial species/strains in this study (Figure 2B), there was an opportunity to assess their similarity/dissimilarity with that of other milk and water kefir studies described in Walsh *et al* [11] and Breselge *et al* (data unpublished), respectively (Figure 5A). Specifically the microbiome composition of the 203 milk and 245 water kefir Kefir4All metagenomes were compared with that of the 256 milk provided by Walsh *et al* **[46]** and 375 water kefir microbial profiles provided by Breselge *et al* (publication in preparation). Walsh *et al* and Breselge *et al* employed a consistent, standardised fermentation technique, with differences in fermentation parameters arising because of the different grains that were employed and duration of the fermentation. Firstly, we applied ordination (Figure 5A) and statistical tests of Bray-Curtis distances at the compositional level [47] between metagenomes including Analysis of Similarity (ANOSIM) [30], Permutation Based Analysis of Variance (PERMANOVA) [36] and Beta-dispersion analyses [38], to identify significant difference in the microbial communities of milk and water kefir, respectively, according to the data source provided.

Analysis of Similarity (ANOSIM) and pairwise ANOSIM provided evidence that there were significant differences in species level microbial compositional profiles (ANOSIM, $R^2 = 0.69$ and p= 0.01), with strong, statistically significant differences in microbial communities in water kefir datasets compared to milk (Pairwise ANOSIM, n=3 , R^2 = 0.76-0.96 and p=0.01-0.01) and weaker statistical differences between datasets of the same kefir type, such as Kefir4All - Milk kefir vs Walsh *et al* - Milk kefir (Pairwise ANOSIM, R^2 = 0.46 and p= 0.01) and Breselge *et al* - Water kefir vs Kefir4All - Water kefir (Pairwise ANOSIM, $R^2 = 0.14$ and p=0.01). The PERMANOVA test identified significant differences between centroids in the data source groups, which explain 30% of the variance observed (ADONIS, p<0.01) and pairwise PERMANOVA confirmed the significant differences between centroids of each group (Pairwise ADONIS, p<0.01). The variance/dispersion, between the metagenomes within the groups, 'Kefir4All - Milk kefir', 'Kefir4All - Water kefir' metagenomes, 'Breselge *et al* - Water kefir' metagenomes and 'Walsh *et al* - Milk kefir' was assessed,

this and further comparisons were restricted to those of the same kefir type based on the output of ANOSIM and PERMANOVA. The average distance to the median ranged from 0.53 in 'Breselge *et al* - Water kefir' group to 0.43 in 'Walsh *et al* - Milk kefir' group. An ANOVA test determined significant differences between overall groups dispersion and Tukey's tests, reported significant differences in groups dispersion between 'Walsh *et al* - Milk kefir' compared to both 'Kefir4All - Milk kefir', with no significant differences between 'Breselge *et al* - Water kefir' and 'Kefir4All - Water kefir'.

We further compared the dominating microbial profiles between all metagenomes considered in this study to determine community type placement (Figure 5A). Walsh *et al* and Breselge *et al* both described kefir metagenomes clustering according to the dominant microbial community in such a pattern to suggest 4 reproducible community types in milk kefir [11] and 3 community types in water kefir. We identified 'Kefir4All - Milk kefir' metagenomes that clustered with 'Walsh *et al* - Milk kefir' metagenomes previously categorised as representing the community types dominated by *L. helveticus*, *L. kefiranofaciens*, *Lla. lactis* and *Lla. cremoris* (Figure 5B) and 'Kefir4All -Water kefir' metagenomes that clustered with 'Breselge *et al* - Water kefir' previously categorised as representing community types dominated by *Lc. paracasei*, *Len. hilgardii* and *Z. mobilis* (Figure 5B). Thus, despite using a single source of milk and water kefir grains, and common microbiomes and community types in each cases, multiple different community types emerged over time.

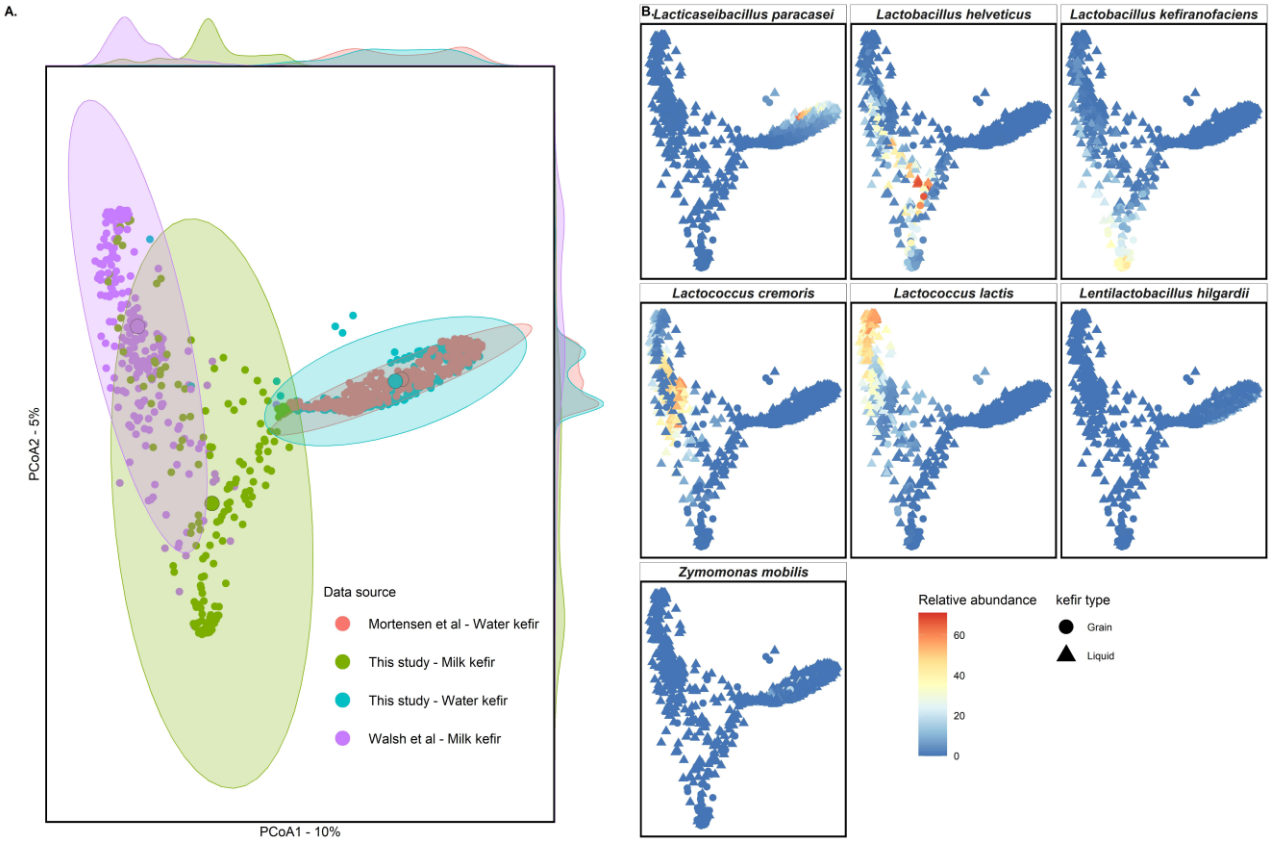


Figure 5: (A.) Principal Coordinate analysis (PCoA) plot of beta diversity measure by Bray-Curtis dissimilarity of the total metagenomes considered in this study, as calculated for species composition obtained using metacache. Metagenomes are coloured according to the datasource, ■ 'Breselge *et al* - Water kefir', ■ 'Kefir4All - Milk kefir', ■ 'Kefir4All - Water kefir' and ■ 'Walsh et al - Milk kefir'. Larger circles ● represent centroid points between each data source. Distribution, skewness, and kurtosis of metagenomes are visualised by the Kernel density estimate displayed on the X and Y-axis **(B.)** Subplots of the PCoA plot of beta diversity separated by dominant species (most abundance), *Lc. paracasei, L. helveticus, L. kefiranofaciens*, *Lla. cremoris*, *Lla. lactis, Len. hilgardii* and *Z. mobilis.* Metagenomes are coloured according to the relative abundance of the metagenome with values ranging from 0% relative abundance (RA) (Blue) to 60% RA (Red). Metagenomes are shaped according the kefir type, ● Grain and▲ Liquid.

5.3.5 Fermentation conditions may influence the compositional profile of both milk and water kefir through the introduction of contaminating microbial species

Within-sample diversity at the species level was compared across metagenomes using alpha diversity measures (Figure 6). Firstly, alpha diversity was compared across all milk and water kefir liquid and grain metagenomes considered in this study (Figure 6A). Both liquid and grain water kefir metagenomes exhibit significantly higher species alpha diversity compared to the milk liquid and grain metagenomes (Kruskal-Wallis, $p < 0.001$ and Dunn test $p < 0.001$), while the liquid metagenomes of both milk and water kefir exhibit significantly higher species alpha diversity compared to their grain counterparts (Kruskal-Wallis, $p < 0.001$

and Dunn test p <0.001) (Figure 6A). We then examined the species alpha diversity between metagenomes derived from ferments produced under laboratory controlled or household conditions. Metagenomes representing the laboratory controlled conditions were sourced from Walsh *et al* - Milk kefir metagenomes and Breselge *et al* - Water kefir metagenomes and metagenomes generated in this study representing the T0 timeframe. The remaining metagenomes generated in this study represented kefir products fermented under household conditions. Metagenomes assigned to the laboratory controlled category exhibit significantly lower species alpha diversity compared to the metagenomes assigned to the household conditions category (Kruskal-Wallis, p <0.001 and Dunn test p <0.001) (Figure 6B). As noted above, a number of prevalent microbes were identified among these metagenomes that were not detected at ≥ 0.1 RA in the initial milk and water kefir grain and liquid metagenomes, and may therefore be environmentally derived or were present at below detectable thresholds in T0 samples. Such prevalent microbes in milk kefir, which were not detected in microbiome of control samples and T0 metagenomes, included *Lc. paracasei* (n= 45), *A. pasteurianus* (n= 24), *B. mycoides* (n= 21*), Br. bruxellensis* (n= 27), *En. faecalis* (n= 26), *Klebsiella grimontii* (n= 18), *Leu. mesenteroides* (n= 56), *P. putida* (n= 26) and *S. thermophilus* (n= 23). Species in water kefir that were not detected in control, substrate, samples included *Sac. cerevisiae* (n= 140), *Lac. rhamnosus* (n= 76), *Zg. mrakii* (n= 32), *A. ascendens* (n= 29), *Br. bruxellensis* (n= 26), and *Lim. fermentum* (n= 26), *L. kefiranofaciens* (n=2). We also identified a number of putative environmentally derived microorganisms that showed patterns of persistence. Persistence was determined through the detection in kefir metagenomes representing multiple sequential time points and derived from the same participant. Such taxa included *Leu. mesenteroides, Br. bruxellensis*, *Clostridium tyrobutyricum*, *K. quasipneumoniae* in milk kefir and *Ac. ursingii*, *Pantoea. vagans* and *Lla. lactis*, in water kefir (Figure 6C).

Other potentially environmentally derived contaminating microbial species, but that did not display a pattern of persistence at ≥ 0.1% RA, included *Escherichia coli* (milk kefir (n) = 26 and water kefir (n) = 9), *Limosilactobacillus fermentum* (milk kefir (n) = 26 and water kefir (n) = 5), *C. beijerinckii* (milk kefir (n) = 18 and water kefir (n) = 5), *Leu. pseudomesenteroides* (milk kefir(n) = 8 and water kefir (n) = 13), *Leu. citreum*, (milk kefir(n) = 10 and water kefir (n) = 11), *Leclercia adecarboxylata* (milk kefir(n) = 5 and water kefir (n) =

4), which could be detected in multiple milk and water metagenomes from different participants.

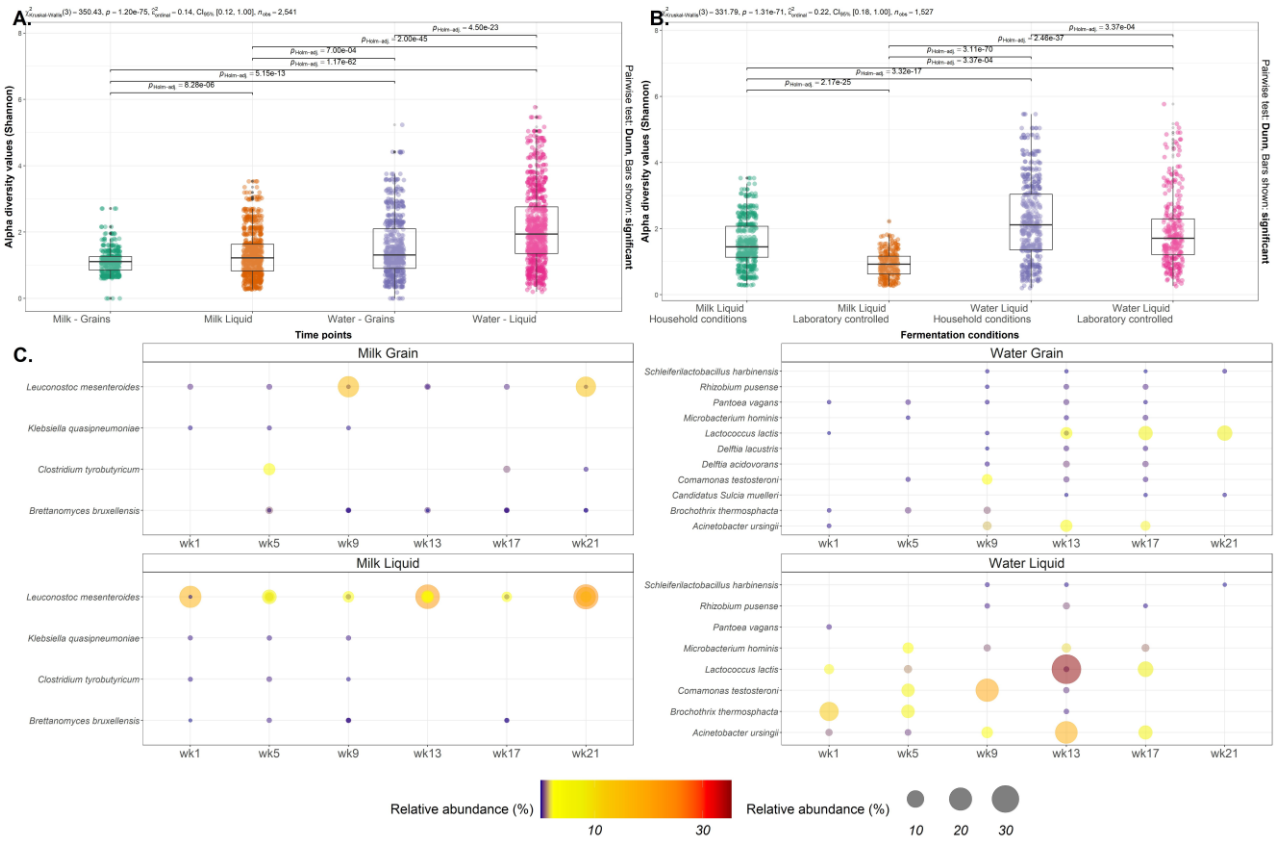


Figure 6: (A.) Alpha-diversity boxplots displaying the Shannon diversity values across kefir types ('Milk- Grains', 'Milk - liquid', 'Water -grains' and 'Water liquid'). Kruskal-Wallis and Dunn test was used to assess differences in alpha diversity between kefir metagenome types, since the data was nonparametric. Corrected p-values indicated as $p < 0.05$, *; $p < 0.01$, **; $p < 0.001$, ***; $p < 0.0001$. ****. **(B.)** Alpha-diversity boxplots displaying the Shannon diversity values across fermentation conditions (Laboratory controlled or household conditions). Kruskal-Wallis and Dunn test was used to assess differences in alpha diversity between fermentation conditions, since the data was nonparametric. **(C.)** Persistence of putative environmental microbial species (Y-axis) in either milk or water kefir grain metagenomes. Persistence was determined through detections at sequential timeframes (X-axis) in metagenomes derived from the same citizen scientists; corresponding liquid samples are also displayed. Size and colour of the bubbles represents the relative abundance of each of the putative environmental microbial species (Y-axis) at the various timeframes (wk01, wk05, wk09, wk13, wk17 and wk21).

5.3.6 Fermentation parameters, and abiotic factors are modest predictors of milk and water kefir compositional profiles

Through our timeseries data, coupled with metadata of the fermentation practises employed by each citizen scientist, it was possible to determine associations between certain fermentation parameters and changes in relative abundance of detectable microbial species. Firstly, we examined the changes in fermentation parameters over the course of the project (wk01-wk21) using the Gower distance (Figure S1E and S1F). The extent of change was determined through pairwise comparisons of dissimilarity values between surveys

collected from the same participant representing different timeframes of the project (wk01-wk21). There was significant differences in Gower dissimilarity values for all periods (Dunn test p<= 0.02), with higher dissimilarity values at the initial periods (wk01 - wk09), compared to the latter timeframes of the study (wk13 - wk17 and wk17 - wk21). However, mean dissimilarity values remained relatively similar throughout (Figure S1E and S1F). To examine the total effect of environmental parameters on microbial composition, we apply the mantel test between the Bray-Curtis dissimilarity matrix of the compositional profiles and the Gower dissimilarity matrix of the metadata collected. Metadata collected included 13 variables (see methods section 2.3). The metadata Gower distance had a significant positive relationship with the species Bray-Curtis dissimilarity distance (Mantel statistic R: 0.25, p value < 0.01). We further performed the Mantel test, restricting the metagenomes compared to both milk and water kefir liquid metagenomes, with the additional variable 'Grain community type', which describes the dominating microbial species in the milk and water kefir grain used to initiate the fermentation (see Results section 5.3.1). Notably, the metadata distance had a stronger significant positive relationship with the species Bray-Curtis dissimilarity distance (Mantel statistic R: 0.31, p value < 0.01).

To further investigate the factors that influence the observed microbial profiles of milk and water kefir, metadata covariates were modelled separately with EnvFit [33] to determine the significance and correlation strength of each covariate with overall microbial communities. 'Presence/absence of other fermented food' (R^2= 0.12, p= 0.02), 'Room temperature' (R^2= 0.12, p = 0.04), 'Grain community type' (R^2= 0.12, p = 0.01), 'Liquid volume' (R^2= 0.1, p = 0.03), 'pH' (R^2= 0.06, p = 0.03), 'Timeframe' (R^2= 0.05, p= 0.04), 'Fermentation characteristics' (R^2= 0.05, p = 0.01) and 'Fermentation categories' (R^2 =0.03, p=0.05), had a weak but significant association with the milk kefir microbiome. The 'Grain weight' (R^2= 0.32, p = 0.03), 'Room temperature' (R^2= 0.13, p = 0.02), 'Liquid volume' (R^2= 0.07, p = 0.02), 'Fermentation characteristics' (R^2= 0.06, p < 0.01), 'Water type' (R^2= 0.06, p < 0.01) and 'pH' (R^2= 0.05, p = 0.01) had a modest to weak significant association with the water kefir microbiome profile. Spearman correlation analysis between species relative abundance and the fermentation parameters 'Timeframe', 'Room temperature', 'pH', 'Liquid volume' and 'Grain weight' revealed a number of statistically significant correlations, following p

value adjustment (Figure 7A). We observed statistically significant correlations between the relative abundance of microbial species within the milk and water kefir microbiome and various fermentation parameters. Correlations between species relative abundance within the milk kefir metagenomes and fermentation parameters included 24 correlations between microbial species and the fermentation parameters 'pH', 23 correlations when considering 'Timeframe', 17 correlations when considering 'Liquid volume' or 'Grain weight', and eigth corrrelations where observed when considering 'Room temperature'. Correlations between species relative abundance within the water kefir metagenomes and fermentation parameters included 72 correlations between microbial species and the fermentation parameter 'pH' volume', a further 39 correlations when considering 'Timeframe', and one corrrelations where observed when considering 'Room temperature'. Notably no correlations were observed for the fermentation parameters 'Liquid volume' and 'Grain weight' in the water kefir microbiome, furthermore in the water kefir cohort the fermentation parameters 'Liquid volume' and 'Grain weight' contained a limited range of values as participants largely adhered to the recommended set ratios of 300ml of water to 12 grams of water kefir grains during the study. However, the 300ml of milk to 12 grams of milk kefir grains was not as readily adopted due to difficulties in grain growth throughout the project. Concerning the prevalent species, we report a negative correlation of *Lla. cremoris* in both milk and water kefir when considering 'Room temperature' and similar trends for the species *Leu. mesenteroides* (Figure 7B) and *Leu. pseudomesenteroides* (Spearman correlation r= -0.32 to -0.34, adjusted p value <=0.02 to 0.04). *Lla. lactis* (Spearman correlation r= -0.43, adjusted p value <=0.01) and *S. thermophilus* (Spearman correlation r= -0.39, adjusted p value <=0.001) in milk kefir exhibited a negative correlation of modest strength with pH, with no positive correlation between prevalent species reported and pH. We observed a number of modest strength negative correlations between prevalent species reported and pH, including *Lim. fermentum* (Spearman correlation r= -0.39, adjusted p value =0.001) in water kefir and *A. pasteurianus* (Spearman correlation r= -0.36, adjusted p value =0.02) in milk kefir, with one positive correlation between the prevalent species *Br. bruxellensis* (Spearman correlation r= 0.32, adjusted p value =0.03) in water kefir. As expected a number of putative environmental microbes were positively correlated with 'pH' such as *B.*

thuringiensis (Spearman correlation r= 0.32, adjusted p value <=0.001) and *B. cereus* (Spearman correlation r= 0.38, adjusted p value =0.01) in milk kefir and *Pa. vagans* (Spearman correlation r= 0.34, adjusted p value =0.01) and *Rahnella aquatilis* (Spearman correlation r= 0.29, adjusted p value =0.04) in water kefir. Similarly many of the putative environmental microbes were negatively correlated with 'Timeframe' (wk01-wk21), including *B. thuringiensis* in milk kefir and *R. aquatilis* (Spearman correlation r= -0.27 to -0.28, adjusted p value <=0.01) in water kefir. Notably no negative correlations were observed for the prevalent species in milk kefir, however a number of modest strength positive correlations we observed for the species *L. crispatus*, *L. amylovorus* and *Lac. plantarum* for both 'Timeframe' and 'Grain weight' (Spearman correlation r= 0.28 to 0.4, adjusted p value <=0.001)(Figure 7A). Notably no positive correlation was observed for any of the prevalent species within the water kefir microbiome, when considering both 'Timeframe' and 'Grain weight', however, several negative correlations were observed for the species *Z. mobilis*, *Zg. mrakii* and *Br. bruxellensis* (Spearman correlation r= -0.27 to -0.59, adjusted p value <=0.001)(Figure 7A), when considering both 'Timeframe' and 'Grain weight'.

We further note the presence of differentially abundant species in a number of fermentation parameters through LEfSe analysis (Figure 7B). Of the putative environmental microbes *Lc. paracasei*, *Br. bruxellensis*, *A. pasteurianus, C. tyrobutyricum*, *Len. hilgardii* detected in the milk kefir microbiome and *Lla. lactis* detected in the water kefir microbiome were differentially abundant with higher abundances when considering the presence of pets. *K. grimontii* and *P. putida*, *K. quasipneumoniae* detected in the milk kefir microbiome and *Leu. gelidum* detected in the water kefir microbiome were differentially abundant with higher abundances on average in participants who ferment other fermented foods (Figure 7B). While sample size was insufficient to determine if the presence of *L. kefiranofaciens* was associated with any fermentation categories, we did observe its presence in a water kefir liquid sample provided by a participant that also fermented injera. *C. tyrobutyricum* - Cow milk (low fat), *C. beijerinckii* - Cow milk (low fat), *K. grimontii* - Cow milk (low fat), *En. Faecalis* - Goat, *Leu. mesenteroides* - Cow milk (low fat) *and P. putida* - Cow milk (low fat), detected in the milk kefir microbiome, and *Lla. lactis* - White sugar-dried apricot and *Lim. fermentum* - White sugar-dried fig, detected in the water kefir microbiome, were differentially

abundant with higher abundances on average when fermented under the outlined categories (Figure 7B).

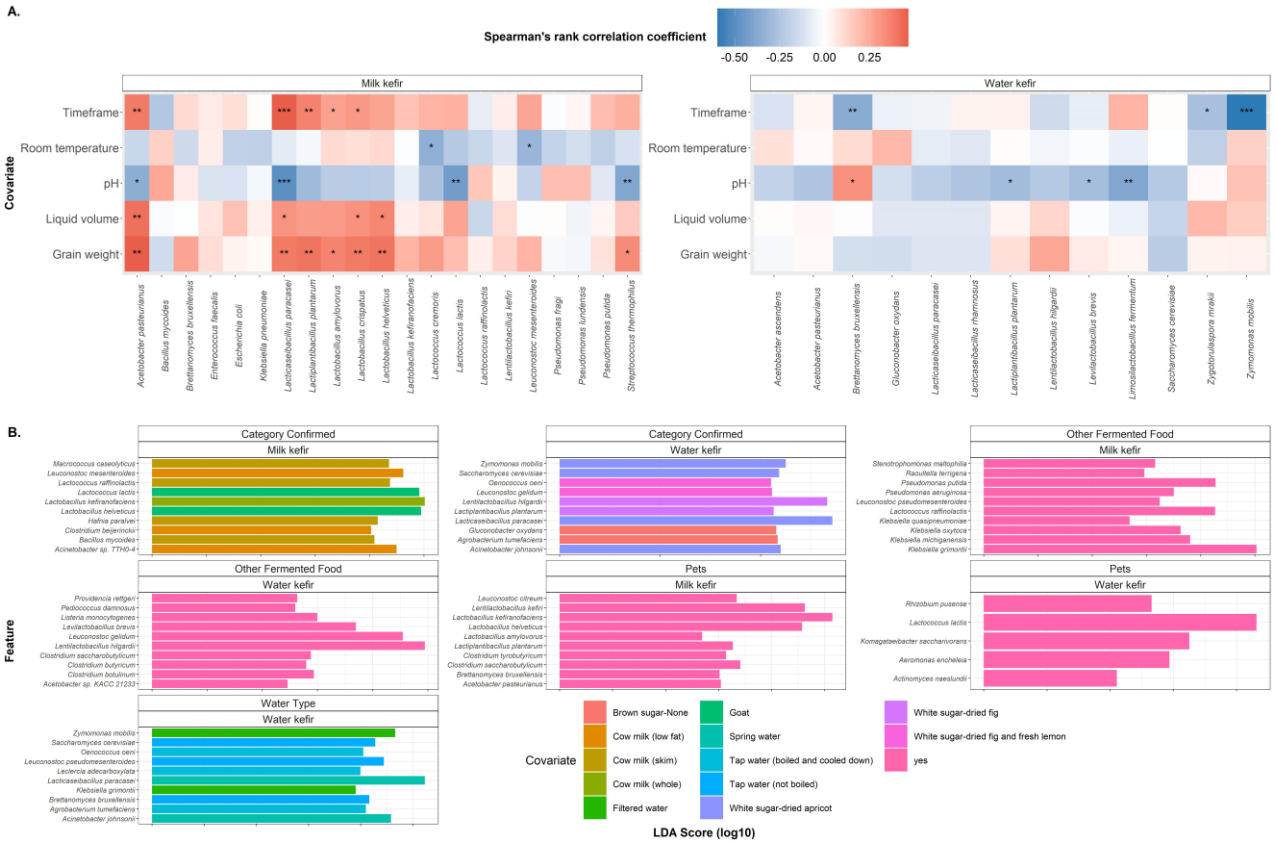


Figure 7: (A.) Heatmap displaying the Spearman rank correlation coefficient between the fermentation parameters 'Total fermentation response frequency', 'Room temperature', 'Grain weight', Liquid volume and 'pH' and the prevalent microbial species (X-axis) detected in milk and water kefir metagenomes (liquids and grains), derived from samples collected from citizen scientists over the course of the Kefir4All project (wk01-wk21) and T0 samples. **(B.)** Lefse covariate analysis reveals the effect size (LDA score) (X-axis) of the fermentation parameters 'Fermentation category', 'Other fermented food', 'Pets' and 'Water type' on the relative abundance of the outlined species (Y-axis), separated according to the kefir type - milk kefir or water kefir, and coloured according to the various options available to each fermentation parameter. For example If the citizen scientists selected milk kefir, they had the option to ferment in Cow milk (whole), Cow milk (low fat), Cow milk (skim) and Goat milk.

5.3.7 Phylogenetic analysis of kefir metagenomes suggests the majority of evolutionary change for most representative species occurs in the initial periods of the project

StrainPhlAn4 [20] profiling of all kefir metagenomes considered in this study detected strains representative of 89 species. Species with the highest number of strains included *Sac. Cerevisiae* (n=461), *Lc. paracasei* (n=460), *L. satsumensis* (n=447) and *Len. hilgardii* (n=370) in water kefir and *L. kefiranofaciens* (n=317), *Len. kefiri* (n=308), *Lla. lactis* subcluster 2 (n=308), and *L. helveticus* in milk kefir (n=248)(Figure 8A and 8B). As

described above, we identified the presence of 2 sub species of *Lc. paracasei*, which can occur in the same water kefir metagenome, and the possible presence of additional subspecies. Therefore, we screened for the presence of multiple strains within the kefir microbiome that would influence downstream results. We initially examined the number of polymorphic sites across alignments to determine strain diversity within the milk and water kefir microbiome (Figure S2) [48]. Polymorphic sites were detected on marker genes for all examined species across the datasets, indicating the presence of multiple strains within the milk and water kefir microbiome. The mean number of polymorphic sites across marker genes for all examined species ranged from 0.04% to 1.9% (Figure S2), suggesting low strain diversity within metagenomes. A number of species specific marker genes derived from metagenomes generated in this study contained higher numbers of polymorphic sites such as *Bacillus cereus* (0.05% to 7.1%), *Glu. oxydans* (0% to 5.8%) and *A. malorum* (0% to 4.6%) (Figure S2), however these were restricted to a limited number of metagenomes and overall only 2% of clade specific marker genes separated across 38 different species contained ≥ 2% polymorphic sites. These combined results suggest the dominant strain accounts for the majority of the strain level diversity for the detectable species within the milk and water kefir metagenomes. We initially measured the distance between a pair of strains of the same species based on their phylogenetic distances. Strain comparisons between strains derived from kefir samples provided by the same participant in the case of the citizen science dataset or derived from the same kefir grain in the case of the global datasets were deemed 'Intra-participant comparisons', comparisons between strains sourced from the citizen science dataset but provided by different participants were deemed 'Inter-participant comparisons' and lastly strain comparisons between datasets or from different kefir grains were termed as 'Unrelated comparisons' (Figure S3). On average, the mean patristic distance (i.e., branch length) was significantly less in 'Intra-participant comparisons', than in 'Inter-participant comparisons' which in turn was also significantly less than 'Unrelated comparisons'. The strains of *Leu. mesenteroides* in water kefir and *Lla. lactis* subcluster 1 in milk kefir were notable exceptions to this general trend (Figure S3).

We then compared nucleotide variation obtained using the --mutation_rates parameter of StrainPhlAn4 in a pairwise manner for each of the recovered strains to a representative strain recovered from the earliest

possible timeframe of the study. The output of the --mutation_rates parameter represent the number of single nucleotide polymorphisms (SNPs) between strains normalised by alignment length of the MSA of all strains of the outlined species. In situations where strains were recovered from metagenomes derived from the same timeframe, the metagenome containing the highest RA for the assigned strain was selected as determined using MetaPhlAn4 output of StrainPhlAn4 [20]. Furthermore, where representative strains were compared with multiple strains recovered at the same timeframe, such as the detection of strains in both liquid and grain samples, the strain with the lowest nucleotide variation was considered. The rate of nucleotide varation ranged from 0 to 0.14 (Figure 8B).

Strains of the species *Sacc. cerevisiae* and *Z. mobilis* subcluster 1 in water kefir presented with the highest divergence from the initial strain as assessed through nucleotide change during the wk0-wk01 period (Figure 8B). While *Leu. pseudomesenteroides* in milk kefir and *Glu. oxydans* in water kefir showed the highest rates of nucleotide after wk21. Species with the lowest nucleotide included the exopolysaccharide producers *L. kefiranofaciens* and *Len. hilgardii* in milk and water kefir, respectively (Figure 8B), though several strains detected across the timeseries dataset had a nuceltide change rate of 0 when compared to their ancestral strain. In water kefir, this was most commonly observed for strains of the species *B. aquikefiri* (84%), *Len. hilgardii* (83%), *O. oeni* (75%) and in milk kefir , this was most commonly observed in strains of the strains of the species, *L. helveticus* (75%), *A. fabarum* (63%), and *Lactococcus lactis* subcluster 1 (60%). We further examined MAGs assembled from the metagenomics data generated in Kefir4All to determine the presence of same strains or strain divergence (Figure 8C) over time using the genome identity of ≥ 0.99 and ANI and the coverage thresholds of ≥ 90% genome coverage, recommended by Feehily *et al* 2023 [49]. Here we confirm the presence of same strains for a number of prevalent species including, *A.fabarum* (97% of strain comparisons), *A. indonesiensis* (54% of strain comparisons), *A. malorum* (89% of strain comparisons), *A. orientalis* (73% of strain comparisons), *A. persici* (38% of strain comparisons), *A. syzygii* (94% of strain comparisons), *Acinetobacter johnsonii* (10% of strain comparisons), *Bi. aquikefiri* (99 % of strain comparisons), *Glu. albidus* (53% of strain comparisons), *Glu. cerinus* (12.4% of strain comparisons), *Kom. hansenii* (94.3% of strain comparisons), *Lc. paracasei* (53.4% of strain comparisons), *L. helveticus*

(26.7% of strain comparisons), *L. kefiranofaciens* (47.1% of strain comparisons), *Lla. lactis* subcluster 1 (30% of strain comparisons), *Lla. lactis* subcluster 2 (40% of strain comparisons), *Lc. raffinolactis* (35.1% of strain comparisons), *Len. hilgardii* (99.8% of strain comparisons), *Len. kefiri* (99.9% of strain comparisons) and *Leu. mesenteroides* (35.1% of strain comparisons)(Figure 8C).

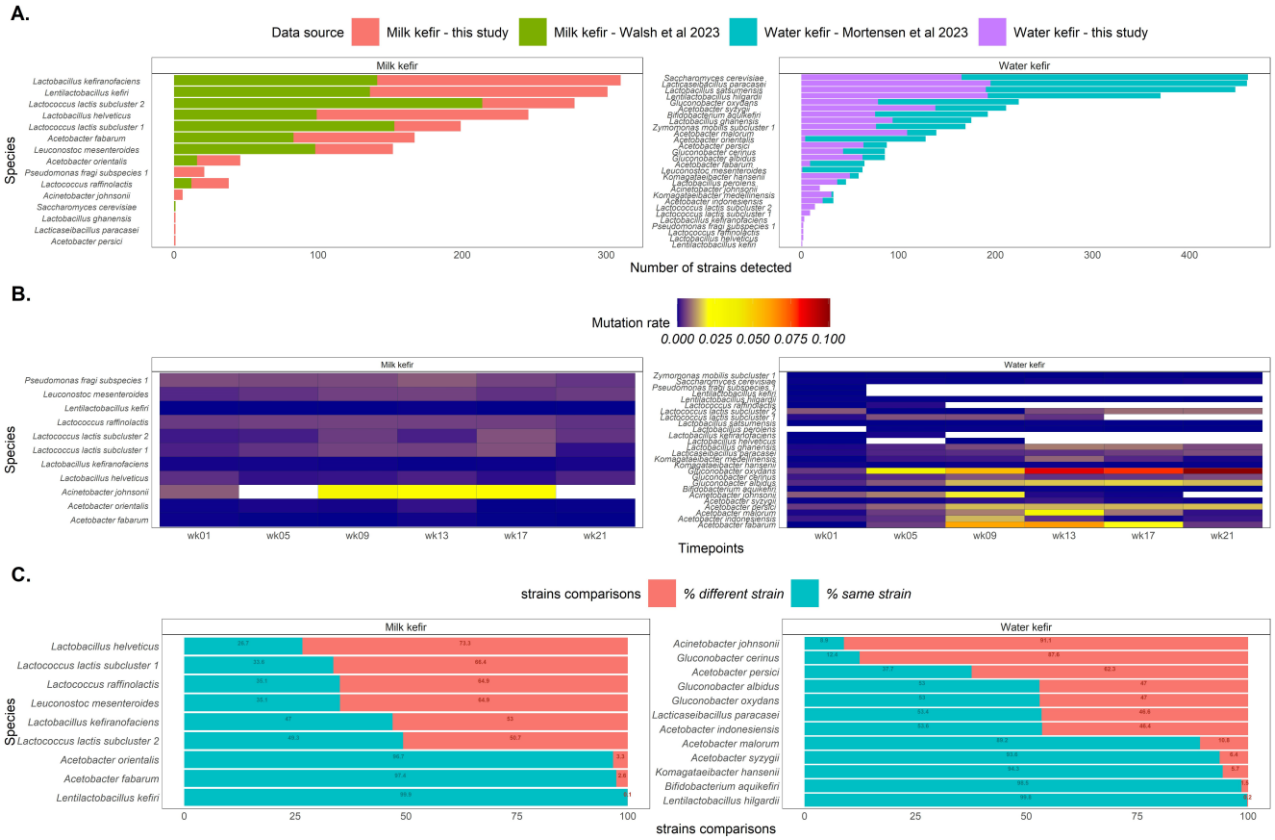


Figure 8: (A.) Bar chart displaying the number (X-axis), of strains detected (Y-axis), separated according to the kefir type - milk kefir or water kefir, and coloured according to the data source; ■ 'Milk kefir - Kefir4All ', ■ 'Water kefir - Kefir4All ', ■'Water kefir - Breselge *et al.*' and ■ 'Milk kefir - Walsh *et al.*'. **(B.)** Heatmap displaying the mean rate of nucleotide change mutation for prevalent microbial species detected in milk and water kefir metagenomes (liquids and grains), derived from samples collected from citizen scientists over the course of the Kefir4All project (wk01-wk21) compared to T0 samples. T0 samples represent the initial microbiome composition of the kefir grains and the resulting liquids produced through fermentation of these grains in a controlled laboratory setting. **(C.)** Bar chart displaying the percentages of metagenome assembled genomes (MAGs) (X-axis) for each of the prevalent species (Y- axis) deemed to be the 'same strains' vs 'different strain'.

Generally, low-abundance species had higher rates of nucledtide change compared to their high-abundance counterparts (Figure S4). We observed significant negative correlations, following p value adjustments, between relative abundance and rate of nucleotide change in strains derived from milk kefir grains and liquid samples for the species *L. helveticus* (R= -0.96, p<0.001 and R= -0.63, p<0.001) and *L. kefiranofaciens* (R= -0.39, p<0.001 and R= -0.59, p<0.001) and further negative correlations between strains of the species

Len. kefiri (R= -0.56, p<0.001) derived from milk kefir grain samples. Similarly, significant negative correlations were observed in strains derived from both the water kefir liquid and grain samples for the species *A. malorum* (R= -0.61, p<0.001 and R= -0.66, p<0.001)*, Len. hilgardii* (R= -0.57, p<0.001 and R= -0.59, p<0.001)*, L. satsumensis* (R= -0.48, p <0.001 and R= -0.57, p<0.001) and *Sacc. cerevisiae* (R= -0.65, p<0.001 and R= -0.66, p<0.001). Further negative correlations were observed between strains derived from water kefir liquid of the species *Z. mobilis* subcluster 1 (R= -0.83, p<0.001), *A. syzygii* (R= -0.59, p< 0.001) and *Komagataeibacter hansenii* (R= -0.65, p<0.001). Finally, positive correlations between relative abundance and rates of nucleotide change were observed in both the water kefir liquid and grain metagenomes for strains of the species *Glu. oxydans* (R= 0.54, p<0.001 and R= 0.71, p<0.001 (Figure S4). We further examined other variables that may affect the rate of nucleotide change including 'Fermentation category' and 'Timeframe'. We observed that strains of the same fermentation category did not cluster together as assessed using ordination and statistical testing when considering evolutionary distance.

Section 5.4 Discussion

In this study, we applied metagenomic-based identification and characterisation of the microorganisms present in milk and water kefir metagenomes. We utilised existing kefir metagenomics data from global sampling initiatives by 'Walsh *et al* - Milk kefir' [11] and 'Breselge *et al* - Water kefir' (Data unpublished) in which 256 milk and 375 water kefir metagenomes were generated from the fermentation of different kefir grains, sourced from different geographical locations. We further generated 203 milk kefir metagenomes and 246 water kefir metagenomes, provided by citizen scientists over the course of 21 weeks of regular fermentation (Figure 1). Firstly, through ordination (Figure 5), statistical testing and community type placement (Figure 5B) (see Result 5.3.4), we clearly show that there are significant differences in the microbial composition at the species and strain level between milk and water kefir Correlations including the relative abundance of 17 species (Figure 2, 5 and 8A). Different community types driven by the dominant microbial species were observed in the milk kefir grain and liquid metagenomes, including *L. helveticus*, *L. kefiranofaciens, Len. kefiri, Lla. cremoris*, *Lla. lactis* and *Leu. mesenteroides* (Figure 2B and 5B), while the

microbial species *Lc. paracasei*, *Len. hilgardii* and *Z. mobilis*, tended to dominate in the water kefir grain and liquid metagenomes (Figure 5C). We identify that compositional profiles generated in this study, and therefore produced by the actions of citizen scientists, mirrored the diversity of the global initiatives when comparing the same kefir type (Milk or water kefir) through the identification of all outlined community types within the citizen science dataset (Figure 5B) [11]. The ANOSIM statistic test provided further evidence of this similarity with an R-value of 0.14 between metagenomes generthis R vated in 'Breselge *et al* - Water kefir' and 'Kefir4All -Water kefir. This R-value value suggests the data sources are similar with some statistically significant differences, and beta dispersion tests further reported no significant differences in variance/dispersion between these two data sources. The metagenomes generated in 'Walsh *et al* - Milk kefir' and 'Kefir4All - Milk kefir had an ANOSIM R value of 0.46 suggesting different data sources with some similarities, which was further captured through the community type analysis and significant differences in variance/dispersion between data sources (Figure 5A). Such results suggest that data sources were significantly different from each other and importantly metagenomes from each data source analysed in this study can provide novel information to extend the current scientific knowledge concerning the milk and water kefir microbiome. For further details concerning the microbiome of milk and water kefir derived from different geographical sources, please refer to Walsh *et al* [46] and Breselge *et al*.

Milk and water kefir metagenomes, generated in this study were analysed primarily to elucidate the extent of compositional (Figure 2 , 3 and S1A-D), functional (Figure 4) and evolutionary change (Figure 8) in the microbiota of milk and water kefir over 21 weeks of fermentation, in separate households and under different conditions. We monitored the changes in the microbial profile of the initial milk and water kefir grains (Figure 2A) over time (Figure 3). Through pairwise comparisons based on dissimilarity values at the compositional and functional level, we observed dissimilarity values for all timeframes, with the greatest change occurring during the initial periods (wk01-wk09) of the project, with little change on average in dissimilarity values thereafter (wk13-wk21) (Figure 3A, 3B, 4 and S1C and S1D). Similarly, comparisons of both milk and water kefir grain and liquid metagenomes provided by the same participants showed a similar pattern where dissimilarity values peaked at wk05 in milk kefir grain metagenomes or wk09 in the case of

the water kefir grain metagenomes (Figure 3A and 3B). Such results further indicate that the greatest period of compositional (Figure 3 and S1) and functional change (Figure 4) occurs at the initial timeframe of the project, which coincides with community type changes (Figure 2) observed. Specifically, through repeat fermentations, milk kefir grains initially dominated by *L. kefiranofaciens* (Figure 2A) transitioned into the different community types identified among 'Walsh *et al* - Milk kefir' and categorised as representing the *L. helveticus*, *L. kefiranofaciens*, *Lla. lactis* and *Lla. cremoris* community type (Figure 2B and 5B) while water kefir grains dominated by *Len. hilgardii* (Figure 2A) transitioned to community types identified among 'Breselge *et al* - Water kefir' metagenomes categorised as representing the *Lc. paracasei* and *Len. hilgardii* community types (Figure 2C and 5B). This is a further reflection of the fact that the milk and water kefir grain microbiome is not stable and can change drastically over time (Figure 2, 3A, 3B and S1), highlighting the need for a pitched alternative in instances where manufacture of a uniform product, with a specific microbial composition or functionalities, is required [50].

We performed numerous statistical testing, including correlation analysis and differential abundance, to determine the various fermentation parameters influencing the compositional profile of both milk and water kefir grain and liquid metagenomes. Using the Mantel test, we reported that the associated metadata, including 'Grain community type', 'pH, 'Timeframe' and 'Fermentation category' (Result: 3.6 for more details) collected in this project were moderately correlated with compositional differences between metagenomes expressed using Bray Curtis. A number of fermentation parameters correlated with increased or decreased relative abundance of certain microorganisms such as the positive correlation of *L. helveticus* with 'Grain weight' (Spearman correlation r= 0.40, adjusted p value <=0.01), which was also more abundant on average in the Goat milk 'fermentation category' (LEfSe LDA= 2.9, adjusted p value =0.05) (Figure 7). Notably, none of the fermentation parameters had a defining effect on the community type observed, as primarily determined using Envfit. Overall, these combined findings suggest an accumulative effect from the fermentation parameters in defining the compositional profile of milk and water kefir. Of note, we provide evidence to suggest the possibility of additional variables having an effect on the community type of milk and water kefir, variables could include the seasonality of the fermentation medium, the presence

of bacteriophage and or microbial interactions, which were unexplored in this study. In general, water kefir is more diverse in terms of species diversity compared to milk kefir (Krustal Wallis, p< 0.01) (Figure 6A) and the microbial profile of both milk and water kefir may be influenced by the household environment (Figure 7C and 7B). Both milk and water kefir metagenomes generated under household conditions have significantly higher alpha diversity values compared to laboratory controlled metagenomes (Kruskal Wallis, p <001 and Dunn test p < 0.001) (Figure 6B). As the majority of the laboratory controlled metagenomes were generated in Walsh *et al* and Breselge *et al*, and were produced through the use of different kefir grains, we speculated that we would observe higher alpha diversity values in those instances. Why this did not occur may relate to the potential exposure to environmental microbes in the case of kefir samples provided by citizen scientists. It is particularly noteworthy that household-produced metagenomes contained a number of microbial species that were not detectable in the initial microbial composition of the milk and water kefir microbiome. Of these microbes, those detected in the majority of liquid samples, such as *Leu. mesenteroides* (n= 56) in the milk kefir metagenomes and *Sac. cerevisiae* (n= 140), *Glu. oxydans* (n= 108), *Lac. rhamnosus* (n= 76) and *A. pasteurianus* (n= 56) in the water kefir metagenomes (Figure 2B), are commonly identified in the associated kefir fermentations and may simply have been present initially at below detection cut-offs. However, several other species such as *Lla. lactis* in water kefir and *E. coli* and *K. grimontii*, in both milk and water kefir, may represent environmentally derived contaminating microbial species, raising, in the latter instances, food safety concerns and highlighting the importance of good hygiene practices in preparing both milk and water kefir. The detection and persistence of *Lla. lactis* using both species and strain detection tools in the water kefir microbiome is particularly notable given that this species was not detected in both T0 metagenomes and metagenomes derived from the global kefir study, in which fermentations were performed in a controlled laboratory setting. The linkage of some of these putative environmental microbes with the fermentation parameters 'Presence/absence of pets', 'Presence/absence of other fermented food' and 'Fermentation category' (Result section 3.6), provides further evidence that some of these microbes may be environmentally derived, as well as giving an insight as to their source. Indeed, in one instance, a connection between the presence of *Lla. lactis* and the water

kefir category 'White sugar-dried apricot' suggests apricot as a potential source.

We document evolutionary changes in a number of prevalent species within the milk and water kefir microbiome (Figure 8), highlighting strains such as exopolysaccharide producers with lower rates of nucleotide change and higher rates of strain retention over time. Examples of strains with lower rates of nucleotide change and higher rates of strain retention over time included *Len. kefiri in* milk kefir and *Len. hilgardii* in water kefir. Other strains with higher rates of nucleotide change and few cases of strain retention included *Glu. cerinus* in water kefir (Figure 8B).

Bacteria have previously demonstrated elevated rates of nucleotide change during adaptation to challenging environments [51]. In this study, we record the greatest period of nucleotide change occurs within the early timeframes of the project, followed by a stasis in the rates of nucleotide change for the majority of studied microorganisms (Figure 8). Previous research efforts in the gut microbiome provide evidence that low abundant microbes may be under increased selective pressure compare to high abundant microbes [52]. Here we provide evidence to suggest that low abundant microbes in the milk and water kefir microbiome are also under increased selective pressure compare to high abundant microbes. Through phylogenetic analysis we provided evidence of strain stability, where on average strains present in samples sourced for a single individual than those from a different person (Figure S3). Our combined results suggest that milk and water kefir microbial communities can undergo a period of change at the compositional, functional and strain level (Figure 2, Figure 4 and Figure 8) as they adapt to external pressures, where on average they remain relatively stable thereafter. We note periods of change can result in changes in community subtype as represented by the dominant species present. Such results should be taken into consideration when utilising milk and or water kefir as a model systems [6] or in commercial production as deviations at the community and strain level can occur after a single week of fermenting under different conditions. Ultimately, our combined findings provide a new insight into how microbial ecosystems change and evolve over successive fermentations.

Section 5.5 References

1. Calatayud, M., et al., *Water Kefir and Derived Pasteurized Beverages Modulate Gut Microbiota, Intestinal Permeability and Cytokine Production In Vitro.* Nutrients, 2021. **13**(11).

2. Nielsen, B., G.C. Gürakan, and G. Unlü, *Kefir: a multifaceted fermented dairy product.* Probiotics Antimicrob Proteins, 2014. **6**(3-4): p. 123-35.

3. Verce, M., L. De Vuyst, and S. Weckx, *Shotgun Metagenomics of a Water Kefir Fermentation Ecosystem Reveals a Novel Oenococcus Species.* Frontiers in Microbiology, 2019. **10**.

4. Bourrie, B.C., B.P. Willing, and P.D. Cotter, *The Microbiota and Health Promoting Characteristics of the Fermented Beverage Kefir.* Front Microbiol, 2016. **7**: p. 647.

5. Laureys, D. and L. De Vuyst, *The water kefir grain inoculum determines the characteristics of the resulting water kefir fermentation process.* Journal of Applied Microbiology, 2017. **122**(3): p. 719-732.

6. Blasche, S., et al., *Metabolic cooperation and spatiotemporal niche partitioning in a kefir microbial community.* Nat Microbiol, 2021. **6**(2): p. 196-208.

7. Patel, S.H., et al., *A temporal view of the water kefir microbiota and flavour attributes.* Innovative Food Science & Emerging Technologies, 2022. **80**: p. 103084.

8. Walsh, A.M., et al., *Microbial Succession and Flavor Production in the Fermented Dairy Beverage Kefir.* mSystems, 2016. **1**(5).

9. Leech, J., et al., *Fermented-Food Metagenomics Reveals Substrate-Associated Differences in Taxonomy and Health-Associated and Antibiotic Resistance Determinants.* mSystems, 2020. **5**(6).

10. Cotter, P., et al., *Comparison of the relative impacts of acute consumption of an inulin-enriched diet, milk kefir or a commercial probiotic product on the human gut microbiome and metabolome*. 2023.

11. Cotter, P., et al., *Analysis of the Milk Kefir Pan-Metagenome Reveals Four Community Types, Core Species and Associated Metabolic Pathways*. 2023.

12. Walsh, A.M., et al., *Species classifier choice is a key consideration when analysing low-complexity food microbiome data.* Microbiome, 2018. **6**(1): p. 1-15.

13. Olm, M.R., et al., *inStrain profiles population microdiversity from metagenomic data and sensitively detects shared microbial strains.* Nature Biotechnology, 2021. **39**(6): p. 727-736.

14. Alekseeva, A.Y., et al., *Eco-Evolutionary Dynamics in Microbial Communities from Spontaneous Fermented Foods.* Int J Environ Res Public Health, 2021. **18**(19).

15. Walsh, L.H., et al., *Bioinformatic approaches for studying the microbiome of fermented food.* Crit Rev Microbiol, 2022: p. 1-33.

16. Andrews, S., et al., *Trim Galore.* Trim Galore, 2015.

17. Langmead, B. and S.L. Salzberg, *Fast gapped-read alignment with Bowtie 2.* Nature methods, 2012. **9**(4): p. 357-359.

18. Müller, A., et al., *MetaCache: context-aware classification of metagenomic reads using minhashing.* Bioinformatics, 2017. **33**(23): p. 3740-3748.

19. Silva, G.G., et al., *SUPER-FOCUS: a tool for agile functional analysis of shotgun metagenomic data.* Bioinformatics, 2016. **32**(3): p. 354-61.

20. Beghini, F., et al., *Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3.* eLife, 2021. **10**: p. e65088.

21. Nurk, S., et al., *metaSPAdes: a new versatile metagenomic assembler.* Genome Res, 2017. **27**(5): p. 824-834.

22. Antipov, D., et al., *MetaviralSPAdes: assembly of viruses from metagenomic data.* Bioinformatics, 2020. **36**(14): p. 4126-4129.

23. Antipov, D., et al., *Plasmid detection and assembly in genomic and metagenomic data sets.* Genome Res, 2019. **29**(6): p. 961-968.

24. Uritskiy, G.V., J. DiRuggiero, and J. Taylor, *MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis.* Microbiome, 2018. **6**(1): p. 158.

25. Chklovski, A., et al., *CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning.* bioRxiv, 2022: p. 2022.07. 11.499243.

26. Manni, M., et al., *BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes.* Mol Biol Evol, 2021. **38**(10): p. 4647-4654.

27. Dröge, J., I. Gregor, and A.C. McHardy, *Taxator-tk: precise taxonomic assignment of metagenomes by fast approximation of evolutionary neighborhoods.* Bioinformatics, 2015. **31**(6): p. 817-824.

28. Chaumeil, P.-A., et al., *GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database.* Bioinformatics, 2019. **36**(6): p. 1925-1927.

29. Olm, M.R., et al., *dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication.* The ISME Journal, 2017. **11**(12): p. 2864-2868.

30. Shaffer, M., et al., *DRAM for distilling microbial metabolism to automate the curation of microbiome function.* Nucleic Acids Research, 2020. **48**(16): p. 8883-8900.

31. Wick, R.R., L.M. Judd, and K.E. Holt, *Performance of neural network basecalling tools for Oxford Nanopore sequencing.* Genome Biology, 2019. **20**(1): p. 129.

32. R Core Team, *R: A language and environment for statistical computing.* 2020.

33. Oksanen, J., et al., *vegan: Community Ecology Package*. 2020.

34. Gower, J.C., *A general coefficient of similarity and some of its properties.* Biometrics, 1971: p. 857-871.

35. D'Orazio, M. and M.M. D'Orazio, *Package 'StatMatch'.* available online at one of the mirror sites shown at, 2022.

36. Anderson, M.J., *Permutational Multivariate Analysis of Variance (PERMANOVA)*, in *Wiley StatsRef: Statistics Reference Online*. p. 1-15.

37. Martinez Arbizu, P., *pairwiseAdonis: Pairwise multilevel comparison using adonis.* R package version 0.4, 2020. **1**.

38. Anderson, M.J., *Distance-based tests for homogeneity of multivariate dispersions.* Biometrics, 2006. **62**(1): p. 245-53.

39. St, L. and S. Wold, *Analysis of variance (ANOVA).* Chemometrics and intelligent laboratory systems, 1989. **6**(4): p. 259-272.

40. Somerfield, P.J., K.R. Clarke, and R.N. Gorley, *A generalised analysis of similarities (ANOSIM) statistic for designs with ordered factors.* Austral Ecology, 2021. **46**(6): p. 901-910.

41. Segata, N., et al., *Metagenomic biomarker discovery and explanation.* Genome biology, 2011. **12**(6): p. R60-R60.

42. Legendre, P. and L. Legendre, *Numerical ecology*. 2012: Elsevier.

43. BioRender, *BioRender.* 2021.

44. Wickham, H., *ggplot2: elegant graphics for data analysis*. 2016: springer.

45. Patil, I., *Visualizations with statistical details: The'ggstatsplot'approach.* Journal of Open Source Software, 2021. **6**(61): p. 3167.

46. Walsh, L.H., et al., *Analysis of the milk kefir pan-metagenome reveals four community types, core species, and associated metabolic pathways.* iScience, 2023. **26**(10): p. 108004.

47. Bray, J.R. and J.T. Curtis, *An ordination of the upland forest communities of southern Wisconsin.* Ecological monographs, 1957. **27**(4): p. 326-349.

48. Truong, D.T., et al., *Microbial strain-level population structure and genetic diversity from metagenomes.* Genome Res, 2017. **27**(4): p. 626-638.

49. Feehily, C., et al., *Detailed mapping of Bifidobacterium strain transmission from mother to infant via a dual culture-based and metagenomic approach.* Nat Commun, 2023. **14**(1): p. 3015.

50. Bourrie, B.C.T., C. Richard, and B.P. Willing, *Kefir in the Prevention and Treatment of Obesity and Metabolic Disorders.* Curr Nutr Rep, 2020. **9**(3): p. 184-192.

51. Maddamsetti, R. and N.A. Grant, *Divergent Evolution of Mutation Rates and Biases in the Long-Term Evolution Experiment with Escherichia coli.* Genome Biol Evol, 2020. **12**(9): p. 1591-1603.

52. Jin, H., et al., *Hybrid, ultra-deep metagenomic sequencing enables genomic and functional characterization of low-abundance species in the human gut microbiome.* Gut Microbes, 2022. **14**(1): p. 2021790.

Section 5.6. Supplementary information

5.6.1 Supplementary Figures

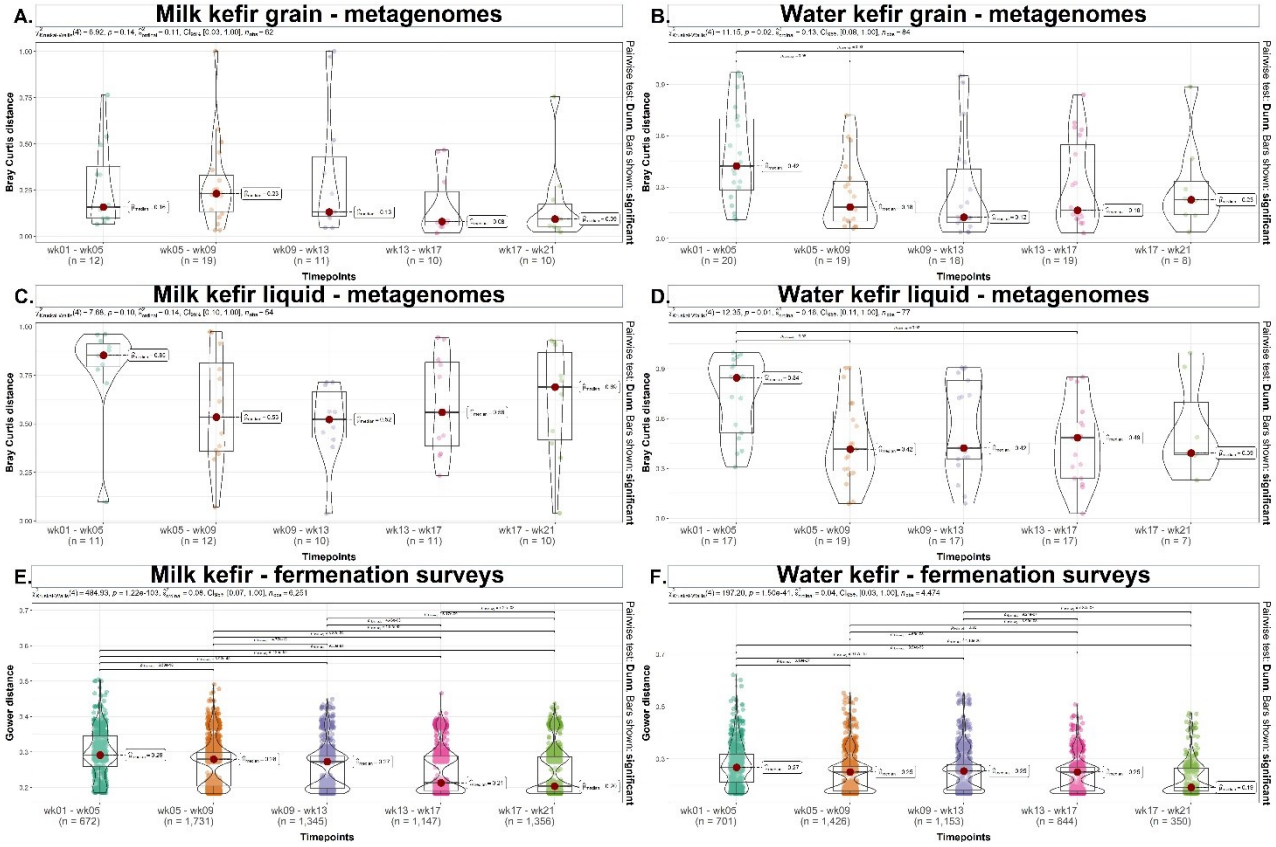


Figure S1: (A.) Beta-diversity boxplots displaying the Bray Curtis values between milk kefir grain metagenomes from the same participant, representing different timeframes of the project ('wk01 vs wk05', 'wk05 vs wk09', 'wk13 vs wk17' and 'wk17 vs wk21'). **(B.)** Beta-diversity boxplots displaying the Bray Curtis values between water kefir grain metagenomes from the same participant, representing different timeframes of the project ('wk01 vs wk05', 'wk05 vs wk09', 'wk13 vs wk17' and 'wk17 vs wk21')**. (C.)** Beta-diversity boxplots displaying the the Bray Curtis between milk kefir liquid metagenomes from the same participant, representing different timeframes of the project ('wk01 vs wk05', 'wk05 vs wk09', 'wk13 vs wk17' and 'wk17 vs wk21'). **(D.)** Beta-diversity boxplots displaying the Bray Curtis values between water kefir liquid metagenomes from the same participant, representing different timeframes of the project ('wk01 vs wk05', 'wk05 vs wk09', 'wk13 vs wk17' and 'wk17 vs wk21'). Kruskal-Wallis and Dunn test was used to assess differences in beta diversity, since the data was nonparametric. Red dot represents the mean Bray Curtis between metagenomes representing different timeframes of the project. **(E.)** Gower-diversity boxplots displaying the Gower distance between milk kefir fermentation surveys completed by the same participant, representing different timeframes of the project ('wk01 vs wk05', 'wk05 vs wk09', 'wk13 vs wk17' and 'wk17 vs wk21'). **(F.)** Gower-diversity boxplots displaying the Gower distance between water kefir fermentation surveys completed by the same participant, representing different timeframes of the project ('wk01 vs wk05', 'wk05 vs wk09', 'wk13 vs wk17' and 'wk17 vs wk21').

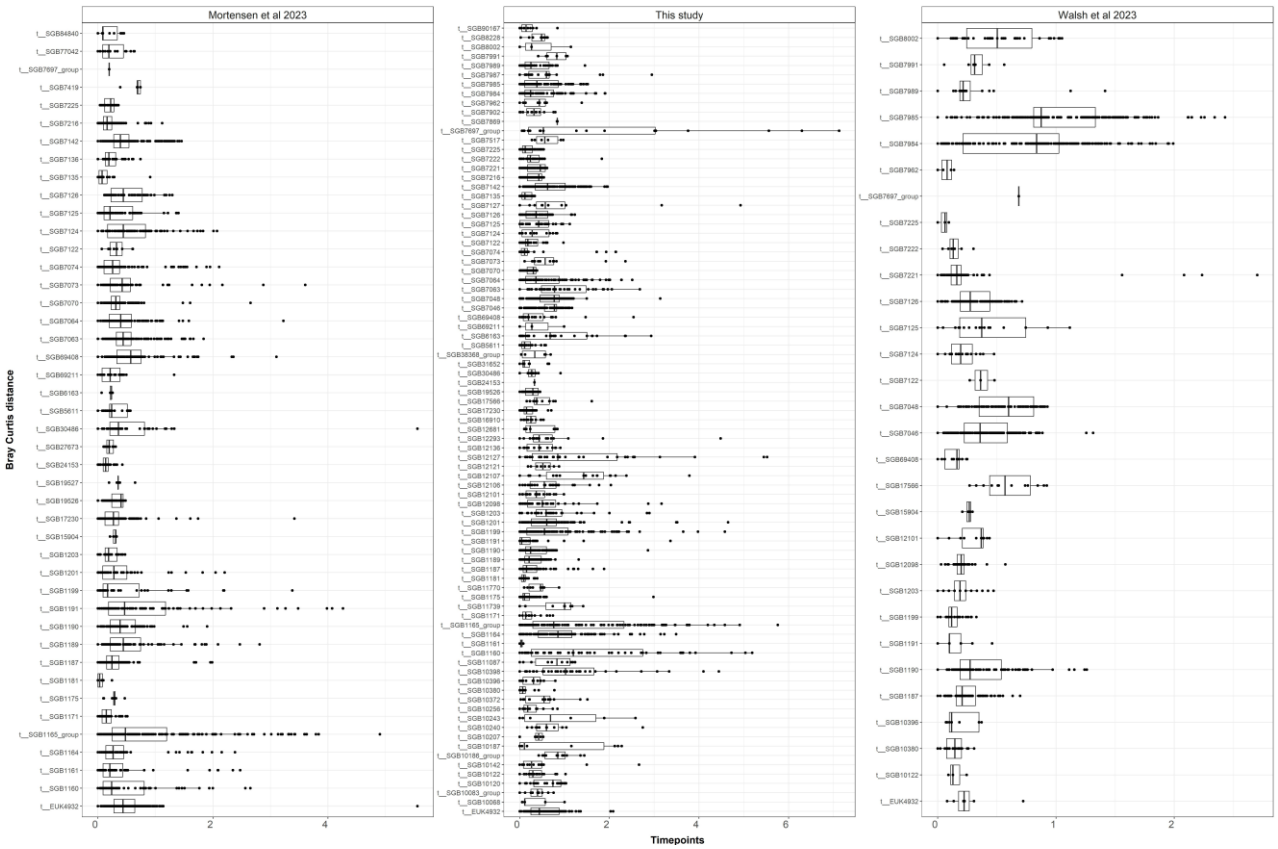


Figure S2: Collective intra-sample strain diversity across metagenomes for each of the outlined bacterial species (X-axis). Each coloured boxplot displays the range of polymorphic sites expressed as percentages (Y-axis); along the MLSA of strains.

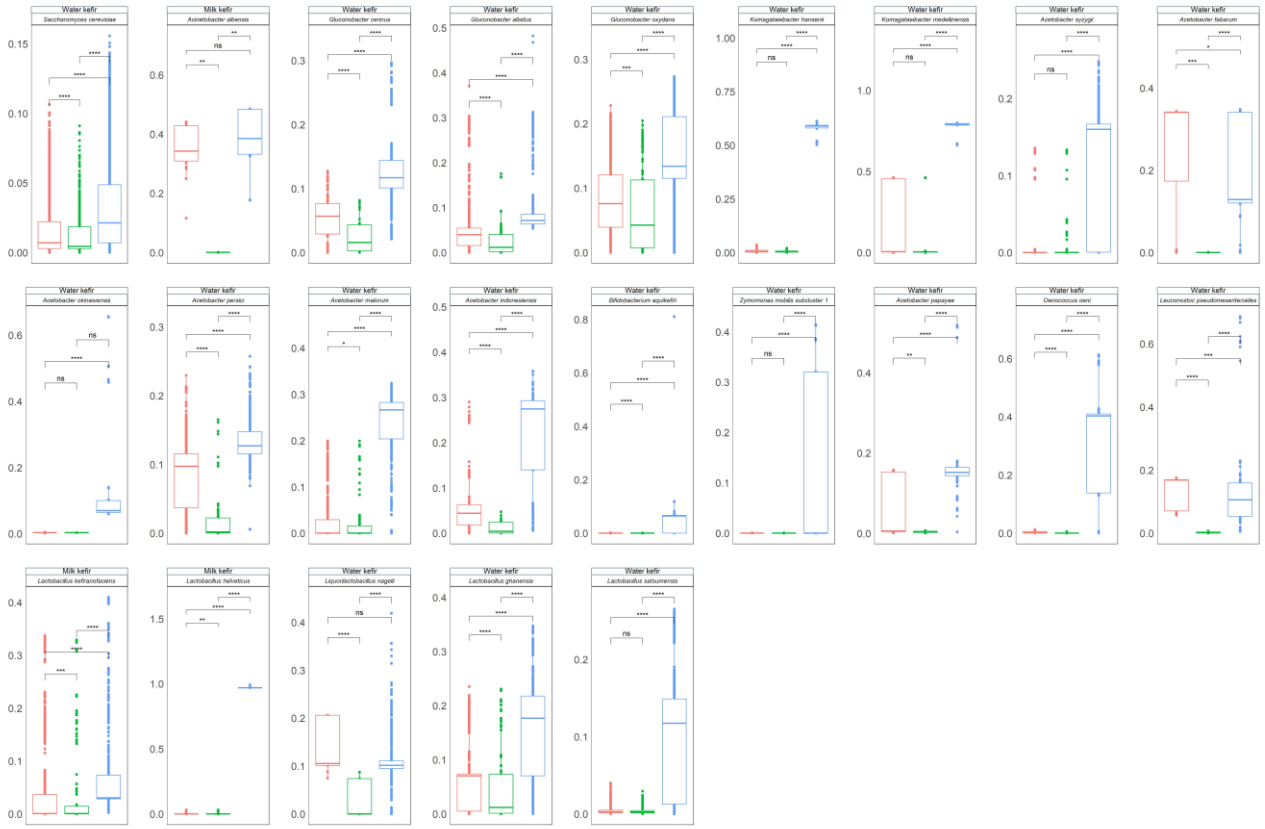


Figure S3: Phylogenetic-diversity boxplots displaying the phylogenetic distance values between Strain comparisons 'Intra-participant comparisons', 'Inter-participant comparisons' and 'Unrelated comparisons', of the outlined species (see title).

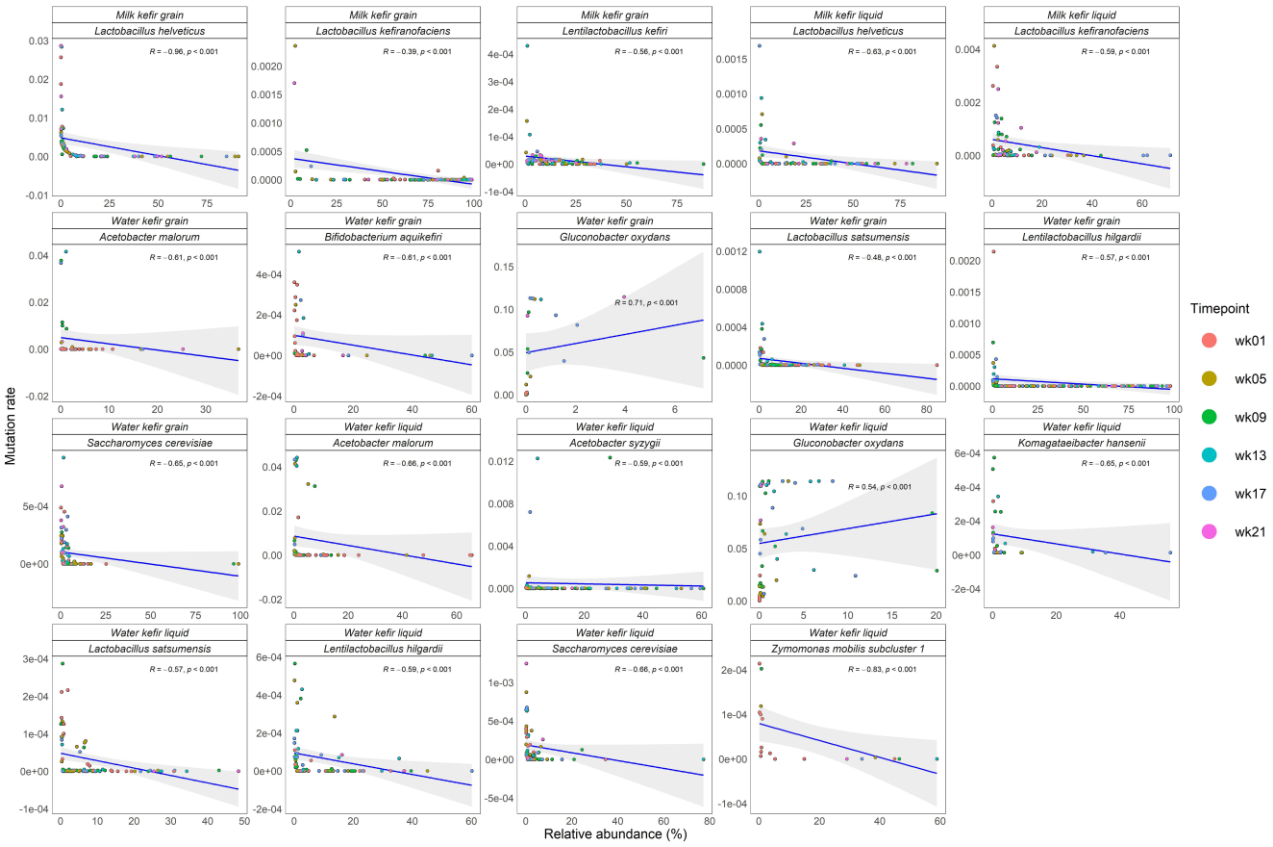


Figure S4: Correlations between species relative abundance and strain rates of nucleotide change.

5.6.2 Supplementary References

1. Wei, T., et al., *Package 'corrplot'.* Statistician, 2017. **56**(316): p. e24.

2. Venables, W. and B. Ripley, *Exploratory multivariate analysis*, in *Modern applied statistics with S*. 2002, Springer. p. 301-330.

3. R Core Team, *R: A language and environment for statistical computing.* 2020.

General Discussion

As discussed in Chapter 1, high-throughput DNA sequencing-based approaches have and will continue to revolutionise our understanding of microbial ecosystems [1]. Metagenomic and metatranscriptomic approaches are state of the art biological profiling methods that have greatly enhanced the field of food microbiology. Whole metagenome shotgun sequencing (WMGS) provides an untargeted sequencing-based approach to assess DNA from biological samples, including starter cultures, probiotics, and foodborne pathogens [2]. Additionally, and of particular relevance to this thesis, WMGS allows culture-independent analysis of the mixed microbial communities, or microbiota, present in fermented foods [3]. In Chapter 1, we discuss how the expanded use of shotgun metagenomics and metatranscriptomics has provided a deep insight into the microbial composition, functional potential and gene expression of fermented foods. However at present many bioinformatic tools use different reference sources and classification algorithms, contributing to performance differences, furthermore there is an absence of gold standard methods for metagenomics analysis that will allow for accurate standardised workflows. In chapter 2-5, we utilise several such bioinformatic tools to extend current knowledge concerning the microbiota of milk and water kefir. Furthermore, we provide representative examples in chapter 2, 3 and 5 into how the results of each tool in isolation can be greatly extended through the application of additional complementary bioinformatic tools. For example in chapter 2, the detection of 3 community types using Kraken 2, was extended through strain specific bioinformatic tools to four community types based on the identification of sub- clusters of *Lla. lactis*. Collectively these insights were used to explain patterns of dissimilarity in the functional potential data obtained using HUMAnN 3, further confirming the presence of 4 reproducible community types.

As stated in chapter 5, the fermented beverages milk kefir and water kefir have a long history of safe consumption. Although traditionally consumed in some regions, the global popularity of, and demand for, both milk and water kefir has increased in recent years in light of the health benefits attributed to these foods. Increasingly starter cultures are applied to kefir production on an industrial scale, to facilitate greater

quality control. Due to limitations in regulations regarding the microbial composition of kefir, some mixtures of starter cultures used in industrial applications are not reflective of the microbial composition within the artisan beverage [4]. Thus it is of considerable industrial and academic value given the usefulness of kefir as model systems [5, 6] to gain deep knowledge about the species, strain and functional composition within the kefir microbiome [7]. In chapter 2 and 4, we utilise metagenomics-based identification and characterisation of the microorganisms present within kefir communities during fermentation to provide a considerable insight into the heterogeneity of these populations. Specifically, in chapter 2, we examined 256 milk kefir metagenomes representing 8 and 24-hour fermentations initiated using 64 grains sourced from 25 different countries. This analysis revealed a consistent pattern of domination by the microbial species *L. helveticus*, *L. kefiranofaciens*, *Lla. lactis* and *Lla. cremoris* . However, even in kefirs where *Lla. lactis* and *Lla. cremoris* did not dominate, these species and 51 associated metabolic pathways were identified across all metagenomes, providing a framework upon which the minimal microbial and functional criteria to define a milk product as a milk kefir can be established. In chapter 4 and 5, we highlight how research relating to the milk kefir and water kefir microbiome was greatly extended through the efforts of the Kefir4All citizen scientists. Again a WMGS approach was applied in Chapter 4 and 5 to characterise milk and water kefir fermentations. From the samples provided by citizen scientists throughout the project, we generated 203 milk kefir metagenomes and 246 water kefir metagenomes, to study compositional change and microbial evolution at the species and strain level. Specifically, through repeat fermentations milk kefir grains initially dominated by *L. kefiranofaciens* transitioned into separate milk kefir metagenomes that clustered with 'Walsh et al - Milk kefir' (chapter 2) metagenomes and categorised as representing the *L. helveticus*, *L. kefiranofaciens*, *Lla. lactis* and *Lla. cremoris* community type, and thus reproduced the same patterns of dominance in milk kefir metagenomes as outlined in chapter 2. Through repeat regular fermentations, water kefir grains initially dominated by *Len. hilgardii* transitioned into distinct water kefir metagenomes that clustered with 'Breselge et al - Water kefir' metagenomes categorised as representing the *Lc. paracasei, Len. hilgardii* and *Z. mobilis* community types. Such community type changes tended to occur at the initial timeframe of the project (wk01-wk09). Similarly, the majority of functional change at the

community level and evolutionary change for most prevalent species within the milk and water kefir microbiome occurred at the initial timeframe of the project (wk01-wk09). Such results should be taken into consideration when utilising milk and or water kefir as a model systems [6] or in commercial production through a traditional method as deviations at the community and strain level can occur after a single week of fermenting under different conditions. In addition to expanding on the current knowledge concerning the milk and water kefir microbiome, we further discuss in chapter 5 how the citizen science project, Kefir4All, represents a powerful platform, to convey concepts and information concerning microbiology, for example by providing an individualised practical example of microbiology in the context of food production. Such platforms are of particular importance as the survey results and comments collected and analysed in chapter 5, indicate an absence of background knowledge concerning the topics of Microbiology and Fermented food in the general public, which was one of the primary factors that limited identified that influenced scientific interest into both topics. Our results further highlight the value of schools as hubs for citizen science projects. We hope that this study will guide other food microbiologists in their efforts to design citizen science experiments.

As outlined in chapter 1, high throughput sequencing continues to be applied to increasing larger datasets. Soon the capacity of such sequencing approaches, will extend past a laboratory scale, as such citizen science initiatives may represent an attractive approach for sample collection due to the potential for citizen science to facilitate sample collection from both broad geographic areas and over extended time periods. It is notably the Kefir4aAll citizen science initiative described in chapter 4 and 5 could reproduce the community types, found in professionally obtain milk and water kefir metagenomic datasets, with only marginal differences between the 'Breselge et al' kefir and Kefir4All -Water kefir' (Pairwise ANOSIM, R2 = 0.14 and p=0.01) in terms of similarity.

While in Chapters 1, 2, 4 and 5 we used shotgun metagenomics to characterise the kefir microbiome; in Chapter 3 we examined if consumption of milk kefir and other foods could impact the gut microbiota and urine metabolome of healthy human participants. In chapter 3, we combined shotgun metagenomics with metabolomics to identify changes in the human gut microbiota and metabolome of healthy participants

following daily consumption, over 28 days, of either a single portion of a commercial FMP containing the probiotic *Lc. casei*, a traditional fermented milk beverage kefir or a diet enriched with the prebiotic inulin. Our findings suggest that kefir consumption in a healthy cohort can have a subtle impact on the urinal metabolome and gut microbiome in a subset of participants. The principle change to the gut microbiome was the detection of *Lla. raffinolactis*. In conclusion, in this thesis, I have demonstrated the value of WMGS based characterisation of fermented foods, the benefits of citizen science for educational and research purposes and examined the impact of regular milk kefir consumption on the host gut microbiota.

References

1. Jo, J., J. Oh, and C. Park, *Microbial community analysis using high-throughput sequencing technology: a beginner's guide for microbiologists.* J Microbiol, 2020. **58**(3): p. 176-192.
2. Quince, C., et al., *Shotgun metagenomics, from sampling to analysis.* Nat Biotechnol, 2017. **35**(9): p. 833-844.
3. Walsh, A.M., et al., *Integrated molecular approaches for fermented food microbiome research.* FEMS Microbiology Reviews, 2023. **47**(2).
4. Nejati, F., et al., *Traditional Grain-Based vs. Commercial Milk Kefirs, How Different Are They?* Applied Sciences, 2022. **12**(8): p. 3838.
5. Walsh, A.M., et al., *Microbial Succession and Flavor Production in the Fermented Dairy Beverage Kefir.* mSystems, 2016. **1**(5).
6. Blasche, S., et al., *Metabolic cooperation and spatiotemporal niche partitioning in a kefir microbial community.* Nat Microbiol, 2021. **6**(2): p. 196-208.
7. Nejati, F., S. Junne, and P. Neubauer, *A Big World in Small Grain: A Review of Natural Milk Kefir Starters.* Microorganisms, 2020. **8**(2).