RESEARCH

Open Access

The curated *Lactobacillus acidophilus* NCFM genome provides insights into strain specificity and microevolution

Meichen Pan[1†], Sarah O'Flaherty[1†], Ashley Hibberd[2], Svetlana Gerdes[2], Wesley Morovic[2] and Rodolphe Barrangou[1*]

Abstract

Background The advent of next generation sequencing technologies has enabled a surge in the number of whole genome sequences in public databases, and our understanding of the composition and evolution of bacterial genomes. Besides model organisms and pathogens, some attention has been dedicated to industrial bacteria, notably members of the *Lactobacillaceae* family that are commonly studied and formulated as probiotic bacteria. Of particular interest is *Lactobacillus acidophilus* NCFM, an extensively studied strain that has been widely commercialized for decades and is being used for the delivery of vaccines and therapeutics.

Results Here, we revisit the *L. acidophilus* genome, which was sequenced twenty years ago, and determined the core and pan genomes of 114 publicly available *L. acidophilus* strains, spanning commercial isolates, academic strains and clones from the scientific literature. Results indicate a predictable high level of homogeneity within the species, but also reveal surprising mis-assemblies. Furthermore, by investigating twenty one available *L. acidophilus* NCFM-derived variants, we document overall genomic stability, with no observed genomic re-arrangement or inversions.

Conclusion This study provides a comparative analysis of the currently available genomes for *L. acidophilus* and examines microevolution patterns for several strains derived from *L. acidophilus* NCFM, which revealed no to very few SNPs with strains sequenced at different points in time using different sequencing technologies and platforms. This re-affirms its suitability for industrial deployment as a probiotic and its use as an engineering chassis and delivery modality for novel biotherapeutics.

Keywords *Lactobacillus acidophilus* NCFM, Probiotic, Comparative genomics, Pan genome, Strain variation, Genome, Single nucleotide polymorphism

Background

Lactic acid bacteria (LAB) constitute a distinct group of microorganisms that produce lactic acid during fermentation. LAB have been isolated from a wide range of habitats including the human gut, plants and fermented food products [1]. Their ability to rapidly generate a substantial amount of lactic acid has been mostly exploited for the purpose of food preservation for centuries, and likely millennia [2]. Among them, *Lactobacillus* species have been studied most extensively for their important human health relevance and implications, especially in the past

[†] Meichen Pan and Sarah O'Flaherty Co-lead authors.

*Correspondence:
Rodolphe Barrangou
rbarran@ncsu.edu
[1] Department of Food, Bioprocessing, & Nutrition Sciences, North Carolina State University, Raleigh, NC, USA
[2] Health and Biosciences, IFF, Madison, WI, USA

few decades [3, 4]. The genus *Lactobacillus,* once comprised of over 250 species that encompassed a relatively high level of diversity in both phenotypes and genotypes, has recently been re-classified into 25 different genera to appropriately reflect the phylogenetic relationship of these diverse microorganisms [1]. The amended *Lactobacillus* genus includes species specifically adapted to vertebrates, notably *Lactobacillus acidophilus* [1].

L. acidophilus is an industrially commercialized species and several strains have been categorized as probiotics, which are defined as "*live microorganisms which, when administered in adequate amounts, confer a health benefit on the host*" [5]. Numerous in vivo and in vitro studies have documented a range of beneficial effects on humans, such as immunomodulation [4, 6, 7], digestive health [8] and protection against the colonization of pathogens such as *Helicobacter pylori* [9]. *L. acidophilus* also produces lactacin B, a bacteriocin contributing to its probiotic activity [10, 11]. Given these reported and purported benefits, *L. acidophilus* strains are often consumed as dietary supplements and widely formulated in dairy products such as milk, yogurt and infant formula. *L. acidophilus* NCFM was the first commercially available probiotic strain and has been industrialized since 1972 [12]. Its complete genome was first sequenced and reported in 2005 [13] and has been characterized in-depth to substantiate its Generally Considered As Safe (GRAS) status by the Food and Drug Administration (FDA, GRAS Notice No 357 and 865). In addition, genome editing tools have been developed [14] and improved upon [15, 16] facilitating the construction of vaccine [17–19] and therapeutic strains of *L. acidophilus* NCFM [20, 21].

Although it was historically challenging to determine and distinguish different species and strains of *Lactobacillus* by sequencing and using classical microbiology methods, technological advances in molecular biology and DNA sequencing over the past few decades have substantially improved and democratized bacterial phylogenetic determination and typing, with notable gains in the accuracy of probiotic strain identification and labeling [22, 23]. Traditional classification methods relying on polyphasic taxonomy, encompassing phenotypic characterizations (such as carbohydrate fermentation profile) and genetic data (such as 16S rRNA and DNA fingerprinting methods) [3] have been supplanted by draft and whole genome sequencing (WGS) [24]. Such unambiguous technologies enable unequivocal phylogenetic determination including comparative genomic analyses of *L. acidophilus* [25] and provide a basis to address nomenclature and naming conventions sometimes exacerbated by commercial trademarks and labeling practices, which are legally and practically troublesome for some companies and consumers, and occasionally prove challenging even for scientists. Within the scientific community, strains are sometimes re-isolated or re-named by various labs (for instance, *L. acidophilus* NCFM has been previously associated with other names such NCK56, NCK45 and RL8K), and are (re)isolated from commercial products with subsequent re-naming and re-sequencing,

Over two decades ago, the first draft genome of NCFM was sequenced using an ABI377 sequencer [13]. Since then, next-generation sequencing technologies have rapidly evolved, rendering whole genome sequencing more affordable and scalable. Indeed, the original ~ $1.6 M price tag has been impressively reduced by 4 orders of magnitude, concurrent with an increase in sequencing throughput. These technological advances provide an opportunity to determine genome sequences with high accuracy, and investigate how precision varies across technologies and time, including Illumina, Nanopore, and MinION platforms, with nearly 500,000 genomes spanning over 60 bacterial phyla now available at NCBI [26, 27].

L. acidophilus NCFM and many of its derivatives have been sequenced over time using different sequencing technologies and contextual settings (academic vs. industrial). This has allowed for single-nucleotide polymorphism (SNP) analysis in strains selected for thermal adaptability [28], response to a simulated vaginal environment [29] and passage through the murine gastrointestinal tract [30]. In this study, we analyze and compare the *L. acidophilus* NCFM genome sequences from various vantage points, encompassing space (commercial vs. academic sources), time (4 decades), sequencing platforms, and isolates. We also examine the publicly available *L. acidophilus* genomes to infer the *L. acidophilus* pan genome and investigate assembly consistency and accuracy. Results establish *L. acidophilus* genomic stability and microevolution, and provide insights into derived regulatory and commercial considerations for the probiotic market.

Results

Core and pan genomes of *L. acidophilus*

The pan genome of 114 *L. acidophilus* genomes publicly available at NCBI was determined. (Table S1). Overall, we observed some variations in genome size (1.98 ± 0.70 Mb, mean ± SD) and GC content (34.6 ± 0.12%, mean ± SD). Most of the genomes are scaffolds or contigs (91 out of 114) with 23 genomes marked as complete. This subset of 23 complete genomes had an average genome size of 2.00 ± 0.27 Mb, mean ± SD and GC content of 34.69 ± 0.07%, mean ± SD. The number of genes predicted in each genome varied by 5.9% (1,992 ± 117).

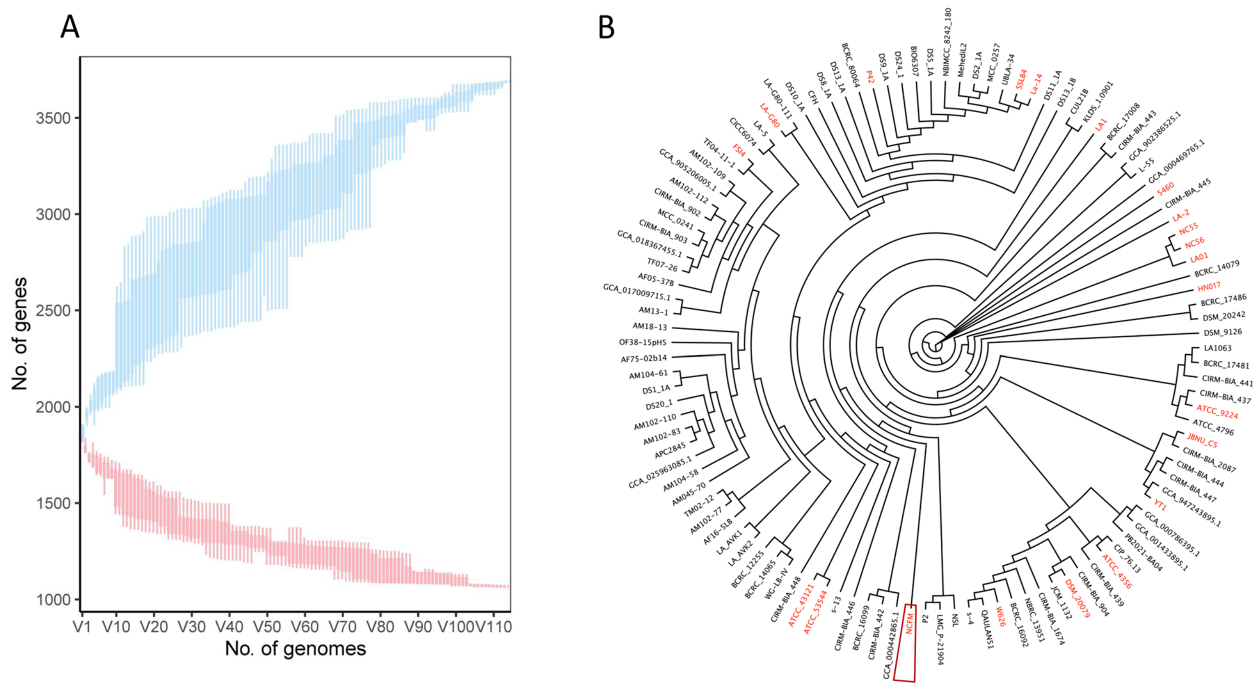

Fig. 1 The core (red) and pan (blue) genome of the 114 *L. acidophilus* strains were depicted using Roary (**A**). The Maximum likelihood phylogenomic tree based on 1062 core genes of 114 *L. acidophilus* genomes (**B**). The complete genomes were annotated in red (*n* = 23)

A total of 3,692 genes were identified across the pan genome of the 114 *L. acidophilus* sequences (Fig. 1A and Fig. 2). We also identified a total of 1,062 core genes that were shared by all the genomes and 1,714 cloud genes that were shared by less than 15% of the genomes. The pan genome of *L. acidophilus* can be considered as open, based on the gamma parameter larger than 0.17 [31] (Fig. 1A and Supplemental Fig. 1). The pan genome of *L. acidophilus* is smaller compared to other *Lactobacillus* species such as *L. crispatus* (12,114 open pan genome) [32] and *L. paragasseri* (6,535, open pan genome) [33]. This is presumably due to the relatively smaller genome size of *L. acidophilus,* but also reflects the low level of within species heterogeneity in *L. acidophilus* genomes as compared to other species. When considering all 114 *L. acidophilus* genomes, only 28% of the total genes were shared by every genome. However, if only complete genomes were taken into account, over 55% of the total genes were found to be shared. A phylogenomic tree based on the 1,062 core genes of the 114 *L. acidophilus* genomes was performed (Fig. 1B), demonstrating the clustering of the *L. acidophilus* genomes.

L. acidophilus strains share highly similar genomes (ANI)

We then focused on the complete genomes of *L. acidophilus* to determine how genetically similar they are to one another. Pairwise comparison at the 95% threshold was performed to calculate the ANI values of 16 closed *L. acidophilus* genomes (Fig. 3A). The ANI value was over 99% for all 16 genomes, reflecting high sequence conservation within the species. In fact, all but two genomes shared over 99.8% ANI values. The ANI value for YT1 was around 99.2% and for JBNU_C5 was around 99.6% across the board. The accepted level of ANI between genomes to be considered as the same species is 95%. These genomes share a high level of similarities that are not usually observed in other *Lactobacillus* species, for example *L. crispatus* and *L. gasseri* [32, 33].

We next performed a mauve alignment of 16 closed genomes which revealed that some genome regions showed inversion or likely miss-assembly compared to the NCFM genome (Fig. 3). Particularly, the two locally collinear blocks (LCBs) located between 45 K bp and 165 K bp are often inversely assembled, likely due to long stretches of repetitive regions located in that region. The mauve alignment also revealed some low coverage regions in genomes YT1 and JBNU_C5, which was consistent with the lower ANI values. We selected eight complete genomes that had an ANI value lower than 99.5% compared to the NCFM genome to perform genome-wide BLAST alignment (Fig. 4B). Overall, when the predicted coding sequences were compared to the NCFM genome, we observed a high level of identity scores. Similar to the mauve alignment, we observed

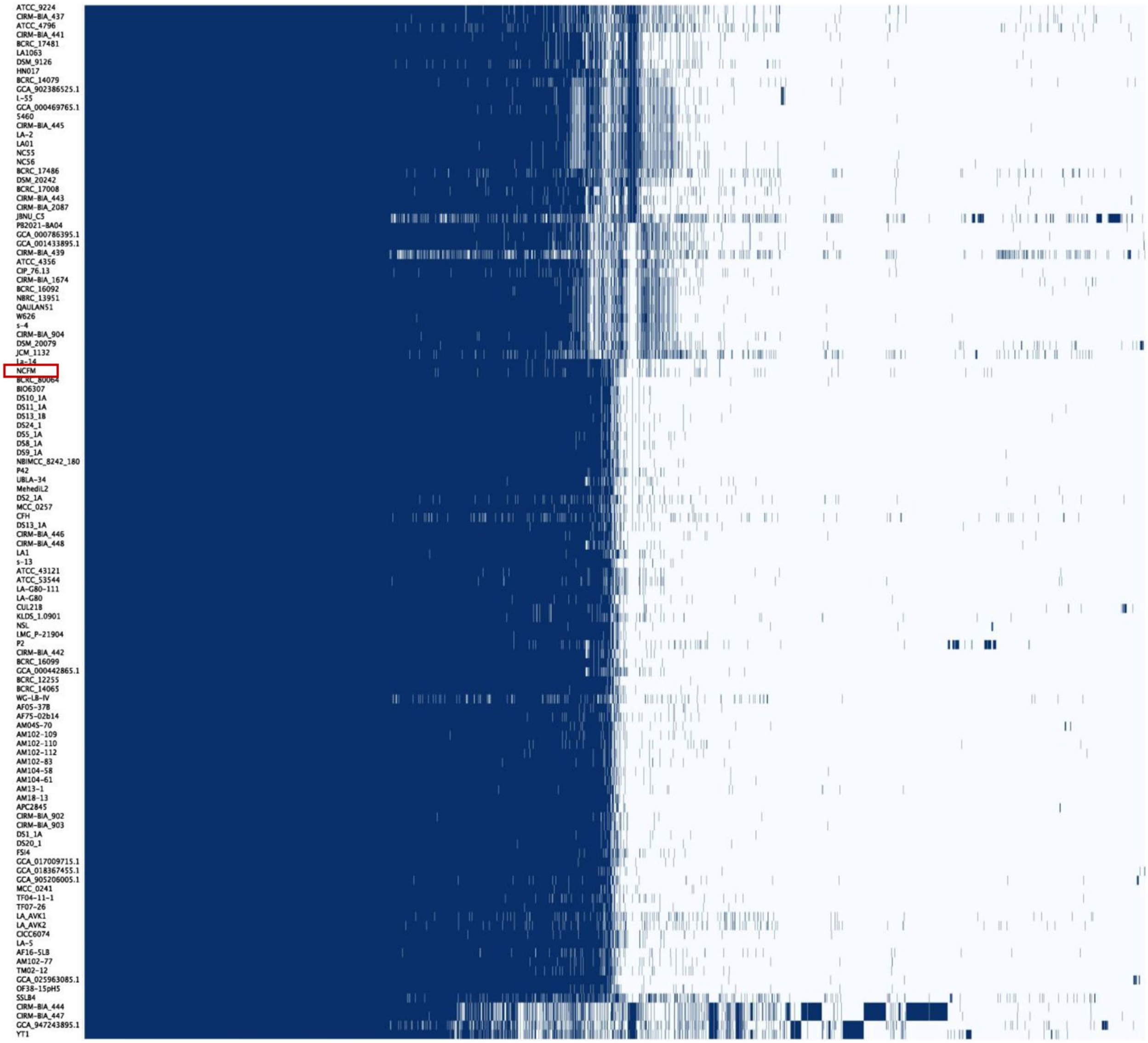

Fig. 2 Heatmap depicting the presence and absence of genes from the *L. acidophilus* genomes. (Dark blue = presence)

more gaps in genomes YT1 and JBNU_C5 (Fig. 4B). Upon closer examination, two additional gaps, interestingly not located in YT1 and JBNU_C5, were identified through a noticeable change in GC content and a low identity score. The first gap was around 2 kb where NCFM encoded *pbpX*1 and a *def*2 gene that are absent in the genomes of HN017, NC55, NC56, LA-2 and 5460. The second gap was around 10 kb. This island encoded genes such as *bglH*, *bglA*, *gmuC*, *gmuR*, *licA*, *celA* and multiple hypothetical proteins. This island was not found in NC55, NC56, LA-2, and 5460 genomes. These four genomes were also missing the first 2 kb gap.

Prophage prediction in *L. acidophilus* genomes

We predicted the presence of prophages in the 114 *L. acidophilus* genomes and determined only four strains (3.5% of genomes) encoded for intact prophages. Four additional strains were determined to encode for incomplete or partial prophage genomes (Table S1). Two of the strains *L. acidophilus* CIRM-BIA 444 and CIRM-BIA 447 were each predicted to have two intact, one questionable and one incomplete prophage. A previous study with 35 *L. acidophilus* strains determined that the occurrence rate for prophages was similar at 2.9% [34].

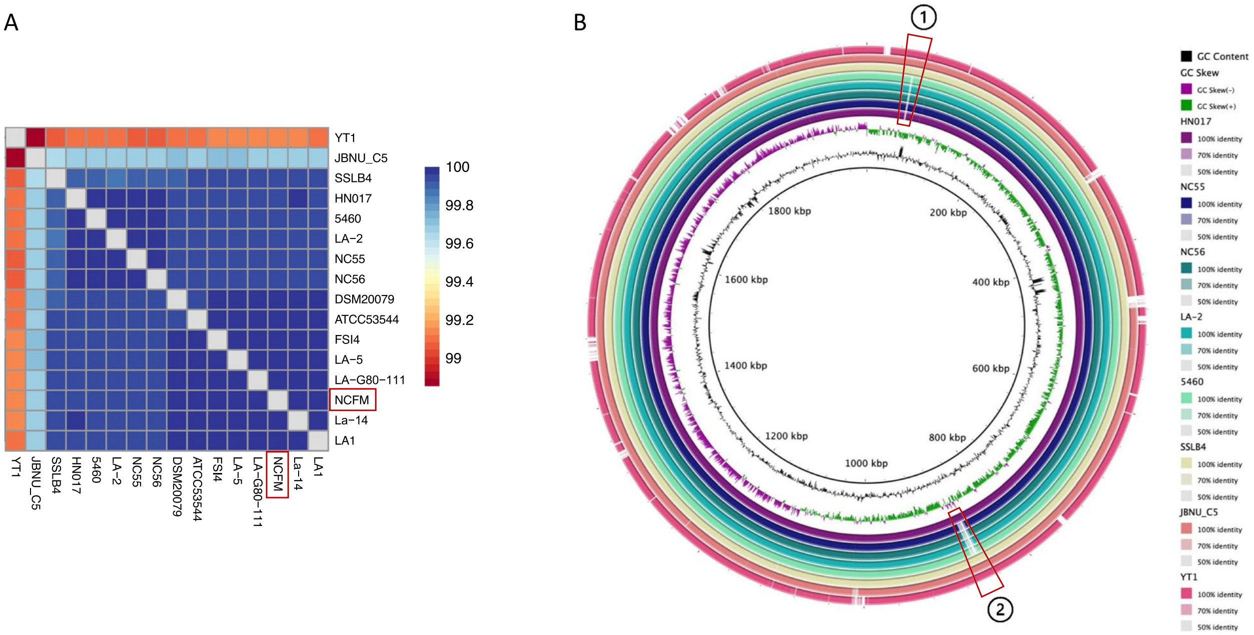

Fig. 3 The ANI heatmap of 16 closed *L. acidophilus* genomes (**A**). BRIG genome alignment of 8 complete genomes that had an ANI value lower than 99.5% compared to the NCFM genome (**B**). Regions that showed genomic variation are labeled as 1 and 2

Next-generation sequencing technologies reveal sequencing errors

Next, we focused our analysis on whole genome comparisons of the *L. acidophilus* NCFM genome. Our initial analysis determined the amount of SNPs between our currently available NCFM$_A$ and the originally submitted genome of *L. acidophilus* NCFM in NCBI [[13] and NC_006814]. This allowed us to compare the output from sequencing technology of the late 1990s with our current NCFM$_A$ genome determined with both long and short read next-generation technologies. Running the SNP analysis on the 2005 (NC_006814) compared to the NCFM$_A$ genome sequence revealed over 100 SNPs including deletions and insertions (Fig. 5A). We anticipated that the majority of these SNPs might be sequencing errors. Therefore, we (re)sequenced the DNA from the same glycerol stock of NCFM (NCK1070) that had been stored as a glycerol stock since the original genome sequencing in the late 1990s. Mapping the raw reads from NCK1070 to NCFM$_A$ and running a SNP analysis remarkably resulted in only one potential validated SNP. However, with a 74% variant frequency we determined this was in fact a mixed genotype (MG1) with an extra T (before position 525,270) co-existing within the bacterial population. This insertion would introduce a frame shift in this gene, which encodes for an inner membrane protein. We next examined the raw reads for NCFM$_A$ and determined that NCFM$_A$ also had MG1 (TT vs TTT) at this coordinate, albeit at a lower minority frequency, since the sequence reports the dominant consensus genotype. Remarkably, under the parameters tested, there was actually no difference in the genomes of NCK1070 and NCFM$_A$ and differences noted in the initial comparison were due to errors in sequencing technology of NCFM (NC_006814) at that time.

Comparison of *L. acidophilus* NCFM with engineering variant genomes

Given our history of working with *L. acidophilus* and numerous studies performed with deletion, knock outs, genome editing and alteration of genes from the genome to determine probiotic function [35–39] and/or insertions for vaccine development [18, 40–42] we also had access to 21 whole genome sequences of engineered *L. acidophilus* NCFM$_A$ variants spanning multiple decades (Table 1 and Table S3). Twenty of the 21 strains also contained the aforementioned mixed genotype MG1. *L. acidophilus* NCK2636, had undergone 45 iterative transfers in simulated vaginal fluid as part of a previous study [29], which did not retain the MG1, but the TT, which would not induce a frame shift. We also confirmed this genotype in raw reads from RNA sequencing data from the same study. In fact, we determined the MG1 genotype was present at generation 0 but lost after 50, 100, 500 and 1,000 generations in simulated vaginal fluid, presumably through a selective bottleneck during iterative passages in these experimental conditions.

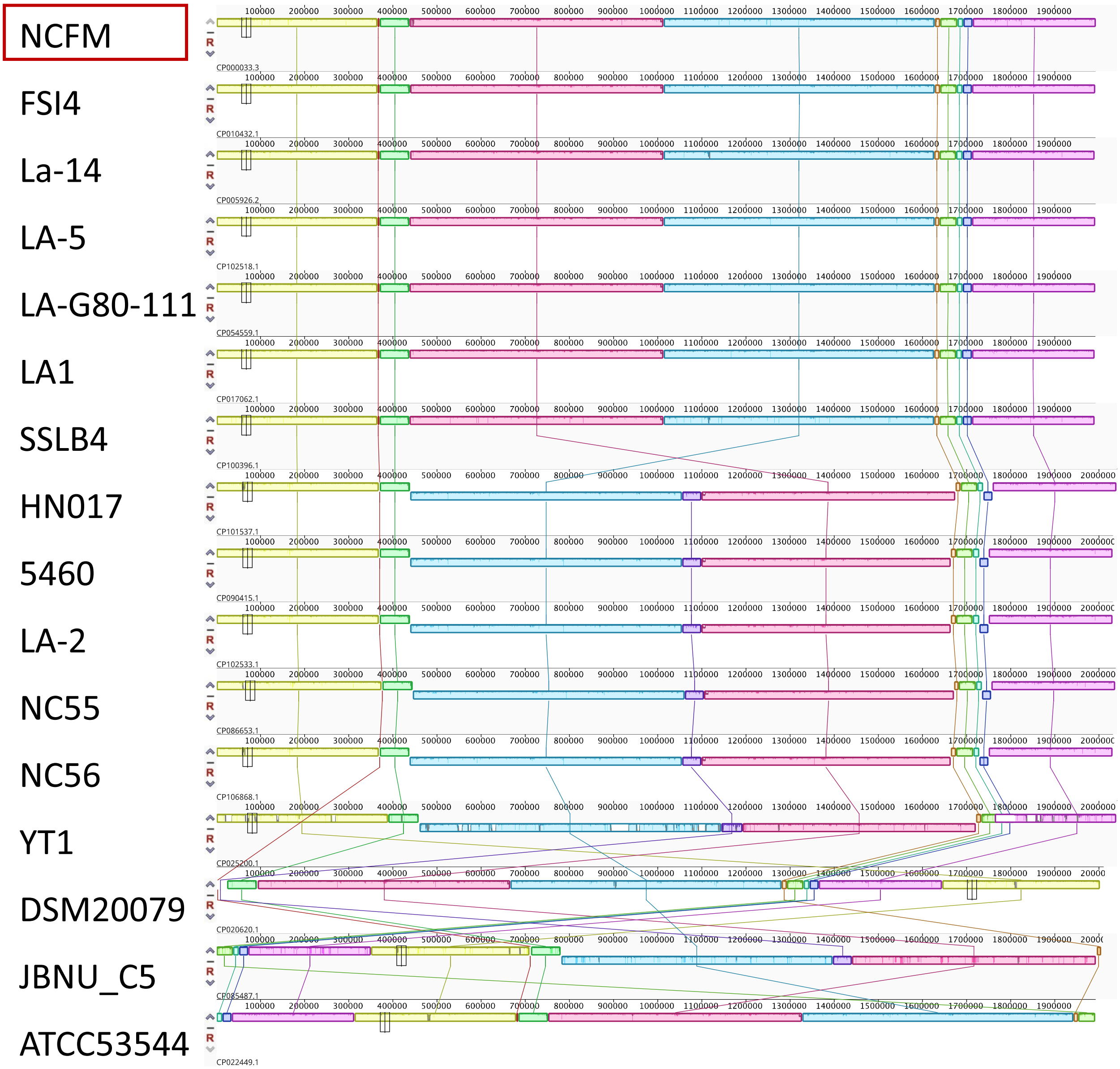

Fig. 4 Mauve alignment of 16 closed *L. acidophilus* genomes with *L. acidophilus* NCFM as reference genome

A second mixed genotype was located in the six strains that were isolated from murine fecal samples. This mixed genotype, a T or C (MG2), resulting in a change in amino acid from a Y to an H in a hydrolase protein (coordinate 119,340) was also located in $NCFM_A$ and NCK1070 but not the other strains, in which the T was constant with the exception of NCK2636. After 45 transfers in simulated vaginal fluid, the C genotype was also fixed in NCK2636. The RNA sequencing raw reads also determined the MG2 co-existing at 0 generations but the fixed C genotype at 50, 100, 500 and 1,000 generations, likely selected for during the aforementioned passages.

We next determined whether there were any SNPs between these 21 strains and $NCFM_A$. We located 22 unique SNPs (Fig. 5B, Tables S3 and S4). Noteworthy, only one of the 22 SNPs were found in more than one strain (Table S3 and S4). Included in the 21 WGS sequences is the strain NCK1909 that was constructed with a deletion in the *upp* gene to serve as a counter selectable marker for subsequent constructs for gene deletions and insertions [15]. Ten additional strains are NCK1909 derivatives (Table 1). All 11 strains (NCK1909 and subsequent derivatives) were determined to contain between one and five SNPs (Table S3). NCK1909

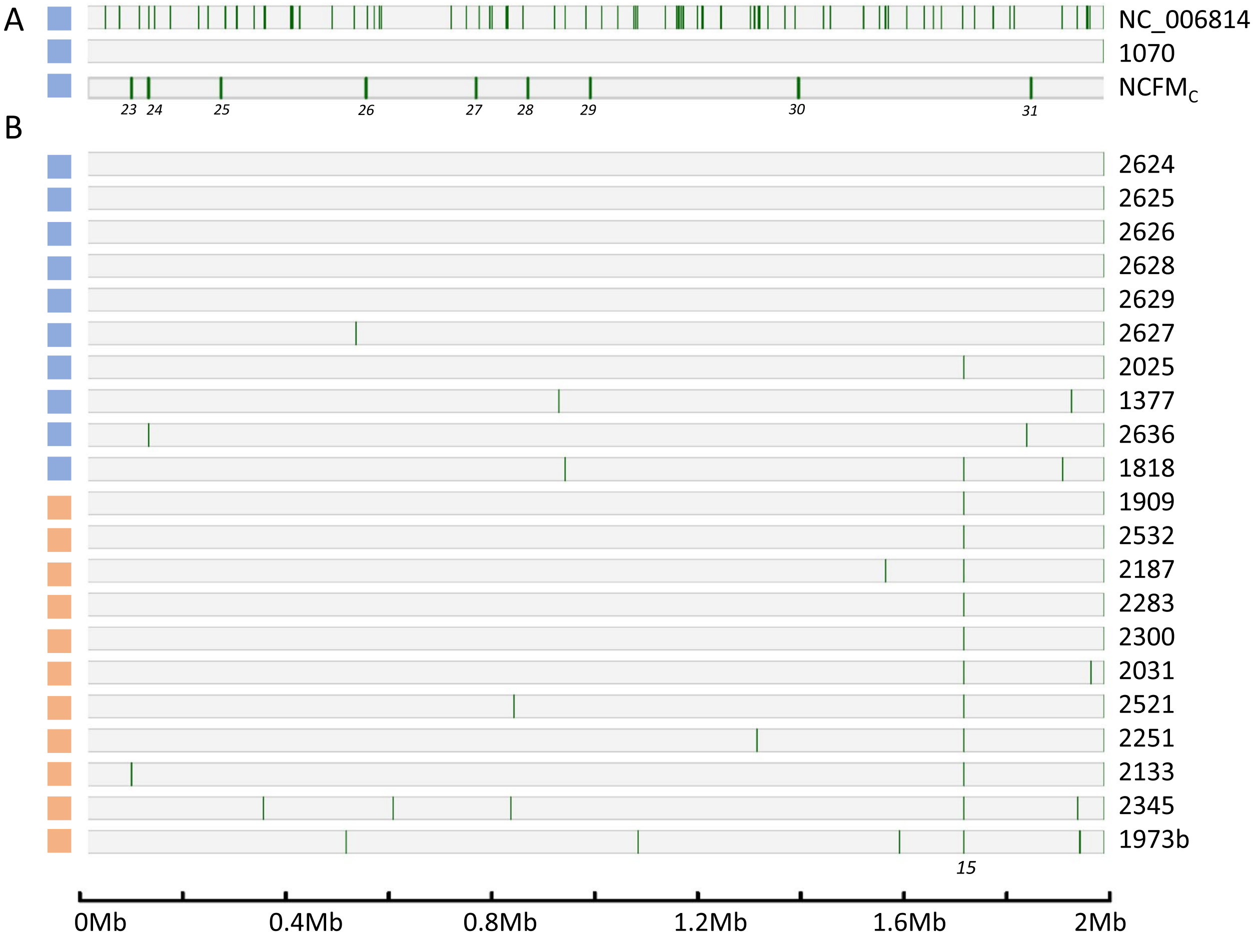

Fig. 5 Schematic of detected SNPs between NCFM$_A$ genome compared to the NCBI publicly available genome (NC_006814), genome of NCK1070 and genome of NCFM$_C$ (**A**). SNPs identified between NCFM$_A$ genome and 21 WGS of *L. acidophilus* NCFM variants (**B**). Green lines; SNPs. Strains without a deletion in the *upp* gene are marked by blue boxes and strains with the *upp* gene deletion are marked by the orange box. Numbers under SNPs in NCFM$_C$ correlate to Table S4

contained one SNP; SNP15 (T to C) in a non-coding region. Our data showed that SNP15 was also located in all 10 derivatives of NCK1909 (Fig. 5B and Table S3), indicating this SNP remained stable through subsequent genetic manipulations. This data demonstrated the overall genetic stability of *L. acidophilus* NCFM given the paucity of SNPs (on average, one SNP per genome, amongst a nearly 2Mbp sequence). Key genes historically related to probiotic efficacy that were previously studied and established are stable, notably for example carbohydrate utilization [36, 50], cell surface composition [37–39, 45], and genes related to stability in the intestinal tract such as bile salt hydrolase genes [49, 51]. Evidently, besides the stability of the actual genomic sequence, it is important to also note, based on the mauve alignment, the stability of the overall chromosomal architecture, with no observed re-arrangement or inversions.

Comparison to the commercial *L. acidophilus* NCFM genome

Given the importance of *L. acidophilus* NCFM as a commercial probiotic strain [12] we also performed a SNP analysis on the genome sequence of the *L. acidophilus* NCFM strain that is commercialized by IFF (https://www.howaru.com/hcp/strains/ncfm) against our NCFM$_A$ genome. Noteworthy, nine SNPs were identified between NCFM$_A$ and the IFF genome, NCFM$_C$ (Table S5), somewhat establishing a benchmark for genomic equivalency between isolates. The two genomes are strictly collinear and differ in length by a single nucleotide (1,991,977 bp versus 1,991,998 bp). No mixed genotype loci were detected in NCFM$_C$ including the absence of MG1 (and hence no frameshift in the corresponding gene) and MG2. Notably, the sequences at both of these positions in NCFM$_C$ correspond to the dominant and subsequently fixed genotype that the NCFM$_A$ strain

Table 1 Details of bacterial strains

Strain NCK Number	Strain Details	*upp* deletion	No of SNPs	Reference
2624	NCFM colonized in mouse GIT; mouse#1 fecal sample	N	0	[30]
2625	NCFM colonized in mouse GIT; mouse#2 fecal sample	N	0	[30]
2626	NCFM colonized in mouse GIT; mouse#3 fecal sample	N	0	[30]
2628	NCFM colonized in mouse GIT; mouse#5 fecal sample	N	0	[30]
2629	NCFM colonized in mouse GIT; mouse#6 fecal sample	N	0	[30]
1070	NCFM glycerol stock used for initial WGS	N	0	[13]
2627	NCFM colonized in mouse GIT; Mouse#4 fecal sample	N	1	[30]
2025	deletion in phosphoglycerol transferase – LTA, *lba0447*	N	1	[21]
1377	Surface layer Protein A gene inactivation with plasmid	N	2	[43]
2636	45X transfer in simulated vaginal fluid	N	2	[29]
1818	Deletion in *luxS*	N	3	[44]
1909	Deletion in *upp*, backround strain for *upp* deletions	Y	1	[15]
2532	Deletion in surface layer protein LBA0695	Y	1	[45]
2187	*lta* (ltaS1), *slpX*, *slpB* deletion	Y	1	[46]
2283	Streptomycin resistant NCK1909	Y	1	unpublished
2300	Rifampicin resistant NCK1909	Y	1	[47]
2031	*slpB, slpX, slpA* deletion/inactivation	Y	2	[48]
2521	deletion in *bshA*	Y	2	[49]
2251	Hypothetical membrane protein (*lba1740*) deletion	Y	2	unpublished
2133	*fbpA, muc* (3 mucin gene) deletions	Y	4	unpublished
2345	vaccine strain with PA-DCPpep and BoNT/A-Hc-Dcpep	Y	5	[42]
1973b	*upp, slpB, slpA* gene deletions/inactivation	Y	5	[15]

N No and *Y* Yes See Table S3 for SNP details

evolved to after 45 transfers in vaginal fluid described above. Four SNPs are located in non-coding regions not expected to have phenotypic effects, and five resulted in amino acid substitutions rather than frameshifts. The NCFM commercial strain has remained genetically stable throughout decades of manufacturing for commercial probiotic usage; accordingly, the current production genome NCFM$_C$ remains genetically identical to the oldest seed vials archived in internal and external culture collections. Remarkably, even the degenerate and nonfunctional CRISPR locus, which is devoid of *cas* genes and contains repeats that have hypervariable sequences, is conserved and devoid of SNPs. This data further confirmed the overall integrity, stability and lack of genetic drift in the *L. acidophilus* genome, across space and time, including maintenance of both the genomic sequence and the overall chromosomal architecture.

Discussion

The advent of next-generation sequencing (NGS) technologies, together with scalable bioinformatic tools has fueled a *bona fide* revolution in genetics and the inherent deposit of vast genomic and other omic data at NCBI. Yet, despite the intriguingly large volume of data available, there may be issues regarding the quality of this quantitatively massive dataset. This challenge is confounded by differences in sequencing technologies (throughput and quality), assembly pipelines (consistency and standards) and users (expertise, sophistication, analytical insights, and extent of curation).

Despite the overall stability of the *L. acidophilus* genomes and their conservation across strains, it is noteworthy that this study revealed perplexing species misattribution and mis-assemblies. It has been nearly two decades since the original publication of the first complete genome sequence of *L. acidophilus* NCFM [13]. Since then, high-quality assemblies have been generated by combining the accuracy of short-read sequencing technologies, typically Illumina, with the gap-filling and assembly capabilities of long-read sequencing technologies such as PacBio and Nanopore. In fact, the first *L. acidophilus* NCFM genome sequencing project showcases this progress firsthand, considering the financial resources, human capital and time required for completion of a curated closed genome. Our analysis showed no difference in the genomes of the originally sequenced NCK1070 and NCFM$_A$ as differences in the initial comparison are due to errors in sequencing technology of NCFM (NC_006814) at that time. Therefore, while this study illustrates how improvements in sequencing technology have enabled accurate WGS, it reinforces the need to re-sequence and update the public database with curated genomes of biologically and industrially relevant strains. This is important as technology advances; for example, precision genome editing using CRISPR-Cas systems typically hinges on a PAM sequence that can be just two nucleotides in length [16].

Genomic data is essential for other purposes such as strain tracking, typing, genetic drift and genetic population diversity [52, 53]. Genomic data also allows for the identification of the virome, a more recent research focus in relation to human health [54]. We confirmed a previous report demonstrating the lack of prophages in *L. acidophilus* genomes. It has been suggested that strain isolation source may play a factor in prophage scarcity *L. acidophilus* genomes [34] but the reason still needs to be determined. Our data is also similar to that reported previously [25] where comparative genomic analysis of 46 strains of *L. acidophilus* also determined high ANI values > 97%, similar GC content (34.66%) and genome size of ~ 2 Mb and an open pan genome [25].

Commercial implications are also important to monitor strain genetic integrity, track strain dissemination and (re)isolation, and document genetic content and stability in regulatory dossiers. *L. acidophilus* NCFM has been commercially produced as a probiotic bacterium since the 1970s, and continues to be one of the most important and widely consumed probiotic strains. Our lab has almost half a century of experience working with *L. acidophilus* NCFM and has generated vast amounts of genomic, transcriptomic, proteomic and phenotypic data on this strain. We previously investigated and demonstrated the genome integrity of *L. acidophilus* NCFM post colonization through the mouse GIT [30] and after transfers in simulated vaginal fluid [29]. We identified zero SNPs and 100% genome-wide sequence conservation when we compared our NCFM$_A$ genome with the re-sequenced genome (NCK1070), and very few inconsequential SNPs in *L. acidophilus* NCFM derivatives and the genome from a commercially available NCFM strain. The identified mixed genotypes were also present in the NCFM$_A$ genome. However, we observed that certain mixed genotypes will undergo genetic drift during passage in simulated vaginal fluid, indicating the importance of updated WGS of strains genomes during extensive passages, industrial fermentation and the manufacturing process. We also highlighted that publicly available assemblies alone can be misleading, and occasionally incorrect. Although base calling has been mostly accurate, when it comes to determining the accuracy of SNPs and the identification of MGs, it is essential to have raw reads accessible for full confidence.

As discussed above and previously reported, *L. acidophilus* is a homogenous species, with the publicly available closed genomes sharing over 99% ANI scores. This brings up the practical question of the definition of a bacterial strain, and the practical challenges to defining the genetic sequence of a specific commercial strain. It has been documented that even a single or small number of SNPs can lead to dramatic phenotypic variation such as carbohydrate metabolism [55] and genome engineering efficiencies [56]. Yet, the commercial context hinges on specific strains, with customized formulations and tailored supplements touting the benefits of very specific probiotic strains related to particular published literature and documented functional attributes. This is further exacerbated by some rampant issues in the dietary supplement industry, in which products are frequently mis-labeled, encompassing species and strain mis-identification, as well as viable cell count variability and shortcomings [22]. Best practices for maintaining genetic integrity within an industrial fermentation setting include the designation and archiving of master seed vials in a culture collection that are verified to be free from contamination and variants. The purity of the master seed vial can be confirmed by traditional streak-plating and WGS sequencing of several individual colonies. A complete genome that has been manually curated for accuracy can then be generated and subsequently serve as a gold standard reference to confirm the genetic integrity of materials to be used in clinical trials, to monitor for genetic drift in fermentation runs and to validate new seed vials. As these species and strains are increasingly used in therapeutic formulations and for vaccine delivery, while their history of use and human consumptions establishes safety, their efficacy will continue to be determined in human clinical trials.

Conclusions

This work emphasizes the importance of monitoring genetic drift and SNP occurrence throughout the industrial manufacturing and clinical trial process, while also stressing the importance to select probiotic strains with stable genomes. Overall, this study affirms *L. acidophilus* NCFM as a valuable, stable and reliable model organism for lactic acid bacteria, as a promising live biotherapeutic and as a useful chassis for the genetic engineering of next-generation therapeutic bacteria.

Methods

Lactobacillus acidophilus genomes

L. acidophilus genomes (drafts and complete genome) were retrieved from the NCBI database in March 2024 ($n = 114$, Supplementary Table S1). The 114 genomes were annotated using Prokka v1.14.6 to predict open reading frames (ORF) for subsequent pan genome analysis [57]. CheckM analysis to check for genome completeness and contamination was ran on the 114 genomes [58] (Table S2). For reference, we used our in house NCFM_Academic (NCFM$_A$) genome (see below). Additionally, the curated industrial NCFM genome, NCFM_HOWARU_Commerical (NCFM$_C$) provided by IFF sequenced using Illumina and Nanopore technology was used as a comparative reference. The genome of NCK1070 was sequenced as described below by Corebiome (www.diversigen.com).

Prophages were predicted in the 114 genomes using the PHASTEST (Phage search tool with enhanced sequences translation) tool [59]. Prophages were determined according to the PHASTEST scoring system and determined to be intact when the score was > 90, questionable with a score between 70 to 90 and incomplete with a score < 70. We determined 2 prophages to be partial with a score > 90 but each only encoded for 7 proteins (Table S1).

Lactobacillus acidophilus NCFM strains sequencing and SNP analysis

The *L. acidophilus* NCFM$_A$ genome was sequenced with methods detailed previously [16]. Briefly, DNA extraction, library preparation, whole-genome sequencing with Illumina and Nanopore technologies, and assembly were performed at the Roy J. Carver Biotechnology Center (University of Illinois at Urbana-Champaign, Urbana, IL). Genomic DNA was extracted with the MasterPure DNA purification kit (Lucigen). For the Illumina platform libraries were prepared with the MasterPure DNA purification kit (Lucigen) and shotgun genomic libraries prepared with the KAPA HyperPrep library construction kit (Roche). Libraries were sequenced (one lane, 251 cycles) on a HiSeq 2500 system using the HiSeq Rapid SBS sequencing kit v2. For long read sequencing, the genomic DNAs were converted into Nanopore libraries with the NBD114 and 1D (SQK-LSK109) library kits which were sequenced in two SpotON R9.4.1 FLO-MIN106 flow cells for 72 h, using a GridION X5 sequencer. Additionally, *L. acidophilus* genomes from our in-house culture collection (Table 1) were sequenced using the StrainView platform by Corebiome (www.diversigen.com) as follows: DNA was extracted from cell pellets with MO Bio PowerFecal (Qiagen) automated for high throughput on QiaCube (Qiagen), with bead beating in 0.1 mm glass bead plates. DNA Quantification Samples were quantified with Qiant-iT Picogreen dsDNA Assay (Invitrogen). Libraries were prepared with the Nextera Library Prep kit (Illumina) and sequenced on an Illumina NextSeq using paired-end 2×150 reads with a NextSeq 500/550 High Output v2 kit (Illumina). DNA sequences were filtered for low quality and length, and adapter sequences were trimmed using cutadapt (v.1.15).

The NCFM$_A$ genome was used to map raw reads for each strain to obtain each genome using the Geneious mapper in Geneious Prime [60]. For the SNP analysis, all samples were compared to the NCFM$_A$ reference genome. SNPs were determined using the find variations/SNPs tool in Geneious, with settings at a minimum coverage of 20 and minimum variant frequency of 0.5. SNPs were then manually curated to ensure accuracy. The RNA sequencing datasets from simulated vaginal tract samples were determined previously [29] are available at NCBI under BioProject PRJNA600659.

Average nucleotide identity analysis

FastANI v1.33 [61] was used to calculate the paired average nucleotide identity (ANI) values matrix among 16 selected closed *L. acidophilus* genomes. The ANI matrix values were depicted using the "pheatmap" package v1.0.12 in RStudio v1.1.463.

Whole genome alignment

All vs. NCFM genome alignments were visualized using the BLAST Ring Image Generator (BRIG) [62], depicting one ring per genome as well as the GC content and GC skew rings. BLASTn was implemented using the following parameters: 90% as the upper identity threshold and 70% as the lower identity threshold. Genomic regions that showed low identity scores repetitively were manually inspected. Mauve alignment was performed in Geneious Prime software V. 11.0.12, using the progressive Mauve algorithm with a minimum LCB score of 100,000 [63].

Pan genome and phylogenomic analysis

The core and pan genome of the 114 *L. acidophilus* strains were determined using Roary v3.12.0, with the flags -env and the standard threshold of 95% BLASTp identity [64]. The core genome was aligned using the PRANK algorithm to generate the multi-FASTA alignment file. The alignment file was then imported into ClustalW v2.1 [65] to generate the phylogenomic tree using the RAxML (Randomized Axelerated Maximum Likelihood) method [66] The core genome phylogenomic tree was depicted with FigTree v1.4.4 (http://tree.bio.ed.ac.uk/software/figtree/). Methods described by Tettelin et al. [31] were used to determine if the pan genome was open or closed. The total number of genes found is plotted against an increasing number of *L. acidophilus* genomes. The exponent gamma > 0 indicates an open pan-genome species. The unique vs. new genes graph and conserved vs total genes graph were generated using the create_pan_genome_plots. R script included in the Roary package. The gene presence/absence heatmap was generated using the roary_plots.py script included in the Roary package.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-024-11177-2.

Supplementary Material 1.

Supplementary Material 2.

Acknowledgements

We acknowledge the history and legacy of Todd R. Klaenhammer and his lab and the contributions of its many members over the past five decades. We highlight the contributions of Yong Jun Goh, Claudio Hidalgo-Cantabrana, Rosemary Sanozky-Dawes and Matthew Poore for their contribution with sample preparation for WGS and analysis. We thank Kendall Malmstrom and Ruben Sanchez-Nieves for assistance with SNP analysis.

Authors' contributions

MP, SOF and RB conceptualized the study. MP, SOF, WM and SG generated and analyzed the data. MP and SOF wrote the paper. RB and AAH edited the paper. All authors approve the final version.

Funding

The authors would like to acknowledge support from NC State University, the NC Ag Foundation and IFF Inc.

Data availability

The genomes of *L. acidophilus* NCFM$_A$ and NCFM$_C$ have been submitted to the NCBI database, accession numbers CP156988 and CP156983, respectively.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

R.B. is a cofounder of Intellia Therapeutics, Locus Biosciences, TreeCo, CRISPR Biotechnologies and Ancilia Biosciences, and a shareholder of Caribou Biosciences, CRISPRQC, KromaTiD, Raleigh Biosciences, Inari Ag, Felix Biotechnologies, Invaio, Tune Therapeutics, Hoofprint Biome, and Provaxus. M.P., S.O.F. and R.B. are co-inventors on patent applications related to probiotic uses. SG and AH are employees of IFF Inc. WM was an employee of IFF Inc at the time of data generation.

Received: 17 June 2024 Accepted: 20 December 2024
Published online: 03 January 2025

References

1. Zheng J, Wittouck S, Salvetti E, Franz C, Harris HMB, Mattarelli P, et al. A taxonomic note on the genus *Lactobacillus*: Description of 23 novel genera, emended description of the genus *Lactobacillus* Beijerinck 1901, and union of Lactobacillaceae and Leuconostocaceae. Int J Syst Evol Microbiol. 2020;70(4):2782–858.
2. Carr FJ, Chill D, Maida N. The Lactic Acid Bacteria: A Literature Survey. Crit Rev Microbiol. 2002;28(4):281–370.
3. O'Toole PW, Marchesi JR, Hill C. Next-generation probiotics: the spectrum from probiotics to live biotherapeutics. Nat Microbiol. 2017;2(5):1–6.
4. Ventura M, O'Flaherty S, Claesson MJ, Turroni F, Klaenhammer TR, van Sinderen D, et al. Genome-scale analyses of health-promoting bacteria: probiogenomics. Nat Rev Microbiol. 2009;7(1):61–71.
5. Hill C, Guarner F, Reid G, Gibson GR, Merenstein DJ, Pot B, et al. Expert consensus document The International Scientific Association for Probiotics and Prebiotics consensus statement on the scope and appropriate use of the term probiotic. Nat Rev Gastroenterol Hepatol. 2014;11(8):506–14.
6. Andreasen AS, Larsen N, Pedersen-Skovsgaard T, Berg RM, Moller K, Svendsen KD, et al. Effects of *Lactobacillus acidophilus* NCFM on insulin sensitivity and the systemic inflammatory response in human subjects. Br J Nutr. 2010;104(12):1831–8.
7. Leyer GJ, Li S, Mubasher ME, Reifer C, Ouwehand AC. Probiotic effects on cold and influenza-like symptom incidence and duration in children. Pediatrics. 2009;124(2):e172–9.
8. Ouwehand AC, DongLian C, Weijian X, Stewart M, Ni J, Stewart T, et al. Probiotics reduce symptoms of antibiotic use in a hospital setting: a randomized dose response study. Vaccine. 2014;32(4):458–63.
9. Chen YH, Tsai WH, Wu HY, Chen CY, Yeh WL, Chen YH, et al. Probiotic Lactobacillus spp. act Against Helicobacter pylori-induced Inflammation. J Clin Med. 2019;8(1):90.
10. Barefoot SF, Klaenhammer TR. Purification and characterization of the *Lactobacillus acidophilus* bacteriocin lactacin B. Antimicrob Agents Chemother. 1984;26(3):328–34.
11. Dobson AE, Sanozky-Dawes RB, Klaenhammer TR. Identification of an operon and inducing peptide involved in the production of lactacin B by *Lactobacillus acidophilus*. J Appl Microbiol. 2007;103(5):1766–78.
12. Sanders ME, Klaenhammer TR. Invited Review: The Scientific Basis of *Lactobacillus acidophilus* NCFM Functionality as a Probiotic. J Dairy Sci. 2001;84(2):319–31.
13. Altermann E, Russell WM, Azcarate-Peril MA, Barrangou R, Buck BL, McAuliffe O, et al. Complete genome sequence of the probiotic lactic acid bacterium *Lactobacillus acidophilus* NCFM. Proc Natl Acad Sci. 2005;102(11):3906–12.
14. Russell WM, Klaenhammer TR. Efficient system for directed integration into the *Lactobacillus acidophilus* and *Lactobacillus gasseri* chromosomes via homologous recombination. Appl Environ Microbiol. 2001;67(9):4361–4.
15. Goh YJ, Azcarate-Peril MA, O'Flaherty S, Durmaz E, Valence F, Jardin J, et al. Development and application of a upp-based counterselective gene replacement system for the study of the S-layer protein SlpX of *Lactobacillus acidophilus* NCFM. Appl Environ Microbiol. 2009;75(10):3093–105.
16. Goh YJ, Barrangou R. Portable CRISPR-Cas9(N) System for Flexible Genome Engineering in *Lactobacillus acidophilus*, *Lactobacillus gasseri,* and *Lactobacillus paracasei*. Appl Environ Microbiol. 2021;87(6):e02669-20.
17. Gilfillan D, Vilander AC, Pan M, Goh YJ, O'Flaherty S, Feng N, et al. *Lactobacillus acidophilus* Expressing Murine Rotavirus VP8 and Mucosal Adjuvants Induce Virus-Specific Immune Responses. Vaccines (Basel). 2023;11(12):1774.
18. Kajikawa A, Zhang L, LaVoy A, Bumgardner S, Klaenhammer TR, Dean GA. Mucosal Immunogenicity of Genetically Modified *Lactobacillus acidophilus* Expressing an HIV-1 Epitope within the Surface Layer Protein. PLoS ONE. 2015;10(10):e0141713.
19. Mohamadzadeh M, Duong T, Sandwick SJ, Hoover T, Klaenhammer TR. Dendritic cell targeting of Bacillus anthracis protective antigen expressed by *Lactobacillus acidophilus* protects mice from lethal challenge. Proc Natl Acad Sci U S A. 2009;106(11):4331–6.
20. Khazaie K, Zadeh M, Khan MW, Bere P, Gounari F, Dennis K, et al. Abating colon cancer polyposis by *Lactobacillus acidophilus* deficient in lipoteichoic acid. Proc Natl Acad Sci U S A. 2012;109(26):10462–7.
21. Mohamadzadeh M, Pfeiler EA, Brown JB, Zadeh M, Gramarossa M, Managlia E, et al. Regulation of induced colonic inflammation by *Lactobacillus acidophilus* deficient in lipoteichoic acid. Proc Natl Acad Sci U S A. 2011;108 Suppl 1(Suppl 1):4623–30.
22. Morovic W, Hibberd AA, Zabel B, Barrangou R, Stahl B. Genotyping by PCR and High-Throughput Sequencing of Commercial Probiotic Products Reveals Composition Biases. Front Microbiol. 2016;7:1747.

23. Pridmore RD, Berger B, Desiere F, Vilanova D, Barretto C, Pittet A-C, et al. The genome sequence of the probiotic intestinal bacterium *Lactobacillus johnsonii* NCC 533. Proc Natl Acad Sci. 2004;101(8):2512–7.
24. Sun Z, Harris HM, McCann A, Guo C, Argimon S, Zhang W, et al. Expanding the biotechnology potential of lactobacilli through comparative genomics of 213 strains and associated genera. Nat Commun. 2015;6:8322.
25. Huang Z, Zhou X, Stanton C, Ross RP, Zhao J, Zhang H, et al. Comparative Genomics and Specific Functional Characteristics Analysis of *Lactobacillus acidophilus*. Microorganisms. 2021;9(9):1992.
26. Sayers EW, Bolton EE, Brister JR, Canese K, Chan J, Comeau DC, et al. Database resources of the national center for biotechnology information. Nucleic Acids Res. 2022;50(D1):D20–6.
27. Pervez MT, Hasnain MJu, Abbas SH, Moustafa MF, Aslam N, Shah S. A Comprehensive Review of Performance of Next-Generation Sequencing Platforms. BioMed Res Int. 2022;2022:3457806.
28. Jeon S, Kim H, Choi Y, Cho S, Seo M, Kim H. Complete Genome Sequence of the Newly Developed *Lactobacillus acidophilus* Strain With Improved Thermal Adaptability. Front Microbiol. 2021;12:697351.
29. Brandt K, Barrangou R. Adaptive response to iterative passages of five *Lactobacillus* species in simulated vaginal fluid. BMC Microbiol. 2020;20(1):339.
30. Goh YJ, Barrangou R, Klaenhammer TR. In Vivo Transcriptome of *Lactobacillus acidophilus* and Colonization Impact on Murine Host Intestinal Gene Expression. mBio. 2021;12(1):e03399-20. https://doi.org/10.1128/mBio.03399-20.
31. Tettelin H, Riley D, Cattuto C, Medini D. Comparative genomics: the bacterial pan-genome. Curr Opin Microbiol. 2008;11(5):472–7.
32. Pan M, Hidalgo-Cantabrana C, Barrangou R. Host and body site-specific adaptation of *Lactobacillus crispatus* genomes. NAR Genomics Bioinform. 2020;2(1):lqaa001.
33. Zhou X, Yang B, Stanton C, Ross RP, Zhao J, Zhang H, et al. Comparative analysis of *Lactobacillus gasseri* from Chinese subjects reveals a new species-level taxa. BMC Genomics. 2020;21(1):119.
34. Pei Z, Sadiq FA, Han X, Zhao J, Zhang H, Ross RP, et al. Comprehensive Scanning of Prophages in *Lactobacillus*: Distribution, Diversity, Antibiotic Resistance Genes, and Linkages with CRISPR-Cas Systems. mSystems. 2021;6(3):e0121120.
35. Andersen JM, Barrangou R, Abou Hachem M, Lahtinen S, Goh YJ, Svensson B, et al. Transcriptional and functional analysis of galactooligosaccharide uptake by lacS in *Lactobacillus acidophilus*. Proc Natl Acad Sci U S A. 2011;108(43):17785–90.
36. Goh YJ, Klaenhammer TR. A functional glycogen biosynthesis pathway in *Lactobacillus acidophilus*: expression and analysis of the glg operon. Mol Microbiol. 2013;89(6):1187–200.
37. Johnson BR, O'Flaherty S, Goh YJ, Carroll I, Barrangou R, Klaenhammer TR. The S-layer Associated Serine Protease Homolog PrtX Impacts Cell Surface-Mediated Microbe-Host Interactions of *Lactobacillus acidophilus* NCFM. Front Microbiol. 2017;8:1185.
38. Klotz C, Goh YJ, O'Flaherty S, Barrangou R. S-layer associated proteins contribute to the adhesive and immunomodulatory properties of *Lactobacillus acidophilus* NCFM. BMC Microbiol. 2020;20(1):248.
39. O'Flaherty SJ, Klaenhammer TR. Functional and phenotypic characterization of a protein from *Lactobacillus acidophilus* involved in cell morphology, stress tolerance and adherence to intestinal cells. Microbiology (Reading). 2010;156(Pt 11):3360–7.
40. Douglas GL, Goh YJ, Klaenhammer TR. Integrative food grade expression system for lactic acid bacteria. Methods Mol Biol. 2011;765:373–87.
41. Kajikawa A, Nordone SK, Zhang L, Stoeker LL, LaVoy AS, Klaenhammer TR, et al. Dissimilar properties of two recombinant *Lactobacillus acidophilus* strains displaying *Salmonella* FliC with different anchoring motifs. Appl Environ Microbiol. 2011;77(18):6587–96.
42. O'Flaherty S, Klaenhammer TR. Multivalent Chromosomal Expression of the *Clostridium botulinum* Serotype A Neurotoxin Heavy-Chain Antigen and the *Bacillus anthracis* Protective Antigen in *Lactobacillus acidophilus*. Appl Environ Microbiol. 2016;82(20):6091–101.
43. Konstantinov SR, Smidt H, de Vos WM, Bruijns SC, Singh SK, Valence F, et al. S layer protein A of *Lactobacillus acidophilus* NCFM regulates immature dendritic cell and T cell functions. Proc Natl Acad Sci U S A. 2008;105(49):19474–9.
44. Buck BL, Azcarate-Peril MA, Klaenhammer TR. Role of autoinducer-2 on the adhesion ability of *Lactobacillus acidophilus*. J Appl Microbiol. 2009;107(1):269–79.
45. Klotz C, Goh YJ, O'Flaherty S, Johnson B, Barrangou R. Deletion of S-Layer Associated Ig-Like Domain Protein Disrupts the *Lactobacillus acidophilus* Cell Surface. Front Microbiol. 2020;11:345.
46. Selle K, Goh YJ, Johnson BR, O'Flaherty S, Andersen JM, Barrangou R, et al. Deletion of Lipoteichoic Acid Synthase Impacts Expression of Genes Encoding Cell Surface Proteins in *Lactobacillus acidophilus*. Front Microbiol. 2017;8:553.
47. Goh YJ, Klaenhammer TR. Insights into glycogen metabolism in *Lactobacillus acidophilus*: impact on carbohydrate metabolism, stress tolerance and gut retention. Microb Cell Fact. 2014;13:94.
48. Zadeh M, Khan MW, Goh YJ, Selle K, Owen JL, Klaenhammer T, et al. Induction of intestinal pro-inflammatory immune responses by lipoteichoic acid. J Inflamm (Lond). 2012;9:7.
49. Foley MH, O'Flaherty S, Allen G, Rivera AJ, Stewart AK, Barrangou R, et al. *Lactobacillus* bile salt hydrolase substrate specificity governs bacterial fitness and host colonization. Proc Natl Acad Sci U S A. 2021;118(6):e2017709118.
50. Barrangou R, Altermann E, Hutkins R, Cano R, Klaenhammer TR. Functional and comparative genomic analyses of an operon involved in fructooligosaccharide utilization by *Lactobacillus acidophilus*. Proc Natl Acad Sci U S A. 2003;100(15):8957–62.
51. Foley MH, Walker ME, Stewart AK, O'Flaherty S, Gentry EC, Patel S, et al. Bile salt hydrolases shape the bile acid landscape and restrict *Clostridioides difficile* growth in the murine gut. Nat Microbiol. 2023;8(4):611–28.
52. Barrangou R, Dudley EG. CRISPR-Based Typing and Next-Generation Tracking Technologies. Annu Rev Food Sci Technol. 2016;7:395–411.
53. Brandt K, Barrangou R. Using glycolysis enzyme sequences to inform *Lactobacillus* phylogeny. Microb Genom. 2018;4(6):e000187.
54. Liang G, Bushman FD. The human virome: assembly, composition and host interactions. Nat Rev Microbiol. 2021;19(8):514–27.
55. Barrangou R, Briczinski EP, Traeger LL, Loquasto JR, Richards M, Horvath P, et al. Comparison of the Complete Genome Sequences of Bifidobacterium animalis subsp lactis DSM 10140 and Bl-04. J Bacteriol. 2009;191(13):4144–51.
56. Pan M, Morovic W, Hidalgo-Cantabrana C, Roberts A, Walden KKO, Goh YJ, et al. Genomic and epigenetic landscapes drive CRISPR-based genome editing in *Bifidobacterium*. Proc Natl Acad Sci. 2022;119(30):e2205068119.
57. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014;30(14):2068–9.
58. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 2015;25(7):1043–55.
59. Wishart DS, Han S, Saha S, Oler E, Peters H, Grant JR, et al. PHASTEST: faster than PHASTER, better than PHAST. Nucleic Acids Res. 2023;51(W1):W443–50.
60. Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, et al. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics. 2012;28(12):1647–9.
61. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun. 2018;9(1):5114.
62. Alikhan N-F, Petty NK, Ben Zakour NL, Beatson SA. BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. BMC Genomics. 2011;12(1):402.
63. Darling ACE, Mau B, Blattner FR, Perna NT. Mauve: Multiple Alignment of Conserved Genomic Sequence With Rearrangements. Genome Res. 2004;14(7):1394–403.
64. 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(22):3691–3.
65. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 1994;22(22):4673–80.
66. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30(9):1312–3.

Publisher's Note