



^b
**UNIVERSITÄT
BERN**

Graduate School for Cellular and Biomedical Sciences
University of Bern

**Decoding Microbial Genomes:
Novel User-Friendly Tools Applied to
Fermented Foods**

PhD Thesis submitted by

Thomas Roder

for the degree of
PhD in Computational Biology

Supervisor

PD Dr. Rémy Bruggmann
Interfaculty Bioinformatics Unit
Faculty of Science of the University of Bern

Co-advisor

Prof. Dr. Stephanie Ganal-Vonarburg
Department for Biomedical Research
Faculty of Medicine of the University of Bern

Co-advisor

PD Dr. Guy Vergères
Research Group Functional Nutritional Biology / Department of Health Sciences and Technology
Agroscope / ETH Zürich



This work is licensed under a Creative Commons Attribution 4.0 International License.

<https://creativecommons.org/licenses/by/4.0/>

This license allows readers to reproduce, disseminate and reuse your work independently of format, medium or purpose, as long as they provide the appropriate copyright and legal information and indicate whether any changes were made.

Accepted by the Faculty of Medicine, the Faculty of Science and the
Vetsuisse Faculty of the University of Bern at the request of the Graduate
School for Cellular and Biomedical Sciences

Bern, Dean of the Faculty of Medicine

Bern, Dean of the Faculty of Science

Bern, Dean of the Vetsuisse Faculty Bern

Abstract

Over the past two decades, the cost of DNA sequencing per base has significantly outpaced Moore's law. Many organizations and research groups have exploited this trend and generated large amounts of genomic data and made it possible to tackle new research questions. This growth also brings challenges, including the need for faster algorithms, more efficient ways to visualize and explore data, more automatized data processing, and systematic data management.

For more than a century, Agroscope collects lactic acid bacteria (LAB) extracted from the Swiss dairy environment. Today, the collection comprises more than 10'000 strains and so far, for about 15% of the strains the genome was sequenced. The over-arching goal of this thesis is to find new ways of exploiting this genetic potential to design new fermented food products including potential additional health benefits, and to understand the underlying mechanisms.

One compound with potential health benefits is indole. Previous experiments have shown that indole compounds modulate the gut immune system via the aryl hydrocarbon receptor (AhR). Our objective was to create a yoghurt enriched in indole metabolites through fermentation, and then to examine whether maternal consumption of this yoghurt would enhance gut immune system maturation in germ-free mice. To reduce the number of strains to test, I developed comparative genomics tools to pre-select strains from the strain collection. This led to the successful development of a yoghurt with significantly increased AhR activation activity. In germ-free mice, we could show the expected effect.

Based on these comparative genomics tools, I developed the software OpenGenomeBrowser to enable biologists, who know their organisms of interest in great detail, to efficiently explore the genomic data by themselves, without bioinformatics skills or the need for a middleman bioinformatician. The foundation of OpenGenomeBrowser is a simple system for transparent data management of microbial genomes which makes the automation of common bioinformatics workflows possible. In addition, I built a user-friendly website based on modern web technologies to facilitate common bioinformatics workflows. Because of OpenGenomeBrowser's solid foundation, it is the first software of its kind that can be self-hosted and is dataset-independent, making it potentially useful for many similar genome datasets.

During the project, we measured thousands of metabolites in yoghurts made using different strains. However, we experienced that no existing tools could adequately connect such a high-dimensional phenotypic dataset to the genomic information, i.e., presence-absence of orthogenes. Finding high-confidence causative links between these datasets is challenging because of the properties of microbial genomes. For instance, clonal reproduction leads to genome-wide linkage disequilibrium, which prohibits the use of techniques developed for human genome-wide association studies (hGWAS). To this end, I developed Scoary2, a complete rewrite and extension of the original microbial GWAS (mGWAS) software Scoary. The key improvements include an implementation of the core algorithm that is orders of magnitude faster and an interactive web-app that enables efficient data exploration of the output, which is crucial given the size of the dataset. With this software, we discovered two previously uncharacterized genes involved in the carnitine metabolism.

Table of Contents

Abstract	IV
Table of Contents	V
Abbreviations	VII
1. Introduction.....	1
1.1. Genomics.....	1
1.1.1. First-generation sequencing.....	1
1.1.2. Second-generation sequencing.....	1
1.1.3. Third-generation sequencing.....	2
1.1.4. Genome assembly process.....	2
1.1.5. Assembly quality control.....	5
1.1.6. Gene prediction in prokaryotes.....	5
1.1.7. Functional gene annotation.....	6
1.1.8. Genome data management	8
1.2. The human gut microbiota.....	9
1.2.1. Environmental influences that shape the microbiota.....	10
1.2.2. How to intervene in the microbiota?.....	11
1.2.3. Yogurt.....	12
1.3. Polyfermenthealth.....	12
1.3.1. A “polydiverse” yoghurt.....	13
1.3.2. Indoles and the aryl hydrocarbon receptor (AhR).....	13
1.3.3. Folate	16
1.4. Linking metabolomics with genomic data.....	18
1.4.1. Existing approaches.....	18
1.4.2. Microbial genome-wide association studies (mGWAS).....	19
1.5. Genome-scale metabolic modeling.....	20
1.5.1. Traditional applications.....	21
1.5.2. Community models.....	22
1.5.3. Community modeling of yoghurt.....	22
2. Aims and Objectives.....	24
3. Results.....	25
3.1. Manuscript 1: Comparative genomics of the Dialact database	25
3.2. Manuscript 2: OpenGenomeBrowser.....	43
3.3. Manuscript 3: Scoary2.....	55
3.4. Manuscript 4: Indole yoghurt.....	77
4. Discussion and Outlook.....	91
4.1. Management of microbial genome databases.....	91
4.2. Polyfermenthealth: strain selection.....	92

4.2.1. Genome-scale metabolic modeling.....	94
4.3. OpenGenomeBrowser.....	95
4.4. Polyfermenthealth: yoghurts.....	96
4.4.1. Polydiverse yoghurt.....	96
4.4.2. AhR-activating “indole” yoghurt.....	97
4.4.3. Folate yoghurt.....	98
4.5. Scoary2.....	99
4.6. Polyfermenthealth.....	100
4.7. Bioinformatics software development.....	101
5. Appendix.....	104
5.1. Manuscript 5: Comparative genomics of <i>Brachyspira hyodysenteriae</i>	105
5.2. Manuscript 6: Frontiers for Young Minds article.....	121
6. References.....	130
7. Acknowledgments.....	143
8. Curriculum Vitae.....	144
9. List of Publications.....	145
10. Declaration of Originality.....	146

Abbreviations

5-Me-THF	L-5-methyltetrahydrofolate
AhR	aryl hydrocarbon receptor
AIP	AhR-interacting protein
ARNT	aryl hydrocarbon receptor nuclear translocator
BBH	bi-directional best hit
BCFA	branched-chain fatty acids
BMI	Body mass index
bp	base pairs
CCS	circular consensus sequencing
CD	Crohn's disease
CDS	coding sequence
c-Src	proto-oncogene tyrosine-protein kinase cellular-sarcoma
dAMP	Adenine monophosphate
dGMP	Guanosine monophosphate
dTMP	Thymidine monophosphate
FAK	focal adhesion kinase
FBA	flux-balance analysis
FFT	fecal filtrate transplant
FMP	fermented milk pellet
FMT	fecal microbiota transplant
FODMAP	fermentable oligosaccharides, disaccharides, monosaccharides and polyols
GC-MS	gas chromatography mass spectrometry
GRAS	generally recognized as safe
GSMM	genome-scale metabolic model
GWAS	genome-wide association studies
hGWAS	human genome-wide association studies
Hsp90	heat shock protein 90
I3A	indole 3-acetate
I3C	indole-3-carbinol
IAC	indoleacrylic acid
IBD	inflammatory bowel disease
IBU	Interfaculty Bioinformatics Unit
IgA	immunoglobulin A
ILC3	type 3 innate lymphoid cells
LAB	lactic acid bacteria
LC-MS	liquid chromatography mass spectrometry
MAPK	mitogen-activated protein kinase
mGWAS	microbial genome-wide association studies
Microbiome	Gut microbiome
Microbiota	Gut microbiota
MS	mass spectrometry
NF- κ B	nuclear factor-kappa B
NK	natural killer cell
OLC	overlap-layout-consensus
ONT	Oxford Nanopore Technologies
PacBio	Pacific Bioscience
PCR	polymerase chain reaction
PXR	pregnane X receptor
SBH	single-directional best hit
SCFA	short-chain fatty acid
SNP	single nucleotide polymorphism

SNV	single nucleotide variant
T2D	type-2-diabetes
TCA	tricarboxylic acid
T _H	T helper cell
THF	tetrahydrofolate
TMA	trimethylamine
T _{reg}	regulatory T cell
UC	ulcerative colitis
XAP2	X-associated protein 2
XME	xenobiotic metabolizing enzyme
XRE	xenobiotic response element

1. Introduction

1.1. Genomics

Knowledge of the genetic code is a cornerstone of the life sciences, enabling scientists to study a wide array of topics, including cellular processes, characterization of biodiversity and evolutionary history. Fundamentally, DNA sequencing is possible because of the simple chemical structure of DNA (a linear molecule composed of four different bases), and the existence of natural enzymes that enable DNA replication (DNA polymerases).

The reduction in sequencing cost per base pair has even outpaced Moore's law. Each significant improvement, be it in price, accuracy, speed, read length, preparation protocol or portability, opens new opportunities, and requires new bioinformatics tools to deal with them. This is reflected in Sanger's rule [1]:

Anytime you get technical development that's two to threefold or more efficient, accurate, cheaper, a whole range of experiments opens up.

1.1.1. First-generation sequencing

The method that defines the first generation of DNA sequencing technology is called Sanger sequencing and was developed by Frederick Sanger in 1977 [2], though in the same year, Maxam and Gilbert [3] also published a similar method. The process begins with the amplification of the target DNA molecule. Early on, phages and plasmids were used for this purpose, later complemented by the more efficient polymerase chain reaction (PCR). Next, the double-stranded DNA is denatured using heat, resulting in single-stranded DNA. Subsequently, a primer complementary to the template strand is allowed to anneal. The DNA polymerase then starts to add new nucleotides to the primer. However, a low percentage of the reaction mixture consists of dideoxynucleotides, which lack a hydroxyl group on the 3' carbon of the sugar moiety, preventing the addition of a next nucleotide. This leads to the synthesis of DNA strands of all possible lengths which can be separated using polyacrylamide gel electrophoresis. Because the dideoxynucleotides are fluorescently labeled with one of four colors encoding their nucleobase identity, the color of each band in the gel reveals the terminal nucleotide. (The original method used four different reaction mixtures, each containing a different dideoxynucleotide labeled with a radioactive isotope.)

The crowning achievement of this technology was the sequencing of most of the human genome in 2001, resulting in a 2.91-billion base pair (bp) consensus sequence [4, 5]. Sanger sequencing is still in use today for affordable, high-quality, low-throughput sequencing of fragments up to 1000 bp [6].

1.1.2. Second-generation sequencing

The technologies that eventually replaced Sanger's method in popularity are based on a different principle. The main difference is that the position of the nucleotides is no longer encoded through fragment length but in temporal sequence. The crucial advantage of this approach is that it is highly parallelizable.

In the case of the market leader Illumina, this is how DNA sequences are read: a 50 to 500 bp fragment of DNA is amplified and primers are added. Next, the DNA polymerase adds a single fluorescently labeled nucleotide to each copy of the DNA fragment. Each nucleotide is labeled with a different fluorescent label. The labels prevent the addition of more nucleotides. After the labels have been added, they are excited by a laser and the cumulative signal of all copies is measured. The labels are then enzymatically removed, and the next nucleotide can be added, initiating the next cycle.

The parallelization is enabled by binding the DNA fragments on a glass slide using adapter sequences. Where they land, they can be amplified in-place such that the glass slide is covered by patches of identical DNA fragments (bridge amplification). Next, after a DNA fragment is read in one direction, the complementary fragment can be produced and read in the same way, resulting in what is termed paired-end reads.

This generation of sequencing technology is defined by the high throughput and sequence accuracy. For instance, the Illumina MiSeq sequencer, released in 2011, can read up to 25 million DNA fragments in one sequencing run, yielding 300 bp long, paired-end reads [6]. In 2014, Illumina's technology reached the symbolic milestone of reducing the price of sequencing the human genome to 1,000 USD [7]. However, there are three caveats to this milestone. First, the sequencing costs have stagnated somewhat since then [8]. Second, with data generated by short-read technologies only incomplete draft-quality assemblies can be made. Third, the price excludes bioinformatics processing, which is essential [9].

1.1.3. Third-generation sequencing

The latest generation of DNA sequencing technologies focuses on generating long reads (up to 60 kbp) from single, non-amplified DNA molecules. The market leaders in this field are Pacific Bioscience (PacBio) and Oxford Nanopore Technologies (ONT), though other companies will enter the market soon [10].

In PacBio sequencing, a modified DNA polymerase is immobilized at the bottom of a nano-scale well ($\varnothing < 100$ nm). The well is filled with the four nucleotides, each labeled with a different fluorophore. Because the well is smaller than the wavelength of fluorophore-activating light coming from below the well, only the very bottom is illuminated. Only when the DNA polymerase adds the fluorescent nucleotides to the DNA of interest do they come close enough to the bottom of the well to be activated and emit a fluorescent signal. Thus, the DNA amplification process can be measured in real-time, at the speed of the slowed-down DNA polymerase (1-3 bases per second [11]). Furthermore, because modified DNA bases slightly slow down DNA polymerase, this technology also allows researchers to capture methylations and other epigenetic modifications [12].

The main drawback of long-read technologies (both PacBio and ONT) is that the accuracy of measuring a DNA molecule with this technique is not very high (75-90%). Since 2019, however, PacBio offers a circular consensus sequencing (CCS) protocol capable of producing long, high fidelity ("HiFi") reads. Using this protocol, the DNA is circularized such that the DNA polymerase continues to read the same molecule over and over again, culminating in 99.9% single molecule read accuracy of the consensus sequence [13]. Moreover, while third-generation technologies used to be significantly more expensive, at least for sequencing many bacteria in bulk, the cost of PacBio HiFi sequencing is almost the same as that of Illumina because up to 96 samples can be sequenced on one SMRTcell.

1.1.4. Genome assembly process

None of these technologies directly yield complete DNA sequences at the scale of even small genomes, except for phages or viruses. Thus, algorithms were developed to combine the multitude of generated sequences into larger assemblies, with the goal of eventually deducing full genomes or plasmids. This problem is difficult, and new methods are still being developed. Which method should be chosen depends on the dataset. Biological factors matter, such as the ploidy of the organism in question, as well as the type of sequencing reads. None of the methods are perfect and usually produce at least slightly different results, and even supposedly simple genome assemblies should be carefully inspected and curated.

1.1.4.1. Assembly of short reads

The *greedy algorithm* is a simple and straightforward solution for the assembly of short reads. It iteratively joins together the reads with the highest quality overlap, removing the used read from the pool of reads, until no further extensions are possible. This approach was successfully used in the Human Genome Project. It is computationally efficient but cannot handle highly repetitive regions (Figure 1 b) [14]. To solve this problem, the greedy approach was superseded by graph-based algorithms.

In the *overlap-layout-consensus (OLC)* method, a graph is created where each read is represented by a node and edges represent overlaps between reads (Figure 1 c). This assembly process has three stages. First, the *overlaps* between reads must be found. Because comparing all reads with each other has quadratic time complexity, instead, an index of k -mers is created which can be efficiently queried. In the *layout* step, the graph is constructed. This involves joining reads that can be unambiguously assembled into contigs. Finally, the *consensus* sequence of each contig is calculated. This method was used by Celera's genome sequencing project that competed with the Human Genome Project. However, this algorithm is still not computationally efficient enough for high-depth data from second-generation sequencing. The reason is that the high number of reads balloons the size of the graph, and solving it in the layout step has quadratic time complexity [14, 15].

The basic principle behind the *de Bruijn graph* method is that each read is decomposed into overlapping k -mers. An edge is added between two k -mers if they are adjacent on the same read (Figure 1 d). This solves the problem of large sequencing depth because adding more copies from the same genome does not introduce new k -mers into the graph. The assembly problem can then be reformulated as an Eulerian path problem, where the goal is to find a path through the graph which visits each edge exactly once. In practice, there usually is no ideal Eulerian solution, so the assemblers create multiple contigs from unambiguous regions of the graph. The number of contigs strongly depends on the k -mer length and the read length. The popular assembler SPADeS [16] automatically determines an optimal k -mer length based on the data, but the read length fundamentally limits the quality of genomes assembled from short reads. However, this algorithm can be implemented efficiently enough to allow the assembly of short-read sequencing data into draft human genomes [14].

The *string graph* method evolved from OLC and is, thus, based on a graph of overlapping reads. Like the de Bruijn graph method, it solves the problem of ballooning graph size, but without k -mers. The basic idea is that before graph construction, redundant reads (i.e., a read that is a sub-read of another) and edges (i.e., edges in unambiguous regions) are collapsed (Figure 1 e) [14].

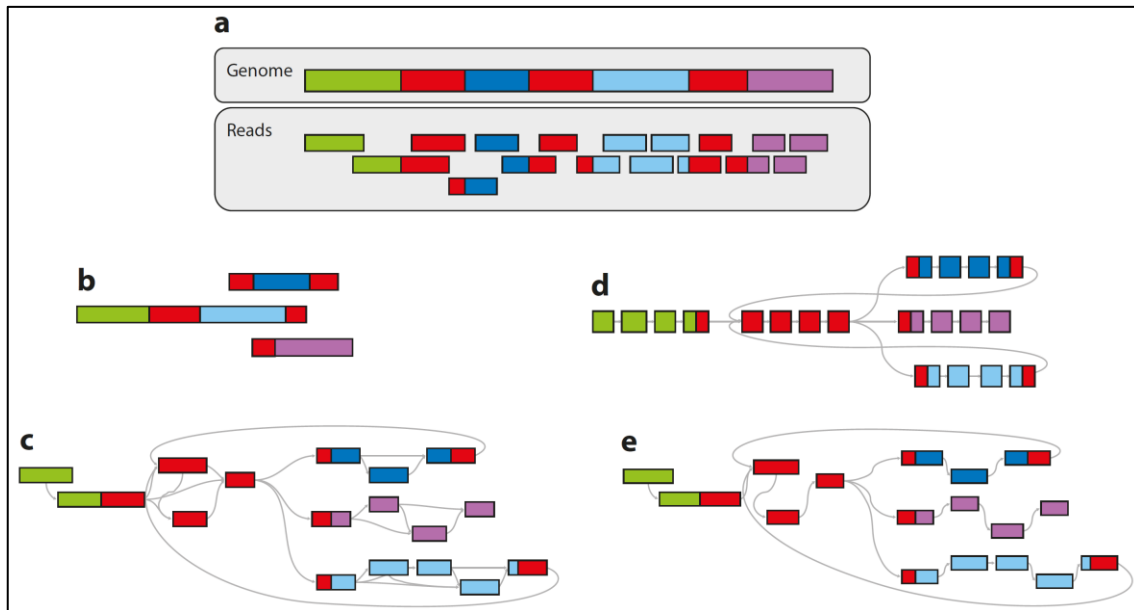


Figure 1: Genome assembly paradigms.

(a) The layout of a genome containing repeated genomic regions (red), along with sequence reads. (b) The solution from the greedy approach. It erroneously connected the green and the light blue region because of local optimization and created isolated contigs for the blue region. (c) The overlap-layout-consensus (OLC) graph, where each read is represented as a nodes in the graph, and edges represent overlaps between reads. (d) In the de Bruijn graph, reads are split into k -mers which are connected by an edge if they are adjacent on a read. Repeated regions are connected to multiple different regions. (e) The string graph is like the OLC graph, only that unnecessary edges and reads were removed. (This example contains no unnecessary reads.) Figure from Simpson and Pop 2015 [14].

1.1.4.2. Assembly of long reads

Even the sequence of relatively simple bacterial genomes can rarely be fully reconstructed using short-read sequencing technologies. This is because many genomes contain repetitive elements, for example, multiple copies of rRNA genes or transposons. Without reads that span these regions, multiple ways of solving the graph persist and ambiguity remains. Short-read assemblies thus consist of many contigs that cannot be connected or confidently identified as chromosomal or plasmid (or contaminant) DNA. This may not be important for many downstream analyses, for instance, species identification or identifying most genes. However, the perfect (or near-perfect) genomes that are now possible with long-read technologies may be essential for other analyses, like detecting mobile elements and structural rearrangements, estimating mutation rates, or studying transmission chains. Even if such analyses are not planned at the outset, it could make sense for sequencing projects to choose long-read assemblies, as such projects tend to be long-term investments and such analyses may become necessary in the future [17, 18].

Many algorithms and tools for the assembly of third-generation sequencing reads have been developed, most building on the approaches developed for short reads. Before the advent of HiFi reads, the major challenge was the high error rate. It poses problems for k -mer-based algorithms, which assume that most k -mers are preserved in multiple reads. Thus, most algorithms are based on the OLC approach (Canu [19], wtdbg2 [20], Miniasm [21] and Smartdenovo [22], Raven [23], NECAT [24]) or string graph (Falcon [25], NextDenovo [26]), and only Flye [27, 28] is based on a generalized variation of de Bruijn graphs. There are two broad strategies: error-correcting the reads before the assembly process (Canu, MECAT, Falcon, NextDenovo), which may lead to more structurally accurate assemblies, and performing the assembly with error-prone reads and polishing afterward (Miniasm, Flye, wtdbg, Smartdenovo, Raven), which is significantly faster [24]. Assemblies may be further improved by polishing them with short reads, resulting in so-called hybrid assemblies, though this is generally not necessary with PacBio's randomly distributed sequencing errors if the sequencing coverage is high enough [29]. HiFi reads offer new possibilities for long-read assemblers, which may bring fast de-Bruijn-based algorithms back into vogue [30, 31].

In general, the field is fast-moving. As the underlying sequencing technology evolves (e.g., HiFi reads), assembly strategies must adapt. While new assemblers are published every year, many established assemblers are under active development and successive versions may include major changes. Choosing the best assembler is not a simple task and may depend on the organism of interest, the sequencing technology, the read quality and depth, planned downstream analyses, and the software version. Compounding this problem, it is not simple to design fair benchmark studies for assemblers, or to compare genome assemblies.

In the most recent benchmark of long-read assemblers for 500 simulated and 120 real bacterial genomes, Wick and Holt conclude that not all tools perform well on all metrics and that Flye, Miniasm, NextDenovo, and Raven performed best overall [32]. This group went on to develop Tricycler, a software that takes as input assemblies based on the same reads made from different assemblers and guides the semi-automated creation of a consensus assembly [33].

1.1.5. Assembly quality control

Since many different, imperfect methods exist to generate genome assemblies, it is important to assess the quality of the result.

The most straightforward metrics to consider are the genome size, the number of contigs and the N50. The genome size should ideally be close to that of reference genomes of the same species or genus. Apart from that, the fewer and larger the contigs, the better the assembly. N50 is “the size of the smallest contig (or scaffold) such that 50% of the genome is contained in contigs of size N50 or larger” [34]. The software QUAST [35, 36] has been developed to calculate these statistics (and others). For long-read-based assemblies, these stats should perhaps be viewed more critically [37] and overruled by whether a chromosome could be circularized or all plasmids were found. Plasmids are a difficult challenge for certain assemblers, as they may have significantly different sequencing depths and unpredictable sizes.

A way to assess the quality of an assembly is to test whether the expected *universal single-copy orthologs* are present using BUSCO [38]. These are genes that occur only once in most genomes of a certain taxa. If many of these genes cannot be found, it suggests that the assembler dropped contigs and the assembly is incomplete. A high number of duplicated genes could reflect an assembly mistake or indicate contamination, e.g. multiple strains of different or the same species.

Particularly in short-read assemblies, certain contigs are sometimes suspected to originate from contamination. In order to screen for this, tools like Kraken [39], GTDB-Tk [40] and ConFindr [41]. The advantage of ConFindr and Kraken is that they can be used to analyze raw reads even before the start of the assembly process.

1.1.6. Gene prediction in prokaryotes

Often, the first step after genome assembly of prokaryotic genomes is to predict the genes in the genome. This process is called structural annotation or gene prediction. It is possible *in silico* because of certain properties of bacterial genomes. They are very gene dense: around 80-90% of the sequence consists of protein-coding sequences (CDS). Their prediction is simplified by the lack of introns. Steven Salzberg, a pioneer of the field and author of GLIMMER [42], put it like this: “Finding genes in bacteria is relatively easy (...). The gene-finding problem [in bacteria] is mostly about deciding which of the six possible reading frames (three in each direction) contains the protein.” [43] The process consists of a variety of steps, for example, accounting for GC-content bias, codon usage patterns using Markov models, learning which start codons are used by the organism and prediction of upstream patterns including promoters and ribosomal binding sites. Currently, the leading algorithms are Prodigal [44] and GeneMarkS-2 [45]. They generally perform very well and with high

agreement, though the results are not identical. In comparative genomics, it is thus important to use the same annotation pipeline for all genomes [46, 47].

There are two major annotation pipelines for structural annotation of prokaryotic genomes: Prokka [48] and NCBI's PGAP [49–51]. These utilize Prodigal and GeneMarkS-2, respectively, and bundle them with other algorithms that can detect non-coding genes like ribosomal and transfer RNAs. These are very conserved genes that can be detected with simpler methods, for example, small databases of Markov models. The advantage of Prokka is that it is very fast: it can be installed in minutes and takes around 5 minutes to process a genome. PGAP, in contrast, takes around 45 minutes per genome and requires downloading over 40 GB of data. The reason for the discrepancy lies in their purpose: Prokka is focused on structural gene prediction with few functional annotations, whereas PGAP also focuses on curating NCBI's genome databases. As such, PGAP performs some of the assembly quality controls described in the previous section and describes each gene in far more detail. In 2021, a pipeline called Bakta was published that also uses Prodigal and promises to be a “well-balanced tradeoff” between PGAP and Prokka [52].

1.1.7. Functional gene annotation

After structural gene prediction, the next task is usually to find out what their biological function might be. Prokka and particularly PGAP already include the first steps in this direction. Both contain a database of reference proteins and, for each predicted protein, use BLASTp to find the most similar one in the database, and then transfer the functional information of this reference protein to the predicted one [48, 49]. Accordingly, the size and quality of the reference database determines the number and quality of the new annotations. This highlights the importance of curated annotation resources such as UniProt/Swiss-Prot [53]. Researchers must keep in mind that the information in the reference database may be wrong, that a similar DNA sequence does not guarantee identical function and the fact that most annotations are strongly biased. Annotations of model organisms like *E. coli*, where most genes have been studied, are much more detailed and complete than annotations of unculturable organisms. Moreover, certain types of genes are better characterized than others, for example, genes of the core metabolism and genes important to humans, for example, virulence factors or antibiotic resistance genes. Overall, around 40-60% of predicted genes in a genome are unknown [54].

1.1.7.1. Types of annotations

There are many types of annotations with different properties. The following list, which includes the most important ones, illustrates the variety:

1. Systematic identifiers

Amongst other annotations, Prokka and PGAP assign to each new gene they find in the database a short description, for instance, “Ig-like domain 2 protein”. These descriptions are very useful as humans can readily interpret them. But they are not systematic: another gene may be called “Ig-like domain (group 2)”, which means the same but could cause problems for bioinformaticians. In contrast, standardized gene names (e.g., *dnaA*) and KEGG [55] orthologs (e.g., K00001) are examples of annotation types with systematic identifiers. In both cases, the identifier has a specific meaning that one can look up on the relevant database resource.

2. Exclusivity

Some annotation types are exclusive, meaning that one gene can only have one identifier of this annotation. KEGG ortholog annotations belong to this category: If a gene belongs to the KEGG orthogroup K00001, it cannot also belong to another KEGG orthogroup. In contrast, multiple KEGG reaction annotations may apply to the same gene, for example, a dihydrofolate reductase (*folA*) can catalyze two reactions: R00936 (Tetrahydrofolate \rightleftharpoons Dihydrofolate) and R00937 (Tetrahydrofolate \rightleftharpoons Folate).

3. Topology

Annotations of certain annotation types have relations with each other, for example, GO-terms [56]. The GO topology has three root nodes: GO:0005575 (cellular component), GO:0008150 (biological process), and GO:0003674 (molecular function). All other GO-terms are descendants of one of them and may themselves have child terms. For example, “adaptive immune response” is a descendant of “immune response”, and both are ultimately descendants of “biological process”. EC-numbers [57] are also organized topologically. They stand for enzymatic reactions and connect educt metabolites with product metabolites, together forming metabolic pathways.

Such topologies can be exploited computationally as they allow genes to be analyzed in context, for instance in enrichment analyses and metabolic models.

4. Orthology

All previously mentioned annotation types are about known biological properties. In comparative genomics, it is possible to group similar genes into orthogroups based solely on their sequence similarity and phylogenetic reconstruction. This process yields orthogroup identifiers without any functional meaning that nevertheless function like annotations. It enables treating all detected genes in an unbiased way.

1.1.7.2. Annotation using sequence similarity

The most common way to assign functional annotations to new genes is to perform a sequence similarity search against a reference database of annotated genes, the single-directional best hit (SBH) strategy. However, this strategy depends on arbitrary cutoffs. For instance, BLAST does not report whether a hit has the same function. Instead, for every hit, BLAST reports how many percent of the query sequence are covered to the hit, the fraction of identical nucleotides or amino acids between the query and the hit, and various other scores that are difficult to interpret. Prokka deals with this by implementing an *e*-value cutoff (10^{-6}) [48]. Early versions of PGAP filter by sequence identity (25%) and coverage (70%) [49]. Because there are no generally optimal cutoff values, newer PGAP versions include “BlastRules” that apply different cutoffs to different reference proteins [50]. One possibility to reduce false positives is to use a bi-directional best hit (BBH) strategy, where hits only count if a sequence similarity search of the hit back onto the new assembly finds the original query gene. This strategy is implemented in KEGG’s KAAS annotation server [58].

1.1.7.3. Annotation using orthogroup inference

Two genes are orthologs of each other if they share a common ancestor (homology) and emerged through a speciation event. In contrast, paralogs emerged through a gene duplication event. While this is not always true, orthologs tend to have conserved functions whereas paralogs tend to diverge in function (“ortholog conjecture”). Orthology-aware functional predictions have higher precision than SBH approaches because they avoid transferring annotations from close paralogs [59]. The BBH strategy is, in fact, a simple strategy to determine orthologous genes between two genomes, but it is not directly applicable to multiple genomes. As a result, different approaches have been proposed to solve the problem of reliably determining gene orthology amongst many genomes in a timely manner [60].

The eggNOG resource (v5.0) is a public database that consists of genes from 25,038 genomes, including 4,445 bacteria. These genes were divided into 4.4 million fine-grained, functionally annotated orthologous groups [61]. The eggNOG-mapper tool makes it possible to efficiently annotate new protein sequences. The first step is to download the 51 GB eggNOG database. Next, each protein is mapped against a representative subset of the proteins in the eggNOG database. The best hit from this search (the “seed ortholog”) is used to narrow down the search space to close orthologs and paralogs. Once the best ortholog is found, its annotations are transferred to the new protein [62]. Though this method does not solve the problem of cutoffs during sequence similarity searches, the orthology-aware algorithm as well as the ever-increasing scale of the database mitigate the problem. EggNOG-mapper performs much better than the sequence-similarity-based methods

BLAST and InterProScan at annotating proteins with GO-terms [63]. On top of providing high-quality annotations, it is easy to use and fast, and provides annotations from many different sources (BiGG, CARD, CAZy, GO, KEGG, PDB, PFAM, SMART and eggNOG ortholog identifiers) so that it is usually sufficient for downstream analyses.

Nevertheless, eggNOG usually does not include all proteins of a new genome. Thus, to ensure that all proteins of any given pangenome are included in a comparative genomics analysis, it is sometimes necessary to apply ortholog inference algorithms to the genomes of interest. In the past few years, OrthoFinder [64] consistently outperformed its competitors in the Quest for Orthologs benchmark service study [65].

1.1.8. Genome data management

Since the revolution of second-generation sequencing, many institutions and research groups have accumulated large genomic datasets. The generation, management, curation, exchange, and efficient analysis of these data is a demanding interdisciplinary challenge that involves biologists, sequencing experts, and bioinformaticians. For instance, when questions arise about the quality of assemblies, expertise in all three of these fields is required to make good decisions. Bioinformaticians are often in the middle of these discussions and can do the most to smoothen the process. I will highlight six aspects:

1. Data organization

When sequencing projects start, it is not always clear that they may last for a very long time. Bioinformaticians may process each batch of new data in a separate folder. Within an organization, multiple groups may generate sequencing data but fail to share it with each other. To maintain order and reproducibility, the data should be systematically organized in only one location. The structure should be well-documented and preferably be the same across different projects to make it easier for new people to start working with the data. A standardized structure also makes it possible to process the data automatically.

2. Metadata

In long-term projects, different biological extraction approaches, different sequencing, assembly, and annotation methods as well as different people will eventually be involved. One cannot rely on one person's memory to remember how data was generated. Thus, meaningful log files should be generated and stored together with the genomic data, and the most important metadata properties should be stored in a format that can be processed automatically. For instance, during data analysis, it may be crucial to know which genomes were assembled using short or long reads.

3. Unique identifiers between genome versions

Whenever a genome is re-sequenced, re-assembled or re-annotated, new gene identifiers are generated. Especially in the long run, it is very important that no confusion between different versions of a genome sequence arises. Genome database managers should ensure that these identifiers are unique and that it is always clear to which version of a genome they belong.

4. Identical processing

If annotations are not provided by bioinformaticians, biologists may waste precious time finding ways of generating them. These may be sub-optimal and will likely not be shared efficiently within an organization. Therefore, this step should be integrated in the bioinformatics pipeline. Moreover, annotations from different tools (or different versions of tools) may differ significantly. To prevent systematic confounders as much as possible, all data should be treated the same way.

5. Provide biologists with the tools they need

Biologists who work with genomic data on a daily basis often discover problems that were overlooked by automated processing. For instance, a newly sequenced strain may not belong to the expected species upon closer inspection. Alternatively, they may discover that a certain genome is most likely contaminated. While not everyone should be able to change the metadata, the curators of the database should be able to make such changes themselves without involving a bioinformatician. More broadly, there are certain common workflows with genomic data that can be automated, making them accessible to non-bioinformaticians in a standardized, convenient, and fast manner. Such tools save a lot of time and money in the long term and greatly amplify the value of the data itself.

6. Generalization

If the same data organization system is used for multiple projects, a script that works on one can be used on another without much adaptation. Similarly, automated workflows, as described above, could benefit multiple projects. While many platforms have been developed to manage specific genome datasets and simplify common workflows, none can easily be re-used for other projects. To be usable in different fields, the software must be flexible enough to work with data from different sequencing technologies, assemblers, and structural annotation pipelines. Annotation types are particularly project-dependent: In a hospital, the most relevant genes in microbial genomes might be related to antibiotic resistance or virulence whereas glycobiochemists are most interested in carbohydrate metabolism. The former may thus choose ABRicate [66] and the latter the CAZY database [67] to annotate their genes. Generalist software should be configurable enough to work with any annotation type.

1.2. The human gut microbiota

The human gut is a key organ, tasked with the digestion and uptake of nutrients. Most of the absorption occurs in the small intestine. For this reason, it has a surface area of approximately 30 m² [68] that, from the point of view of the immune system, is the most exposed and vulnerable frontier to the “outside” of our body. It is colonized by trillions of microbes, collectively known as the human gut microbiota, comprised of bacteria, archaea, fungi, protozoa, helminths, phages, and viruses. To keep the gut microbiota in check, the immune system delegates most of its immune cells there [69]. Hereafter, the “gut microbiota” is just referred to as “microbiota” and the term “microbiome” refers to their collective genome [70].

The microbiota is not just a challenge, it also performs many important functions for the host. These include digestion of host-indigestible fibers, vitamin synthesis, promotion of normal gastrointestinal function, out-competition of pathogens, and contribution to the maturation and education of the immune system. The interactions between the microbiota and its host are numerous and complex, having co-evolved for millions of years [71], and the microbiota has been described as a virtual endocrine organ [72] because many of its metabolic outputs influence the host. For instance, certain bacteria can affect the host’s mood or appetite while others modulate immune and inflammatory responses [71–74].

Some metabolites produced by the microbiota are also harmful, for instance, trimethylamine (TMA) which is formed by bacterial catabolism of precursors like choline or carnitine and is associated with cardiovascular diseases [74]. Many diseases are associated with an altered *composition* of the microbiota, including Crohn’s disease (CD), ulcerative colitis (UC), type-2-diabetes (T2D), coeliac disease, asthma, anxiety, depression, and certain cancers [75]. (By contrast, microbiota *density*, which is harder to measure, often gets overlooked.) It is a difficult challenge, to understand and precisely define microbiota homeostasis or dysbiosis, because of the large inter-individual variation of microbiota composition. In other words, there are different ways in which the microbiota can be homeostatic or dysbiotic, making it impossible to define these terms simply and with confidence by

the presence, absence, or ratio of specific microorganisms [76, 77]. Meanwhile, it is generally agreed that a diverse microbiota, i.e. one that contains many different species, is healthy: not because diversity is a priori beneficial, but because diversity provides robustness through functional redundancy and coverage of more functional niches [78]. Species richness, also termed alpha diversity, can be measured using various indices, such as the Shannon index [79] or Simpson index [80]. Shannon entropy (H) is calculated as follows:

$$H = - \sum_{i=1}^P p_i \ln p_i$$

where p_i is the proportion of the i th species in the sample. However, it is crucial to understand whether the microbiota is the cause or merely an effect of the disease in question, though this is often neglected [81]. In some instances, the causal relationship has been established. For example, Turnbaugh et al. [82, 83] transplanted fecal microbiota from obese or lean donor mice to germ-free mice and found that mice colonized with an “obese microbiota” developed significantly more body fat than their littermates colonized with a “lean microbiota”. In the same way, adiposity is even transmissible from humans to mice [84]. Similar experiments show that the “dysbiotic” microbiomes associated with inflammatory diseases such as UC [85] and CD [86] are also transferrable and alone sufficient to cause the disease. This raises the question of which factors shape the microbiota.

1.2.1. Environmental influences that shape the microbiota

These appear to be mainly environmental rather than genetic factors. Household sharing is a better predictor of microbiome composition similarity than relatedness, and the heritability of inter-person microbiome variability has been estimated to be between 1.9% and 8.1% while over 20% could be explained by factors related to diet and lifestyle. The remaining variability remains unexplained [87]. Since the microbiota is important for human health and the microbiota is primarily shaped by the environment, what are some of the key environmental influences?

The intestinal lumen is separated from the host by structural barriers: the epithelial cell surface and an inner as well as an outer, looser mucus layer. The microbiota is controlled via the secretion of α -defensins (small, antimicrobial peptides), RegIII γ (an antibacterial peptidoglycan-binding protein) and IgA [71]. Moreover, the host controls the availability of the terminal electron acceptors with high redox potentials in the small intestine: oxygen and nitrate. The duodenum and ileum are characterized by higher oxygen and nitrate concentrations, respectively, favoring facultative anaerobes which can use these molecules for respiration, which yields more energy than fermentation. Conversely, in the colon, both molecules are scarce, favoring obligate anaerobes that efficiently produce energy through fermentation. Weakened host control mechanisms are associated with dysbiosis and can be influenced by the diet [77].

In utero, humans are sterile, i.e., they do not possess a microbiota and acquire it through environmental exposure during and after birth. A key factor appears to be the early development, the so-called “neonatal window of opportunity”. In this period, the microbiota evolves together with the newborn’s immune system, and, for better or worse, external influences have a greater and more durable impact than during adult life. Adult-like microbiota diversity is only reached by approximately age 2-3. The early microbiome is shaped by factors like nutrition (breast milk or formula, then solid diet), delivery (vaginal or cesarean), presence of siblings, pets or farm animals, antibiotics, and genetics. While some studies in this field produced spurious results due to the many factors involved (as well as publication biases), many of the observed trends are congruent with the “hygiene hypothesis” [88]. It states that microbial exposure is critically important for the development of a healthy and robust immune system. It could explain the correlation between modern lifestyles, characterized by enhanced hygiene, limited exposure to animals, fewer infectious diseases, higher antibiotic use, and increasing modern immune-mediated and allergic diseases [89, 90].

Another major influence on the gut ecosystem are electron donors, which come from the diet. Many host-digestible nutrients are efficiently absorbed in the small intestine. However, host-indigestible nutrients such as fiber (plant-derived polymers like cellulose, resistant starch, or beta-glucans) and poorly absorbed carbohydrates (FODMAP) remain available as the main source of electron donors for the colonic microbiota [77]. The major products of the digestion of these nutrients are short-chain fatty acids (SCFAs) like acetate, propionate, and butyrate. SCFA affect the intestinal immune and inflammatory responses and are taken up by, resulting in far-reaching effects in the host, including on the brain [73]. Gut microbes also digest proteins and ferment amino acids, producing SCFA, but also branched-chain fatty acids (BCFA) and potentially toxic substances such as TMA, depending on the type of protein. Fats, food additives (sweeteners and emulsifiers), micronutrients (vitamins and minerals), polyphenols, and salt were also shown to affect the microbiome. Since not all microbes thrive equally well on these chemically diverse nutrients, the microbial composition changes with the nutritional environment [91].

Changes in diet can result in rapid and dramatic shifts in the microbiota, but if the change is merely a short-term intervention, the microbiota tends to return to baseline within a few days. Instead, habitual, long-term dietary patterns appear to play a role in shaping the microbiota [92]. Diverse, fiber-rich diets, particularly diets that consist of many different types of plants, are associated with improved health measures as well as microbiota diversity [93]. Though diets high in fermented foods are less well studied [94], a 17-week prospective study that compared a high-fiber and a high-fermented-food diet found that the latter increased microbiome diversity and decreased markers of inflammation [95]. Conversely, calorie-dense “Western” diets (characterized by prepared meals, processed food, higher intakes of refined carbohydrates, added sugars, fats, and animal-source foods), are associated with obesity and loss of microbiome diversity [83, 96, 97].

1.2.2. How to intervene in the microbiota?

Since microbiota dysbiosis contributes to certain diseases, it makes sense to develop strategies to reinforce microbiota homeostasis and host control or to mitigate or prevent certain diseases. These strategies are reviewed in Hitch et al., 2022 [78] and summarized in Figure 2.

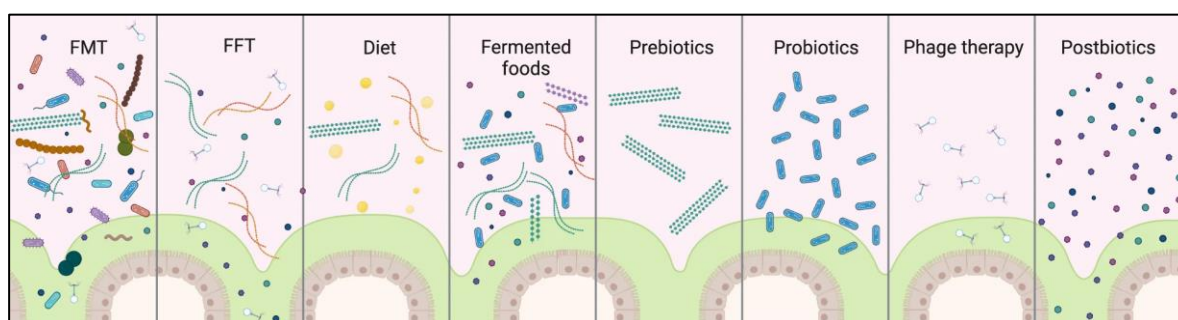


Figure 2: Overview of microbiota-based intervention methods, ordered on the complexity of the interactions with microbiota and immune system. Most complex (left) to least complex (right). FMT means fecal microbiota transplant, FFT means fecal filtrate transplant. **Legend:** Intestinal epithelial cells (bottom) surrounded in mucus (light green), prebiotic polymers (dark green, dark red, violet), lipids (yellow). Figure adapted from Hitch et al., 2022 [78].

In 2006, probiotics were defined as “live microbes which, when administered in adequate amounts, confer a health benefit to the host” in a joint report by the Food and Agriculture Organization of the United Nations and the World Health Organization [98]. Probiotics are associated with some of the following attributes: (i) human origin, (ii) generally recognized as safe (GRAS) status, (iii) survive gastrointestinal transit, (iv) adherence to intestinal cells/mucus, (v) antagonism towards pathogenic microbes, (vi) positive effect on host immune system, and (vii) positive effect on microbiome [99].

1.2.3. Yogurt

Yoghurt offers a particularly intriguing strategy to influence the microbiota, ticking most boxes in Figure 2: it is a culturally accepted, fermented source of microbes that may combine high doses of probiotics (healthy, live microbes), prebiotics (food for probiotics), postbiotics (healthy, inanimate microbe-produced metabolites) and potentially phages. Moreover, especially dairy-related species of LAB are well-studied and fit many attributes of probiotics listed above.

Yoghurt is generally acknowledged to be healthy. In their systematic review, Savaiano and Hutkins conclude that the strongest yoghurt-associated health benefits are related to improved lactose digestion and tolerance, but yoghurt is also associated with various benefits including reduced risk of type 2 diabetes, weight maintenance, and improved gastrointestinal and cardiovascular health. At the same time, probably because yoghurt is such a complex food, they state that multiple mechanisms could explain these findings [100].

Typically, the bacteria used to create yoghurt are LAB, which today means belonging to the order *Lactobacillales*. Most LAB are aero-tolerant anaerobes, lacking a complete tricarboxylic acid (TCA) cycle and often an efficient, iron-dependent electron transfer chain [101]. Instead, characteristically, they utilize carbohydrates to generate lactate as main metabolic product, though some also produce acetate, acetoin, or ethanol. Some LAB are excellent at growing rapidly and inhibiting competitors through accumulation of lactate and acetate, making them ideal food fermenters [102]. Many LAB have exacting food requirements that are associated with nutrient-rich habitats, including food, plants and animals. LAB are a particularly phylogenetically and metabolically diverse and may be free-living, nomadic or host-adapted [103]. Some species evolved a symbiotic lifestyle which is associated with substantial genome decay [104]: while the genome size of the free-living *Lentilactobacillus parakefiri* is 4.91 Mb, that of the host-adapted *Lactobacillus iners* is only 1.27 Mb [103]. In species such as *S. thermophilus*, who have many new pseudogenes, this process of decay is ongoing [104]. Certain strains have adapted to human-made habitats such as fermented foods despite the recent evolutionary emergence of these habitats. The ancestral ecological niche of these microorganisms remains to be determined [103]. According to their need for a nutrient-rich environment, their anaerobicity and their aero-tolerance, in the human gastrointestinal tract, LAB generally inhabit the small intestine [105–109]. Because of the difficulty sampling the small intestine and the fact that LAB are continuously taken up through the diet, it is difficult to determine which species are resident or transient inhabitants. Similarly, the proportion of LAB relative to other microbes is difficult to determine but considered to be small [109].

Humans have been making yoghurt and other fermented milk products for thousands of years. Our ancestors added yoghurt to boiled milk to create more yoghurt, unaware of the existence of microbes. Accordingly, traditional yoghurts differed by geographic origin and were evolving ecosystems consisting of different strains and species of LAB, but also yeasts, and possibly other microbes [110, 111]. Today, most industrially produced yoghurts comprise merely two starter strains, both LAB, one belonging to the species *Streptococcus salivarius subsp. thermophilus*, the other to *Lactobacillus delbrueckii subsp. bulgaricus* [110, 111]. Fortunately, some of the traditional repertoire of dairy bacteria has been preserved in bacterial strain collections such as the Agroscope Liebefeld collection which comprises more than 10,000 isolates collected over a century of dairy research [112]. By mid-2018, 631 of these bacteria were sequenced and made available in a digital genome repository termed *Dialact*.

1.3. Polyfermenthealth

The goal of the Polyfermenthealth project is to exploit the genetic potential contained in collections of lactic acid bacteria (LAB) to design new yoghurts with possible additional health benefits and to understand the underlying mechanisms [113]. Hence the name “Polyfermenthealth”,

referring to the creation of yoghurts *fermented* with multiple (*poly*) bacterial strains to create a *healthy* yoghurt.

The project is an interdisciplinary collaboration between Agroscope, Inselspital, and the University of Bern and consists of multiple steps: (1) genetic screening of bacterial strains, determination of promising strategies, and strain pre-selection; (2) production of initial polyfermented yoghurts, phenotypic characterization, and selection of final polyfermented yoghurts; (3) testing of the final yoghurts in mice, for example regarding fate of the metabolites, integration into the microbiome, microbiota composition, the transcriptome, and the immune system; (4) integrative data analysis of the metabolomic and genomic data.

The project proposal already included a concrete strategy for one yoghurt, namely the targeting of the aryl hydrocarbon receptor (AhR). A second strategy, the maximization of yoghurt metabolic diversity, was hinted at, and the remaining strategies were left open. The following subsections introduce the background for the three different strategies that we chose to design new yoghurts.

1.3.1. A “polydiverse” yoghurt

We decided to design a yoghurt with high taxonomic as well as metabolomic diversity, to exert the strongest possible impact on the microbiome. We decided to add 5 strains, each from a different species to the yoghurt in addition to the starter culture. We selected these additional strains based on two criteria: (i) We wanted to increase the chance of the strains being able to engraft in the gut as long as possible. To this end, we prioritized strains with genes that are known survival factors, for example, genes that confer bile or acid resistance, or adherence to mucus [114]. (ii) New strains would be added iteratively to maximize the metabolic diversity of the yoghurt. The final yoghurt was called “polydiverse” because it consists of many different species. We were motivated by the two following reasons:

First, LAB are generally associated with beneficial health outcomes. While many LAB from fermented foods reach the gut microbiome alive, they tend not to persist because they are not adapted to the environment. Continuous ingestion of LAB via fermented foods such as yoghurts may thus be the source of these bacteria found in the microbiome [109]. Since most modern yoghurts only consist of two species, an opportunity to positively contribute to microbial diversity might be missed.

Second, because they do not generally persist in the gut, it has been suggested that the metabolites produced by LAB, produced in high concentration during fermentation, may be the mechanism by which they benefit their host [78, 115].

1.3.2. Indoles and the aryl hydrocarbon receptor (AhR)

Indoles are among the microbiota-produced metabolites with immunomodulatory properties. One known pathway that leads to indole formation is the deamination of tryptophan by microbial tryptophanases. In the liver, indole is then converted into 3-indoxylsulphate, which can bind the aryl hydrocarbon receptor (AhR). The AhR is expressed in many organs, including the placenta, lung, heart, pancreas, liver, kidney, brain, muscle, and the intestine. Amongst other roles, at physiological barriers such as the intestine, the receptor is involved in the catabolism of xenobiotics and defensive immune signaling [116]. In the adaptive immune system, the AhR is expressed by dendritic cells and certain T cells, including T_{H17} and T_{reg} but excluding naive $CD4^+$ T cells, T_{H1} , and T_{H2} . In the intestine, it is expressed by intestinal epithelial cells, intraepithelial lymphocytes, and innate lymphoid cells (ILC), particularly the subclass ILC3 [117]. The AhR can induce a plethora of pathways (Figure 3).

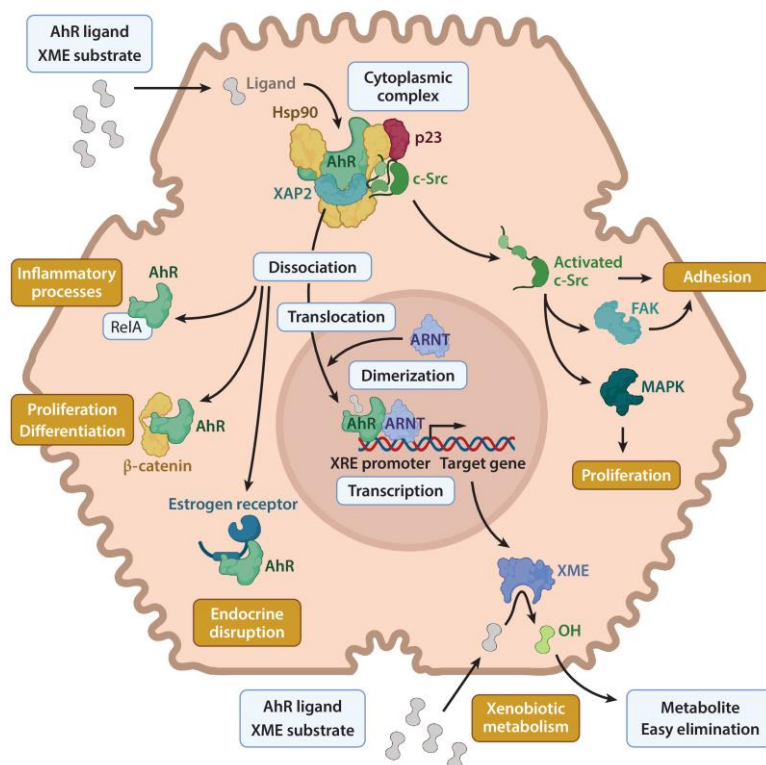


Figure 3: AhR signaling pathways.

The AhR is normally located in the cytoplasm. When a ligand binds, the complex dissociates, and may activate one of the following pathways: (i) translocation of ligand-AhR-complex into the nucleus, activating xenobiotic response elements (XREs) involved in detoxification; activation of c-Src, impact on (ii) cellular adhesion and (iii) proliferation; release of the ligand from AhR, interaction with (iv) NF- κ B members (RelA), (v) β -catenin, or (vi) estrogen receptor, impacting of adhesion, proliferation, differentiation, inflammation and endocrine disruption. Figure from Larigot et al. 2022 [116].

However, 3-indoxylsulphate is only one of many AhR ligands [118], and the activity of the AhR depends on the type of ligand. For example, superagonists cause toxic effects, including skin damage and, through long-term exposure, birth defects, reduced fertility, immunotoxicity, and thyroid disruption. The most famous example is 2,3,7,8-tetrachlorodibenzo-*p*-dioxin, a contaminant of the herbicide Agent Orange used in the Vietnam war. However, most diet-derived ligands only activate the AhR with low affinity [116]. Different molecules may have different or even opposite effects: *in vitro* experiments on human colon cancer cells revealed that tryptamine and indole 3-acetate (I3A) are AhR agonists whereas indole is an AhR antagonist [119]. If AhR is activated weakly, it seems to induce a pro-inflammatory response, while a stronger activation supports immune tolerance. Moreover, host metabolism of the activating ligands may be crucial as it determines the duration of the activation [120]. Multiple molecules can bind to the receptor at the same time, sometimes producing synergistic effects [121]. Interestingly, non-indole compounds such as vitamin B12, folic acid and glucose are also AhR antagonist [122, 123]. A further complication is that the human AhR and the mural AhR have different ligand binding profiles as well as different downstream effects [124].

Certain vegetables are known to contain ample amounts of AhR ligands, for instance, broccoli which contains indole-3-carbinol (I3C). The essential amino acid tryptophan is another source, but indirectly. Most of the tryptophan that humans take up is catabolized via the kynurenine pathway, generating AhR ligands. Similarly, gut microbes are known to transform tryptophan into AhR ligands [117].

The activation of AhR appears to have both positive and negative effects, and they may depend on the cell type expressing AhR, co-morbidities and other idiosyncrasies. In *in vitro* experiments, AhR antagonist indole had both pro- and anti-inflammatory effects and decreased the attachment of pathogenic *E. coli* to HCT-8 cells [125]. Blood serum AhR activity was correlated with parameters of

insulin resistance and significantly higher in T2D compared to non-T2D subjects [126] and AhR expression is higher in T2D patients [127]. Conversely, in the feces of patients with inflammatory bowel disease (IBD), tryptophan catabolites and AhR activity are decreased compared to healthy patients [128], and the microbiota from *Card9^{-/-}* mice that fails to metabolize tryptophan into AhR ligands increases the susceptibility of germ-free mice to colitis compared to microbiota from wild-type mice. These symptoms could be reduced by the administration of AhR agonist-producing *Lactobacillus* strains. Furthermore, AhR agonists like indole-3-carbinol may be crucial in the development of a mature immune system [129, 130]. The AhR is also key for the development and homeostasis of the liver and is required to adjust liver regeneration after acute toxic injury. Interestingly, *Ahr^{-/-}* mice recovered better after severe liver damage [131].

Notably, indoles can also affect humans in other ways: they are among the compounds that can activate the pregnane X receptor (PXR). PXR is, like AhR, involved in xenobiotics degradation, regulation of the intestinal barrier and inflammation, but also metabolism of lipids, glucose, and bile acids [132]. Some indoles can inhibit K^+ channels and impact the secretion of GLP-1, a peptide which effects insulin and glucagon release that is relevant in T2D. Indole can modulate the IFN1 pathway in mice, 3-methylindole can induce cell death in intestinal epithelial cells, tryptamine modulates intestinal ionic flux and indoleacrylic acid (IAC) can suppress inflammation via the Nrf2 transcription factor [133]. This illustrates that many indole metabolites are bioactive and may act through multiple different pathways on systems with relevance to human health. Almost certainly, many indoles and various mechanisms of action are still unknown.

1.3.2.1. A yoghurt to activate the AhR

Accordingly, there is interest in AhR-agonists such as indoles as potential anti-inflammatory drugs to combat various conditions either as small molecule drugs or via fermented foods (Figure 4) [134, 135].

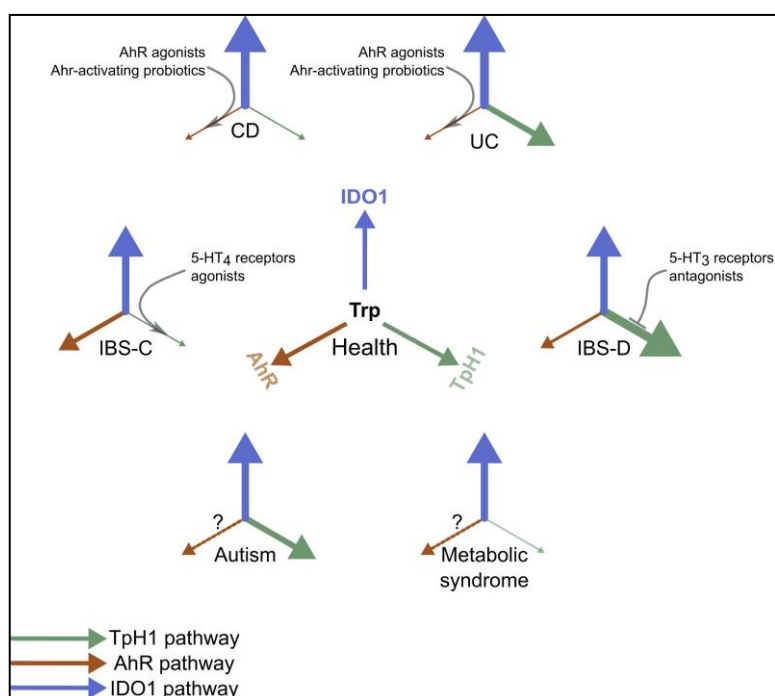


Figure 4: Perturbations to tryptophan metabolism in diseases and potential for therapeutic strategies.

The three major pathways of tryptophan metabolism are tightly interconnected and differentially affected in diseases. IDO1 is the rate-limiting enzyme that leads to the kynurenine pathway. TpH1 is the enzyme required for over 90% of serotonin production. Diagrams indicate repartitioning of tryptophan fluxes in diseases. Weights of arrows indicate the strength of pathway activation. The restoration of disrupted equilibrium using molecules or probiotics represents a promising therapeutic strategy. **Legend:** AhR, aryl hydrocarbon receptor; IDO1, indoleamine 2,3-dioxygenase 1; TpH1, tryptophan hydroxylase 1; 5-HT, 5-hydroxytryptamine; CD, Crohn disease; UC, ulcerative colitis; IBS-C, irritable bowel syndrome with constipation; IBS-D, irritable bowel syndrome with diarrhea. Figure from Agus et al. 2018 [135].

Milk is a particularly tryptophan-rich food, and the fermentation of yoghurt increases the amount of free tryptophan in the substrate [136]. Moreover, many LAB have the metabolic capability to transform tryptophan into AhR ligands [137]. The changes in tryptophan and indole catabolites are also known to propagate into blood upon consumption: In a human trial conducted at Agroscope, the postprandial effects of a yoghurt fermented with a probiotic *Lactobacillus rhamnosus* strain were compared to acidified milk. The researchers found significant differences in serum tryptophan and tryptophan catabolites between the two groups as well as significantly lower AhR expression in the yoghurt group [138–140]. This suggests that yoghurt fermented with specific LAB might be an ideal vector to affect tryptophan metabolism.

Various publications describe beneficial effects of certain potentially probiotic strains where the AhR was discussed as a plausible mechanism of action [125, 141–143]. Despite this, there appear to be very few published attempts [144] to create yoghurts with maximally increased AhR activating capacity using indole-producing strains. One reason for this may be that many AhR ligands, their binding profiles as well as the genes that produce them are unknown [137], making more extensive phenotypic screening necessary.

A key advance to study the effect of the microbiota on the immune system was the invention of reversibly (or transiently) colonized germ-free mice. It consisted of the development of the *E. coli* K12 mutant HA107 which is auxotrophic for meso-diaminopimelic acid and D-alanine. Mammals do not depend on either metabolite, but they are key building blocks of peptidoglycan, the polysaccharide that constitutes the bacterial cell wall. As a result, germ-free mice can be gavaged with HA107, which without supplementation of both metabolites disappear within 1-3 days, leaving the mice germ-free again. Reversible colonization is achieved by regular gavage of HA107 or supplementation of the two metabolites. This technique was key to characterizing the intestinal IgA response, which is also adaptive and has a short-term memory (in mice with a normal microbiota), thus requiring interrupted phases of colonization [145]. It was also key to study the influence of the maternal microbiota on their pups. Persistent colonization of the mother mouse would have led to the colonization of the pups during birth, making it impossible to separate the effect from the maternal microbiota from the effect of endogenous colonization of the offspring. In a series of experiments, pregnant mice were transiently colonized with HA107 and gave birth to germ-free pups, which they then nursed. These pups, compared to pups of non-colonized mothers, exhibited increased numbers of certain innate immune cells in the small intestine (NKp46⁺ ILC3s and F4/80⁺CD11c⁺ intestinal mononuclear cells), a different intestinal transcription profile, and better resistance to *Bacteroides fragilis* colonization. Some of these changes persisted up to 60 days after birth. Moreover, it was found that IgG-containing serum from gestation-only colonized mice sufficed to induce the increase of ILC3, and further experiments showed that indole-3-carbinol (I3C) feeding to the dams during pregnancy has the same effect on the ILC3s in the offspring, likely through AhR signaling (though I3C is also a PXR ligand) [130].

In this context, we hypothesized that our indole-rich, AhR activating yoghurt would have a similar effect on the pups when fed to mothers during pregnancy. These experiments were conducted with germ-free mice, and the goal was to separate yoghurt-derived metabolite signaling from commensal effects. Hereby, one hurdle was the sterilization of the yoghurts as yoghurts are densely populated with live LAB and the sterilization process should not destroy the indole metabolites.

1.3.3. Folate

Folate, also known as vitamin B₉, is an essential nutrient that cannot be produced by the human body. The term “folate” refers to multiple chemically similar molecules that can be interconverted (Figure 5). Biochemically, the main function of folate is as a coenzyme that transfers “one-carbon groups”, primarily methyl (CH₃), methylene (CH₂), and formyl (CHO) groups. These reactions are essential in many major cellular processes, such as the synthesis of nucleotides (dAMP, dGMP, and

dTMP), amino acid homeostasis (methionine, serine, and glycine), epigenetics (specifically, DNA methylation) and redox homeostasis. Folate is circulated in the body in monoglutamate form. To retain folate within the cell (or the mitochondrion), folate is polyglutamated to penta- and hexaglutamate forms, which are less readily transported across membranes [113, 146, 147].

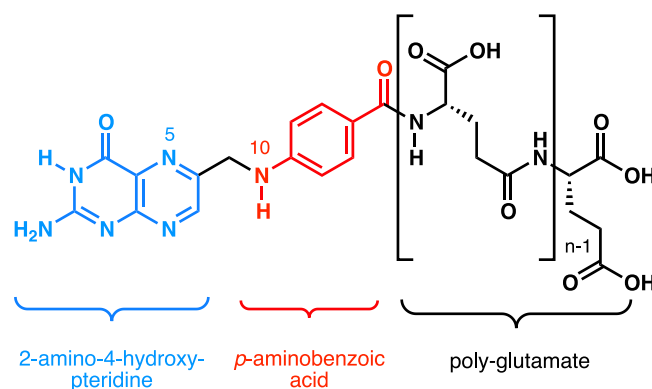


Figure 5: Chemical structure of folate.

Commercial folate (termed folic acid) only has one glutamate residue ($n=1$). Methyl groups are added to the N5 and/or N10 atoms. The main metabolically active form of folate, L-5-methyltetrahydrofolate (5-Me-THF), carries a methyl group on N5 and has a reduced pteridine moiety. The demethylated form of 5-Me-THF is termed tetrahydrofolate (THF). Illustration by Boghog 2019 [148], distributed under the CC BY-SA 4.0 license.

Folate deficiency in pregnant women can cause neural tube defects in the fetus. For this reason, folic acid supplementation is advised before and during pregnancy. More generally, folate deficiency can lead to macrocytic anemia, Alzheimer's disease, cardiovascular diseases, and may increase the risk of cancer [147]. Other than pregnant women, groups that are particularly at risk are elderly people, characterized by lower food intake, and children who consume a limited variety of food [99]. Sufficiently high levels of dietary folate may be important in long-term genome stability and health [147]. However, studies that examined the effect of folate supplementation on cardiovascular disease and cancer development found no strong benefits [146, 147]. While over 80 countries (including the US and UK) have introduced mandatory fortification of flour with folic acid, as of July 2021, neither Switzerland nor any EU member state have taken this measure [149, 150].

1.3.3.1. A yoghurt rich in folates

Thus, the development of folate-enriched yoghurt may benefit certain at-risk groups. While milk contains relatively low amounts of folate (20-50 $\mu\text{g/l}$ [151]), fermented dairy products such as yoghurt have long been known to contain higher amounts. But just how high should a yoghurt's folate content be to make a meaningful contribution? The recommended daily dose for adults is 400 μg per day according to the World Health Organization, and a "good" source of folate is one that provides at least 10-20% thereof [152]. Under the assumption that one portion of yoghurt is 200 ml, that would mean a concentration of around 200-400 $\mu\text{g/l}$.

The folate concentration of yoghurt can be increased manyfold by selecting appropriate starter strains [152] or additional strains [99]. Combinations of strains may produce synergistic effects: Folate biosynthesis consists of three parts, corresponding to the three substructures shown in Figure 5. *In vitro* experiments have shown that high amounts of the middle sub-structure, *p*-aminobenzoic acid, stimulates folate production [151]. It follows that one strain might produce large amounts of this compound, stimulating another strain to produce more folate.

Many commercial yoghurts are artificially fortified with chemically produced folic acid. However, there appear to be health issues associated with this form of folate. The human enzyme (dihydrofolate reductase) that converts folic acid into its active form (THF) has a very low activity. For this reason, taking supplements can cause a build-up of unmetabolized folic acid in the body. This excess then competes with active forms of folate, ironically having opposite effects as intended

[153, 154]. In contrast to synthetic folate, the majority of microbe-produced folate is in an active form. There are other differences compared to folic acid, though. Many natural folates are polyglutamated, resulting in a lowered bioavailability of around 80%. In yoghurt, this may be balanced out by folate-binding protein from milk, which protects folate from absorption by the microbiota [99].

Folate has manifold effects on the immune system. Folate deficiency inhibits the proliferation of CD8⁺ T-cells [155] and reduces the cytotoxicity of natural killer cells (NK) [153], increasing susceptibility to viral infections and potentially tumors. Colonic T_{reg} express folate receptor 4 and require folic acid to survive. Reduction in T_{reg} population has a pro-inflammatory effect and favors induced colitis [156]. Folic acid also reduced the lipopolysaccharide-induced inflammatory response in THP-1 monocytes through methylation of proinflammatory genes [157]. Moreover, since folate is particularly important in fast-growing tissues, antifolates (molecules that disrupt folate metabolism) have been developed for chemotherapy. However, it was found that antifolates also kill immune cells, increasing side-effects and limiting the doses that can be administered to patients [147].

With this in mind, we wanted to create a yoghurt rich in folate and test its effect on gnotobiotic mice as well as germ-free mice, with a particular focus on immune cells. From the bioinformatics side, the goal was to leverage the genomic information about the strains to reduce the number of strains whose folate-producing ability we would have to measure [158]. Measurement of folate is complicated by its sensitivity to light, oxidation, heat and acid [99]. In this project, a microbiological assay was used, where the growth of the auxotrophic *Lactocaseibacillus rhamnosus* is limited by amount of folate in the sample. The transmittance of the mixture after incubation indicates the amounts of folate produced [158]. In contrast, folate has a low irradiation sensitivity, enabling sterilization for germ-free experiments [159].

1.4. Linking metabolomics with genomic data

Today, researchers can measure both genotypes as well as phenotypes, for example transcriptomics, proteomics, and metabolomics, with high throughput. However, integrating these data to generate meaningful knowledge or useful hypotheses remains challenging. Perhaps counterintuitively, high dimensionality, for instance, large numbers of genetic differences measured per biological sample, makes it harder to analyze the data. This phenomenon is termed “curse of dimensionality” [160]. It occurs because as the number of features increases, the amount of data needed to accurately model the relationships between those features increases exponentially. In other words, the key to success in “big” data is not just a dataset with many *features*, but primarily a dataset with many *samples* to achieve enough statistical power.

1.4.1. Existing approaches

Most published algorithms that attempt to achieve holistic, integrative omics analysis are based on finding reliable patterns between groups of samples, i.e., in the simplest case, a control and a treatment group. One application of this is the field of biomarker discovery, where the goal is to find robust metabolic indicators of intake of certain foods, independent of genetic factors. Should groups not be known in advance, for example in disease subtyping, unsupervised clustering approaches are generally used to generate them [161]. One of the most advanced frameworks is mixOmics/DIABLO, which is capable of both supervised and unsupervised approaches [162].

These algorithms are focused on applications related to human disease and nutrition and thus profit from the following factors: (i) the subjects all belong to the same species with relatively little population structure; (ii) the genes are the same between subjects, thus transcriptomics can be used; (iii) the genes are well-studied and annotated; (iv) the metabolomic samples (blood, urine) are relatively simple, and the metabolites are well-studied and many can be identified; (v) the focus is on characterizing groups of samples.

In the context of microbes, there are important differences. (i) many microbes reproduce clonally, which may result in strong, confounding population structure effects; (ii) the most relevant genetic differences between the samples are binary data (e.g., presence-absence of orthologs, *k*-mers, unitigs or SNPs); (iii) the genes of many microbes are not well studied, and if multiple species are involved, there could be an annotation bias; (iv) while the metabolic sample might be simpler (i.e., a minimal medium), it could also be more complex and harder to measure (e.g., milk or yoghurt); (v) the focus is on finding causative relationships between genes and metabolites. These differences necessitate the development of novel methods.

1.4.2. Microbial genome-wide association studies (mGWAS)

To determine causative relationships between human genes and metabolites, approaches from human genome-wide association studies (hGWAS) have been adapted for metabolomics [163]. Unfortunately, due to the differences between human and microbial genetics, hGWAS methods cannot simply be applied to microbes. The main differences are that humans are very closely related compared to microbes, and that humans reproduce sexually, with a diploid genome that undergoes recombination. Recombination leads to “linkage equilibrium” between most genes in the human genome, a state where frequencies of different alleles are not correlated with one another. In contrast, the entire genome of haploid, clonally reproducing microbes is in linkage disequilibrium and population structure becomes a major confounding factor that leads to statistical problems (pseudoreplication, see Figure 6) [164, 165].

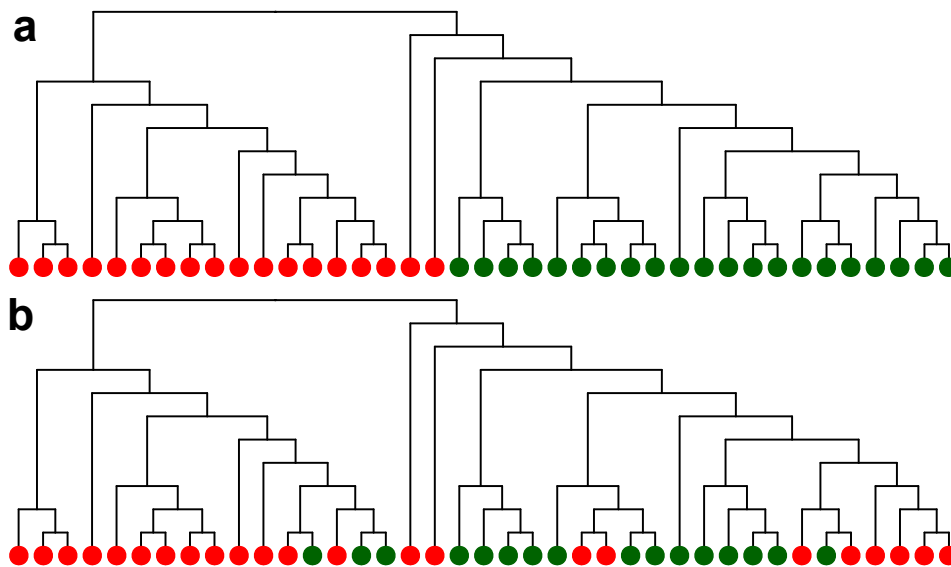


Figure 6: The problem of population structure and pseudoreplication in mGWAS.

The distribution of a trait on a phylogenetic tree determines how likely a correlating gene is to cause the trait. In these phylogenetic trees, each terminal node represents a bacterial strain. If the terminal node is green, it means that the strain has trait T as well as gene G. **(a)** Maximally stratified trait. All changes that have accumulated between the two major clades strongly correlate with T, even if only one is causally responsible. Only a single evolutionary transition is required to explain this correlation. Despite perfect correlation of G and T, this constitutes only very weak evidence for a causal relationship. The many leaves can be considered uninformative pseudoreplicates as they are not statistically independent due to their shared evolutionary history. **(b)** Moderately stratified trait. Six evolutionary transitions are required to explain the correlation, constituting much more convincing evidence.

Various tools have been developed to overcome these issues [166], but so far, none have been adapted to metabolomics data. To the best of my knowledge, only one published study applies GWAS to microbial metabolomics. In 2019, 49 metabolites were associated to the orthogenes of 56 strains, each from a different LAB species, using the Wilcoxon rank sum test [167]. In other words, without any control for population structure and with so few metabolites that it is feasible to analyze the full output manually.

There are three main challenges in applying a GWAS approach to metabolomics data: (i) the GWAS algorithm must be able to handle large numbers of traits, which can be in the tens of thousands; (ii) similar traits should be grouped together in the output; and (iii) automated post-GWAS workflows, such as searching protein sequences on databases and comparing gene loci, become essential.

Currently, the output of many mGWAS tools consists only of a list of orthogene identifiers per trait, without gene identifiers or functional annotations. This makes it difficult to interpret the results and proceed with post-GWAS workflows. To effectively analyze large metabolomics datasets, as much as possible must be automated. As noted by San et al., this automation would also “immensely” benefit regular single-trait mGWAS analyses [166]. Such tools are generally needed in multi-omics studies, where data processing, integration, and visualization usually takes far more time and effort than data generation [168, 169].

1.5. Genome-scale metabolic modeling

Splitting metabolism into separate pathways is a useful abstraction for human comprehension. In fact, pathways are interconnected and defining a boundary is arbitrary. Arguably, a more elegant and holistic solution is to construct a network that depicts the entire known metabolism in a genome-scale metabolic model (GSMM). GSMM are created in a four-step process: (1) construction of a draft model based on genome annotations of the organism in question and existing metabolic models; (2) manual curation of the draft model based on multiple databases and literature; (3) conversion into a mathematical model (the stoichiometric matrix, see Figure 7); (4) iterative testing and adjustment the model using experiments until the desired quality is reached. This process requires extensive lab work and expert knowledge of the organism of interest and metabolism. According to Gudmundsson et al, step 1 takes from hours to days, step 2 from weeks to months, step 3 from hours to days, and step 4 from weeks to months [170]. While recent developments in genome annotation and automatization of draft model generation (e.g. eggNOG [171] and CarveMe [172]) improved step 1, the opportunities to automatize the remaining steps are limited [173].

GSMM can then be used for a variety of simulations. The classical example is flux-balance analysis (FBA, Figure 7). FBA is a computational method used to predict the metabolic fluxes, or the flow of molecules through a metabolic network, of an organism in steady state. The input to FBA consists of (i) a stoichiometric matrix, which encodes the organism’s metabolism, (ii) influx- and efflux “reactions” that encode which nutrients the organism imports from and exports to the environment and (iii) an objective function that represents the goal of the model. This results in a large, under-determined system of linear equations. The fluxes are then optimized to maximize the objective function using linear programming [174].

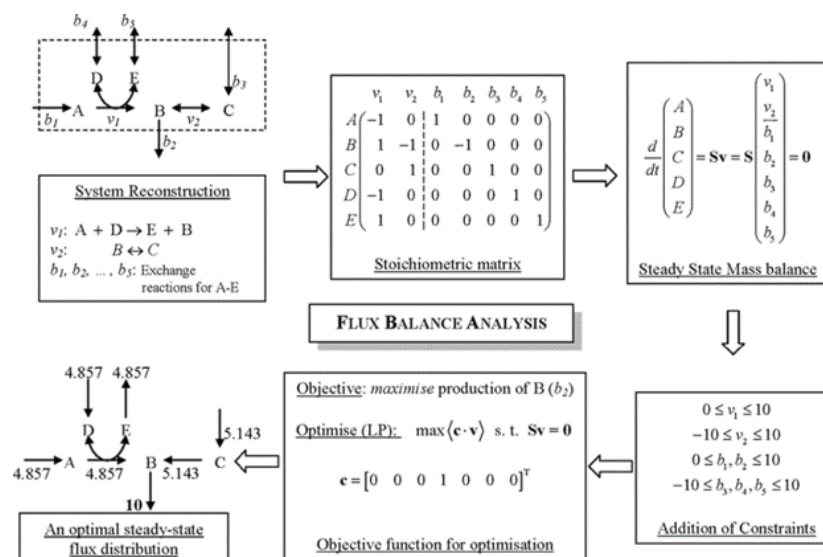


Figure 7: Genome-scale metabolic models (GSMM) and flux balance analysis (FBA).

The stoichiometric matrix S encodes an organism's exchange reactions and metabolism, i.e., its metabolites (A-E) and enzymes (arrows). The goal of FBA is to determine the fluxes through the enzymes (v_1 and v_2) and fluxes in- and out of the organism (b_1 - b_5). These fluxes are modeled as vector v . In FBA, fluxes are constrained by the steady-state assumption $Sv=0$, meaning that the concentration of the metabolites stays constant, as well as additional flux-limiting constraints. This results in a large but conceptually simple system of linear equations that are under-determined, meaning multiple solutions are possible. The fluxes v are optimized to maximize an objective function using linear programming. Classically, the objective function is the maximization of biomass, which could be represented as the sum of the fluxes through the reactions involved in growth. Figure from Raman and Chandra 2009 [174].

It is not possible to automatically generate useful GSMM based on genomic data alone, particularly for non-model organisms. GSMM strongly dependent on the quality of annotations, which can vary greatly between species. Obviously, GSMM only contain known genes as well as pathways but not most secondary metabolites, which may comprise the most interesting intra-species variety. The genes of the core metabolism may be the most conserved and best studied, but even semi-curated GSMM still failed to predict the metabolic needs of bacteria of the microbiota [175]. Tools like CarveMe improve the workflow of integrating experimental data (e.g., growth requirements) in the generation of GSMM using automatic gap-filling [172], but this arguably merely illustrates the importance of extensive experimental work in the creation of GSMM.

The usefulness of GSMM hinges on the quality of the metabolic reconstruction and whether the constraints and objective function are reasonable. For instance, if a GSMM incorrectly models an important pathway or a key nutrient (influx reaction), the fluxes may be badly off. Similarly, the objective function determines the outcome: if it reflects biomass creation, it may simulate many fluxes realistically. However, branches of the pathway that lead to secondary metabolites which do not contribute to growth will end up with zero flux. If the objective function is to maximize the production of a certain compound, it may help bioengineers determine which genes to modify but is unlikely to reflect biological reality [174].

1.5.1. Traditional applications

Genome-scale modeling enables researchers to understand an organism's core metabolism, i.e., the creation of a high-quality metabolic model may itself be the goal. A good metabolic model can predict which substrates are essential for an organism to grow, where the metabolic bottlenecks are, and which genes are essential. This may give insight in the ecology and evolution of an organism. GSMM may help improve gene annotations, as it may indicate the presence of certain pathways that were not predicted by existing knowledge. GSMM can be of great help for metabolic engineering, for instance to find promising drug targets (essential genes) or to determine which genes to manipulate (bottlenecks) to force an organism to efficiently produce certain metabolites [174]. GSMM can be improved or extended with omics-data, including transcriptomics, interactions, epigenomics,

metabolomics and proteomics. Of these, transcriptomics data is most commonly used to extend GSMM, under the assumption that changes in gene expression correlate with changes in fluxes. Thus, RNA sequencing data from multiple conditions can be used to create condition-specific metabolic models. Metabolomics is usually used to confirm predictions and is limited by the number of metabolites that can be identified, particularly in complex media [169, 176, 177].

1.5.2. Community models

GSMM from different organisms can be combined to community metabolic models. This entails modeling exchange reactions between the microbes and makes determination of appropriate objective function even more challenging [178]. The increase in complexity makes experimental validation of most specific predictions impossible. Nevertheless, such models are used in bioengineering to split biosynthesis pathways between different microbes (division of labor), which may improve robustness and yield [179]. Community models also may provide interesting insight into diverse ecosystems through simulations, even though the particular fluxes may be wrong and cannot be verified experimentally. For instance, Freilich et al. simulated community models from pairs of 118 GSMM. They found that most cooperative interactions only favor one of the strains, but that in bigger communities, these interactions compound to benefit all [180]. Simulations also suggest that in resource-restricted environments, the release of metabolites that do not incur a fitness cost on the producer enable cross-feeding, driving interspecies interaction [181]. Community models may show that two organisms can utilize different resources or produce metabolites useful to the other, thus predicting mutualism [182]. Overall, the field of community modeling is very active. A recent review by San León and Nogales describes the main strategies to design microbial communities. All of them rely on high-quality GSMM of each strain [179].

1.5.3. Community modeling of yoghurt

The use of community metabolic models for yoghurt was speculated about already in 2008 by Sieuwerts et al. [183]:

With respect to mixed-culture fermentations, it will be interesting to see whether it is possible to connect genome-scale metabolic models of the individual components of mixed cultures through a limited number of interactions. Such multigenome scale models should be effective tools for the optimization of mixed-culture performance with respect to growth and metabolite production.

The first kinetic yoghurt community model was published in 2015 but uses whey as substrate [184]. In 2021, Özcan et al. were the first to model communities of cheese-related LAB using dynamic FBA but using a chemically defined medium [185]. (In dynamic FBA, growth and uptake rates are used to update the subsequent FBA constraints, enabling temporal simulations.) The development of truly yoghurt-based models has so far been hampered because of the complexity of yoghurt as a substrate, the complexity of correctly modeling the interactions between *S. thermophilus* and *L. bulgaricus*, and unknowns regarding the dynamics of the proteolysis process [186]. Moreover, Somerville et al. mention the difficulty in identification and quantification of metabolites in complex media, which is critical to establish exchange bounds, as well as FBA being prone to predict optimal metabolisms (i.e., respiration or acetate formation instead of lactate fermentation). Finally, the illegality of genetical engineering for food is another limitation in Europe [187]. Hanemaaijer et al. describe the modeling of such communities as a theoretical exercise rather than a useful tool in practice. They suggest that conceptual innovations, namely more coarse-grained metabolic models that collapse entire pathways into single “reactions” and only model pathways of particular interest in detail, may be required to usefully model such communities [188].

2. Aims and Objectives

The aim of the Polyfermenthealth project is to translate the biochemical potential contained in the Agroscope Liebefeld collection into potentially beneficial metabolic profiles *in vivo*. This is to be achieved using yoghurt fermented with additional LAB strains, termed *polyfermented* yoghurts which are tested in mice. The project consists of multiple steps: (1) genetic screening of bacterial strains, determination of promising strategies and strain pre-selection; (2) production of initial polyfermented yoghurts, phenotypic characterization, and selection of final polyfermented yoghurts; (3) testing of the final yoghurts in mice; (4) integrative data analysis of the metabolomic and genomic data.

This thesis addresses the bioinformatics challenges posed by the Polyfermenthealth project.

Aims:

- i. Comparative genomics analysis of the sequenced strains of Agroscope's Dialact database and determination of promising strategies for functional foods.
- ii. Creation of an interactive comparative genomics platform for collaborative strain selection. This evolved into the development of a general purpose, user-friendly and dataset-independent platform for comparative genomics.
- iii. Linking the genotype to the phenotype, i.e., analyzing the metabolomes of polyfermented yoghurts in relation to their respective genomes with consideration of population structure.

3. Results

3.1. Manuscript 1: Comparative genomics of the Dialact database

***In Silico* Comparison Shows that the Pan-Genome of a Dairy-Related Bacterial Culture Collection Covers Most Reactions Annotated to Human Microbiomes**

Thomas Roder, Daniel Wüthrich, Comelia Bär, Zahra Sattari, Ueli von Ah, Francesca Ronchi, Andrew J. Macpherson, Stephanie C. Ganal-Vonarburg, Rémy Bruggmann, Guy Vergères

Status:

Published in *Microorganisms* 2020, 8(7):966 [189]

Statement of contribution:





Conceptualization, D.W. and G.V.; methodology, D.W.; software, T.R. and D.W.; validation, T.R.; formal analysis, T.R. and D.W.; investigation, T.R. and D.W.; resources, G.V. and R.B.; data curation, T.R. and D.W.; writing—original draft preparation, T.R., G.V., C.B.; writing—review and editing, T.R., C.B., Z.S., U.v.A., F.R., A.J.M., S.C.G.-V., G.V., R.B.; visualization, T.R.; supervision, G.V., R.B.; project administration, G.V.; funding acquisition, G.V., R.B. All authors have read and agreed to the published version of the manuscript.

Research objective:

To gain an overview of the diversity and biochemical potential of the sequenced strains of Agroscope's Liebefeld collection which consists of bacteria that originate from the Swiss dairy environment. Moreover, to compare the biochemical potential of the strains to human microbiomes and to find promising strategies to create functional foods.

Article

In Silico Comparison Shows that the Pan-Genome of a Dairy-Related Bacterial Culture Collection Covers Most Reactions Annotated to Human Microbiomes

Thomas Roder ^{1,2,3} , Daniel Wüthrich ^{1,2,3}, Cornelia Bär ³, Zahra Sattari ^{3,4}, Ueli von Ah ³ ,
Francesca Ronchi ⁴, Andrew J. Macpherson ⁴, Stephanie C. Ganal-Vonarburg ⁴ ,
Rémy Bruggmann ^{1,2}  and Guy Vergères ^{3,*}

¹ Interfaculty Bioinformatics Unit, University of Bern, Baltzerstrasse 6, CH-3012 Bern, Switzerland; thomas.roder@bioinformatics.unibe.ch (T.R.); danielwue@hotmail.com (D.W.); remy.bruggmann@bioinformatics.unibe.ch (R.B.)

² Swiss Institute of Bioinformatics, University of Bern, CH-3012 Bern, Switzerland

³ Agroscope, Schwarzenburgstrasse 161, CH-3003 Bern, Switzerland; cornelia.baer@agroscope.admin.ch (C.B.); zahra.sattari@dbmr.unibe.ch (Z.S.); ueli.vonah@agroscope.admin.ch (U.v.A.)

⁴ Department for Biomedical Research (DBMR), University Clinic for Visceral Surgery and Medicine, Bern University Hospital, University of Bern, Murtenstrasse 35, CH-3008 Bern, Switzerland; francesca.ronchi@dbmr.unibe.ch (F.R.); andrew.macpherson@dbmr.unibe.ch (A.J.M.); stephanie.ganal@dbmr.unibe.ch (S.C.G.-V.)

* Correspondence: guy.vergeres@agroscope.admin.ch

Received: 18 May 2020; Accepted: 25 June 2020; Published: 27 June 2020



Abstract: The diversity of the human microbiome is positively associated with human health. However, this diversity is endangered by Westernized dietary patterns that are characterized by a decreased nutrient variety. Diversity might potentially be improved by promoting dietary patterns rich in microbial strains. Various collections of bacterial cultures resulting from a century of dairy research are readily available worldwide, and could be exploited to contribute towards this end. We have conducted a functional in silico analysis of the metagenome of 24 strains, each representing one of the species in a bacterial culture collection composed of 626 sequenced strains, and compared the pathways potentially covered by this metagenome to the intestinal metagenome of four healthy, although overweight, humans. Remarkably, the pan-genome of the 24 strains covers 89% of the human gut microbiome's annotated enzymatic reactions. Furthermore, the dairy microbial collection covers biological pathways, such as methylglyoxal degradation, sulfate reduction, γ -aminobutyric (GABA) acid degradation and salicylate degradation, which are differently covered among the four subjects and are involved in a range of cardiometabolic, intestinal, and neurological disorders. We conclude that microbial culture collections derived from dairy research have the genomic potential to complement and restore functional redundancy in human microbiomes.

Keywords: dairy microbiome; human gut microbiome; diversity; health

1. Introduction

Modern genomic technologies have deeply impacted the understanding of the functionality of microbial communities in humans [1]. In particular, the gut microbiome modulates the balance between health and disease in a large array of biological phenomena, including the supply of nutrients and energy, intestinal motility, immunity, cardiovascular function, cancer, infections, and many more [2]. These properties are fueled by the diversity of the gut microbiome, which encodes about 150 times

more genes [3] than the human genome, including genes for biochemical pathways that are not covered by the human genome, but are relevant for human physiology and maintenance [4].

Biodiversity is a critical aspect of functioning ecosystems, providing them with stability and the ability to respond to external stimuli in a more resilient manner [5]. The same holds true for the human microbiome, as anticipated by microbiologist R. J. Dubos in 1966 [6]. Despite this, the last 50 years witnessed the rise of “Westernized” lifestyles, which are characterized by a decreased diversity in nutrient intake. Concomitantly, many chronic pathologies, such as obesity, type 2 diabetes, and inflammatory bowel disease have become prominent public health issues [7]. A shared feature of these disorders is a reduction in the diversity of the gut microbiome, an observation that has been linked to changes in dietary patterns [8].

People with very different gut microbial composition can be equally healthy, suggesting that the intestinal ecosystem contributes to host homeostasis with a large degree of functional redundancy. However, dysbiosis can appear if functional redundancy is decreased, by the disappearance of taxa that significantly contribute to this homeostasis [9], including so far unknown taxa, which cannot be cultured outside of the intestinal environment. However, functional redundancy could possibly be replenished by direct colonization with the missing taxa, or even taxa other than the originally extinct ones, or by the provision of metabolites promoting the reintroduction of the original species or functional redundant species. As a proof of concept, the extinction of microbial taxa in humanized mice fed with a Western diet over several generations could be reversed by oral administration of the missing taxa [10]. The fact that the oral route taken in this study was successful provides support for human nutritional strategies which aim to deliver bacteria and products of their metabolism to maintain or restore gut homeostasis.

The impact of fermented foods on human health was put forward by Metchnikoff in the early 1900s when he proposed that yoghurt may extend human lifespan [11]. This early work triggered significant scientific efforts during the last century to promote health by using prebiotics and probiotics.

However, the reductionist view of pre-omics science in the 1980s has led food scientists and microbiologists to concentrate their efforts on a narrow range of pre- and probiotic strains rather than on fermented foods with complex microbial composition [12], thus taking the risk to negatively contribute to microbial diversity in human diets. With food being complex, as well as the gut microbiome being primarily characterized by their complexity, it is not a surprise that these narrow strategies were met with moderate success at improving health [13,14].

Meanwhile, the consumption of fermented milk products introduces a diverse range of microbes, which may positively contribute to the restoration of gut microbial diversity [10,15]. Humans have been fermenting milk for almost ten thousand years, primarily to increase its shelf life, but also to be able to tolerate it better and improve its taste. This has likely promoted a close interaction between the ecological niche of lactic acid bacteria (LAB) in dairy environments and the human gut microbiome. Indeed, LAB are acidophilic organisms growing well at pH 3.5–6.5 and constituting about 0.01 to 1.8% of the total bacterial community in the gut. In this respect, recent research indicates that lactic acid bacteria populating the gastrointestinal tract are originating from fermented foods [16]. Studies in mice [17,18] and piglets [19] have revealed that *Lactobacillus* are present in the small intestine. Given the importance of the small intestine for nutrient absorption, targeting functional activities of LAB in this part of the gastrointestinal tract appears to be an interesting nutritional strategy. However, according to the current state of knowledge, only few LAB species seem to stably integrate into the gut microbiome and most require continuous uptake through the diet [20]. Nonetheless, a functional response may be achieved by the high proportion of fermented food and beverage ingested by humans, that is estimated to be between 5% and 40% [15]. Moreover, temporary colonization may be advantageous, by allowing a better control of the health risks associated with the consumption of these organisms.

Apart from potentially being able to complement the microbiome, microbes from fermented foods may also produce or consume compounds with relevance to human health or provide substrates for the resident gut microbiome, leading to metabolites that in turn affect the health of the host. For instance,

some *Lactobacillus* strains such as *L. rhamnosus* can digest lactose, which may be advantageous for lactose intolerant individuals [21]. Some LABs can produce folate (vitamin B9), which is a particularly relevant nutrient for pregnant women, as it can help to prevent birth defects such as *spina bifida* [22]. Furthermore, *Lactobacillus* strains such as *L. reuteri* are known for their ability to metabolize tryptophan and produce indole derivatives, which can bind and activate the human aryl hydrocarbon receptor (AhR) [23]. AhR is a transcription factor that has an important role in cellular proliferation and differentiation and adaptive and innate immune response, as well as detoxification [24].

The consortium of bacteria present in dairy culture collections established over decades may thus contain some of the genomic diversity that was lost in modern food production. In this article, we hypothesize that such collections share biochemical functions with the human gut microbiome and might thus represent a strategically interesting source of bacteria to contribute to the establishment of a healthy gut microbiome. As of March 2019, Agroscope, the Swiss center of excellence for agricultural research, had sequenced and annotated 869 strains of its “Liebefeld collection”, containing more than 10,000 isolates, mostly LAB, originating from the Swiss dairy environment. For this analysis, we confine ourselves to one strain per species that might conceivably be used in food production. These 24 strains will hereafter be referred to as the “Liebefeld selection”. To evaluate the potential of the Liebefeld collection to contribute to the functionality of the human gut microbiome, we used *in silico* methods to compare the coverage of MetaCyc superpathways [25] encoded by the genomes of the Liebefeld selection with the published gut microbiome of four healthy overweight individuals [26]. Given the role of the gut microbiome composition and diversity in obesity [27,28], variations in the genomic content of the gut microbiome of these subjects could provide hints on the replenishing potential of the Liebefeld collection.

2. Materials and Methods

2.1. Genome Sequencing

The strains were either sequenced using PacBio (19) or Illumina (612) technologies. Library preparation, sequencing and genome assembly were performed as described in Wüthrich et al. [29]. In brief, Illumina reads were trimmed with Trimmomatic [30] and assembled with SPAdes [31], whereas PacBio reads were assembled using the HGAP 3 pipeline [32]. Assembly and annotation statistics for the Liebefeld selection strains are reported in Table S1.

2.2. Annotation of the Genome Assemblies

Before sequencing, the strains were taxonomically classified using MALDI-TOF fingerprinting, as described previously [33]. After sequencing, the taxonomic classification was adapted, when appropriate, based on 16S analysis and assembly similarity to genomes available at NCBI [34]. The *de novo* assemblies were structurally annotated using Prokka (version 1.9) [35]. The predicted coding sequences (CDSs) were blasted against Swiss-Prot [36,37], and the hits (e -value $< 10^6$) were clustered based on alignment identity and query coverage using the machine learning algorithm DBSCAN [38]. The gene ontology (GO) terms of the cluster containing the best hit were assigned to the CDS. In addition, the GO terms of all found protein families (Pfam) (e -value $< 10^6$) [39] were assigned to the CDSs. The identified GO terms were then mapped to enzyme commission (EC) numbers. This algorithm, and a more detailed explanation thereof, is available on GitHub [40].

2.3. Selection of LAB

From the 869 entries in our sequencing database, we removed legally restricted strains, strain mixtures and strains which could not confidently be categorized taxonomically. Furthermore, assemblies with more than 500 scaffolds and less than 90% BUSCO v3 [41] single-copy-completeness were discarded. Some strains were sequenced multiple times. In this case, duplicates were removed in favor of the better assembly. Each of the 869 entries were shown to describe a unique genome and

this report therefore refers to each of them as a “strain” and not an “isolate”, in agreement with van Rossum et al. [42]

From the 31 species, 7 were excluded because they are not relevant to dairy product development: *Brevibacterium linens*, *Corynebacterium variabilis*, *Desemzia incerta* and *Glutamicibacter arilaitensis* occur only in the cheese rind. *Anaerospaera aminiphila* strains were first isolated from swine manure and are most likely contaminants of raw milk. Furthermore, the genus *Listeria* is associated with disease, and thus, strains that were identified as *Listeria monocytogenes* or *Listeria innocua* were excluded. The 626 remaining strains will hereafter be termed “Liebefeld collection”.

From each of the 24 species remaining in the Liebefeld collection, the strain with the highest number of unique EC numbers (uEC) was selected. These 24 strains will hereafter be referred to by their species name and collectively referred to as “Liebefeld selection”.

2.4. Selection of Human Microbiomes

Sequences of four human microbiomes (MH0001-4) were downloaded from GutCyc [26]. The microbiomes belong to middle-aged overweight Danes, two male and two female, living in the northern part of the Copenhagen region. The microbiomes were measured as inflammatory bowel disease (IBD) control subjects for the MetaHIT study. At the original recruitment, the individuals had normal fasting plasma glucose and normal 2-hour plasma glucose, following an oral glucose tolerance test. At the time of fecal sampling, all were examined in the fasting state and had non-diabetic fasting plasma glucose levels below 7.0 mmol/L [3]. The available metadata are summarized in Tables S2 and S3. Information about their diet is not available.

Twenty-four genomes were randomly chosen from the 1520 human gut bacteria published by Zou et al. in 2019 [43], and annotated using Prokka (version 1.11) [35]. They were then annotated with EC-numbers, as described in Section 2.2. Their assembly and annotation statistics are reported in Table S4.

2.5. Calculation of Core- and Pan-Genomes

The pan-genome comprises the set of EC-numbers present in any of the respective strains or microbiomes. Similarly, the core-genome comprises the set of EC-numbers that occur in all respective strains or microbiomes.

2.6. Calculation of Superpathway Coverage

As a heuristic to assess biochemical potential, we chose MetaCyc over the Kyoto Encyclopedia of Genes and Genomes (KEGG), because MetaCyc has a stronger focus on biochemistry and reactions compared to KEGG, which is more medicine- and compound-oriented. MetaCyc contained more reactions (14,039 EC-annotated reactions) compared to KEGG (11,381 reactions), as of early 2020. MetaCyc superpathways consist of connected sub-pathways providing—compared to KEGG—a more comprehensive biochemical context with regard to scale and purpose [44]. MetaCyc superpathways are annotated with an “expected taxonomic range”, which can be very broad (e.g. “pyrimidine ribonucleosides degradation” in *Archaea*, *Bacteria*, and *Metazoa*), or, on the other hand, consist of rarely studied genes, which are only known to occur in a single species (e.g. “tetrathionate reduction” in *Salmonella typhimurium*). Thus, some superpathways are expected to be covered by most strains, while others may not be covered by any strain or microbiome.

Superpathway coverage was calculated by mapping the annotated EC-numbers (Section 2.2) to the bacterial superpathways of the MetaCyc database (version 22.6) [25], using a python script [45].

Because superpathways consist of multiple pathways, complete superpathway coverage is not always required for the synthesis of biologically important molecules. Since the main functions of the human gut microbiome include energy production from non-digestible carbohydrates, the deconjugation and dehydroxylation of bile acids, the biosynthesis of vitamins and isoprenoids,

cholesterol reduction, and the metabolism of amino acids and xenobiotics [4]; much of the functionality of the gut microbiome should be covered by the superpathways.

Furthermore, some superpathway reactions are not annotated with EC-numbers, and can therefore not be completely covered in this analysis, even though the relevant gene may be present.

The superpathway coverage of the “average Liebefeld collection strain” is defined as the average coverage value obtained using every strain from the Liebefeld collection. The superpathway coverage of the “Liebefeld random selection” is defined as the average of 1000 bootstrapped random selections of 24 strains from the Liebefeld collection, without regard to taxonomic classification.

3. Results and Discussion

3.1. Liebefeld Collection Overview

Figure 1 presents, for each strain, the number of genes, as well as the EC annotation rate. It provides information on the scale (number of strains), taxonomic composition (number of species), and diversity (differing numbers of genes and annotation rates between species) of the Liebefeld collection.

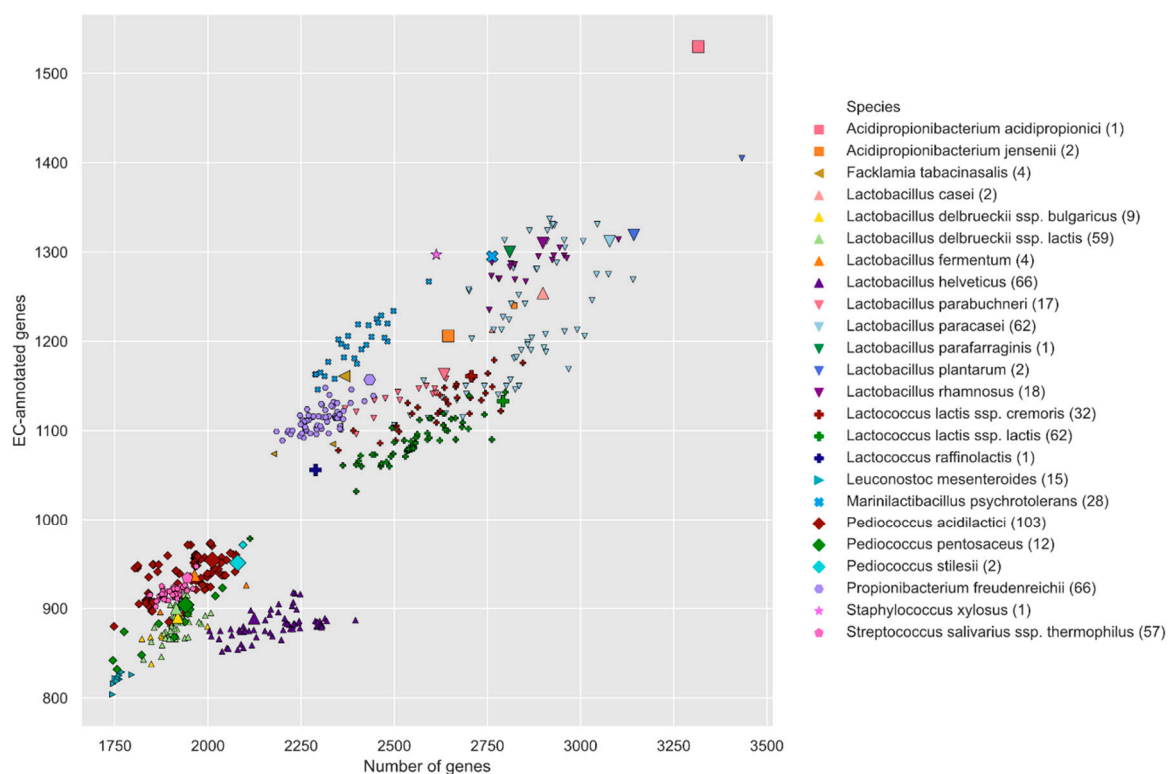


Figure 1. Relationship between the number of identified genes and the number of enzyme commission (EC)-annotated genes per Liebefeld collection strain. Strains from the Liebefeld selection are highlighted with a larger symbol. Each species is represented by a unique symbol and color. A plot that includes the 24 human gut bacteria randomly selected from Zou et al. [43] is available in Figure S1.

The Liebefeld collection strains have fewer genes than the 24 randomly selected gut microbes (Figure S1). The gut microbes have, on average, 3497 genes, which is slightly more than the Liebefeld collection strain, with the highest number of genes. Further, the distributions of the annotation rate are significantly different (Mann–Whitney U test, p -value = 1.68×10^{-4}), the Liebefeld selection demonstrating higher annotation rates (Figure S2).

Unsurprisingly, the number of genes correlates strongly with genome size ($R^2 = 90\%$) and, consequently, with the species. While genomes from the same species have similar numbers of genes, there is a large range within the genus *Lactobacillus*. The number of genes correlates with the

number of EC-annotated genes, but the EC annotation rate also depends on the species. For example, *Lactobacillus helveticus* strains, despite having a similar number of genes as *Propionibacterium freudenreichii* strains, have a significantly lower annotation rate.

3.2. Comparison of Superpathway Coverage

As a display of the biochemical potential and functional diversity of the Liebefeld collection, Figure 2 shows the superpathway coverage of the individual strains of the Liebefeld selection in comparison with the four human microbiomes. The average superpathway coverage of the Liebefeld selection strains ranged from 36% (*Lactobacillus helveticus*) to 53% (*Acidipropionibacterium acidipropionici*). The average Liebefeld collection strain covered 42% of the superpathways. Remarkably, the pan-genome of the Liebefeld selection covered 65% of the superpathways, and the human microbiomes between 60% (MH0004) and 67% (MH0003), indicating that 24 strains suffice to reach a coverage similar to the human microbiome.

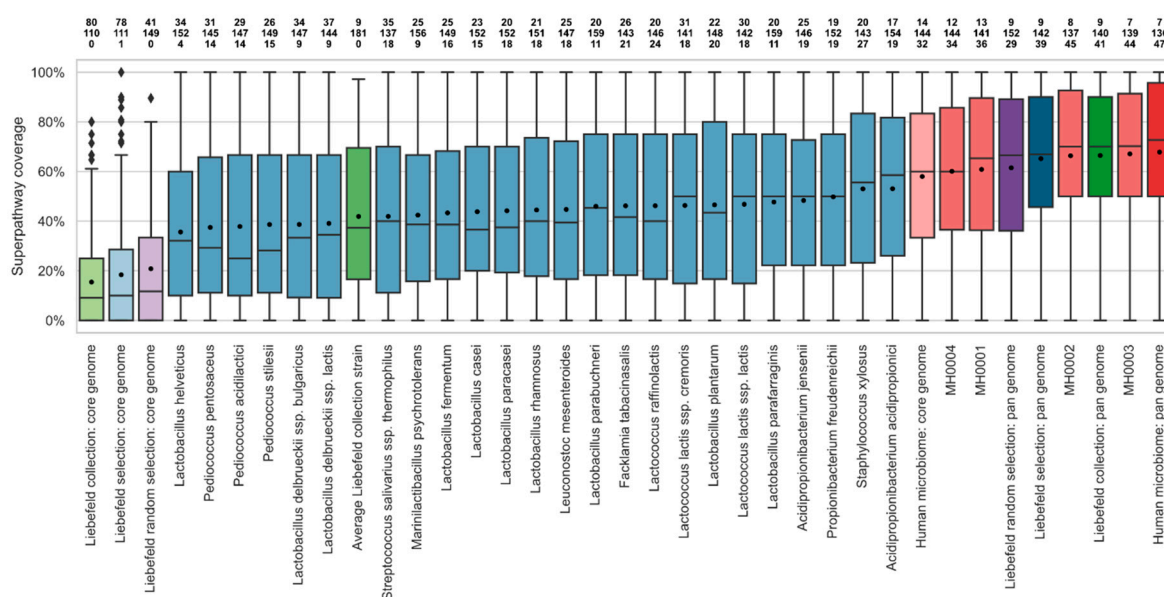


Figure 2. Boxplot of the coverage of the 190 MetaCyc superpathways by each of the 24 strains of the Liebefeld selection (blue, referred to by their species name) and the four human microbiomes (red, MH0001-4), the Liebefeld collection (green), and the Liebefeld random selection (violet). Core-genomes are colored in a lighter and pan-genomes in a darker shade of the corresponding color. The strains or sets of strains are sorted in ascending order according to their mean superpathway coverage, indicated by a black dot. Above each boxplot, three numbers indicate how many superpathways are not covered (top row), partially covered (middle row) and completely covered (bottom row). An analogous plot comparing the Liebefeld selection strains to the 24 human gut bacteria randomly selected from Zou et al. [43] is available in Figure S3.

Furthermore, the average superpathway coverage of the 24 Liebefeld selection strains (Liebefeld selection: pan genome) was significantly higher than that of 24 randomly selected strains (Liebefeld random selection: pan genome) (p -value under normal distribution = 1.58×10^{-4}).

In all strains and microbiomes, the majority of superpathways remained only partially covered. No single strain covered more than 27 superpathways completely. Together, the Liebefeld selection covered 39 superpathways completely, placing it within the range of human microbiomes (34 to 44).

The 24 randomly selected genomes of human gut bacteria have a slightly higher mean superpathway coverage than the Liebefeld selection strains (Figure S3), but the difference in their distribution is not significant (Mann–Whitney U test, p -value = 0.32).

The detailed coverage of each superpathway, as well as dendrograms, is available in Figure 3.

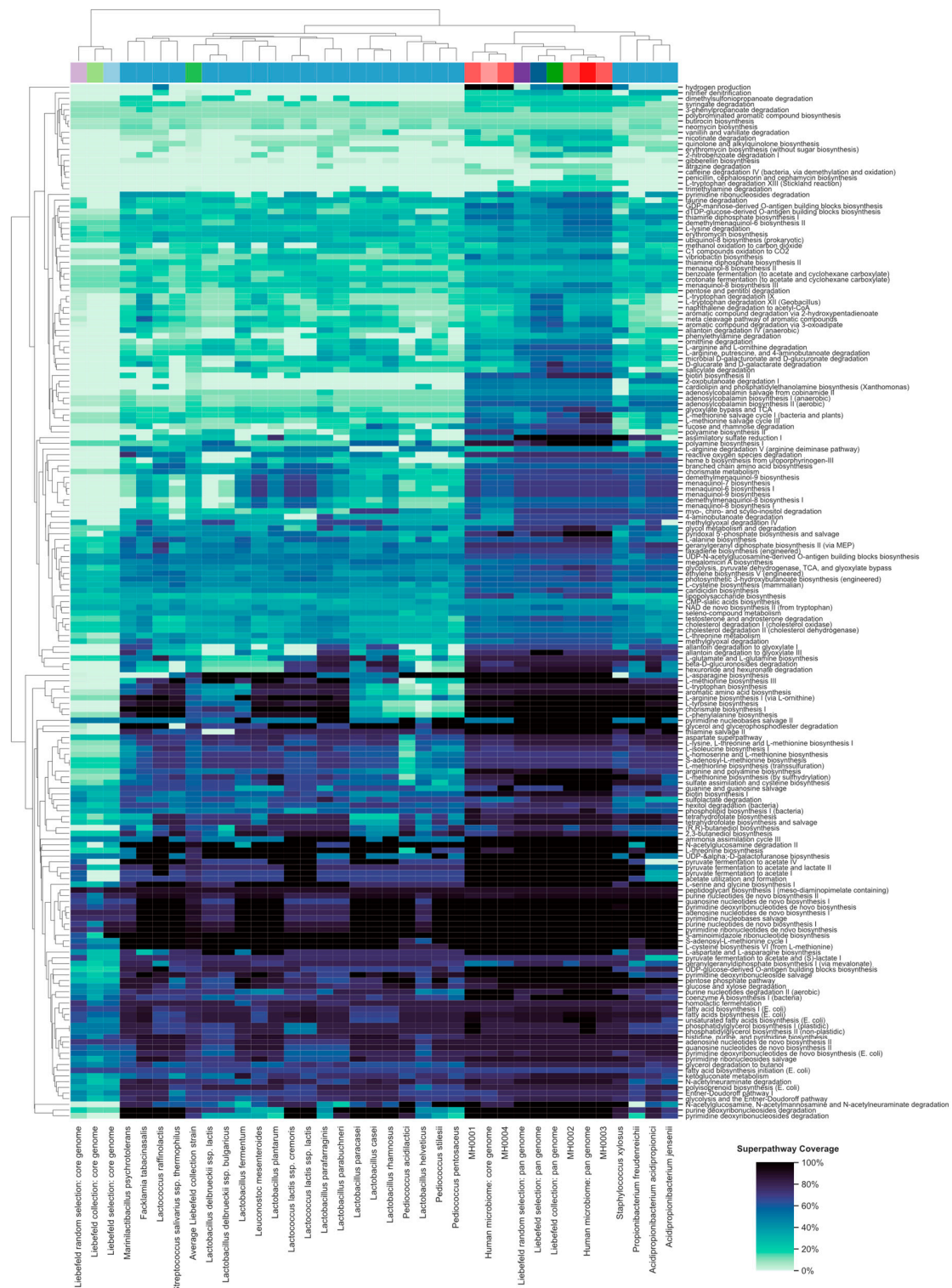


Figure 3. Overview of the biochemical potential of the 24 strains of Liebefeld selection (blue, referred to by their species name), the four human microbiomes (red, MH0001-4), the Liebefeld collection (green), and the Liebefeld random selection (violet). Core-genomes are colored in a lighter- and pan-genomes in a darker shade of the corresponding color. The Y-axis denotes the 190 superpathways of MetaCyc. The dendrogram of both axes resulted from hierarchical clustering. The colors of the heatmap denote superpathway coverage and range from white (0%) to black (100%). An analogous plot comparing the Liebefeld selection strains to the 24 human gut bacteria randomly selected from Zou et al. [43] is available in Figure S4.

3.3. Comparison of Unique EC Numbers (uECs)

Figure 4 gives an overview of the shared genomic content between the strains of the Liebefeld selection and the four human microbiomes. The number of unique EC-numbers (uECs) that were annotated to the strains of the Liebefeld selection range from 727 (*Lactobacillus helveticus*) to 1161 (*Acidipropionibacterium acidipropionici*). Since it would be beyond the scope of this study to go into detail about the enzymes and reactions covered, this analysis is limited to a numerical comparison. On average, a single strain of the Liebefeld selection covers 53% of the uECs of the human microbiomes (MH0001: 59%; MH0002: 49%; MH0003: 49%; MH0004: 56%). Taken together, the 24 strains of the Liebefeld selection contain 1676 uECs, a number comparable to that of the studied human microbiomes, their uECs ranging from 1367 (MH0001) to 1811 (MH0003). On average, the Liebefeld selection pan-genome covers 89% of the uECs of the four studied human microbiomes (MH0001: 92%; MH0002: 87%; MH0003: 87%; MH0004: 90%). Taken together, the 626 strains of the Liebefeld collection contain 1728 uECs, i.e., more than two of the four studied microbiomes, and they cover 91% of the uECs of the four human microbiomes (MH0001: 94%; MH0002: 89%; MH0003: 89%; MH0004: 92%). Conversely, if the 24 strains of the Liebefeld selection were added to the four human microbiomes, the latter would gain, on average, 17% uECs (MH0001: 31%; MH0002: 8%; MH0003: 6%; MH0004: 24%). However, because of the variability of the four human microbiomes, only few uECs in the Liebefeld collection (5%) and the Liebefeld selection (4%) cannot be found in any of the four studied human microbiomes. Figure 5 graphically illustrates the large overlap of uECs between the human microbiome and the Liebefeld selection, as well as the Liebefeld collection.

In addition, the 24 Liebefeld selection strains (Liebefeld selection: pan genome) have significantly more uECs than 24 randomly selected strains (Liebefeld random selection: pan genome) (p -value under normal distribution = 2.53×10^{-6}). Even though we assume that organisms have a large number of similar enzymes in common, the biochemical potential of the Liebefeld collection is remarkable. The number of uECs present in its pan-genome exceeds that of two out of the four studied human microbiomes and resulted in a MetaCyc superpathway coverage greater than three of them. Even after restricting the number of strains to 24, each from a different species (Liebefeld selection), the pan-genome showed a higher superpathway coverage and more uECs than any of the 1'000 randomly chosen combinations of 24 strains (Liebefeld random selection). Furthermore, their superpathway coverage, as well as the number of uECs, was well within the range of the human microbiomes analyzed in this study. Thus, the results of our in silico study on the biochemical potential of the strains that originate from the dairy environment are promising. However, these results could be improved, as only one strain per species was selected to build the Liebefeld selection.

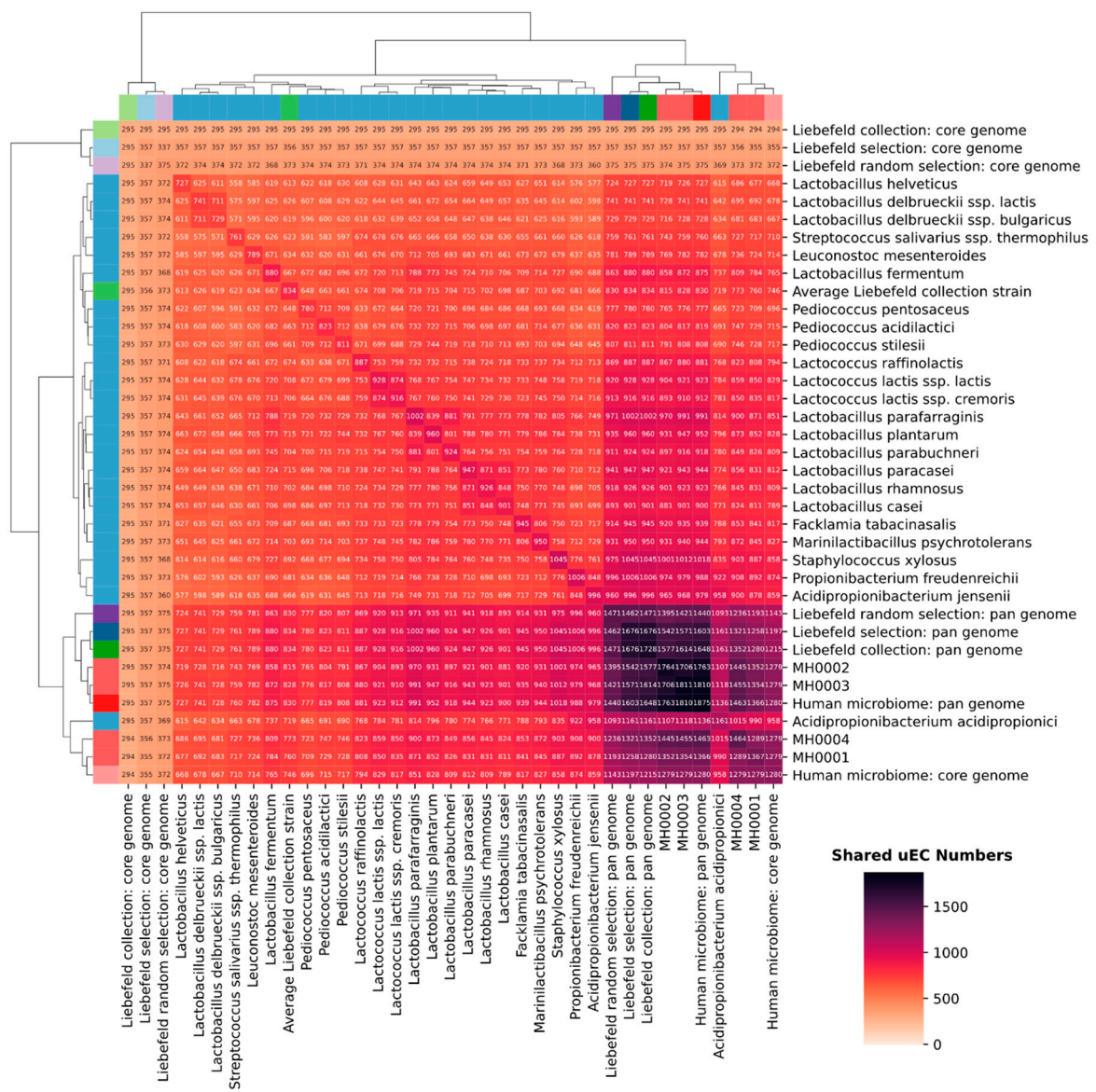


Figure 4. Clustered heatmap of the number of shared unique EC numbers (uECs) of 24 strains of Liebfeld selection (blue, referred to by their species name), the four human microbiomes (red, MH0001-4), the Liebfeld collection (green), and the Liebfeld random selection (violet). Core-genomes are colored in a lighter- and pan-genomes in a darker shade of the corresponding color. The total number of uECs annotated to a species/metagenome can be read from the anti-diagonal line, where the same species/metagenome intersect. The dendrogram of both axes resulted from hierarchical clustering. An analogous plot comparing the Liebfeld selection strains to the 24 human gut bacteria randomly selected from Zou et al. [43] is available in Figure S5.

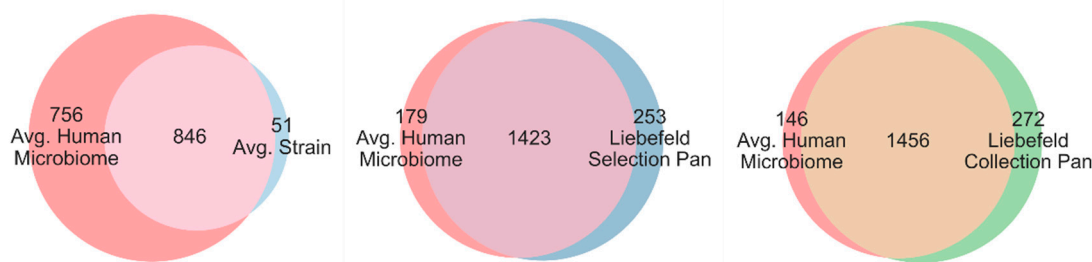


Figure 5. Venn diagrams of the shared unique EC numbers (uECs) between the average human microbiome (red) and the average Liebefeld selection strain (light blue), the Liebefeld selection pan-genome (dark blue) and the Liebefeld collection pan-genome (green).

3.4. Functional Properties of the Gut Microbiome, Which Might be Enriched by the Liebefeld Collection

The four human subjects selected in this report were healthy, although overweight. Individual differences in the coverage of their MetaCyc superpathways may be associated with their metabolic or health status, possibly indicating the onset of dysbiosis. To illustrate our approach, we have searched for such superpathways in the microbiomes of the four subjects and addressed whether the underrepresented microbial functions might theoretically be supported by the Liebefeld selection or strains thereof. The fifteen superpathways with the highest variance amongst the human microbiomes are presented in Figure 6, and some of them are discussed in their biological context below.

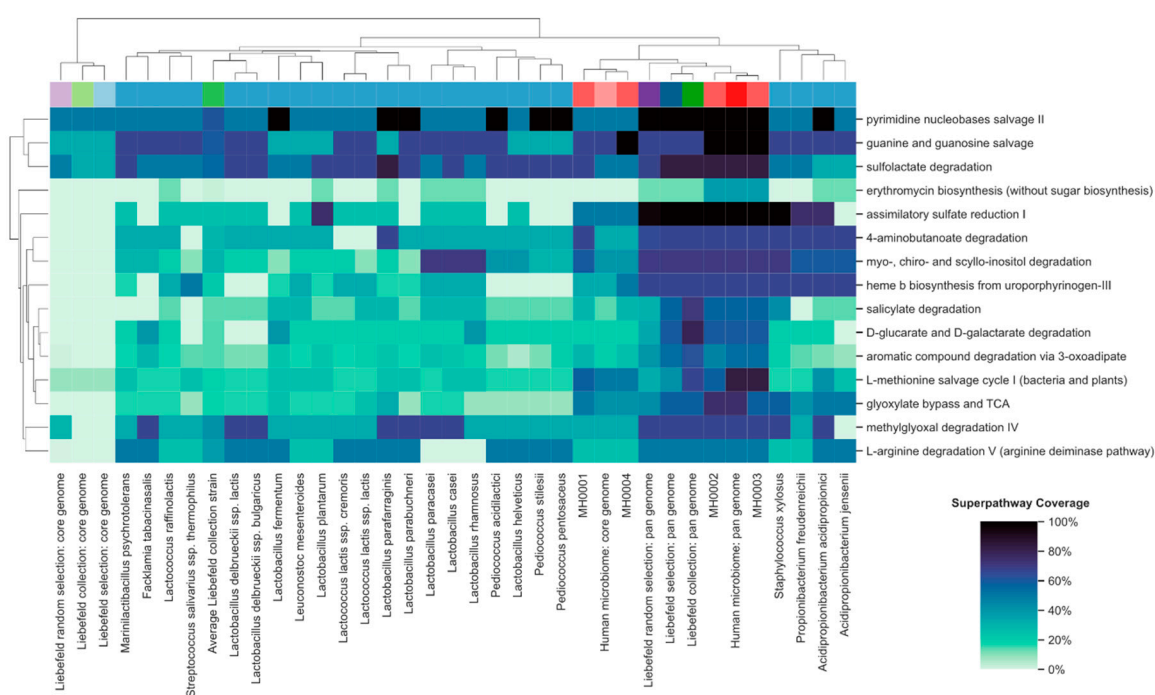


Figure 6. Overview of the biochemical potential of the 24 strains of Liebefeld selection (blue, referred to by their species name) and the four human microbiomes (red, MH0001-4), the Liebefeld collection (green), and the Liebefeld random selection (violet), for the fifteen superpathways with the highest variance amongst the human microbiomes. Core-genomes are colored in a lighter and pan-genomes in a darker shade of the corresponding color. The Y-axis denotes these 15 superpathways. The dendrogram of both axes resulted from hierarchical clustering. The colors of the heatmap denote superpathway coverage and range from white (0%) to black (100%). An analogous plot comparing the Liebefeld selection strains to the 24 human gut bacteria randomly selected from Zou et al. [43] is available in Figure S6.

3.4.1. Methylglyoxal Degradation IV Superpathway

The superpathway “methylglyoxal degradation IV” has a low coverage in subjects MH0001 and MH0004 (Figure 6). Methylglyoxal is a product of glucose and glycine metabolism that increases the activity of the gut microbial trimethylamine (TMA)-lyase [46]. TMA lyase catalyzes the transformation of dietary choline and carnitine to TMA, which in turn is metabolized by the liver to trimethylamine-N-oxide (TMAO), a potential risk factor for cardiovascular diseases [47]. The pan-genome of the Liebefeld selection covers the superpathway “methylglyoxal degradation IV” similarly to subjects MH0002 and MH0003. Among the 24 species of the selection, nine strains demonstrate a high coverage of the superpathway (*L. parafarraginis*, *L. parabuchneri*, *L. delbrueckii ssp. lactis*, *L. delbrueckii ssp. bulgaricus*, *Facklamia tabacinasalis*, *L. paracasei*, *L. casei*, *Staphylococcus xylosum*, and *Acidopropionibacterium acidopropionici*). These strains could enhance the methylglyoxal degradation capability in the subjects MH0001 and MH0004, and potentially lower their TMAO levels. That the fermentation of food products with LAB can redirect the transformation of precursors of TMAO was recently demonstrated by Burton et al. [48], who showed that the fermentation of milk to yoghurt decreases TMAO in urine and plasma.

3.4.2. Assimilatory Sulfate Reduction I Superpathway

The superpathway “Assimilatory sulfate reduction I” has a lower coverage in subjects MH0001 and MH0004 (Figure 6). Intestinal microorganisms use sulfate to synthesize cysteine via the assimilatory sulfate reduction pathway [49,50]. Sulfate can, however, also be metabolized via the dissimilatory sulfate reduction pathway to produce hydrogen sulfide (H₂S). H₂S is a toxic molecule associated, among others, with IBD. On the other hand, recent research has revealed that, similar to nitric oxide (NO), H₂S is an important signaling molecule, with therapeutic potential in a range of diseases, in particular oxidative stress-induced neurodegenerative diseases [51]. The pan-genome of the Liebefeld selection covers the superpathway “Assimilatory sulfate reduction I”, similarly to subjects MH0002 and MH0003, and could thus potentially shift the activity of the gut microbiome towards assimilatory sulfate reduction. As H₂S inhibits the growth of LAB strains, hampering their development as probiotics [52], LAB strains diverting sulfate metabolism towards the assimilatory pathway could be interesting components of probiotic products. Among the 24 species of the selection, one strain demonstrates a high coverage of this superpathway (*Lactobacillus plantarum*).

3.4.3. 4-Aminobutanoate Degradation (GABA) Degradation Superpathway

The superpathway “4-aminobutanoate degradation” has a lower coverage in subject MH0004 compared to the other three subjects (Figure 6). The inhibitory neurotransmitter 4-aminobutanoate (GABA), which is also synthesized by microbes in the intestine, is known for balancing the stimulation of synapses by glutamate in the brain. Although, to our knowledge, there are no studies which show direct evidence for the effect of gut-derived GABA on the human CNS, an in vivo study in mice with GABA-producing *Lactobacillus rhamnosus* (JB-1) revealed changes in the mRNA of GABA receptors B1b and A2, as well as reduced anxiety- and depression-related behavior [53]. Specific GABA/glutamate antiporters mainly achieve homeostasis between glutamate and GABA, although the degradation of GABA also plays a role. The pan-genome of the Liebefeld selection covers the superpathway “GABA degradation”, similarly to the microbiome of subjects MH0001, MH0002, and MH0003. Among the 24 species of the selection, five strains demonstrate a high coverage of this superpathway (*Staphylococcus xylosum*, *L. parafarraginis*, *Propionibacterium freudenreichii*, *Acidopropionibacterium acidopropionici*, and *Acidopropionibacterium jensenii*).

3.4.4. Salicylate Degradation Superpathway

The superpathway “salicylate degradation” has a lower coverage in subject MH0001 and MH0004 compared to the other two subjects (Figure 6). Salicylates are known for their analgesic,

antipyretic, antithrombotic and anti-inflammatory effects. The main mechanism by which these effects are achieved is the inhibition of cyclooxygenases, which are responsible for the biosynthesis of prostaglandins. However, as recent research shows, salicylates also induce the activation of the adenosine monophosphate-activated protein kinase (AMPK) [54], which plays an important role in cellular energy homeostasis and immunity, promoting the generation of Tregs [55]. However, salicylates can also activate the AhR [56], which is involved in cellular proliferation and differentiation and has a major role in adaptive and innate immune response [57]. Furthermore, salicylates influence the intestinal microbiome itself by decreasing the expression of adherence factors and biofilm formation [58]. The degradation of salicylates therefore contributes to the salicylate homeostasis and thus influences different biological functions regulated by AMPK and AhR activation, as well as the composition and structure of the intestinal microbiome. The pan-genome of the Liebefeld selection covers the superpathway “salicylate degradation” similarly to subjects MH0002 and MH0003. Among the 24 species of the selection, only one strain, *Staphylococcus xylosus*, shows a higher coverage as MH0001 and MH0004, but seven further strains demonstrate a comparable high coverage of the superpathway (*L. fermentum*, *Leuconostoc mesenteroides*, *Lactococcus lactis* ssp. *lactis*, *L. parafarraginis*, *L. paracasei*, *L. casei*, *L. rhamnosus*).

3.5. Limitations

The usefulness of the bioinformatic analysis in this study depends on the quality of the annotations. Because ubiquitous genes are better studied and annotated than less common genes, the extent of overlap between the Liebefeld strains and the microbiomes is likely overestimated. In addition, the human microbiome has received little attention from researchers until recent years. Since it consists of mostly non-cultivable species, the available information on the composition and dynamics of the species present is still limited. As the genomes of well-studied species are likely better annotated (see Figure S2 for a comparison of annotation rate between the Liebefeld selection and the 24 randomly selected gut bacteria), it is possible that the superpathway coverage of the four human microbiomes is underestimated.

To achieve a sustainable effect on the human microbiome, a strain must either be able to integrate stably into the microbiome or be supplied continuously. For stable integration, factors that favor inclusion into the gut microbiome must be considered, such as resistance to low pH, ability to tolerate bile salts, pancreatin, pepsin, lysozyme, and H₂S [52], the ability to compete with other gut bacteria, and the ability to adhere to intestinal epithelial cells or mucus [59,60]. It may be possible to predict some of these factors by studying the genomes of the strains. For example, genes that belong to the mucus-binding (MUB) protein family, which has first been discovered in *Lactobacillus reuteri* and *acidophilus*, could be indicators of adaptation to the gut environment. Pili have also been shown to mediate the gut adhesion of *Bifidobacteria*, but these structures have a far wider applicability and are probably less sensitive predictors [60]. In this context, although most species in the Liebefeld selection are not classified as common residents, most have been detected in the human gastrointestinal microbiome (Table S1) [61–63].

However, the presence of such genes does not necessarily translate into their expression in the relevant environment, and as most relevant genes are unknown in the first place, biological experiments are indispensable. This study focuses on the presence or absence of annotations without regard to gene copy numbers and similar nuances. As gene duplication often increases gene dosage [64] and contributes to the diversification of microbes in the gut [65], it could be interesting to add this parameter to the panel of criteria to identify interesting bacteria. This analysis goes, however, beyond the scope of this report.

Although Section 3.4 provides an illustration of the functional potential of the Liebefeld collection, these examples are limited to healthy overweight individuals. This approach could be extended to a comparison of the microbiomes of healthy and dysbiotic patients. Such an analysis would, however, also require a large dataset, in order to move from the illustrative cases presented in Section 3.4 to a

clinically meaningful approach. In addition, the complementation with RNA-seq data would also be desirable to measure whether, and in what proportion, the genes of interest are actually expressed.

4. Conclusions

Our in silico study shows that the Liebefeld collection, which consists of strains of LAB that were collected in the Swiss dairy environment during a period of almost a century, has significant biochemical potential. In particular, the Liebefeld collection offers a strategic opportunity to design food products, which could possibly mitigate the negative effects of Westernized diets on health, by restoring the functional redundancy of the gut microbiome, which is often lost in the context of chronic diseases like obesity, diabetes, and inflammatory bowel disease. Currently, the potential of these strains is explored for various applications in food production. In this context, bioinformatic investigations are a powerful tool for identifying the most promising candidates at the genetic level. In particular, the selection of combinations of several strains that complement each other at the biochemical level is a novel and highly interesting concept for the food industry. The approach shown in this study thus has the potential to revolutionize the production in the field of fermented foods and lead to a completely new product palette that meets the growing demand for health supporting foods.

Supplementary Materials: Supplementary materials can be found at <http://www.mdpi.com/2076-2607/8/7/966/s1>. Table S1: Assembly and annotation statistics for the Liebefeld selection strains, Table S2: Information about the individuals whose metagenomes were analyzed, Table S3: Information about sequencing data of the metagenomes, Table S4: Assembly and annotation statistics for the genomes of 24 human gut bacteria, Figure S1: Relationship between the number of identified genes and the number of EC-annotated genes for the Liebefeld collection strains and 24 human gut bacteria, Figure S2: Histogram of the annotation rates for the Liebefeld selection strains and the 24 human gut bacteria, Figure S3: Boxplot of the coverage of the 190 MetaCyc superpathways by the Liebefeld selection strains and 24 human gut bacteria, Figure S4: Overview of the biochemical potential of the Liebefeld selection strains and 24 human gut bacteria, Figure S5: Clustered heatmap of the number of shared unique EC numbers (uECs) of the Liebefeld selection strains and 24 human gut bacteria, Figure S6: Overview of the biochemical potential of Liebefeld selection strains and 24 human gut bacteria for the fifteen superpathways with the highest variance amongst the human microbiomes.

Author Contributions: Conceptualization, D.W. and G.V.; methodology, D.W.; software, T.R. and D.W.; validation, T.R.; formal analysis, T.R. and D.W.; investigation, T.R. and D.W.; resources, G.V. and R.B.; data curation, T.R. and D.W.; writing—original draft preparation, T.R., G.V., C.B.; writing—review and editing, T.R., C.B., Z.S., U.v.A., F.R., A.J.M., S.C.G.-V., G.V., R.B.; visualization, T.R.; supervision, G.V., R.B.; project administration, G.V.; funding acquisition, G.V., R.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Gebert RUF Stiftung within the program “Microbials”, grant number GRS-070/17.

Acknowledgments: We thank the Liebefeld Culture Collection Team of Agroscope, Bern, Switzerland (Noam Shani, Emmanuelle Arias, Monika Haueter) for organizing, maintaining and providing access to the strain collection; Hélène Berthoud for DNA extraction; the Next Generation Sequencing (NGS) Platform of the University of Bern for sequencing; and Simone Oberhansli for maintaining the in silico database.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Morgan, X.C.; Segata, N.; Huttenhower, C. Biodiversity and functional genomics in the human microbiome. *Trends Genet.* **2013**, *29*, 51–58. [[CrossRef](#)]
2. Cani, P.D. Human gut microbiome: Hopes, threats and promises. *Gut* **2018**, *67*, 1716–1725. [[CrossRef](#)] [[PubMed](#)]
3. Qin, J.; Li, R.; Raes, J.; Arumugam, M.; Burgdorf, K.S.; Manichanh, C.; Nielsen, T.; Pons, N.; Levenez, F.; Yamada, T.; et al. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **2010**, *464*, 59–65. [[CrossRef](#)] [[PubMed](#)]
4. Yadav, M.; Verma, M.K.; Chauhan, N.S. A review of metabolic potential of human gut microbiome in human nutrition. *Arch. Microbiol.* **2018**, *200*, 203–217. [[CrossRef](#)]

5. Balvanera, P.; Pfisterer, A.B.; Buchmann, N.; He, J.S.; Nakashizuka, T.; Raffaelli, D.; Schmid, B. Quantifying the evidence for biodiversity effects on ecosystem functioning and services. *Ecol. Lett.* **2006**, *9*, 1146–1156. [[CrossRef](#)]
6. Dubos, R.; Schaedler, R.W. The digestive tract as an ecological system. *Rev. Immunol. Ther. Antimicrob.* **1966**, *30*, 247–252. [[PubMed](#)]
7. Bach, J.-F. The Effect of Infections on Susceptibility to Autoimmune and Allergic Diseases. *N. Engl. J. Med.* **2002**, *347*, 911–920. [[CrossRef](#)]
8. Heiman, M.L.; Greenway, F.L. A healthy gastrointestinal microbiome is dependent on dietary diversity. *Mol. Metab.* **2016**, *5*, 317–320. [[CrossRef](#)]
9. Lozupone, C.A.; Stombaugh, J.I.; Gordon, J.I.; Jansson, J.K.; Knight, R. Diversity, stability and resilience of the human gut microbiota. *Nature* **2012**, *489*, 220–230. [[CrossRef](#)]
10. Sonnenburg, E.D.; Smits, S.A.; Tikhonov, M.; Higginbottom, S.K.; Wingreen, N.S.; Sonnenburg, J.L. Diet-induced extinctions in the gut microbiota compound over generations. *Nature* **2016**, *529*, 212–215. [[CrossRef](#)]
11. Metchnikoff, E. Quelques remarques sur le lait aigri. In *Revue Générale de Chimie Pure et Appliquée*; Bureau de la Revue: Paris, France, 1907; Volume 10, pp. 77–85.
12. Miquel, S.; Beaumont, M.; Martin, R.; Langella, P.; Braesco, V.; Thomas, M. A proposed framework for an appropriate evaluation scheme for microorganisms as novel foods with a health claim in Europe. *Microb. Cell Fact.* **2015**, *14*, 48. [[CrossRef](#)] [[PubMed](#)]
13. Katan, M.B. Why the European Food Safety Authority was right to reject health claims for probiotics. *Benef. Microbes* **2012**, *3*, 85–89. [[CrossRef](#)] [[PubMed](#)]
14. Kumar, H.; Salminen, S.; Verhagen, H.; Rowland, I.; Heimbach, J.; Banares, S.; Young, T.; Nomoto, K.; Lalonde, M. Novel probiotics and prebiotics: Road to the market. *Curr. Opin. Biotechnol.* **2015**, *32*, 99–103. [[CrossRef](#)] [[PubMed](#)]
15. Borresen, E.C.; Henderson, A.J.; Kumar, A.; Weir, T.L.; Ryan, E.P. Fermented foods: Patented approaches and formulations for nutritional supplementation and health promotion. *Recent Pat. Food Nutr. Agric.* **2012**, *4*, 134–140. [[CrossRef](#)] [[PubMed](#)]
16. Pasoli, E.; De Filippis, F.; Mauriello, I.E.; Cumbo, F.; Walsh, A.M.; Leech, J.; Cotter, P.D.; Segata, N.; Ercolini, D. Large-scale genome-wide analysis links lactic acid bacteria from food with the gut microbiome. *Nat. Commun.* **2020**, *11*, 2610. [[CrossRef](#)]
17. Li, C.; Bei, T.; Niu, Z.; Guo, X.; Wang, M.; Lu, H.; Gu, X.; Tian, H. Adhesion and Colonization of the Probiotic *Lactobacillus rhamnosus* Labeled by Dsred2 in Mouse Gut. *Curr. Microbiol.* **2019**, *76*, 896–903. [[CrossRef](#)]
18. Xing, Z.; Tang, W.; Yang, Y.; Geng, W.; Rehman, R.U.; Wang, Y. Colonization and Gut Flora Modulation of *Lactobacillus kefirifaciens* ZW3 in the Intestinal Tract of Mice. *Probiotics Antimicrob. Proteins* **2017**, *10*, 374–382. [[CrossRef](#)]
19. Nemcova, R.; Bomba, A.; Herich, R.; Gancarcikova, S. Colonization capability of orally administered *Lactobacillus* strains in the gut of gnotobiotic piglets. *DTW. Dtsch. Tierarztl. Wochenschr.* **1998**, *105*, 199–200.
20. Pessione, E. Lactic acid bacteria contribution to gut microbiota complexity: Lights and shadows. *Front. Cell Infect. Microbiol.* **2012**, *2*, 86. [[CrossRef](#)]
21. Arnold, J.W.; Simpson, J.B.; Roach, J.; Bruno-Barcena, J.M.; Azcarate-Peril, M.A. Prebiotics for Lactose Intolerance: Variability in Galacto-Oligosaccharide Utilization by Intestinal *Lactobacillus rhamnosus*. *Nutrients* **2018**, *10*. [[CrossRef](#)] [[PubMed](#)]
22. Saubade, F.; Hemery, Y.M.; Guyot, J.P.; Humblot, C. Lactic acid fermentation as a tool for increasing the folate content of foods. *Crit. Rev. Food Sci. Nutr.* **2017**, *57*, 3894–3910. [[CrossRef](#)] [[PubMed](#)]
23. Heeney, D.D.; Gareau, M.G.; Marco, M.L. Intestinal *Lactobacillus* in health and disease, a driver or just along for the ride? *Curr. Opin. Biotechnol.* **2018**, *49*, 140–147. [[CrossRef](#)] [[PubMed](#)]
24. Stockinger, B.; Di Meglio, P.; Gialitakis, M.; Duarte, J.H. The aryl hydrocarbon receptor: Multitasking in the immune system. *Annu. Rev. Immunol.* **2014**, *32*, 403–432. [[CrossRef](#)]
25. Caspi, R.; Altman, T.; Dreher, K.; Fulcher, C.A.; Subhraveti, P.; Keseler, I.M.; Kothari, A.; Krummenacker, M.; Latendresse, M.; Mueller, L.A.; et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.* **2012**, *40*, D742–D753. [[CrossRef](#)]
26. Hahn, A.S.; Altman, T.; Konwar, K.M.; Hanson, N.W.; Kim, D.; Relman, D.A.; Dill, D.L.; Hallam, S.J. GutCyc: A Multi-Study Collection of Human Gut Microbiome Metabolic Models. *BioRxiv* **2016**. [[CrossRef](#)]

27. Shahid, S.U.; Irfan, U. The gut microbiota and its potential role in obesity. *Future Microbiol.* **2018**, *13*, 589–603.
28. Le Chatelier, E.; Nielsen, T.; Qin, J.; Prifti, E.; Hildebrand, F.; Falony, G.; Almeida, M.; Arumugam, M.; Batto, J.M.; Kennedy, S.; et al. Richness of human gut microbiome correlates with metabolic markers. *Nature* **2013**, *500*, 541–546. [[CrossRef](#)]
29. Wuthrich, D.; Berthoud, H.; Wechsler, D.; Eugster, E.; Irmeler, S.; Bruggmann, R. The Histidine Decarboxylase Gene Cluster of *Lactobacillus parabuchneri* Was Gained by Horizontal Gene Transfer and Is Mobile within the Species. *Front. Microbiol.* **2017**, *8*, 218. [[CrossRef](#)]
30. Bolger, A.M.; Lohse, M.; Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **2014**, *30*, 2114–2120. [[CrossRef](#)]
31. Bankevich, A.; Nurk, S.; Antipov, D.; Gurevich, A.A.; Dvorkin, M.; Kulikov, A.S.; Lesin, V.M.; Nikolenko, S.I.; Pham, S.; Pribelski, A.D.; et al. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **2012**, *19*, 455–477. [[CrossRef](#)] [[PubMed](#)]
32. Chin, C.S.; Alexander, D.H.; Marks, P.; Klammer, A.A.; Drake, J.; Heiner, C.; Clum, A.; Copeland, A.; Huddleston, J.; Eichler, E.E.; et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* **2013**, *10*, 563–569. [[CrossRef](#)] [[PubMed](#)]
33. Pfrunder, S.; Grossmann, J.; Hunziker, P.; Brunisholz, R.; Gekenidis, M.T.; Drissner, D. *Bacillus cereus* Group-Type Strain-Specific Diagnostic Peptides. *J. Proteome Res.* **2016**, *15*, 3098–3107. [[CrossRef](#)] [[PubMed](#)]
34. Daniel Wüthrich, R.B. Bacterial Genome Contamination Analysis. Available online: https://github.com/danielwuethrich87/contamination_analysis_of_assembly (accessed on 27 January 2020).
35. Seemann, T. Prokka: Rapid prokaryotic genome annotation. *Bioinformatics* **2014**, *30*, 2068–2069. [[CrossRef](#)] [[PubMed](#)]
36. Gasteiger, E.; Jung, E.; Bairoch, A. SWISS-PROT: Connecting biomolecular knowledge via a protein database. *Curr. Issues Mol. Biol.* **2001**, *3*, 47–55.
37. Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403–410. [[CrossRef](#)]
38. Ester, M.; Kriegel, H.-P.; Sander, J.; Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of the KDD (Knowledge Discovery and Data Mining), Portland, OR, USA, 2–4 August 1996; pp. 226–231.
39. El-Gebali, S.; Mistry, J.; Bateman, A.; Eddy, S.R.; Luciani, A.; Potter, S.C.; Qureshi, M.; Richardson, L.J.; Salazar, G.A.; Smart, A.; et al. The Pfam protein families database in 2019. *Nucleic Acids Res.* **2019**, *47*, D427–D432. [[CrossRef](#)]
40. Daniel Wüthrich, R.B. PfastGO. Available online: <https://github.com/danielwuethrich87/pfastGO> (accessed on 27 January 2020).
41. Waterhouse, R.M.; Seppey, M.; Simao, F.A.; Manni, M.; Ioannidis, P.; Klioutchnikov, G.; Kriventseva, E.V.; Zdobnov, E.M. BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol. Biol. Evol.* **2017**. [[CrossRef](#)]
42. Van Rossum, T.; Ferretti, P.; Maistrenko, O.M.; Bork, P. Diversity within species: Interpreting strains in microbiomes. *Nat. Rev. Microbiol.* **2020**. [[CrossRef](#)]
43. Zou, Y.; Xue, W.; Luo, G.; Deng, Z.; Qin, P.; Guo, R.; Sun, H.; Xia, Y.; Liang, S.; Dai, Y.; et al. 1520 reference genomes from cultivated human gut bacteria enable functional microbiome analyses. *Nat. Biotechnol.* **2019**, *37*, 179–185. [[CrossRef](#)] [[PubMed](#)]
44. Altman, T.; Travers, M.; Kothari, A.; Caspi, R.; Karp, P.D. A systematic comparison of the MetaCyc and KEGG pathway databases. *BMC Bioinform.* **2013**, *14*, 112. [[CrossRef](#)] [[PubMed](#)]
45. Roder, T. MetaCycSuperPathwayParser. Available online: <https://binfgitlab.unibe.ch/troder/metacycsuperpathwayparser> (accessed on 3 March 2020).
46. Li, Q.; Chen, H.; Zhang, M.; Wu, T.; Liu, R.; Zhang, Z. Potential Correlation between Dietary Fiber-Suppressed Microbial Conversion of Choline to Trimethylamine and Formation of Methylglyoxal. *J. Agric. Food Chem.* **2020**, *67*, 13247–13257. [[CrossRef](#)] [[PubMed](#)]
47. Zhu, Y.; Li, Q.; Jiang, H. Gut Microbiota in Atherosclerosis: Focus on Trimethylamine N-Oxide. *APMIS* **2020**. [[CrossRef](#)]
48. Burton, K.J.; Kruger, R.; Scherz, V.; Munger, L.H.; Picone, G.; Vionnet, N.; Bertelli, C.; Greub, G.; Capozzi, F.; Vergeres, G. Trimethylamine-N-Oxide Postprandial Response in Plasma and Urine Is Lower After Fermented Compared to Non-Fermented Dairy Consumption in Healthy Adults. *Nutrients* **2020**, *12*. [[CrossRef](#)]

49. Kushkevych, I.; Cejnar, J.; Treml, J.; Dordevic, D.; Kollar, P.; Vitezova, M. Recent Advances in Metabolic Pathways of Sulfate Reduction in Intestinal Bacteria. *Cells* **2020**, *9*. [[CrossRef](#)]
50. Carbonero, F.; Benefiel, A.C.; Alizadeh-Ghamsari, A.H.; Gaskins, H.R. Microbial pathways in colonic sulfur metabolism and links with health and disease. *Front. Physiol.* **2012**, *3*, 448. [[CrossRef](#)]
51. Tabassum, R.; Jeong, N.Y. Potential for therapeutic use of hydrogen sulfide in oxidative stress-induced neurodegenerative diseases. *Int. J. Med. Sci.* **2019**, *16*, 1386–1396. [[CrossRef](#)]
52. Kushkevych, I.; Kotrsova, V.; Dordevic, D.; Bunkova, L.; Vitezova, M.; Amedei, A. Hydrogen Sulfide Effects on the Survival of Lactobacilli with Emphasis on the Development of Inflammatory Bowel Diseases. *Biomolecules* **2019**, *9*. [[CrossRef](#)]
53. Bravo, J.A.; Forsythe, P.; Chew, M.V.; Escaravage, E.; Savignac, H.M.; Dinan, T.G.; Bienenstock, J.; Cryan, J.F. Ingestion of Lactobacillus strain regulates emotional behavior and central GABA receptor expression in a mouse via the vagus nerve. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 16050–16055. [[CrossRef](#)] [[PubMed](#)]
54. Hawley, S.A.; Fullerton, M.D.; Ross, F.A.; Schertzer, J.D.; Chevtzoff, C.; Walker, K.J.; Pegg, M.W.; Zibrova, D.; Green, K.A.; Mustard, K.J.; et al. The ancient drug salicylate directly activates AMP-activated protein kinase. *Science* **2012**, *336*, 918–922. [[CrossRef](#)] [[PubMed](#)]
55. Michalek, R.D.; Gerriets, V.A.; Jacobs, S.R.; Macintyre, A.N.; MacIver, N.J.; Mason, E.F.; Sullivan, S.A.; Nichols, A.G.; Rathmell, J.C. Cutting edge: Distinct glycolytic and lipid oxidative metabolic programs are essential for effector and regulatory CD4+ T cell subsets. *J. Immunol.* **2011**, *186*, 3299–3303. [[CrossRef](#)] [[PubMed](#)]
56. Sridharan, G.V.; Choi, K.; Klemashevich, C.; Wu, C.; Prabakaran, D.; Pan, L.B.; Steinmeyer, S.; Mueller, C.; Yousofshahi, M.; Alaniz, R.C.; et al. Prediction and quantification of bioactive microbiota metabolites in the mouse gut. *Nat. Commun.* **2014**, *5*, 5492. [[CrossRef](#)] [[PubMed](#)]
57. Bock, K.W. Aryl hydrocarbon receptor (AHR): From selected human target genes and crosstalk with transcription factors to multiple AHR functions. *Biochem. Pharmacol.* **2019**, *168*, 65–70. [[CrossRef](#)]
58. Damman, C.J. Salicylates and the Microbiota: A New Mechanistic Understanding of an Ancient Drug's Role in Dermatological and Gastrointestinal Disease. *Drug Dev. Res.* **2013**, *74*, 344–352. [[CrossRef](#)]
59. Alp, D.; Kuleasan, H. Adhesion mechanisms of lactic acid bacteria: Conventional and novel approaches for testing. *World J. Microbiol. Biotechnol.* **2019**, *35*, 156. [[CrossRef](#)]
60. Nishiyama, K.; Sugiyama, M.; Mukai, T. Adhesion Properties of Lactic Acid Bacteria on Intestinal Mucin. *Microorganisms* **2016**, *4*. [[CrossRef](#)]
61. Rajilic-Stojanovic, M.; de Vos, W.M. The first 1000 cultured species of the human gastrointestinal microbiota. *FEMS Microbiol. Rev.* **2014**, *38*, 996–1047. [[CrossRef](#)]
62. De Paepe, K.; Verspreet, J.; Rezaei, M.N.; Hidalgo Martinez, S.; Meysman, F.; Van de Walle, D.; Dewettinck, K.; Raes, J.; Courtin, C.; Van de Wiele, T. Isolation of wheat bran-colonizing and metabolizing species from the human fecal microbiota. *PeerJ* **2019**, *7*, e6293. [[CrossRef](#)]
63. Wang, Y.; Guan, M.; Zhao, X.; Li, X. Effects of garlic polysaccharide on alcoholic liver fibrosis and intestinal microflora in mice. *Pharm. Biol.* **2018**, *56*, 325–332. [[CrossRef](#)] [[PubMed](#)]
64. Treangen, T.J.; Rocha, E.P. Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. *PLoS Genet.* **2011**, *7*, e1001284. [[CrossRef](#)] [[PubMed](#)]
65. Xu, J.; Mahowald, M.A.; Ley, R.E.; Lozupone, C.A.; Hamady, M.; Martens, E.C.; Henrissat, B.; Coutinho, P.M.; Minx, P.; Latreille, P.; et al. Evolution of symbiotic bacteria in the distal human intestine. *PLoS Biol.* **2007**, *5*, e156. [[CrossRef](#)] [[PubMed](#)]



3.2. Manuscript 2: OpenGenomeBrowser

OpenGenomeBrowser: A versatile, dataset-independent and scalable web platform for genome data management and comparative genomics

Thomas Roder, Simone Oberhänsli, Noam Shani, Rémy Bruggmann

Status:

Published in BMC Genomics (2022), 23(1):855 [190]

Statement of contribution:

TR, SO and RB conceived the project. TR programmed the software. SO, RB and NS contributed conceptually and with feedback to the software. TR and RB wrote the manuscript. All authors edited, read, and approved the final manuscript.

Research objective:

To design a user-friendly, dataset-independent platform for genome data management and comparative genomics.

SOFTWARE

Open Access



OpenGenomeBrowser: a versatile, dataset-independent and scalable web platform for genome data management and comparative genomics

Thomas Roder¹, Simone Oberhänsli¹, Noam Shani² and Rémy Bruggmann^{1*}

Abstract

Background: As the amount of genomic data continues to grow, there is an increasing need for systematic ways to organize, explore, compare, analyze and share this data. Despite this, there is a lack of suitable platforms to meet this need.

Results: OpenGenomeBrowser is a self-hostable, open-source platform to manage access to genomic data and drastically simplifying comparative genomics analyses. It enables users to interactively generate phylogenetic trees, compare gene loci, browse biochemical pathways, perform gene trait matching, create dot plots, execute BLAST searches, and access the data. It features a flexible user management system, and its modular folder structure enables the organization of genomic data and metadata, and to automate analyses. We tested OpenGenomeBrowser with bacterial, archaeal and yeast genomes. We provide a docker container to make installation and hosting simple. The source code, documentation, tutorials for OpenGenomeBrowser are available at opengenomebrowser.github.io and a demo server is freely accessible at opengenomebrowser.bioinformatics.unibe.ch.

Conclusions: To our knowledge, OpenGenomeBrowser is the first self-hostable, database-independent comparative genome browser. It drastically simplifies commonly used bioinformatics workflows and enables convenient as well as fast data exploration.

Keywords: Genome database, Genome browser, Comparative genomics, Open-source, Self-hosted

Background

Driven by advances in sequencing technologies, many organizations and research groups have accumulated large amounts of genomic data. As sequencing projects progress, the organization of such genomic datasets becomes increasingly difficult. Systematic ways of storing data and metadata, tracking and denoting changes in

assemblies or annotations, and enabling easy access are key challenges. While standardized data formats and free software are widely used in the field to process genomic data, data exploration is often still cumbersome. This is especially true for non-bioinformaticians, although numerous platforms have been developed to simplify data access.

Most of these platforms have different user interfaces and sometimes limited functionality. The reason for this heterogeneity is that most of them have been developed independently, i.e., each one for a specific genomic dataset. Such platforms exist for many well-studied organisms, such as *Pseudomonas* spp. [1], but also for

*Correspondence: remy.bruggmann@bioinformatics.unibe.ch

¹ Interfaculty Bioinformatics Unit and Swiss Institute of Bioinformatics, University of Bern, 3012 Bern, Switzerland
Full list of author information is available at the end of the article



© The Author(s) 2022. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

non-model species such as ginseng [2] and cork oak [3]. These platforms share a set of core features: access to data, sequence similarity searches (like BLAST [4]), and limited annotation searches. The most advanced of these platforms, such as CoGe [5], MicrobesOnline [6], WormBase [7], Genomicus [8], MicroScope [9] and ChlamDB [10], include additional functions to answer a wide range of questions.

However, these platforms tend to be tied to the characteristics of a specific dataset and adapting their software to other projects would be extremely difficult. This is surprising given that the underlying data are essentially the same: genome assemblies, genes, proteins, and their annotations. Fortunately, this information is stored in standardized data formats across many fields, which in principle would allow code reuse and collaborative development. Even while some degree of purpose-built software tools may still be necessary for certain projects, independent development comes at a significant initial cost as well as a long-term maintenance cost and a higher risk of becoming outdated.

We addressed these issues by developing OpenGenomeBrowser, a self-hostable, open-source software based on the Python web framework Django [11]. OpenGenomeBrowser runs on all modern browser engines (Firefox, Chrome, Safari). It contains more features than most similar platforms, is highly user-friendly and *dataset-independent* – i.e., not bound to any specific genomic dataset. A comparison of OpenGenomeBrowser and similar platforms is available in Table S1.

Implementation

To enable automated processing of genomic data, as in OpenGenomeBrowser, it is essential that the data is stored in a systematic fashion. We present our solution to this problem in detail in the section “*folder structure*”. The subsequent section “*OpenGenomeBrowser tools*” describes a set of scripts that simplify the handling of the aforementioned folder structure.

Folder structure

Every sequencing project faces an important challenge: systematic storage of data and metadata according to the FAIR principles [12]. These principles enable reproducibility, automation, data interoperability and sharing. Especially in long-term projects, it is crucial to know when and how the data was generated, and to have a transparent way of handling different genome and annotation versions. Different versions are the result of organism re-sequencing, raw data re-assembly or assembly re-annotation. Importantly, each version of a gene must have a unique identifier, and legacy data should be kept instead of being overwritten.

To address these problems, we developed a modular folder structure (Fig. 1A). The *organisms* folder contains a directory for each biological entity, e.g., a bacterial strain. Each of these folders must contain a metadata file, *organism.json* (Fig. 1A, center), describing the biological entity, and a folder named *genomes*. The *genomes* folder contains one folder for each genome version. One of these genomes must be designated as the *representative* genome of the biological entity in *organism.json*. This allows project maintainers to update an assembly transparently, by designating the new version as *representative* without removing the old one.

Each genome folder must contain a metadata file, *genome.json* (Fig. 1A), and the actual data: an assembly FASTA file, a GenBank file, and a gff3 (general feature format version 3) file. While not strictly required but strongly recommended, annotation files in tab-separated format which map gene identifiers to annotations, may be provided. OpenGenomeBrowser supports several annotation types by default, such as Enzyme Commission numbers, KEGG [13] genes and KEGG reactions, Gene Ontology terms [14, 15], and annotations from EggNOG [16]. Additional annotation types can be easily configured. Files that map annotations to descriptions (e.g., EC:1.1.1.1 → alcohol dehydrogenase) can be added to a designated folder.

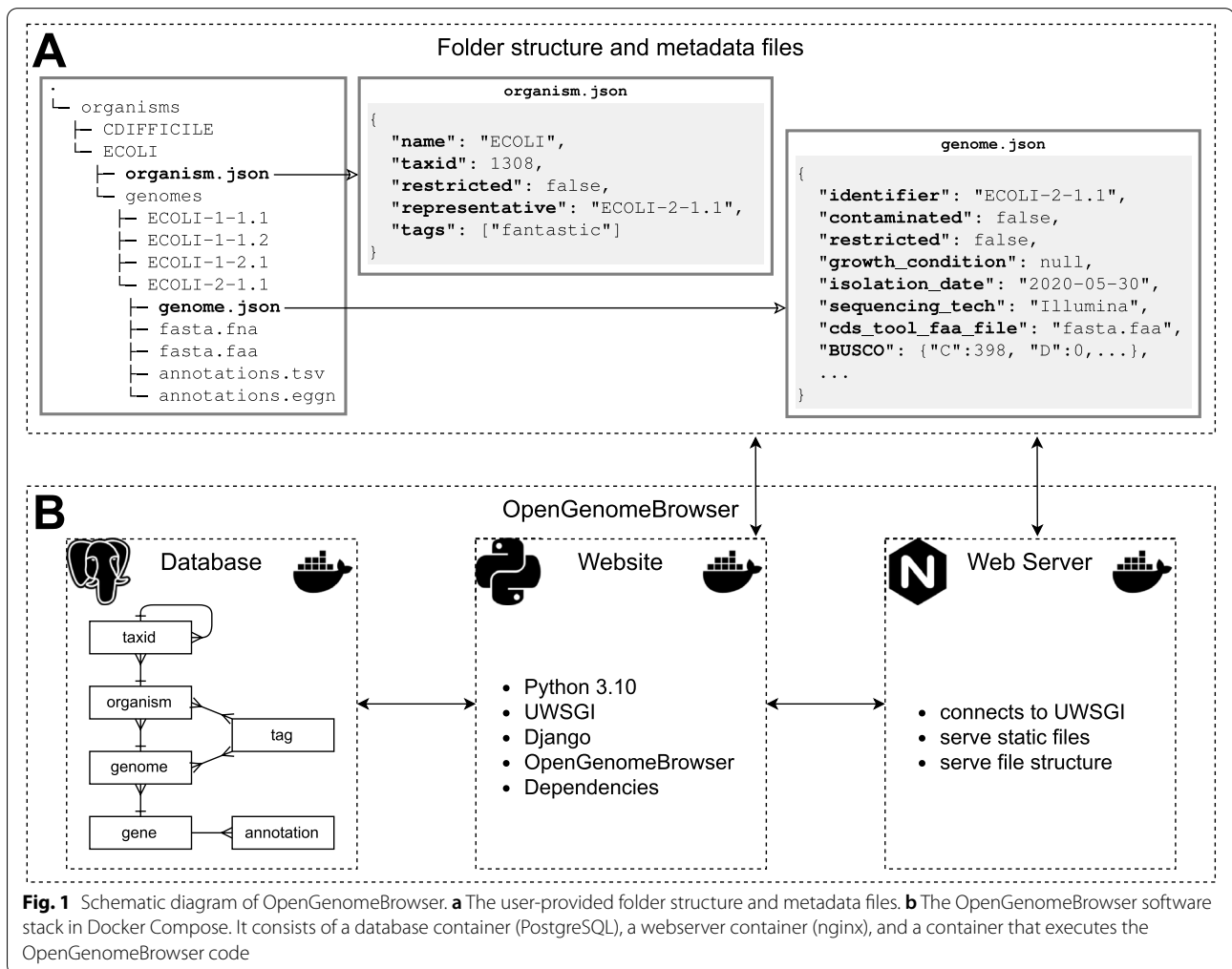
OpenGenomeBrowser tools

A set of scripts called *OpenGenomeBrowser Tools* simplifies the creation of the previously described folder structure and the incorporation of new genomes. As shown below, a functional folder structure that contains one genome can be set up with only four commands.

```
#!/bin/bash
# Install OpenGenomeBrowser Tools (requires Python 3.10+)
pip install opengenomebrowser-tools
# Set desired location of the folder structure
export FOLDER_STRUCTURE=/path/to/folder_structure
# Create a bare-bone folder structure
# Download annotation descriptions for default annotation types
init_folder_structure
# Add a genome to the folder structure. The import-dir must at least contain:
# - an assembly FASTA (.fna)
# - a GenBank file (.gbk)
# - a general feature format file (.gff)
# The output directories of Prokka [17] and PGAP [18] are directly compatible.
import_genome --import-dir=/path/to/genomic/files
```

Software architecture

OpenGenomeBrowser itself is distributed as a Docker container [19]. Using Docker Compose, the container is combined with a database and a webserver to create a production-ready software stack (Fig. 1B).



Results and discussion

The following section describes the main features of OpenGenomeBrowser. The reader may try them out at opengenomebrowser.bioinformatics.unibe.ch, where a freely accessible demo server with 70 bacterial genomes is hosted. Notably, on most pages, users may click on *Tools*, then *Get help with this page* to be redirected to a site that explains how the tool works and how to use it. Moreover, advanced configuration options are available on some pages. They can be accessed via a sidebar that opens when one clicks on the settings wheel (⚙) at the top right corner of the page.

Genomes table

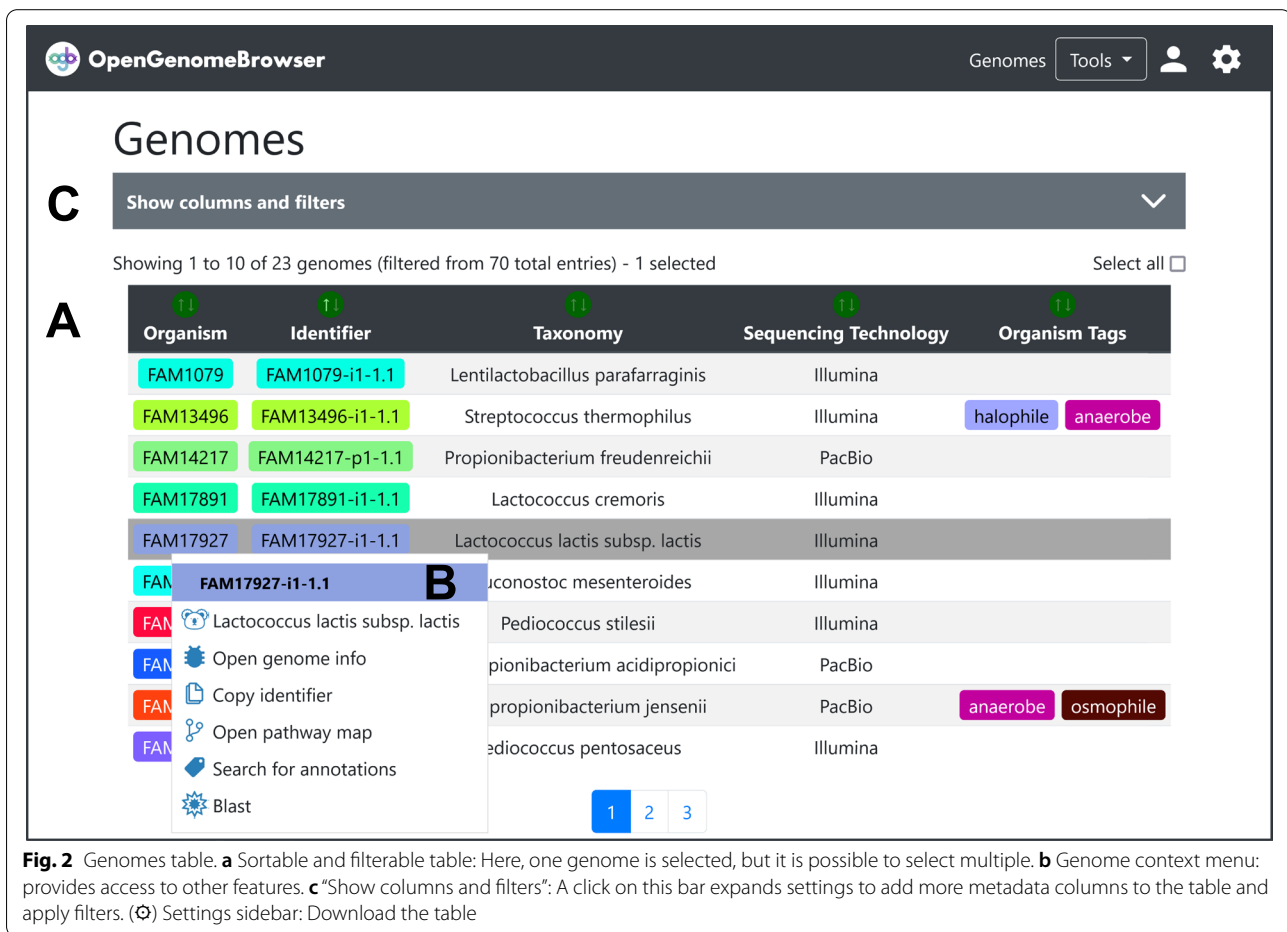
Especially in large sequencing projects, it is vital that the data can be filtered and sorted according to metadata. This is the purpose of the *genomes table view* (Fig. 2) which serves as the entry point of OpenGenomeBrowser. By default, only the *representative* genomes are listed and only the name of

the organism, the genome identifier, the taxonomic name, and the sequencing technology are shown as columns. Furthermore, there are over forty additional metadata columns available that can be dynamically added to the table. All columns can be used to filter and sort the data, which makes this view the ideal entry point for an analysis.

Detail views

The *genome detail view* (Fig. S1A) shows all available metadata of the respective genome and allows the user to download the associated files.

The *gene detail view* (Fig. S1B) is designed to facilitate easy interpretation of the putative functions of genes. It shows all annotations, their descriptions, the nucleotide- and protein sequences, metadata from the GenBank file and an interactive gene locus visualization facilitated by DNA features viewer [20]. If the gene is annotated with a gene ontology term that represents a subcellular location, this location will be highlighted on a SwissBioPics image [21].



Genomes in OpenGenomeBrowser can be labelled with tags, i.e., a short name (e.g., “*halophile*”) and a description (e.g., “*extremophiles that thrive in high salt concentrations*”). The *tag detail view* (Fig. S1C) shows the description of the tag and the genomes that are associated with it. Tags are particularly useful to quickly select groups of genomes in many tools of OpenGenomeBrowser. For example, to select all genomes with the tag “*halophile*”, the syntax “@tag:halophile” can be used.

Similarly, the *TaxId detail view* (Fig. S1D) shows all genomes that belong to the respective NCBI Taxonomy identifier (TaxId) [22], as well as the parent TaxId. Similar to tags, TaxIds can be used to select all genomes that belong to a certain TaxId, like this: “@taxphylum:Firmicutes”, or simply “@tax:Firmicutes”.

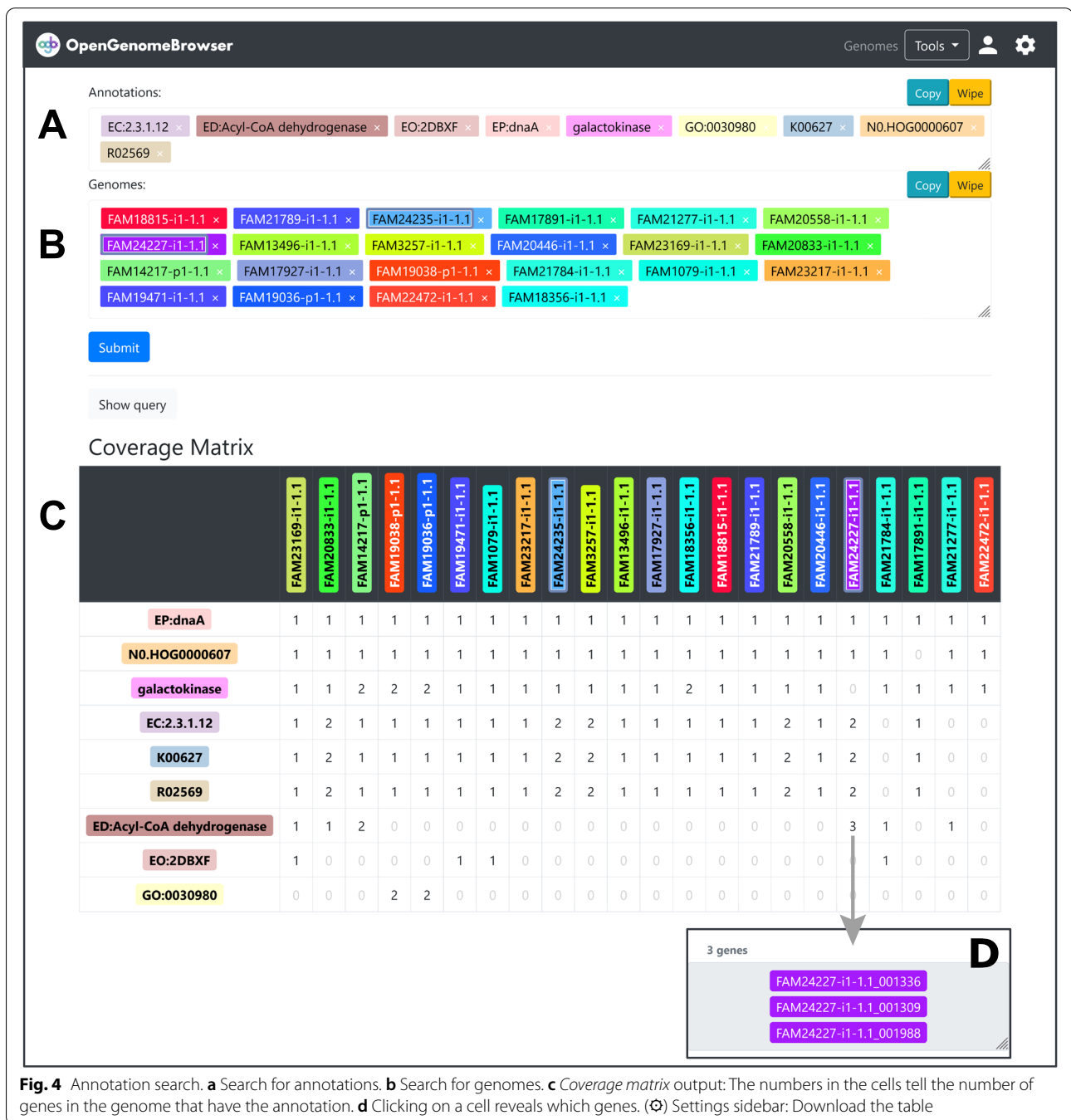
Gene comparison

The *gene comparison view* (Fig. 3) enables users to easily compute multiple sequence alignments and to compare gene loci side-by-side. Currently, Clustal Omega [23], MAFFT [24] and MUSCLE [25] are supported alignment algorithms. Alignments are visualized using MSViewer

[26] (Fig. 3B). Furthermore, the genomic regions around the genes of interest can be analyzed using a customized implementation of DNA features viewer [20] (Fig. 3C). Figure 3 shows an alignment of all genes on the demo server that contain the annotation *K01610* (phosphoenolpyruvate carboxykinase; from the pyruvate metabolism pathway). The gene loci comparison reveals that in all queried *Lactocaseibacilli*, the genes are located in syntenic regions, i.e., next to the same orthologous genes.

Annotation search

Despite conceptually and technically straightforward, searching for annotations in a set of genomes can be tedious or even impossible for non-programmers. In OpenGenomeBrowser, annotation search is quick and easy, thanks to the PostgreSQL backend that allows fast processing of annotation information. In the *annotation search view* (Fig. 4), users can search for annotations in genomes, resulting in a *coverage matrix* (Fig. 4C) with one column per genome and one row per annotation. The numbers in the cells show how many genes in the genome have the same annotation. Clicking on these cells shows the relevant genes (Fig. 4D), while

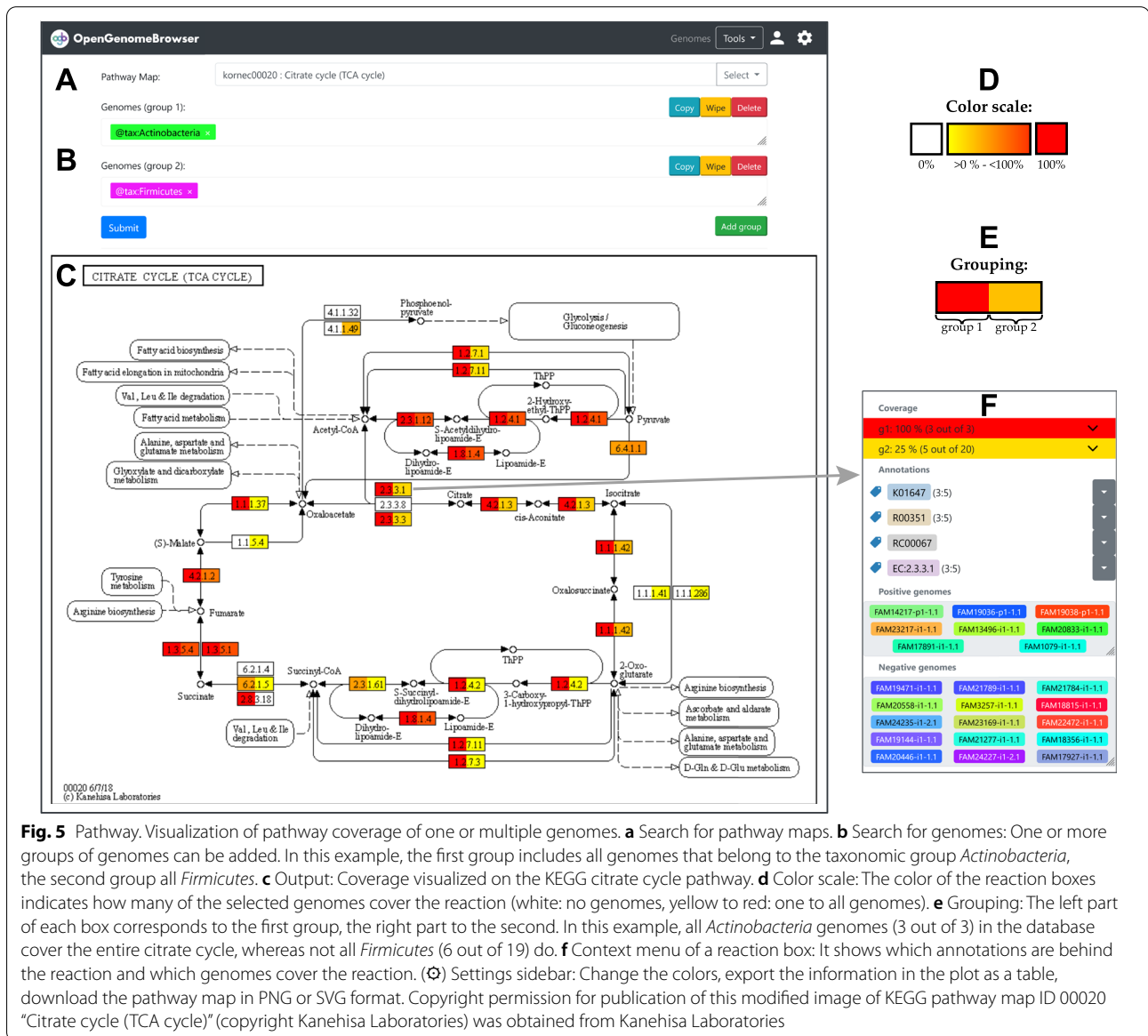


clicking on an annotation enables users to compare the corresponding genes (*gene comparison view*).

Pathways

Pathway maps, particularly the ones from the KEGG [27], are valuable tools to understand the metabolism of an organism. However, using them may be cumbersome. Commonly, biologists upload sequences to a service like BlastKOALA [28]. This service is designed to

process one organism at a time, and calculation times can last multiple hours. Because each genome must be submitted individually, it becomes cumbersome when multiple organisms must be processed. Furthermore, it is not trivial to visualize multiple genomes on a pathway map. In OpenGenomeBrowser, this process is straightforward (Fig. 5A-C), user-friendly, and fast, as the annotations are pre-calculated and loaded into the database beforehand. Pathway maps are interactive, which allows



the user to explore this information in great detail (Fig. 5D-F). For example, to investigate the genes that are involved in a certain enzymatic step, one needs only to click on the enzyme box, then on an annotation of interest, and finally on “compare the genes” to be redirected to *gene comparison view*.

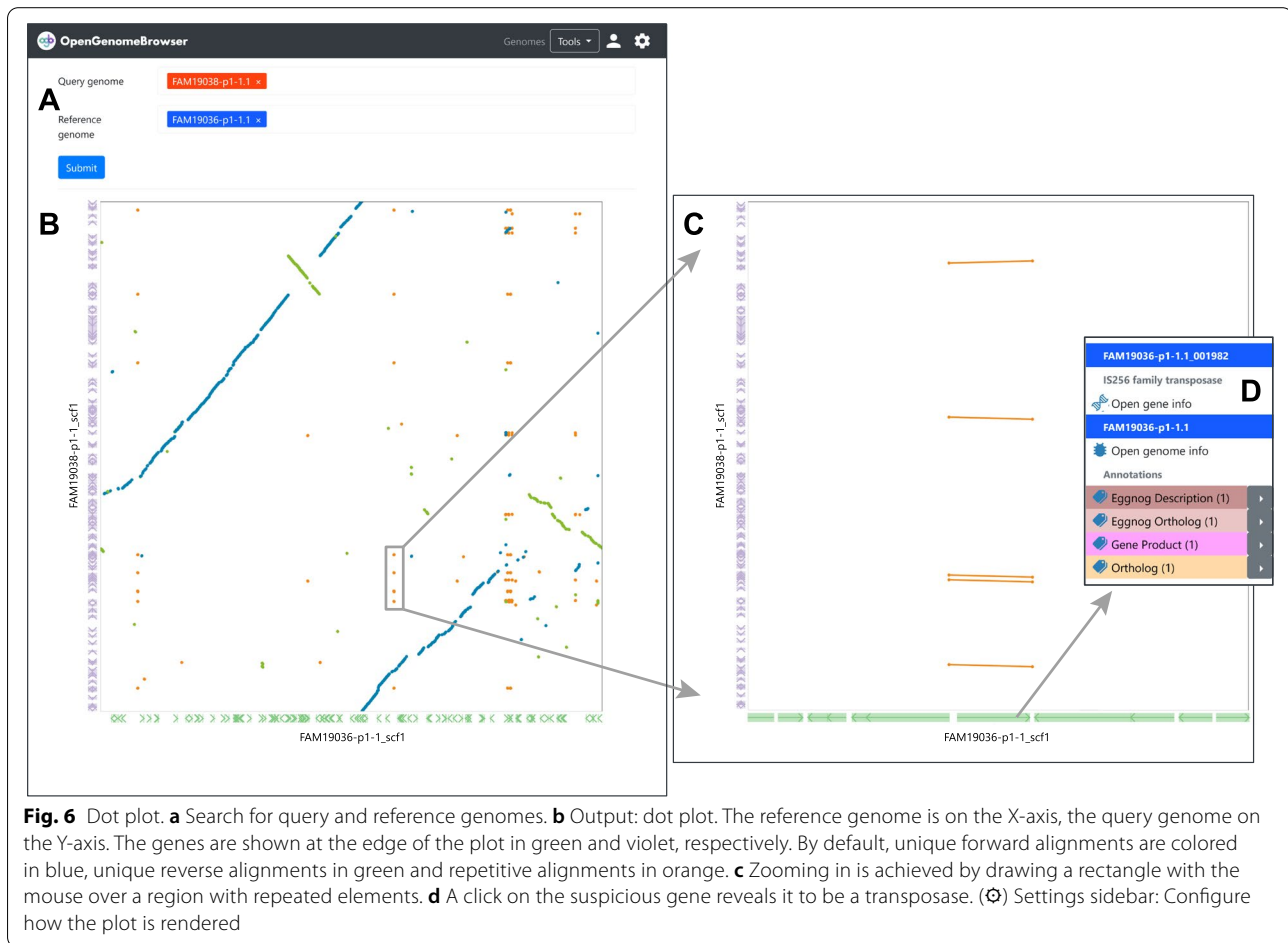
While OpenGenomeBrowser does not include KEGG maps for licensing reasons, users with appropriate rights can generate them using a separate program [29]. The pathway maps do not necessarily have to be from KEGG. Pathway maps in a custom Scalable Vector Graphics (SVG) may be added to a designated folder in the folder structure (not shown in Fig. 1).

Blast

OpenGenomeBrowser allows users to perform a local alignment of protein and nucleotide sequences using BLAST [4]. The results are visualized using the BlasterJS [30] library.

Trees

OpenGenomeBrowser computes three kinds of phylogenetic trees. The fastest type of tree is based on the NCBI taxonomy ID which is registered in the metadata. It is helpful to get a quick taxonomic overview, but it entirely depends on the accuracy of the metadata.



The second type of tree is based on genome similarity. The assemblies of the selected genomes are compared to each other using GenDisCal-PaSiT6, a fast, hexanucleotide-frequency-based algorithm with similar accuracy as average nucleotide identity (ANI) based methods [31]. This algorithm yields a similarity matrix from which a dendrogram is calculated with the unweighted pair group method with arithmetic mean (UPGMA) algorithm [32]. We recommend this type of tree as a good compromise between speed and accuracy, specifically if many genomes are to be compared.

The third type of tree is based on the alignment of single-copy orthologous genes. This type of tree is calculated using the OrthoFinder [33] algorithm. Of all proposed tree type algorithms it is the most time- and computation-intensive and requires pre-computed all-vs-all DIAMOND [34] searches.

Dot plot

Dot plot is a simple and established [35] method of comparing two genome assemblies. It allows the discovery of insertions, deletions, and duplications, especially in

closely related genomes sequenced with long-read technologies. In OpenGenomeBrowser's implementation of dot plot, the assemblies are aligned against each other using MUMmer [36] and visualized using the *Dot* library [37]. The resulting plot (Fig. 6) is interactive, i.e., the user can zoom in on regions of interest by drawing a rectangle with the mouse and clicking on a gene which then opens the context menu with detailed information.

Gene trait matching

The *gene trait matching view* enables users to find annotations that correlate with a (binary) phenotypic trait. The input must consist of two non-intersecting sets of organisms that differ in a trait. OpenGenomeBrowser applies a Fisher's exact test for each orthologous gene and corrects for multiple testing ($\alpha = 10\%$) using the Benjamini-Hochberg method [38, 39]. The multiple testing parameters can be adjusted in the settings sidebar. The test can be used on orthogenes as well as any other type of annotation, such as KEGG-gene annotation. The gene candidates that may be causing the trait can easily be

further analyzed, for example by using the *compare genes view*.

Flower plot

The *flower plot view* provides the users with a simple overview of the shared genomic content of multiple genomes. The genomes are displayed as petals of a flower. Each petal indicates the number of annotations that are unique to this genome and the number of genes that are shared by some but not all others. The number of genes shared by all genomes is indicated in the center of the flower. (The code is also available as a standalone Python package [40]).

Downloader

The *downloader view* facilitates the convenient download of multiple raw data files, for example all protein FASTA files for a set of organisms.

Admin panel

OpenGenomeBrowser has a powerful user authentication system and admin interface, inherited from the Django framework. Instances of OpenGenomeBrowser can be configured to require a login or to allow basic access to anonymous users. Users can be given specific permissions, for example to create other user accounts, to edit metadata of organisms, genomes, and tags, and even to upload new genomes through the browser.

Resource requirements

OpenGenomeBrowser is not resource intensive. An instance containing over 1400 bacterial genomes runs on a computer with 8 CPU-cores (2.4GHz) and 20GB of RAM. The Docker container is about 3GB in size and the Postgres database takes 21GB of storage (SSD recommended).

Conclusions

OpenGenomeBrowser is, to our knowledge, the first comparative genome browser that is not tied to a specific dataset. It automates commonly used bioinformatics workflows, enabling convenient and fast data exploration, particularly for non-bioinformaticians, in an intuitive and user-friendly way.

The software has minimal hardware requirements and is easy to install, host, and update. OpenGenomeBrowser's folder structure enforces systematic yet flexible storage of genomic data, including associated metadata. This folder structure (i) enables automation of analyses, (ii) guides users to maintain their data in a coherent and structured way, and (iii) provides version tracking, a precondition for reproducible research.

OpenGenomeBrowser is flexible and scalable. It can run on a local machine or on a public server, access may be open for anyone or restricted to authenticated users. Annotation types can be customized, and ortholog-based features are optional. While the demo server only holds 70 genomes, the performance scales and is still outstanding even when hosting over 1400 microbial genomes [41].

We believe that our software will be useful to a large community since sequencing microbial and other genomes has become a commodity. Therefore, researchers performing new sequencing projects can directly benefit from OpenGenomeBrowser by saving development costs, making their data potentially FAIR, and adapting the browser for their purposes. It could also replace older, custom-made platforms which may be outdated and more difficult to maintain. Because our software is open-source, adaptations of OpenGenomeBrowser and new features will be available for the whole community under the same conditions. The open-source model also allows problems to be identified and quickly fixed by the community, making OpenGenomeBrowser a sustainable platform.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12864-022-09086-3>.

Additional file 1: Table S1. Comparison of OpenGenomeBrowser's features with alternative software platforms. Legend: ✓: feature present; ⚠: feature present, but with limitations; ✗: feature absent. Features were inferred to the best of our knowledge.

Additional file 2: Fig. S1. Detail views. (A) Genome detail view: Shows genome-associated metadata. (B) Gene detail view: Displays a gene's annotations, nucleotide- and protein sequence, metadata extracted from the GenBank file, as well as an interactive plot that shows the adjacent genes. (C) Tag detail view: Shows the tag's name, its description and the organisms and genomes that have it. (D) TaxId detail view: Shows the NCBI TaxId, its taxonomic rank, its parent TaxId and the organisms and their genomes that belong to it.

Acknowledgements

We are grateful to Darja Studer for designing the logo, Lars Vögtlin for his advice on containerization, Linda Studer for her advice on the manuscript, to Kimberly Gilbert for proofreading the article, and Pierre Berthier for his support in hosting OpenGenomeBrowser. We thank Emmanuelle Arias-Roth, Remo Schmidt, Cornelia Bär, Ueli von Ah und Guy Vergères (Agroscope) for their support and feedback on this project.

Availability and requirements

Project name: OpenGenomeBrowser.
Project home page: <https://opengenomebrowser.github.io/>
Operating system(s): Linux (hosting); platform independent (usage).
Programming language: Python, JavaScript.
Other requirements: Docker.
License: GPL-3.
Any restrictions to use by non-academics: GPL-3.

Authors' contributions

TR, SO and RB conceived the project. TR programmed the software. SO, RB and NS contributed conceptually and with feedback to the software. TR and

RB wrote the manuscript. All authors edited, read, and approved the final manuscript.

Funding

This research was funded by Gebert RUF Stiftung within the program "Microbi-als", grant number GRS-070/17 and the Canton of Bern to RB. The funding bodies played no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Availability of data and materials

The data used to generate the figures in this study are included in the published article Roder et al., 2020 [41] where the GenBank accession numbers are listed in Supplementary Table S1. <https://doi.org/10.3390/microorganisms8070966>.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

¹Interfaculty Bioinformatics Unit and Swiss Institute of Bioinformatics, University of Bern, 3012 Bern, Switzerland. ²Methods Development and Analytics, Agroscope, Schwarzenburgstrasse 161, CH-3003 Bern, Switzerland.

Received: 15 August 2022 Accepted: 14 December 2022

Published online: 27 December 2022

References

- Winsor GL, Lam DKW, Fleming L, Lo R, Whiteside MD, Yu NY, et al. Pseudomonas Genome Database: Improved comparative analysis and population genomics capability for Pseudomonas genomes. *Nucleic Acids Res.* 2011 Jan;39(SUPPL. 1).
- Jayakodi M, Choi BS, Lee SC, Kim NH, Park JY, Jang W, et al. Ginseng genome database: an open-access platform for genomics of Panax ginseng. *BMC Plant Biol.* 2018 Apr;12:18(1).
- Arias-Baldrich C, Silva MC, Bergeretti F, Chaves I, Miguel C, Saibo NJM, et al. CorkOakDB-the cork oak genome database portal. *Database.* 2020;2020.
- Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. *BMC Bioinformatics.* 2009 Dec;10:10.
- Nelson ADL, Haug-Baltzell AK, Davey S, Gregory BD, Lyons E. EPIC-CoGe: managing and analyzing genomic data. *Bioinformatics.* 2018;34(15):2651–3.
- Dehal PS, Joachimiak MP, Price MN, Bates JT, Baumohl JK, Chivian D, et al. MicrobesOnline: An integrated portal for comparative and functional genomics. *Nucleic Acids Res.* 2009 Nov;38(SUPPL. 1).
- Harris TW, Arnaboldi V, Cain S, Chan J, Chen WJ, Cho J, et al. WormBase: a modern model organism information resource. *Nucleic Acids Res.* 2020 Jan 1;48(D1):D762–7.
- Nguyen NTT, Vincens P, Crollius HR, Louis A. Genomicus 2018: karyotype evolutionary trees and on-the-fly synteny computing. *Nucleic Acids Res.* 2018 Jan 1;46(D1):D816–22.
- Vallenet D, Calteau A, Dubois M, ... PAN acids, 2020 undefined. MicroScope: an integrated platform for the annotation and exploration of microbial gene functions through genomic, pangenomic and metabolic comparative analysis. [academic.oup.com](https://academic.oup.com/nar/article-abstract/48/D1/D579/5606622) [Internet]. [cited 2022 Nov 23]; Available from: <https://academic.oup.com/nar/article-abstract/48/D1/D579/5606622>
- Pillonel T, Tagini F, Bertelli C, Greub G. ChlamDB: a comparative genomics database of the phylum Chlamydiae and other members of the Planctomycetes-Verrucomicrobiae-Chlamydiae superphylum. *Nucleic Acids Res.* 2020;48(D1):D526–34.
- Django Software Foundation. Django [Internet]. Lawrence, Kansas: Django Software Foundation; 2013 [cited 2021 Jan 1]. Available from: <https://djangoproject.com/>
- Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR guiding principles for scientific data management and stewardship. *Sci Data.* 2016;3(1):1–9.
- Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes [Internet]. Vol. 28, *Nucleic Acids Research.* 2000. Available from: <http://www.genome.ad.jp/kegg/>
- Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology the gene ontology consortium* [Internet]. 2000. Available from: <http://www.flybase.bio.indiana.edu>
- Carbon S, Douglass E, Good BM, Unni DR, Harris NL, Mungall CJ, et al. The gene ontology resource: enriching a GOld mine. *Nucleic Acids Res.* 2021 Jan 8;49(D1):D325–34.
- Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, et al. EggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* 2019;47(D1):D309–14.
- Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics.* 2014 Jul 15;30(14):2068–9.
- Li W, O'Neill KR, Haft DH, Dicuccio M, Chetverin V, Badretdin A, et al. RefSeq: expanding the prokaryotic genome annotation pipeline reach with protein family model curation. *Nucleic Acids Res.* 2021 Jan 8;49(D1):D1020–8.
- Merkel D. Docker: lightweight linux containers for consistent development and deployment. *Linux journal.* 2014;2014(239):2.
- Zulkower V, Rosser S. DNA features viewer: a sequence annotation formatting and plotting library for Python. *Bioinformatics.* 2020 Aug 1;36(15):4350–2.
- Bolleman J, Bansal P, Redaschi N. SwissBioPics [Internet]. <https://www.swissbiopics.org/>. 2021 [cited 2021 Sep 1]. Available from: <https://www.swissbiopics.org/>
- Schoch CL, Ciufo S, Domrachev M, Hottot CL, Kannan S, Khovanskaya R, et al. NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database [Internet].* 2020 Jan 1;2020:baaa062. Available from: <https://doi.org/10.1093/database/baaa062>.
- Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal omega. *Mol Syst Biol.* 2011;7.
- Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol.* 2013 Apr;30(4):772–80.
- Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 2004;32(5):1792–7.
- Yachdav G, Wilzbach S, Rauscher B, Sheridan R, Sillitoe I, Procter J, et al. MSAViewer: interactive JavaScript visualization of multiple sequence alignments. *Bioinformatics.* 2016 Nov 15;32(22):3501–3.
- Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* 2016;44(D1):D457–62.
- Kanehisa M, Sato Y, Morishima K. BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. *J Mol Biol [Internet].* 2016;428(4):726–31 <https://www.sciencedirect.com/science/article/pii/S002228361500649X>.
- Roder T. KeggMapWizard [Internet]. Bern: GitHub; 2021. <https://github.com/MrTomRod/kegg-map-wizard>
- Blanco-Míguez A, Fdez-Riverola F, Sánchez B, Lourenço A. BlasterJS: a novel interactive JavaScript visualisation component for BLAST alignment results. *PLoS One.* 2018 Oct;13(10).
- Goussarov G, Goussarov G, Cleenwerck I, Mysara M, Leys N, Monsieurs P, et al. PaSiT: a novel approach based on short-oligonucleotide frequencies for efficient bacterial identification and typing. *Bioinformatics.* 2020 Apr 15;36(8):2337–44.
- Kunzmann P, Hamacher K. Biotite: a unifying open source computational biology framework in Python. *BMC Bioinformatics.* 2018 Oct;1:19(1).
- Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* 2019 Nov;14:20(1).

34. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. In: Vol. 12, Nature Methods: Nature Publishing Group; 2014. p. 59–60.
35. Gibbs AJ, McIntyre GA. The diagram, a method for comparing sequences its use with amino acid and nucleotide sequences. *Eur J Biochem.* 1970;16.
36. Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. MUMmer4: A fast and versatile genome alignment system. *PLoS Comput Biol.* 2018 Jan;14(1).
37. Maria Nattestad. Dot - an interactive dot plot viewer for genome-genome alignments. <https://github.com/MariaNattestad/dot>. 2021.
38. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods.* 2020 Mar 1;17(3):261–72.
39. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B Methodol.* 1995;57(1):289–300.
40. Thomas Roder. flower-plot [Internet]. GitHub. 2021 [cited 2022 Jan 1]. Available from: <https://github.com/MrTomRod/flower-plot>
41. Roder T, Wüthrich D, Bär C, Sattari Z, von Ah U, Ronchi F, et al. In Silico comparison shows that the Pan-genome of a dairy-related bacterial culture collection covers Most reactions annotated to human microbiomes. *Microorganisms.* 2020;8(7):966.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions



3.3. Manuscript 3: Scoary2

Scoary2: Rapid association of large phenotypic datasets to microbial pan-genomes

Thomas Roder, Grégory Pimentel, Pascal Fuchsmann, Mireille Tena Stern,
Ueli von Ah, Guy Vergères, Stephan Peischl, Ola Brynildsrud,
Rémy Bruggmann, Cornelia Bär

Status:

Manuscript submitted to BMC Genome Biology [191]

Statement of contribution:

GV, RB, CB, UvA, GP and TR conceived the project. GV and RB provided funding to the project. UvA and GP produced the yoghurts. GP measured the LC-MS yoghurt metabolome. PF and MTS measured the GC-MS volatiles yoghurt metabolomes. GP identified the carnitine metabolites. SP and TR explored different ways of analyzing the data. TR programmed the Scoary2 software with advice from OB. TR, CB and GP analyzed the data. TR, CB and OB wrote the manuscript with input from all authors. TR created the figures. All authors read and approved the final manuscript.

Research objective:

To create a software that enables the analysis of an entire metabolomics dataset using a population-structure-aware microbial GWAS approach.

Article

Scoary2: Rapid association of phenotypic multi-omics data with microbial pan-genomes

Thomas Roder¹, Grégory Pimentel², Pascal Fuchsmann³, Mireille Tena Stern³, Ueli von Ah³, Guy Vergères³, Stephan Peischl¹, Ola Brynildsrud⁴, Rémy Bruggmann^{1a*}, Cornelia Bär^{2a}

¹ Interfaculty Bioinformatics Unit and Swiss Institute of Bioinformatics and Graduate School for Cellular and Biomedical Sciences, University of Bern, 3012 Bern, Switzerland

² Methods development and analytics, Agroscope, Schwarzenburgstrasse 161, CH-3003 Bern, Switzerland

³ Food microbial systems, Agroscope, Schwarzenburgstrasse 161, CH-3003 Bern, Switzerland

⁴ Norwegian Institute of Public Health, Oslo and Norwegian University of Life Science, Ås, Norway

^a These authors share senior authorship

* Correspondence: Tel: +41 31 631 48 99; Email: remy.bruggmann@unibe.ch

Present address: [Rémy Bruggmann] Interfaculty Bioinformatics Unit, Baltzerstrasse 6, CH-3012 Bern, Switzerland.

Abstract

Genomic screening of bacteria is common practice to select strains with desired properties. However, 40-60% of all bacterial genes are still unknown, making capturing the phenotype an important part of the selection process. While omics-technologies collect high-dimensional phenotypic data, it remains challenging to link this information to genomic data to elucidate the impact of specific genes on phenotype. To this end, we present Scoary2, an ultra-fast software for microbial genome-wide association studies (mGWAS), enabling integrative data exploration. As proof of concept, we explore the metabolome of 44 yogurts with different strains of *Propionibacterium freudenreichii*, discovering two genes affecting carnitine metabolism.

Keywords

prokaryote, bacteria, pan-genome, metabolite, microbial genome-wide association studies, GWAS, BGWA, genotype-phenotype association, fermented food, omics

Background

The emergence of large-scale whole-genome sequencing, coupled with rapid development of tools for analyzing and sharing data, presents unprecedented opportunities to understand microbial genomics, to establish connections between genetic variations and functions, both at the level of individual organisms and within complex microbial communities. In the field of fermented foods, research focuses not only on studying the common characteristics of bacteria, but also on the individual abilities of certain bacteria to produce specific components in a product. Metabolic models can be used to gain deep insights into bacterial physiology, which is one possible approach to address these questions. However, useful models are challenging to develop, particularly for interacting microbial communities such as those present in yogurt and cheese [1]. These models make strong assumptions and are inherently limited to established and curated networks of genes and metabolites. Since the function of many bacterial genes (around 40-60% [2]) is not yet known, a more straightforward approach to learn about the function and interaction of genes is to capture the bacterial phenotype and correlate it with genomic data. While this may not lead to a holistic understanding of the microbes, it allows to select bacteria appropriate for the composition of specific fermented foods. In this context, individual genes can be key for acquiring a desired characteristic.

Although it is not an easy task to capture the full abundance of substances that make up the final nutritional composition of fermented food, recent developments in omics technologies such as mass spectrometry (MS) make it possible to capture massive metabolic profiles [3].

Even though sequencing and omics technologies are advancing rapidly, linking these data to gain an understanding of functional relationships remains a major challenge. Numerous and conceptually different approaches have been developed to integrate omics datasets, with a strong focus on human genetics and disease-related tasks such as disease subtyping or biomarker prediction [4]. Among these, only a few attempt to directly link phenotypes measured by omics technologies to genes using established human genome-wide association (hGWAS) or quantitative trait loci (QTL) methods [5]. Unfortunately, due to the differences between human and microbial genomes, hGWAS methods cannot be readily applied to microbes.

Microbial genome-wide association studies (mGWAS), sometimes termed bacterial genome-wide association (BGWA), are still a new area of research with the goal of finding genetic explanations to bacterial phenotypes [6]. The reason why the well-established methods of hGWAS cannot simply be adapted lies in the plasticity of bacterial genomes. In human, the genetic diversity is very low. This is not surprising given that the generation time of humans is around 25 years [7], population bottlenecks occurred only around 70,000 years ago [8] (around 2,800 generations) and founder effects during migration further reduced diversity [9]. In contrast, bacteria commonly have reproduction times measured in minutes and can be much older: the last common ancestor of *E. coli* K-12 and *E. coli* O157:H7 lived about 4.5 million years ago [10]. To illustrate the difference, most of our genes still have chimpanzee orthologs, and only 0.6% of bases [11] in a typical human genome differ from the human reference genome. Meanwhile, the core_{97%} genome of 10,667 *E. coli* genomes represents only 1.96% of the total pangenome [12]. As a result, hGWAS is typically performed by aligning reads to a human reference genome and focuses almost exclusively on single nucleotide polymorphisms (SNPs), amounting to more than 99.9% of human genomic variants [11]. In mGWAS, on the other hand, researchers more often focus on gene-presence-absence, copy-number-variants, unitigs or *k*-mers. Moreover, humans reproduce sexually, and the genome is diploid. Because of recombination, genetic variants that are in proximity have a higher chance of being co-inherited, a phenomenon termed “linkage disequilibrium” that can lead to false positives in GWAS. As bacteria reproduce clonally, the entire genome is in linkage disequilibrium and population structure becomes a strong confounding factor (pseudoreplication) [13, 14]. For this reason, classical dimension-reducing techniques popular in hGWAS, such as multidimensional scaling (MDS), might not sufficiently control false positives. Finally, bacterial genomes are very diverse, with varying numbers of circular or linear DNA molecules, sometimes with plasmids or phages, and recombination and mutation rates that may vary considerably between and even within species. While bacteria do not exchange genetic material through meiosis, recombination of DNA can happen in many species through the processes of transformation, transduction or conjugation [15, 16].

A good overview of existing mGWAS software can be found in San et al. [6]. Among the tools presented, Scoary was the most-cited software (as of February 2023), undoubtedly due to its simplicity and user-friendliness. Scoary scores binary genomic features (i.e., presence/absence of orthogenes, SNPs, unitigs or *k*-mers) for associations to a binary phenotype using Fisher’s test and accounts for population structure using a post-hoc label-switching permutation test. This post-hoc permutation test is based on the pairwise comparisons algorithm [17, 18]. A major advantage of this permutation test is that users do not need to experiment with ill-informed mutation rate parameters or inform the program about population structure [19].

According to San et al. [6], many mGWAS solutions are limited in that they lack data pre-processing functionality as well as post-GWAS methods. Moreover, Scoary was not designed to handle large sample sizes and requires binning for quantitative phenotypes. In addition, the use of

mGWAS was hitherto mostly limited to single phenotypes, usually to pathogenicity and to drug resistance.

In the here presented study, 182 strains belonging to 20 different (sub-)species were selected from the strain collection of Agroscope, the Swiss center of excellence for agricultural research, which comprises over 10,000 isolates of lactic acid bacteria - a valuable legacy from a century of cheese research. Over the past decade, more than 1,300 of these isolates were sequenced and the data collected in the Dialact database. Of these selected strains, 182 yogurts were produced, each by combining one strain from the strain collection with the same starter culture. The metabolomes of these yogurts were measured using liquid chromatography MS (LC-MS) and gas chromatography MS (GC-MS), the latter measuring the yogurt's volatile metabolites (volatiles). The aim of this study was to investigate the effect of the pan-genome of the added bacterial strains on the phenotype of the yogurts.

Here we present Scoary2, a complete re-write and extension of the original Scoary software, developed to efficiently link phenotypic multi-omics data of yogurt to microbial genomes using mGWAS and enable integrative data exploration of yogurt metabolomes. Scoary2 is significantly faster and can thus be applied to more traits as well as isolates. Moreover, the pre-processing (binning) of continuous phenotypes is now integrated and the types of genomic input-data permitted are expanded. Crucial for efficient post-GWAS data exploration of large datasets, Scoary2 includes a simple frontend implemented in HTML/JavaScript that visually and interactively integrates the data as well as optional metadata describing isolates, traits and orthogenes.

Results

The Scoary2 software

Overview

Scoary2 retains all features that are already familiar to users of original Scoary [19]. As in Scoary, the two basic inputs are i) a table that describes the genotypes (orthogenes, SNPs, *k*-mers, unitigs) present in all isolates and ii) a table containing the trait(s) of the isolates. These function as explanatory and response variables, respectively. Optionally, metadata files describing the genotypes, traits, and isolates can be added, greatly facilitating the exploration of the output. Like in original Scoary, the output is a list of significant genes per trait. A manual [20] as well as a tutorial [21] detailing how to use Scoary2 is available on GitHub. Below, we describe the improvements over original Scoary.

Scoary2: performance enhancements

The original Scoary software only had one software dependency (SciPy [22]) and the entire software was implemented using Python-native data structures (i.e., lists and dictionaries) only. In general, Scoary2 uses the efficient NumPy [23] and pandas [24] libraries to load and process the data. Most importantly, the pairwise comparison algorithm was reimplemented, drastically reducing the number of phylogenetic tree traversals. Gene-presence-absence and trait-presence-absence data are now represented as Boolean NumPy arrays, enabling just-in-time compilation of the pairwise comparison algorithm using Numba [25]. The new implementation of this most time-consuming step is around 40x faster than original Scoary. In addition, confidence intervals in the permutation test only depend on the topology of the gene and the number of isolates with the trait. In a dataset with many traits, the same confidence intervals may be used many times. Thus, caching confidence intervals in an SQLite database [26] reduces the number of times this expensive algorithm is executed. The modular software design makes it possible to import the pairwise comparison from the Scoary2 Python module and re-use the algorithm in different programs. Another substantial speed boost comes from enabling true multiprocessing during binarization and analysis of traits using the

producer/consumer software architecture pattern. Also, Scoary2 uses a just-in-time-compiled implementation of Fisher's test (available as a standalone Python library [27, 28]) which is orders of magnitudes faster than the reference implementation in SciPy. Moreover, original Scoary is limited to analyzing datasets with less than 3,000 isolates due to Python's recursion limit. By dynamically adjusting this limit, Scoary2 can now analyze datasets with up to 13,000 isolates.

Using equivalent settings (*permute* = 1000, *correction* I, *p-value-cutoff* = 0.1 / *n-permut* = 1000, *multiple-testing*=native:0.1), Scoary2 is about 63 times faster at analyzing 100 randomly selected traits from the dataset described in this paper (44 isolates, 9051 genes). Scoary2 can analyze a dataset with 44 isolates and 20,000 continuous traits in around 30 minutes on an Intel i7-8565U laptop (4 cores, 8 threads, 1.80-4.60 GHz), something that was not feasible with original Scoary.

Scoary2: software distribution

Scoary2 can be installed using the python package manager (pip) or used through an official docker container, where all dependencies are bundled, guaranteeing easy installation far into the future, thus ensuring reproducibility.

Binning of continuous phenotypes

The core algorithm of Scoary is based on binary genotype and phenotype data. Scoary2 is newly capable of automatically pre-processing continuous phenotypes into binary ones. For this purpose, two Scikit-learn [29] methods, *k*-means and Gaussian mixture model, are available. The former will classify all isolates as having or lacking the trait. The Gaussian mixture model seeks to fit two Gaussian distributions and calculates the probability of each isolate having or not having the trait. By default, isolates that are classified with less than 85% predicted posterior probability are ignored from further analysis. The fitting of Gaussian mixture models can fail, and the user can decide whether to skip such traits or use *k*-means as a backup instead. In the data exploration app, the original continuous values are used again to calculate a histogram.

OrthoFinder support

The name Scoary was chosen in homage to the orthology inference software Roary [30], which transformed bacterial comparative genomics in 2015 thanks to its speed and user-friendliness [31]. However, Roary does not seem to be under active development anymore and was not included in recent *Quest for Orthologs* benchmark studies [32]. Today, OrthoFinder is the most accurate ortholog inference method according to this benchmark [32, 33]. It is under continued development and is among the most used tools in the field. As input, original Scoary uses Roary's *gene-count* table, which indicates how many genes per orthogroup each genome has. However, this makes it cumbersome to find the relevant genes of an interesting orthogroup. While Scoary2 is still compatible with the *gene-count* table, it is highly recommended to use the *gene-list* table, produced by both Roary and OrthoFinder, where cells contain a list of gene identifiers. This way, the gene names will be shown in the data exploration app.

Output and data exploration app

Scoary2 produces similar tables as output as original Scoary. As San et al. [6] indicated, the ability to add annotations to orthogroups would "contribute immensely" to the utility of mGWAS tools. Therefore, Scoary2 does not just allow to add metadata to orthogroups, but also to traits and isolates. In addition, Scoary2 contains a simple data exploration app for easy inspection of the results. It was built using the JavaScript libraries Bootstrap, Papa Parse, Slim Select, DataTables, Plotly and Phylocanvas [34–39] and consists of two pages.

The first page, *overview.html* (Figure 1), shows a dendrogram of all traits that were analyzed. The dendrogram is calculated based on how the traits split the isolates into two sets. The distance metric

used a “symmetric” Jaccard index which ensures that highly correlated and highly anti-correlated traits end up close to each other in the dendrogram. The negative logarithms of the corrected p -value from Fisher's test, the p -value from the permutation test, and the product of the two values are presented next to the dendrogram. These plots, created with SciPy and matplotlib [22, 40], can show at least 20,000 traits. When the mouse pointer hovers over a trait, the associated metadata is presented.

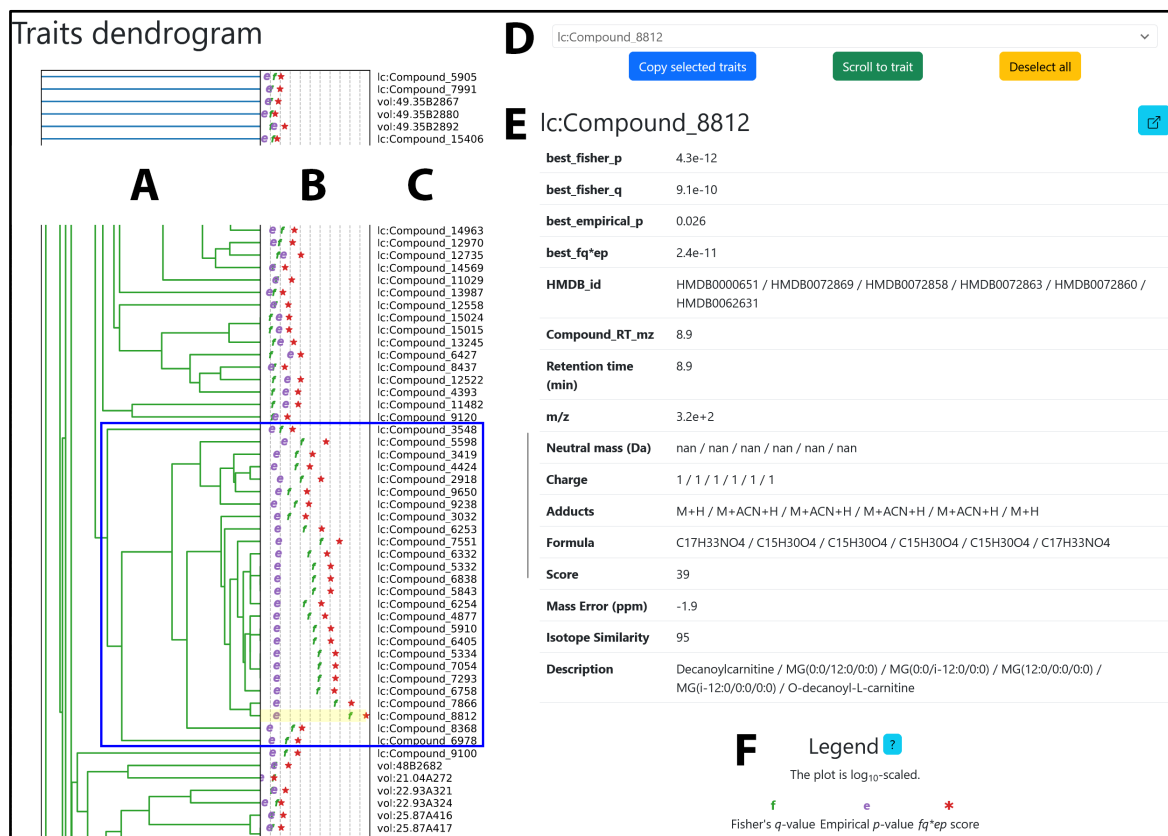


Figure 1: The first page (overview.html) of the Scoary2 data exploration tool. (A): Dendrogram of traits. The blue rectangle surrounds a cluster of carnitine-related genes. (B): Negative logarithms of the p -values calculated by Scoary2: p -values range from high (left) to low (right); f stands for the p -value from Fisher's test, e for the p -value from the post-hoc test and * for the product of the two values. (C): Trait names. (D): Trait search and navigation tool (E): Trait metadata. It is updated when the mouse hovers over the traits in the dendrogram. (F): Plot legend.

The second page, *trait.html* (Figure 2), allows users to further investigate each trait. This page includes a phylogenetic tree of the isolates, where color bars indicate which isolates have the trait and which have a selected orthogroup. In addition, a pie chart shows the fraction of isolates that have the trait and how many of these have the gene. If the trait data is continuous, a histogram is also displayed. These plots are updated whenever the user clicks on an orthogroup. Below the phylogenetic tree, there are two tables. The first displays the Scoary statistics and, if present, metadata for each orthogroup. The second table is a *coverage matrix*, which shows the number of genes each isolate has from each orthogroup. If the isolates have metadata, this information is also displayed in this table. If Scoary2 uses an OrthoFinder-style *gene-list* table as input, clicks on these numbers reveal the gene identifiers. Moreover, the data exploration app can be configured to generate hyperlinks, such that clicks on gene identifiers forward the user to a certain URL, for example one where more information about the gene is available, such as its sequence and annotations. Clicks on orthogroups can also be configured to redirect to custom URLs, for example to enable a comparison of the genes.

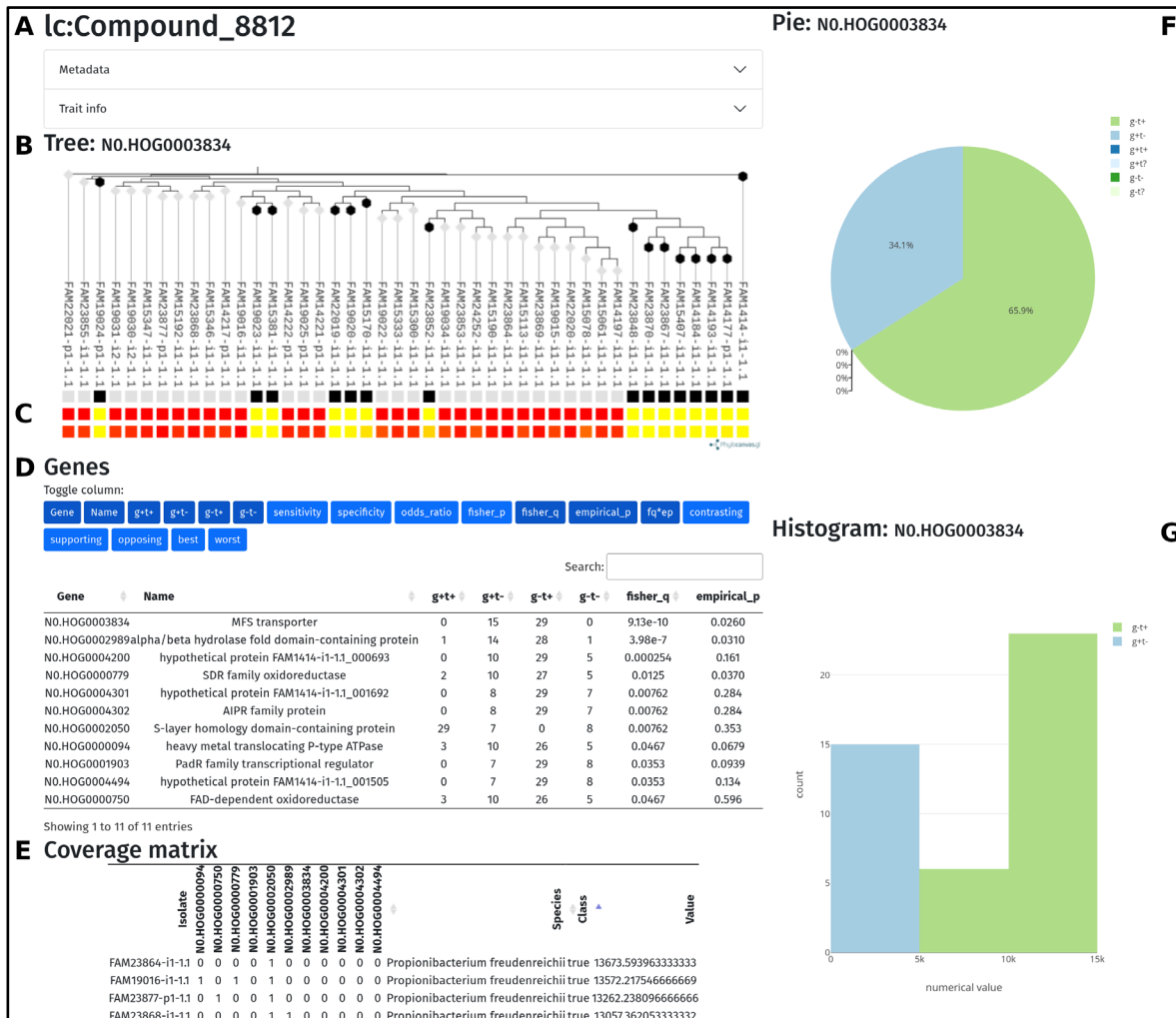


Figure 2: The second page (trait.html) of the Scoary2 data exploration tool. (A): Trait name. (B): Phylogenetic tree of the isolates. (C): Top row: presence (black) / absence (white) of orthogene. Middle row: binarized trait. Bottom row: continuous trait. (D): List of best candidate orthogenes with associated *p*-values. (E): Coverage matrix: The numbers in the cells tell the number of genes in the genome that have the annotation. (F): Pie chart that shows how the orthogene and the trait intersect in the dataset. (G): Histogram of the continuous values, colored by whether each isolate has the orthogene (*g*+/*g*-) and the trait (*t*+/*t*-).

Scoary2 analysis of yogurt dataset

Dataset overview

Figure 3 A/B shows a 2D embedding of the LC-MS and GC-MS volatiles datasets that was generated using uniform manifold approximation and projection (UMAP) [41]. Notably, yogurts made with closely related strains tend to cluster together. Both datasets show one cluster dominated by yogurts made with strains from the order *Propionibacteriaceae*, and another dominated by *Lactobacillales*. Interestingly, the control yogurts which contain only the starter strains cluster with the former in the LC-MS dataset, but with the latter in the GC-MS volatiles dataset.

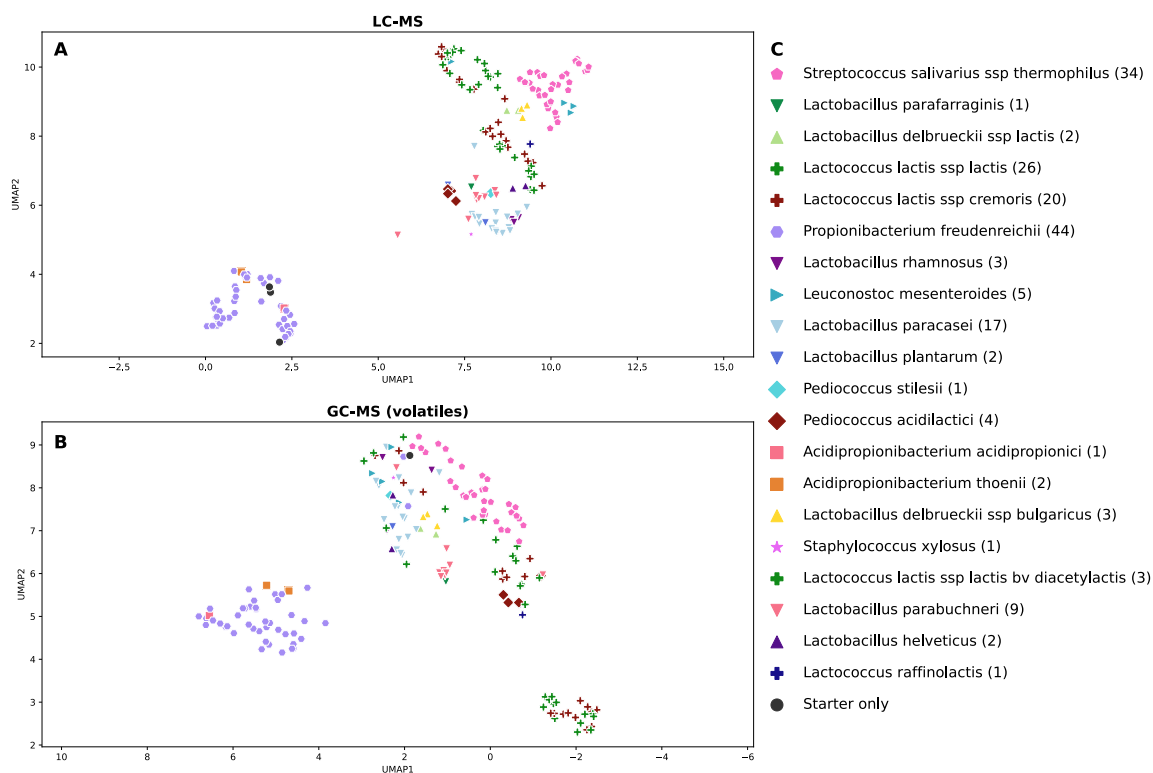


Figure 3: UMAP projections of mass spectrometry datasets. Each symbol represents one yogurt that was made with a different bacterial strain in addition to the starter culture YC-381. (A): LC-MS dataset: 2,348 metabolites. (B): GC-MS volatiles dataset: 1,541 metabolites. (C): Legend: each (sub)species has a unique combination of color and symbol. The number in brackets indicates the number of yogurts made using the respective (sub-)species.

Scoary2 results

With the parameters $n_cpus = 8$, $multiple_testing = bonferroni:0.1$, $n_permut = 1000$, $worst_cutoff = 0.1$, $max_genes = 50$, Scoary2 took 22 minutes to analyze the full dataset (3,889 traits, 182 isolates, 10,358 hierarchical orthogroups). As the analysis of this full dataset would go beyond the scope of this publication, as proof of concept, we show results that can be replicated by restricting the dataset to the *Propionibacterium freudenreichii* isolates. Scoary2 took only one minute to process this reduced dataset (3,889 traits, 44 isolates, 1,466 hierarchical orthogroups), with the parameters $n_cpus = 8$, $n_permut = 1000$. In comparison, original Scoary took 7.65 min to process ten traits with analogous parameters ($p_value_cutoff = 0.1$, $permute = 1000$), or approximately 50 h for the entire dataset.

The output consists of 1,256 metabolites (those with Fisher's test q -values > 0.999 are automatically removed). One cluster of metabolites, highlighted with a blue rectangle in Figure 1, had particularly high p -values (e.g., *Compound_8812*: q -value Fisher's test: 9.1×10^{-10} , p -value from post-hoc test: 0.026). Because similar metabolites are clustered together (including the anti-correlated metabolites like *Compound_3032*) and each metabolite's metadata is available in *overview.html*, we noticed straightforwardly that the MS database found hits with molecules with carnitine in their names for 14 out of 21 of the metabolites in this cluster (Figure 4 D). Looking at the results in more detail using *trait.html*, shown in Figure 2, we found that two genes correlate strongly with these metabolites: an MFS transporter and an α/β -hydrolase fold domain-containing protein. A closer look at the gene identifiers suggests that the two genes are adjacent. Furthermore, the gene loci (Figure 4 E/F) were compared using OpenGenomeBrowser [42] via custom URLs as mentioned earlier, revealing that the two genes are indeed adjacent and located in a syntenic gene cluster, one gene away from an L-carnitine CoA transferase (*caiA*). In the isolates which lack the two genes, many of the clusters were seemingly disjoined by transposases and other genes on the cluster were pseudogenized (Figure 4 E/F). Interestingly, the cluster includes the *caiABC* and *fixABCX* genes, which are associated with the anaerobic metabolism of carnitine [43]. To discover these patterns,

summarized in Figure 4, all these pieces of information needed to be integrated, highlighting the importance and convenience of the data exploration app.



Figure 4: Abundance of the metabolites that correlate with the putative carnitine transporter and corresponding gene loci of three yogurts made from starter cultures only and 44 yogurts made with additional *Propionibacterium freudenreichii* isolates. (A): Heat map of the scaled metabolite abundances. Scale: blue (low) to average (white) to red (high). (B): Scale factor of each metabolite. (C): Sum of absolute intensities. (D): Color bar that indicates i) whether the mass spectrometry database suggested a match with carnitine in the name (green) or not (grey), ii) whether the suggestion could be confirmed (green) or not (red). (E): Comparison of the associated gene cluster spanning from the MFS transporter (red) to *fixX* (dark blue). (F): Annotations of the orthogenes. Orthologs are highlighted in the same color. The putative carnitine transporter highlighted in red, the *caiABC* genes in shades of green and the *fixABCX* genes in shades of blue.

Confirmation of identities for carnitine compounds

The identities of five metabolites (decanoylcarnitine, octanoylcarnitine, hexanoylcarnitine, carnitine and acetylcarnitine), assigned to the cluster detected by Scoary2 (Figure 4 D) were subsequently confirmed by LC-MS analysis of pure analytical standard solutions (Table 1).

Table 1: List of MFS-transporter-associated metabolites that were confirmed by standard injection.

Metabolite	Measured m/z	Database match	CAS n°	Mass error [ppm]	Retention time error [%]
Ic:Compound_8812	316.24764	Decanoylcarnitine	3992-45-8	< 1	0.67
Ic:Compound_7866	288.21635	Octanoylcarnitine	25243-95-2	2.02	0.47
Ic:Compound_6838	260.18515	Hexanoylcarnitine	22671-29-0	< 1	2.13
Ic:Compound_3548	162.11237	Carnitine	541-15-1	< 1	8.69
Ic:Compound_4877	204.12298	Acetylcarnitine	3040-38-8	< 1	9.66

Compared to yogurt made from starter cultures only, we found that two thirds of the *Propionibacterium freudenreichii* isolates did not strongly affect the composition of the carnitine-related metabolites shown in Figure 4. These yogurts are characterized by high amounts of certain acylcarnitines. In contrast, the presence in isolates of the two genes identified by Scoary2 (MFS transporter and α/β -hydrolase fold domain-containing protein), did influence the abundance of those acylcarnitines. Yogurts prepared using such isolates contain depleted amounts of acylcarnitines, particularly octanoyl- and decanoylcarnitine, and are characterized by higher amounts of carnitine, γ -butyrobetaine (putative), and certain other (putative) acylcarnitines.

Discussion

Translation of bacterial genomes into the metabolome of fermented foods

Although the yogurts produced in this study are multi-strain mixtures, the metabolomes of the yogurts show a clear correlation between the genetic relatedness of the added strains and the metabolomic profile of the yogurts produced. This is illustrated by the fact that the metabolomes of yogurts produced with closely related strains tend to cluster together in both MS datasets (Figure 3). Genomic differences were thus successfully translated into the metabolome of the yogurts, even though the standard manufacturing procedure does not correspond to the preferred growth conditions, such as temperature or growth time, of each species and strain. Hence, the inclusion of strains and species that differ at the genetic level in the phenotypic screening, even when done under standardized conditions, is a promising strategy to influence the composition of fermented foods. However, this taxonomy-based clustering of the metabolomic data poses a major problem when trying to find causal connections between orthogenes and metabolites, as the strongest correlations in the dataset are between the many metabolites and orthogenes that also strongly correlate with the population structure. Though these orthogenes may be good predictors of metabolism, most are not causally related to metabolites. In order to avoid spurious associations in this scenario, and to pinpoint real causal relationships, mGWAS methods such as Scoary's pairwise comparisons are essential.

Relevance of Carnitines in Yogurts

Carnitine and bacteria

L-carnitine is a ubiquitous quaternary amine compound that can be found in all kingdoms of life and is ubiquitous in many environments [44]. In bacteria, carnitine can serve multiple roles in the core metabolism, including as terminal electron acceptor and as carbon, nitrogen and energy source. Moreover, carnitine is a compatible solute and osmolyte, and is used by some bacteria as osmo- and cryoprotectant or to increase bile tolerance. Accordingly, variations in bacterial carnitine metabolism

may have implications for yogurt or probiotics regarding refrigeration and gastrointestinal transit. Although many bacteria are known to be able to metabolize L-carnitine in different ways, to date, only *Sinorhizobium meliloti* is known to be able to synthesize it *de novo* [45]. In contrast, most bacteria import this molecule or direct precursors from the environment [44, 46]. Thus, transporters may be key to carnitine metabolism in bacteria. For example, the *caiT* carnitine/ γ -butyrobetaine antiporter is known to be involved in anaerobic carnitine metabolism, as it imports carnitine and exports the fermentation product and is not involved in osmotic stress response. On the other hand, if a transporter is one-way only, the purpose may more likely be osmoregulation or a different metabolic route [47, 48].

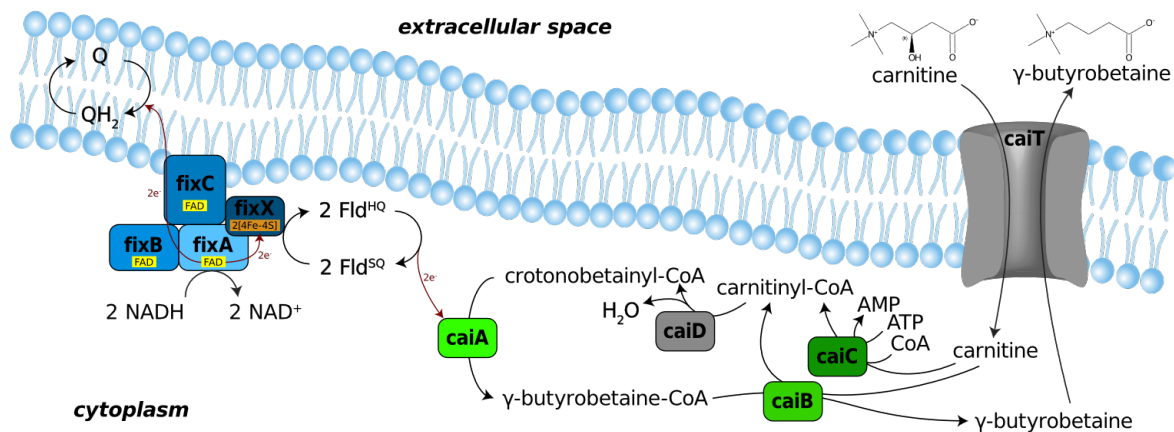


Figure 5: Anaerobic Carnitine Reduction in *Escherichia coli*, adapted from Walt 2002 [49], Bernal 2008 [50] and Ledbetter 2017 [51]. Proteins that were not found in any of the tested *Propionibacterium freudenreichii* strains are colored grey. *caiT*: antiport of carnitine and γ -butyrobetaine. *caiC*: generates initial carnitinylyl-CoA. *caiD*: dehydration of carnitinylyl-CoA to crotonobetainyl-CoA. *caiA*: reduction of crotonobetainyl-CoA to γ -butyrobetainyl-CoA using electrons from *fixABCX*. *caiB*: recycles CoA moiety. *fixABCX*: Oxidation of NADH is coupled to reduction of ubiquinone (Q) and flavodoxin semiquinone (Fld^{SQ}), which then delivers electrons to a terminal electron acceptor.

Legend: FAD: flavin adenine dinucleotide; [4Fe-4S]: iron-sulfur cluster; Fld^{SQ} : flavodoxin, semiquinone form; Fld^{HQ} : flavodoxin, hydroquinone form; Q: ubiquinone; QH_2 : ubiquinol

caiABC, *fixABCX*, and the potential function of the identified genes

Homologs of *fixABCX* were originally characterized in *Rhizobium meliloti* where they function as a respiratory chain, providing electrons for nitrogen fixation [52]. The genes *caiABC* were first identified as part of the *E. coli* *caiTABCDE* operon, which is close to and co-expressed with the *fixABCX* operon and together ferment carnitine to γ -butyrobetaine in anaerobic conditions and absence of preferred substrates [43, 49, 53]. Briefly, *caiABCD* converts carnitine to γ -butyrobetaine, aided by the respiratory chain *fixABC* which provides two electrons in the reduction step (Figure 5).

However, the selected *Propionibacterium freudenreichii* isolates are lacking homologs of the crotonobetainyl-CoA hydratase *caiD* and the carnitine/ γ -butyrobetaine antiporter *caiT*. Instead, between *caiABC* and *fixABCX*, we find an MFS transporter and an enoyl-CoA hydratase (Figure 4), which might fill these gaps in the pathway. On the other hand, the two genes identified by Scoary2 are also an MFS transporter and a hydrolase, and since only the strains with these genes have a strong impact on the carnitine composition of the yogurt (Figure 4), it appears that the full operon is required to permit efficient import of precursors and fermentation of carnitine in *P. freudenreichii*. This is supported by the apparent degradation of the gene cluster through transposases and pseudogenization in many genomes where the two genes were lost.

One piece of evidence suggests that the MFS transporter identified by Scoary2 is an acylcarnitine importer: In the *P. freudenreichii* strain FAM23848, a transposase has split *fixABC* from the rest of the

operon. This appears to result in continued acylcarnitine import and deacylation, but inhibited reduction of carnitine to γ -butyrobetaine, leading to significant accumulation of carnitine (Figure 4).

If our interpretation is correct and the *P. freudenreichii* strains with the complete operon can indeed use carnitine as terminal electron acceptor, it could enable them to persist better in the human gut, where such redox reactions are a key ecological pressure [54].

Carnitine and humans

Even though humans can synthesize L-carnitine endogenously from the essential amino acids L-methionine and L-lysine, 75% is obtained through diet [55], which is why it has been termed a “conditionally essential nutrient”. One of the richest sources of carnitine is red meat, but bovine milk (about 169 $\mu\text{mol/L}$ [56]) and milk products also contain carnitine [57]. Humans require L-carnitine to transport long- to short-chain fatty acids in and out of the mitochondrion [44]. This “carnitine shuttle” is rate-limiting for fatty acid oxidation (FAO) [58].

There are two known pathways that link carnitine to human diseases via microbiota metabolism: First, two isobaric ($m/z = 160.133$) microbe-produced structural analogs of L-carnitine (3-methyl-4-(trimethylammonio)butanoate and 4-(trimethylammonio)pentanoate) have recently been shown to hamper FAO in the mural brain by inhibiting the carnitine shuttle. Molecules with matching m/z have been linked to type 2 diabetes, preeclampsia, and nephropathy in type 1 diabetes. This is plausible because these diseases are associated with mitochondrial dysfunction or incomplete FAO [59]. However, no molecules with matching m/z were found in the dataset presented in this paper. Second, certain gut microbes, for example *Acinetobacter*, metabolize dietary L-carnitine to trimethylamine (TMA), which is oxidized in the liver to trimethylamine-*N*-oxide (TMAO). These metabolites are well-known risk factors of cardiovascular disease, though controversy exists as to which of these metabolites is the real culprit [60–62].

Either way, the metabolism of carnitine by the microbiota is of growing scientific interest and may be influenced by the metabolites in yogurt or the large amounts of bacteria it contains. This is indicated by two independent experiments by Burton et al. [63], where ingestion of a probiotic yogurt resulted in a lower postprandial TMAO response in urine and plasma, compared to non-fermented milk. In addition, recently discovered enzymes from the MttB superfamily of the gut bacterium *Eubacterium limosum* were found to demethylate L-carnitine and other TMA precursors and may deplete TMA/TMAO levels through precursor competition [64, 65]. In addition to impacting the postprandial TMAO response, Burton et al. also found that the ingestion of probiotic yogurt resulted in a different production of several bile acids [66], indicating that dietary fat metabolism in humans can be modulated through fermented foods via pathways involving carnitine and bile salts [63, 66].

Potential use in microbial specialized metabolites discovery

To the best of our knowledge, Scoary2 is the first software that makes the study of large phenotypic multi-omics datasets using mGWAS feasible. This approach may constitute a novel discovery strategy for microbial metabolites, thereby providing the potential to accelerate progress in microbiology, drug discovery, and targeted production of functional fermented food to support human health [67, 68]. After all, as outlined in van der Hoof et al. [69], traditional methods are based on established knowledge and labor-intensive experiments, such as activity-guided fractionation of metabolite extracts. These were complemented by genome and metabolome mining approaches. More recently, a “metabologenomic integration” approach was developed that combines high throughput metabolomics with genomics [69]. However, this approach does not take population structure into account and is limited to biosynthetic gene clusters (BGCs), which are challenging to predict, and depend on high quality genome sequences as well as existing knowledge [70–73]. Scoary2 on the other hand, is conceptually simpler and therefore applicable to a wider range of data, in addition to being easier to use. It is fast enough to process entire metabolomes, cannot just take

BGCs but all orthogenes into account, is aware of population structure and does not rely on existing knowledge and thus represents a valid alternative in that context.

Post-GWAS methods enable analysis of large phenotypic datasets

We strongly agree with San et al. on the immense utility of post-GWAS methods [6]. To our knowledge, Scoary2's post-GWAS data exploration app stands out amongst other mGWAS tools, being able to integrate (i) the detected associations between traits and genes, (ii) relations between traits, (iii) relations between isolates and (iv) metadata describing traits, genes and isolates.

These innovations are very convenient for small datasets, but an absolute necessity for datasets with many traits. The dendrogram of traits in *overview.html* helps discover groups of (anti-)correlated traits, and the *p*-values plots help to prioritize them. The presence of trait metadata enabled us to notice quickly that many metabolites of one cluster were annotated as carnitines. Navigating to *traits.html* with only one click allows us to see the phylogeny of the isolates as well as the distribution of the selected trait and the highest-scoring orthogene. The orthogene annotations may also be insightful here. The gene IDs in the *coverage matrix* may reveal that certain orthogenes are often close to each other on the genome, indicating an operon. If the trait is numeric, the histogram may be useful to gauge how strongly the trait varies in the dataset and whether the data points contradicting the hypothesis might just have been incorrectly classified during binarization. If the app is connected to external comparative genomics tools, it becomes easy to study the candidate gene in more detail. In our example, OpenGenomeBrowser [42] enabled us to discover that the two genes most strongly associated with carnitines are located on the same gene cluster and near an *L-carnitine CoA transferase*, providing more evidence for a causal relationship.

Because the output of most mGWAS tools is structurally similar, i.e., consisting of coefficients for genes and traits, this app offers the possibility to be adapted to other tools or even to develop standardized output formats. Thus, a single generic data exploration app could be developed and used by many mGWAS tools.

Comparison with existing mGWAS approaches

The field of mGWAS software is very diverse. Various conceptually different approaches have been developed and refined, and as a result, different tools require different input types and yield conceptually different outputs. The main result from LASSO and Random Forest is the model's predictive performance. The model itself may be harder to understand, as LASSO may randomly choose one of multiple highly correlated genes and drop the others, and Random Forest does not yield correlation coefficients for the genes at all. Linear mixed models yield a straightforward *p*-value for each gene but are based on hard-to-verify assumptions about bacterial evolution. Homoplasmy based methods like treeWAS [74] and Scoary give multiple *p*-values for different types of association scenarios, arguably requiring more careful interpretation. Consequently, tools based on different approaches are difficult to compare. Moreover, benchmarks are often carried out based on simulated datasets, and it is difficult to tell how closely they imitate bacterial evolution and real datasets. We noticed that Scoary and treeWAS were evaluated using simulations that emphasized the evolutionary scenarios they were designed to detect [19, 74], while the simulations from Saber et al. [14], benchmarking linear-model-based tools, did not investigate the effect of homoplastic mutations. We recommend that future research should compare the various approaches using realistic simulations and real datasets and flesh out guidelines as to which approach and tool is recommended in which scenario.

Limitations of the Scoary2 algorithm

Fisher's test

Fisher's test is a simple and fast test that measures how strongly a gene and a trait correlate. To determine a *causal* relationship in mGWAS, however, its assumptions are violated, and the resulting *p*-values should rather be interpreted as scores. For users who simply want to learn which traits are associated with specific clusters in a tree without any assumptions on causal relation, Fisher's test is nonetheless useful.

Pairwise comparisons

To be as generalizable and widely applicable as possible, the pairwise comparisons algorithm is devoid of any explicitly defined models of evolution and sacrifices some statistical power. For example, a gene whose presence is one hundred percent correlated with a particular phenotype might not be considered significant if the variant-phenotype combination is clustered on a single branch, in other words, if it can be traced back to a single event in the phylogenetic history of the input data. However, we prefer the pairwise comparisons algorithm to explicitly defined models because in our opinion, the mutation rates at every branch in the tree are most often unknown or unavailable. Thus, in Scoary2, only the branching pattern of the phylogenetic tree matters. This means that any errors in its topology could confound results.

A clear downside to the pairwise comparisons algorithm is that it can only deal with binary phenotypic events and not continuous or Brownian motion-type transitions. In Scoary2, phenotypes measured on a continuous scale are automatically binarized with either k-means or a Gaussian Mixture Model. For the former, there is a risk of improper phenotypic classification, and the latter discards values that do not clearly fit either of the gaussian means, leading to a reduced dataset to draw conclusions on. The latter issue is partially mitigated by manual inspection of the numerical values in *traits.html*.

Future directions

In the future, tests that can better exploit numerical data, can detect several types of evolutionary scenarios, or have higher statistical power could be added to Scoary2. Possible candidates are the three tests from treeWAS [74], though there is still room for the development of new algorithms [75].

Conclusions

We expanded Scoary's applicability to datasets containing tens of thousands of traits by significantly increasing the performance of the algorithm. Moreover, we added a novel interactive data exploration tool that combines trait, genotype, and isolate metadata, greatly facilitating the interpretation of results and crucial for timely exploration of large datasets. We illustrated Scoary2's capabilities by applying the software to a large MS dataset of 44 yogurts made from different strains of *Propionibacterium freudenreichii*, allowing us to identify novel genes involved in carnitine metabolism. Scoary2 is, to the best of our knowledge, the first software that makes it feasible to study large phenotypic multi-omics datasets using mGWAS. It enables and facilitates the discovery of previously unknown bacterial genotype-phenotype associations and can thus help overcome a major bottleneck in microbial research, namely the unknown role of many genes and their impact on the phenotype. Therefore, it may significantly contribute to fermented food research, accelerating and facilitating the development of fermented food products with specific properties. In addition, Scoary2 has the potential for broader application, for example in basic microbial research, drug discovery and clinical research, and could thus considerably impact microbiological science in the future.

Methods

Yogurt production

Lactose-free, homogenized, pasteurized, semi-skimmed (1.5%) milk purchased from a local retailer was used for yogurt production (Aha! IP Suisse, Migros, Switzerland). Fermentation was carried out overnight (16 hours) at 37 °C using the yogurt culture Yoflex® YC-381 (Chr. Hansen, Denmark) containing *Lactobacillus delbrueckii* subsp. *bulgaricus* and *Streptococcus thermophilus*, as well as one of the selected strains from the Liebefeld culture collection (Table S1). The yogurts were stored at -20 °C until analysis.

GC-MS (volatiles) dataset

Untargeted volatile analysis was carried out using an Agilent 7890B gas chromatography (GC) system coupled with an Agilent 5977B mass selective detector (MSD) (Agilent Technology, Santa Clara, CA, USA). For volatile analysis, 250 mg of yogurt containing 25 µl ISTD (Paraldehyde 0.5 ppm, Tetradecane 0.25 ppm and D4-Decalactone 0.5 ppm) diluted in water were placed in 20 mL HS vials (Macherey-Nagel), hermetically sealed (blue silicone/Teflon septum (Macherey-Nagel)) and measured in a randomized order. After incubation of the samples for 10 min at 60 °C, the headspace was extracted for 5 min at 60°C under vacuum (5 mbar) as described by Fuchsmann et al., 2019 [76], using the Vacuum transfer in trap extraction method. The trap used was a Tenax TA (2/3 bottom) / Carbosieve S III (1/3 top) (BGB analytics). The temperature of the trap was fixed at 35°C and the temperature of the syringe at 100°C. The sorbent and syringe were dried for 20 min under a nitrogen stream of 220-250 mL min⁻¹. Desorption of the volatiles took place for 2 min at 300°C under a nitrogen flow of 100 mL min⁻¹. For this purpose, the programmable temperature vaporization injector (PTV) was cooled at 10°C for 2 min, heated up to 250°C at a rate of 12°C sec⁻¹ and held for 20 minutes in solvent vent mode. After 2 min the purge flow to split vent was set to 100 mL min⁻¹. The separation was carried out on a polar column OPTIMA FFAPplus fused silica capillary column 60m x 0.25mm x 0.5µm (Macherey-Nagel) with helium as the carrier gas at a flowrate of 1.5 mL min⁻¹ (25.3 cm sec⁻¹). The oven temperature was held for 5 min at 40°C, followed by heating up to 240°C at a rate of 5°C min⁻¹ with a final holding time of 55 min. The trap was reconditioned after injection at a nitrogen flow of 100 mL min⁻¹ for 15 min at 300°C. The spectra were recorded in SCAN mode at a mass range between m/z 30 to m/z 350 with a gain at 10 with a solvent delay of 4 minutes. The samples were measured twice in random order. Only compounds that were detected in > 50% of QCs were retained (1,541 metabolites).

LC-MS dataset

Untargeted metabolomic analysis was performed using an UltiMate 3000 HPLC system (Thermo Fisher Scientific) coupled to maXis 4G+ quadrupole time-of-flight mass spectrometer (MS) with electrospray interface (Bruker Daltonik GmbH). Chromatographic separation was conducted on a C18 hybrid silica column (Acquity UPLC HSS T3 1.8 µm 2.1 x 150 mm, Waters, UK), reversed phase at a flow rate of 0.4 mL min⁻¹. The mobile phase consisted in ultrafiltered water (Milli-Q® IQ 7000, Merck, Germany) containing 0.1% formic acid (Fluka™, Honeywell, USA) (A), and acetonitrile (Supelco®, Merck, Germany) with 0.1% formic acid (B), with the following elution gradient (A:B): 95:5 at 0 min to 5:95 at 10 min; 5:95 from 10 to 20 min; 95:5 from 20 to 30 min. The spectra were recorded from m/z 75 to m/z 1500 in positive ion mode. Detailed MS settings were as follows: collision-induced dissociation: 20 to 70 eV, electrospray voltage: 4.5 kV, endplate offset: 500 V, capillary voltage: 3400 V, nitrogen flow: 4 mL min⁻¹ at 200°C, spectra acquisition rate: 1 Hz in profile mode, resolution: 80,000 FWHM. The yogurt samples were measured in triplicates in random order and the values averaged afterwards.

The QC-based robust locally estimated scatterplot smoothing signal correction method was applied for signal drift correction [77] using R (v.3.1.2) [78]. Metabolites with poor repeatability, i.e., detected in < 50% of QCs, were removed, as well as metabolites with a relative standard deviation > 30% in the QC samples. Features that had a median in the QC samples that was < 3 times higher than the median calculated for the blanks were also excluded. This reduced the number of metabolites from 17,310 to 2,348.

Identification of carnitines

The Human Metabolome Database (27) was used with a 5-ppm mass accuracy threshold for the identification of a selection of metabolites. Identity suggestions from databases were then confirmed by MS fragmentation data (when available) and with the injection of pure standards solutions. All standards were purchased at Sigma-Aldrich (Sigma-Aldrich Chemie GmbH, Buchs, Switzerland).

OrthoFinder

Hierarchical orthogroups were called using OrthoFinder [33] version 2.5.4 with default parameters.

Locus plots

The gene locus plots (Figure 4) were generated using OpenGenomeBrowser [42], which utilizes DNA Features Viewer [79], and modified using Adobe Illustrator [80].

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

Software

The Scoary2 source code is publicly available at <https://github.com/MrTomRod/scoary-2/> under the MIT License [81]. An official docker container ([troder/scoary-2](https://hub.docker.com/r/troder/scoary-2/)) is available on Docker Hub [82].

Datasets

The genomes of the 44 *Propionibacterium freudenreichii* used in this study were uploaded to NCBI GenBank and available under BioProject PRJNA946676. This data is also available via the OpenGenomeBrowser demo instance hosted at <https://opengenomebrowser.bioinformatics.unibe.ch/> [83]. The combined LC-MS and GC-MS datasets and the hierarchical orthologs file generated by OrthoFinder are available in the Mendeley Data repository (doi: [10.17632/yytybr3t4y.1](https://doi.org/10.17632/yytybr3t4y.1)) under the CC BY 4.0 license [84].

Competing interests

The authors declare that they have no competing interests.

Funding

This research was funded by Gebert R uf Stiftung within the program "Microbials", grant number GRS-070/17 and the Canton of Bern.

Authors' contributions

GV, RB, CB, UvA, GP and TR conceived the project. GV and RB provided funding to the project. UvA and GP produced the yogurts. GP measured the LC-MS yogurt metabolome. PF and MTS measured the GC-MS volatiles yogurt metabolomes. GP identified the carnitine metabolites. SP and TR explored different ways of analyzing the data. TR programmed the Scoary2 software with advice from OB. TR, CB and GP analyzed the data. TR, CB and OB wrote the manuscript with input from all authors. TR created the figures. All authors read and approved the final manuscript.

Acknowledgments

We thank Arthur Volant for adding Boschloo's exact test to scipy [22] at our request, even though we ended up not using it. Furthermore, we thank Zahra Sattari for her contribution to the production of the initial yogurts.

References

1. Somerville V, Grigaitis P, Battjes J, Moro F, Teusink B. Use and limitations of genome-scale metabolic models in food microbiology. *Current Opinion in Food Science*. 2022;43:225–31.
2. Vanni C, Schechter MS, Acinas SG, Barberán A, Buttigieg PL, Casamayor EO, et al. Unifying the known and unknown microbial coding sequence space. *eLife*. 2022;11.
3. Zeki ÖC, Eylem CC, Reçber T, Kır S, Nemitlu E. Integration of GC-MS and LC-MS for untargeted metabolomics profiling. *J Pharm Biomed Anal*. 2020;190:113509.
4. Subramanian I, Verma S, Kumar S, Jere A, Anamika K. Multi-omics Data Integration, Interpretation, and Its Application. *Bioinform Biol Insights*. 2020;14:1177932219899051.
5. Akiyama M. Multi-omics study for interpretation of genome-wide association study. *J Hum Genet*. 2021;66:3–10.
6. San JE, Baichoo S, Kanzi A, Moosa Y, Lessells R, Fonseca V, et al. Current Affairs of Microbial Genome-Wide Association Studies: Approaches, Bottlenecks and Analytical Pitfalls. *Front Microbiol*. 2019;10:3119.
7. Nachman MW, Crowell SL. Estimate of the mutation rate per nucleotide in humans. *Genetics*. 2000;156:297–304.
8. Gibbons A. Human ancestors were an endangered species. *ScienceNow*. 2010.
9. Manica A, Amos W, Balloux F, Hanihara T. The effect of ancient population bottlenecks on human phenotypic variation. *Nature*. 2007;448:346–8.
10. Reid SD, Herbelin CJ, Bumbaugh AC, Selander RK, Whittam TS. Parallel evolution of virulence in pathogenic *Escherichia coli*. *Nature*. 2000;406:64–7.
11. 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. *Nature*. 2015;526:68–74.
12. Abram K, Udaondo Z, Bleker C, Wanchai V, Wassenaar TM, Robeson MS, et al. Mash-based analyses of *Escherichia coli* genomes reveal 14 distinct phylogroups. *Commun Biol*. 2021;4:117.
13. Earle SG, Wu C-H, Charlesworth J, Stoesser N, Gordon NC, Walker TM, et al. Identifying lineage effects when controlling for population structure improves power in bacterial association studies. *Nat Microbiol*. 2016;1:16041.
14. Saber MM, Shapiro BJ. Benchmarking bacterial genome-wide association study methods using simulated genomes and phenotypes. *Microb Genom*. 2020;6.
15. Epstein B, Abou-Shanab RAI, Shamseldin A, Taylor MR, Guhlin J, Burghardt LT, et al. Genome-Wide Association Analyses in the Model *Rhizobium Ensifer meliloti*. *mSphere*. 2018;3.
16. Hanage WP. Not so simple after all: bacteria, their population genetics, and recombination. *Cold Spring Harb Perspect Biol*. 2016;8.
17. Read AF, Nee S. Inference from binary comparative data. *J Theor Biol*. 1995;173:99–108.
18. Maddison WP. Testing character correlation using pairwise comparisons on a phylogeny. *J Theor Biol*. 2000;202:195–204.
19. Brynildsrud O, Bohlin J, Scheffer L, Eldholm V. Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. *Genome Biol*. 2016;17:238.

20. Roder T. Usage · MrTomRod/scoary-2 Wiki. Scoary2 Usage. 2022. <https://github.com/MrTomRod/scoary-2/wiki/Usage>. Accessed 16 Mar 2023.
21. Roder T. Tutorial · MrTomRod/scoary-2 Wiki. Scoary2 Tutorial. 2022. <https://github.com/MrTomRod/scoary-2/wiki/Tutorial>. Accessed 16 Mar 2023.
22. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods*. 2020;17:261–72.
23. Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, et al. Array programming with NumPy. *Nature*. 2020;585:357–62.
24. The pandas development team. pandas-dev/pandas: Pandas 1.0.3. Zenodo. 2020.
25. Lam SK, Pitrou A, Seibert S. Numba: A LLVM-based Python JIT compiler. In: *Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC - LLVM '15*. New York, New York, USA: ACM Press; 2015. p. 1–6.
26. Allen G, Owens M. *The definitive guide to sqlite*. Berkeley, CA: Apress; 2010.
27. Roder T. GitHub - MrTomRod/fast-fisher: A fast, precise, pure Python implementation of Fisher's exact test. <https://github.com/MrTomRod/fast-fisher>. Accessed 30 May 2022.
28. painyeph. painyeph/FishersExactTest: A fast, precise, pure Python implementation of Fisher's exact test. <https://github.com/painyeph/FishersExactTest>. Accessed 30 May 2022.
29. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*. 2011;12:2825–30.
30. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, et al. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics*. 2015;31:3691–3.
31. Seemann T. Torsten Seemann tweets: “[Roary] transformed bacterial species pan genome analysis.” Twitter. 2018. <https://twitter.com/torstenseemann/status/1061079556356923394>. Accessed 27 May 2022.
32. Nevers Y, Jones TEM, Jyothi D, Yates B, Ferret M, Portell-Silva L, et al. The Quest for Orthologs orthology benchmark service in 2022. *Nucleic Acids Res*. 2022.
33. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol*. 2019;20:238.
34. opencollective.com/bootstrap. Bootstrap · The most popular HTML, CSS, and JS library in the world. <https://getbootstrap.com/>. Accessed 30 May 2022.
35. [papaparse.com](https://www.papaparse.com/). Papa Parse - Powerful CSV Parser for JavaScript. <https://www.papaparse.com/>. Accessed 30 May 2022.
36. slimselectjs.com. Slim Select. <https://slimselectjs.com/>. Accessed 30 May 2022.
37. datatables.net. DataTables | Table plug-in for jQuery. <https://datatables.net/>. Accessed 30 May 2022.
38. Plotly Technologies Inc. Plotly - Collaborative data science. 2015.
39. Centre for Genomic Pathogen Surveillance. PhyloCanvas.gi: Interactive tree visualisation for the web. <https://www.phylocanvas.gi/>. Accessed 30 May 2022.

40. Hunter JD. Matplotlib: A 2D Graphics Environment. *Comput Sci Eng.* 2007;9:90–5.
41. Sainburg T, McInnes L, Gentner TQ. Parametric UMAP embeddings for representation and semisupervised learning. *Neural Comput.* 2021;33:2881–907.
42. Roder T, Oberhänsli S, Shani N, Bruggmann R. OpenGenomeBrowser: a versatile, dataset-independent and scalable web platform for genome data management and comparative genomics. *BMC Genomics.* 2022;23:855.
43. Eichler K, Buchet A, Bourgis F, Kleber H-P, Mandrand-Berthelot M-A. The fix Escherichia coli region contains four genes related to carnitine metabolism. 1995.
44. Meadows JA, Wargo MJ. Carnitine in bacterial physiology and metabolism. *Microbiology (Reading, Engl).* 2015;161:1161–74.
45. Bazire P, Perchat N, Darii E, Lechaplais C, Salanoubat M, Perret A. Characterization of l-Carnitine Metabolism in *Sinorhizobium meliloti*. *J Bacteriol.* 2019;201.
46. Ghonimy A, Zhang DM, Farouk MH, Wang Q. The impact of carnitine on dietary fiber and gut bacteria metabolism and their mutual interaction in monogastrics. *Int J Mol Sci.* 2018;19.
47. Jung H, Buchholz M, Clausen J, Nietschke M, Revermann A, Schmid R, et al. CaiT of *Escherichia coli*, a new transporter catalyzing L-carnitine/gamma -butyrobetaine exchange. *J Biol Chem.* 2002;277:39251–8.
48. Ziegler C, Bremer E, Krämer R. The BCCT family of carriers: from physiology to crystal structure. *Mol Microbiol.* 2010;78:13–34.
49. Walt A, Kahn ML. The fixA and fixB genes are necessary for anaerobic carnitine reduction in *Escherichia coli*. *J Bacteriol.* 2002;184:4044–7.
50. Bernal V, Arense P, Blatz V, Mandrand-Berthelot MA, Cánovas M, Iborra JL. Role of betaine:CoA ligase (CaiC) in the activation of betaines and the transfer of coenzyme A in *Escherichia coli*. *J Appl Microbiol.* 2008;105:42–50.
51. Ledbetter RN, Garcia Costas AM, Lubner CE, Mulder DW, Tokmina-Lukaszewska M, Artz JH, et al. The Electron Bifurcating FixABCX Protein Complex from *Azotobacter vinelandii*: Generation of Low-Potential Reducing Equivalents for Nitrogenase Catalysis. *Biochemistry.* 2017;56:4177–90.
52. Corbin D, Barran L, Ditta G. Organization and expression of *Rhizobium meliloti* nitrogen fixation genes. *Proc Natl Acad Sci USA.* 1983;80:3005–9.
53. Buchet A, Nasser W, Eichler K, Mandrand-Berthelot MA. Positive co-regulation of the *Escherichia coli* carnitine pathway cai and fix operons by CRP and the CaiF activator. *Mol Microbiol.* 1999;34:562–75.
54. Lee J-Y, Tsolis RM, Bäumlner AJ. The microbiome and gut homeostasis. *Science.* 2022;377:eabp9960.
55. Longo N, Frigeni M, Pasquali M. Carnitine transport and fatty acid oxidation. *Biochim Biophys Acta.* 2016;1863:2422–35.
56. Penhaligan J, Poppitt SD, Miles-Chan JL. The Role of Bovine and Non-Bovine Milk in Cardiometabolic Health: Should We Raise the “Baa”? *Nutrients.* 2022;14.
57. Penn D, Dolderer M, Schmidt-Sommerfeld E. Carnitine concentrations in the milk of different species and infant formulas. *Biol Neonate.* 1987;52:70–9.

58. Foster DW. The role of the carnitine system in human metabolism. *Ann N Y Acad Sci.* 2004;1033:1–16.
59. Hulme H, Meikle LM, Strittmatter N, van der Hooft JJJ, Swales J, Bragg RA, et al. Microbiome-derived carnitine mimics as previously unknown mediators of gut-brain axis communication. *Sci Adv.* 2020;6:eaax6328.
60. Jaworska K, Bielinska K, Gawrys-Kopczynska M, Ufnal M. TMA (trimethylamine), but not its oxide TMAO (trimethylamine-oxide), exerts haemodynamic effects: implications for interpretation of cardiovascular actions of gut microbiome. *Cardiovasc Res.* 2019;115:1948–9.
61. Ahmed H, Leyrolle Q, Koistinen V, Kärkkäinen O, Layé S, Delzenne N, et al. Microbiota-derived metabolites as drivers of gut-brain communication. *Gut Microbes.* 2022;14:2102878.
62. Papandreou C, Moré M, Bellamine A. Trimethylamine N-Oxide in Relation to Cardiometabolic Health—Cause or Effect? *Nutrients.* 2020;12.
63. Burton KJ, Krüger R, Scherz V, Münger LH, Picone G, Vionnet N, et al. Trimethylamine-N-Oxide Postprandial Response in Plasma and Urine Is Lower After Fermented Compared to Non-Fermented Dairy Consumption in Healthy Adults. *Nutrients.* 2020;12.
64. Kountz DJ, Behrman EJ, Zhang L, Krzycki JA. MtcB, a member of the MttB superfamily from the human gut acetogen *Eubacterium limosum*, is a cobalamin-dependent carnitine demethylase. *J Biol Chem.* 2020;295:11971–81.
65. Ellenbogen JB, Jiang R, Kountz DJ, Zhang L, Krzycki JA. The MttB superfamily member MtyB from the human gut symbiont *Eubacterium limosum* is a cobalamin-dependent γ -butyrobetaine methyltransferase. *J Biol Chem.* 2021;297:101327.
66. Pimentel G, Burton KJ, von Ah U, Bütikofer U, Pralong FP, Vionnet N, et al. Metabolic footprinting of fermented milk consumption in serum of healthy men. *J Nutr.* 2018;148:851–60.
67. Avalon NE, Murray AE, Baker BJ. Integrated Metabolomic-Genomic Workflows Accelerate Microbial Natural Product Discovery. *Anal Chem.* 2022;94:11959–66.
68. Krause J. Applications and Restrictions of Integrated Genomic and Metabolomic Screening: An Accelerator for Drug Discovery from Actinomycetes? *Molecules.* 2021;26.
69. van der Hooft JJJ, Mohimani H, Bauermeister A, Dorrestein PC, Duncan KR, Medema MH. Linking genomics and metabolomics to chart specialized metabolic diversity. *Chem Soc Rev.* 2020;49:3297–314.
70. Doroghazi JR, Albright JC, Goering AW, Ju K-S, Haines RR, Tchalukov KA, et al. A roadmap for natural product discovery based on large-scale genomics and metabolomics. *Nat Chem Biol.* 2014;10:963–8.
71. Goering AW, McClure RA, Doroghazi JR, Albright JC, Haverland NA, Zhang Y, et al. Metabologenomics: Correlation of Microbial Gene Clusters with Metabolites Drives Discovery of a Nonribosomal Peptide with an Unusual Amino Acid Monomer. *ACS Cent Sci.* 2016;2:99–108.
72. Blin K, Shaw S, Kloosterman AM, Charlop-Powers Z, van Wezel GP, Medema MH, et al. antiSMASH 6.0: improving cluster detection and comparison capabilities. *Nucleic Acids Res.* 2021;49:W29–35.
73. Navarro-Muñoz JC, Selem-Mojica N, Mullowney MW, Kautsar SA, Tryon JH, Parkinson EI, et al. A computational framework to explore large-scale biosynthetic diversity. *Nat Chem Biol.* 2020;16:60–8.

74. Collins C, Didelot X. A phylogenetic method to perform genome-wide association studies in microbes that accounts for population structure and recombination. *PLoS Comput Biol.* 2018;14:e1005958.
75. Maddison WP, FitzJohn RG. The unsolved challenge to phylogenetic correlation tests for categorical characters. *Syst Biol.* 2015;64:127–36.
76. Fuchsmann P, Tena Stern M, Bischoff P, Badertscher R, Breme K, Walther B. Development and performance evaluation of a novel dynamic headspace vacuum transfer “In Trap” extraction method for volatile compounds and comparison with headspace solid-phase microextraction and headspace in-tube extraction. *J Chromatogr A.* 2019;1601:60–70.
77. Dunn WB, Broadhurst D, Begley P, Zelena E, Francis-McIntyre S, Anderson N, et al. Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. *Nat Protoc.* 2011;6:1060–83.
78. R Core Team. *R: A Language and Environment for Statistical Computing.* 2022.
79. Zulkower V, Rosser S. DNA Features Viewer: a sequence annotation formatting and plotting library for Python. *Bioinformatics.* 2020;36:4350–2.
80. Adobe Inc. Adobe Illustrator.
81. Roder T. MrTomRod/scoary-2: Calculate associations between genes and traits. 2022. <https://github.com/MrTomRod/scoary-2/>. Accessed 16 Mar 2023.
82. Roder T. troder/scoary-2 - Docker Image | Docker Hub. 2022. <https://hub.docker.com/r/troder/scoary-2/>. Accessed 16 Mar 2023.
83. Roder T. OpenGenomeBrowser Demo Server. Home. 2022. <https://opengenomebrowser.bioinformatics.unibe.ch/>. Accessed 16 Mar 2023.
84. Roder T. Metabolomics dataset of 44 *Propionibacterium freudenreichii* for Scoary2. Mendeley Data. 2023.

3.4. Manuscript 4: Indole yoghurt

Maternal consumption of yoghurt that activates the aryl hydrocarbon receptor increases intestinal group 3 innate lymphoid cells in the offspring

Grégory Pimentel, Thomas Roder, Cornelia Bär, Sandro Christensen, Zahra Sattari, Ueli von Ah, Nerea Fernandez Trigo, Rémy Bruggmann, Andrew J. Macpherson, Stephanie C. Ganal-Vonarburg, Guy Vergères

Status:

Manuscript in preparation, to be submitted to PNAS

Statement of contribution:

AJM, CB, GP, GV, RB, SCGV, TR, UvA and ZS conceived the project. TR wrote the software for strain selection. CB, GV, TR, UvA and ZS selected the strains. TR and ZS designed the AhR assays. NFT and ZS performed the AhR assays. NFT, SCGV, TR and ZS analyzed the AhR assays.

Research objective:

To design a yoghurt using additional strains to maximally activate the AhR and to test whether feeding this yoghurt to pregnant, germ-free mice will affect the pups' immune system in a similar fashion to transient colonialization and indole-3-carbinol [130].

Maternal consumption of yoghurt that activates the aryl hydrocarbon receptor increases intestinal group 3 innate lymphoid cells in the offspring.

Grégory Pimentel¹, Thomas Roder², Cornelia Bär¹, Sandro Christensen³, Zahra Sattari^{1,3}, Ueli von Ah¹, Nerea Fernandez Trigo³, Rémy Bruggmann², Andrew J. Macpherson³, **Stephanie C. Ganal-Vonarburg³**, **Guy Vergères¹** (shared)

¹ Agroscope, Schwarzenburgstrasse 161, CH-3003 Bern, Switzerland.

² Interfaculty Bioinformatics Unit, University of Bern, Baltzerstrasse 6, CH-3012 Bern, Switzerland.

³ Department of Visceral Surgery and Medicine, Inselspital, Bern University Hospital, Department for BioMedical research (DBMR), University of Bern, Murtenstrasse 35, CH-3008 Bern, Switzerland.

Abstract

Indole derivatives, as metabolites of the tryptophan pathway, are bioactive microbial compounds with a role in gut immune homeostasis. They bind to the aryl hydrocarbon receptor (AhR), thereby modulating development of intestinal group 3 innate lymphoid cells (ILC3) and subsequent interleukin-22 production. In mice, indole derivatives of the maternal microbiota present in maternal milk drove early postnatal intestinal ILC3 development. Apart from the gut microbiota, lactic acid bacteria (LAB) also produce indole compounds during milk fermentation. The aim of our study was to test, in germ-free mice, if maternal intake of a dairy product enriched in indole metabolites through fermentation could boost maturation of the intestinal innate immune system in the offspring.

631 LAB from Agroscope's collection were genetically screened in regard to their potential to produce indole compounds. 135 different yoghurts were produced, and their ability to activate AhR was evaluated *in vitro* using the HepG2-AhR-Luc cell line. The most efficient indole yoghurt or a control yoghurt were fed to germ-free dams during pregnancy and lactation. Analysis of the offspring on postnatal day 14 by flow cytometry revealed an increase in the frequency of small intestinal lamina propria NKp46+ ILC3s in the pups born to the mothers that had consumed the indole yoghurt compared to the control yoghurt.

The selection of specific LABs based on their ability to produce a fermented dairy able to activate AhR appears to be an effective approach to produce a yoghurt with immunomodulatory properties in young mice.

Introduction

Fermented foods make 20 to 40% of the current world food supply [1, 2]. Although rests of grapes and fermentation markers dating from 4300 y BC were discovered in a jar in Greece [3], genetic evidence suggests that adaptation of hominids to naturally fermented foods, in particular ethanol in ripen fruits, may have taken place already 10 million years ago [4]. The selective advantage food fermentation provides to humans includes food preservation [5], modification of the organoleptic character of the food matrix [6], but also health benefits [7]. Fermentation transforms the food constituents (e.g. breakdown of proteins to produce bioactive peptides or increase the bioavailability of free amino acids), synthesizes bioactive and nutritive compounds (e.g. B vitamins), and delivers commensal microbes as well probiotics to the gastrointestinal tract, thereby promoting the establishment of a healthy gut microbiome [8]. These metabolic properties of fermented foods have in turn been associated with potential benefits for a broad range of organ systems including the digestive, cardiovascular, and nervous systems.

The field of microbiology has been revolutionized during the last decade based on the technological breakthroughs in DNA sequencing and biocomputing. These technologies have contributed to validate the importance of the gut microbiota in human health [9] as well as the key role that nutrition has in modulating the structure and dynamics of the gut microbiota [10] and the immune system. Interestingly, many of the metabolites that are key to human health are derived from metabolization of dietary components by the gut microbiota. These include the transformation of choline to trimethylamine-N-oxide (TMAO), the transformation of primary bile salts to secondary and tertiary bile salts, the digestion of indigestible dietary fibre to produces short-chain fatty acids (SCFA), as well as the production of immunomodulatory indoles from tryptophan [11]. Interestingly, these nutrients are similarly metabolized during the fermentation of food by microorganisms, suggesting that the consumption of fermented foods may modulate the delivery of bioactive nutrients otherwise produced by the human gut microbiome, as shown for methylamines [12], bile acids [13], and indoles [13]. In line with these findings, we [14] recently analyzed *in silico* the pan-genome of a collection of >600 genomically annotated lactic acid bacteria (LAB) and found that a subset of 24 of these strains, each from a different species, covered 89% of the enzymatic reactions of the human gut microbiome. Taken together, these results indicate that the targeted selection of bacteria for food fermentation has a significant potential for the production and delivery of bioactive substances to the human organism. Indoles are tryptophan metabolites that represent an interesting group of bioactive molecules to target in biotechnology. Indoles contribute, through activation of the aryl hydrocarbon receptor (AhR) and the pregnane X receptor (PXR), to intestinal health and immune regulation [15] as well as to an array of additional properties associated with diabetes mellitus or vascular regulation [16, 17]. Hence, one

interesting strategy to modulate immune responses is the use of nutritional AhR ligands [18], possibly produced as bioactive indoles through food fermentation [13]. Murine studies showed that the amount of AhR ligands present in diets affected the formation of intestinal lymphoid follicles by effecting expansion of small intestinal AhR-expressing type 3 innate lymphoid cells (ILC3) [19] and the maintenance of intraepithelial lymphocytes [20], both of which are important contributors to the host-microbial mutualism and intestinal homeostasis. ILC3s produce the cytokine IL-22 acting on IL-22-receptor expressing intestinal epithelial cells which in turn respond with the production of anti-microbial peptides [21, 22]. ILC3s have also been shown to prevent translocation of intestinal microbes to systemic sites [23] and to the defense against the enteric pathogen *Citrobacter rodentium* [19, 24]. Interestingly, we could show with a model of reversible colonization of pregnant germ-free mice that indole metabolites produced by the maternal microbiota can reach the offspring via breast milk and can drive early postnatal expansion of a subset of intestinal ILC3s in the offspring [25].

In this report we addressed whether LAB could be selected based on their genomic content to produce indole metabolites and to produce a yoghurt with increased level of indoles that is able to activate the AhR. By feeding pregnant and lactating germ-free mice the indole-enriched yoghurt, we assessed if this could modulate innate immune system development of the offspring.

Results

Selection of strains for production of indole-rich yoghurt. For the selection of lactic acid bacteria (LAB) that potentially synthesize AhR-activating ligands, the genomes of 663 strains from the strain selection of Agroscope, the Swiss center of excellence for agricultural research, were screened regarding their potential to produce indole and indole derivatives as described in the methods section. 128 strains were chosen, and each used as a supplement to a standard starter culture to ferment lactose-free cow milk to produce the test yoghurts. A control yoghurt was produced by fermenting milk with a standard yoghurt starter culture of two different bacteria only. The 128 test yoghurts were screened *in vitro* as to their ability to activate AhR using a the HepG2-AhR-Luc cell line in which the luciferase gene is controlled by an AHR-response element. Luminescence signals showed that nineteen yoghurts induced a significant AhR activation as compared to the negative control (cell culture medium, Wilcoxon Signed-rank test $p < 0.05$). Based on these results, a yoghurt with expected high AhR activity (Indole Yoghurt) was designed. The Indole Yoghurt consisted of milk fermented with four bacterial species: the two normal yoghurt starter strains (*Lactobacillus delbrueckii* ssp. *Bulgaricus* and *Streptococcus salivarius* ssp. *Thermophilus*), associated with the strain that gave the highest AhR activation signal and a strain chosen according to AhR activation and species compatibility considerations. The 16 yoghurts with highest AHR activation in the first round and a Control yoghurt were then compared to the Indole Yoghurt, the

latter showing a 2.76-fold increase in AhR activation compared to the Control yoghurt (Wilcoxon Signed-rank test $p = 0.024$, Figure 1).

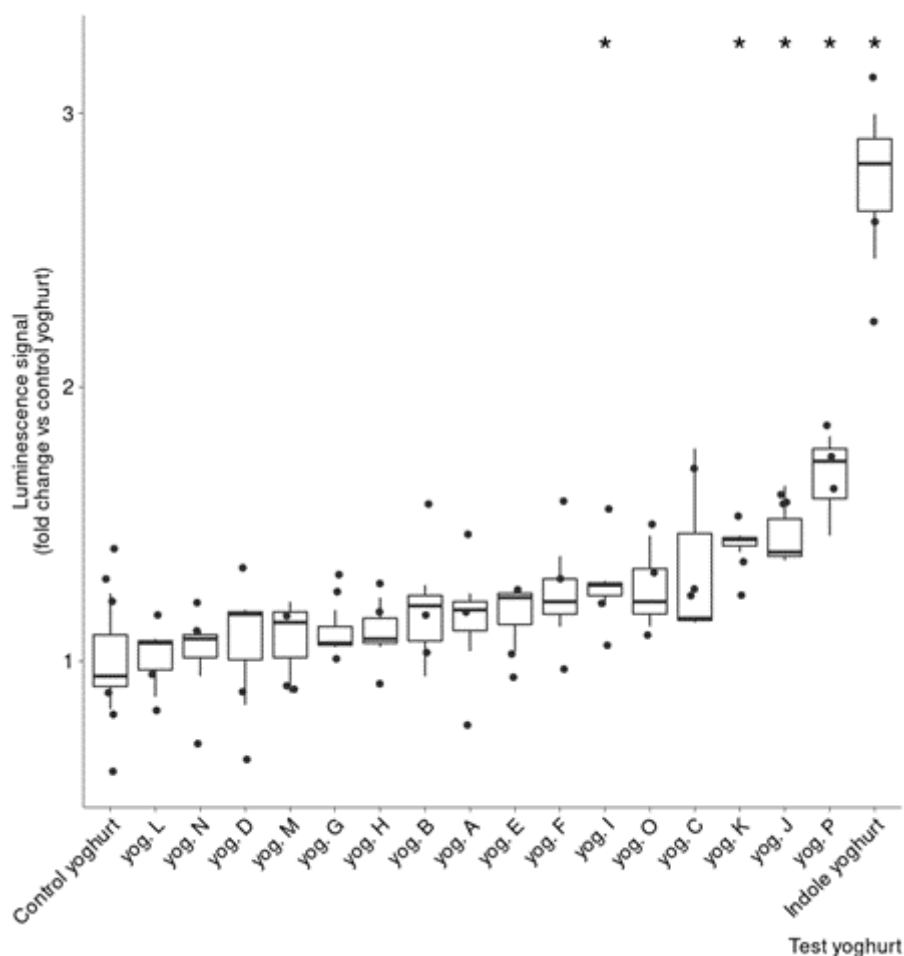


Figure 1: In vitro AhR activation assay to screen indole content of produced yoghurts.

Metabolomic analysis of the yoghurts. As a further read-out to quantify indole metabolites and derivatives in the yoghurts, we performed UHPLC-MS of the Indole and Control yoghurt as well of the other produced yoghurts and of unfermented milk. Of the 37 compounds belonging to the tryptophan pathway, 13 were detected in the test products (table XX), with seven of these showing a significant difference between milk and/or control yoghurt and/or Indole Yoghurt ($p < 0.05$, Figure 2) and we decided to continue our study with the designed Indole Yoghurt.

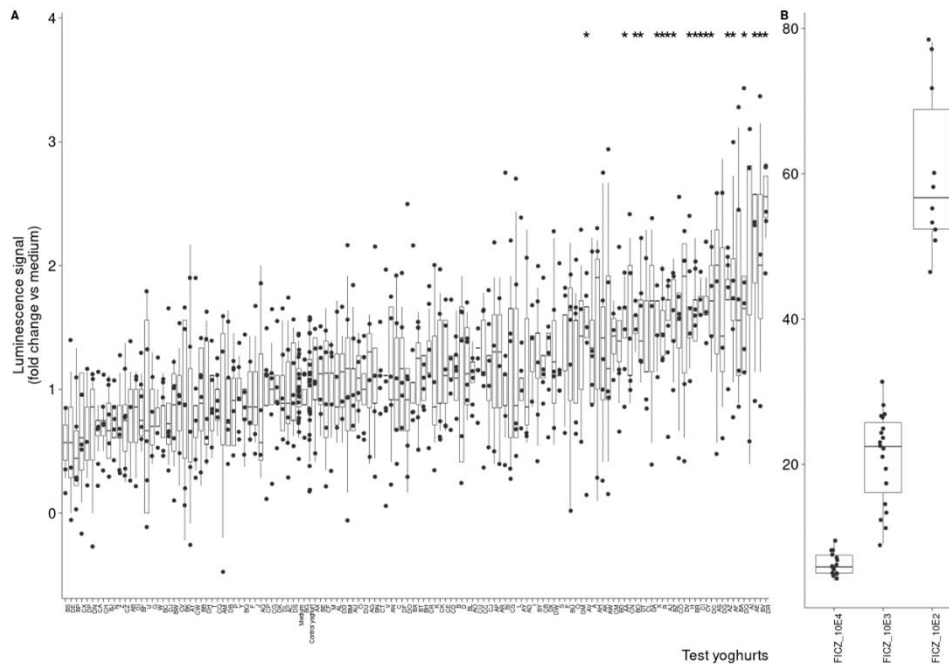


Figure 2: Metabolomic analysis reveals increased levels of tryptophan metabolites in Indole Yoghurt compared to Control Yoghurt and milk.

Production of yoghurt containing diets for mice. To test the bioactivity of the Indole Yoghurt on the immune system *in vivo*, we incorporated the Indole Yoghurt into murine diet. The produced yoghurt was lyophilized and incorporated at 40% (w/w) into an open standard purified diet at ResearchDiets Inc. (USA). A control diet was produced using the Control Yoghurt. A major quality control following production was the sterility of these diets as yoghurt displays a high bacterial load. Absence of any live microbe was a pre-requisite to our murine studies to be able to study the direct impact of the indole metabolites on the host immune system without the confounding factor of live microbes that can directly be sensed by and influence the intestinal immune system. The diets were sterilized by irradiation and tested for sterility by microbiological culture-dependent and independent methods and by *in vivo* testing in germ-free mice which were fed with the diets for 4 weeks and maintained their sterile status.

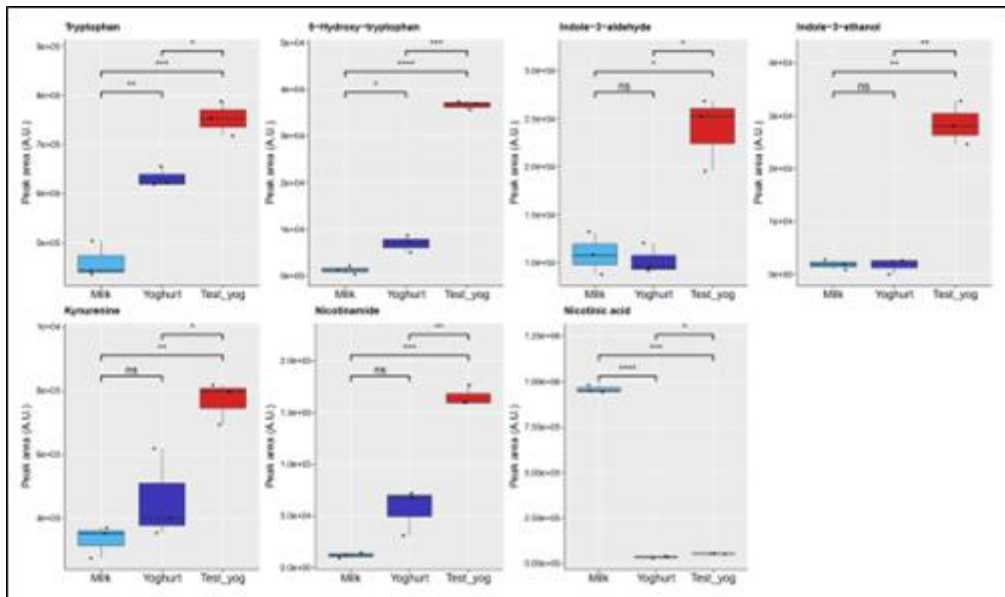


Figure 3: Metabolomic analysis of yoghurt-containing diet pellets.

Mice born to germ-free dams consuming Indole Yoghurt harbor increased numbers of ILC3s in the intestine. Germ-free pregnant mice were switched to either of the two different yoghurt-containing diets at day 7 post conception and kept on this diet until postnatal day 10. On postnatal day 14, the immune compartment in the small intestinal lamina propria of the offspring was analyzed by flow cytometry (Figure 4). Germ-free mice were used as a model to exclude a secondary impact of an altered maternal or infant microbiota after yoghurt consumption. The effect of AhR ligands, which were fed to the pregnant dam and transferred to the pups via breast milk, on the frequency of ILC3 in the offspring small intestine was previously demonstrated using germ-free mice and timed colonization or application of AhR ligands during pregnancy (Gomez et al., 2016). As previously, AhR ligands fed to the dams, in our case via consumption of the Indole Yoghurt compared to the Control Yoghurt, led to an increase in the frequency of NKp46+ ILC3s in the small intestinal lamina propria of the offspring at postnatal day 14 (Figure 4B and C). This effect was specific to ILC3s, as ILC2 and T cell populations were not altered in frequency (Supplementary figure X) between the experimental groups. We also collected serum and breast milk of the dams on day 10 post birth and metabolically analyzed the samples by mass spectrometry. XX and YY compounds were detected in milk and serum samples respectively. (More results to follow.)

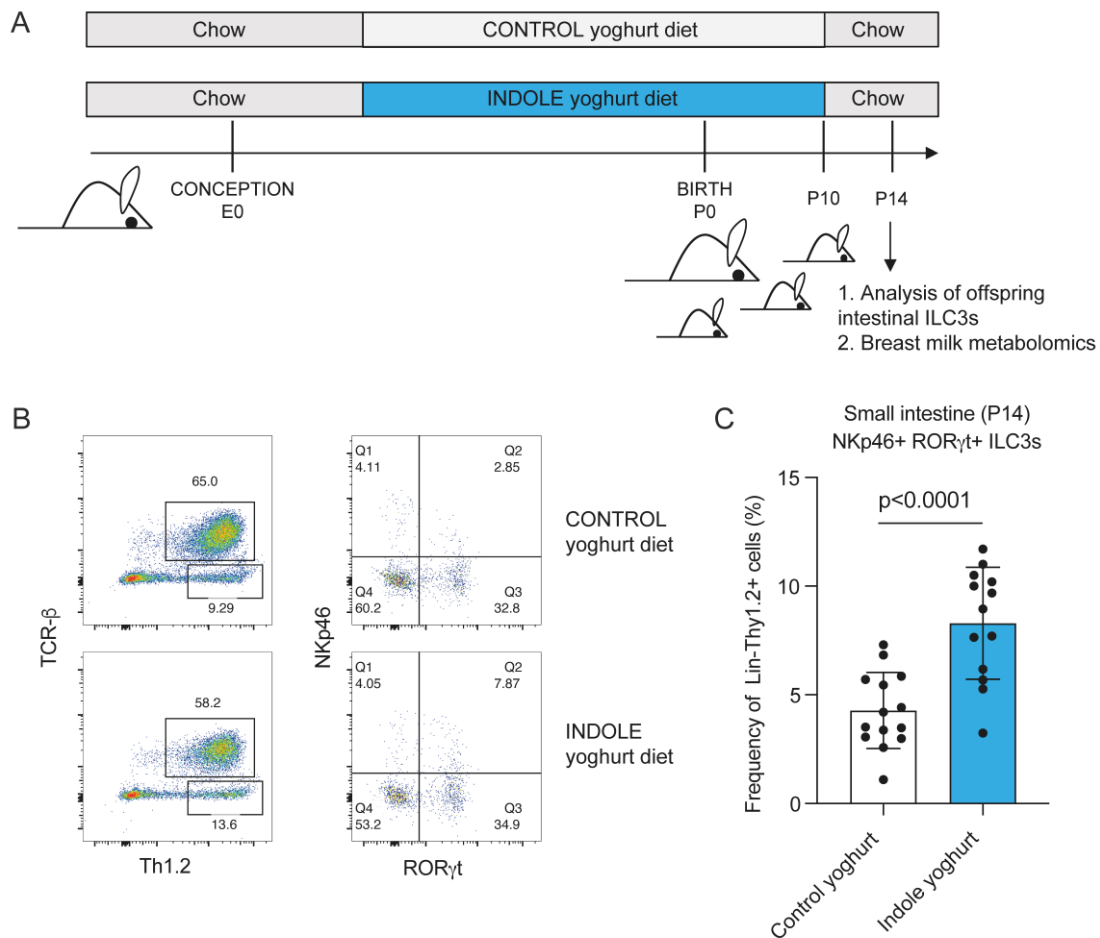


Figure 4: Feeding germ-free dams with a yoghurt activating the aryl hydrocarbon receptor increases intestinal group 3 innate lymphoid cells in the offspring. Time pregnant germ-free C57BL/6 dams were fed purified diets containing Indole Yoghurt or Control Yoghurt starting on day 7 post conception until postnatal day 10. The offspring in each group was analyzed on postnatal day 14 by flow cytometry of the intestinal lamina propria (A). Representative dot plots gated on CD19⁻ (left) and CD19-TCRb-Thy1.2⁺ (right) small intestinal lamina propria lymphocytes. (B) Relative frequency of small intestinal NKp46⁺RORγt⁺ ILC3 at postnatal day 14. Data represent mean ± SD, n=14 pups (Control Yoghurt), n=13 pups (Indole Yoghurt) pooled from two independent experiments.

Discussion

- Importance of food for health/immune system
- Processed food, debate on reintroduction of safe microbes into industrialized food (Marco et al., 2020).
- Fermentation and targeted approach was successful
- Importance of environmental impacts on immune development in early life
- Compare AHR signaling/ligand profile mouse/human. Discuss potential negative effects of AhR activation.

- Possibly: discuss other bioactive ligands?

Methods

1. Genomic selection of bacterial strains

For the selection of lactic acid bacteria that potentially synthesize AhR-activating ligands, the genomes of 663 strains from the strain selection of Agroscope, the Swiss center of excellence for agricultural research, were annotated using KAAS [26], which allowed us to identify the key genes of the phenylalanine, tyrosine and tryptophan biosynthesis pathway. Genes that (i) metabolize tryptophan into indoles and (ii) genes that were shown to provide some advantage at surviving the gastrointestinal transit (specifically bile resistance, acid resistance, gut adherence) were identified through literature search. Finally, the data was integrated into a web-app, that allowed to display the orthogenes present in each strain and the respective copy-number, grouping strains with an identical coverage pattern together. For each group, the pathway coverage could be visualized on the KEGG website.

In addition, the app enabled comparison of orthologous genes for each group of genomes. The analysis included a phylogenetic tree of the genomes based on whole-genome OrthoANI-similarity (v. 1.40) [27], an alignment of the genes calculated using Clustal Omega (v. 1.2.4) [28] and visualized using MView (v. 1.64) [29], and a phylogenetic tree based on said alignment calculated with IQ-TREE (v. 1.6.10) [30]. Most of these functions evolved into the software OpenGenomeBrowser [31].

Based on this information, 128 strains belonging to seven different genus of lactic acid bacteria were selected for yoghurt production and phenotypic characterization.

2. Test yoghurts production

All 128 test yoghurts were produced at Agroscope, Federal Research Station for Agriculture (Bern) to industrial standards in accordance with Swiss food legislation. Lactose free, full-fat (3.5 %), homogenised, pasteurized milk (Aha! IP Suisse, Migros, Switzerland) was used. The 128 tested strains were precultured 16 to 24 h at 30 °C or 37 °C in their respective growth medium, before being added to milk along with a classical yoghurt starter culture consisting of a mix of *Lactobacillus delbrueckii* ssp. *bulgaricus* and *Streptococcus salivarius* ssp. *thermophilus* (Yoflex® YC-381, Chr. Hansen A/S, Denmark). One yoghurt contained only the starter culture without any additional strain. Milk was fermented during 16 h at 37 °C, cooled down to 4 °C and stored at – 20 °C prior analysis.

3. In vitro AhR reporter system

Test yoghurts' ability to activate AhR was evaluated in vitro using the HepG2-AhR-Luc cell line. Human HepG2 liver carcinoma AhR-Lucia reporter (HepG2-Lucia™ AhR) cells were purchased from InvivoGen and grown in Eagle's minimal essential medium (EMEM, ThermoFisher), supplemented

with 10% (v/v) heat-inactivated (30 min at 56 °C) fetal bovine serum (ThermoFisher), 1X non-essential amino acids (ThermoFisher), 100U/ml Pen-Strep (ThermoFisher), and 100 µg/ml Normocin™ (InvivoGen).

At the start of the assay, cells were rinsed with PBS, detached with trypsin, centrifuged (300 g, 5 min, 20°C), counted and resuspended 1.1×10^5 cells/ml in test medium.

20 µl of yoghurt sample were added to a well of a flat-bottom 96-well plate. An AhR agonist (e.g. FICZ in EMEM at 1 µg/ml final concentration) was used as positive control and endotoxin free water as a negative control. 180 µl of cell suspension (~20,000 cells) per well were added to each well and incubated at 37 °C in a CO2 incubator for 72 h. 40 µl of HepG2-Lucia™ AhR stimulated cell supernatant were transferred into a 96-well white (opaque) or black plate. 100 µl of QUANTI-Luc™ solution was added quickly to all wells before immediate measurement at the TECAN Reader Infinite 200 (reading time (integration time) was 500 milliseconds and the end point measurement (settle time) was set at 1000.

4. Yoghurt-containing diets

Two test yoghurts were prepared at Agroscope for the in vivo studies. A control, containing only the starter strains, and a test yoghurt containing the starter and two additional strains selected from the 128 previously tested in vitro. The yoghurts were lyophilized and shipped to Research Diets Inc. (New Brunswick, NJ, USA) where they were incorporated at 40 % (w/w) into an open standard diet (diet D11112201 for mice). The diet pellets were sterilized by two rounds of gamma-ray irradiation (10-20 kGy) and one round of X-ray irradiation (30-60 kGy). They were imported into a surgical isolator after disinfection of the bag with 2% peracetic acid. The sterility of the diets was assessed prior to each experiment by aerobic and anaerobic cultures and Sytox and Gram stainings of the caecal contents of germ-free mice that were fed the purified diets for 7 days and switched back to autoclaved chow for another 14 days.

5. In vivo studies

Germ-free (GF) C57BL/6 were bred and maintained in flexible-film isolators at the Clean Mouse Facility, University of Bern, Switzerland as previously described [32]. Germ-free status was routinely monitored by culture-dependent and independent methods. All mouse experiments were performed in accordance with Swiss Federal and Cantonal regulations under the cantonal license BE104/20. Mice were born and raised while fed conventional chow (3307, Kliba Nafag).

GF females at the age of 8-10 weeks were time-mated (vaginal plug check) in experimental sterile isolators and switched to either FAMIX or control yoghurt-containing diets 7 days post mating and kept on the experimental diet until 10 days post giving birth. Mice were then switched back to chow to prevent the growing pups from being exposed to yoghurt-containing diets that may have fallen into the cage. Dams and offspring were exported from the experimental isolator on postnatal day 14 and

analyzed as described. Sterility of the mice was confirmed continuously throughout the experiment by culture-dependent and independent methods.

6. Isolation of small intestinal lamina propria lymphocytes

The intestines were removed from the mouse and placed in ice-cold DPBS (Gibco). Residual fat and Peyer's patches were removed. The intestine was opened longitudinally before cut into 2 cm segments. The tissue was washed once in ice-cold DPBS followed by four washes of 8 min in 15 ml of DPBS (5 mM EDTA, 10 mM HEPES) with shaking at 37°C to remove epithelial cells. Residual tissue was then washed in 15 ml of IMDM containing 10% FCS (IMDM/FCS) at 37°C for 8 min before being minced and digested in 15 ml of IMDM containing 0.5 mg/ml collagenase type VIII (Sigma) and 10 U/ml DNase I (Roche) with shaking at 37°C for 20-30 min (small intestine) or 30-40 min (colon). The obtained cell suspension was passed through a cell strainer (100 μ m) and washed with IMDM/FCS. Cells were centrifuged (600g, 7 min, 4°C) and resuspended in FACS buffer (PBS, 2 % FCS, 2mM EDTA, 0.01 % NaN₃) for staining for flow cytometry analysis.

7. Flow cytometry

Cells were washed once with DPBS before being stained with fixable viability dye (eBioscience) and anti-mouse-CD16/CD32 Fc-receptor block (93, Biolegend) diluted in DPBS for 30 min on ice. Single cell suspensions were sequentially incubated with primary/biotin- and fluorescence-coupled antibodies diluted in FACS buffer for 15 min on ice. For intracellular stainings, cells were fixed and permeabilized using the Transcription Factor Staining Buffer Set (eBioscience). Antibodies for intracellular staining were diluted in the permeabilization buffer from the Transcription Factor Staining Buffer Set and incubated at 4°C overnight. The following mouse-specific conjugated antibodies were used: CD19 (6D5, Biolegend), CD4 (RM4-5, Biolegend), CD44 (IM7, Biolegend), CD62L (MEL-14), CD8a (53-6.7, BD Bioscience), Foxp3 (FJK-16s, ThermoFisher), Gata-3 (TWAJ, Biolegend), Helios (22F6, Biolegend), NKp46 (19A1.4, Biolegend), ROR γ t (B2D, ThermoFisher), T-bet (4B10, ThermoFisher), TCR- β (H57-597, Biolegend), TCR-gd (GL3, BD Bioscience), Thy1.2 (53-2.1, Biolegend). Data were acquired on a LSRFortessa (BD Biosciences) and analyzed using FlowJo software version 10.6.2 (Tree Star Inc.). In all experiments, FSC-H versus FSC-A was used to gate on singlets with dead cells excluded using. Where lineage exclusion was performed, TCR and CD19 expressing cells were removed from further analysis.

8. Collection of breast milk

14 days after delivery, the nursing dams were separated from their pups for four hours before milking. Dams were anesthetized with isoflurane according to standard operating procedures and 1U of oxytocin (Syntocinon) was injected intraperitoneally. The collection of breast milk was started within 5 min using a custom-made vacuum pump-based collection device. Aliquots of milk were frozen in liquid nitrogen

and stored at -80 °C until further analysis.

9. UHPLC-MS analysis

Protein precipitation of mice milk samples and test products was obtained with the addition 1:4 (vol/vol) of acetonitrile containing 1% (vol/vol) formic acid, and centrifugation at 12000 RPM for 15 min. Supernatant was filtered through a phospholipids filter membrane to limit ion suppression (Phree®, Phenomenex Inc., Torrance, California, USA). The filtrate was then injected into the UHPLC/MS system consisting in an UltiMate 3000 HPLC (Thermo Fisher Scientific) coupled to a maXis 4G+ quadrupole time-of-flight mass spectrometer (MS) with electrospray interface (Bruker Daltonik GmbH, Bremen, Germany). Chromatographic separation was performed on a C18 hybrid silica column (Acquity UPLC HSS T3 1.8 µm 2.1 × 150 mm, Waters, UK), reversed phase at a flow rate of 0.4 ml/min. The mobile phase consisted in ultrafiltered water (Milli-Q® IQ 7000, Merck, Germany) containing 0.1% formic acid (Fluka™, Honeywell, USA) (A), and acetonitrile (Supelco®, Merck, Germany) with 0.1 % formic acid (B), with the following elution gradient (A:B): 95:5 at 0 min to 5:95 at 10 min; 5:95 from 10 to 20 min; 95:5 from 20 to 30 min. The spectra were recorded from m/z 75 to m/z 1500 in positive ion mode. Detailed MS settings were as follows: collision-induced dissociation: 20 to 70 eV, electrospray voltage: 4.5 kV, endplate offset: 500 V, capillary voltage: 3400 V, nitrogen flow: 4 ml/min at 200 °C, spectra acquisition rate: 1 Hz in profile mode, resolution: 80,000 FWHM.

The process of identification focused on 37 metabolites of the tryptophan pathway, and in particular indole derivatives. Their presence in tests products, mice milk, and mice serum was investigated by performing collision-induced dissociation (5–70 eV collision energies) and with the use of pure standards (list investigated compounds and suppliers in Supplemental). The amount of each compound was assessed by using the signal intensity given by Bruker Compass DataAnalysis (Bruker Daltonik, GmbH). Intensities between normal yoghurt and test yoghurt were compared by a Wilcoxon test ($p < 0.05$ as significance threshold) using R (v.4.2.1; R Foundation for Statistical Computing, Vienna, Austria).

Acknowledgements

We are grateful for support by the Clean Mouse Facility which is supported by the Genaxen Foundation, Inselspital and the University of Bern. This work was supported by a grant from the Gebert Rűf foundation (Polyfermenthealth – Microbials 2017). S.C.G.V. was supported through a Peter Hans Hofschneider Professorship provided by the Stiftung Molekulare Biomedizin.

References

1. Li KJ, Brouwer-Brolsma EM, Burton KJ, Vergères G, Feskens EJM. Prevalence of fermented foods in the Dutch adult diet and validation of a food frequency questionnaire for estimating their intake in the NQplus cohort. *BMC Nutr.* 2020;6:69.
2. Campbell-Platt G. Fermented foods — a world perspective. *Food Res Int.* 1994;27:253–7.
3. Garnier N, Valamoti SM. Prehistoric wine-making at Dikili Tash (Northern Greece): Integrating residue analysis and archaeobotany. *J Archaeol Sci.* 2016;74:195–206.
4. Carrigan MA, Uryasev O, Frye CB, Eckman BL, Myers CR, Hurley TD, et al. Hominids adapted to metabolize ethanol long before human-directed fermentation. *Proc Natl Acad Sci USA.* 2015;112:458–63.
5. Naghmouchi K, Belguesmia Y, Bendali F, Spano G, Seal BS, Drider D. *Lactobacillus fermentum*: a bacterial species with potential for food preservation and biomedical applications. *Crit Rev Food Sci Nutr.* 2020;60:3387–99.
6. Zhao CJ, Schieber A, Gänzle MG. Formation of taste-active amino acids, amino acid derivatives and peptides in food fermentations - A review. *Food Res Int.* 2016;89 Pt 1:39–47.
7. O’Leary K. Health benefits of fermented foods. *Nat Med.* 2021.
8. Marco ML, Heeney D, Binda S, Cifelli CJ, Cotter PD, Foligné B, et al. Health benefits of fermented foods: microbiota and beyond. *Curr Opin Biotechnol.* 2017;44:94–102.
9. Gilbert JA, Quinn RA, Debelius J, Xu ZZ, Morton J, Garg N, et al. Microbiome-wide association studies link dynamic microbial consortia to disease. *Nature.* 2016;535:94–103.
10. Gentile CL, Weir TL. The gut microbiota at the intersection of diet and human health. *Science.* 2018;362:776–80.
11. Neves AL, Chilloux J, Sarafian MH, Rahim MBA, Boulangé CL, Dumas M-E. The microbiome and its pharmacological targets: therapeutic avenues in cardiometabolic diseases. *Curr Opin Pharmacol.* 2015;25:36–44.
12. Burton KJ, Krüger R, Scherz V, Münger LH, Picone G, Vionnet N, et al. Trimethylamine-N-Oxide Postprandial Response in Plasma and Urine Is Lower After Fermented Compared to Non-Fermented Dairy Consumption in Healthy Adults. *Nutrients.* 2020;12.
13. Pimentel G, Burton KJ, von Ah U, Bütikofer U, Pralong FP, Vionnet N, et al. Metabolic footprinting of fermented milk consumption in serum of healthy men. *J Nutr.* 2018;148:851–60.
14. Roder T, Wüthrich D, Bär C, Sattari Z, Ah U von, Ronchi F, et al. In Silico Comparison Shows that the Pan-Genome of a Dairy-Related Bacterial Culture Collection Covers Most Reactions Annotated to Human Microbiomes. *Microorganisms.* 2020;8.
15. Stockinger B, Di Meglio P, Gialitakis M, Duarte JH. The aryl hydrocarbon receptor: multitasking in the immune system. *Annu Rev Immunol.* 2014;32:403–32.
16. Ye X, Li H, Anjum K, Zhong X, Miao S, Zheng G, et al. Dual role of indoles derived from intestinal microbiota on human health. *Front Immunol.* 2022;13:903526.
17. Macpherson AJ, de Agüero MG, Ganai-Vonarburg SC. How nutrition and the maternal microbiota shape the neonatal immune system. *Nat Rev Immunol.* 2017;17:508–17.

18. De Juan A, Segura E. Modulation of immune responses by nutritional ligands of aryl hydrocarbon receptor. *Front Immunol.* 2021;12:645168.
19. Kiss EA, Vonarbourg C, Kopfmann S, Hobeika E, Finke D, Esser C, et al. Natural aryl hydrocarbon receptor ligands control organogenesis of intestinal lymphoid follicles. *Science.* 2011;334:1561–5.
20. Li Y, Innocentin S, Withers DR, Roberts NA, Gallagher AR, Grigorieva EF, et al. Exogenous stimuli maintain intraepithelial lymphocytes via aryl hydrocarbon receptor activation. *Cell.* 2011;147:629–40.
21. Zheng Y, Valdez PA, Danilenko DM, Hu Y, Sa SM, Gong Q, et al. Interleukin-22 mediates early host defense against attaching and effacing bacterial pathogens. *Nat Med.* 2008;14:282–9.
22. Sanos SL, Vonarbourg C, Mortha A, Diefenbach A. Control of epithelial cell function by interleukin-22-producing ROR γ t⁺ innate lymphoid cells. *Immunology.* 2011;132:453–65.
23. Sonnenberg GF, Monticelli LA, Alenghat T, Fung TC, Hutnick NA, Kunisawa J, et al. Innate lymphoid cells promote anatomical containment of lymphoid-resident commensal bacteria. *Science.* 2012;336:1321–5.
24. Sonnenberg GF, Monticelli LA, Elloso MM, Fouser LA, Artis D. CD4(+) lymphoid tissue-inducer cells promote innate immunity in the gut. *Immunity.* 2011;34:122–34.
25. Gomez de Agüero M, Ganal-Vonarburg SC, Fuhrer T, Rupp S, Uchimura Y, Li H, et al. The maternal microbiota drives early postnatal innate immune development. *Science.* 2016;351:1296–302.
26. Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M. KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res.* 2007;35 Web Server issue:W182–5.
27. Lee I, Kim YO, Park S-C, Chun J. OrthoANI: An improved algorithm and software for calculating average nucleotide identity. *Int J Syst Evol Microbiol.* 2015.
28. Sievers F, Higgins DG. Clustal Omega for making accurate alignments of many protein sequences. *Protein Sci.* 2018;27:135–45.
29. Brown NP, Leroy C, Sander C. MView: a web-compatible database search or multiple alignment viewer.
30. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermin LS. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods.* 2017;14:587–9.
31. Roder T, Oberhänsli S, Shani N, Bruggmann R. OpenGenomeBrowser: a versatile, dataset-independent and scalable web platform for genome data management and comparative genomics. *BMC Genomics.* 2022;23:855.
32. Smith K, McCoy KD, Macpherson AJ. Use of axenic animals in studying the adaptation of mammals to their commensal intestinal microbiota. *Semin Immunol.* 2007;19:59–69.

4. Discussion and Outlook

4.1. Management of microbial genome databases

The management of large genome databases is a challenge because so many fields move quickly: sequencing technologies, assembly algorithms and annotation resources. Ideally, all genomes would be measured using the same technology and processed using the same software, avoiding biases in downstream data analyses. As this is not possible, reasonable compromises must be sought.

The most basic problem is the fact that most genome databases grow over time, and thus contain older draft genomes based on short reads, fragmented into hundreds of scaffolds, as well as perfect circularized genomes based on very accurate long reads, and everything in-between. This issue cannot be solved without expensive re-sequencing efforts, so it can only be mitigated by making information about assembly quality and how the data was generated easily accessible to downstream analysts through the meticulous creation of metadata.

Re-assembling short read data using modern assemblers may result in a marginal improvement, but the quality of the assembly would still be limited by fundamental read properties. Quality control of fragmented assemblies is also difficult. In organisms with genome decay, tools like BUSCO may falsely indicate a misassembly. The output of contamination analysis tools like Kraken, GTDB-Tk, and ConFindr is difficult to interpret and may lead to false interpretations in case of understudied species or horizontal gene transfer. Today, if a problem is suspected, instead of performing difficult, non-decisive analyses, the organism should be re-sequenced using long read technologies. Using HiFi long reads and Tricycler [33], it is already possible to create near-perfect assemblies. However, Tricycler depends on multiple assemblers and requires manual intervention, which is often too labor intensive. In the next few years, with the maturation of long-read sequencing and assemblers, this may cease to be a factor.

Since the problem of structural annotations in bacteria has been solved for most purposes, the main consideration is to use the same pipeline for all genomes. While Prokka [48] or Bakta [52] suffice for most bioinformatics analyses and are preferred by many bioinformaticians because of their speed, simplicity, and open development, for genome database management, PGAP [49] may be the better choice. PGAP was developed for this purpose, with useful quality controls and more detailed gene descriptions. Moreover, as an established long-term project, it is a more secure choice.

While it may not be necessary to regularly re-sequence, re-assemble, or structurally re-annotate genomes, new knowledge about the functions of genes continues to be generated and existing genomes should be updated accordingly. Functional annotations are collected and curated by different projects, making this transfer more difficult. The eggNOG database simplifies this task considerably by collecting annotations from the most important sources in an ortholog-aware manner. A new version is released every 2-3 years, which is a reasonable timeframe for re-annotation. EggNOG is growing quickly; version 6 has been released in late 2022 [171], increasing the number of orthogroups from 4.4 M to 17 M and the number of bacteria from 4,445 to 10,756 compared to version 5.

Ortholog information should always be available as many comparative genomics analyses depend on finding orthologs, which is difficult to do using manual similarity-based searches. Regarding ortholog inference, there is still room for improvement. The software OrthoFinder has high accuracy and reasonable performance up to around one thousand genomes, but it cannot be scaled much further because the algorithm is based on an all-versus-all gene comparison, which has quadratic time complexity, i.e., $O(n^2)$. Alternative algorithms have been developed: for instance,

Domainoid [192], exploits protein domains and achieves comparable accuracy as OrthoFinder, DeepNOG [193] is an alignment-free method based on convolutional neural networks, SwiftOrtho [194] is based on k -mers, and PEPPAN [195] has a Linclust-based [196] pre-clustering step. OrthoFinder remains the most complete solution to date for datasets with less than one thousand genomes. It is not clear which software should be used for larger datasets, as these alternatives were either not included in the recent OrthoBench benchmark or their long-term development is uncertain. One exciting development is the OrthoFinder-based phylogeny- and ortholog-aware protein sequence search software SHOOT [197]. It may solve the problems associated with sequence-similarity-search-based searches mentioned in the introduction (Section 1.1.7.2).

4.2. Polyfermenthealth: strain selection

The first step in the Polyfermenthealth project, the strain selection for the functional yoghurts, demanded a quick progression from genomic screening to the identification of functional endpoints and the final selection of the strains to be phenotypically screened. To include the interpretations and opinions from all experts in our team during the initial screening of strains, it was necessary to be able to efficiently go through the data together. We decided to search for interesting variety and potential functional targets using KEGG pathway maps, on which, for each enzyme, a color gradient indicates how many strains possess a gene with the respective annotation. The plan was to find interesting genomic variety directly in its biochemical context and easy to follow up on. Unfortunately, we did not see very much intra-species variety. This is most likely explained by the fact that known genes are more likely to belong to the core metabolism, which overestimates the similarity between strains and only provides limited insight as to the unique metabolic potential of the strains. Nevertheless, this approach brought the folate metabolism (see Section 1.3.3) to our attention, as it seemed to harbor promising variation.

The next challenge was to select strains based on their genetic variety with respect to the pathway of interest, i.e., folate and indoles, to narrow down the number of strains to biologically screen. As in the previous step, the goal was to take advantage of the knowledge of our team members, some of which know these strains intimately. Thus, a website was developed to interactively explore the genomic information at strain level (Figure 8). Again, it allowed users to focus on biology rather than technicalities. It made it possible to search for annotations (KEGG orthologs, GO-terms, and manually curated annotations) amongst the genomes and to investigate phylogenetic relatedness as well as protein-level variation.

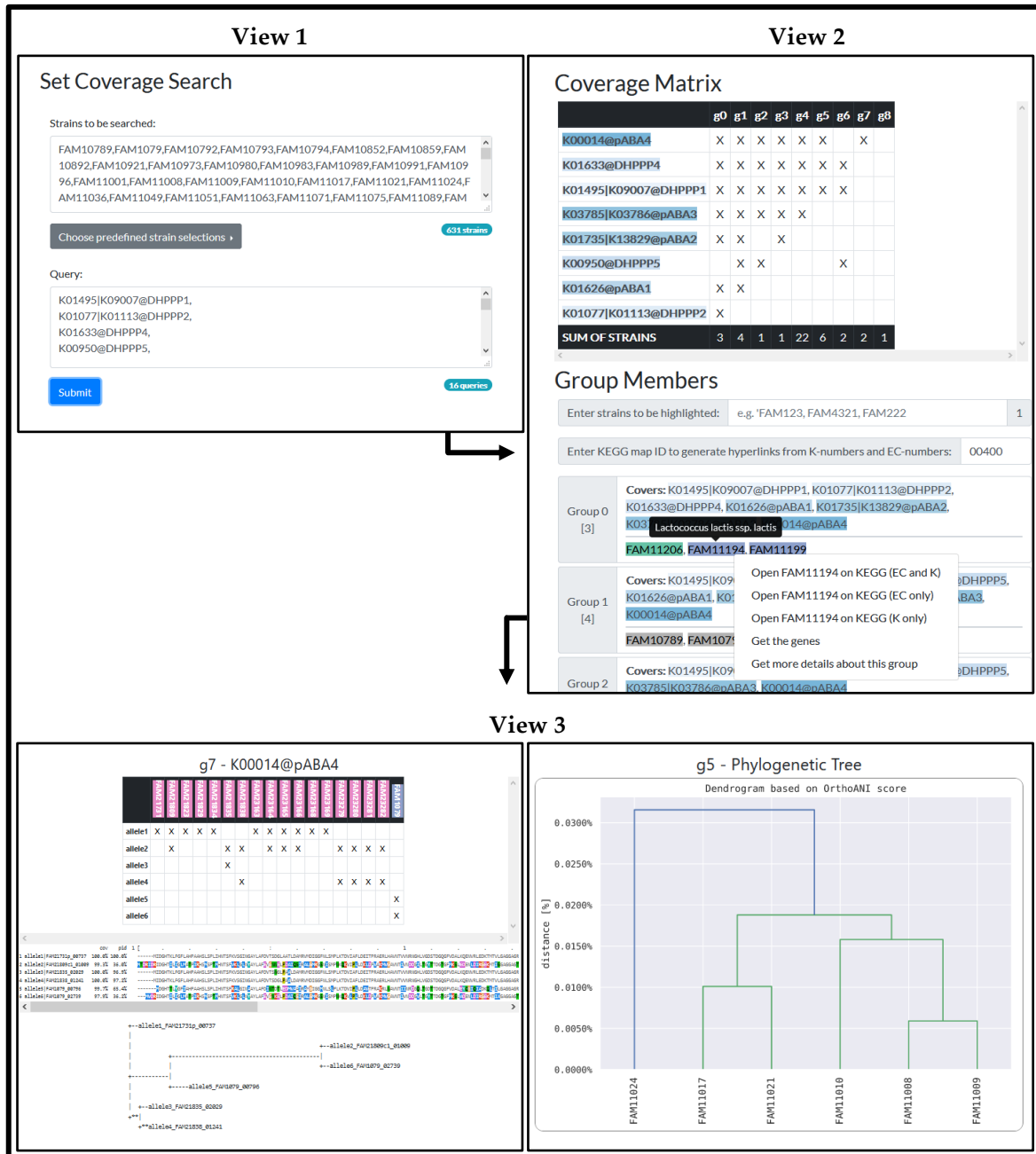


Figure 8: The original strain selection tool. Strains are colored according to the species to which they belong to. (**View 1**) In the first view, strains (top) and annotations (bottom) can be selected. (**View 2**) In the second view, the strains are grouped according to which annotations they cover (a binary 'coverage matrix', top) and there is a summary for each group (bottom). Strains can be opened on the KEGG website, showing the genes of interest in the respective biochemical pathways. (**View 3**) Details about each group are available in the third view, where an ANI-based phylogenetic tree of the group's strains is shown (top). Additionally, for every queried annotation, all different alleles, a multiple sequence alignment and an alignment thereof are shown.

The strain selection website enabled our team of experts from diverse fields to explore the data together and in real-time and make a first selection based on (i) genetic coverage of the pathways of interest, including reactions that branch off; (ii) phylogenetic relatedness; (iii) gene variants; (iv) gene copy numbers; (v) genetic coverage of other annotations, such as antibiotic resistance or gut adherence genes. As shown in Figure 5, the biosynthesis of folate requires two precursors, a pteridine moiety and *p*-aminobenzoic acid. Using this method, we discovered that certain strains in the Dialact database should be able to produce both precursors (e.g., *Propionibacterium freudenreichii*), some neither (e.g., *Lactocaseibacillus paracasei*), some only the pteridine moiety (e.g., *Pediococcus acidilacti*), and some only *p*-aminobenzoic acid (e.g., certain *Lactobacillus delbrueckii subsp. bulgaricus*).

Assessing the success of this approach is challenging. We originally selected strains that were apparently lacking the ability to produce either precursor as negative controls, but to minimize the workload, they were later dropped from the experiment. Although it is possible that even better strains existed outside of the experiment, we did discover strains with high production of folate and indoles.

Using this strain selection strategy, 183 yoghurts were made that contain one strain in addition to the starter culture. These were screened for the phenotypes of interest using an *in vitro* assay to measure AhR activation (see Manuscript 4, Section 3.4), a microbiological assay to measure folate production [158], and untargeted (MS) mass spectrometry to guide the creation of a maximally metabolically diverse yoghurt.

4.2.1. Genome-scale metabolic modeling

Could genome-scale modeling have been used for Polyfermenthealth strain selection?

Everywhere in the literature, it is highlighted that while automatically generated GSMM are a great starting point for metabolic modeling, dedicated experiments and manual curation are essential for good predictions. The generated GSMM from the AGORA pipeline published in 2017, for example, had a prediction sensitivity of merely 32% and a specificity of 92% for gene essentiality of gut microbes. Curation improved these statistics to 68% and 98%, respectively [175]. Models generated using CarveMe, published in 2018, predicted phenotype arrays of *E. coli* with a sensitivity of around 75% and a specificity of around 60% [172]. These constitute far easier tasks than predicting production rates of secondary metabolites or community relations in a complex medium like yoghurt.

The following is a list of compounding problems I see for the use of metabolic modeling for Polyfermenthealth-like projects:

- GSMM are not ideally suited for high throughput screening because manual curation and dedicated experiments are required [187].
- Strain-level metabolic models are a challenge because much of what is important lies in the secondary metabolism [187]. The relevant genes and pathways must be known in advance to be modeled, and such fluxes may be harder to interpret. If a metabolite of interest is modeled as not contributing towards the objective function and competes with another metabolite that does, its flux will be set to zero. If the metabolite does not compete, then the model may predict its production. But in reality, most metabolites probably have both a cost and a benefit which simply are not known in advance, and *in silico* modeling cannot determine it.
- Yoghurt is a very complex, nutrient-rich medium. Metabolic models are generally made using simple, controlled media where nutrients can be added or removed in experiments.
- Yoghurt is a dynamic medium (e.g., pH, metabolites, viscosity). It requires more than steady-state FBA and additional constraints to reflect this.
- As mentioned in the introduction, fermentation is harder to model compared to more efficient forms of metabolism. Yoghurt fermentation is relatively unstable: slight environmental changes like temperature can change the outcome.
- Community models add an additional layer of complexity, and polyfermented yoghurts involve three strains or more by definition.
- GSSM only account for the presence or absence of genes. LAB have particularly many pseudogenes [104], which are difficult to correctly detect as such. For this reason, testing strains with abnormal gene variants should be considered.

GSSM have been used to increase folate production, but as a tool for genetic engineering. The initial strain selection requires screening [198]. The main limitation to creating yoghurts with

increased AhR activity is the fact that most relevant pathways and genes are not characterized [135]. Similarly, for increasing metabolic diversity, GSSM seem to be the wrong tool as these models depict the core metabolism. For instance, a manually curated metabolic model of *Lactococcus lactis* subsp. *cremoris* MG1363 consists of 518 genes and 650 metabolites [199], which amounts to about 20% of its genes and a tiny fraction of its metabolome. For metabolic diversity maximization, the remaining genes and metabolites are arguably more interesting. Given the difficulty of creating good semi-automated GSSM for single strains on simple media and the other challenges listed above, it is very unlikely that community modeling approaches.

In summary, metabolic modeling is a sophisticated instrument that is difficult to use. It is suited for certain applications, like understanding metabolism and genetic engineering, but has clear limitations. Convincingly solving even one or two of the problems listed above would be a major step forward but requires years of research. In this context, it was not a suitable tool for strain selection in the Polyfermenthealth project, which foresaw only 3 months for this step.

4.3. OpenGenomeBrowser

The strain selection website was an *ad hoc* solution, only intended to be used for strain selection in the context of Polyfermenthealth. Despite not being very user-friendly, researchers at Agroscope requested for the tool to be kept online and the inclusion of additional strains. To illustrate why, let us examine how biologists at Agroscope used to investigate a genome's coverage of a KEGG pathway: (1) upload the genome to the KAAS server [58], (2) wait a few hours for the annotation to complete, (3) browse the pathway maps on the server. This method is cumbersome, slow, and does not allow for many genomes to be uploaded at the same time. The server is designed for analyzing only one genome in isolation. While it is possible to compare a few genomes manually, this is cumbersome and not feasible for many genomes. Moreover, the result cannot easily be shared with co-workers, though they may be working on the same organism, and after an unspecified period, the results are removed from the server and the genome must be re-uploaded.

However, long-term maintenance of the strain selection website was not practical because of architectural shortcuts that were taken to make it usable quickly. These shortcuts also meant that the software, if published, would be of little use to other researchers with their own data. Such problems are likely the reason why comparative genomics software are generally not reusable. Realizing there was a great need for automatization of even simple comparative genomics workflows, we decided to design a database-independent software from the ground up. The resulting software, OpenGenomeBrowser, incorporates lessons learned from our group's experience with genome data management, summarized in Section 1.1.8. Over time, the functionalities of the strain selection website were integrated as the *annotation search* tool, the *gene comparison* tool, and the *trees* tool.

Of course, there is great potential for additional tools and other improvements. These include many incremental improvements, for instance, the last feature of the strain selection website that is still missing: the ability to create a phylogenetic tree based on the multiple sequence alignment. Larger upgrades are also possible, for example, igv.js [200], a browser-based classical genome browser that would enable users to inspect reads aligned to the genome assembly. It may also become necessary to add a plugin infrastructure to enable the creation of custom OpenGenomeBrowsers for specific needs. For instance, multi-locus sequence typing (MLST) [201], a procedure for characterizing bacteria using a set of house-keeping genes, is commonly used in the context of hospitals but rarely in other contexts.

I am very much encouraged to see the interest of the community in OpenGenomeBrowser. Some users are very enthusiastic and provide valuable feedback and suggestions. Colleagues at Agroscope have been using it since late 2018, with a dataset that has grown to over 1,500 genomes. Agroscope considers setting up another instance for a different sequencing project. At the SIB days 2022

conference, I was permitted to pitch OpenGenomeBrowser, and won a poster award. At our group, the Interfaculty Bioinformatics Unit, it is incorporated as the last step in the bacterial sequencing pipeline and serves as a value multiplier for five sequencing projects. In 2021, despite not being published at the time, OpenGenomeBrowser drew significant attention from a company, resulting in the formation of a spinoff company to commercialize the software and advance its development.

Though these are encouraging signs, whether OpenGenomeBrowser is a success or not will at least in part be determined by continued maintenance and development. Sequencing projects are often long-term, yet the fields of sequencing, assembly, annotation, and comparative genomics are very fast-moving. Accordingly, the software must be able to keep up with recent developments or risks becoming outdated. Moreover, it is sometimes difficult to put oneself in the shoes of a user as a programmer of a software. Therefore, it is important to maintain contact with users and give them the necessary knowledge to communicate their suggestions effectively and without hurdles.

4.4. Polyfermenthealth: yoghurts

Early in the project, we foresaw several challenges to overcome regarding yoghurt production for the mice trials. They were addressed in parallel to and after strain selection. First, we had to use lactose-free milk for the yoghurt production, to prevent digestive problems in the mice, as mice are lactose-intolerant. Second, because yoghurt contains large amounts of live bacteria, germ-free mice cannot be fed with regular yoghurt without becoming contaminated. Instead, *fermented milk pellets* (FMPs) were developed, and we had to test and confirm that the germ-free mice remained sterile after consuming them.

4.4.1. Polydiverse yoghurt

The goal for this subproject was to create a yoghurt with increased diversity. Diversity can be interpreted in multiple ways, for instance strains could be chosen either to maximize the number of species or orthogenes, or to maximize the diversity of metabolites produced during fermentation (metabolic diversity). We decided to focus on metabolic diversity, but implicitly increased species and orthogene diversity, too. To design a maximally diverse yoghurt using the metabolomics dataset of the 183 yoghurts produced with one additional strain each, we devised two metrics of diversity: Shannon entropy and number of new metabolites.

In the first approach, yoghurts were combined *in silico* by taking the mean value of each metabolite, and from this, the expected entropy was calculated. The greedy algorithm was used to suggest promising combinations of five strains. In the second approach, the number of new metabolites in each yoghurt was calculated (i.e., metabolites not present in yoghurt made from starter only). These approaches were combined to select the 9 most promising strains, from which 127 different yoghurts consisting of 5 strains each were made. Incidentally, this process usually recommended the addition of a strain from a species not previously included, so no additional steps had to be taken to ensure species diversity.

For the selection of the final “polydiverse” yoghurt, the diversity of the 127 yoghurts was evaluated using the two metrics outlined above. In addition, the strains were screened for known resistance genes and survival factors (Section 1.3.1) using OpenGenomeBrowser, and the results of this were taken into consideration.

As it turned out, one strain, FAM22081 from the species *Lactobacillus helveticus*, had a particularly strong effect on metabolomic diversity, possibly because of its known strong proteolytic activity. In fact, additional strains did not substantially increase the metabolomic diversity much further, which is the main reason why we decided not to add more than five strains.

The addition of five strains to the three starters brought the total number of strains in the yoghurt to eight, and even more strains would have increased the risk of malfermentation. Indeed, this already had an impact on production, as the fermentation temperature had to be decreased from 42 to 37 °C and the incubation time increased from 4-6 h to 16 h. For potential industrial production, this may necessitate fermenting the yoghurt directly in the container, in which it is sold, creating a solid yoghurt. While this would meet the current preferences of many consumers, it would also increase the manufacturing costs, which may make it necessary to target a higher price range in marketing. Alternatively, additional experiments could be performed to test whether similar metabolic diversity could be achieved under more established production conditions.

All yoghurts produced yielded edible yoghurts with yoghurt-like texture, with noticeable differences in acidity and taste. This suggests that perhaps, even more strains could be added without risking malfermentation. However, we did not monitor the growth of the added strains during fermentation, so it is possible that some strains did not grow well and had minimal impact on the final yoghurt metabolome. As a result, these strains may be present in such small amounts that they do not significantly affect the microbiota. In the mice trial, the final “polydiverse” yoghurt was compared to a yoghurt with only starter strains, and preliminary transcriptomics results in mice suggest that it has an inflammatory effect on the epithelial immune system, possibly causing proliferation of epithelial tissue.

The “polydiverse” yoghurt was a rather courageous strategy that goes against the trend of most research, which is generally focused on a narrow range of strains [202]. This could be because fermentation using fewer strains simplifies industrial production, but also because scientists prefer to isolate variables and make simpler experiments to better understand the biochemical mechanisms. Any effects of the “polydiverse” yoghurt are not readily explainable, as they may be caused by one of the strains in isolation, by the interaction of a subset of them, or as an emergent property of a more complex ecosystem, that furthermore interacts with the mouse organism and microbiome. However, such experiments may still generate interesting results that might not otherwise emerge. Moreover, the focus on isolated variables may have larger negative consequences. Focusing on what scientists can easily study and engineers can efficiently produce when creating food products may result in an excessive emphasis on micronutrients and processed foods. This, in turn, could lead to oversimplified and ultimately pseudo-scientific nutritional advice promoted by amateur nutritionists who are susceptible to catchy headlines and marketing tactics [203]. Similarly, a “polydiverse” yoghurt, should it ever come to market, may be a safe way to increase diversity in our food and train our immune system. However, customers should be aware of the bigger picture of nutritional advice, minimizing the risk that they might, ironically, reduce the overall diversity of their diet in favor of a more diverse yoghurt.

4.4.2. AhR-activating “indole” yoghurt

Numerous and structurally very different ligands can bind to the AhR (1.4.2). The effects the ligands trigger through binding were sometimes found to be contradictory, and for most of the ligands, the underlying mechanisms and factors of influence have not been fully clarified. Furthermore, most AhR ligands, let alone their binding profiles or associated genes are still unknown [137]. For these reasons, we decided to screen more strains than originally planned in an *in vitro* luciferase-based assay. The assay of 136 yoghurts proved to be difficult to analyze as the luminescent readout was weak, partly because of the complex and turbid yoghurt medium. Different strategies were tried to overcome this problem for subsequent assays, including increasing the incubation time, the addition of higher amounts of luciferin, and removal of milk protein using centrifugation. Ultimately, we successfully created a strongly AhR-activating yoghurt that, when fed as FMP to germ-free mice, had measurable effects on the immune system of the pups (see Manuscript 4, Section 3.4). Unfortunately, the original measurements of 136 yoghurts were not sufficiently precise to enable the discovery of novel tryptophan catabolizing genes or other genetic explanations later-on.

Nevertheless, the experiment was insightful. Studying the dataset, I noticed that the plate on which the samples were grouped introduced a clear bias. This had to be factored in during statistical analysis and illustrated the importance of quality controls in data analysis. Furthermore, the spectrometer that records the experiment stores the output into Excel files, each file containing the measurements of one 96-well-plate. In the experiment with 136 yoghurts, this resulted in nine files as each yoghurt was measured in triplicate and each plate contained additional controls. In these files, the data is represented as a two-dimensional array, reflecting the layout of the samples on the plate. The process of correctly reconstructing which sample belongs to which well, and combining and analyzing these nine files within Excel is tedious and error-prone. I simplified this process programmatically, resulting in a single table which is less likely to contain errors and can more easily be analyzed. Overall, to me, the experiment illustrated the dependence of data science on the data itself, the need for close collaboration between experimentalists and data scientists as well as the importance of automatization.

It was thrilling to discover that our AhR-activating FMP was successful in inducing the desired effect in the pups of germ-free mice. The results of human trials, which are currently being planned, promise to be even more exciting. However, given the complex and diverse actions of AhR, it will be crucial to carefully select appropriate indicators of health and conduct thorough safety testing of the yoghurt in subjects with intestinal issues before it can be recommended to any individual.

4.4.3. Folate yoghurt

Compared to selecting strains to increase indole metabolites in yoghurt, selecting strains that produce folate was relatively straightforward as the genes and pathways responsible for folate biosynthesis are well understood. However, for many strains belonging to the same species, the genes linked to the folate pathway were often found to be identical. Moreover, folate has many roles and is used in the metabolism of all strains. Thus, properties like kinetics of synthesis, catabolism, and transport across membranes may be key to predicting folate production, but they cannot be determined from genomic data alone. Additionally, the *S. thermophilus* strains from the starter culture themselves have all the necessary genes to synthesize folate, so another possible mechanism to increase folate production is through inter-strain interactions, which currently are impossible to predict.

The folate production of 163 yoghurts fermented with one additional strain was measured using a microbiological assay (Section 1.3.3.1). Milk and yoghurt made from starter strains only contained very similar amounts of folate. Interestingly, most of the yoghurts with additional strains showed at least slightly elevated amounts of folate. The amounts of folate in eight yoghurts were increased at least threefold, with three of them exhibiting a fivefold increase.

Four of these eight yoghurts, including the best two, were made with a strain from the subspecies *Lactococcus lactis subsp. lactis*. This is not particularly surprising, as they have all the genes necessary to synthesize and combine the two precursors. However, while all have the genes to combine the two precursors, none of the other four strains appear to have the genes to synthesize the precursor *p*-aminobenzoic acid, and one of them, a *Lacticaseibacillus paracasei*, even lacks the genes for the other precursor, 2-amino-4-hydroxy-pteridine. This illustrates the limitations of gene-based strain selection and justifies our more cautious approach, which takes additional factors like phylogenetic relatedness into account (Section 4.2) and focuses more on phenotypic screening than originally planned.

However, we were not able to determine the genetic cause of these findings. Since in the case of folate, many of the relevant genes are known, we could in principle have focused on finding causative SNVs. However, the dataset does not provide enough statistical power to make a strong association:

there are not enough samples, particularly strong folate-producers, and there is too much genetic variation. One way to find an explanation would be to focus on one species, for instance *L. l. subsp. lactis*, reducing genetic variation, and measuring the folate production of more strains, in combination with measurements of gene expression.

Meanwhile, as these yoghurts showed absorption beyond the linear range of the standard curve, at present, we were unable to determine the absolute amounts of folate and conclusively evaluate the significance of the amounts produced. As mentioned in the introduction (Section 1.3.3.1), a concentration of 200-400 $\mu\text{g/l}$ would make yoghurt a “good” source of folate. Given that milk contains approximately 20-50 $\mu\text{g/l}$ folate and in our experiment, the yoghurt made from starters only was very similar to milk, the fivefold increase in our best yoghurts indicates that they contain approximately 100-250 $\mu\text{g/l}$. Finally, the data from the mice trials have yet to be fully evaluated.

4.5. Scoary2

As mentioned in Section 4.2, the metabolomes of all yoghurts were measured to use this information for the creation of a maximally metabolically diverse yoghurt. This dataset comprises 183 yoghurts, each made from a different strain in addition to starter culture, and consists of 3,889 metabolites, 2,348 detected using liquid chromatography MS (LC-MS) and 1,541 using gas chromatography MS (GC-MS) of volatile compounds. Linking the genetic information about the strains to the metabolites in this dataset, as demanded by the third aim of this thesis (see Section 2), posed various challenges.

Within the framework of this objective, the first challenge was to develop the right research question and to set the focus accordingly. Established methods inspired by disease subtyping could have been used to discover robust metabolomic biomarkers and genetic indicators of major metabolome clusters (Section 1.4.1). However, the yoghurts clustered based on taxonomy (see Figure 3 in Manuscript 3, Section 3.3). On one hand, this is a success for the project, as it demonstrates that the genetic potential of the strains can successfully be translated into the metabolomic profile of the yoghurts, despite the combination with three other starter strains. On the other hand, it makes linking genes and metabolites more difficult. Established methods could be used to find metabolic and genetic biomarkers that robustly differentiate clusters, i.e. species, but the resulting metabolites could merely help distinguish them from each other. This is already a mature, established field (microbial fingerprinting) [204]. The resulting genes would be similarly uninteresting, as better methods exist for classifying genetic material [40].

There were other challenges related to the properties of the dataset. The second challenge was the inability of many existing tools to process binary gene presence-absence data, as they are designed to handle more powerful continuous transcriptomic data, which was not collected as part of this project. Next, the chosen strains are taxonomically very diverse, belonging to 18 different species and two different classes (Actinomycetia and Bacilli). Thus, the third challenge was the curse of dimensionality: the dataset comprises a vast number of *features*, i.e. genetic differences that could explain a given phenotype (9,051 orthogroups and far more SNPs or *k*-mers), and only 183 *samples* (yoghurts made from different strains). The fourth, also caused by this taxonomic diversity, was that the genetic differences, as well as the metabolites, are strongly correlated, meaning that population structure is a major confounding factor. Moreover, the dataset is unbalanced, with most strains belonging to only four species, representing the fifth challenge. The sixth was that some species have been studied more extensively than others, resulting in the fact that the fraction of known genes differs considerably between species (see Figure 1 in Manuscript 1, Section 3.1). As a seventh challenge, only 48 metabolites could be confidently identified using LC-MS reference standards.

Eventually, we concluded that finding causative links between metabolites and genetic variation would be the goal and that the most appropriate approach for this are mGWAS methods because

they can deal with population structure. Because of the vast number of genetic differences, it was clear that the analysis would be focused on the more granular presence-absence of orthologs rather than k -mers, unitigs, or SNPs. We could not find similar datasets or methods that could be directly applied to our dataset in the literature as existing mGWAS tools were designed for few traits and not optimized for speed or efficient post-GWAS data exploration. Therefore, the final, eighth challenge was to adapt mGWAS to large omics datasets.

The development of Scoary2 thus represents the first solution that enables the study of large phenotypic datasets using mGWAS. In addition, its data exploration app will also be useful for conventional, single-phenotype mGWAS workflows. While Scoary2 can successfully find real associations between phenotypes and genetic variation, it has limitations that represent opportunities for future development. The pairwise comparisons algorithm sacrifices statistical power by focusing exclusively on evolutionary transitions and by only being able to deal with binary phenotypes. Next, the two p -values generated by Scoary2 are not easy to interpret. On the second page of Scoary2's data exploration app (*trait.html*), genes that are in close proximity to each other on most genomes could be highlighted as potential gene clusters. While the dendrogram of metabolites on the first page of the data exploration app (*overview.html*) constitutes a way of combining several phenotypes in the analysis, metabolite set enrichment analysis (MSEA) [205, 206] or a similar method could be used to make more predictions about the metabolites in each cluster. Finally, to make Scoary2 even more user-friendly, it could be integrated into OpenGenomeBrowser or combined with a gene prediction and an ortholog inference software, or a k -mer or unitigs caller, to create a complete pipeline.

Scoary2 and possible successors could lead to a new approach of studying microbial genomes. The metabolomes and genomes of strains could be measured in high throughput to discover metabolomic variety with genetic explanations, in an experiment that is independent of and not biased by existing knowledge. This way, metabolites and genes that have so far been uncharacterized may be found in a comparably inexpensive and efficient manner. Existing knowledge about metabolites may be used to derive the function of the genes, and vice versa. Concomitantly, similar datasets may emerge in the context of different experiments and could be further exploited using this approach.

Scoary2 was presented only once outside the University of Bern, by Ola Brynildsrud, author of the original Scoary software and co-author of Scoary2, at the prokaryotic genome evolution network meeting of the European Society for Evolutionary Biology at the Milner Center of Evolution in Bath, UK. According to him, Scoary2 “received a very positive response,” “people were particularly excited about the JavaScript output and all the possibilities with OpenGenomeBrowser,” and “a lot of people [said] they want to start using it right away.”

4.6. Polyfermenthealth

The Polyfermenthealth project was ambitious, broad, and interdisciplinary. Even apart from the questions that were deliberately left open in the original grant proposal, it developed in unforeseen directions and required difficult strategic decisions, such as determining the best strategies for the yoghurts, how to analyze the metabolomics dataset and the prioritization of OpenGenomeBrowser. Because the project is so broad, encompassing the fields of nutrition, dairy science, microbial genomics, immunology, and microbiota, it was challenging to fully comprehend the available methods and their limitations. This breadth also made the communication of technicalities and management of expectations across fields difficult. On the other hand, this also offered a great learning experience for me and my colleagues, leading to new ideas for subsequent projects.

Regarding the first and second goals of the thesis, the use of comparative genomics to determine promising strategies for functional foods and strain selection, my impression is that the original expectations were not fully met. From the start, we wanted to target the indole pathway, and while

we did identify the folate pathway as an interesting target during screening, it was already mentioned in the original proposal as a candidate. In addition, the fact that the most relevant tryptophan catabolizing genes are unknown presented an unavoidable obstacle for strain selection for this yoghurt. In the case of the folate yoghurt, the homogenous distribution of folate-related genes within species meant that strain selection could not be done entirely rationally, purely based on known genes. It is therefore slightly ironic that the software that came out of the screening and selection process is arguably the most impactful achievement of this thesis. Nevertheless, because we realized the limitations of gene-based selection early enough, we incorporated more conservative indicators such as phylogenetic relatedness, gene variants, and gene copy numbers, and relied more heavily on phenotypic screening. Ultimately, this approach was successful, and the AhR-activating yoghurt had a measurable impact not only *in vitro* but also *in vivo* in mice. Furthermore, we managed to find strains that were able to increase the amount of folate in yoghurt fivefold.

Future genome-based yoghurt design will be dramatically easier thanks to OpenGenomeBrowser, as it automates many bioinformatics steps employed during this project, enabling biological experts to explore the data by themselves, even without significant bioinformatics skills. This may be a key advantage, as accessing the data indirectly through a bioinformatician slows down the process of data exploration, and the bioinformatician may miss interesting details because of their lack of domain-specific knowledge. One such project is already being planned at Agroscope, focused on increasing Vitamin K2 production in yoghurt.

The third aim of the thesis, linking the genotype to the phenotype, was not clearly defined at the start of the project. Holistic GSMM-based approaches faced too many obstacles, summarized in Section 4.2.1, and we were unable to find a genetic explanation for increased AhR activation or folate production. The main problem is the genome sequence alone is of limited use: the function of many genes (let alone gene variants) is not known, and it is not possible to predict their expression or regulation. The chosen strategy, to extend and apply the Scoary mGWAS algorithm to our metabolomics dataset, however, was successful. Scoary2 enabled the discovery of a gene cluster that, if complete, enables *Propionibacterium freudenreichii* to metabolize carnitine. This approach may constitute a new approach to design yoghurts and other fermented foods. Instead of designating a strategic target beforehand, the metabolomes of many strains might be measured, interesting differences could be identified using Scoary2, and promising ones further investigated. For instance, the ability to metabolize carnitine could enable the strains to survive better in the gut, as it is an anaerobic environment where redox capabilities are very important [77]. In addition, carnitine, and its metabolism themselves also play a role in human health.

Overall, the Polyfermenthealth project succeeded in its core goal, the creation of yoghurts with potential health benefits. Of course, these must be validated in well-designed human trials, as not all interventions have the intended effect. For instance, in one experiment, a multi-strain probiotic actually impaired the recovery of human microbiomes after antibiotics treatment [207]. Before any health claim, large and robust preregistered studies would be required for a confident, justified recommendation and fair media advertisement according to the definition of the word probiotic: "Live microorganisms which when administered in adequate amounts confer a health benefit on the host." In the fields of microbiota as well as probiotics, limited findings are being generalized and oversold, something that may well be detrimental to these fields as well as to public trust in science more generally [81, 208].

4.7. Bioinformatics software development

The reception to OpenGenomeBrowser and Scoary2 by the community has been overwhelmingly positive, highlighting the need for user-friendly and efficiency-boosting tools, attributes that have arguably been neglected in bioinformatics software development. I am still surprised and do not fully understand why a database-independent comparative genome

management software, like OpenGenomeBrowser, had not been created earlier. From the point of view of a computer scientist, the key building blocks have long existed: PostgreSQL was first released in 1996, Biopython in 2000, Nginx in 2004, the Django framework in 2005, Docker in 2013, and Docker Compose in 2014.

One part of the explanation may lay in the extraordinary breadth of bioinformatics, which is very interdisciplinary, has many different applications in most fields in biology, and may require knowledge in statistics, algorithmics, and computer science. Most bioinformaticians have strong roots in one of these fields and may not feel fully at home with others. The creation of OpenGenomeBrowser required a good idea of what biologists need and what can be automated bioinformatically, things that computer scientists may not know. On the other hand, it also required an overview of and an affinity towards software development, because the project required the creation of a front-end using HTML and JavaScript, a back-end programmed in Python that interacts with a database using SQL, and containerization using Docker. While these things may come naturally to a computer scientist, none of them are ever mentioned in a biology course and not all are covered by the bioinformatics curriculum.

Another reason may be structural and cultural. Despite being a critical part of research infrastructure, there is generally poor funding in academia for software development and maintenance. Software development is often funded through grants, where it is described as research focused on biological discovery, and funding software maintenance is even more difficult. Journals may also contribute to this problem, as they may insist on novel biological insights accompanying new software and are less likely to publish updates to existing software [209]. Ironically, sustained software development appears to be the strongest indicator of high-quality bioinformatics software [210].

Thus, too often, tools are developed to solve a concrete problem and published in a less-than-ideal state. During my PhD, I have encountered problems with bioinformatics tools that are far less common with software developed by computer scientists. There are difficulties with installation, documentation, user-friendliness, maintenance, and even code availability. In effect, the cost of fixing these issues is transferred to the users of the software and amplified considerably, a phenomenon known as technical debt in computer science. While it may take time and experience, it is interesting, educational and may lead to higher standards that would benefit everyone. In my opinion, such optimizations should be insisted upon by editors and reviewers. As an incentive for researchers, it should be widely known that these factors are likely an important predictor of the citation-based metrics of a paper [210].

The development of OpenGenomeBrowser illustrates this very well. I would be surprised if the code that I developed to select strains (Section 4.2) would have been useful to anyone else had it been published. In contrast, OpenGenomeBrowser appears to be of relatively broad utility. Moreover, more time for development enabled the creation of isolated modules which can be used outside of OpenGenomeBrowser. This makes development and particularly code re-use easier. For instance, the flower plot is externalized as a Python module with minimal dependencies [211], the functionality of the gene comparison tool is also mostly contained in a separated module [212] and has been used as such to create locus plot illustrations for at least one paper [213], and the pathways tool is available as its own application [214] comprising the code necessary to convert KEGG maps into smart vector graphics and a minimal JavaScript library to make to make them interactive in a browser.

5. Appendix

This chapter consists of two publications, one not related to the main subject of this thesis and a paper for a kid's journal.

Besides work on Polyfermenthealth, Scoary2 and OpenGenomeBrowser, I had the pleasure to collaborate with Dr. Ana Belén García-Martín on her hybrid sequencing and comparative genomics project of *Brachyspira hyodysenteriae* isolates, which resulted in an additional research paper (Manuscript 5, Section 5.1).

To communicate our research to the public, as intended by the Gebert Rűf foundation that funded the Polyfermenthealth project, we decided to reformulate the content of our "position paper" (Manuscript 1, Section 3.1) for the *Frontiers for Young Minds* journal aimed at aged 8-11-year-olds. The article was published in the collection "New ways to understand how foods affect me and my health!" (Manuscript 6, Section 5.2).

5.1. Manuscript 5: Comparative genomics of *Brachyspira hyodysenteriae*

Whole-genome analyses reveal a novel prophage and cgSNPs-derived sublineages of *Brachyspira hyodysenteriae* ST196

Ana Belén García-Martín, Thomas Roder, Sarah Schmitt, Friederike Zeeh, Rémy Bruggmann, Vincent Perreten

Status:

Published in BMC Genomics (2022), 23(1):131 [215]

Statement of contribution:

VP was in charge of funding acquisition, project supervision and coordination and first revision of the draft manuscript. VP and ABGM designed the study. RB co-supervised, provided super-computation resources and revised the draft manuscript. ABGM performed laboratory procedures, data acquisition and curation, all bioinformatics analyses, created images, wrote the draft manuscript and edited it with inputs from all authors. TR was responsible of installing and maintaining the NCBI-PGAP Singularity container, wrote the custom Python scripts and revised the draft manuscript. SS provided bacterial isolates (BHZ isolates) and epidemiological information and revised the draft manuscript. FZ provided bacterial isolates (B 114_09C, B 115_02A and Bh743-7 isolates) and epidemiological information and revised the draft manuscript. All authors read and approved the final manuscript.

RESEARCH

Open Access



Whole-genome analyses reveal a novel prophage and cgSNPs-derived sublineages of *Brachyspira hyodysenteriae* ST196

Ana Belén García-Martín^{1,2} , Thomas Roder^{2,3} , Sarah Schmitt⁴, Friederike Zeeh⁵, Rémy Bruggmann³ and Vincent Perreten^{1,6*}

Abstract

Background: *Brachyspira (B.) hyodysenteriae* is a fastidious anaerobe spirochete that can cause swine dysentery, a severe mucohaemorrhagic colitis that affects pig production and animal welfare worldwide. In Switzerland, the population of *B. hyodysenteriae* is characterized by the predominance of macrolide-lincosamide-resistant *B. hyodysenteriae* isolates of sequence type (ST) ST196, prompting us to obtain deeper insights into the genomic structure and variability of ST196 using pangenome and whole genome variant analyses.

Results: The draft genome of 14 *B. hyodysenteriae* isolates of ST196, sampled during a 7-year period from geographically distant pig herds, was obtained by whole-genome sequencing (WGS) and compared to the complete genome of the *B. hyodysenteriae* isolate Bh743-7 of ST196 used as reference. Variability results revealed the existence of 30 to 52 single nucleotide polymorphisms (SNPs), resulting in eight sublineages of ST196. The pangenome analysis led to the identification of a novel prophage, *pphBhCH20*, of the *Siphoviridae* family in a single isolate of ST196, which suggests that horizontal gene transfer events may drive changes in genomic structure.

Conclusions: This study contributes to the catalogue of publicly available genomes and provides relevant bioinformatic tools and information for further comparative genomic analyses for *B. hyodysenteriae*. It reveals that Swiss *B. hyodysenteriae* isolates of the same ST may have evolved independently over time by point mutations and acquisition of larger genetic elements. In line with this, the third type of mobile genetic element described so far in *B. hyodysenteriae*, the novel prophage *pphBhCH20*, has been identified in a single isolate of *B. hyodysenteriae* of ST196.

Keywords: Bioinformatics, Horizontal-gene transfer, Pangenome, Structural variations, Singletons, Swine dysentery, WGS

Background

Brachyspira (B.) hyodysenteriae is a fastidious anaerobe spirochete that can cause swine dysentery (SD), a severe mucohaemorrhagic colitis that affects pig farm industry and animal welfare worldwide [1]. In the last years, SD has been controlled by using i.a. antimicrobial

agents of the pleuromutilin, macrolide, lincosamide and tetracycline classes, which has resulted in the selection and emergence of multidrug-resistant *B. hyodysenteriae* strains in many pig producing countries, including Switzerland [1]. Consequently, and considering the reduced arsenal of antimicrobial agents that are authorized and effective against *B. hyodysenteriae*, the treatment and control of SD have turned into new challenges.

As in other bacteria, the *B. hyodysenteriae* genome is evolving through mutations and recombination, paving the way for the formation of new genetic lineages that

*Correspondence: vincent.perreten@vetsuisse.unibe.ch

⁶ Institute of Veterinary Bacteriology, University of Bern, Länggassstrasse 122, CH-3012 Bern, Switzerland

Full list of author information is available at the end of the article



© The Author(s) 2022. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

might have acquired certain advantages for virulence and environmental adaptation like surviving antimicrobial exposure. Specific mutations in hemolysin genes have been shown to be associated with weak and strong hemolytic *B. hyodysenteriae* strains [2]. In addition, differential transcriptional patterns underlying different hemolytic phenotypes in *B. hyodysenteriae* have been reported [3].

Concerning antimicrobial resistance in *B. hyodysenteriae*, resistance to ribosomal-targeting drugs of the macrolide, lincosamide and tetracycline classes has been linked to the presence of single point mutations on the 23 S rRNA and 16 S rRNA, respectively [4–8].

Recently, gene transfer into the genome of *B. hyodysenteriae* has been also shown to contribute to antimicrobial resistance. So far, two acquired antimicrobial resistance genes, the lincosamide resistance gene *lnu(C)* and the tiamulin-valnemulin resistance gene *tva(A)*, have been identified in *B. hyodysenteriae* [4, 5, 9, 10]. These findings indicate that *B. hyodysenteriae* can acquire antimicrobial resistance genes, such as the *lnu(C)*, associated to the transposon MnT_{Sag1} originally found in *Streptococcus agalactiae* by horizontal gene transfer (HGT) mediated by mobile genetic elements (MGEs) [11]. To date, a single MGE, i.e. the defective prophage VSH-1 of *B. hyodysenteriae*, has been shown to mediate intraspecific HGT in *in vitro* experiments, indicating that such gene transfer agent may also play a role in gene acquisition in *B. hyodysenteriae* [12, 13]. MGEs can contribute to the acquisition of elements conveying advantages associated with antimicrobial resistance, virulence and environmental adaptation that can be further fixed and spread by clonal expansion [14, 15].

A powerful comparative approach to detect acquired novel MGEs, acquired antimicrobial resistance genes and putative virulence factors is the pangenome analysis [16, 17]. At higher resolution, single nucleotide polymorphisms (SNPs) are usually analysed to understand their contribution to the expansion of both specific clonal lineages and sublineages. In this line, studies focused on bacterial epidemiology, genetic diversity and population structure are frequently based on non-recombinant core genome SNPs (cgSNPs) [18, 19].

In Switzerland, over nearly the last decade, SD has been caused mainly by a specific predominant macrolide-lincosamide-resistant *B. hyodysenteriae* belonging to sequence type ST196 [20], only reported in Swiss pig herds so far [4, 5, 20]. This fact prompted us to perform whole-genome sequencing (WGS), pangenome and SNP analyses to get deeper insights into the genome structure and variability of different ST196 isolates. Our findings shed light on the genetic diversity of *B. hyodysenteriae* ST196 and revealed the presence of the novel prophage

pphBhCH20 in a single isolate of *B. hyodysenteriae* of ST196.

Methods

Isolates information, bacterial culture and DNA extraction

Fourteen isolates of *B. hyodysenteriae* of ST196, isolated from geographically distant Swiss pig herds without known epidemiological links (Additional Fig. S1) between 2010 and 2016, were sequenced in this study and compared to the closed genome of the *B. hyodysenteriae* isolate Bh743-7 (GenBank accession numbers CP046932 (chromosome), CP046933 (plasmid)) of ST196 isolated in 2017 and used as reference [21]. All 15 isolates included in this study, except two (isolates B 114_09C and B 115_02A), were obtained from pigs with SD [20]. All except one, contained the A2058T mutation in the 23 S rRNA associated with the macrolide-lincosamide resistance phenotype [20]. None of the isolates harbored any known acquired antimicrobial resistance genes such as the *lnu(C)* and *tva(A)* [20]. High-quality genomic DNA of *B. hyodysenteriae* was extracted from the bacterial lawn superficially grown on trypticase soy agar plates containing 5% (v/v) sheep blood (TSA-SB, Becton Dickinson), using a DNeasy[®] Blood & Tissue kit (Qiagen) following the manufacturer's instructions. For each sample, the bacterial lawn of at least two TSA-SB plates were collected using a 10 µL plastic loop and resuspended in 300 µL of resuspension buffer three times to wash away remnants of material from the agar plates, before continuing with the protocol. All DNA samples were RNase (20 mg mL⁻¹) treated for 60 min at 37 °C, purified using AMPure[®] XP magnetic beads (Beckmann Coulter) and quantified using a Qubit 3.0 fluorometer (Life Technologies).

Genome sequencing, assembly and annotation

Standard genomic libraries, containing unique dual indexes, were prepared from genomic DNA obtained from all *B. hyodysenteriae* isolates of ST196, including the isolate Bh743-7 for which we had previously generated its complete genome by Oxford Nanopore Technologies sequencing and hybrid assembly (CP046932 and CP046933) [21]. The libraries were sequenced using the Illumina[®] HiSeq platform (Eurofins Genomics GmbH, Germany) in the sequencing mode NovaSeq[™] 6000 2-PE 2 × 150 bp. Short reads were checked for quality using FastQC v0.11.7 [22] and quality control output files were combined into a single report using MultiQC v1.8 [23]. Illumina adapters, nucleotides at both ends with an average Phred score < 15 over a 4 bp sliding window, reads shorter than 36 bp and low quality bases (average Phred score < 33) were removed using Trimmomatic v0.36 [24].

Illumina paired-end short reads were assembled into contigs using the multi-kmer *de Bruijn* graph-based assembler SPAdes v3.12 in the *--careful* mode [25]. All draft genomes were filtered for contigs larger than 500 bp with depth coverage above 100X, using a custom Python script (Additional file 1), and their quality was assessed using QUAST v4.6.0 [26]. Genome annotation was done locally using the NCBI-PGAP pipeline v4.12 [27].

SNP variants analysis of *Brachyspira hyodysenteriae* of ST196

Core genome single nucleotide polymorphisms (cgSNPs) were called using Snippy v4.5 [28], with default parameters, providing the complete genome of the *B. hyodysenteriae* isolate Bh743-7 as a reference genome for WGS alignment. A phylogenetic tree, based on non-recombinant cgSNPs filtered using Gubbins with default parameters [29], was constructed with FastTree [30], and visualized and edited with iTOL v5.7 [31]. The non-recombinant cgSNP alignment was converted into a pairwise distant matrix using snp-dists v.0.7.0 (<https://github.com/tseemann/snp-dists>). A complementary heatmap displaying distances across genomes was generated with the R package “gplots” (<https://github.com/talgalili/gplots>). The functional effect of SNPs was investigated by SnpEff v4.3T [32], via Snippy. A custom Python parser script (Additional file 2) was used to combine and extract the information relative to all SNPs that were common to all ST196 isolates.

Pangenome analysis of *Brachyspira* spp.

Similarity across genomes of *B. hyodysenteriae* of ST196 was also analysed at nucleotide sequence level by calculating whole-genome average nucleotide identity (ANI) scores, using the Python module PyANI v0.2.0 [33] via Anvi'o v6.2 [34].

The pangenome analysis of all *B. hyodysenteriae* ST196 genomes, extended to a total of 90 genomes, including those corresponding to other *B. hyodysenteriae* STs (ST6, ST66 and ST197) circulating in Switzerland and also those belonging to other *Brachyspira* species, was computed using Anvi'o v6.2 [34].

GenBank-formatted public genomes and linked meta-data (Additional file 3) were downloaded and processed following a Snakemake [35] workflow in Anvi'o (<http://merenlab.org/2019/03/14/ncbi-genome-download-magic/>). For downstream analyses, the unpublished NCBI-PGAP annotation files corresponding to the complete genome of *B. hyodysenteriae* isolate Bh743-7 and draft genomes of fourteen additional isolates of ST196 were reformatted prior importation into Anvi'o, using the Bioinformatics Tools (Bit) package v1.4.71 (https://github.com/AstroBioMike/bioinf_tools) [36]. Additional

structural and functional annotations for each genome were done using Prodigal [37] and the Cluster of Orthologous Groups (COGs) database [38], respectively, as part of the Anvi'o workflow. A contigs database containing information about contig number, sequence composition, structural and functional annotation was generated for each genome and provided to Anvi'o to generate a genome-storage database using the *--external-genome* flag. Next, a pangenome analysis was computed using the *anvi-pan* program with parameters *--min-bit* 0.5 (default) and *--mcl-inflation* 10 (recommended for genetically closely related genomes), as in [39]. Genomes were organized based on shared gene clusters using Euclidean distances and Ward linkage, the number of genomes that contributed to each gene cluster (*number of genomes has hits*) and, when required, by forcing synteny. Genes clusters were grouped into bins containing core genes (common to all the isolates), soft-core genes (could be present/absent) and singletons (only present in a single genome), and saved as a default collection for subsequent summary analysis. In addition, a homogeneity index, which takes 1 as highest value and provides an idea of shared sequence identity, was obtained after calculating both functional (amino acid residue conservation without considering sequence gaps) and geometric (sequence gaps and amino acid residue patterns) indexes using Anvi'o (see “An Anvi'o workflow for microbial pangenomics – Meren Lab”). Details on programs and parameters are presented in the additional file 4. Hemolysin genes obtained from the pangenome analysis were screened for presence of mutations described previously [2] by multiple sequence alignment using Clustal Omega and the genome of the *B. hyodysenteriae* WA1 strain (NC_012225.1) as a reference.

Prophage Hunter analysis

The complete chromosome of *B. hyodysenteriae* isolate Bh743-7 of ST196 (CP046932) was interrogated using the Prophage Hunter web server (<https://pro-hunter.genomics.cn/>) [40], in order to detect prophage elements by similarity comparison to elements already deposited in a reference database. The probability of a predicted prophage being active was provided by the activity score ranging from 0 to 1.

Basic alignment, schematic gene map representation and visualization

Basic alignment visualization of the annotated prophage elements and genomic context analyses were based on the application of the progressive algorithm MAUVE v2.4.0 [41]. Comparison of phage-like regions and schematic gene map were done using Easyfig v2.0 [42]. Final figures were edited using the open-source vector graphics editor Inkscape v1.0 (<https://inkscape.org/>).

Results

Genome sequencing, assembly and annotation

The de novo draft assemblies of all isolates of ST196 were obtained through assembly of Illumina paired-end reads using SPAdes v3.12 (Additional Table S1). These assemblies were characterized by a low G+C content (~27.1%) and a variable number of contigs (Additional Table S1). They contained between 25 and 36 contigs larger than 500 bp and with a minimum coverage of 100X. The N50 metric was above 292,250 bp for all assemblies. Assembly lengths ranged from ~3.01 Mbp (BHZ333) up to ~3.07 Mbp (Bh743-7). The average assembly size was 3.04 Mbp reflecting a high quality in terms of completion compared to the complete *B. hyodysenteriae* genomes of the strains WA1 (CP001357.1), B-78^T (NZ_CP015910.2), BH718 (CP019600.1) and Bh743-7 (CP046932 and CP046933) publicly available. On average, 2605 total CDS, 2583 coding genes and 40 rRNAs were obtained. Contigs containing plasmid-encoded genes were found in 12 of the 14 additional ST196 isolates using BlastN. Their size ranged between 31,271 and 32,625 bp and their G+C content varied between 22.3% and 22.5%. All draft genomes, including the one without plasmid and the one from which the plasmid could not be reconstructed from a single contig, were further analysed.

Genetic diversity across closely related genomes of *Brachyspira hyodysenteriae* of ST196

According to the Snippy analysis, on average, the core genome alignment had a length of ~3.02 Mbp (STDEV = 8912 bp; minimum (2,985,653 bp, 96.77% of aligned bases) and maximum alignment (3,020,968 bp, 97.92% of aligned bases)). A total of 153 different non-recombinant cgSNPs were detected from the 15 isolates of *B. hyodysenteriae* of ST196. Pairwise comparisons revealed that the isolates differed by at least one and at most 52 SNPs, generating eight phylogenetic sublineages (I - VIII) (Fig. 1 A and 1 B). These sublineages consisted of either clusters of highly related isolates or singletons differing from the *B. hyodysenteriae* reference isolate Bh743-7 by a minimum of 30 and up to 52 cgSNPs (Fig. 1 B). The singletons (I - IV) consisted of four isolates which were all obtained from different herds in different years (Fig. S1). With respect to the reference isolate Bh743-7 (from 2017), the isolate BHZ26 (2010) differed by 30 SNPs, and

the isolate BHZ153 (2011) by 34 SNPs. The most divergent isolate (BHZ231), sampled in 2012, harboured 52 different SNPs. The average SNP distance calculated for the isolates contained in clusters V-VIII was 41 SNPs (with a minimum of 31 SNPs and a maximum of 50 SNPs) (Fig. 1 A). Cluster V contained five isolates differing by a minimum of 43 SNPs and a maximum of 50 SNPs (average SNP distance of 45 SNPs). Three of them (BHZ660, BHZ684, BHZ695) were highly related (one to two SNPs, Fig. 1 B) and were isolated in 2014 from three different pig herds. The other two (BHZ777, BHZ819), differing by 10 SNPs, were sampled from two different pig herds in 2015, and differed by a minimum of three and up to eight SNPs from the previous isolates (Fig. 1 B and Fig. S1). Each of the three remaining clusters (VI - VIII) contained two isolates sampled in different years from several herds. The isolates of these clusters differed by a minimum of 31 SNPs and a maximum of 44 SNPs (average SNP distance of 37 SNPs) compared to the reference genome (Fig. 1 A and 1 B). Overall, sublineages conformation was not associated neither by year nor region of isolation.

Regarding the variants, 26 were found in all ST196 that were compared to the reference genome (Additional file 5). Of these, 23 were found in protein-coding sequences and the remaining three in non-coding sequences. From the 23 SNPs in protein-coding sequences, 20 were translated into non-synonymous including one classified as frameshift, three as stop-lost, and six as missense. The missense SNPs were found in genes involved in chemotaxis, transport and metabolism of ions, carbohydrate and inorganic substrates, cell wall/membrane/envelope biogenesis, signal transduction, transcription and translation. In addition, carbon starvation proteins and different enzymes, involved in general metabolism pathways, were represented among the variants-containing genes (Additional file 5).

High conservation of the core genome in *Brachyspira hyodysenteriae*

As shown by the previous cgSNPs and Anvi'o pangenome analyses, the genomes of the ST196 isolates mainly differed from each other by SNPs, being nearly identical throughout their entire lengths (ANiB scores above 99%) (Fig. 2 A). However, lower ANiB scores (97 and 98%)

(See figure on next page.)

Fig. 1 Core-genome SNP-based phylogeny of *Brachyspira hyodysenteriae* isolates of ST196. **A** The relationship among isolates of ST196 are shown according to the approximately-maximum-likelihood phylogenetic tree. Labels containing names of the isolates and regions (from Western to Eastern Switzerland: R1 to R4) of isolation are colored according to the years of isolation. Sublineages are indicated with roman numerals. Numbers of different non-recombinant cgSNPs identified respect to the reference genome are indicated for each isolate. Average (Avg) SNPs distance is also indicated for clusters V, VI, VII and VIII. **B** Isolates are clustered according to the distance matrix of pairwise differences calculated from the non-recombinant cgSNPs. Number of different cgSNPs used to identify genetic distances across all genomes are indicated in each cell of the heatmap

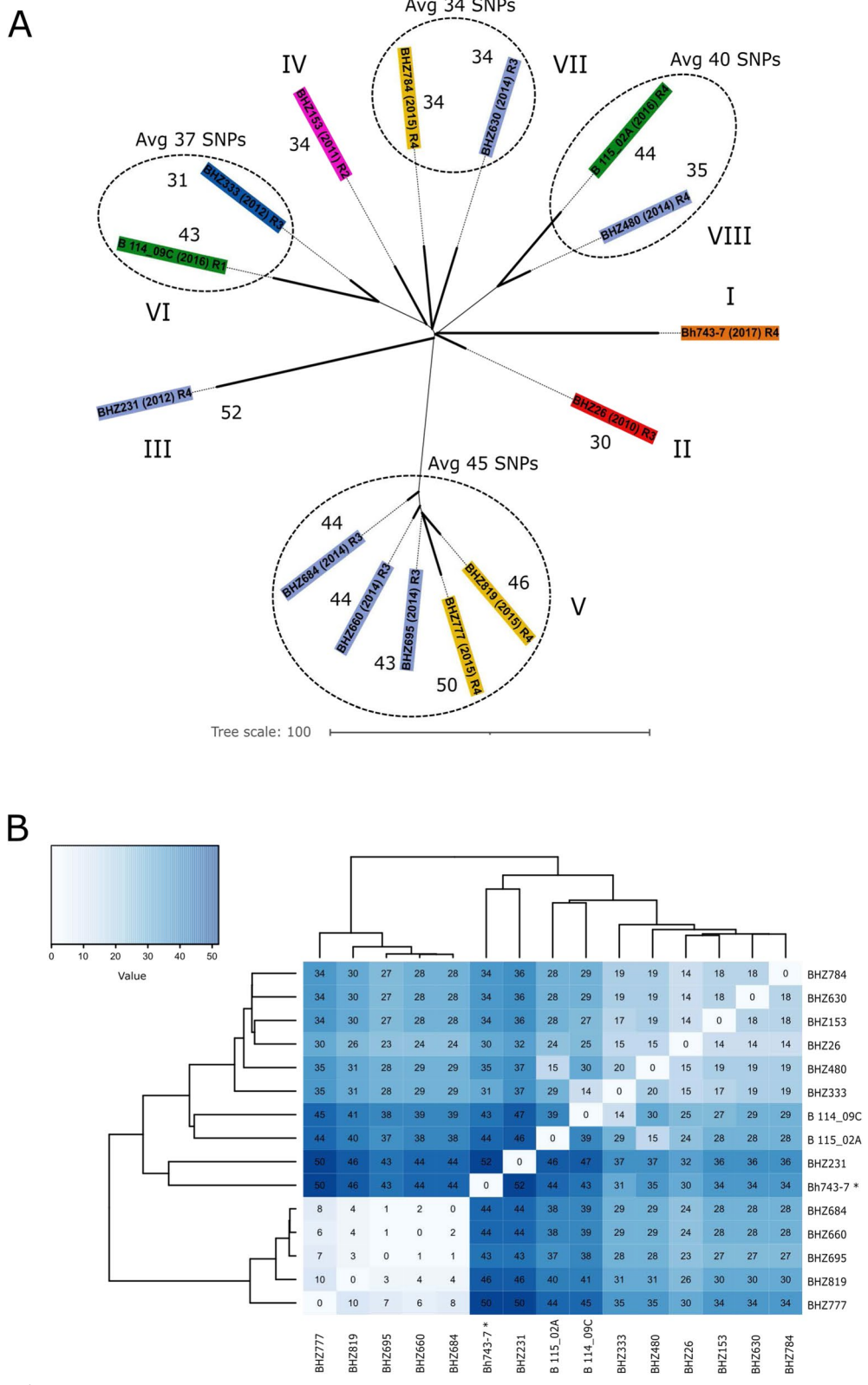
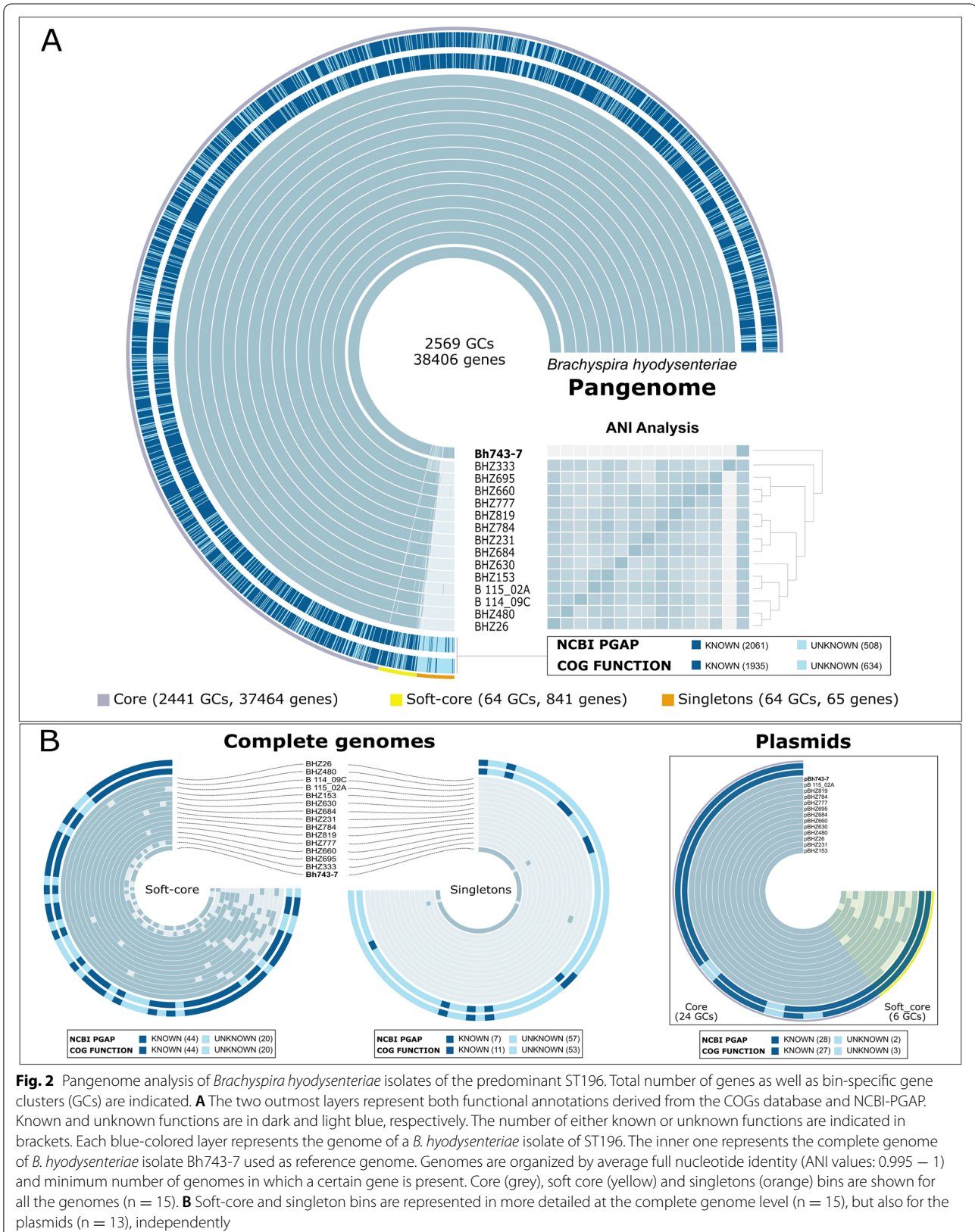


Fig. 1 (See legend on previous page.)



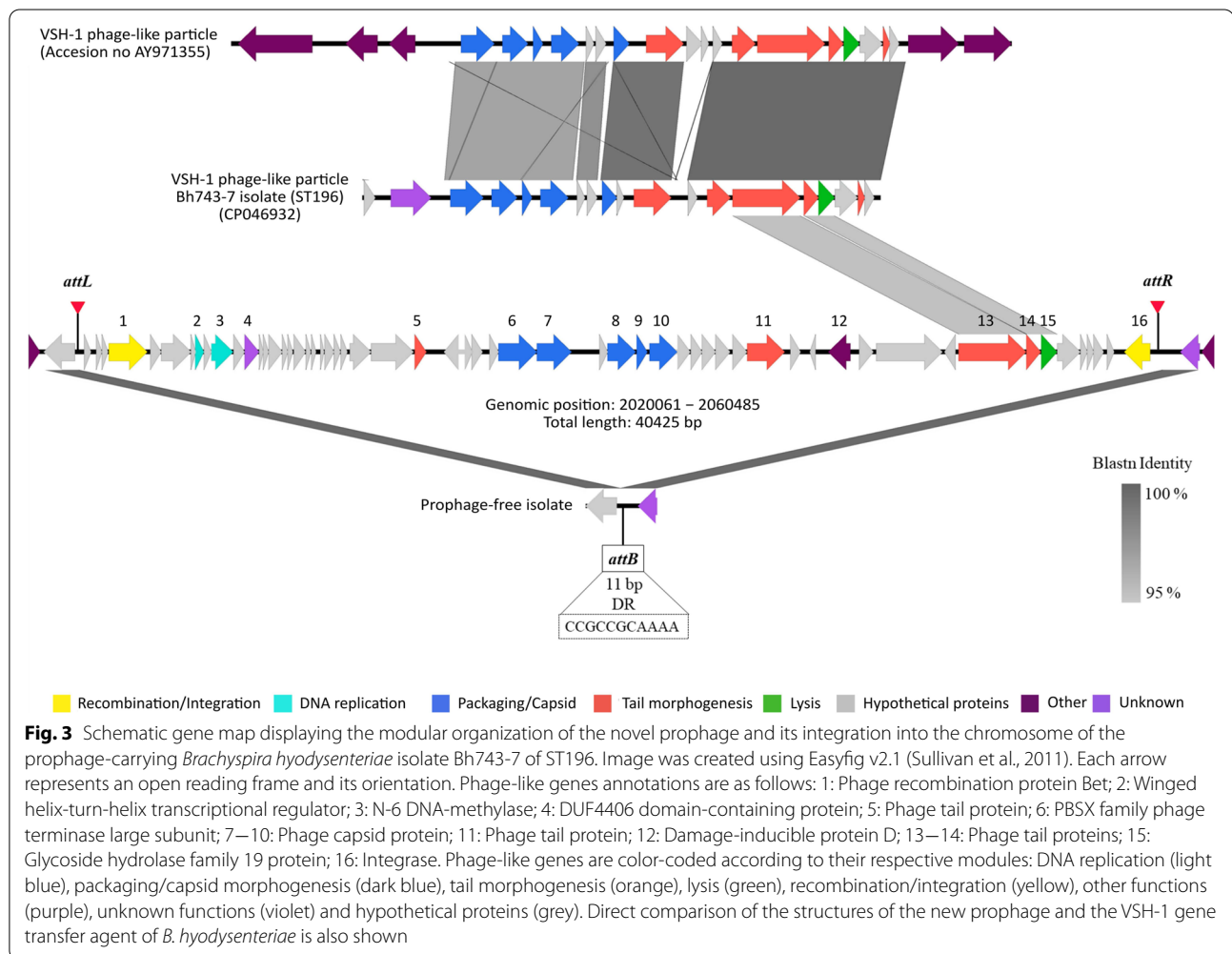
were obtained when the shortest and largest genomes of *B. hyodysenteriae* isolates BHZ333 and Bh743-7 were considered for pairwise comparison (Fig. 2 A). The first pangenome analysis was computed considering the whole set of coding genes of the 15 Swiss isolates of ST196 analysed, revealing the presence of core, soft-core and singletons gene clusters (Fig. 2 A and 2 B). A total of 2569 gene clusters (GCs) containing 38,406 genes were binned into core (2441 (95%) GCs, 37,464 genes), soft-core (64 (2.5%) GCs, 841 genes) and singletons (64 (2.5%) GCs, 65 genes) (Fig. 2 A). The plasmids were nearly identical and contributed to the pangenome with only 30 GCs; of those, 24 were core GCs and six were soft-core GCs. No singletons were identified (Fig. 2 B). Most of the plasmid-encoded genes were classified as players of cellular processes and signalling, followed by transport and metabolism of coenzymes, nucleotides and carbohydrates and motility (Additional file 6). Only two GCs that contained uncategorized genes were found in both bins core and soft-core GCs of the plasmids (Additional file 6). At the complete genome level, out of the 2569 GCs, 2489 were shared among 14 isolates and 2505 GCs were shared by a maximum of four isolates. COGs functions and categories, among which cellular and signalling processes, amino acid transport and metabolism and poorly characterized categories were the most abundant, were assigned to each protein-coding gene (Additional file 6). All eight hemolysin genes were identical at the DNA level and were classified within the core genome (Additional file 6). Four, two and one non-synonymous mutations were detected in the hemolysin III, hemolysin and hemolysin activation protein encoding genes, respectively. The soft-core bin contained 64 GCs, of which 48 were shared among 14 isolates. Forty-two of these GCs were categorized mainly as cellular and signalling genes, but also as poorly characterized ones (Additional file 6). The other 16 GCs were present in a varying number of isolates, and 11 of them contained proteins annotated as acetyl and glycosyl transferases, HAMP domain-containing proteins, radical SAM proteins, alpha-1,2-fucosyltransferase, tetratricopeptide repeat-containing protein and methyl accepting chemotaxis proteins. Concerning the singleton bin, three singletons were found in *B. hyodysenteriae* isolates BHZ660, BHZ695 and B 115_02A (Fig. 2B). Sixty-two singletons were present exclusively in the chromosome of *B. hyodysenteriae* isolate Bh743-7 (Fig. 2 A and 2 B). While most singletons did not have functional annotation, 13 of them were annotated as phage-like proteins (Additional file 6). Among these phage-like genes, only seven were classified into the COGs categories cellular and signalling processes and defense mechanisms (Additional file 6). The organization of the gene clusters based on the synteny of the complete genome of

B. hyodysenteriae isolate Bh743-7 revealed that these phage-like elements were located in the same chromosomal region (Additional Fig. S2).

A novel prophage integrated in the genome of *Brachyspira hyodysenteriae* Bh743-7

The analysis of the chromosome of *B. hyodysenteriae* isolate Bh743-7 using Prophage Hunter revealed the presence of a 40,425 bp insert, which corresponded to the phage-like region mentioned above (Fig. 3). This insert was compatible with a predicted active prophage, with an activity score above 0.8 and was characterized by a low G+C content (27.4%). The modular organization of its open reading frames resembled the structure of tailed and double-stranded DNA bacteriophages of the *Siphoviridae* family (e.g. *Streptococcus agalactiae* phage LYGO9 (JX409894)). By comparing the insert-containing isolate Bh743-7 with insert-free isolates of the same ST, we could identify a novel prophage, named *pphBhCH20*. This prophage was integrated in the chromosome between positions 2,020,061 and 2,060,485, between a hypothetical protein and a DUF-domain containing protein, and was flanked by two 11 bp direct repeats (DR) (5'-CCGCCGCAAAA-3') (Fig. 3). The prophage *pphBhCH20* contained 58 genes of which 16 were annotated as phage-like genes and the rest as hypothetical proteins (Fig. 3). The 16 annotated genes consisted of one phage recombination protein Bet, one winged helix-turn-helix transcriptional regulator, one N-6 DNA-methylase, one DUF4406 domain-containing protein, three phage tail proteins, one PBSX family phage terminase large subunit, five phage capsid proteins, one damage-inducible protein D, one glycoside hydrolase family 19 protein and one integrase, which were organized in different modules according to their function (Fig. 3). Comparative analysis of the nucleotide sequences of *pphBhCH20* and VSH-1 (AY971355) revealed that, with an alignment coverage above 99%, three ORFs of the new prophage shared more than 93% sequence identity with protein-coding genes of VSH-1 (Fig. 3). Specifically, two tail proteins and the glycosidase hydrolase (lysin) family 19 protein differed from both Hvp101 and Hvp28 tail proteins and the lysin of VSH-1 by only 32, 11 and six amino acids, respectively (Fig. 3).

To determine whether other *Brachyspira* carried the novel prophage *pphBhCH20* of *B. hyodysenteriae* isolate Bh743-7, as well as other new prophages, a pangenome analysis was conducted with additional *Brachyspira* species genomes (Additional Fig. S3). More than half of the genomes used in this analysis belonged to the species *B. hyodysenteriae* (n = 54). The others were from *B. aalborgi* (n = 17), *B. pilosicoli* (n = 8), *B. hamptonii* (n = 7), *B. murdochii* (n = 2), *B.*



intermedia (n = 1), and *B. suanatina* (n = 1). The new pangenome analysis comprised 7401 GCs containing 225,366 gene calls, of which more than 50% had neither COGs nor NCBI-PGAP functional annotations (Fig. S3). In general, *B. hyodysenteriae* genomes had more functional annotations assigned than other *Brachyspira* species genomes (Additional file 7). Most genes were not shared by all genomes and were binned into either soft-core (5029 GCs, 68.0%) or singleton (1556 GCs, 21.0%) bins. Singletons were less abundant in the *B. hyodysenteriae* genomes than in those of the other *Brachyspira* species. Less singletons in *B. hyodysenteriae* may arise from the greater number of genomes available for analysis. The core genome represented by all the genes shared among all *Brachyspira* genomes comprised only 816 GCs (11%) containing 76,427 gene calls, many of them with unknown function. Functional annotation of soft-core genes revealed the presence of phage-like genes in the genomes of different

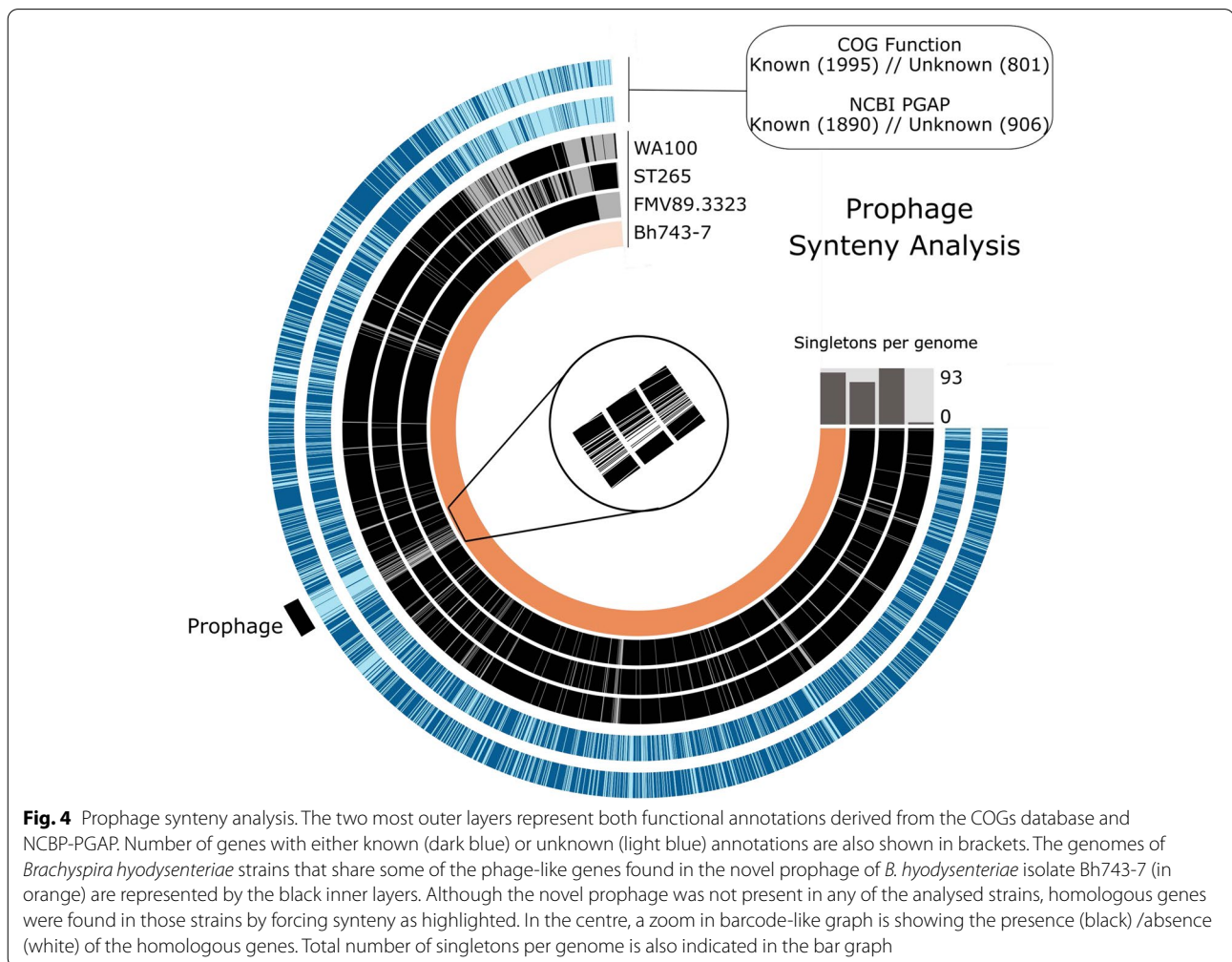
Brachyspira species. Although containing a high number of singletons without assigned function, the singleton bin was enriched in genes associated with CRISPR/Cas system, endonucleases, transposases, cell wall and lipopolysaccharide synthesis, outer membrane protein, proteases, efflux pump, and transcriptional regulators (Additional file 7). Some phage-like genes were identified in *B. intermedia* strain PWS-A (NC_017243.1), and *B. pilosicoli* strains SP16 (NZ_AFQM01000000.1), B2904 (NC_018607.1), and WesB (NC_018604.1), exclusively. The genome of *B. hyodysenteriae* isolate Bh743-7 contributed to the singleton bin with only 10 genes, eight of which had no functional annotation. The other two annotated genes represented the phage capsid protein and the phage recombination protein Bet of the novel prophage *pphBhCH20*. The reduction in the number of singletons counted for the genome of isolate Bh743-7 suggested the existence of elements shared with other genomes. For instance, prophage elements highly similar (combined homogeneity index

ranging from 0.65 to < 1), or even identical (combined homogeneity index equal to 1), to those found in *pphB-hCH20* were also identified, in a varying number, in the genomes of different *Brachyspira* species including *B. hyodysenteriae*, *B. intermedia*, *B. pilosicoli*, *B. hampsonii*, *B. murdochii*, and *B. aalborgi* (Additional file 7). However, variations in amino acids (see functional homogeneity index) and protein sequence length (see geometric homogeneity index) were observed (Additional file 7). The highest number of homologous phage-like genes was shared among *B. hyodysenteriae* strains FMV89.3323 (JXNB00000000.1), ST265 (NZ_JXNQ00000000.1), WA100 (NZ_JXNS00000000.1) and Bh743-7 (CP046932 and CP046933). Despite those elements seem to be part of a prophage region, as shown by forcing synteny according to the gene organization of the complete genome of *B. hyodysenteriae* isolate Bh743-7, they co-occurred randomly and synteny was only partially resembled (Fig. 4). The complete novel prophage *pphBhCH20* was not found in any of the

other genomes analysed here, but the results indicated that other prophages or parts of them are present in *Brachyspira* spp. (Additional file 7).

Discussion

This study presents a unique comparative analysis of 15 genomes of *B. hyodysenteriae* belonging to the same ST and reveals different sublineages, as well as a new prophage. All assemblies obtained here were of high quality in terms of completeness (as compared to the ST196 reference genome) and accuracy (supported by high depth of coverage and high similarity among sequences at both nucleotide and protein sequence levels). All genomes consisted of a pair of replicons corresponding to one chromosome and one plasmid in concordance with the known structure of *B. hyodysenteriae* genomes [4, 5, 21, 43–45], except for one genome that lacked the plasmid and another for which the plasmid could not be fully reconstructed. The genomes were characterized by a low G+C content (~27.0%) and a high frequency of long



homopolymeric regions and tandem repeats, features that complicate the sequencing process [46].

Despite the fragmentation and inherent limitations of the draft assemblies of the genomes sequenced by Illumina reads, they were similar in the number of contigs and length. Compared to other *B. hyodysenteriae* genomes [47], our nearly complete assemblies gained quality in terms of reduced number of contigs and increased contiguity thanks to advances in sequencing technology and assemblers. Complete plasmids with sizes of ~32 kb could be reconstructed from single contigs in 12 isolates. While the plasmid of the isolate B 114_09C could be reconstructed from two contigs, no plasmid was found in the isolate BHZ333. The role of this *B. hyodysenteriae* plasmid is still not well understood. The absence of plasmid has been reported in some *B. hyodysenteriae* strains not associated with SD [47–49]. In a follow-up publication, it was reported that the absence of four plasmid-encoded genes was predictive of a reduced pathogenic potential in *B. hyodysenteriae* [50]. However, in our study, the isolate BHZ333 lacking the plasmid was obtained from a pig with SD [20], while the isolate B115_02A, harboring all the plasmid-encoded genes described previously [50], was isolated from a pig in which SD was not diagnosed [20]. Previous studies associated pathogenicity with the presence of other chromosomal virulence genes (e.g. ankyrin proteins, outer membrane proteins, proteins associated with chemotaxis and motility, and hemolysins) [43, 51, 52]. Moderately to weakly hemolytic strains of *B. hyodysenteriae* have been reported in different countries [2, 3, 53, 54]. All eight hemolysin genes regarded as important in the pathogenesis of SD [3] were identified in all isolates of ST196 including in the two isolated from pigs with subclinical infections. The two genes known to be associated with the strong hemolytic phenotype were also present [2]. Although none of the mutations reported by Card and collaborators [2] were identified, several non-synonymous mutations were detected in three hemolysin genes. Regarding the hemolytic phenotype of all 15 *B. hyodysenteriae* isolates of ST196, not only changes in both strength and extension of hemolysis, but also loss of hemolytic activity were observed even for a single isolate following repeated culture passages. Whether such phenotypic differences were due to differential gene expression and/or post-transcriptional events, as suggested previously [2, 3], remain unclear and require further investigations. While our data suggest that pathogenic isolates generally harbor intact plasmids, as well as all the eight hemolysin genes, we consider that we are still far from understanding all genetic and environmental factors that orchestrate both mild and full pathogenicity in *B. hyodysenteriae*. Our comparative analysis of 15 genomes

of *B. hyodysenteriae* of ST196 provides new insights into the genomic structure and variability within *B. hyodysenteriae* isolate of a same ST over time. Despite observing a high degree of genomic stability, supported by pangenome and ANI analyses, cgSNPs analysis revealed genetic differences across the genomes of *B. hyodysenteriae* isolates of ST196 and the presence of sublineages within the same ST. Most of such differences occurred randomly. Those frameshift, stop-lost, or missense mutations that occurred in protein-coding sequences have not been investigated for change in function.

An association between the different sublineages and both date and region of isolation was not found, suggesting that the different lineages of ST196 have evolved in an independent manner and persisted in Switzerland over nearly a decade. Finding isolates of ST196 over time suggests not only the existence of few common sources, as previously thought [20], but points out towards other factors that should be considered for successful control of SD in Switzerland (e.g. herd management, transportation and biosecurity practices) [55]. Furthermore, the fact of having found macrolide-lincosamide resistance isolates of ST196 being predominant over other STs (ST6, ST66 and ST197) in Switzerland, could be associated with the acquisition and widespread dissemination of point mutations linked to the decreased susceptibility to such antimicrobials [20]. Although WGS and cgSNPs analyses have been used for high resolution comparison and outbreak investigations of other pathogenic bacterial species, such as *Klebsiella pneumoniae* [56, 57], this combination of analyses has only been recently applied to characterize and to assess persistence of *B. hyodysenteriae* [9]. Specifically, it was found that *B. hyodysenteriae* isolates of various STs can persist over time. Previously, persistence of *B. hyodysenteriae* isolates was assessed by multilocus sequence typing analyses. In fact, *B. hyodysenteriae* isolates of ST56, to which the *B. hyodysenteriae* type strain B-78^T (NZ_CP015910.2) belongs, were shown to persist in North America over long periods of time [58]. Nonetheless, the authors did not report the existence of sublineages within the same ST56, probably due to the less resolutive power of their seven-housekeeping-genes-based multilocus sequence typing analysis, (see La and collaborators [59]) compared to both WGS and cgSNP analyses. Also supporting our findings, minor differences across Australian *B. hyodysenteriae* isolates within the same ST as well as persistence of certain STs over time were reported in 2016 [60]. Nonetheless, such isolates displayed different antimicrobial susceptibility profiles that were not observed in our previous study [20].

At the pangenome level, considering both core and singleton bins, our results also revealed a highly conserved but flexible genomic structure that can be shaped by the

integration of larger genetic elements, such as the novel prophage *pphBhCH20* found in the chromosome of a single *B. hyodysenteriae* isolate of ST196. The 40,425 bp *pphBhCH20* constitutes the third different type of mobile genetic element described so far in *B. hyodysenteriae*. It resembles the structure of the tailed and double-stranded DNA *Streptococcus agalactiae* phage LYGO9 (JX409894) of the *Siphoviridae* family [61]. In line with this, the *lnu(C)*-carrying transposon *MTnSag1* recently detected in *B. hyodysenteriae*, was originally identified in *S. agalactiae* [11], demonstrating the capability of *B. hyodysenteriae* to acquire foreign genetic material from other bacterial species through various HGT mechanisms [4, 5, 62]. Although other prophages have been identified in *B. intermedia*, *B. murdochii* and *B. pilosicoli* [48, 63], to our knowledge, only the defective prophage VSH-1 has been shown to play a role in HGT in *B. hyodysenteriae* in *in vitro* experiments [62]. Compared to the 16.3 kb genome of VSH-1, which also exhibits the structure of bacteriophages of the *Siphoviridae* family [62], the new prophage *pphBhCH20* was more than double in size and carried early function genes involved in DNA replication (i.e. N-6 DNA-methylase and winged helix-turn-helix transcriptional regulator). Moreover, three late function genes (two tail proteins and a glycosidase hydrolase family 19 protein) were shared between *pphBhCH20* and VSH-1, a phenomenon also observed in *Brachyspira* species phages that suggests that VSH-1 is responsible for HGT [63]. So far, a possible role of *pphBhCH20* in the *B. hyodysenteriae* isolate Bh743-7 could not be elucidated based on the genes present within its structure, since most of the non-phage related genes coded for hypothetical proteins. The fact that this prophage was found in a single Swiss *B. hyodysenteriae* isolate sampled in 2017 could suggest a recent integration event. However, it seems not to provide any advantage on persistence over time so far.

By extending the pangenome analysis to other *Brachyspira* species, we have shown that many genes still remain poorly characterized, as previously reported [48], thus, needing further investigation. The lack of functional annotation clearly reflects the difficulties (e.g. specialized growth requirements, absence of means for genetic manipulation) associated with studies on the slow-growing *Brachyspira* [64]. Despite this absence of information, the generated pangenome data set may serve as a basis for identifying candidate acquired genes that could play a role in antimicrobial resistance like e.g. those coding for putative MATE family efflux transporter and glyoxalase/bleomycin resistance protein/dioxygenase superfamily protein which were present as singletons in specific strains (Additional file 7). Also we were able to identify phage-like genes, likely belonging to phages already described

in other genomes of *B. intermedia*, *B. murdochii* and *B. pilosicoli* [48, 63]. Nevertheless, although identical and highly similar genes to those phage-like genes of the novel prophage *pphBhCH20* were found in three other *B. hyodysenteriae* strains, the prophage structure was not resembled entirely. Of note, these *B. hyodysenteriae* strains were found in Europe (*B. hyodysenteriae* of ST265), Australia (*B. hyodysenteriae* WA100) and Canada (*B. hyodysenteriae* FMV89.3323). It has been shown that bacterial strains can develop different strategies to avoid phage infection and consequently, turn into phage-resistant bacterial strains [65]. Whether all the *B. hyodysenteriae* strains sharing the highest number of homologous phage-like genes are more sensitive to phage infections compared to other strains remains unclear. Alternatively, potential technical limitations that apply to phage genomics (e.g., sequencing process, viral genome assembly and bioinformatics analyses) could compromise the possibility of finding either such phage-like genes or entire prophages [66]. Likely, WGS of *B. hyodysenteriae* strains using long-read sequencing platforms, as well as improvements in phage-like genes annotations, will help to explain our observations. In agreement with other authors [13, 48], the identification of different phage-like genes in different *Brachyspira* species suggests that prophages might be more common than previously thought and that horizontal gene transfer events mediated by prophages play an important role in *Brachyspira* biology and evolution.

Conclusions

With this study we contribute to the *B. hyodysenteriae* genomes pool with 14 new genomes. Moreover, we highlighted the power of WGS-based analyses to identify genetic differences across *B. hyodysenteriae* isolates within the same ST. Regarding the variability of Swiss *B. hyodysenteriae* of ST196, this in-depth WGS analysis revealed the existence of different sublineages that seem to have evolved independently and persisted in Switzerland over nearly a decade. The implications of such findings need to be considered for epidemiological projects aiming to trace back specific clones to successfully control SD in Switzerland. Moreover, we showed how horizontal gene transfer events in *Brachyspira* spp. can be detected by pangenome analyses, and found a novel prophage *pphBhCH20* integrated into the chromosome of the *B. hyodysenteriae* isolate Bh743-7. This study paves the way for further research into WGS comparative analysis and into the functionality of the highly abundant poorly characterized genes of *Brachyspira* spp., as well as into phage diversity and phages-*Brachyspira* species interactions.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12864-022-08347-5>.

Additional file 1: This file contains python instructions.

Additional file 2: This file contains python instructions.

Additional file 3: This file contains BioProjects and information of all analysed genomes.

Additional file 4: This file contains the commands followed for pangenome analysis using Anvi'o v6.2.

Additional file 5: This file contains information regarding core genome single nucleotide polymorphisms, their location and their effect.

Additional file 6: This file contains detailed information about the genes that represent the pangenome of *Brachyspira hyodysenteriae* of sequence type ST196.

Additional file 7: This file contains detailed information about the entire set of genes that constitute the *Brachyspira* pangenome at the genus level.

Additional file 8: Figure S1. Geographical distribution of *B. hyodysenteriae* isolates across Switzerland overtime. Each diamond represents one isolate sampled per herd. Different colors are used to identify samples obtained from geographically distant Swiss pig herds in different years between 2010 and 2017. The basemap indicating density of pig population (grey rings) was obtained from the Federal Statistical Office (<http://www.bfs.admin.ch>), ThemaKart.

Additional file 9: Figure S2. Synteny analysis. The two most outer layers represent both functional annotations, derived from the COGs database and NCBI-PGAP, and number of genes with either known (dark blue) or unknown (light blue) annotations are also shown. Genomes are organized according to the synteny of the reference genome (in orange). The genome of the novel prophage consisting of phage-like genes organized sequentially and integrated into the chromosome is highlighted in black. Total number of gene clusters (GCs) and genes are indicated for core (dark blue), soft-core (light pink) and singleton (dark pink) bins.

Additional file 10: Figure S3. Pangenome analysis of genomes of different *Brachyspira* species. (A) General organization and visualization of 90 *Brachyspira* genomes based on the presence/absence of genes and their contribution to the bins core, soft-core and singleton. Total number of gene clusters (GCs) as well as the number of gene calls falling into each bin are shown in brackets. A graph bar displaying the number of singletons per genome, varying from 0 to 174, is also included. The two outmost layers represent both functional annotations derived from the NCBI-PGAP and COGs database. Known and unknown functions are in dark and light blue, respectively. Total number of both known and unknown functions are indicated in brackets. Genomes are colored-coded according to the seven different species they belong to except Swiss *B. hyodysenteriae* genomes that are colored in orange to facilitate their visualization. (B) Visualization of the soft-core gene clusters containing a high number of accessory genes. (C) Singletons present in each genome are highlighted and information regarding functional annotation is indicated in brackets.

Additional file 11: Alignment scores of *B. hyodysenteriae* ST196 isolates obtained with Snippy.

Additional file 12: Table S1. Metrics and genetic features of the de novo draft assemblies of *B. hyodysenteriae* isolates of ST196. File containing a table with metrics and genetic features of the de novo draft assemblies of 15 *B. hyodysenteriae* of ST196.

Acknowledgements

We thank to the teams of the Competency Centre in Bioinformatics and Computational Biology (VITAL-IT, <https://www.vital-it.ch/>), and the high-performance computation clusters of the University of Bern (UBELIX, <http://www.id.unibe.ch/hpc> and IBU, www.bioinformatics.unibe.ch/services) for providing software support and computing resources.

Authors' contributions

VP was in charge of funding acquisition, project supervision and coordination and first revision of the draft manuscript. VP and ABGM designed the study. RB co-supervised, provided super-computation resources and revised the draft manuscript. ABGM performed laboratory procedures, data acquisition and curation, all bioinformatics analyses, created images, wrote the draft manuscript and edited it with inputs from all authors. TR was responsible of installing and maintaining the NCBI-PGAP Singularity container, wrote the custom Python scripts and revised the draft manuscript. SS provided bacterial isolates (BHZ isolates) and epidemiological information and revised the draft manuscript. FZ provided bacterial isolates (B 114_09C, B 115_02A and Bh743-7 isolates) and epidemiological information and revised the draft manuscript. All authors read and approved the final manuscript.

Funding

This study was financed by grants no. 1.16.04 and 1.19.05 of the Swiss Federal Food Safety and Veterinary Office (SFVO).

Availability of data and materials

All supporting data, scripts and workflow commands are provided as additional material. Raw sequencing datasets are deposited at the SRA database and accessible under SRA study SRP306907. Genome assemblies obtained in this study belong to the BioProject PRJNA697976 and are available under the accession numbers JAFCS0000000000 (BHZ26), JAFCS0000000000 (BHZ153), JAFCSR0000000000 (BHZ231), JAFCSQ0000000000 (BHZ333), JAFCSPO0000000000 (BHZ480), JAFCSO0000000000 (BHZ630), JAFCSN0000000000 (BHZ2660), JAFCSM0000000000 (BHZ684), JAFCSL0000000000 (BHZ695), JAFCSK0000000000 (BHZ777), JAFCSJ0000000000 (BHZ784), JAFCSI0000000000 (BHZ819), JAFCSV0000000000 (B 114_09C) and JAFCSU0000000000 (B 115_02A). Accession information is also provided for each genome assembly retrieved from GenBank as additional material (see additional file 3).

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

None of the authors of this paper have a financial or personal relationship with other people or organizations that could inappropriately influence or bias the content of the paper.

Author details

¹Division of Molecular Bacterial Epidemiology and Infectious Diseases, Institute of Veterinary Bacteriology, Vetsuisse Faculty, University of Bern, Bern, Switzerland. ²Graduate School for Cellular and Biomedical Sciences, University of Bern, Bern, Switzerland. ³Interfaculty Bioinformatics Unit and Swiss Institute of Bioinformatics, University of Bern, Bern, Switzerland. ⁴Section of Veterinary Bacteriology, Institute for Food Safety and Hygiene, Vetsuisse Faculty, University of Zurich, Zurich, Switzerland. ⁵Clinic for Swine, Department of Clinical Veterinary Medicine, Vetsuisse Faculty, University of Bern, Bern, Switzerland. ⁶Institute of Veterinary Bacteriology, University of Bern, Länggassstrasse 122, CH-3012 Bern, Switzerland.

Received: 28 May 2021 Accepted: 25 January 2022

Published online: 15 February 2022

References

- Hampson D, Lugsomya K, La T, Dale Phillips N, J. Trott D, Abraham S. Antimicrobial resistance in *Brachyspira* – An increasing problem for disease control. *Vet Microbiol.* 2019;229:59-71.
- Card RM, La T, Burrough ER, Ellis RJ, Núñez-García J, Thomson JR, Mahu M, Phillips ND, Hampson DJ, Rohde J, et al. Weakly haemolytic variants of

- Brachyspira hyodysenteriae* newly emerged in Europe belong to a distinct subclade with unique genetic properties. *Vet Res.* 2019;50(1):21.
3. Joerling J, Willems H, Ewers C, Herbst W. Differential expression of hemolysin genes in weakly and strongly hemolytic *Brachyspira hyodysenteriae* strains. *BMC Vet Res.* 2020;16(1):169.
 4. De Luca S, Nicholson P, Magistrali CF, García-Martín AB, Rychener L, Zeeh F, Frey J, Perreten V. Corrigendum to "Transposon-associated lincosamide resistance *Inu(C)* gene identified in *Brachyspira hyodysenteriae* ST83" [*Vet Microbiol.* 214 (2018) 51–55]. *Vet Microbiol.* 2018;220:113.
 5. De Luca S, Nicholson P, Magistrali CF, García-Martín AB, Rychener L, Zeeh F, Frey J, Perreten V. Transposon-associated lincosamide resistance *Inu(C)* gene identified in *Brachyspira hyodysenteriae* ST83. *Vet Microbiol.* 2018;214:51–5.
 6. Hidalgo Á, Carvajal A, Vester B, Pringle M, Naharro G, Rubio P. Trends towards lower antimicrobial susceptibility and characterization of acquired resistance among clinical isolates of *Brachyspira hyodysenteriae* in Spain. *Antimicrob Agents Chemother.* 2011;55(7):3330–7.
 7. Hillen S, Willems H, Herbst W, Rohde J, Reiner G. Mutations in the 50S ribosomal subunit of *Brachyspira hyodysenteriae* associated with altered minimum inhibitory concentrations of pleuromutilins. *Vet Microbiol.* 2014;172(1–2):223–9.
 8. Karlsson M, Fellström C, Heldtander M, Johansson K-E, Franklin A. Genetic basis of macrolide and lincosamide resistance in *Brachyspira* (Serpulina) *hyodysenteriae*. *FEMS Microbiol Lett.* 2006;172:255–60.
 9. Card RM, Stubberfield E, Rogers J, Núñez-García J, Ellis RJ, AbuOun M, Strugnell B, Teale C, Williamson S, Anjum MF. Identification of a new antimicrobial resistance gene provides fresh insights into pleuromutilin resistance in *Brachyspira hyodysenteriae*, aetiological agent of swine dysentery. *Front Microbiol.* 2018;9:1183.
 10. García-Martín AB, Schwendener S, Perreten V. The *tva(A)* gene from *Brachyspira hyodysenteriae* confers decreased susceptibility to pleuromutilins and streptogramin A in *Escherichia coli*. *Antimicrob Agents Chemother.* 2019;63(9):e00930-19.
 11. Achard A, Villers C, Pichereau V, Leclercq R. New *Inu(C)* gene conferring resistance to lincosycin by nucleotidylation in *Streptococcus agalactiae* UCN36. *Antimicrob Agents Chemother.* 2005;49(7):2716–9.
 12. Ritchie A, Robinson I, Joens L, Kinyon J. A bacteriophage for *Treponema hyodysenteriae*. *Vet Rec.* 1978;103(2):34.
 13. Stanton TB. Prophage-like gene transfer agents—novel mechanisms of gene exchange for *Methanococcus*, *Desulfovibrio*, *Brachyspira*, and *Rhodobacter* species. *Anaerobe.* 2007;13(2):43–9.
 14. Anani H, Zgheib R, Hasni I, Raouf D, Fournier P-E. Interest of bacterial pangenome analyses in clinical microbiology. *Microb Pathog.* 2020;149:104275.
 15. Jamroz D, Coll F, Mather AE, Harris SR, Harrison EM, MacGowan A, Karas A, Elston T, Török ME, Parkhill J. Evolution of mobile genetic element composition in an epidemic methicillin-resistant *Staphylococcus aureus*: temporal changes correlated with frequent loss and gain events. *BMC Genomics.* 2017;18(1):1–12.
 16. Kaas RS, Friis C, Ussery DW, Aarestrup FM. Estimating variation within the genes and inferring the phylogeny of 186 sequenced diverse *Escherichia coli* genomes. *BMC Genomics.* 2012;13(1):577.
 17. Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV, Crabtree J, Jones AL, Durkin AS. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome." *Proc National Acad Sci.* 2005;102(39):13950–5.
 18. Labbé G, Kruczkiewicz P, Mabon P, Robertson J, Schonfeld J, Kein D, Rankin MA, Gopez M, Hole D, Son Det al. Rapid and accurate SNP genotyping of clonal bacterial pathogens with BioHansel. *bioRxiv* 2020:2020.2001.2010.902056.
 19. Pightling AW, Pettengill JB, Luo Y, Baugher JD, Rand H, Strain E. Interpreting whole-genome sequence analyses of foodborne bacteria for regulatory applications and outbreak investigations. *Front Microbiol.* 2018;9:1482.
 20. García-Martín AB, Perreten V, Rossano A, Schmitt S, Nathues H, Zeeh F. Predominance of a macrolide-lincosamide-resistant *Brachyspira hyodysenteriae* of sequence type 196 in Swiss pig herds. *Vet Microbiol.* 2018;226:97–102.
 21. García-Martín AB, Schmitt S, Zeeh F, Perreten V. Complete circular genome sequences of *Brachyspira hyodysenteriae* isolates of the four different sequence types causing swine dysentery in Switzerland. *Microbiol Resour Announc.* 2021;10(39):e00847-00821.
 22. Andrews S. FastQC: a quality control tool for high throughput sequence data. In: Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom; 2010.
 23. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics.* 2016;32(19):3047–8.
 24. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics.* 2014;30(15):2114–20.
 25. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Pribelski AD et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Computational Biol.* 2012;19(5):455–77.
 26. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics.* 2013;29(8):1072–5.
 27. Tatusova T, DiCuccio M, Badretdin A, Chetvernin V, Nawrocki EP, Zaslavsky L, Lomsadze A, Pruitt KD, Borodovsky M, Ostell J. NCBI prokaryotic genome annotation pipeline. *Nucleic Acids Res.* 2016;44(14):6614–24.
 28. Seemann T. Snippy: rapid haploid variant calling and core genome alignment. 2015.
 29. Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, Parkhill J, Harris SR. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res.* 2015;43(3):e15.
 30. Price MN, Dehal PS, Arkin AP. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol Biol Evol.* 2009;26(7):1641–50.
 31. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res.* 2019;47(W1):256–9.
 32. Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, Lu X, Ruden DM. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly.* 2012;6(2):80–92.
 33. Pritchard L, Glover RH, Humphris S, Elphinstone JG, Toth IK. Genomics and taxonomy in diagnostics for food security: soft-rotting enterobacterial plant pathogens. *Analytical Methods.* 2016;8(1):12–24.
 34. Eren AM, Esen Ö C, Quince C, Vineis JH, Morrison HG, Sogin ML, Delmont TO. Anvi'o: an advanced analysis and visualization platform for omics data. *PeerJ* 2015;3:e1319.
 35. Köster J, Rahmann S. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics.* 2012;28(19):2520–2.
 36. Lee M. Bioinformatics Tools (Bit) package v1.4.71. 2018. Accessed 27 May 2021. <https://github.com/AstroBioMike/bit>.
 37. Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics.* 2010;11(1):119.
 38. Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, Kiryutin B, Galperin MY, Fedorova ND, Koonin EV. The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.* 2001;29(1):22–8.
 39. Delmont TO, Eren AM. Linking pangenomes and metagenomes: the *Prochlorococcus* metapangenome. *PeerJ.* 2018;6:e4320–e4320.
 40. Song W, Sun H-X, Zhang C, Cheng L, Peng Y, Deng Z, Wang D, Wang Y, Hu M, Liu W. Prophage Hunter: an integrative hunting tool for active prophages. *Nucleic Acids Res.* 2019;47(W1):W74–80.
 41. Darling AE, Mau B, Perna NT. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS ONE* 2010;5(6):e11147.
 42. Sullivan MJ, Petty NK, Beatson SA. Easyfig: a genome comparison visualizer. *Bioinformatics.* 2011;27(7):1009–10.
 43. Bellgard MJ, Wanchanthuek P, La T, Ryan K, Moolhuijzen P, Albertyn Z, Shaban B, Motro Y, Dunn DS, Schibeci Det al. Genome sequence of the pathogenic intestinal spirochete *Brachyspira hyodysenteriae* reveals adaptations to its lifestyle in the porcine large intestine. *PLoS One* 2009;4(3):e4641.
 44. Mirajkar NS, Johnson TJ, Gebhart CJ. Correction for Mirajkar et al., Complete genome sequence of *Brachyspira hyodysenteriae* type strain B78 (ATCC 27164). *Genome Announc.* 2017;5(3):e01453-16.
 45. Mirajkar NS, Johnson TJ, Gebhart CJ. Complete genome sequence of *Brachyspira hyodysenteriae* type strain B-78 (ATCC 27164). *Genome Announc.* 2016;4(4):e00840-16.

46. Sarkozy P, Jobbágy Á, Antal P. Calling homopolymer stretches from raw nanopore reads by analyzing *k*-mer dwell times. In: EMBEC & NBC 2017. Springer. 2017;241–244.
47. Black M, Moolhuijzen P, Barrero R, La T, Phillips N, Hampson D, Herbst W, Barth S, Bellgard M. Analysis of multiple *Brachyspira hyodysenteriae* genomes confirms that the species is relatively conserved but has potentially important strain variation. PLoS One. 2015;10(6):e0131050.
48. Häfström T, Jansson DS, Segerman B. Complete genome sequence of *Brachyspira intermedia* reveals unique genomic features in *Brachyspira* species and phage-mediated horizontal gene transfer. BMC Genomics. 2011;12:395.
49. La T, Phillips ND, Wanchanthuek P, Bellgard MI, O'Hara AJ, Hampson DJ. Evidence that the 36 kb plasmid of *Brachyspira hyodysenteriae* contributes to virulence. Vet Microbiol. 2011;153(1–2):150–5.
50. La T, Phillips ND, Thomson JR, Hampson DJ. Absence of a set of plasmid-encoded genes is predictive of reduced pathogenic potential in *Brachyspira hyodysenteriae*. Vet Res. 2014;45:131.
51. Barth S, Gömmel M, Baljer G, Herbst W. Demonstration of genes encoding virulence and virulence life-style factors in *Brachyspira* spp. isolates from pigs. Vet. Microbiol. 2011;155(2–4):438–443.
52. Álvarez-Ordóñez A, Martínez-Lobo FJ, Argüello H, Carvajal A, Rubio P. Swine Dysentery: aetiology, pathogenicity, determinants of transmission and the fight against the disease. Int J Environ Res Public Health. 2013;10(5):1927–47.
53. Mahu M, De Pauw N, Vande Maele L, Verlinden M, Boyen F, Ducatelle R, Haesebrouck F, Martel A, Pasmans F. Variation in hemolytic activity of *Brachyspira hyodysenteriae* strains from pigs. Vet Res. 2016;47(1):66.
54. Clothier KA, Kinyon JM, Frana TS, Naberhaus N, Bower L, Strait EL, Schwartz K. Species characterization and minimum inhibitory concentration patterns of *Brachyspira* species isolates from swine with clinical disease. J Vet Diagn Invest. 2011;23(6):1140–5.
55. Zeeh F, Vidondo B, Nathues H. Risk factors for the infection with *Brachyspira hyodysenteriae* in pig herds. Prev Vet Med. 2019;174:104819.
56. Marsh JW, Mustapha MM, Griffith MP, Evans DR, Ezeonwuka C, Pasculle AW, Shutt KA, Sundermann A, Ayres AM, Shields RK. et al. Evolution of outbreak-causing carbapenem-resistant *Klebsiella pneumoniae* ST258 at a tertiary care hospital over 8 years. mBio 2019;10(5):e01945–01919.
57. Miro E, Rossen JWA, Chlebowicz MA, Harmsen D, Brisse S, Passet V, Navarro F, Friedrich AW, García-Cobos S. Core/whole genome multilocus sequence typing and core genome SNP-based typing of OXA-48-producing *Klebsiella pneumoniae* clinical isolates from Spain. Front Microbiol. 2020;10:2961.
58. Mirajkar NS, Gebhart CJ. Understanding the molecular epidemiology and global relationships of *Brachyspira hyodysenteriae* from swine herds in the United States: a multi-locus sequence typing approach. PLoS One. 2014;9(9):e107176.
59. La T, Phillips ND, Harland BL, Wanchanthuek P, Bellgard MI, Hampson DJ. Multilocus sequence typing as a tool for studying the molecular epidemiology and population structure of *Brachyspira hyodysenteriae*. Vet Microbiol. 2009;138(3–4):330–8.
60. La T, Phillips ND, Hampson DJ. An investigation into the etiological agents of swine dysentery in Australian pig herds. PLoS One. 2016;11(12):e0167424.
61. Sanz-Gaitero M, Seoane-Blanco M, van Raaij MJ. Structure and function of bacteriophages. In: Harper DR, Abedon ST, Burrows BH, McConville ML, editors. Bacteriophages: Biology, Technology, Therapy. Cham: Springer International Publishing; 2019. p. 1–73.
62. Matson EG, Thompson MG, Humphrey SB, Zuerner RL, Stanton TB. Identification of genes of VSH-1, a prophage-like gene transfer agent of *Brachyspira hyodysenteriae*. J Bacteriol. 2005;187(17):5885–92.
63. Mappley LJ, Black ML, AbuOun M, Darby AC, Woodward MJ, Parkhill J, Turner AK, Bellgard MI, La T, Phillips ND et al. Comparative genomics of *Brachyspira pilosicoli* strains: genome rearrangements, reductions and correlation of genetic complement with phenotypic diversity. BMC Genomics. 2012;13:454.
64. Wanchanthuek P, Bellgard MI, La T, Ryan K, Moolhuijzen P, Chapman B, Black M, Schibeci D, Hunter A, Barrero R et al. The complete genome sequence of the pathogenic intestinal spirochete *Brachyspira pilosicoli* and comparison with other *Brachyspira* genomes. PLoS One. 2010;5(7):e11455.
65. Labrie SJ, Samson JE, Moineau S. Bacteriophage resistance mechanisms. Nat Rev Microbiol. 2010;8(5):317–27.
66. Klumpp J, Fouts DE, Sozhamannan S. Next generation sequencing technologies and the changing landscape of phage genomics. Bacteriophage. 2012;2(3):190–9.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions



5.2. Manuscript 6: Frontiers for Young Minds article

Can Eating Bacteria In Dairy Products Support Your Health?

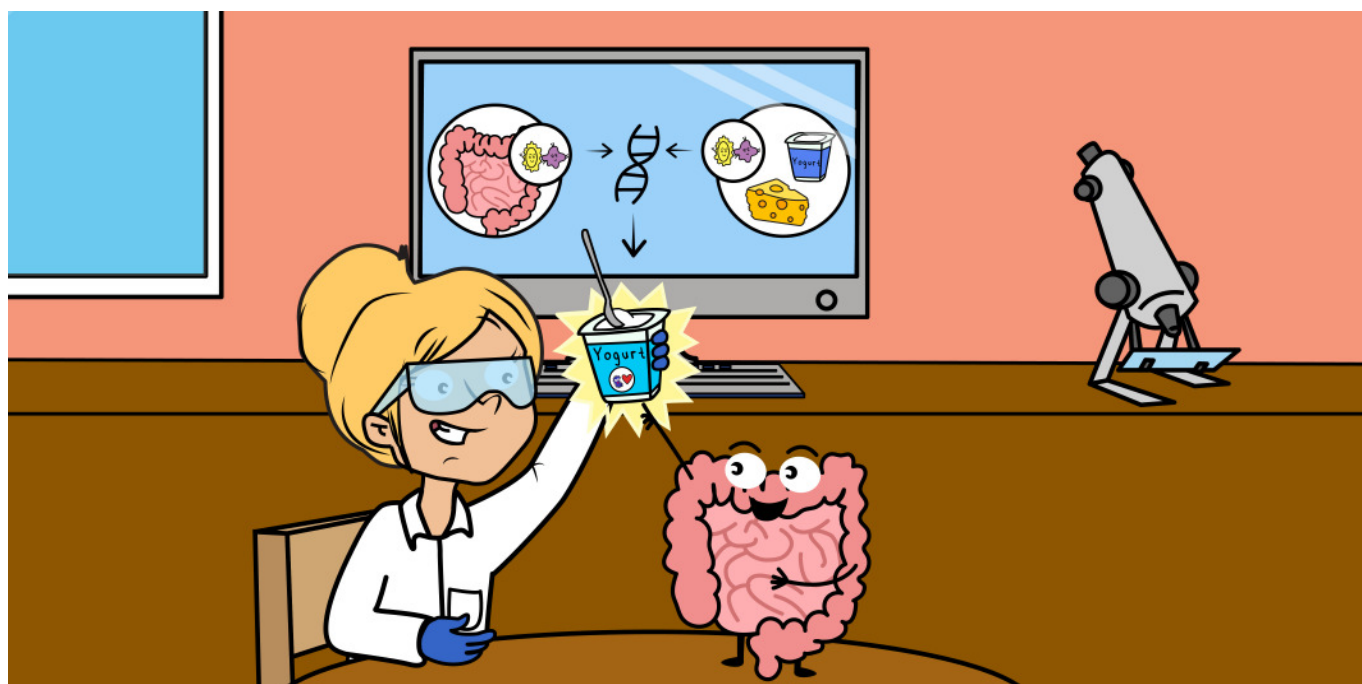
Thomas Roder, Grégory Pimentel, Cornelia Bär, Ueli von Ah,
Rémy Bruggmann, Guy Vergères

Status:

Published in Frontiers for Young Minds (2022), 10 [216]

Statement of contribution:

TR, CB and GP wrote the manuscript. All authors edited, read, and approved the final manuscript.



CAN EATING BACTERIA IN DAIRY PRODUCTS SUPPORT YOUR HEALTH?

Thomas Roder^{1,2,3*}, Grégory Pimentel³, Cornelia Bär³, Ueli von Ah³, Rémy Bruggmann^{1,2} and Guy Vergères³

¹Interfaculty Bioinformatics, University of Bern, Bern, Switzerland

²Swiss Institute of Bioinformatics (SIB), Lausanne, Switzerland

³Agroscope, Zurich, Switzerland

YOUNG REVIEWERS:



CAMERON

AGE: 10



ELLIOT

AGE: 11



EVE

AGE: 11

SPECIES

A group of organisms, like bacteria, that behave similarly and have very similar genomes.

Huge numbers of bacteria live in the human gut. We know those bacteria are important to our health, so we need to treat them well. We wanted to know whether it was possible to design new yogurts that can introduce special bacteria into the gut, to improve our well-being. We studied hundreds of types of bacteria isolated from cheese and yogurt and found that 24 of these bacterial species can perform most of the important bacterial functions that happen in the human gut. Therefore, there is exciting potential for designing new, gut-healthy yogurts.

BACTERIA KEEP US HEALTHY

Bacteria were among the first life forms to appear on Earth. They are extremely small organisms, consisting of a single cell. There are many **species** (major types) of bacteria that can be incredibly different from

each other: some prosper deep below the ocean, at temperatures higher than that of boiling water, while others live happily in Antarctic ice. Bacteria populate the entire surface of Earth. So, it is not surprising that some bacterial species live with humans—a vast number of bacteria live in and on our bodies. Unfortunately, humans commonly think of bacteria as bad, because some types of bacteria can make us sick. Yes, there are bad guys—but there are also good guys! In fact, most bacterial species do *not* make us sick, and some even help us stay healthy [1, 2]. Most of the bacteria that live inside humans are found in the gut. These good bacteria protect us against disease-causing bacteria, help us digest food, and produce vitamins that our bodies need but cannot produce on their own [1]. Recently, researchers found that gut bacteria can even influence mental health [1]. Together, this collection of gut bacteria is known as the **gut microbiome**.

GUT MICROBIOME

All the bacteria that live in the human gut.

WHAT MAKES A HEALTHY GUT MICROBIOME?

How can we nurture good gut bacteria without strengthening the bad ones? One way is to feed our good gut bacteria their favorite foods. Fast foods, for example, even though they are very tasty, are good neither for us nor for our gut bacteria. Fast foods are full of sugar and low in vitamins and fiber. Most good gut bacteria prefer to be fed fiber, and if there is not enough, they can starve! It is important for the gut microbiome to contain many different types of bacteria, because a high **diversity** in the gut microbiome makes us more resistant to infections and to the effects of the occasional fast-food treat. Why might that be?

DIVERSITY

The variety of living things in a particular habitat. The more species, the higher the diversity.

Suppose the bacteria in the gut are very diverse. Some like to eat sugar, some fat, some fiber, and some protein. In that case, when there is no sugar around, only the sugar-eating bacteria become weak—the others are fine, which means that, as a whole, the community is still strong. However, if most bacteria only ate sugar, the microbiome would become weak in the absence of sugar, making it much easier for bad bacteria to conquer the gut and cause illness. This is one reason why a balanced diet is important.

There is another reason why a lack of diversity in the gut microbiome may be bad. When gut bacteria are diverse, multiple types of bacteria can perform the same job in the bacterial community, like digesting milk sugar or vitamin production. If one species of bacteria is lost, another might be able to take its place. In a less diverse gut microbiome, this might not be possible, and the gut microbiome may be weakened. One strategy for strengthening the gut microbiome, and thus human health, could be ensuring that there are several different bacterial species that can perform the same functions. We can achieve this by supplementing those bacteria through our diets.

CAN YOGURT OR CHEESE SUPPORT THE GUT MICROBIOME?

Bacteria can be found in many foods, particularly in dairy products like cheese or yogurt, which have been consumed for ages. Ten thousand years ago, when most people still lived as hunter-gatherers, all humans were lactose intolerant, which means they could drink milk when they were young, but if they did so as adults, they would feel bloated and sick. When humans started farming sheep, cows, and other milk-producing animals, they discovered ways to process milk so that it tasted different, lasted longer, and, most importantly, could be digested without making them feel sick [3]. We call this process **fermentation**, and it led to the production of the first yogurts and cheeses. What these early communities did not know was that bacteria were responsible for fermentation. Specific species of bacteria live and grow in milk, changing its taste, texture, and composition. After fermentation, a cup of yogurt (200 g) contains more bacteria (at least 20 billion) than there are humans on Earth (8 billion)!

Early yogurts contained many distinct species of bacteria, and every yogurt was different. In contrast, most modern yogurts only contain two species of bacteria, selected for fast and reliable mass production. Could we make yogurt from a different cocktail of bacteria, optimized not just for mass production but also to help the gut microbiome and improve human health (Figure 1)?

FERMENTATION

The production of foods such as yogurt, cheese, bread, kimchi, beer, and wine with the help of yeast or bacteria.

Figure 1

Can yogurt or cheese support the gut microbiome? **(A)** Unbalanced diets can lower the diversity of the gut microbiome, meaning that some species are present in very low numbers or even extinct. This can reduce the strength of the gut microbiome. **(B,C)** We are trying to find out if a dairy product fermented with selected bacteria can support or restore the functions of a healthy gut microbiome. **(D)** These functions include providing us with useful vitamins, sugars, and fats, and helping us to digest fiber from fruits and vegetables.

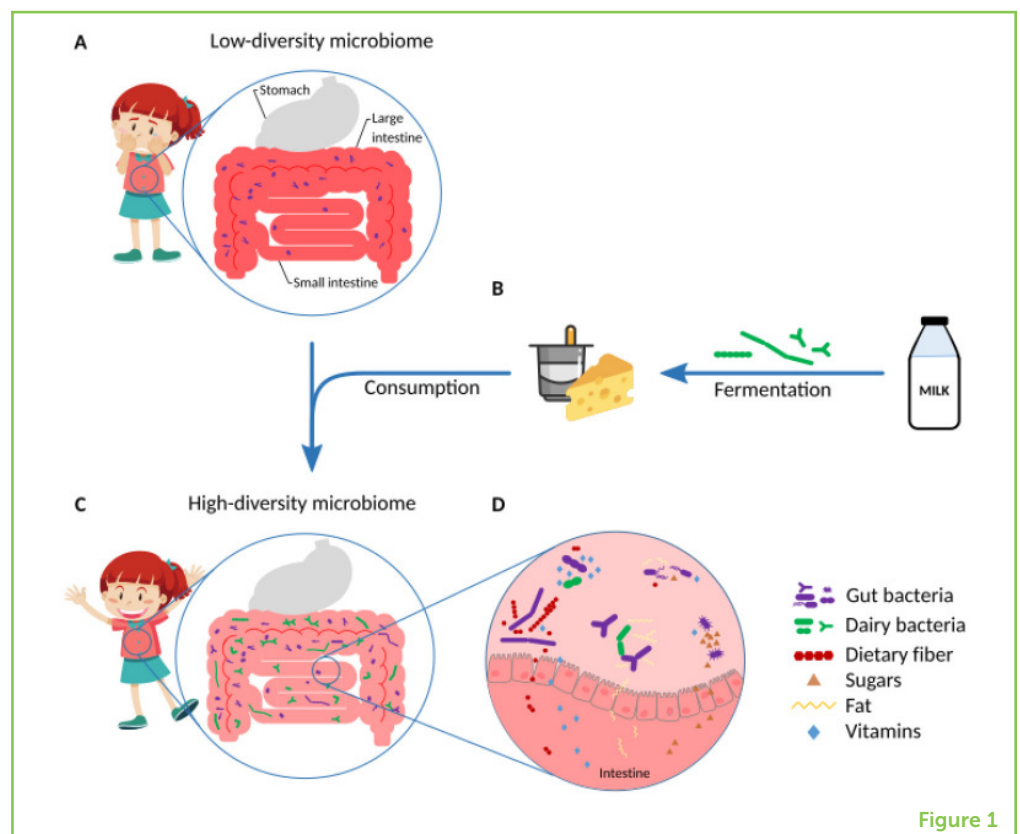


Figure 1

SEQUENCING

The process of studying the DNA composition of an organism.

GENE

A segment of DNA that determines a specific characteristic, capability, or function of a life form.

THE QUEST FOR THE RIGHT BACTERIA

Where could we get suitable bacteria for making gut-healthy yogurt? Switzerland has a proud tradition of cheese and yogurt making. Agroscope, the Swiss center of excellence for agricultural research, collects, stores, and investigates the bacteria found in yogurt and cheese. So far, Agroscope has collected over 10,000 different bacteria from dozens of species!

With such a large bacterial bank, how do we decide which are the best for a healthy yogurt? Conveniently, scientists at Agroscope have studied the DNA of close to 1,000 of the bacteria they have collected, using a process called **sequencing**, to see which **genes** the bacteria have. Genes are short segments of DNA that code for functions that allow bacteria to survive. For example, certain genes enable bacteria to split into two, and others give them the ability to swim around.

SUPPLEMENTING THE GUT MICROBIOME

Our idea was to see if yogurt bacteria could support gut bacteria in their work, and possibly even increase the diversity of the gut microbiome. We used DNA sequences from four human microbiomes, which contained a mix of genes from many different gut bacteria. Then, we used a computer program to determine the functions of these genes, using the same method we used for the dairy bacteria (Figure 2).

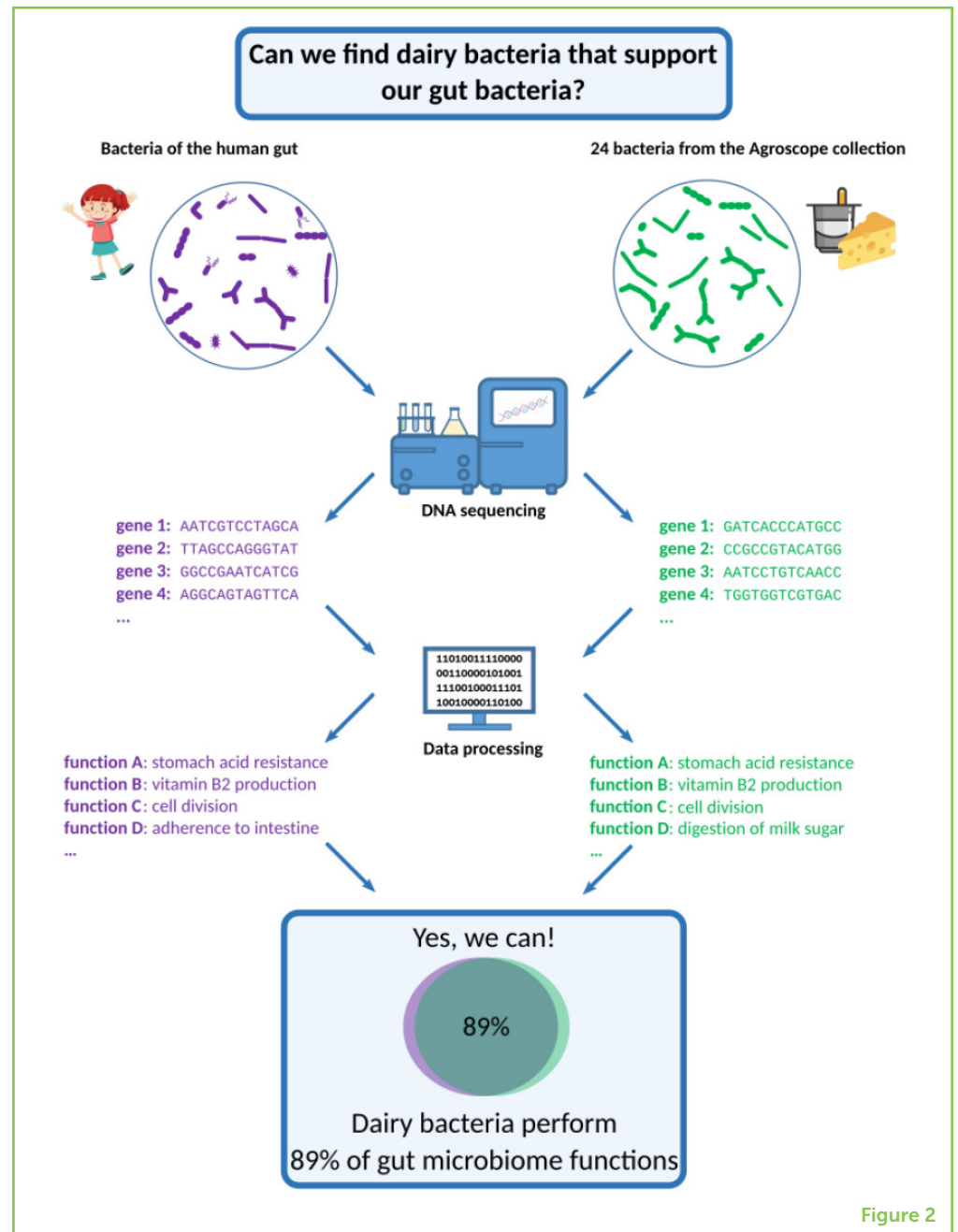
When we compared the dairy bacteria to bacteria from the human microbiome, we were surprised: each individual bacterial species from the milk could perform about half of the functions of the human gut microbiome, even though dairy bacteria are from a completely different environment and are rarely found in the human gut. Combined, our 24 species of dairy bacteria covered 89% of the functions of the human gut microbiome [4]! We also noticed that some human microbiomes lack certain functions compared to other human microbiomes. Some of those functions are present in dairy bacteria, meaning we may be able to develop a yogurt that can supplement missing or lost functions in a human microbiome, making it more resilient and thus helping people to be healthier.

WHAT IS NEXT?

Our study shows that bacteria from cheese or yogurt have similar functions to those of human gut bacteria. We think this knowledge will be very useful. For example, people with certain diseases, like obesity, lack specific types of gut bacteria. So, in future studies, we could recruit study participants with a known disease who lack the gut bacteria

Figure 2

The quest for the right bacteria. Using DNA sequencing followed by computer analysis of the data, we found that 24 dairy bacteria from Agroscope's bacteria bank (right) can perform most of the functions of the human gut microbiome (left). This finding may help us to design a yogurt that supports the functions of the gut microbiome, making it more resilient and thus promoting human health.



that perform certain functions. Then we could search Agroscope's collection for milk bacteria that have the missing functions and produce a special yogurt with them. Next, we could feed this yogurt to the participants and evaluate the effects on their health.

Humans have been raising cattle and eating fermented foods for millennia, but only recently have we gained the understanding and the tools to develop health-promoting dairy foods based on scientific data. This process is long, but worthwhile! If scientists continue to explore the potential of bacteria to improve human health, we may eventually be able to help many people with diseases like

obesity and diabetes by feeding those people foods that contain helpful bacteria.

FUNDING

This research was funded by Gebert R uf Stiftung within the program Microbials, Grant No. GRS-070/17.

REFERENCES

1. Wang, H., Wei, C. X., Min, L., and Zhu, L. Y. 2018. Good or bad: gut bacteria in human health and diseases. *Biotechnol. Biotechnol. Equipment*. 32:1075–80. doi: 10.1080/13102818.2018.1481350
2. Marco, M. L., Sanders, M. E., G anzle, M., Arrieta, M. C., Cotter, P. D., de Vuyst, L., et al. 2021. The international scientific association for probiotics and prebiotics (ISAPP) consensus statement on fermented foods. *Nat. Rev. Gastroenterol. Hepatol.* 18:96–208. doi: 10.1038/s41575-020-00390-5
3. Curry, A. 2013. The milk revolution. *Nature* 500:20–2. doi: 10.1038/500020a
4. Roder, T., W uthrich, D., B ar, C., Sattari, Z., von Ah, U., Ronchi, F., et al. 2020. *In silico* comparison shows that the pan-genome of a dairy-related bacterial culture collection covers most reactions annotated to human microbiomes. *Microorganisms*. 8:966. doi: 10.3390/microorganisms8070966

SUBMITTED: 07 June 2021; **ACCEPTED:** 26 May 2022;

PUBLISHED ONLINE: 21 June 2022.

EDITOR: Lorraine Brennan, University College Dublin, Ireland

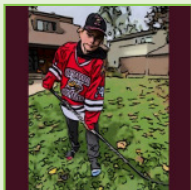
SCIENCE MENTORS: Emilio Isaac Alarcon and Wendy E. Huddleston

CITATION: Roder T, Pimentel G, B ar C, von Ah U, Bruggmann R and Verg eres G (2022) Can Eating Bacteria In Dairy Products Support Your Health? *Front. Young Minds* 10:721939. doi: 10.3389/frym.2022.721939

CONFLICT OF INTEREST: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

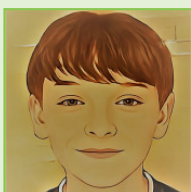
COPYRIGHT   2022 Roder, Pimentel, B ar, von Ah, Bruggmann and Verg eres. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

YOUNG REVIEWERS



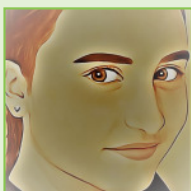
CAMERON, AGE: 10

I am 10 years old, I like to play sports, especially hockey. I like to read and play video games in my spare time. In the summer I like going to my cottage to swim and play baseball with my friends.



ELLIOT, AGE: 11

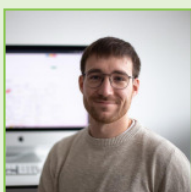
Elliot loves to read, play soccer and camp with his boy scout troop. He also has a blast going on adventures with his friend, Eve. Together, they like to rock climb, complete high ropes courses and downhill ski. Basically, they like to play hard and laugh hard!



EVE, AGE: 11

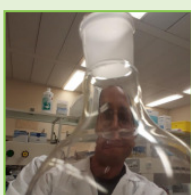
Eve loves to read, play softball and hang out with her friends. She also has a blast going on adventures with her friend, Elliot. Together, they like to rock climb, complete high ropes courses and downhill ski. Basically, they like to play hard and laugh hard!

AUTHORS



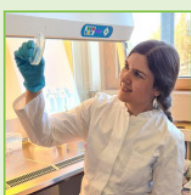
THOMAS RODER

Thomas is a Ph.D. student in bioinformatics at the university of Bern, Switzerland. He is trying to design new yogurts by combining various bacteria. At the same time, he is developing a website that makes comparing bacterial genomes easier. He works on these projects using computers, but prior to this project, he studied the interaction between plants and root-eating larvae in the lab. *thomas.roder@bioinformatics.unibe.ch



GRÉGORY PIMENTEL

Grégory is a researcher at the Functional Nutritional Biology Group at Agroscope. He specializes in the analysis of dairy products and other biological fluids (blood or urine) using a technique called metabolomics, which allows the detection of thousands of small compounds present in a sample. Metabolomics can help scientists better understand the chemical reactions happening in milk during fermentation, and it can be used to investigate the health effects of eating fermented dairy products. Grégory holds master's degrees in food science, engineering, and nutrition, and his Ph.D. is from the University of Lausanne in Switzerland, in cardiovascular biology and metabolism.



CORNELIA BÄR

Cornelia is a scientist in the Biochemistry of Milk and Microorganisms group at Agroscope, Switzerland. She was always driven by the desire to put scientific knowledge into practice and she earned her Ph.D. studying fortified foods. A postdoc studying food composition followed, which sparked her interest in how bacterial metabolism changes the composition of food. Trained in microbiology, immunobiology, and protein biochemistry, Cornelia is particularly interested

in which bacteria and proteins are responsible for food transformation, how bacteria interact in food, and how the consumption of these foods affects human health.



UELI VON AH

The combination of technology and biology has always been of great interest to Dr. Ueli von Ah. After studying food sciences, he earned a Ph.D. in food biotechnology. His work focuses on the use of lactic acid bacteria in food applications. In addition to finding the optimal growth conditions for these bacteria, he is also interested in understanding how genome information relates to the functions of bacteria in food. Ueli von Ah is now head of the Biotechnology research group at Agroscope, and he teaches a class in food biotechnology at a Swiss university of applied sciences.



RÉMY BRUGGMANN

Rémy is the head of the Bioinformatics Unit and director of studies of the master's of science program called Bioinformatics and Computational Biology at the University of Bern. A molecular biologist by training, he has always been interested in computer science, and in bioinformatics he found the ideal combination of his two passions. He has sequenced hundreds of genomes from bacteria and higher organisms and wants to better understand how genomic information is translated into a functioning organism.



GUY VERGÈRES

During the last three decades, Dr. Guy Vergères conducted research in several scientific disciplines, including chemistry, biochemistry, molecular biology, physical chemistry, pharmaceutical sciences, and microbiology—always with the aim of linking important molecules to their impact on human health. This combination naturally led him to conduct nutritional research on fermented foods. Guy Vergères is now heading the Functional Nutritional Biology research group at Agroscope and teaching the science of nutrigenomics (modern nutrition research) at Swiss universities.

6. References

1. Schlessinger D. NHGRI's Oral History Collection: Interview with David Schlessinger. 2018. https://youtu.be/N_0SUvzMTQ0?t=4489#https://www.genome.gov/sites/default/files/media/files/2019-05/david_schlessinger_transcript.pdf. Accessed 17 Dec 2022.
2. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci USA*. 1977;74:5463–7.
3. Maxam AM, Gilbert W. A new method for sequencing DNA. *Proc Natl Acad Sci USA*. 1977;74:560–4.
4. International Human Genome Sequencing Consortium, Whitehead Institute for Biomedical Research, Center for Genome Research, Lander ES, Linton LM, Birren B, Nusbaum C, et al. Initial sequencing and analysis of the human genome. *Nature*. 2001;409:860–921.
5. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. *Science*. 2001;291:1304–51.
6. Slatko BE, Gardner AF, Ausubel FM. Overview of Next-Generation Sequencing Technologies. *Curr Protoc Mol Biol*. 2018;122:e59.
7. Hayden EC. Technology: The \$1,000 genome. *Nature*. 2014;507:294–5.
8. NHGRI. The Cost of Sequencing a Human Genome. <http://genome.gov/sequencingcosts>. Accessed 17 Dec 2022.
9. Mardis ER. The \$1,000 genome, the \$100,000 analysis? *Genome Med*. 2010;2:84.
10. Marx V. Method of the year: long-read sequencing. *Nat Methods*. 2023;20:6–11.
11. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, et al. Real-time DNA sequencing from single polymerase molecules. *Science*. 2009;323:133–8.
12. Heather JM, Chain B. The sequence of sequencers: The history of sequencing DNA. *Genomics*. 2016;107:1–8.
13. Wenger AM, Peluso P, Rowell WJ, Chang P-C, Hall RJ, Concepcion GT, et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat Biotechnol*. 2019;37:1155–62.
14. Simpson JT, Pop M. The theory and practice of genome sequence assembly. *Annu Rev Genomics Hum Genet*. 2015;16:153–72.
15. Compeau PEC, Pevzner PA, Tesler G. How to apply de Bruijn graphs to genome assembly. *Nat Biotechnol*. 2011;29:987–91.
16. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol*. 2012;19:455–77.
17. Wick RR, Judd LM, Holt KE. Assembling the perfect bacterial genome using Oxford Nanopore and Illumina sequencing. 2022.
18. Koren S, Harhay GP, Smith TPL, Bono JL, Harhay DM, Mcvey SD, et al. Reducing assembly complexity of microbial genomes with single-molecule sequencing. *Genome Biol*. 2013;14:R101.

19. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* 2017;27:722–36.
20. Ruan J, Li H. Fast and accurate long-read assembly with wtdbg2. *Nat Methods.* 2020;17:155–8.
21. Li H. Minimap and miniiasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics.* 2016;32:2103–10.
22. Liu H, Wu S, Li A, Ruan J. SMARTdenovo: A de novo Assembler Using Long Noisy Reads. 2020.
23. Vaser R, Šikić M. Time- and memory-efficient genome assembly with Raven. *Nat Comput Sci.* 2021.
24. Chen Y, Nie F, Xie S-Q, Zheng Y-F, Dai Q, Bray T, et al. Efficient assembly of nanopore reads via highly accurate and intact error correction. *Nat Commun.* 2021;12:60.
25. Chin C-S, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, et al. Phased diploid genome assembly with single-molecule real-time sequencing. *Nat Methods.* 2016;13:1050–4.
26. Nextomics. Nextomics/NextDenovo: Fast and accurate de novo assembler for long reads. 2019. <https://github.com/Nextomics/NextDenovo>. Accessed 20 Dec 2022.
27. Lin Y, Yuan J, Kolmogorov M, Shen MW, Chaisson M, Pevzner PA. Assembly of long error-prone reads using de Bruijn graphs. *Proc Natl Acad Sci USA.* 2016;113:E8396–405.
28. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol.* 2019;37:540–6.
29. Chin C-S, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat Methods.* 2013;10:563–9.
30. Ekim B, Berger B, Chikhi R. Minimizer-space de Bruijn graphs: Whole-genome assembly of long reads in minutes on a personal computer (mdBG). *Cell Syst.* 2021;12:958-968.e6.
31. Bankevich A, Bzikadze AV, Kolmogorov M, Antipov D, Pevzner PA. Multiplex de Bruijn graphs enable genome assembly from long, high-fidelity reads. *Nat Biotechnol.* 2022;40:1075–81.
32. Wick RR, Holt KE. Benchmarking of long-read assemblers for prokaryote whole genome sequencing. *F1000Res.* 2021;8:2138.
33. Wick RR, Judd LM, Cerdeira LT, Hawkey J, Méric G, Vezina B, et al. Trycycler: consensus long-read assemblies for bacterial genomes. *Genome Biol.* 2021;22:266.
34. Salzberg SL, Phillippy AM, Zimin A, Puiu D, Magoc T, Koren S, et al. GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome Res.* 2012;22:557–67.
35. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUASt: quality assessment tool for genome assemblies. *Bioinformatics.* 2013;29:1072–5.
36. Mikheenko A, Pribelski A, Saveliev V, Antipov D, Gurevich A. Versatile genome assembly evaluation with QUASt-LG. *Bioinformatics.* 2018;34:i142–50.
37. Wang J, Chen K, Ren Q, Zhang Y, Liu J, Wang G, et al. Systematic comparison of the performances of de novo genome assemblers for oxford nanopore technology reads from piroplasm. *Front Cell Infect Microbiol.* 2021;11:696669.

38. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*. 2015;31:3210–2.
39. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. *Genome Biol*. 2019;20:257.
40. Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics*. 2019;36:1925–7.
41. Low AJ, Koziol AG, Manninger PA, Blais B, Carrillo CD. ConFindr: rapid detection of intraspecies and cross-species contamination in bacterial whole-genome sequence data. *PeerJ*. 2019;7:e6995.
42. Salzberg SL, Delcher AL, Kasif S, White O. Microbial gene identification using interpolated Markov models. *Nucleic Acids Res*. 1998;26:544–8.
43. Salzberg SL. Next-generation genome annotation: we still struggle to get it right. *Genome Biol*. 2019;20:92.
44. Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*. 2010;11:119.
45. Lomsadze A, Gemayel K, Tang S, Borodovsky M. Modeling leaderless transcription and atypical genes results in more accurate gene prediction in prokaryotes. *Genome Res*. 2018;28:1079–89.
46. Korandla DR, Wozniak JM, Campeau A, Gonzalez DJ, Wright ES. AssessORF: combining evolutionary conservation and proteomics to assess prokaryotic gene predictions. *Bioinformatics*. 2020;36:1022–9.
47. Dimonaco NJ, Aubrey W, Kenobi K, Clare A, Creevey CJ. No one tool to rule them all: Prokaryotic gene prediction tool annotations are highly dependent on the organism of study. *Bioinformatics*. 2021;38:1198–207.
48. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*. 2014;30:2068–9.
49. Tatusova T, DiCuccio M, Badretdin A, Chetvernin V, Nawrocki EP, Zaslavsky L, et al. NCBI prokaryotic genome annotation pipeline. *Nucleic Acids Res*. 2016;44:6614–24.
50. Haft DH, DiCuccio M, Badretdin A, Brover V, Chetvernin V, O'Neill K, et al. RefSeq: an update on prokaryotic genome annotation and curation. *Nucleic Acids Res*. 2018;46:D851–60.
51. Li W, O'Neill KR, Haft DH, DiCuccio M, Chetvernin V, Badretdin A, et al. RefSeq: expanding the Prokaryotic Genome Annotation Pipeline reach with protein family model curation. *Nucleic Acids Res*. 2021;49:D1020–8.
52. Schwengers O, Jelonek L, Dieckmann MA, Beyvers S, Blom J, Goesmann A. Bakta: rapid and standardized annotation of bacterial genomes via alignment-free sequence identification. *Microb Genom*. 2021;7.
53. Poux S, Arighi CN, Magrane M, Bateman A, Wei C-H, Lu Z, et al. On expert curation and scalability: UniProtKB/Swiss-Prot as a case study. *Bioinformatics*. 2017;33:3454–60.
54. Vanni C, Schechter MS, Acinas SG, Barberán A, Buttigieg PL, Casamayor EO, et al. Unifying the known and unknown microbial coding sequence space. *eLife*. 2022;11.
55. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res*. 2016;44:D457–62.

56. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology The Gene Ontology Consortium*. 2000.
57. Webb EC. Enzyme nomenclature 1992: recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the nomenclature and classification of enzymes. Academic Press. 1992.
58. Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M. KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res.* 2007;35 Web Server issue:W182-5.
59. Altenhoff AM, Studer RA, Robinson-Rechavi M, Dessimoz C. Resolving the ortholog conjecture: orthologs tend to be weakly, but significantly, more similar in function than paralogs. *PLoS Comput Biol.* 2012;8:e1002514.
60. Altenhoff AM, Glover NM, Dessimoz C. Inferring orthology and paralogy. *Methods Mol Biol.* 2019;1910:149–75.
61. Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* 2019;47:D309–14.
62. Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Mol Biol Evol.* 2021;38:5825–9.
63. Huerta-Cepas J, Forslund K, Coelho LP, Szklarczyk D, Jensen LJ, von Mering C, et al. Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. *Mol Biol Evol.* 2017;34:2115–22.
64. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* 2019;20:238.
65. Nevers Y, Jones TEM, Jyothi D, Yates B, Ferret M, Portell-Silva L, et al. The Quest for Orthologs orthology benchmark service in 2022. *Nucleic Acids Res.* 2022.
66. Seemann T. ABRicate: mass screening of contigs for antibiotic resistance genes. ABRicate. 2014. <https://github.com/tseemann/abricate>. Accessed 29 Dec 2022.
67. Drula E, Garron M-L, Dogan S, Lombard V, Henrissat B, Terrapon N. The carbohydrate-active enzyme database: functions and literature. *Nucleic Acids Res.* 2022;50:D571–7.
68. Helander HF, Fändriks L. Surface area of the digestive tract - revisited. *Scand J Gastroenterol.* 2014;49:681–9.
69. Chassaing B, Kumar M, Baker MT, Singh V, Vijay-Kumar M. Mammalian gut immunity. *Biomed J.* 2014;37:246–58.
70. Davis CD. The gut microbiome and its role in obesity. *Nutr Today.* 2016;51:167–74.
71. Hooper LV, Littman DR, Macpherson AJ. Interactions between the microbiota and the immune system. *Science.* 2012;336:1268–73.
72. Clarke G, Stilling RM, Kennedy PJ, Stanton C, Cryan JF, Dinan TG. Minireview: Gut microbiota: the neglected endocrine organ. *Mol Endocrinol.* 2014;28:1221–38.
73. Cryan JF, O’Riordan KJ, Cowan CSM, Sandhu KV, Bastiaanssen TFS, Boehme M, et al. The Microbiota-Gut-Brain Axis. *Physiol Rev.* 2019;99:1877–2013.

74. Neves AL, Chilloux J, Sarafian MH, Rahim MBA, Boulangé CL, Dumas M-E. The microbiome and its pharmacological targets: therapeutic avenues in cardiometabolic diseases. *Curr Opin Pharmacol.* 2015;25:36–44.
75. Jackson MA, Verdi S, Maxan M-E, Shin CM, Zierer J, Bowyer RCE, et al. Gut microbiota associations with common diseases and prescription medications in a population-based cohort. *Nat Commun.* 2018;9:2655.
76. Magne F, Gotteland M, Gauthier L, Zazueta A, Pesoa S, Navarrete P, et al. The firmicutes/bacteroidetes ratio: A relevant marker of gut dysbiosis in obese patients? *Nutrients.* 2020;12.
77. Lee J-Y, Tsolis RM, Bäumlér AJ. The microbiome and gut homeostasis. *Science.* 2022;377:eabp9960.
78. Hitch TCA, Hall LJ, Walsh SK, Leventhal GE, Slack E, de Wouters T, et al. Microbiome-based interventions to modulate gut ecology and the immune system. *Mucosal Immunol.* 2022;15:1095–113.
79. Shannon CE. A mathematical theory of communication. *Bell System Technical Journal.* 1948;27:379–423.
80. Simpson EH. Measurement of Diversity. *Nature.* 1949;163:688–688.
81. Cani PD. Gut microbiota - at the intersection of everything? *Nat Rev Gastroenterol Hepatol.* 2017;14:321–2.
82. Turnbaugh PJ, Ley RE, Mahowald MA, Magrini V, Mardis ER, Gordon JI. An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature.* 2006;444:1027–31.
83. Turnbaugh PJ, Bäckhed F, Fulton L, Gordon JI. Diet-induced obesity is linked to marked but reversible alterations in the mouse distal gut microbiome. *Cell Host Microbe.* 2008;3:213–23.
84. Ridaura VK, Faith JJ, Rey FE, Cheng J, Duncan AE, Kau AL, et al. Gut microbiota from twins discordant for obesity modulate metabolism in mice. *Science.* 2013;341:1241214.
85. Garrett WS, Lord GM, Punit S, Lugo-Villarino G, Mazmanian SK, Ito S, et al. Communicable ulcerative colitis induced by T-bet deficiency in the innate immune system. *Cell.* 2007;131:33–45.
86. Schaubeck M, Clavel T, Calasan J, Lagkouvardos I, Haange SB, Jehmlich N, et al. Dysbiotic gut microbiota causes transmissible Crohn’s disease-like ileitis independent of failure in antimicrobial defence. *Gut.* 2016;65:225–37.
87. Rothschild D, Weissbrod O, Barkan E, Kurilshikov A, Korem T, Zeevi D, et al. Environment dominates over host genetics in shaping human gut microbiota. *Nature.* 2018;555:210–5.
88. Strachan DP. Hay fever, hygiene, and household size. *BMJ.* 1989;299:1259–60.
89. Torow N, Homef MW. The Neonatal Window of Opportunity: Setting the Stage for Life-Long Host-Microbial Interaction and Immune Homeostasis. *J Immunol.* 2017;198:557–63.
90. Hornef MW, Torow N. “Layered immunity” and the “neonatal window of opportunity” - timed succession of non-redundant phases to establish mucosal host-microbial homeostasis after birth. *Immunology.* 2020;159:15–25.
91. Rinninella E, Cintoni M, Raoul P, Lopetuso LR, Scaldaferri F, Pulcini G, et al. Food components and dietary habits: keys for a healthy gut microbiota composition. *Nutrients.* 2019;11.

92. Wu GD, Chen J, Hoffmann C, Bittinger K, Chen Y-Y, Keilbaugh SA, et al. Linking long-term dietary patterns with gut microbial enterotypes. *Science*. 2011;334:105–8.
93. Leeming ER, Johnson AJ, Spector TD, Le Roy CI. Effect of diet on the gut microbiota: rethinking intervention duration. *Nutrients*. 2019;11.
94. Gille D, Schmid A, Walther B, Vergères G. Fermented Food and Non-Communicable Chronic Diseases: A Review. *Nutrients*. 2018;10.
95. Wastyk HC, Fragiadakis GK, Perelman D, Dahan D, Merrill BD, Yu FB, et al. Gut-microbiota-targeted diets modulate human immune status. *Cell*. 2021;184:4137–4153.e14.
96. Sonnenburg ED, Smits SA, Tikhonov M, Higginbottom SK, Wingreen NS, Sonnenburg JL. Diet-induced extinctions in the gut microbiota compound over generations. *Nature*. 2016;529:212–5.
97. Zinöcker MK, Lindseth IA. The Western Diet-Microbiome-Host Interaction and Its Role in Metabolic Disease. *Nutrients*. 2018;10.
98. Nutrition Division FAO/WHO. Probiotics in food: Health and nutritional properties and guidelines for evaluation. Joint FAO/WHO Expert Consultation. ... of the United Nations and World ...; 2006.
99. Homayouni Rad A, Yari Khosroushahi A, Khalili M, Jafarzadeh S. Folate bio-fortification of yoghurt and fermented milk: a review. *Dairy Sci Technol*. 2016;96:427–41.
100. Savaiano DA, Hutkins RW. Yogurt, cultured fermented milk, and health: a systematic review. *Nutr Rev*. 2021;79:599–614.
101. Zotta T, Parente E, Ricciardi A. Aerobic metabolism in the genus *Lactobacillus*: impact on stress response and potential applications in the food industry. *J Appl Microbiol*. 2017;122:857–69.
102. Gänzle MG. Lactic metabolism revisited: metabolism of lactic acid bacteria in food fermentations and food spoilage. *Current Opinion in Food Science*. 2015;2:106–17.
103. Duar RM, Lin XB, Zheng J, Martino ME, Grenier T, Pérez-Muñoz ME, et al. Lifestyles in transition: evolution and natural history of the genus *Lactobacillus*. *FEMS Microbiol Rev*. 2017;41 Supp_1:S27–48.
104. Makarova KS, Koonin EV. Evolutionary genomics of lactic acid bacteria. *J Bacteriol*. 2007;189:1199–208.
105. Gu S, Chen D, Zhang J-N, Lv X, Wang K, Duan L-P, et al. Bacterial community mapping of the mouse gastrointestinal tract. *PLoS ONE*. 2013;8:e74957.
106. Li C, Bei T, Niu Z, Guo X, Wang M, Lu H, et al. Adhesion and Colonization of the Probiotic *Lactobacillus rhamnosus* Labeled by Dsred2 in Mouse Gut. *Curr Microbiol*. 2019;76:896–903.
107. Xing Z, Tang W, Yang Y, Geng W, Rehman RU, Wang Y. Colonization and Gut Flora Modulation of *Lactobacillus kefirifaciens* ZW3 in the Intestinal Tract of Mice. *Probiotics Antimicrob Proteins*. 2018;10:374–82.
108. Nemcová R, Bomba A, Herich R, Gancarcíková S. Colonization capability of orally administered *Lactobacillus* strains in the gut of gnotobiotic piglets. *DTW Dtsch Tierarztl Wochenschr*. 1998;105:199–200.
109. De Filippis F, Pasolli E, Ercolini D. The food-gut axis: lactic acid bacteria and their link to food, the gut microbiome and human health. *FEMS Microbiol Rev*. 2020;44:454–89.

110. Aryana KJ, Olson DW. A 100-Year Review: Yogurt and other cultured dairy products. *J Dairy Sci.* 2017;100:9987–10013.
111. Yu J, Wang HM, Zha MS, Qing YT, Bai N, Ren Y, et al. Molecular identification and quantification of lactic acid bacteria in traditional fermented dairy foods of Russia. *J Dairy Sci.* 2015;98:5143–54.
112. Ruegg M. The swiss federal dairy research station. *Int J Dairy Technol.* 2003;56:2–5.
113. Vergères G. Nutrigenomics – Linking food to human metabolism. *Trends Food Sci Technol.* 2013;31:6–12.
114. Papadimitriou K, Zoumpopoulou G, Foligné B, Alexandraki V, Kazou M, Pot B, et al. Discovering probiotic microorganisms: in vitro, in vivo, genetic and omics approaches. *Front Microbiol.* 2015;6:58.
115. Taylor BC, Lejzerowicz F, Poirel M, Shaffer JP, Jiang L, Aksenov A, et al. Consumption of Fermented Foods Is Associated with Systematic Differences in the Gut Microbiome and Metabolome. *mSystems.* 2020;5.
116. Larigot L, Benoit L, Koual M, Tomkiewicz C, Barouki R, Coumoul X. Aryl hydrocarbon receptor and its diverse ligands and functions: an exposome receptor. *Annu Rev Pharmacol Toxicol.* 2022;62:383–404.
117. Rothhammer V, Quintana FJ. The aryl hydrocarbon receptor: an environmental sensor integrating immune responses in health and disease. *Nat Rev Immunol.* 2019;19:184–97.
118. Denison MS, Nagy SR. Activation of the aryl hydrocarbon receptor by structurally diverse exogenous and endogenous chemicals. *Annu Rev Pharmacol Toxicol.* 2003;43:309–34.
119. Jin U-H, Lee S-O, Sridharan G, Lee K, Davidson LA, Jayaraman A, et al. Microbiome-derived tryptophan metabolites and their aryl hydrocarbon receptor-dependent agonist and antagonist activities. *Mol Pharmacol.* 2014;85:777–88.
120. Ehrlich AK, Pennington JM, Bisson WH, Kolluri SK, Kerkvliet NI. TCDD, FICZ, and Other High Affinity AhR Ligands Dose-Dependently Determine the Fate of CD4⁺ T Cell Differentiation. *Toxicol Sci.* 2018;161:310–20.
121. Stepankova M, Bartonkova I, Jiskrova E, Vrzal R, Mani S, Kortagere S, et al. Methylindoles and methoxyindoles are agonists and antagonists of human aryl hydrocarbon receptor. *Mol Pharmacol.* 2018;93:631–44.
122. Kim DJ, Venkataraman A, Jain PC, Wiesler EP, DeBlasio M, Klein J, et al. Vitamin B12 and folic acid alleviate symptoms of nutritional deficiency by antagonizing aryl hydrocarbon receptor. *Proc Natl Acad Sci USA.* 2020;117:15837–45.
123. Dabir P, Marinic TE, Krukovets I, Stenina OI. Aryl hydrocarbon receptor is activated by glucose and regulates the thrombospondin-1 gene promoter in endothelial cells. *Circ Res.* 2008;102:1558–65.
124. Flaveny CA, Murray IA, Perdew GH. Differential gene regulation by the human and mouse aryl hydrocarbon receptor. *Toxicol Sci.* 2010;114:217–25.
125. Bansal T, Alaniz RC, Wood TK, Jayaraman A. The bacterial signal indole increases epithelial-cell tight-junction resistance and attenuates indicators of inflammation. *Proc Natl Acad Sci USA.* 2010;107:228–33.

126. Roh E, Kwak SH, Jung HS, Cho YM, Pak YK, Park KS, et al. Serum aryl hydrocarbon receptor ligand activity is associated with insulin resistance and resulting type 2 diabetes. *Acta Diabetol.* 2015;52:489–95.
127. Zhao R-X, He Q, Sha S, Song J, Qin J, Liu P, et al. Increased AHR Transcripts Correlate With Pro-inflammatory T-Helper Lymphocytes Polarization in Both Metabolically Healthy Obesity and Type 2 Diabetic Patients. *Front Immunol.* 2020;11:1644.
128. Lamas B, Richard ML, Leducq V, Pham H-P, Michel M-L, Da Costa G, et al. CARD9 impacts colitis by altering gut microbiota metabolism of tryptophan into aryl hydrocarbon receptor ligands. *Nat Med.* 2016;22:598–605.
129. Kiss EA, Vonarbourg C, Kopfmann S, Hobeika E, Finke D, Esser C, et al. Natural aryl hydrocarbon receptor ligands control organogenesis of intestinal lymphoid follicles. *Science.* 2011;334:1561–5.
130. Gomez de Agüero M, Ganal-Vonarburg SC, Fuhrer T, Rupp S, Uchimura Y, Li H, et al. The maternal microbiota drives early postnatal innate immune development. *Science.* 2016;351:1296–302.
131. Rejano-Gordillo CM, González-Rico FJ, Marín-Díaz B, Ordiales-Talavera A, Nacarino-Palma A, Román AC, et al. Liver regeneration after partial hepatectomy is improved in the absence of aryl hydrocarbon receptor. *Sci Rep.* 2022;12:15446.
132. Illés P, Krasulová K, Vyhliđalová B, Poulíková K, Marcalíková A, Pečínková P, et al. Indole microbial intestinal metabolites expand the repertoire of ligands and agonists of the human pregnane X receptor. *Toxicol Lett.* 2020;334:87–93.
133. Dvořák Z, Sokol H, Mani S. Drug mimicry: promiscuous receptors PXR and ahr, and microbial metabolite interactions in the intestine. *Trends Pharmacol Sci.* 2020;41:900–8.
134. Puccetti M, Pariano M, Costantini C, Giovagnoli S, Ricci M. Pharmaceutically active microbial ahr agonists as innovative biodrugs in inflammation. *Pharmaceuticals (Basel).* 2022;15.
135. Agus A, Planchais J, Sokol H. Gut microbiota regulation of tryptophan metabolism in health and disease. *Cell Host Microbe.* 2018;23:716–24.
136. Bertazzo A, Ragazzi E, Visioli F. Evolution of tryptophan and its foremost metabolites' concentrations in milk and fermented dairy products. *PharmaNutrition.* 2016;4:62–7.
137. Montgomery TL, Eckstrom K, Lile KH, Caldwell S, Heney ER, Lahue KG, et al. *Lactobacillus reuteri* tryptophan metabolism promotes host susceptibility to CNS autoimmunity. *Microbiome.* 2022;10:198.
138. Burton KJ, Rosikiewicz M, Pimentel G, Bütikofer U, von Ah U, Voirol M-J, et al. Probiotic yogurt and acidified milk similarly reduce postprandial inflammation and both alter the gut microbiota of healthy, young men. *Br J Nutr.* 2017;117:1312–22.
139. Burton KJ, Pimentel G, Zangger N, Vionnet N, Draï J, McTernan PG, et al. Modulation of the peripheral blood transcriptome by the ingestion of probiotic yoghurt and acidified milk in healthy, young men. *PLoS ONE.* 2018;13:e0192947.
140. Pimentel G, Burton KJ, Pralong FP, Vionnet N, Portmann R, Vergères G. The postprandial metabolome — a source of Nutritional Biomarkers of Health. *Current Opinion in Food Science.* 2017;16:67–73.
141. Hou Q, Ye L, Liu H, Huang L, Yang Q, Turner JR, et al. *Lactobacillus* accelerates ISCs regeneration to protect the integrity of intestinal mucosa through activation of STAT3 signaling pathway induced by LPLs secretion of IL-22. *Cell Death Differ.* 2018;25:1657–70.

142. Takamura T, Harama D, Fukumoto S, Nakamura Y, Shimokawa N, Ishimaru K, et al. *Lactobacillus bulgaricus* OLL1181 activates the aryl hydrocarbon receptor pathway and inhibits colitis. *Immunol Cell Biol.* 2011;89:817–22.
143. Fang Z, Pan T, Li L, Wang H, Zhu J, Zhang H, et al. *Bifidobacterium longum* mediated tryptophan metabolism to improve atopic dermatitis via the gut-skin axis. *Gut Microbes.* 2022;14:2044723.
144. Hasegawa Y, Raghuvanshi R, Bolling B. Fermentation increases AhR ligands in yogurt that may prevent inflammatory intestinal barrier dysfunction. *DIH.* 2022.
145. Hapfelmeier S, Lawson MAE, Slack E, Kirundi JK, Stoel M, Heikenwalder M, et al. Reversible microbial colonization of germ-free mice reveals the dynamics of IgA immune responses. *Science.* 2010;328:1705–9.
146. Tibbetts AS, Appling DR. Compartmentalization of Mammalian folate-mediated one-carbon metabolism. *Annu Rev Nutr.* 2010;30:57–81.
147. Ducker GS, Rabinowitz JD. One-Carbon Metabolism in Health and Disease. *Cell Metab.* 2017;25:27–42.
148. Boghog. File:Folate family.svg - Wikimedia Commons. File:Folate family.svg - Wikimedia Commons. 2019. https://commons.wikimedia.org/wiki/File:Folate_family.svg. Accessed 8 Dec 2022.
149. EU JRC. Fortification with folic acid could help prevent severe birth defects in at least 1000 pregnancies per year. News announcement. 2021.
150. Morris JK, Addor M-C, Ballardini E, Barisic I, Barrachina-Bonet L, Braz P, et al. Prevention of neural tube defects in europe: A public health failure. *Front Pediatr.* 2021;9:647038.
151. Sybesma W, Starrenburg M, Tijsseling L, Hoefnagel MHN, Hugenholtz J. Effects of cultivation conditions on folate production by lactic acid bacteria. *Appl Environ Microbiol.* 2003;69:4542–8.
152. Laiño JE, Juarez del Valle M, Savoy de Giori G, LeBlanc JGJ. Development of a high folate concentration yogurt naturally bio-enriched using selected lactic acid bacteria. *LWT - Food Science and Technology.* 2013;54:1–5.
153. Troen AM, Mitchell B, Sorensen B, Wener MH, Johnston A, Wood B, et al. Unmetabolized folic acid in plasma is associated with reduced natural killer cell cytotoxicity among postmenopausal women. *J Nutr.* 2006;136:189–94.
154. Menezo Y, Elder K, Clement A, Clement P. Folic Acid, Folinic Acid, 5 Methyl TetraHydroFolate Supplementation for Mutations That Affect Epigenesis through the Folate and One-Carbon Cycles. *Biomolecules.* 2022;12.
155. Courtemanche C, Elson-Schwab I, Mashiyama ST, Kerry N, Ames BN. Folate deficiency inhibits the proliferation of primary human CD8+ T lymphocytes in vitro. *J Immunol.* 2004;173:3186–92.
156. Yamaguchi T, Hirota K, Nagahama K, Ohkawa K, Takahashi T, Nomura T, et al. Control of immune responses by antigen-specific regulatory T cells expressing the folate receptor. *Immunity.* 2007;27:145–59.
157. Samblas M, Martínez JA, Milagro F. Folic Acid Improves the Inflammatory Response in LPS-Activated THP-1 Macrophages. *Mediators Inflamm.* 2018;2018:1312626.
158. Iyer R, Tomar SK. Determination of folate/folic acid level in milk by microbiological assay, immuno assay and high performance liquid chromatography. *J Dairy Res.* 2013;80:233–9.

159. Kilcast D. Effect of irradiation on vitamins. *Food Chem.* 1994;49:157–64.
160. Bellman R. *Dynamic Programming*. First Edition. Princeton University Press; 1957.
161. Subramanian I, Verma S, Kumar S, Jere A, Anamika K. Multi-omics Data Integration, Interpretation, and Its Application. *Bioinform Biol Insights.* 2020;14:1177932219899051.
162. Singh A, Shannon CP, Gautier B, Rohart F, Vacher M, Tebbutt SJ, et al. DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays. *Bioinformatics.* 2019;35:3055–62.
163. Kastenmüller G, Raffler J, Gieger C, Suhre K. Genetics of human metabolism: an update. *Hum Mol Genet.* 2015;24:R93–101.
164. Brynildsrud O, Bohlin J, Scheffer L, Eldholm V. Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. *Genome Biol.* 2016;17:238.
165. Maddison WP, FitzJohn RG. The unsolved challenge to phylogenetic correlation tests for categorical characters. *Syst Biol.* 2015;64:127–36.
166. San JE, Baichoo S, Kanzi A, Moosa Y, Lessells R, Fonseca V, et al. Current Affairs of Microbial Genome-Wide Association Studies: Approaches, Bottlenecks and Analytical Pitfalls. *Front Microbiol.* 2019;10:3119.
167. Buron-Moles G, Chailyan A, Dolejs I, Forster J, Mikš MH. Uncovering carbohydrate metabolism through a genotype-phenotype association study of 56 lactic acid bacteria genomes. *Appl Microbiol Biotechnol.* 2019;103:3135–52.
168. Palsson B, Zengler K. The challenges of integrating multi-omic data sets. *Nat Chem Biol.* 2010;6:787–9.
169. Fondi M, Liò P. Multi -omics and metabolic modelling pipelines: challenges and tools for systems microbiology. *Microbiol Res.* 2015;171:52–64.
170. Gudmundsson S, Agudo L, Nogales J. Applications of genome-scale metabolic models of microalgae and cyanobacteria in biotechnology. In: *Microalgae-Based Biofuels and Bioproducts*. Elsevier; 2017. p. 93–111.
171. Hernández-Plaza A, Szklarczyk D, Botas J, Cantalapiedra CP, Giner-Lamia J, Mende DR, et al. eggNOG 6.0: enabling comparative genomics across 12 535 organisms. *Nucleic Acids Res.* 2022.
172. Machado D, Andrejev S, Tramontano M, Patil KR. Fast automated reconstruction of genome-scale metabolic models for microbial species and communities. *Nucleic Acids Res.* 2018;46:7542–53.
173. Mendoza SN, Olivier BG, Molenaar D, Teusink B. A systematic assessment of current genome-scale metabolic reconstruction tools. *Genome Biol.* 2019;20:158.
174. Raman K, Chandra N. Flux balance analysis of biological systems: applications and challenges. *Brief Bioinformatics.* 2009;10:435–49.
175. Magnúsdóttir S, Heinken A, Kutt L, Ravcheev DA, Bauer E, Noronha A, et al. Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota. *Nat Biotechnol.* 2017;35:81–9.
176. Rosato A, Tenori L, Cascante M, De Atauri Carulla PR, Martins Dos Santos VAP, Saccenti E. From correlation to causation: analysis of metabolomics data using systems biology approaches. *Metabolomics.* 2018;14:37.

177. Fang X, Lloyd CJ, Palsson BO. Reconstructing organisms in silico: genome-scale models and their emerging applications. *Nat Rev Microbiol.* 2020;18:731–43.
178. Diener C, Gibbons SM, Resendis-Antonio O. MICOM: Metagenome-Scale Modeling To Infer Metabolic Interactions in the Gut Microbiota. *mSystems.* 2020;5.
179. San León D, Nogales J. Toward merging bottom-up and top-down model-based designing of synthetic microbial communities. *Curr Opin Microbiol.* 2022;69:102169.
180. Freilich S, Zarecki R, Eilam O, Segal ES, Henry CS, Kupiec M, et al. Competitive and cooperative metabolic interactions in bacterial communities. *Nat Commun.* 2011;2:589.
181. Pacheco AR, Moel M, Segrè D. Costless metabolic secretions as drivers of interspecies interactions in microbial ecosystems. *Nat Commun.* 2019;10:103.
182. Zomorodi AR, Maranas CD. OptCom: a multi-level optimization framework for the metabolic modeling and analysis of microbial communities. *PLoS Comput Biol.* 2012;8:e1002363.
183. Sieuwerts S, de Bok FAM, Hugenholtz J, van Hylckama Vlieg JET. Unraveling microbial interactions in food fermentations: from classical to genomics approaches. *Appl Environ Microbiol.* 2008;74:4997–5007.
184. Aghababae M, Khanahmadi M, Beheshti M. Developing a kinetic model for co-culture of yogurt starter bacteria growth in pH controlled batch fermentation. *J Food Eng.* 2015;166:72–9.
185. Özcan E, Seven M, Şirin B, Çakır T, Nikerel E, Teusink B, et al. Dynamic co-culture metabolic models reveal the fermentation dynamics, metabolic capacities and interplays of cheese starter cultures. *Biotechnol Bioeng.* 2021;118:223–37.
186. Hanemaaijer MJ. The effect of experimental evolution of a yoghurt culture on growth and metabolic interactions. Doctoral dissertation. Vrije Universiteit Amsterdam; 2016.
187. Somerville V, Grigaitis P, Battjes J, Moro F, Teusink B. Use and limitations of genome-scale metabolic models in food microbiology. *Current Opinion in Food Science.* 2022;43:225–31.
188. Hanemaaijer M, Röling WFM, Olivier BG, Khandelwal RA, Teusink B, Bruggeman FJ. Systems modeling approaches for microbial community studies: from metagenomics to inference of the community structure. *Front Microbiol.* 2015;6:213.
189. Roder T, Wüthrich D, Bär C, Sattari Z, Ah U von, Ronchi F, et al. In Silico Comparison Shows that the Pan-Genome of a Dairy-Related Bacterial Culture Collection Covers Most Reactions Annotated to Human Microbiomes. *Microorganisms.* 2020;8.
190. Roder T, Oberhänsli S, Shani N, Bruggmann R. OpenGenomeBrowser: a versatile, dataset-independent and scalable web platform for genome data management and comparative genomics. *BMC Genomics.* 2022;23:855.
191. Roder T, Pimentel G, Fuchsmann P, Tena Stern M, von Ah U, Vergères G, et al. *Scoary2*: Rapid association of phenotypic multi-omics data with microbial pan-genomes. *BioRxiv.* 2023.
192. Persson E, Kaduk M, Forslund SK, Sonnhammer ELL. Domainoid: domain-oriented orthology inference. *BMC Bioinformatics.* 2019;20:523.
193. Feldbauer R, Gosch L, Lüftinger L, Hyden P, Flexer A, Rattei T. DeepNOG: Fast and accurate protein orthologous group assignment. *Bioinformatics.* 2020;36:5304–12.
194. Hu X, Friedberg I. SwiftOrtho: A fast, memory-efficient, multiple genome orthology classifier. *Gigascience.* 2019;8.

195. Zhou Z, Charlesworth J, Achtman M. Accurate reconstruction of bacterial pan- and core genomes with PEPPAN. *Genome Res.* 2020;30:1667–79.
196. Steinegger M, Söding J. Clustering huge protein sequence sets in linear time. *Nat Commun.* 2018;9:2542.
197. Emms DM, Kelly S. SHOOT: phylogenetic gene search and ortholog inference. *Genome Biol.* 2022;23:85.
198. Yang H, Zhang X, Liu Y, Liu L, Li J, Du G, et al. Synthetic biology-driven microbial production of folates: Advances and perspectives. *Bioresour Technol.* 2021;324:124624.
199. Flahaut NAL, Wiersma A, van de Bunt B, Martens DE, Schaap PJ, Sijtsma L, et al. Genome-scale metabolic model for *Lactococcus lactis* MG1363 and its application to the analysis of flavor formation. *Appl Microbiol Biotechnol.* 2013;97:8729–39.
200. Robinson JT, Thorvaldsdottir H, Turner D, Mesirov JP. igv.js: an embeddable JavaScript implementation of the Integrative Genomics Viewer (IGV). *Bioinformatics.* 2023;39.
201. Maiden MC, Bygraves JA, Feil E, Morelli G, Russell JE, Urwin R, et al. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc Natl Acad Sci USA.* 1998;95:3140–5.
202. Miquel S, Beaumont M, Martín R, Langella P, Braesco V, Thomas M. A proposed framework for an appropriate evaluation scheme for microorganisms as novel foods with a health claim in Europe. *Microb Cell Fact.* 2015;14:48.
203. Goldacre B. Chapter 6: “The Nonsense du Jour.” In: *Bad Science*. London: Fourth Estate; 2008. p. 288.
204. Ogier JC, Lafarge V, Girard V, Rault A, Maladen V, Gruss A, et al. Molecular fingerprinting of dairy microbial ecosystems by use of temporal temperature and denaturing gradient gel electrophoresis. *Appl Environ Microbiol.* 2004;70:5628–43.
205. Xia J, Wishart DS. MSEA: a web-based tool to identify biologically meaningful patterns in quantitative metabolomic data. *Nucleic Acids Res.* 2010;38 Web Server issue:W71-7.
206. Li S, Park Y, Duraisingham S, Strobel FH, Khan N, Soltow QA, et al. Predicting network activity from high throughput metabolomics. *PLoS Comput Biol.* 2013;9:e1003123.
207. Suez J, Zmora N, Zilberman-Schapira G, Mor U, Dori-Bachash M, Bashiardes S, et al. Post-Antibiotic Gut Mucosal Microbiome Reconstitution Is Impaired by Probiotics and Improved by Autologous FMT. *Cell.* 2018;174:1406-1423.e16.
208. Reid G, Gadir AA, Dhir R. Probiotics: reiterating what they are and what they are not. *Front Microbiol.* 2019;10:424.
209. Siepel A. Challenges in funding and developing genomic software: roots and remedies. *Genome Biol.* 2019;20:147.
210. Gardner PP, Paterson JM, McGimpsey S, Ashari-Ghomi F, Umu SU, Pawlik A, et al. Sustained software development, not number of citations or journal choice, is indicative of accurate bioinformatic software. *Genome Biol.* 2022;23:56.
211. Roder T. GitHub - MrTomRod/flower-plot: A Python function that makes flower plots. GitHub. 2021. <https://github.com/MrTomRod/flower-plot>. Accessed 1 Jan 2022.

212. Roder T. GitHub - MrTomRod/gene-loci-comparison: Create fancy (bokeh) gene locus plots from GenBank files! 2021. <https://github.com/MrTomRod/gene-loci-comparison>. Accessed 6 Feb 2023.

213. Turgay M, Falentin H, Irmeler S, Fröhlich-Wyder M-T, Meola M, Oberhaensli S, et al. Genomic rearrangements in the *aspA-dcuA* locus of *Propionibacterium freudenreichii* are associated with aspartase activity. *Food Microbiol.* 2022;106:104030.

214. Roder T. GitHub - MrTomRod/kegg-map-wizard: Downloads pathway maps from KEGG, creates Python objects, converts them into SVG, allows processing in modern browsers. Computer software. Bern: GitHub; 2021. <https://github.com/MrTomRod/kegg-map-wizard>. Accessed 1 Jan 2022.

215. García-Martín AB, Roder T, Schmitt S, Zeeh F, Bruggmann R, Perreten V. Whole-genome analyses reveal a novel prophage and cgSNPs-derived sublineages of *Brachyspira hyodysenteriae* ST196. *BMC Genomics.* 2022;23:131.

216. Roder T, Pimentel G, Bär C, von Ah U, Bruggmann R, Vergères G. Can eating bacteria in dairy products support your health? *Front Young Minds.* 2022;10.

7. Acknowledgments

First and foremost, I would like to thank Rémy Bruggmann for giving me the opportunity to do my PhD at the Interfaculty Bioinformatics Unit (IBU) despite my lack of significant formal (bio-)informatics training. I am grateful for his guidance, encouragement, and supportiveness, and for defending my interest in doing a purely *in silico* PhD. He shared my enthusiasm and vision for OpenGenomeBrowser and enabled its development by securing additional funding.

I am also grateful to my co-advisors, Guy Vergères, who skillfully led the Polyfermenthealth project with so much energy and passion, Andrew Macpherson, whose expertise and resources were invaluable, and Stephanie Ganal-Vonarburg, for her creative inputs, experienced experimental contributions and knowledgeable interpretations.

Moreover, I thank Rory Johnson and Thomas Rattei for being my mentor and external co-referee, respectively. I would like to thank the Graduate School for Cellular and Biomedical Sciences (GCB) for providing me with a comfortable environment to pursue a PhD at the University of Bern.

Many thanks to the tireless Polyfermenthealth team, namely yoghurt guru Ueli von Ah, metabolomics master Grégory Pimentel and especially Cornelia Bär for her considerable support of my writing, her encouragement and enthusiasm, and for helping me keep my head straight. Special thanks go to Zahra Sattari for her unrewarded efforts.

I also want to thank all the current and past members of IBU for the *töggele*, the banter, the camaraderie, fruitful discussions, and advice. I am grateful to Stephan Peischl, for his stewardship during my machine-learning-related adventures, and particularly to Simone Oberhänsli, for her enthusiasm and support of OpenGenomeBrowser, her compassion and encouragement, and her friendship.

I am grateful to the spontaneous, helpful, and generous bioinformaticians I had the pleasure of meeting. Due to the COVID pandemic and the 2019 NuGO conference taking place in Bern, I did not visit a single conference outside of Switzerland during my PhD. Nevertheless, I had fruitful exchanges with fellow bioinformaticians. Most significantly Ola Brynildsrud, the author of original Scoary, with whom I collaborated on the development of Scoary2. Caitlin Collins, spontaneously and very generously, answered my questions about her mGWAS software. Christophe Dessimoz enabled me to arrange a fruitful meeting with Sina Majidian and Adrian Altenhoff on pressing questions regarding ortholog inference. Despite being virtual, all these meetings with top scientists in their respective fields were arranged casually in less than a week, fun, productive, and motivating.

Warmest thanks go to my parents, Franziska and Bernhard, my little brother Marcel, and all my friends for their tremendous support. Finally, I would like to thank Linda for enduring me during some pretty dark days, for her patience, support and love.

8. Curriculum Vitae

REDACTED

9. List of Publications

Giannini, F., Geiser, L., Paul, L. E., Roder, T., Therrien, B., Süss-Fink, G., & Furrer, J. (2015). Tuning the in vitro cell cytotoxicity of dinuclear arene ruthenium trithiolato complexes: Influence of the arene ligand. *Journal of organometallic chemistry*, 783, 40-45.

Roder, T., Wüthrich, D., Bär, C., Sattari, Z., von Ah, U., Ronchi, F., ... & Vergères, G. (2020). In Silico Comparison Shows that the Pan-Genome of a Dairy-Related Bacterial Culture Collection Covers Most Reactions Annotated to Human Microbiomes. *Microorganisms*, 8(7), 966.

Huber, M., Roder, T., Irmisch, S., Riedel, A., Gablenz, S., Fricke, J., ... & Erb, M. (2021). A beta-glucosidase of an insect herbivore determines both toxicity and deterrence of a dandelion defense metabolite. *Elife*, 10, e68642.

García-Martín, A. B., Roder, T., Schmitt, S., Zeeh, F., Bruggmann, R., & Perreten, V. (2022). Whole-genome analyses reveal a novel prophage and cgSNPs-derived sublineages of *Brachyspira hyodysenteriae* ST196. *BMC genomics*, 23(1), 1-14.

Roder T., Pimentel G., Bär C., von Ah U., Bruggmann R. & Vergères G. (2022). Can Eating Bacteria In Dairy Products Support Your Health? *Front. Young Minds*. 10:721939.

Roder, T., Oberhänsli, S., Shani, N., & Bruggmann, R. (2022). OpenGenomeBrowser: A versatile, dataset-independent and scalable web platform for genome data management and comparative genomics. *BMC genomics*, 23(1), 1-11.

Roder, T., Pimentel G., Fuchsmann P., Tena Stern M., von Ah U., Vergères G., Peischl S., Bruggmann R. & Bär C. (2023). Scoary2: Rapid association of phenotypic multi-omics data with microbial pan-genomes. Available on bioRxiv, submitted to BMC Genome Biology.

Pimentel G., Roder, T., Bär C., Christensen S., Sattari Z., von Ah U., Fernandez Trigo N., Bruggmann R., Macpherson A., Ganal-Vonarburg S. & Vergères G. (2023). Maternal consumption of yoghurt that activates the aryl hydrocarbon receptor increases intestinal group 3 innate lymphoid cells in the offspring. In preparation.

10. Declaration of Originality

Last name, first name: Roder, Thomas

Matriculation number: 11-121-803

I hereby declare that this thesis represents my original work and that I have used no other sources except as noted by citations.

All data, tables, figures and text citations which have been reproduced from any other source, including the internet, have been explicitly acknowledged as such.

I am aware that in case of non-compliance, the Senate is entitled to withdraw the doctorate degree awarded to me on the basis of the present thesis, in accordance with the "Statut der Universität Bern (Universitätsstatut; UniSt)", Art. 69, of 7 June 2011.

Place, date

Bern, 28.4.2023

Signature