ABSTRACT

BRANDT, KATELYN. Genomics and Phylogeny of Probiotic Lactic Acid Bacteria Used in Industry. (Under the direction of Dr. Rodolphe Barrangou).

Lactic acid bacteria (LAB) are known for their widespread occurrence and diverse commercial use. Several species of LAB are industrially formulated to enhance human health through probiotics and modulation of the microbiome, as well as food production as starter cultures. LAB are as genetically and functionally diverse as their many uses, and thus have been extensively investigated. Here, we use functional genomics and molecular biology to develop a new phylogenetic technique to improve genotyping and develop frameworks to assess potential industrial cultures and inform strain selection for commercial formulation.

First, we developed a phylogenetic approach based on genes encoding enzymes driving the glycolysis pathway. As a universal biochemical hallmark, it has appeal with regards to widespread occurrence such as ribosomal RNA sequences, but also has the advantage of hinging on several genes encoding multiple enzymes that can be concatenated to increase resolution and sequence variability. We show that this technique can be used to assess the phylogeny of *Bifidobacterium* and *Lactobacillus* and show that this method is valuable in cases of well-established relationships and complicated phylogeny, respectively. Conveniently, analyses can be carried out on the whole pathway, portions thereof, or hypervariable genes, showcasing both consistency with and higher resolution than 16S rRNA sequences. We also show how these genes reflect the evolutionary path of the rest of the genome more extensively with regards to GC content.

Next, we used the aforementioned method to group select species of *Lactobacillus* and investigated an under-studied phylogenetic cluster, focusing on *Lactobacillus fermentum*. We determined the genome of the type strain ATCC 14931 and used comparative genomic analyses

to investigate genetic diversity amongst 37 other strains within this species. Several transposon-enriched hypervariable islands were identified; we determined the occurrence of CRISPR-Cas systems, revealing enrichment and diversity across the species, with focus on and characterization of a widespread Type I CRISPR-Cas system.

Finally, we carried out a long-term experiment for five distinct *Lactobacillus* species cultured for one thousand generations in a simulated vaginal fluid. We selected species known to dominate the human vaginal microbiome, notably *Lactobacillus crispatus, Lactobacillus jensenii,* and *Lactobacillus gasseri;* we compared their genome evolution and transcriptional response to those of an intestinal species, *Lactobacillus acidophilus,* and a food culture, *Lactobacillus fermentum.* Growth patterns revealed vaginal species outperforming non-vaginal species. Genetically, we found that non-vaginal strains were more likely to acquire SNPs than vaginal strains. Whole transcriptome comparisons reflected the shift from a lab-rich growth medium to a more nutrient-constrained synthetic proxy reflecting *in vivo* conditions. Results determined that vaginal strains were more aptly equipped for growth and survival in a simulated vaginal environment, predictably.

Overall, these studies provide a methodological framework for the functional genomic investigation of industrial lactic acid bacteria, shedding light on their phylogenetic relationships and genetic adaptation and transcriptional responses to various environments. This work will aid in the selection and development of novel probiotic strains for human gastrointestinal and vaginal health.

Genomics and Phylogeny of Probiotic Lactic Acid Bacteria Used in Industry

by
Katelyn Brandt

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Functional Genomics

Raleigh, North Carolina
2019

APPROVED BY:

_____	_____
Dr. Rodolphe Barrangou Committee Chair	Dr. Jeffrey L. Thorne
_____	_____
Dr. Matthew Koci	Dr. Alison Motsinger-Reif

DEDICATION

To my family: Bill, Jeannie, and Sarah Brandt. Thank-you for your never-ending support, patience, and advice as I have pursued my dreams.

BIOGRAPHY

Katelyn Brandt was born in Philadelphia, PA, to parents Bill and Jeannie Brandt on May 3, 1992. She was raised outside of Philadelphia along with her younger sister, Sarah, and graduated Valedictorian from Oxford Area High School in 2010. Katelyn then enrolled at Pennsylvania State University. While there, she was heavily involved with Penn State's THON, spreading awareness and raising funds for childhood cancer. Katelyn also performed undergraduate research under the direction of Dr. Allen T. Phillips. In 2013, Katelyn graduated from the Eberly College of Science at Penn State with a B.S. in Biochemistry and Molecular Biology. From here, Katelyn joined the Functional Genomics Program at North Carolina State University to pursue a doctoral degree under the direction of Dr. Rodolphe Barrangou. While at NC State University, she continued her passion of spreading awareness of childhood illness with Dance Marathon at NC State, where she served as Vice President of Business. Katelyn also had the opportunity to serve as a student representative through the Chancellor's Aides Program.

ACKNOWLEDGMENTS

First, I would like to first acknowledge Ms. Jennifer Pusey for being an incredible middle school science teacher. Without your inspiration, I may never have found my passion for science.

I would like to acknowledge my committee: Dr. Rodolphe Barrangou, Dr. Jeffrey Thorne, Dr. Matthew Koci, Dr. Alison Motsinger-Reif, and Dr. Casey Theriot. Thank-you for your support and guidance over the years. To Dr. Rodolphe Barrangou, thank-you for your belief in my success, your trust in my abilities, and your patience as I pushed the boundaries in my research and professional development.

The foundation of a great research project is the support provided from a lab, and the CRISPRlab is no exception. A special thanks to Dr. Todd Klaenhammer, whose efforts first established our group, and who provided us with a great legacy to continue. A heartfelt thank-you to the staff members of the lab: Rosemary Sanozky-Dawes, Dr. Yong Jun Goh, Dr. Sarah O'Flaherty, Dr. Claudio Hidalgo Cantabrana, and Natalia Cobián Fernández. Thank-you for your constant support and willingness to answer questions and solve problems. Thank-you to my fellow students for being a great support system and sounding board: Dr. Courtney Klotz, Matthew Nethery, Meichen (Echo) Pan, Elice Kitchen-McKinley, and Allison Fulp. Also, thank-you to all past members in the lab who have aided and supported me on this journey: Evelyn Durmaz, Dr. Emily Henriksen, Katheryne Daughtry, Dr. Brant Johnson, Dr. Kurt Selle, Jeffrey Hymes, Dr. Joakim Andersen, Dr. Mia Theilmann, Emily Stout, Cassandra Cañez, Dr. Alexandra B. Crawley, and Stefanie Andersen.

Finally, I would like to acknowledge my family: Bill, Jeannie, and Sarah Brandt for their unwavering support. This journey would not have been as successful or fulfilling without you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: FUNCTIONAL GENOMICS OF INDUSTRIAL LACTIC ACID BACTERIA

1.1. CONTRIBUTION TO THE WORK

Katelyn Brandt is the main author on the following chapter, she wrote, edited, and designed figures. Rodolphe Barrangou aided in editing. Sarah Brandt aided in graphical representation.

1.2. ABSTRACT

Lactic acid bacteria (LAB) are one of the most common industrially employed and studied groups of microorganisms. Derived from a shared common ancestor and connected by their retained ability to produce lactic acid, this complex is phylogenetically expansive, along with being functionally and genetically diverse. This diversity has led LAB to make homes of various habitats and hosts, typically representing important members of the microbiomes of humans, animals, and plants. Accordingly, they are widely used in the food industry. There is a rich history of LAB interacting with and affecting their environment, just as the environment interacts with and affects LAB species. This interplay has shaped several areas within the industry, such as the food, human health, and agricultural sectors. Here, we analyze the genomic trends of LAB, arising from long-term adaptation to industrially relevant environments. Next, we examine how specific LAB species and strains have uniquely adapted to their environments. We examine how these adaptations have allowed LAB species to colonize, compete, and thrive in various settings, such that they have become widely formulated across various industrial segments, notably the food industry, across the supply chain. We then reflect on how these changes are driving industry decisions via the impact of LAB genomics.

1.3. INTRODUCTION

Lactic Acid Bacteria (LAB) are one of the most ubiquitous and utilized groups of microbes. These microorganisms are a part of the normal microbiota of humans, animals, and plants. They are also heavily used in the food industry as starter cultures for foods, such as fermented dairy, meat, and vegetable products. Due to their pervasiveness and documented history of safe use, many LAB species are labelled with the Qualified Presumption of Safety (QPS) status from the European Union (EU), or Generally Regarded as Safe (GRAS) status in the United States (US). For these reasons, these industry-relevant organisms have been the focus of many studies over the years.

The LAB group is defined as microaerophilic, Gram-positive bacteria that produce lactic acid as their primary fermentation end product [1]. The complex consists of several genera, including *Enterococcus, Pediococcus, Streptococcus, Oenococcus, Lactococcus, Leuconostoc, Weissella,* and *Lactobacillus* [2]. Although not technically considered a *bona fide* member of the LAB group, the *Bifidobacterium* genus is often considered in parallel due to its similar habitats (notably the human gastrointestinal tract (GIT)) and applications (as widely used commercial probiotics). LAB are common members of the human and animal microbiomes and multiple studies have linked their presence to a healthy status [3-5]. In addition, LAB are the common strains found in probiotic products. In the food industry, strains are responsible for the effective fermentation of many different food products, such as sauerkraut, pickles, cheese, and sausages [6-9].

The advent of omics-based technologies in the past decades, from genomics to transcriptomics and now metagenomics, enabled the characterization of LAB functionality in host environments. Mainly, research has focused on the potential benefits LAB provide in human

health (modulating gut homeostasis), agriculture (increased yields), and food manufacturing (genesis of desirable texture attributes and production of flavor compounds) [10-12]. Mechanistic studies have determined how each species is suited to their environment. LAB habitats are typically nutrient-rich and have heavily shaped the genomic trajectory of the group. Species of LAB have adapted to best suit their environments in numerous ways. While some of the adaptations and evolutionary paths are similar between species and genera (such as carbohydrate uptake), there is also a great deal of variation, such as the ability to produce exopolysaccharide (EPS) layers. Indeed, environment adaptation can be found on a genus-, species-, and strain-specific level. The interplay between host and strain has had a profound effect on the evolutionary history of LAB. Each habitat has provided its own conditions for species to adapt to, and beyond a few universal, well-characterized adaptations, each species has adapted in its own way. Yet, despite being nutrient-rich, human, animal, plant, and food habitats are not without their own set of stress conditions. For instance, in human and animal microbiomes, organisms must be able to survive GIT transit and ideally adhere to epithelial cells. Oftentimes, this means resisting acidic conditions and earning a competitive edge over other possible commensals and pathogens. While these challenges and resultant unique adaptations make evolutionary history in LAB difficult to parse, it also reveals the great flexibility of the group. The ability to uniquely adapt to various environments is what has enabled so many LAB species to become cornerstones in their community. Additionally, these abilities have shown great benefit to the agricultural, food manufacturing, and consumer industries. In this review, we will see how LAB have become valuable for these industry sectors, and most importantly, how they are adding benefit and value.

1.4. LACTIC ACID BACTERIA GENOMICS

Rising accessibility of sequencing technologies has led to thorough genomic studies of LAB, determining their genetic content to understand the genetic basis for their phenotypic features of interest. This has enabled a greater understanding of phylogeny, adaptation, and functional attributes of LAB.

Comparative genomics show that LAB, specifically the *Lactobacillales* order, have evolved from a common ancestor primarily through systematic gene loss and the occasional burst of gene acquisition [13]. This gene loss is a hallmark of LAB growing in nutrient-rich environments, as those found in the food manufacturing, agricultural, and consumer industries, with no selective pressure to maintain a large genome. Gene acquisition is lineage-dependent and occurs primarily through horizontal gene transfer (HGT); it is used to gain niche-specific adaptations for a competitive advantage in an organism's preferred environment. Specifically, LAB are marked by a loss of carbohydrate degradation and cofactor biosynthesis, but also an increase in peptidases and transporters [13, 14]. In fact, in a typical LAB genome, transporters make up 15% of the coding material [15], which allows them to make use of the nutrients provided by their environment.

In this review, we focus on select examples across diverse commercial applications to illustrate unique LAB adaptations to host environments. Our analyses focus on relevant species of *Lactobacillus* (10), *Streptococcus* (1), *Lactococcus* (3), and *Bifidobacterium* (3). Descriptions of their genomic statuses can be found in Table 1.1, while visual representation of their genomes can be found in Figure 1.1. The genomes range in size from 1.67Mb (*Lactobacillus jensenii*) to 3.31Mb (*Lactobacillus plantarum*). GC content varies from 32.9% (*Lactobacillus salivarius*) to 60.5% (*Bifidobacterium animalis* subsp. *lactis*). A few of the included species are known to

carry plasmids: *Lb. plantarum*, *Lb. salivarius*, *Bifidobacterium longum*, and *Lactococcus lactis* subsp. *cremoris.* Figure 1.1 compares genome size of the select species, as well as a visual representation of genomic traits of interest, such as the transporters mentioned above. Bacteriocin genes, EPS genes, and Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) sites are also highlighted and will be discussed below.

Genome size, gene content, plasmids, and other genomic features may be applied in order to reveal general trends in the adaptation of LAB species. As one of the most widely utilized genera of LAB, *Lactobacillus* has been the focus of several recent genomic studies and serves as guide to trends in other LAB genera, since genomic trends of lactobacilli and LAB mimic each other [16]. A common school of thought is that smaller genomes belong to organisms with narrow habitat ranges, while larger genomes belong to those with varied habitats. *Lb. jensenii*, the organism with the smallest genome in this study, is primarily found in the human vaginal microbiome [17]. In contrast, *Lb. plantarum*, the organism with the largest genome in this study, has numerous habitats, showing that *Lb. plantarum* strains do not show specialized adaptations to a single environment [18]. Rather, strains maintain several genomic islands that carry the necessary genes to survive in various habitats [19]. This contrasts with other lactobacilli whose strains carry unique islands for their specialized habitat, such as was found in *Lactobacillus rhamnosus* [20]. Such flexibility of *Lactobacillus* is exemplified by its larger-than-average pan-genome [21], and has allowed lactobacilli to become cornerstone members in several industry sectors.

Lactobacillus is not the only LAB genus to have been shaped at a genomic level by their environment. For example, a study of bifidobacteria analyzed how the host's glycan source shaped the evolution of bifidobacteria species [22]. In addition, it has been proposed that due to

its growth in dairy environments, *Streptococcus thermophilus* underwent genome decay and lost much of its virulence capabilities [23].

1.5. PHYLOGENY OF LACTIC ACID BACTERIA

LAB are considered part of the order *Lactobacillales,* of the *Bacilli* class and *Firmicutes* phylum [24]. Historically, LAB have typically been split into two groups: homofermentative organisms that produce lactic acid in fermentation, and heterofermentative species that produce lactic acid, ethanol, and carbon dioxide during fermentation [25]. Members of *Lactococcus* and *Pediococcus* belong to the homofermentative group while members of *Leuconostoc* and *Weissella* belong in the heterofermentative group; *Lactobacillus* has members in both groups [25]. *Lactococcus* and *Streptococcus* share a close phylogenetic relationship [23]. Bifidobacteria, although traditionally considered LAB, are phylogenetically distant to other genera of LAB [23]. A phylogenetic tree based on 16S rRNA can be found in Figure 1.1. It shows *Lactobacillus* as one group that is distinct from the other genera and also illustrates the close relationship between *Streptococcus* and *Lactococcus.*

Lactobacillus is the largest and most diverse genus of LAB [23]. This paraphyletic clade contains over 200 species and subspecies, as well as several other genera including the *Pediococcus*, *Leuconostoc*, *Oenococcus*, *Weissella*, and *Fructobacillus* genera [26]. Originally, lactobacilli were phylogenetically defined by their fermentation capabilities, but this proved to be less than satisfactory and genomic approaches began to be utilized [26, 27]. Since then, several approaches have been developed.

A recent phylogenetic study used niche association to develop an evolution model for *Lactobacillus,* with the common ancestor being a free-living organism from which species have adapted specifically to their host, the host falling into one of four categories: vertebrate, insect, nomadic, or free-living [28]. Of our analyzed species -- for those with available data -- most are vertebrate-adapted. In line with the model, the smaller genome of *Lb. jensenii* would be a result

of its narrow habitat range. In contrast, the large genome of *Lb. plantarum* is in line with its nomadic lifestyle. Its habitat range includes humans, animals, plants, and food products. This approach appears promising as niche adaptations have been noted as a driving force in genomic changes, as mentioned above for *Bifidobacterium* and *Streptococcus*. Additionally, this methodology still allows for the grouping of heterofermentative and homofermentative species [28], a split that has led to much discussion in recent phylogenetic studies of *Lactobacillus* [29]. Overall, it is apparent that LAB environments shape LAB evolution and genomics as much as LAB shape their environment.

1.6. APPLICATIONS OF LACTIC ACID BACTERIA IN AGRICULTURE

Due to their ubiquitous nature, GRAS and QPS status, and health-promoting benefits, there has long been an assumed connection between LAB and agriculture. This ideology was enforced when LAB were continuously isolated from crops and livestock. Soon, the practice of adding LAB to feeds and fertilizers began, and this application has only increased with rising concerns over antibiotic and hormone use. As consumers become more concerned with where their food is coming from and how it is grown, the market has adapted to employ nature-made solutions – rather than man-made – in order to promote growth, add functional benefits, and to eliminate pathogens. This has led to a push for probiotic alternatives in the agricultural sector. Probiotics have been defined as "live microorganisms which when administered in adequate amounts confer a health benefit on the host" [30]. For this to occur, species must be able to survive in their host's environment, specifically utilize the available nutrients, tolerate stress conditions, and interact with the host in such a way to provide a benefit. While probiotic research has historically focused primarily on humans as hosts, these recent trends in agriculture have served as the impetus to expand their use in livestock and plants as additional hosts, and thus products. Initial research in livestock microbiomes and probiotics have revealed similar trends to those as seen in humans (see below). However, mechanistic and genomic studies meant to unravel the interplay between hosts and LAB in the agriculture sector are still in their infancy.

1.6.1. Uses of Lactic Acid Bacteria in the Livestock Industry

With a desire to better understand animal health and to introduce non-antibiotic measures such as probiotics, there has been an increasing amount of studies focusing on microbiome and probiotic usage in livestock. These recent studies are revealing that many animal trends mimic

those already known in humans; this allows us to use tools developed for other industries and knowledge derived from humans to improve livestock health and production. For instance, in addition to maintaining microflora balance in the intestine, LAB are capable of affecting animal immunity through their influence on intestinal permeability and their ability to produce bioactive molecules [31]. Furthermore, GIT homeostasis is important to animal health, and bifidobacteria have been identified as important members of the animal GIT, their presence being indicative of a healthy status [31]. Research has also shown that animal microbiomes change and adapt as the host ages, and that proposed benefits and adaptations are on a strain level, not a species level. For example, strains of *Lactobacillus ruminis* from infant bovine and porcine showed beta-galactosidase activity indicative of milk (lactose) being part of the diet, but strains from adult racehorses did not show this activity [32]. This also means that the species and strain selected to be used as potential probiotics will most likely vary based on the animal, age, lifestyle, and goal [8].

Beyond general trends, studies have already begun to look at the effect LAB has on various livestock species. There is evidence that LAB would be effective probiotics in calves, pigs, and chickens [31]. Probiotics are being examined in poultry because it is believed that their use will prevent pathogens and promote growth of the host without the use of antibiotics [33]. Specifically, *Lactobacillus* species have been shown to inhibit the growth of several foodborne pathogens [33]. Additionally, LAB have been linked to increased egg production in chickens and milk yields in ruminants [31]. LAB species already employed as animal probiotics include *Lc. lactis*, *Pediococcus acidilactici*, and *Lactobacillus acidophilus* [8]. In addition, both *Lb. jensenii* and *Lc. lactis* have shown genomic attributes to improve immune response in animals (Table 1.1).

One way to introduce LAB strains to livestock is through their feed. Silage is a common feed for livestock in certain climates [34], and is often fermented using LAB species. It is believed that silage fermentation begins spontaneously with LAB genera such as lactococci and pediococci, before finishing with lactobacilli. As such, commercial products typically incorporate *Lb. plantarum, Lactobacillus casei,* and *Lb. rhamnosus* [34]. LAB have specifically been added for their antifungal properties in order to protect grass silage during storage [35].

1.6.2. Use of Lactic Acid Bacteria in Crop Manufacturing

In contrast to animal studies, there has been few mechanistic studies on LAB in crops. LAB and plants are less studied since LAB are relatively minor members of the plant microbiome. Most LAB plant species are considered epiphytes, with some evidence to show species serving as endophytes [36]. *Weissella cibaria/Weissella confusa* and *Lb. plantarum* are the most commonly found LAB species, while other members of the *Lactobacillus, Leuconostoc, Pediococcus,* and *Enterococcus* genera have also been identified occasionally [37]. Overall, LAB have been isolated from herbs, sugarcane, grass, cereals, fruits, and vegetables [38].

Plant environments differ greatly from the animal, food manufacturing, and human environments from which LAB are most typically isolated. Growth conditions provide a whole new set of challenges and, as such, LAB must uniquely adapt, typically through the formation of niche-specific genomic islands. *Lb. plantarum* and *Lb. casei*, two of the larger genomes analyzed in our set (Figure 1.1), both have genomic islands that allow them to adapt to numerous niches, including plants [39, 40]. Plant-specific adaptations that LAB need to thrive in this environment are the abilities to handle the phenolic compounds that are often a part of the metabolic pathways of plants and to have antimicrobial properties [36]. LAB plant strains have adapted to utilize

phenolic compounds in order to remove them from their environment; many strains use these compounds as external electron acceptors [41]. After determining which species and strains can grow in this different environment, studies began to look at how LAB interacted with their host. Most studies focused on the benefit the plants receive and evaluate the LAB's ability to act as growth promoters, provide pathogenic resistance, and alter the host's expression profile [42]. LAB, such as *Lb. plantarum, Lactobacillus delbrueckii*, *Lactobacillus fermentum,* and *B. longum*, have been sprayed on plants and soil as EM (effective microorganisms) to increase plant growth and pathogen resistance [35].

1.6.3. Agricultural Genomic Adaptations

Since probiotics in livestock are proposed as an alternative to antibiotic use, most studies have focused on a proposed species ability to outcompete or eliminate pathogens. As it has already been employed as a probiotic in livestock, *Lc. lactis* has been the focus of numerous *in vivo* studies, specifically looking at the prevention of mastitis in bovine. One study showed that the administration of *Lc. lactis* was able to clear mastitis infection as well as antibiotics did [43]. Further studies have shown that *Lc. lactis* is able to modulate the host's immune response to better prime the host to combat infection [44, 45]. These studies, combined with the known fact that *Lc. lactis* produces the bacteriocin lacticin, have led to the proposal that *Lc. lactis* is a suitable alternative to antibiotics (Figure 1.2) [45].

Like the studies completed to date in livestock, very few shed light on the genomic attributes and mechanisms of action of LAB in plants. The studies that have focused on LAB genomics and their effect on plants focus, again, mainly on pathogenic protection. Exclusion of pathogens not only provides a benefit to the host, but also limits competition for LAB. *Lb.*

plantarum encodes several bacteriocins that are predicted to help mitigate the effects of pathogenic microorganisms on their hosts (Table 1.1). Several strains of *Lb. plantarum* coding for a ClassIIb bacteriocin were able to combat the effects of fire blight on pear and apple plants (Figure 1.2) [46].

1.7. USE OF LACTIC ACID BACTERIA FOR FOOD MANUFACTURING

Humankind has used LAB in the food production process for millennia. This is, in part, due to their ubiquitous presence -- LAB species have been isolated from tomatoes, pineapples, peppers, carrots, eggplants, cucumbers and several other fruits and vegetables [37]. It is also due to LAB's ability to spontaneously ferment many of the listed fruits and vegetables. LAB have uniquely adapted to be reliant on fermentation, despite it being a high-energy process, which enables the production of various fermented vegetables, as well as several dairy and meat products [47]. In fact, the production of fermented foods uses LAB as its most common starter [48].

Food fermentation is meant to achieve the development of flavor and texture, provide food preservation, add fortification, and decrease cooking time [48]. Depending on the starting material, desired flavor and textures, and the final product, strains are selectively used to drive the food manufacturing process, thus leading to specialized genomic adaptations, as discussed below. While fermentation is their main role in food manufacturing, LAB also add flavors and textures to food products [25]. LAB are universal in the food manufacturing process from causing fermentation, adding flavor and texture, providing protection from spoilage, and in some cases, acting as agents of spoilage [8, 49].

Several species have functional attributes beyond their fermentation capabilities that make them desirable starter cultures in the food manufacturing industry, such as having the ability to produce EPS for texture, protect against phages that cause spoilage, and produce desired sugars for taste. Beyond the development of food products, LAB bacteriocins such as nisin have been proposed as and used as natural food preservatives [37]. Species that have been used in the food manufacturing process include *Lb. casei, Lb. fermentum, Lb. delbrueckii* subsp.

bulgaricus, Lb. salivarius, S. thermophilus, Lc. lactis subsp. *lactis,* and *Lc. lactis* subsp. *cremoris* (Table 1.1). *S. thermophilus* and *Lc. lactis* subsp. *cremoris* have a well-documented history in the food manufacturing process, specifically in the production of yogurt.

1.7.1. Lactic Acid Bacteria for Fermented Dairy Products

LAB are well known for their ability to ferment dairy products including milk, cheeses, and yogurt. Starter cultures of LAB greatly affect the outcome of the final product, as they not only provide flavor compounds, but also texturizing compounds, including consistency, syneresis, and firmness. Depending on the species and strains used, LAB impact the flavor and development of dairy products through their metabolism of lactate, their lipolysis properties, and proteolysis capabilities [25]. Milk flavors come from the production of acetic acid, acetaldehyde, and diacetyl, produced by species such as *Lc. lactis, S. thermophilus,* and *Lb. delbrueckii* subsp. *bulgaricus* [50]. Ripened cheese flavors are produced through proteolysis performed by *Lc. lactis, S. thermophilus,* and *Lb. delbrueckii* subsp. *bulgaricus* [50].While fermentation is vital to the production of fermented dairy products, other LAB adaptations are also important to their development, such as EPS and antimicrobial agents.

The ability to produce EPS is important for the development of several fermented dairy products. EPS' effect differs depending on the species and the underlying structure. The EPS of certain LAB species have been associated with several properties of yogurt, including texture, stability, and sensory properties [51]. Because EPS improves sensory and rheological properties, there is less of a need to add chemicals to the final product [52]. LAB species in yogurt production that produce an EPS include *Lc. lactis* subsp. *cremoris* and *Lb. delbrueckii* subsp. *bulgaricus*. Additional adaptations important to yogurt production include the ability to prevent

pathogens. Fermented dairy products are susceptible to pathogens, which can cause spoilage and thus raises the cost of production. The main adaptation contributing to LAB's ability to act as an antimicrobial is the fact that LAB acidify their environment during fermentation, preventing the growth of pathogens [25]. An additional method to prevent pathogen spoilage is CRISPR (discussed below).

While individual strains have adapted to the fermented milk environment, two strains, *S. thermophilus* and *Lb. delbrueckii* subsp. *bulgaricus*, adapted in tandem and are best known for their protocooperative growth. As mentioned earlier, growth in a rich medium often leads to genome decay in LAB species. For these two species, their decay has occurred in a complimentary fashion. In other words, they each provide the other with essential growth factors that their partner can no longer generate [53]. Intriguingly, yogurt production is better when both organisms are used as compared to either one alone [25].

1.7.2. Lactic Acid Bacteria and Other Fermented Foods

Perhaps best known for their work in fermented dairy products, LAB are also commonly used to ferment other foods such as vegetables, fruits, and meats. LAB have been used in the manufacturing of several fermented vegetable and fruit products including sauerkraut, kimchi, capers, cabbage, radishes, cucumbers, and beets [37, 48]. The main LAB used in vegetable fermentations are *Lactobacillus*, *Leuconostoc*, and *Pediococcus* [48]. LAB are also used in meat production. Specifically, homofermentative LAB are preferred in meat fermentation [54]. They are preferred because their fermentation prevents spoilage [50] and their ability to produce bacteriocins combat foodborne pathogens [54]. *Lactobacillus* is primarily used in fermented meats: *Lactobacillus sakei* is the most commonly used species; however, *Lb. plantarum* has also

been utilized [54]. Finally, LAB have been used to create bread products, specifically used as common starters in sourdough. *Lb. plantarum* and *Lb. delbrueckii* have both been used in sourdough production, yet *Lactobacillus sanfranciscensis* is perhaps the most well-known [55]. The fermentation processes allows for a more aerated bread, and lactobacilli also contribute through proteolysis, anti-mold production, and the production of volatile compounds [55].

1.7.3. Genomic Adaptation to Industrial Manufacturing Environments

LAB strains have specifically adapted to food environments. Some adaptations are universal -- such as their fermentation ability -- but other adaptations are specific to either species or strains. One such adaptation is the ability to produce EPS. As mentioned above, EPS is able to affect various rheological properties during dairy product fermentation. Interestingly, while many LAB strains have the ability to produce EPS, the exact structure of the EPS differs by strain, and therefore, the effect of the EPS is strain-specific. One study on EPS focused on *Lc. lactis* subsp. *cremoris,* a species commonly used in yogurt production. An analysis of several of the EPS-producing strains of *Lc. lactis* subsp. *cremoris* determined that it is the molecular makeup of the EPS that determines its overall impact on viscosity, not just the production of EPS (Figure 1.2). In other words, the more structured and viscous the EPS structure of the strain is, the more viscous the final fermented dairy product will be [56].

Also as stated above, an additional benefit of using LAB in food manufacturing is the ability to act as an antimicrobial. LAB have several ways of preventing phage infection and spoilage, however the most popular would be CRISPR. While most well-known for its genetic engineering capabilities and the “CRISPR craze,” CRISPR was actually first identified as an adaptive immune system in *S. thermophilus* [57, 58]. As bacteria and phage have been engaged

in an arms race for millennia, the identification of CRISPR in *S. thermophilus* revealed a new widespread defense system in bacteria. It was demonstrated that when *S. thermophilus* cultures were challenged with phage, the host was able to fend off the attack if the strain had a CRISPR system (Figure 1.2). Additionally, it was shown that the strain was able to genetically capture a memory of the infection in order to fend off subsequent infections. This means that culturing using *S. thermophilus* strains with functional CRISPR loci are preferred in manufacturing, as they will be less likely to spoil due to phage infection.

1.8. PROBIOTIC LACTIC ACID BACTERIA AND PROMOTING CONSUMER HEALTH

While each of the previously discussed industrial sectors have benefitted from advances in microbiome studies, the greatest advances thus far have arguably been achieved in the human microbiome. Scientists have discovered a community much more numerous, diverse, and complex than originally expected. Most importantly, researchers are beginning to understand how large a role microbiomes play in human health. LAB have been at the forefront of many of these studies due to their ubiquitous presence across human microbiome sites: skin, oral, GIT, and vaginal.

Although ubiquitous, the majority of studies on the interaction between LAB and humans have primarily focused on the GIT and, more recently, on the vaginal microbiomes. The role of LAB in the gut microbiome is well-studied due to the use of LAB as probiotics. Scientists now know that the human gut is sterile at birth, is rapidly colonized, and then transitions to an adult microbiome in early childhood [59]. From a mechanistic standpoint, researchers have evaluated how LAB adhere to host cells, utilize the available nutrients, and overcome various stress conditions to not only grow, but to also provide a benefit to the consumer. LAB are noted as health-promoting in consumer microbiomes for their ability to act as an antimicrobial via the lowering of pH in their environment and the production of organic acids [38].

1.8.1. Lactic Acid Bacteria and the Host Gastrointestinal Tract

The best studied microbiome is arguably that of the human GIT. Studies have shown how important the normal microflora is to establishing homeostasis. It is important for beneficial bacteria to maintain colonization to prevent pathogens, absorb bacterial metabolites, and interact

with the immune system of the host [60]. The main LAB genera in the GIT are *Bifidobacterium, Lactobacillus, Enterococcus,* and *Streptococcus* [61]. *Lb. acidophilus, Lactobacillus gasseri, Lb. rhamnosus, Lb. casei, Lb. salivarius, B. animalis* subsp. *lactis, Bifidobacterium breve, B. longum,* and *Lc. lactis* subsp. *lactis* all have functional attributes that benefit consumer gut health (Table 1.1). Research into how these organisms colonize, interact with, and benefit the host have led to better understanding of how the GIT microbiome functions and maintains a healthy state throughout aging.

A decrease of lactobacilli and bifidobacteria are linked to health risks at all ages [61]. In infants, a high incidence of *Bifidobacterium* is considered a healthy state. Since infant guts are sterile, the colonization of bifidobacteria excludes the colonization of pathogenic bacteria. *Bifidobacterium* first gain a competitive advantage by being introduced to the gut through vertical transmission, passing directly from mother to child [62].

Research has largely focused on the benefit of adding LAB to the GIT through the use of probiotics. Several clinical studies show that LAB have a positive impact on the health status of the host [61]. *Lactobacillus* and *Bifidobacterium* are heavily studied in the gut due to their prevalence and their potential as probiotics [61]. These studies have revealed functional attributes that are necessary for strains to be able to survive and provide a probiotic effect. Briefly, in order to survive passage, probiotic strains need to have tolerance to acid, gastric juice, and bile [38]. Adhesion has been shown to be important for immune modulation and pathogen exclusion [38], and is a way for strains to interact with and grow in the host. Studies have shown that one benefit of *Lactobacillus* probiotics is that they have been able to improve barrier function in the intestinal tract [63]. A final interesting note about probiotic function is that

lactobacilli species represent both allochthonous and autochthonous species in the GIT, however even allochthonous species are able to impart health benefits to the host [61].

1.8.2. Applications of Lactic Acid Bacteria to Promote Women's Health

Vaginal microbiome studies have gained attention in recent years due to the unique differences of the vaginal microbiome as compared to other human microbiomes. Unlike other human microbiomes, the vaginal microbiome is mainly composed of the *Lactobacillus* genus. One study found that 73% of its participants had a community dominated by one or two *Lactobacillus* species assigned to over 50% of the sequences [64]. In fact, a lack of lactobacilli is indicative of health issues. *Lactobacillus crispatus, Lb. gasseri,* and *Lb. jensenii* are all vaginal lactobacilli (Table 1.1). It has become evident that *Lactobacillus* vaginal species are uniquely suited for their environment, and while it is a relatively new area of research, several adaptation theories have been proposed. Unlike in the GIT, it is not clear that lactobacilli species can preferentially use the primary carbohydrate source, glycogen, over competing organisms [65]. Instead, LAB of the vaginal microbiome are known for their role in protecting the host from pathogens [38]. In general, the production of lactic acid by LAB vaginal species lowers the pH of the vaginal environment to levels that are inhibitive for competitors [66]. However, there are some species-specific adaptations that also inhibit pathogens (see below).

1.8.3. Genomic Adaptations to the Consumer

As mentioned above, bifidobacteria are early colonizers of the infant gut. They gain competitive advantage by being one of the first microorganisms present, however infant gut *Bifidobacterium* have also uniquely adapted to the available nutrients in order to gain a greater

advantage. Specifically, bifidobacteria utilize the niche-specific nutrients before competing microorganisms can. In general, gut LAB have adapted to use the readily available oligosaccharides found in the GIT [67]. *Bifidobacterium* have specifically adapted to utilize human milk oligosaccharides (HMO)—the oligosaccharides found in mothers' milk [68]. They achieve this through a catabolism gene, such as a lacto-*N*-biosidase, to breakdown HMOs in order to access their sugars (Figure 1.2) [69]. By being the first colonizers and by utilizing the available nutrient source first, *Bifidobacterium* become dominant members of the infant gut microbiome.

As mentioned above, lactobacilli are the dominant members of healthy vaginal microbiomes. While there is still debate over how *Lactobacillus* species come to dominate and interact with the host, it is widely held that these species are important in preventing the growth of pathogens. While the production of lactic acid has been shown to be an inhibitive, further research into vaginal species has discovered other possible mechanisms of action. For instance, a comparative genomic study identified several genes of *Lb. crispatus* that are predicted to interfere with the adherence of competitors. One gene, the *Lactobacillus* epithelium adhesion (LEA) gene, is predicted to interfere with or block the binding ability of *Gardnerella vaginalis* -- an indicator species of bacterial vaginosis (BV) (Figure 1.2) [70]. BV is a dysbiosis of the vaginal microbiome and the reason for 50% of the clinical visits by women [71].

1.9. CONCLUSION

Genomic studies have provided the framework for understanding how LAB have integrated themselves into our everyday lives. These studies have been enhanced by the growing microbiome field. As technologies become more widespread across the food supply chain from farm to fork, researchers will be able to elucidate how LAB function in a microbiome setting and additionally how the microbiome affects LAB evolution. The knowledge of how these microorganisms adapt to, interact with, and benefit their environment is essential to better understand how we can fully leverage their unique abilities. A great deal of effort has already been made in understanding how LAB affect human health, specifically as probiotics. This understanding will allow easier selection of potential strains to develop into probiotics. In fact, experts have recently curated a set of probiotic effector molecules to be used as criteria for selection [72]. Expanding on our current understanding of how probiotics interact with the consumer, efforts are underway to capitalize on the use of probiotics by co-opting them in vaccinations and biotherapeutics as an alternative to antibiotics [73, 74]. Similarly, knowledge of what products LAB produce that influence food texture, flavor, and stability will enhance the manufacturing process. Increased knowledge on how LAB function in the food manufacturing process has given rise to a new emerging trend in the food industry: functional foods. Functional foods are foods that provide a benefit beyond nutrients, such as probiotics or bioactive compounds. In place of synthetically derived compounds, there is a push to use the natural LAB found in the food manufacturing process to produce functional properties such as probiotics, vitamins, and EPS [52]. Advances in the consumer and food manufacturing industries will also benefit the agricultural industry. The food industry is facing a deadline in the year 2050 to support the world's population, and traditional methods are either costly, unreliable, or time-

consuming [75, 76]. Knowledge of how to enhance the health and yield of livestock and crops without using antibiotics, hormones, and engineering will have a profound effect. Finally, there is work to expand LAB's effect beyond their natural habitats. There have been suggestions to use LAB as cell factories in industrial chemistry due to their unique abilities and simple genomes [77]. Overall, LAB provide a resource of highly specialized, streamlined genomes for use and manipulation in several industrial settings.

1.10. ACKNOWLEDGEMENTS

We would like to thank the CRISPR lab for support during this project. We also thank Sarah Brandt for graphical assistance.

1.11. REFERENCES

1. Pot, B. et al. (1994) Taxonomy of Lactic Acid Bacteria. In Bacteriocins of Lactic Acid Bacteria: Microbiology, Genetics and Applications (De Vuyst, L. and Vandamme, E.J. eds), pp. 13-90, Springer US.

2. Stiles, M.E. and Holzapfel, W.H. (1997) Lactic acid bacteria of foods and their current taxonomy. International Journal of Food Microbiology 36 (1), 1-29.

3. Conlon, M.A. and Bird, A.R. (2014) The impact of diet and lifestyle on gut microbiota and human health. Nutrients 7 (1), 17-44.

4. Gerritsen, J. et al. (2011) Intestinal microbiota in human health and disease: the impact of probiotics. Genes & Nutrition 6 (3), 209-240.

5. Feng, X.-B. et al. (2016) Role of intestinal flora imbalance in pathogenesis of pouchitis. Asian Pacific Journal of Tropical Medicine 9 (8), 786-790.

6. Zagorec, M. and Champomier-Vergès, M.-C. (2017) *Lactobacillus sakei:* A Starter for Sausage Fermentation, a Protective Culture for Meat Products. Microorganisms 5 (3), 56.

7. Smit, G. et al. (2005) Flavour formation by lactic acid bacteria and biochemical flavour profiling of cheese products. FEMS Microbiology Reviews 29 (3), 591-610.

8. Vignolo, G.M. et al. (2016) Biotechnology of Lactic Acid Bacteria : Novel Applications, Wiley-Blackwell.

9. Plengvidhya, V. et al. (2007) DNA Fingerprinting of Lactic Acid Bacteria in Sauerkraut Fermentations. Applied and Environmental Microbiology 73 (23), 7697.

10. Herich, R. and Levkut, M. (2002) Lactic acid bacteria, probiotics and immune system. VETERINARNI MEDICINA-PRAHA- 47 (6), 169-180.

11. Murthy, K.N. et al. (2013) Lactic acid bacteria (LAB) as plant growth promoting bacteria (PGPB) for the control of wilt of tomato caused by *Ralstonia solanacearum*. Pest Management In Horticultural Ecosystems 18 (1), 60-65.

12. van Kranenburg, R. et al. (2002) Flavour formation from amino acids by lactic acid bacteria: predictions from genome sequence analysis. International Dairy Journal 12 (2), 111-121.

13. Makarova, K. et al. (2006) Comparative genomics of the lactic acid bacteria. Proceedings of the National Academy of Sciences of the United States of America 103 (42), 15611-15616.

14. Salvetti, E. and O'Toole, P.W. (2017) The Genomic Basis of Lactobacilli as Health-Promoting Organisms. Microbiol Spectr 5 (3).

15. Makarova, K.S. and Koonin, E.V. (2007) Evolutionary Genomics of Lactic Acid Bacteria. Journal of Bacteriology 189 (4), 1199.

16. Klaenhammer, T.R. et al. (2005) Genomic features of lactic acid bacteria effecting bioprocessing and health. FEMS Microbiology Reviews 29 (3), 393-409.

17. Borgogna, J.-L.C. and Yeoman, C.J. (2017) Chapter 3 - The Application of Molecular Methods Towards an Understanding of the Role of the Vaginal Microbiome in Health and Disease. In Methods in Microbiology (Harwood, C. ed), pp. 37-91, Academic Press.

18. Martino, M.E. et al. (2016) Nomadic lifestyle of *Lactobacillus plantarum* revealed by comparative genomics of 54 strains isolated from different habitats. Environmental Microbiology 18 (12), 4974-4989.

19. Siezen, R.J. et al. (2008) Genome-Scale Genotype-Phenotype Matching of Two *Lactococcus lactis* Isolates from Plants Identifies Mechanisms of Adaptation to the Plant Niche. Applied and Environmental Microbiology 74 (2), 424.

20. Douillard, F.P. et al. (2013) Comparative Genomic and Functional Analysis of 100 *Lactobacillus rhamnosus* Strains and Their Comparison with Strain GG. PLOS Genetics 9 (8), e1003683.

21. Inglin, R.C. et al. (2018) Clustering of Pan- and Core-genome of *Lactobacillus* provides Novel Evolutionary Insights for Differentiation. BMC Genomics 19 (1), 284.

22. Milani, C. et al. (2016) Genomics of the Genus *Bifidobacterium* Reveals Species-Specific Adaptation to the Glycan-Rich Gut Environment. Applied and Environmental Microbiology 82 (4), 980.

23. Sun, Z. et al. (2014) Phylogenesis and Evolution of Lactic Acid Bacteria. In Lactic Acid Bacteria: Fundamentals and Practice (Zhang, H. and Cai, Y. eds), pp. 1-101, Springer Netherlands.

24. Papadimitriou, K. et al. (2016) Stress Physiology of Lactic Acid Bacteria. Microbiology and Molecular Biology Reviews 80 (3), 837.

25. Ameen, S.M. and Caruso, G. (2017) Lactic Acid and Lactic Acid Bacteria: Current Use and Perspectives in the Food and Beverage Industry. In Lactic Acid in the Food Industry (Ameen, S.M. and Caruso, G. eds), pp. 33-44, Springer International Publishing.

26. Sun, Z. et al. (2015) Expanding the biotechnology potential of lactobacilli through comparative genomics of 213 strains and associated genera. Nature Communications 6, 8322.

27. Claesson, M.J. et al. (2008) *Lactobacillus* phylogenomics – towards a reclassification of the genus. International Journal of Systematic and Evolutionary Microbiology 58 (12), 2945-2954.

28. Duar, R.M. et al. (2017) Lifestyles in transition: evolution and natural history of the genus *Lactobacillus*. FEMS Microbiol Rev 41 (Supp_1), S27-s48.

29. Salvetti, E. et al. (2018) Comparative Genomics of the Genus *Lactobacillus* Reveals Robust Phylogroups That Provide the Basis for Reclassification. Applied and Environmental Microbiology 84 (17), e00993-18.

30. Hill, C. et al. (2014) The International Scientific Association for Probiotics and Prebiotics consensus statement on the scope and appropriate use of the term probiotic. Nature Reviews Gastroenterology &Amp; Hepatology 11, 506.

31. Cai, Y. et al. (2014) Application of Lactic Acid Bacteria for Animal Production. In Lactic Acid Bacteria: Fundamentals and Practice (Zhang, H. and Cai, Y. eds), pp. 443-491, Springer Netherlands.

32. O' Donnell, M.M. et al. (2015) *Lactobacillus ruminis* strains cluster according to their mammalian gut source. BMC Microbiology 15 (1), 80.

33. Chichlowski, M. et al. (2007) Metabolic and physiological impact of probiotics or direct-fed-microbials on poultry: a brief review of current knowledge. Int J Poult Sci 6 (10), 694-704.

34. Parvin, S. et al. (2010) Effects of inoculation with lactic acid bacteria on the bacterial communities of Italian ryegrass, whole crop maize, guinea grass and rhodes grass silages. Animal Feed Science and Technology 160 (3), 160-166.

35. Oliveira, P.M. et al. (2014) Cereal fungal infection, mycotoxins, and lactic acid bacteria mediated bioprotection: From crop farming to cereal products. Food Microbiology 37, 78-95.

36. Filannino, P. et al. (2018) Metabolic and functional paths of lactic acid bacteria in plant foods: get out of the labyrinth. Current Opinion in Biotechnology 49, 64-72.

37. Di Cagno, R. et al. (2013) Exploitation of vegetables and fruits through lactic acid fermentation. Food Microbiology 33 (1), 1-10.

38. Teneva-Angelova, T. et al. (2018) Chapter 4 - Lactic Acid Bacteria—From Nature Through Food to Health. In Advances in Biotechnology for Food Industry (Holban, A.M. and Grumezescu, A.M. eds), pp. 91-133, Academic Press.

39. Siezen, R.J. and van Hylckama Vlieg, J.E. (2011) Genomic diversity and versatility of *Lactobacillus plantarum*, a natural metabolic engineer. Microbial Cell Factories 10 (1), S3.

40. Broadbent, J.R. et al. (2012) Analysis of the *Lactobacillus casei* supragenome and its influence in species evolution and lifestyle adaptation. BMC Genomics 13 (1), 533.

41. Filannino, P. et al. (2014) Hydroxycinnamic Acids Used as External Acceptors of Electrons: an Energetic Advantage for Strictly Heterofermentative Lactic Acid Bacteria. Applied and Environmental Microbiology 80 (24), 7574.

42. Lamont, J.R. et al. (2017) From yogurt to yield: Potential applications of lactic acid bacteria in plant production. Soil Biology and Biochemistry 111, 1-9.

43. Klostermann, K. et al. (2008) Intramammary infusion of a live culture of *Lactococcus lactis* for treatment of bovine mastitis: comparison with antibiotic treatment in field trials. Journal of Dairy Research 75 (3), 365-373.

44. Crispie, F. et al. (2008) Intramammary infusion of a live culture for treatment of bovine mastitis: effect of live lactococci on the mammary immune response. Journal of Dairy Research 75 (3), 374-384.

45. Beecher, C. et al. (2009) Administration of a live culture of *Lactococcus lactis* DPC 3147 into the bovine mammary gland stimulates the local host immune response, particularly IL-1β and IL-8 gene expression. Journal of Dairy Research 76 (3), 340-348.

46. Roselló, G. et al. (2013) Biological control of fire blight of apple and pear with antagonistic *Lactobacillus plantarum*. European Journal of Plant Pathology 137 (3), 621-633.

47. Papadimitriou, K. et al. (2015) How microbes adapt to a diversity of food niches. Current Opinion in Food Science 2, 29-35.

48. Liu, S.-n. et al. (2011) Lactic acid bacteria in traditional fermented Chinese foods. Food Research International 44 (3), 643-651.

49. Johanningsmeier, S.D. et al. (2012) Influence of Sodium Chloride, pH, and Lactic Acid Bacteria on Anaerobic Lactic Acid Utilization during Fermented Cucumber Spoilage. Journal of Food Science 77 (7), M397-M404.

50. Matthews, K.R. et al. (2017) Food Microbiology: An Introduction, Fourth Edition, American Society of Microbiology.

51. Mende, S. et al. (2016) Influence of exopolysaccharides on the structure, texture, stability and sensory properties of yoghurt and related products. International Dairy Journal 52, 57-71.

52. Linares, D.M. et al. (2017) Lactic Acid Bacteria and Bifidobacteria with Potential to Design Natural Biofunctional Health-Promoting Dairy Foods. Frontiers in Microbiology 8 (846).

53. Sieuwerts, S. et al. (2010) Mixed-Culture Transcriptome Analysis Reveals the Molecular Basis of Mixed-Culture Growth in *Streptococcus thermophilu*s and *Lactobacillus bulgaricus*. Applied and Environmental Microbiology 76 (23), 7775.

54. Ammor, M.S. and Mayo, B. (2007) Selection criteria for lactic acid bacteria to be used as functional starter cultures in dry sausage production: An update. Meat Science 76 (1), 138-146.

55. Chavan, R.S. and Chavan, S.R. (2011) Sourdough Technology—A Traditional Way for Wholesome Foods: A Review. Comprehensive Reviews in Food Science and Food Safety 10 (3), 169-182.

56. Ruas-Madiedo, P. et al. (2002) Role of exopolysaccharides produced by *Lactococcus lactis subsp. cremoris* on the viscosity of fermented milks. International Dairy Journal 12 (8), 689-695.

57. Barrangou, R. et al. (2007) CRISPR provides acquired resistance against viruses in prokaryotes. Science 315 (5819), 1709-12.

58. LaManna, C.M. and Barrangou, R. (2018) Enabling the Rise of a CRISPR World. The CRISPR Journal 1 (3), 205-208.

59. Quercia, S. et al. (2014) From lifetime to evolution: timescales of human gut microbiota adaptation. Frontiers in microbiology 5, 587-587.

60. Mikelsaar, M. et al. (2016) Biodiversity of Intestinal Lactic Acid Bacteria in the Healthy Population. In Advances in Microbiology, Infectious Diseases and Public Health: Volume 4 (Donelli, G. ed), pp. 1-64, Springer International Publishing.

61. Kwok, L.-y. (2014) Lactic Acid Bacteria and the Human Gastrointestinal Tract. In Lactic Acid Bacteria: Fundamentals and Practice (Zhang, H. and Cai, Y. eds), pp. 375-441, Springer Netherlands.

62. Milani, C. et al. (2015) Exploring Vertical Transmission of Bifidobacteria from Mother to Child. Applied and Environmental Microbiology 81 (20), 7078.

63. Sarojini, S. (2018) Chapter 1 - Gut Microbes: The Miniscule Laborers in the Human Body. In Diet, Microbiome and Health (Holban, A.M. and Grumezescu, A.M. eds), pp. 1-31, Academic Press.

64. Ravel, J. et al. (2011) Vaginal microbiome of reproductive-age women. Proceedings of the National Academy of Sciences 108 (Supplement 1), 4680.

65. Vaneechoutte, M. (2017) The human vaginal microbial community. Research in Microbiology 168 (9), 811-825.

66. Tachedjian, G. et al. (2017) The role of lactic acid production by probiotic *Lactobacillus* species in vaginal health. Research in Microbiology 168 (9), 782-792.

67. Barrangou, R. et al. (2003) Functional and comparative genomic analyses of an operon involved in fructooligosaccharide utilization by *Lactobacillus acidophilus*. Proceedings of the National Academy of Sciences 100 (15), 8957.

68. Sánchez, B. et al. (2013) Adaptation of bifidobacteria to the gastrointestinal tract and functional consequences. Pharmacological Research 69 (1), 127-136.

69. Shani, G.I. et al. (2018) Chapter 9 - Interactions Between Bifidobacteria, Milk Oligosaccharides, and Neonate Hosts. In The Bifidobacteria and Related Organisms (Mattarelli, P. et al. eds), pp. 165-175, Academic Press.

70. Ojala, T. et al. (2014) Comparative genomics of *Lactobacillus crispatus* suggests novel mechanisms for the competitive exclusion of Gardnerella vaginalis. BMC Genomics 15 (1), 1070.

71. Huang, B. et al. (2014) The Changing Landscape of the Vaginal Microbiome. Clinics in laboratory medicine 34 (4), 747-761.

72. Lebeer, S. et al. (2018) Identification of probiotic effector molecules: present state and future perspectives. Current Opinion in Biotechnology 49, 217-223.

73. Flaherty, S. and Klaenhammer, T.R. (2016) Multivalent Chromosomal Expression of the *Clostridium botulinum* Serotype A Neurotoxin Heavy-Chain Antigen and the *Bacillus anthracis* Protective Antigen in *Lactobacillus acidophilus*. Applied and Environmental Microbiology 82 (20), 6091.

74. Klotz, C. and Barrangou, R. (2018) Engineering Components of the *Lactobacillus* S-Layer for Biotherapeutic Applications. Frontiers in microbiology 9, 2264-2264.

75. Acquaah, G. (2009) Principles of plant genetics and breeding, John Wiley & Sons.

76. Telugu, B.P. et al. (2017) Genome editing and genetic engineering in livestock for advancing agricultural and biomedical applications. Mamm Genome 28 (7-8), 338-347.

77. Sauer, M. et al. (2017) The Efficient Clade: Lactic Acid Bacteria for Industrial Chemical Production. Trends in Biotechnology 35 (8), 756-769.

78. Johnson, B. et al. (2013) Identification of extracellular surface-layer associated proteins in *Lactobacillus acidophilus* NCFM. Microbiology 159 (11), 2269-2282.

79. Selle, K. and Klaenhammer, T.R. (2013) Genomic and phenotypic evidence for probiotic influences of *Lactobacillus gasseri* on human health. FEMS Microbiology Reviews 37 (6), 915-935.

80. Gorbach, S. et al. (2017) Chapter 7 - *Lactobacillus rhamnosus* GG. In The Microbiota in Gastrointestinal Pathophysiology (Floch, M.H. et al. eds), pp. 79-88, Academic Press.

81. Suda, Y. et al. (2014) Immunobiotic *Lactobacillus jensenii* as immune-health promoting factor to improve growth performance and productivity in post-weaning pigs. BMC Immunology 15 (1), 24.

82. Wu, R. et al. (2011) Proteomic analysis of responses of a new probiotic bacterium *Lactobacillus casei* Zhang to low acid stress. International Journal of Food Microbiology 147 (3), 181-187.

83. Wu, C. et al. (2012) *Lactobacillus casei* combats acid stress by maintaining cell membrane functionality. Journal of Industrial Microbiology & Biotechnology 39 (7), 1031-1039.

84. Vrancken, G. et al. (2008) Kinetic analysis of growth and sugar consumption by *Lactobacillus fermentum* IMDO 130101 reveals adaptation to the acidic sourdough ecosystem. International Journal of Food Microbiology 128 (1), 58-66.

85. Messaoudi, S. et al. (2013) *Lactobacillus salivarius:* Bacteriocin and probiotic activity. Food Microbiology 36 (2), 296-304.

86. Candela, M. et al. (2010) DnaK from *Bifidobacterium animalis subsp. lactis* is a surface-exposed human plasminogen receptor upregulated in response to bile salts. Microbiology 156 (6), 1609-1618.

87. Connell Motherway, M. et al. (2011) Functional genome analysis of *Bifidobacterium breve* UCC2003 reveals type IVb tight adherence (Tad) pili as an essential and conserved host-colonization factor. Proceedings of the National Academy of Sciences 108 (27), 11217.

88. Hwanhlem, N. et al. (2017) Inhibition of food-spoilage and foodborne pathogenic bacteria by a nisin Z-producing *Lactococcus lactis subsp. lactis* KT2W2L. LWT - Food Science and Technology 82, 170-175.

Table 1.1 | Genomic characteristic of select LAB strains. Select set of 16 LAB species and subspecies used in this study.

Genus	Species	Subspecies	No. Genomes	Complete Genomes	Reference Strain	Reference Genome	Genome Size (Mb)	GC %	Genes	Plasmids	Functional Attribute	Market	Ref
Lactobacillus	*acidophilus*		38	7	NCFM	NC_006814	1.99	34.7	1927		Surface Layer Proteins	Consumer (GIT)	[78]
	crispatus		61	1	ST1	NC_014106	2.04	36.9	2036		Pathogen exclusion	Consumer (Vaginal)	[70]
	gasseri		31	3	ATCC 3323	NC_008530	1.89	35.3	1891		Degradation of Oxalate	Consumer (GIT), Consumer (Vaginal)	[79]
	rhamnosus		153	17	GG	NC_013198	3.01	46.7	3062		Surface Molecules (Pili, EPS, etc.)	Consumer (GIT)	[80]
	jensenii		18	1	ASM193623v1	NZ_CP018809	1.67	34.4	1626		Immunobiotic Properties	Agriculture (Animals), Consumer (Vaginal)	[81]
	casei		23	6	BL23	NC_010999	3.08	46.3	3236		Acid stress resistance	Consumer (GIT), Food (Manufacturing)	[82, 83]
	fermentum		56	16	IFO 3956	NC_010610	2.1	51.5	2134		Mannitol production	Food (Manufacturing)	[84]
	plantarum		316	76	WCFS1	NC_004567	3.31	44.5	3124	3	Antimicrobial properties	Agriculture (Plants)	[46]
	delbrueckii	*bulgaricus*	57	21	ATCC 11842	NC_008054	1.87	49.7	1961		Peptide Synthesis	Food (Manufacturing)	[53]
	salivarius		84	8	UCC118	NC_007929	1.83	32.9	1864	3	Bacteriocins	Consumer (GIT), Food (Manufacturing)	[85]
Streptococcus	*thermophilus*		53	26	JIM 8232	NC_017581	1.93	38.9	2033		Phage Protection	Food (Manufacturing)	[53, 57]
Bifidobacterium	*animalis*	*lactis*	54	22	DSM 10140	NC_012815	1.94	60.5	1655		Bile Acid Resistance	Consumer (GIT)	[86]
	breve		95	40	DSM 20213	NZ_AP012324	2.27	58.9	2039		Surface Molecules (Pili)	Consumer (GIT)	[87]
	longum		185	20	NCC2705	NC_004307	2.26	60.1	1797	2	HMO utilization	Consumer(GIT)	[68]
Lactococcus	*lactis*	*lactis*	72	21	Il 1403	NC_002662	2.37	35.3	2406		Antimicrobial properties	Consumer (GIT), Food (Manufacturing), Agriculture (Plants), Agriculture (Livestock)	[19, 43, 44, 88]
	lactis	*cremoris*	40	13	158	NZ_CP015894	2.25	35.9	2255	7	EPS	Food (Manufacturing)	[56]

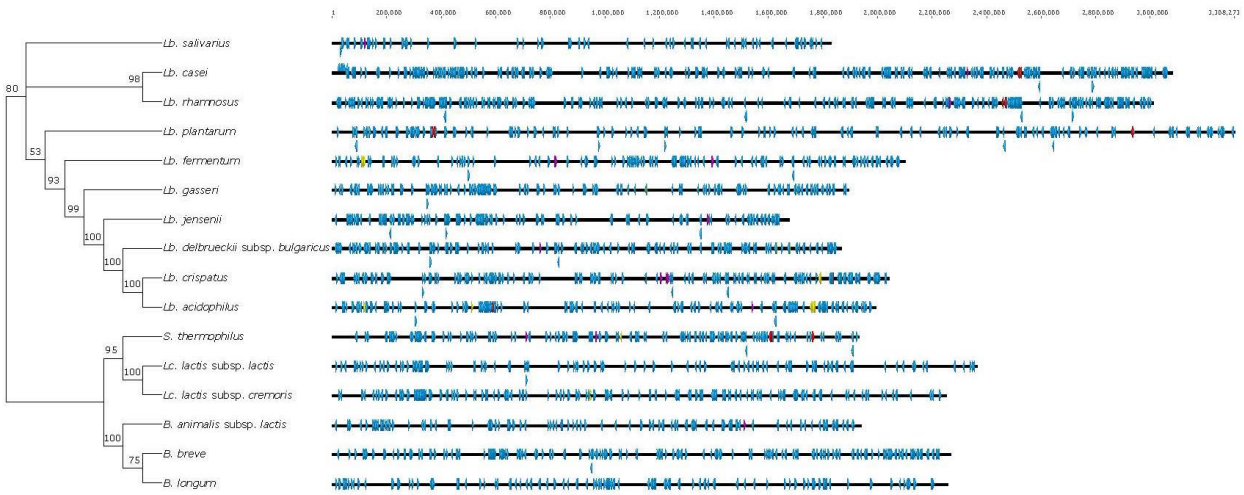


Figure 1.1 | Phylogenetic and Genomic Comparison of LAB Strains. (Left) A phylogenetic tree of 16 LAB strains from Table 1.1. Tree was generated using RAxML and is based on the nucleotide alignment of the 16S rRNA sequence. To the right are representations of each genome. Transporter (blue), EPS (yellow), CRISPR (purple), and bacteriocin (red) loci are highlighted.

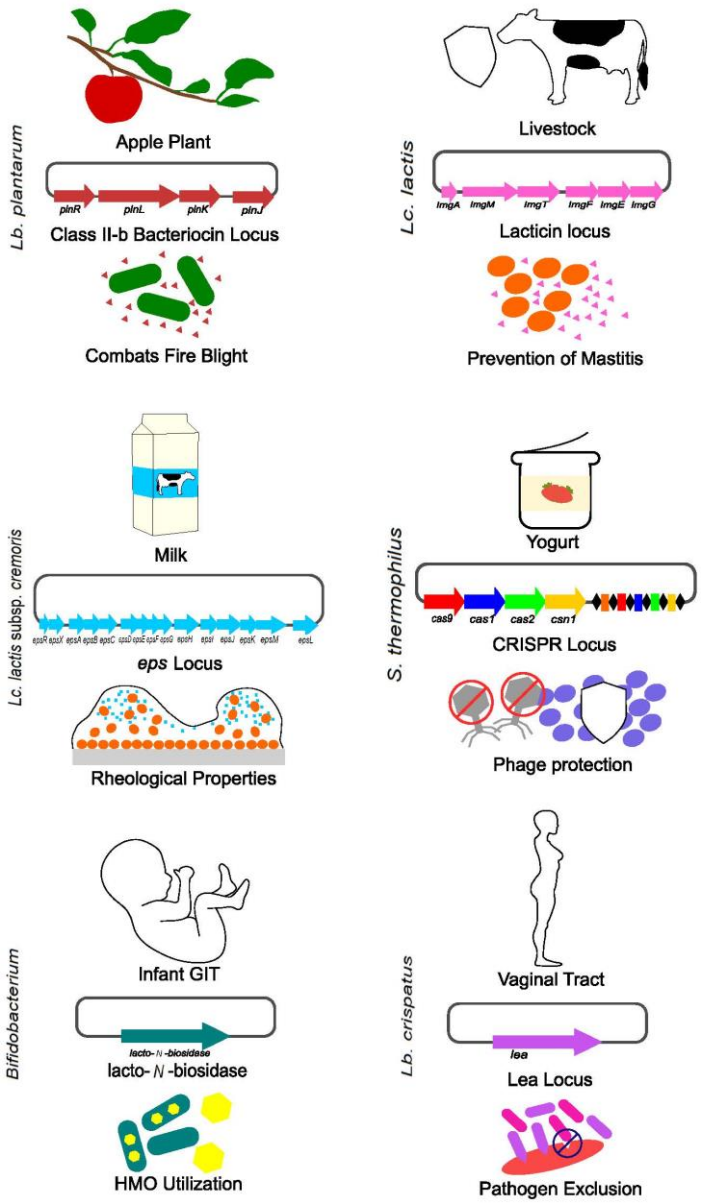


Figure 1.2 | Functional Genomics of LAB. *Lb. plantarum* (top left) contains a bacteriocin that combats firelight in apple plants. *Lc. lactis* (top right) produces lacticin to prevent mastitis in cows. *Lc. lactis* subsp. *cremoris* (middle left) produces EPS for the production of fermented milk products. *S. thermophilus* (middle right) utilizes its CRISPR systems to prevent phage spoilage in yogurt production. *Bifidobacterium* (bottom left) encode lacto-*N*-biosidases for the utilization of HMOs in the infant gut. *Lb. crispatus* (bottom right) uses its LEA protein to inhibit the adhesion of pathogenic *G. vaginalis* in the human vaginal microbiome.

CHAPTER 2: PHYLOGENETIC ANALYSIS OF THE *BIFIDOBACTERIUM* GENUS USING GLYCOLYSIS ENZYME SEQUENCES

2.1. CONTRIBUTION TO WORK

Katelyn Brandt is first author on "Phylogenetic Analysis of the *Bifidobacterium* Genus Using Glycolysis Enzyme Sequences" published in *Frontiers in Microbiology*. She planned and carried out experiments. She and Rodolphe Barrangou analyzed the data and wrote the manuscript. The following chapter is from Brandt and Barrangou, *Frontiers in Microbiology* 7:657.

2.2. ABSTRACT

Bifidobacteria are important members of the human gastrointestinal tract that promote the establishment of a healthy microbial consortium in the gut of infants. Recent studies have established that the *Bifidobacterium* genus is a polymorphic phylogenetic clade, which encompasses a diversity of species and subspecies that encode a broad range of proteins implicated in complex and non-digestible carbohydrate uptake and catabolism, ranging from human breast milk oligosaccharides, to plant fibers. Recent genomic studies have created a need to properly place *Bifidobacterium* species in a phylogenetic tree. Current approaches, based on core-genome analyses come at the cost of intensive sequencing and demanding analytical processes. Here, we propose a typing method based on sequences of glycolysis genes and the proteins they encode, to provide insights into diversity, typing, and phylogeny in this complex and broad genus. We show that glycolysis genes occur broadly in these genomes, to encode the machinery necessary for the biochemical spine of the cell, and provide a robust phylogenetic marker. Furthermore, glycolytic sequences-based trees are congruent with both the classical 16S rRNA phylogeny, and core genome-based strain clustering. Furthermore, these glycolysis markers can also be used to provide insights into the adaptive evolution of this genus, especially with regards to trends towards a high GC content. This streamlined method may open new avenues for phylogenetic studies on a broad scale, given the widespread occurrence of the glycolysis pathway in bacteria, and the diversity of the sequences they encode.

2.3. INTRODUCTION

Bifidobacterium species are an important component of the human gastrointestinal tract (GIT) microbiome, and exert critical functional roles, especially during the establishment of gut microbial composition early in life. Consequently, they are the subject of extensive microbiological and genetics studies, to investigate their probiotic phenotypes, and genotypes, respectively. Actually, many studies are investigating the genetic basis for their health-promoting functionalities, both in industry and academia. This genus is often found in the GIT of animals (Ventura et al., 2014), and is the predominant phylogenetic group early in human life (Turroni et al., 2012a). Indeed, a mounting body of evidence has established vertical transmission between the mother and infants (Milani et al., 2015), notably through the selective nurture of bifidobacteria through diverse non-digestible human-milk oligosaccharides (HMOs) that are a critical component of breast milk (Sela, 2011). These HMOs selectively drive the colonization of the infantile GIT by species that encode prebiotic transporters and hydrolases (Turroni et al., 2012b). Recently, a dichotomy has been established between healthy term babies with a normal gut microbiome, and preterm infants whom have not been colonized by *Bifidobacterium* species (Arboleya et al., 2015). Several studies have implicated the expansive carbohydrate uptake and catabolism gene repertoire of bifidobacteria as the key driver of adaptation of this genus to the infant diet (Milani et al., 2014). In fact, several species of bifidobacteria have shown unique genome composition adaptation trajectories in their carbohydrate utilization machinery, rendering them competitive in this environment (Pokusaeva et al., 2011, Ventura et al., 2012).

To better understand how these organisms have emerged as potent early-life colonizers, there has been a surge in genome sequencing in recent years. At the time of writing, 47 established species and subspecies have been sequenced (Milani et al., 2016), providing a wealth

of genomic information, which serves as a valuable tool for understanding the species and strain diversity within this polymorphic genus, as well as unraveling the key elements that drive health-promoting and colonization phenotypes in humans. However, given the democratization of sequencing technologies in general, and genome and microbiome sequencing in particular, it is imperative that tools and methods be available to analyze this high-throughput data, and specifically allow experimentalists to parse out the complex phylogeny of this broad genus. Indeed, basic questions being addressed regarding the occurrence, diversity and functions of various *Bifidobacterium* species in the human GIT will require the ability to accurately and consistently assign phylogeny.

Fundamentally, as new sequences become available, it is important to know where to place strains on the phylogenetic tree of *Bifidobacterium.* Whereas the affordability, accessibility and ability to generate high-throughput data have become somewhat straightforward, a key challenge lies in the analysis of these sequences, regarding assembly, comparative analyses and phylogenetic assignments. Historically, 16S rRNA sequences have been used across the phylogenetics field for classification and sequence tree-based assignments, but there are growing concerns about the adequacy and sustainability of this method (Fox et al., 1992), notably with regards to the availability of proper references (Clarridge, 2004), and the actual levels of conservation of sequences targeted by "universal" primers (Baker et al., 2003). Because of this, new approaches have been suggested, ranging from multi-locus approaches, using housekeeping genes (Eisen, 1995), to core-genome analyses (Medini et al., 2005). For *Bifidobacterium*, efforts have been focused on creating a phylogeny based on whole and/or conserved genomic sequences, namely the pan-genome and the core-genome, respectively (Lukjancenko et al., 2011, Lugli et al., 2014). While the core-genome is arguably comprehensive, core-genome assembly is

time consuming and computationally intense. Alternative methods need to be developed, to allow rapid and convenient phylogenetic screening of new and potentially unknown sequences. Preferably, such a method would provide high resolution, low-throughput, robust, accurate, and affordable information.

Notwithstanding phenotypic diversity between organisms that have specialized metabolic pathway combinations, and the corresponding genomic complement, there are core biochemical pathways and processes that are broadly distributed across the Tree of Life. Noteworthy, glycolysis is a fundamental process for most cells, and may be construed as the biochemical backbone of most, if not all, living organisms. Indeed, this process allows for the genesis of energy through the catabolism of simple carbohydrates. This pathway is, at least partially, present in all genomes (Fothergill-Gilmore and Michels, 1993) and consequently constitutes a promising biochemical, and thus genetic, marker for phylogenetic studies. Because these genes are important, they are typically members of the house-keeping genomic set, and are widely dispersed across the Tree of Life. However, they are likely subject to less selective pressure than other phylogenetic markers (i.e. ribosomal sequences), and thus afford a more diverse set of sequences to encompass a broad range of assorted sequences (Fothergill-Gilmore, 1986). Therefore, we set out to assess the potential of glycolytic genes, and the sequences of the proteins they encode, for bifidobacteria phylogenetic studies. In particular, we determined the occurrence and diversity of these glycolytic enzyme genes in the genomes of bifidobacteria, and compared and contrasted sequence alignment-based trees with one another, and to those derived from alternative sequences, notably the core-genome, and the 16S rRNA-based reference tree. Our results show how the glycolysis protein sequences can be used as suitable markers to create a phylogeny of *Bifidobacterium* that is as accurate as the core-genome based phylogeny, but

much less computationally demanding. We also explore how basic features of the genetic sequences of glycolysis can reveal trends and patterns of evolution among the different *Bifidobacterium* species and the genus as a whole.

2.4. MATERIALS AND METHODS

2.4.1. Genetic Sequences Sampling and Reference Genomes

We used sequences derived from a total of 48 *Bifidobacterium* genomes from distinct species and subspecies, as listed in Table 2.1. *Bifidobacterium stercoris* was included in this analysis, as a separate species, but it was recently renamed as a strain of *Bifidobacterium adolescentis* (Killer et al., 2013). Our results (see below) show that *B. stercoris* is always a close neighbor of *B. adolescentis*, consistent with the newest findings. These genomes were mined for the presence of glycolytic enzymes using Geneious version 9.0.5(Kearse et al., 2012). We selectively elected to pursue a scheme based on canonical glycolysis genes, as to generate a broadly applicable method. Nevertheless, the classical glycolysis genes do not universally occur in bacterial genomes. Furthermore, some organisms do carry alternative pathways, such as the bifid shunt in *Bifidobacterium,* which could prove valuable, but are not widely distributed. The nine canonical glycolysis enzymes from bifidobacteria (de Vries and Stouthamer, 1967) were found in each genome. Four reference species (*Bifidobacterium longum* subsp. *longum*, *B. adolescentis*, *Bifidobacterium animalis sub. lactis*, and *Bifidobacterium breve*) were used to make a database of the nine genes. The Annotate from Database feature was used (with 40% nucleotide sequence similarity cut-off) to identify glycolytic orthologs in the other genomes. As all genomes had been previously annotated, we confirmed the original annotation to the database annotation manually to validate this method of mining. In cases where multiple hits were obtained, BLAST (Altschul et al., 1990) analyses were carried out to select the correct homolog. Translated sequences were confirmed using ExPasy (Gasteiger et al., 2003). For the 16S rRNA analysis, the 16S rRNA sequences were extracted manually from each genome. In case of multiple hits, BLAST analyses were carried out to select the right sequences. For increased

robustness, the glycolysis enzyme sequences were concatenated in order of occurrence in the glycolysis pathway. (Lang et al., 2013).

2.4.2. Genesis of Sequence Alignment-Based Trees

Five different alignments were made for each tree using Geneious version 9.0.5. ClustalW (Larkin et al., 2007) was used, with the BLOSUM scoring matrix, and settings of gap creation at -10 cost, and gap extension at -0.1 cost per element. For the 16S rRNA alignment, ClustalW was set so that the cost matrix was IUB, with a gap opening penalty of 15, and gap extension cost of 6.66. MUSCLE (Edgar, 2004) was used with the setting of 8 maximum number of iterations for the amino acid sequences and the 16S rRNA alignments. The Geneious Pairwise Alignment was set so that the alignment type was global alignment with free end gaps and the cost matrix was BLOSUM62 for the amino acid sequences. For the 16S rRNA gene analysis, the alignment type was global alignment with free end gaps and a cost matrix of 65% similarity (5.0/-4.0). MAFFT(Katoh et al., 2002) was used twice, for both the amino acid sequences and the 16S rRNA sequences. For the amino acid sequences the first alignment had an algorithm setting of auto, a scoring matrix of BLOSUM62, a gap open penalty of 1.53, and an offset value of 0.123. The second alignment had an algorithm setting of auto, a scoring matrix of BLOSUM80, a gap open penalty of 1.53, and an offset value of 0.123. For the first 16S rRNA alignment, the algorithm was set to auto, the scoring matrix was set to 100PAM/k=2, the gap open penalty was set to 1.53, and the offset value was set to 0.123. The second alignment for the 16S rRNA was set so that the algorithm was auto, the scoring matrix was 200PAM/k=2, the gap open penalty was 1.53, and the offset value was 0.123. trimAl (Capella-Gutiérrez et al., 2009) was used to select a consistent alignment between the five alignments. The parameters were

compareset and automated1. Using Geneious, trees were made from the respective consistent alignments. The trees were generated using RaxML version 7.2.8 (Stamatakis, 2006b, Stamatakis, 2014). For the protein based trees the parameters were set so that the model was CAT (Lartillot and Philippe, 2004) BLOSUM62, the algorithm was Bootstrap using rapid hill climbing with random seed 1, and the number of bootstrap replicates was 100 (Stamatakis, 2006a). For the 16S rRNA tree, the nucleotide model was GTR CAT, the algorithm was Bootstrap using rapid hill climbing with random seed 1, and the number of bootstrap replicates was 100. A consensus tree was then built using the consensus builder in Geneious, at a 50% support threshold. The consensus tree was used in all further analyses. The sums of branch lengths for each tree were found by adding the branch lengths together in Mega6 (Tamura et al., 2013).

2.4.3. Statistical Analyses

All statistical analyses were carried out using R version 3.2.2 (R Core Team, 2015). This software was also used to generate plots, graphs and display quantitative data throughout the manuscript.

2.5. RESULTS

2.5.1. Glycolytic Enzyme Sequence-Based Phylogeny

Bifidobacteria contain nine of the 10 traditional enzymes (Figure 2.1) commonly found in the glycolysis pathway (de Vries and Stouthamer, 1967). Phylogenetic analyses were carried out using the amino acid sequences of the proteins encoded by the aforementioned glycolysis genes. A comprehensive tree based on sequence alignment of the concatenated sequences of the glycolytic enzymes found in *Bifidobacterium* is shown in Figure 2.2. Six separate phylogenetic groups were identified, as previously established from the core-genome (Milani et al., 2016). These groups are: the *B. longum* group (orange), the *B. adolescentis* group (green), the *Bifidobacterium pseudolongum* group (purple), the *Bifidobacterium pollurom* group (blue-green), the *Bifidobacterium boum* group (blue), and the *Bifidobacterium asteroides* group (red) (Bottacini et al., 2014). The number of individuals in each group varied between 3 and 11, with the *B. longum* group being the most diverse. *Bifidobacterium angulatum* and *Bifidobacterium merycicum* were moved to the *B. adolescentis* group due to a high bootstrap value in the concatenated tree. The concatenated tree has bootstrap values that range from 52 to 100. We observe a total of 34 bootstrap values of 70 and above (Figure 2.S1). Trees based on sequence alignments of the individual enzymes of glycolysis can be found in Figures 2.S2 to 2.S10. Interestingly, all of the individual trees resolved the phylogenetic groups found in the core-genome with only the Gap and Eno trees providing alternative locations for a few branches, notably *Bifidobacterium magnum, Bifidobacterium gallicum* and *Bifidobacterium thermacidophilum sub. thermacidophilum*. Table 2.2 shows the sum of branch lengths for each tree. The 16S rRNA tree has the largest sum at 204.99, while the concatenated tree had the smallest sum at 99.56. The consistent clustering into these six phylogenetic trees illustrates how

robust and valuable the glycolytic sequences are with regards to phylogenetic information. It also shows that this method is congruent with the core-genome.

2.5.2. 16S rRNA-Based Reference Phylogeny

A reference phylogeny was generated using the 16S rRNA sequences of each of the 48 species and sub-species included in this study (Figure 2.3). The six phylogenetic groups are identified and colored the same as in the concatenated tree. We elected to assign the *B. angulatum* and *B. merycicum* from the *B. longum* group to the *B. adolescentis* group, consistent with the concatenated tree. Noteworthy, the tree has bootstrap values that range from 51 to 100, with 17 nodes at values of 70 and above, which is half the amount found in the concatenated tree (Figure 2.S1). With regards to size, we point out that the concatenated tree is based on overall sequences ranging between 3,205 amino acids and 3,479 amino acids, which quantitatively compares as approximately twice the amount to the 16S rRNA ~1,600nt range, in terms of input-information amounts.

2.5.3. Genome-Wide Analyses

The overall genome sizes in this study ranged from 1.73Mb for *Bifidobacterium indicum* to 3.26Mb for *Bifidobacterium biavatii,* with an average of 2.28Mb and a median of 2.17Mb. The GC content ranged from 52.8% for *Bifidobacterium tsurumiense* to 65.5% for *Bifidobacterium choerinum*, with an average of 60.4% and a median of 60.2%. This substantiates the perception that bifidobacteria are generally categorized as high-GC content organisms, at the genome-wide level (Ventura et al., 2007). However, a thorough analysis of GC content across the phylogenetic groups revealed that even among these high-GC organisms there are three

distinct subsets of high, medium, and low-GC bifidobacteria (Figure 2.4a). Most of the species fall in the upper medium-GC range, with the low-GC range being the least populated. There are some noteworthy groupings between the phylogenetic groups, specifically the *B. pullorum* and the *B. boum* groups, for which the entire groups are packed tightly in the high GC region and the medium GC region, respectively. All of the other groups, except the *B. longum* group, span two of these subsets. For the *B. longum* group, *Bifidobacterium saguini* lies just at the border between the low and medium GC subsets. This group has the largest spread, consistent with being the most diverse in the concatenated and 16S rRNA trees.

Next, we looked at how the GC content varied across the trees. Figure 2.4b shows boxplots of the GC content of each tree and the total GC content. Except for the 16S rRNA and *tpi* trees, all other trees had median GC values with strong evidence of being higher than the median total GC content (Chambers, 1983). Looking on an individual basis, over half of the genomes have 16S rRNA and *tpi* GC values below their total GC, while the other genes are either above or close to their total GC (Figure 2.5). Again, the *B. pullorum* and *B. boum* groups are tightly packed in regards to their GC spread amongst their glycolysis genes, 16S rRNA, and total GC. In contrast, the *B.longum* group has the largest spread, a parallel to its higher diversity in the phylogenetic trees.

2.6. DISCUSSION

Bifidobacterium is a diverse genus of human intestinal beneficial microbes that provide health-promoting functionalities, as illustrated by their broad use as probiotics in foods and dietary supplements (Turroni et al., 2014). Recently, extensive genomic analyses of diverse species, subspecies and phylogenetic groups have provided insights into their adaptation to the human gut, notably with regards to their ability to colonize the intestinal cavity in general, and utilize non-digestible carbohydrates in particular (Milani et al., 2016). Studies investigating the use of human breast milk oligosaccharides illustrate the important contribution of these probiotics in establishing the human gut microbiome at the early stages of life (Sela, 2011). Yet, these studies also reveal that there are many distinct and diverse *Bifidobacterium* species and phylogenetic groups that colonize the human GIT, perhaps with idiosyncratic genomic attributes, and their corresponding functionalities (Chaplin et al., 2015). These organisms have specifically adapted to their environment to competitively utilize available nutrients (Sánchez et al., 2013). In the human gut, these consist of non-digestible complex oligosaccharides that are not adsorbed, nor broken down in the upper GIT. Whereas plant-based fibers are important in the adult diet, human milk oligosaccharides are important components of the infant diet. Furthermore, *Bifidobacterium* have even been successful in helping each other through cross-feeding (Turroni et al., 2015). Thus, we addressed the need to establish practical means to allocate phylogeny with minimalistic information based on sequences that encode glycolysis, the biochemical spine of most cells.

Here, we have shown that a multigene approach using glycolysis sequences can be used to uncover genomic trends and to make an accurate phylogenetic tree, based on a relatively small amount of information. The concatenated glycolysis tree in Figure 2.2 is congruent with both the

16S rRNA tree and the established core-genome-based tree (Milani et al., 2016). The only notable exception is the placement of *B. merycicum* and *B. angulatum*. However, the relocation was between two neighboring phylogenetic groups in the concatenated and core-genome based trees. The glycolysis pathway is perhaps as, if not more, robust and accurate than the 16S rRNA tree. Compared to the 16S rRNA, the bootstrap values of the concatenated tree were higher on average. This leads to more confidence in the placement of species and the identification of phylogenetic groups, which in comparison, can appear arbitrarily located on the 16S rRNA. The concatenated tree is able to identify groups as well as the core-genome based tree. In fact, all of the phylogenetic groups from the core-genome were consistently found across the glycolytic pathway based trees. However, the glycolysis-based trees have the advantage of being much less labor intensive than the core-genome approach. This allows for accurate phylogenetic mapping of new strains or species, possibly encompassing unknown species, in less time and with less data than a core-genome. This approach is high resolution, low throughput, affordable, and accurate. Part of the success of this approach is the universality of glycolysis. Glycolysis is the biochemical backbone of the cell, and as such all organisms have at least some part of the glycolysis pathway represented (Fothergill-Gilmore and Michels, 1993). Even though these are slower-evolving genes, the changes that are made are enough to make an accurate phylogeny (Fothergill-Gilmore, 1986), evidenced from the congruence between our trees and the core-genome based tree. Even though the glycolysis enzymes are considered "slow evolvers", our data shows they are evolving at different rates amongst themselves. This can be explained by the fact that the glycolysis pathway is adapted by organisms to best fit their own unique environment and requirements (Bar-Even et al., 2012), as seen here in the *Bifidobacterium* and their bifid shunt (Sela et al., 2010). Some of the genes have specialized secondary functions, such as

enolase acting as a cell surface receptor in *Bifidobacterium* (Candela et al., 2009*).* All of this makes the glycolysis pathway an excellent phylogenetic marker candidate. The various rates in evolution and moonlighting abilities also allow for further applications in recognizing adaptive trends.

The functional diversity of bifidobacteria is underpinned by multi-dimensional variety in their genomes, including overall content, organization, sequence diversity, and others. In extreme cases, even a two-fold difference in genome size can be observed. Despite being generally perceived as high GC organisms, they vary enough to have distinct relative classes of high, middle, and low–GC, amongst themselves (Figure 2.4a). Yet, there are non-random patterns and phenomena that drive these differences. The phylogenetic groups are clustered in specific regions of the GC continuum. Some groups are more tightly packed than others. A general trend that is observed across the genus is an evolutionary movement towards a high(er) GC content. The higher end of the spectrum is more densely populated then the lower end of the spectrum, indicative of an upward trend. This is reflected by the increased GC content in the individual glycolysis genes, when compared to the total GC content. Of the glycolysis genes, only one, *tpi*, does not show strong evidence for being different from the genome-wide (total) GC content. Critically, all of the other genes are above the total GC content. When we combine the overall genomic data with the GC-content groupings and trends discovered using glycolysis as phylogenetic markers, we posit the hypothesis that, over time, the GC content within the genomes of bifidobacteria increases, as to deviate further away from the 50% value, as the organisms adapt, and their genomes evolve accordingly.

Because of the broad occurrence of the glycolysis pathway in the Tree of Life, it is a suitable candidate marker to use in phylogenetic studies, likely beyond its application in

bifidobacteria. In addition to being conserved genes that capture genetic diversity, glycolysis genes are consistently amongst the most highly expressed in not only *Bifidobacterium* (Turroni et al., 2015), but other organisms as well (Barrangou et al., 2006). This reflects both the importance of these sequences genetically (as illustrated by GC content drift), and functionally (as illustrated by their propensity for high levels of constitutive transcription). Because of this, it may be possible to correlate transcriptional data to phylogenetic studies on a broader scale. From here, it could be feasible to assign species and map data to known references using transcriptomic, genomic, or meta-data. Indeed, as the democratization of metagenomic technologies continues, and the need to assign phylogenetic information to partial genomic information increases, we propose that this method be used to provide insights into the phylogeny of un-assigned contigs. Overall, this approach allows for accurate phylogenetic mapping, congruent with a core-genome and more robust than the 16S rRNA phylogenetic approach, as well as inference on genomic adaptation, using either genomic, transcriptomic, or meta-data in a timely fashion and with minimal computation.

2.7. ACKNOWLEDGEMENTS

We would like to thank the Dr. Todd Klaenhammer lab and the CRISPR lab for providing insights and support during this project.

2.8. REFERENCES

Altschul, S. F., W. Gish, W. Miller, E. W. Myers and D. J. Lipman (1990). "Basic local alignment search tool." J Mol Biol **215**(3): 403-410 DOI: 10.1016/s0022-2836(05)80360-2.

Arboleya, S., B. Sánchez, C. Milani, S. Duranti, G. Solís, N. Fernández, C. G. de los Reyes-Gavilán, M. Ventura, A. Margolles and M. Gueimonde (2015). "Intestinal Microbiota Development in Preterm Neonates and Effect of Perinatal Antibiotics." The Journal of Pediatrics **166**(3): 538-544 DOI: http://dx.doi.org/10.1016/j.jpeds.2014.09.041.

Baker, G. C., J. J. Smith and D. A. Cowan (2003). "Review and re-analysis of domain-specific 16S primers." Journal of Microbiological Methods **55**(3): 541-555 DOI: http://dx.doi.org/10.1016/j.mimet.2003.08.009.

Bar-Even, A., A. Flamholz, E. Noor and R. Milo (2012). "Rethinking glycolysis: on the biochemical logic of metabolic pathways." Nat Chem Biol **8**(6): 509-517.

Barrangou, R., M. A. Azcarate-Peril, T. Duong, S. B. Conners, R. M. Kelly and T. R. Klaenhammer (2006). "Global analysis of carbohydrate utilization by *Lactobacillus acidophilus* using cDNA microarrays." Proceedings of the National Academy of Sciences of the United States of America **103**(10): 3816-3821 DOI: 10.1073/pnas.0511287103.

Bottacini, F., M. Ventura, D. van Sinderen and M. O'Connell Motherway (2014). "Diversity, ecology and intestinal function of bifidobacteria." Microbial Cell Factories **13**(Suppl 1): S4-S4 DOI: 10.1186/1475-2859-13-S1-S4.

Candela, M., E. Biagi, M. Centanni, S. Turroni, M. Vici, F. Musiani, B. Vitali, S. Bergmann, S. Hammerschmidt and P. Brigidi (2009). "Bifidobacterial enolase, a cell surface receptor for human plasminogen involved in the interaction with the host." Microbiology **155**(10): 3294-3303 DOI: doi:10.1099/mic.0.028795-0.

Capella-Gutiérrez, S., J. M. Silla-Martínez and T. Gabaldón (2009). "trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses." Bioinformatics **25**(15): 1972-1973 DOI: 10.1093/bioinformatics/btp348.

Chambers, J. M. (1983). "Notched box plots," in *Graphical Methods for Data Analysis*, (Belmont, CA: Wasworth International Group), 60-63.

Chaplin, A. V., B. A. Efimov, V. V. Smeianov, L. I. Kafarskaia, A. P. Pikina and A. N. Shkoporov (2015). "Intraspecies Genomic Diversity and Long-Term Persistence of *Bifidobacterium longum*." PLoS ONE **10**(8): e0135658 DOI: 10.1371/journal.pone.0135658.

Clarridge, J. E. (2004). "Impact of 16S rRNA Gene Sequence Analysis for Identification of Bacteria on Clinical Microbiology and Infectious Diseases." Clinical Microbiology Reviews **17**(4): 840-862 DOI: 10.1128/CMR.17.4.840-862.2004.

de Vries, W. and A. H. Stouthamer (1967). "Pathway of Glucose Fermentation in Relation to the Taxonomy of Bifidobacteria." Journal of Bacteriology **93**(2): 574-576.

Edgar, R. C. (2004). "MUSCLE: multiple sequence alignment with high accuracy and high throughput." Nucleic Acids Research **32**(5): 1792-1797 DOI: 10.1093/nar/gkh340.

Eisen, J. A. (1995). "The RecA protein as a model molecule for molecular systematic studies of bacteria: Comparison of trees of RecAs and 16S rRNAs from the same species." Journal of Molecular Evolution **41**(6): 1105-1123 DOI: 10.1007/BF00173192.

Fothergill-Gilmore, L. A. (1986). "The evolution of the glycolytic pathway." Trends in Biochemical Sciences **11**(1): 47-51 DOI: http://dx.doi.org/10.1016/0968-0004(86)90233-1.

Fothergill-Gilmore, L. A. and P. A. M. Michels (1993). "Evolution of glycolysis." Progress in Biophysics and Molecular Biology **59**(2): 105-235 DOI: http://dx.doi.org/10.1016/0079-6107(93)90001-Z.

Fox, G. E., J. D. Wisotzkey and P. Jurtshuk (1992). "How Close Is Close: 16S rRNA Sequence Identity May Not Be Sufficient To Guarantee Species Identity." International Journal of Systematic and Evolutionary Microbiology **42**(1): 166-170 DOI: doi:10.1099/00207713-42-1-166.

Gasteiger, E., A. Gattiker, C. Hoogland, I. Ivanyi, R. D. Appel and A. Bairoch (2003). "ExPASy: the proteomics server for in-depth protein knowledge and analysis." Nucleic Acids Research **31**(13): 3784-3788 DOI: 10.1093/nar/gkg563.

Katoh, K., K. Misawa, K. i. Kuma and T. Miyata (2002). "MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform." Nucleic Acids Research **30**(14): 3059-3066 DOI: 10.1093/nar/gkf436.

Kearse, M., R. Moir, A. Wilson, S. Stones-Havas, M. Cheung, S. Sturrock, S. Buxton, A. Cooper, S. Markowitz, C. Duran, T. Thierer, B. Ashton, P. Meintjes and A. Drummond (2012). "Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data." Bioinformatics **28**(12): 1647-1649 DOI: 10.1093/bioinformatics/bts199.

Killer, J., I. Sedláček, V. Rada, J. Havlík and J. Kopečný (2013). "Reclassification of *Bifidobacterium stercoris* Kim et al. 2010 as a later heterotypic synonym of *Bifidobacterium adolescentis*." International Journal of Systematic and Evolutionary Microbiology **63**(11): 4350-4353 DOI: doi:10.1099/ijs.0.054957-0.

Lang, J. M., A. E. Darling and J. A. Eisen (2013). "Phylogeny of Bacterial and Archaeal Genomes Using Conserved Genes: Supertrees and Supermatrices." PLoS ONE **8**(4): e62510 DOI: 10.1371/journal.pone.0062510.

Larkin, M. A., G. Blackshields, N. P. Brown, R. Chenna, P. A. McGettigan, H. McWilliam, F. Valentin, I. M. Wallace, A. Wilm, R. Lopez, J. D. Thompson, T. J. Gibson and D. G.

Higgins (2007). "Clustal W and Clustal X version 2.0." Bioinformatics **23**(21): 2947-2948 DOI: 10.1093/bioinformatics/btm404.

Lartillot, N. and H. Philippe (2004). "A Bayesian Mixture Model for Across-Site Heterogeneities in the Amino-Acid Replacement Process." Molecular Biology and Evolution **21**(6): 1095-1109 DOI: 10.1093/molbev/msh112.

Lugli, G. A., C. Milani, F. Turroni, S. Duranti, C. Ferrario, A. Viappiani, L. Mancabelli, M. Mangifesta, B. Taminiau, V. Delcenserie, D. van Sinderen and M. Ventura (2014). "Investigation of the Evolutionary Development of the Genus *Bifidobacterium* by Comparative Genomics." Applied and Environmental Microbiology **80**(20): 6383-6394 DOI: 10.1128/aem.02004-14.

Lukjancenko, O., D. W. Ussery and T. M. Wassenaar (2011). "Comparative Genomics of *Bifidobacterium, Lactobacillus* and Related Probiotic Genera." Microbial Ecology **63**(3): 651-673 DOI: 10.1007/s00248-011-9948-y.

Medini, D., C. Donati, H. Tettelin, V. Masignani and R. Rappuoli (2005). "The microbial pan-genome." Current Opinion in Genetics & Development **15**(6): 589-594 DOI: http://dx.doi.org/10.1016/j.gde.2005.09.006.

Milani, C., G. A. Lugli, S. Duranti, F. Turroni, F. Bottacini, M. Mangifesta, B. Sanchez, A. Viappiani, L. Mancabelli, B. Taminiau, V. Delcenserie, R. Barrangou, A. Margolles, D. van Sinderen and M. Ventura (2014). "Genomic Encyclopedia of Type Strains of the Genus *Bifidobacterium*." Applied and Environmental Microbiology **80**(20): 6290-6302 DOI: 10.1128/aem.02308-14.

Milani, C., L. Mancabelli, G. A. Lugli, S. Duranti, F. Turroni, C. Ferrario, M. Mangifesta, A. Viappiani, P. Ferretti, V. Gorfer, A. Tett, N. Segata, D. van Sinderen and M. Ventura (2015). "Exploring Vertical Transmission of Bifidobacteria from Mother to Child." Applied and Environmental Microbiology **81**(20): 7078-7087 DOI: 10.1128/aem.02037-15.

Milani, C., F. Turroni, S. Duranti, G. A. Lugli, L. Mancabelli, C. Ferrario, D. van Sinderen and M. Ventura (2016). "Genomics of the Genus *Bifidobacterium* Reveals Species-Specific Adaptation to the Glycan-Rich Gut Environment." Applied and Environmental Microbiology **82**(4): 980-991 DOI: 10.1128/aem.03500-15.

Pokusaeva, K., G. F. Fitzgerald and D. Sinderen (2011). "Carbohydrate metabolism in Bifidobacteria." Genes & Nutrition **6**(3): 285-306 DOI: 10.1007/s12263-010-0206-6.

R Core Team (2015). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Sánchez, B., L. Ruiz, M. Gueimonde, P. Ruas-Madiedo and A. Margolles (2013). "Adaptation of bifidobacteria to the gastrointestinal tract and functional consequences." Pharmacological Research **69**(1): 127-136 DOI: http://dx.doi.org/10.1016/j.phrs.2012.11.004.

Sela, D. A. (2011). "Bifidobacterial utilization of human milk oligosaccharides." International Journal of Food Microbiology **149**(1): 58-64 DOI: http://dx.doi.org/10.1016/j.ijfoodmicro.2011.01.025.

Sela, D. A., N. P. J. Price and D. Mills (2010). Metabolism of Bifidobacteria. Bifidobacteria: Genomics and Molecular Aspects. B. Mayo and D. van Sinderen, Caister Academic Press.

Stamatakis, A. (2006a). Phylogenetic models of rate heterogeneity: a high performance computing perspective. Parallel and Distributed Processing Symposium, 2006. IPDPS 2006. 20th International.

Stamatakis, A. (2006b). "RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models." Bioinformatics **22**(21): 2688-2690 DOI: 10.1093/bioinformatics/btl446.

Stamatakis, A. (2014). "RAxML Version 8: A tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies." Bioinformatics DOI: 10.1093/bioinformatics/btu033.

Tamura, K., G. Stecher, D. Peterson, A. Filipski and S. Kumar (2013). "MEGA6: Molecular Evolutionary Genetics Analysis version 6.0." Molecular Biology and Evolution DOI: 10.1093/molbev/mst197.

Turroni, F., S. Duranti, F. Bottacini, S. Guglielmetti, D. Van Sinderen and M. Ventura (2014). "*Bifidobacterium bifidum* as an example of a specialized human gut commensal." Frontiers in Microbiology **5** DOI: 10.3389/fmicb.2014.00437.

Turroni, F., E. Özcan, C. Milani, L. Mancabelli, A. Viappiani, D. van Sinderen, D. Sela and M. Ventura (2015). "Glycan cross-feeding activities between bifidobacteria under in vitro conditions." Frontiers in Microbiology **6** DOI: 10.3389/fmicb.2015.01030.

Turroni, F., C. Peano, D. A. Pass, E. Foroni, M. Severgnini, M. J. Claesson, C. Kerr, J. Hourihane, D. Murray, F. Fuligni, M. Gueimonde, A. Margolles, G. De Bellis, P. W. O?Toole, D. van Sinderen, J. R. Marchesi and M. Ventura (2012a). "Diversity of Bifidobacteria within the Infant Gut Microbiota." PLoS ONE **7**(5): e36957 DOI: 10.1371/journal.pone.0036957.

Turroni, F., F. Strati, E. Foroni, F. Serafini, S. Duranti, D. van Sinderen and M. Ventura (2012b). "Analysis of Predicted Carbohydrate Transport Systems Encoded by *Bifidobacterium bifidum* PRL2010." Applied and Environmental Microbiology **78**(14): 5002-5012 DOI: 10.1128/AEM.00629-12.

Ventura, M., C. Canchaya, A. Tauch, G. Chandra, G. F. Fitzgerald, K. F. Chater and D. van Sinderen (2007). "Genomics of *Actinobacteria*: Tracing the Evolutionary History of an Ancient Phylum." Microbiology and Molecular Biology Reviews : MMBR **71**(3): 495-548 DOI: 10.1128/MMBR.00005-07.

Ventura, M., F. Turroni, G. A. Lugli and D. van Sinderen (2014). "Bifidobacteria and humans: our special friends, from ecological to genomics perspectives." Journal of the Science of Food and Agriculture **94**(2): 163-168 DOI: 10.1002/jsfa.6356.

Ventura, M., F. Turroni, M. O. C. Motherway, J. MacSharry and D. van Sinderen (2012). "Host–microbe interactions that facilitate gut colonization by commensal bifidobacteria." Trends in Microbiology **20**(10): 467-476 DOI: http://dx.doi.org/10.1016/j.tim.2012.07.002.

Table 2.1 | Species and Genome List. List of the 48 species and subspecies used in this study. Accession numbers and naming conventions included.

Genus	Species	Subspecies	Strain	Accession Number	Naming Convention	Locus Tag
Bifidobacterium	*actinocoloniiforme*		DSM 22766	NZ_CP011786	B_actinocoloniiforme	AB656
Bifidobacterium	*adolescentis*		ATCC 15703	NC_008618	B_adolescentis	BAD
Bifidobacterium	*angulatum*		LMG 11039	NZ_JGYL00000000	B_angulatum	BIANG
Bifidobacterium	*animalis*	*animalis*	ATCC 22527	NC_017834	B_animalis_a	BANAN
Bifidobacterium	*animalis*	*lactis*	DSM 10140	NC_012815	B_animalis_l	BALAT
Bifidobacterium	*asteroides*		PRL 2011	NC_018720	B_asteroides	BAST
Bifidobacterium	*biavatii*		DSM 23969	NZ_JDUU00000000	B_biavatti	OU23
Bifidobacterium	*bifidum*		LMG 13200	NZ_JSEB00000000	B_bifidum	LMG13200
Bifidobacterium	*bohemicum*		DSM 22767	NZ_JDUS00000000	B_bohemicum	OU21
Bifidobacterium	*bombi*		DSM 19703	NZ_JDTS00000000	B_bombi	OT95
Bifidobacterium	*boum*		LMG 10736	NZ_JGYQ00000000	B_boum	BBOU
Bifidobacterium	*breve*		UCC 2003	NC_020517	B_breve	Bbr
Bifidobacterium	*callitrichos*		DSM 23973	NZ_JGYS00000000	B_callitrichos	BCAL
Bifidobacterium	*catenulatum*		JCM 1194	NZ_AP012325	B_catenulatum	BBCT
Bifidobacterium	*choerinum*		LMG 10510	NZ_JGYU00000000	B_choerinum	BCHO
Bifidobacterium	*coryneforme*		LMG 18911	NZ_CP007287	B_coryneforme	BCOR
Bifidobacterium	*crudilactis*		LMG 23609	NZ_JHAL00000000	B_crudilactis	DB51
Bifidobacterium	*cuniculi*		LMG 10738	NZ_JGYV00000000	B_cuniculi	BCUN
Bifidobacterium	*dentium*		Bd1	NC_013714	B_dentium	BDP
Bifidobacterium	*gallicum*		DSM 20093	NZ_ABXB00000000	B_gallicum	BIFGAL
Bifidobacterium	*gallinarum*		LMG 11586	NZ_JGYX00000000	B_gallinarum	BIGA
Bifidobacterium	*indicum*		LMG 11587	NZ_CP006018	B_indicum	BINDI
Bifidobacterium	*kashiwanohense*		JCM 15439	NZ_AP012327	B_kashiwanohense	BBKW
Bifidobacterium	*longum*	*longum*	NCC 2705	NC_004307	B_longum	BL
Bifidobacterium	*longum*	*infantis*	ATCC 15697	NC_011593	B_longum_i	Blon
Bifidobacterium	*longum*	*suis*	LMG 21814	NZ_JGZA00000000	B_longum_s	BLSS
Bifidobacterium	*magnum*		LMG 11591	NZ_JGZB00000000	B_magnum	BMAGN
Bifidobacterium	*merycicum*		LMG 11341	NZ_JGZC00000000	B_merycicum	BMERY
Bifidobacterium	*minimum*		LMG 11592	NZ_JGZD00000000	B_minimum	BMIN
Bifidobacterium	*mongoliense*		DSM 21395	NZ_JGZE00000000	B_mongoliense	BMON
Bifidobacterium	*moukalabense*		DSM 27321	NZ_AZMV00000000	B_moukalabense	BMOU
Bifidobacterium	*pseudocatenulatum*		JCM 1200	NZ_AP012330	B_pseudocatenulatum	BBPC
Bifidobacterium	*pseudolongum*	*globosum*	LMG 11569	NZ_JGZG00000000	B_pseudolongum_g	BPSG
Bifidobacterium	*pseudolongum*	*pseudolongum*	LMG 11571	NZ_JGZH00000000	B_pseudolongum_p	BPSP
Bifidobacterium	*psychraerophilum*		LMG 21775	NZ_JGZI00000000	B_psychraerophilum	BPSY
Bifidobacterium	*pullorum*		LMG 21816	NZ_JGZJ00000000	B_pullorum	BPULL
Bifidobacterium	*reuteri*		DSM 23975	NZ_JGZK00000000	B_reuteri	BREU
Bifidobacterium	*ruminantium*		LMG 21811	NZ_JGZL00000000	B_ruminantium	BRUM
Bifidobacterium	*saeculare*		LMG 14934	NZ_JGZM00000000	B_saeculare	BSAE
Bifidobacterium	*saguini*		DSM 23967	NZ_JGZN00000000	B_saguini	BISA
Bifidobacterium	*scardovii*		LMG 21589	NZ_JGZO00000000	B_scardovii	BSCA
Bifidobacterium	*stellenboschense*		DSM 23968	NZ_JGZP00000000	B_stellenboschense	BSTEL
Bifidobacterium	*stercoris*		DSM 24849	NZ_JGZQ00000000	B_stercoris	BSTER
Bifidobacterium	*subtile*		LMG 11597	NZ_JGZR00000000	B_subtile	BISU
Bifidobacterium	*thermacidophilum*	*porcinum*	LMG 21689	NZ_JGZS00000000	B_thermacidophilum_p	BPORC
Bifidobacterium	*thermacidophilum*	*thermacidophilum*	LMG 21395	NZ_JGZT00000000	B_thermacidophilum_t	THER5
Bifidobacterium	*thermophilum*		JCM 7027	-	B_thermophilum	BTHER
Bifidobacterium	*tsurumiense*		JCM 13495	NZ_JGZU00000000	B_tsurumiense	BITS

Table 2.2 | Sum of Branch Lengths for each tree. Sum of branch lengths for each tree. EC number for each enzyme is also listed.

Gene	E. C. Number	Sum
Phosphoglucomutase (pgm,1)	5.4.2.2	125.03
Glucose-6-phosphota isomerase (pgi,2)	5.3.1.9	153.43
Fructose bisphosphate aldolase (fba, 4)	4.1.2.13	151.76
Triose phosphate isomerase (tpi, 5)	5.3.1.1	170.61
Glyceraldehyde 3-phosphate dehydrogenase (gap, 6)	1.2.1.12	103.07
Phsophoglycerate kinase (pgk, 7)	2.7.2.3	132.41
Phosphoglycerate mutase (gpm, 8)	5.4.2.11	174.7
Enolase (eno, 9)	4.2.1.11	145.06
Pyruvate Kinase (pyk, 10)	2.7.1.40	107.56
Concatenated	-	99.56
16S rRNA	-	204.99

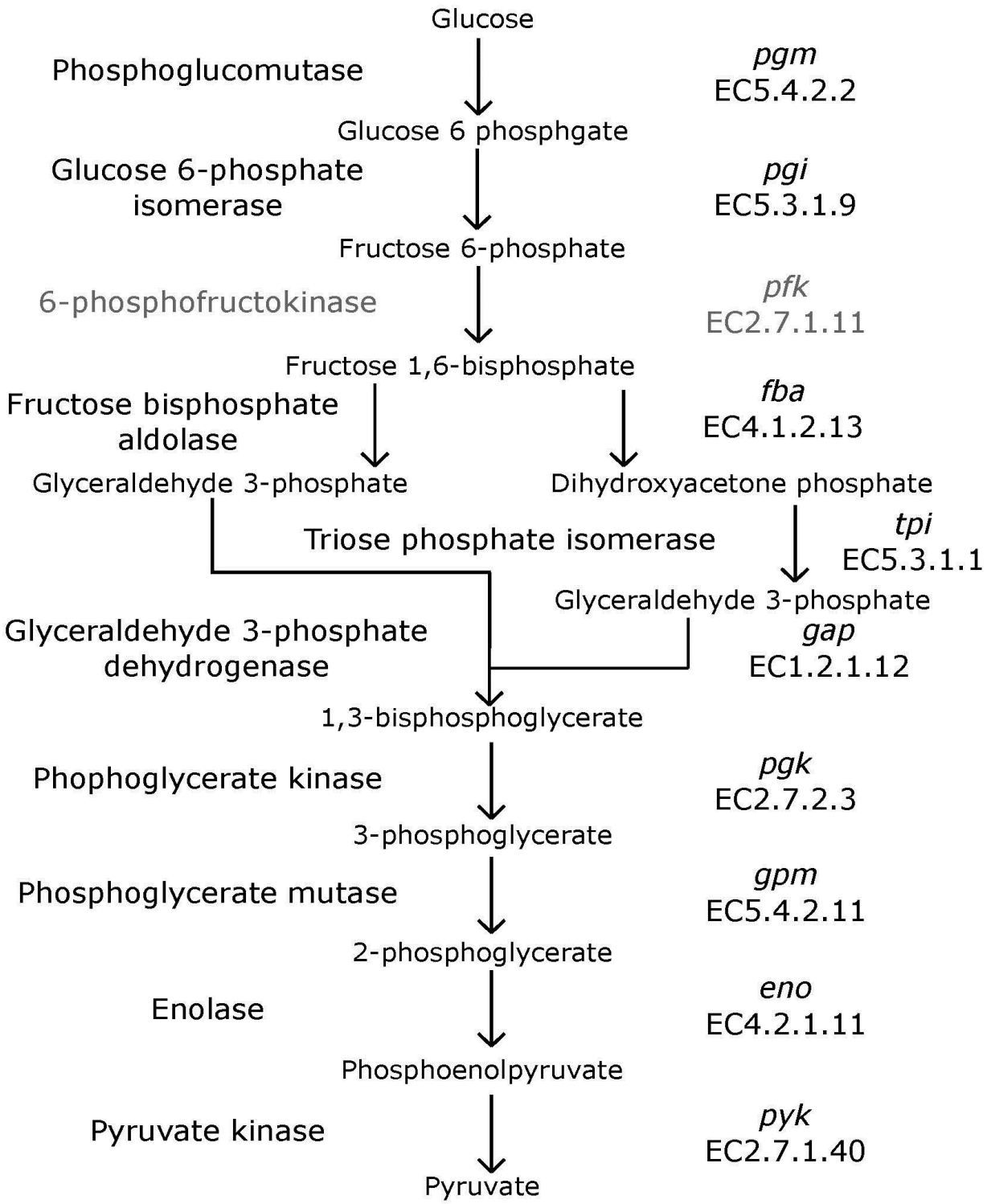


Figure 2.1 | Glycolysis Pathway. Traditional biochemical pathway of glycolysis. Enzyme names are listed to left of arrows, and gene names and EC numbers are shown on the right. 6-phosphofructokinase is faded to represent its absence in Bifidobacterium.

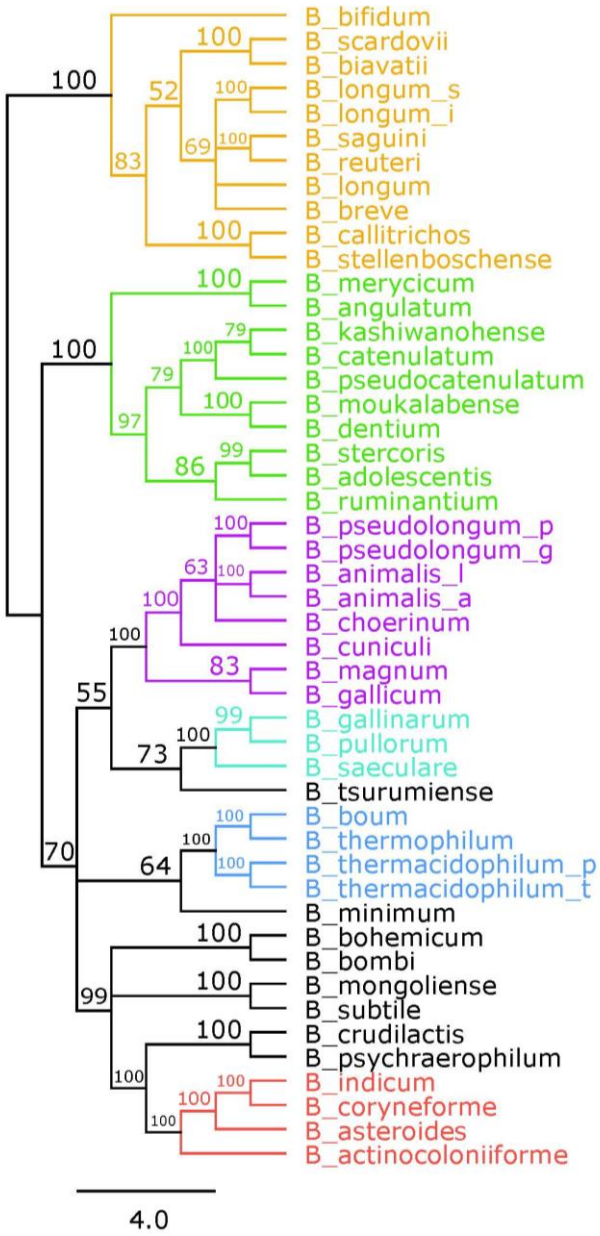


Figure 2.2 | Glycolytic proteins concatenated Tree. Consensus tree based on alignment of the concatenated amino acid sequences of the glycolysis pathway found in Bifidobacterium. Trees were made using RaxML. Bootstrap values are found on each node. Phylogenetic groups are colored as follows: Bifidobacterium longum is orange, Bifidobacterium adolescentis is green, Bifidobacterium psdeudolongum is purple, Bifidobacterium pollorum is blue-green, Bifidobacterium boum is blue, and Bifidobacterium asteroides is red. Species names follow the naming convention from Table 2.1.

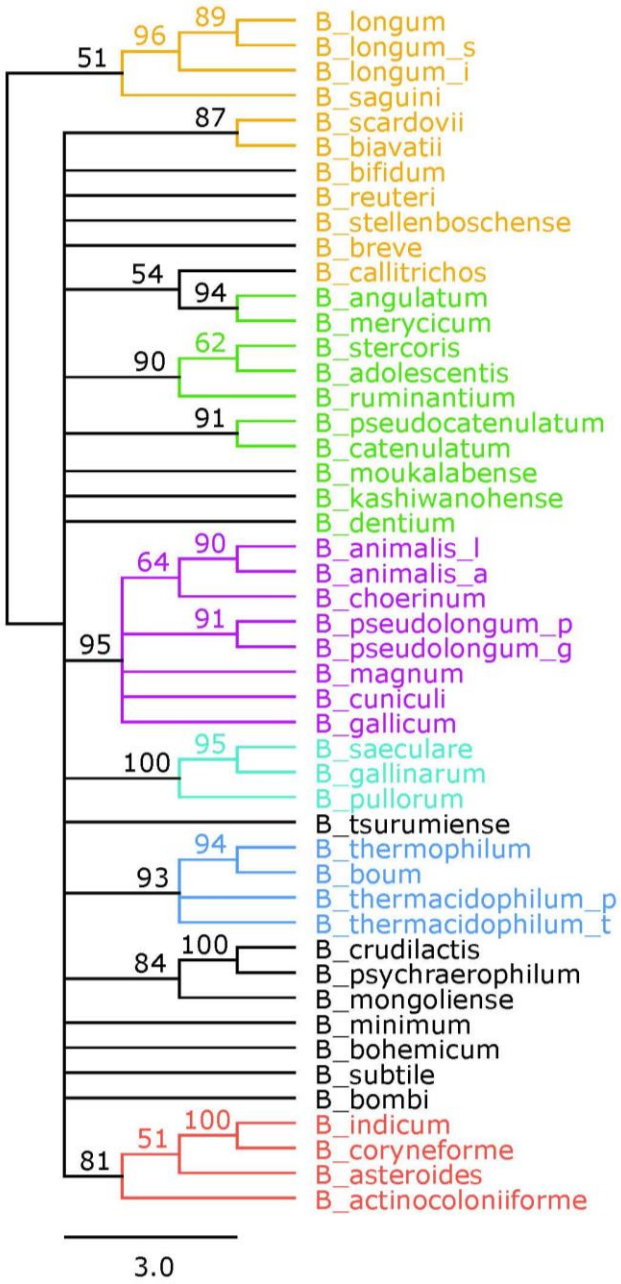


Figure 2.3 | 16S rRNA phylogenetic Tree. Consensus tree based on alignment of the 16S rRNA sequences. Trees were made using RaxML. Bootstrap values are found on each node. Phylogenetic groups are colored as follows: *Bifidobacterium longum* is orange, *Bifidobacterium adolescentis* is green, *Bifidobacterium psdeudolongum* is purple, *Bifidobacterium pollorum* is blue-green, *Bifidobacterium boum* is blue, and *Bifidobacterium asteroides* is red. Species names follow the naming convention from Table 2.1.

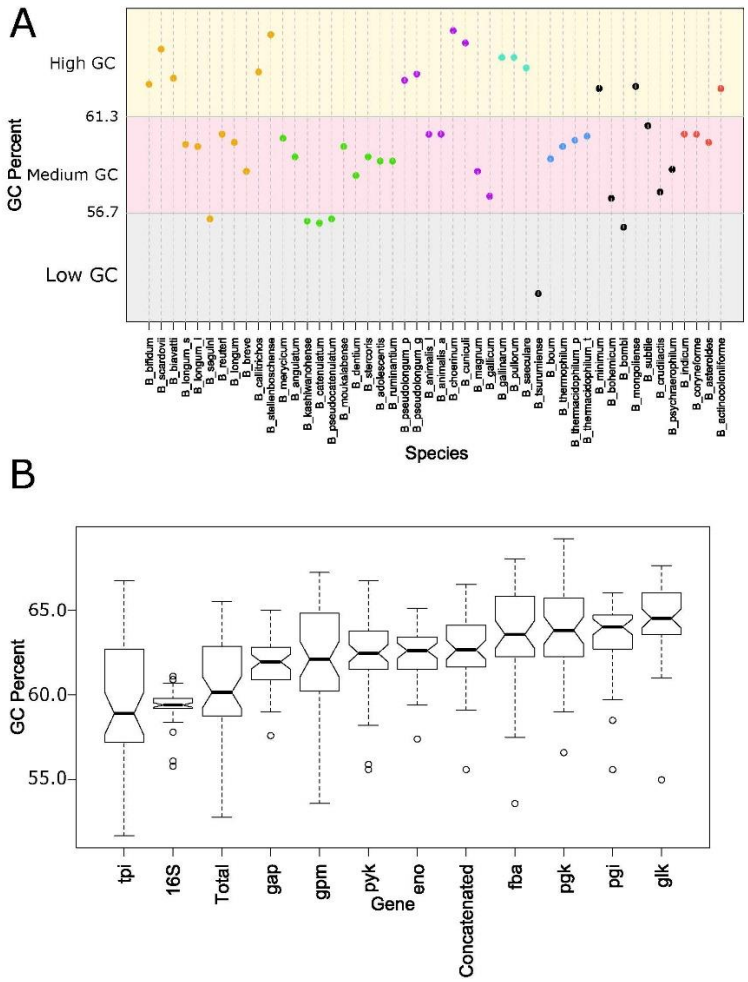


Figure 2.4 | GC content by species and glycolytic genes. (A) shows the total GC content of each species organized by the glycolytic concatenated tree. Spectrum is split into three groups: Low GC from 0.52 to 0.567 (gray), Medium GC from 0.567 to 0.613 (pink), and High GC from 0.613 to 0.66 (yellow). Phylogenetic groups are colored as follows: Bifidobacterium longum is orange, Bifidobacterium adolescentis is green, Bifidobacterium psdeudolongum is purple, Bifidobacterium pollorum is blue-green, Bifidobacterium boum is blue, and Bifidobacterium asteroides red. Species names following the naming convention from Table 2.1. (B) contains notched boxplots of the GC values of each gene and total GC. Boxes are ranked in order of median. Notches that do not overlap are indicative of strong evidence of difference between two medians.

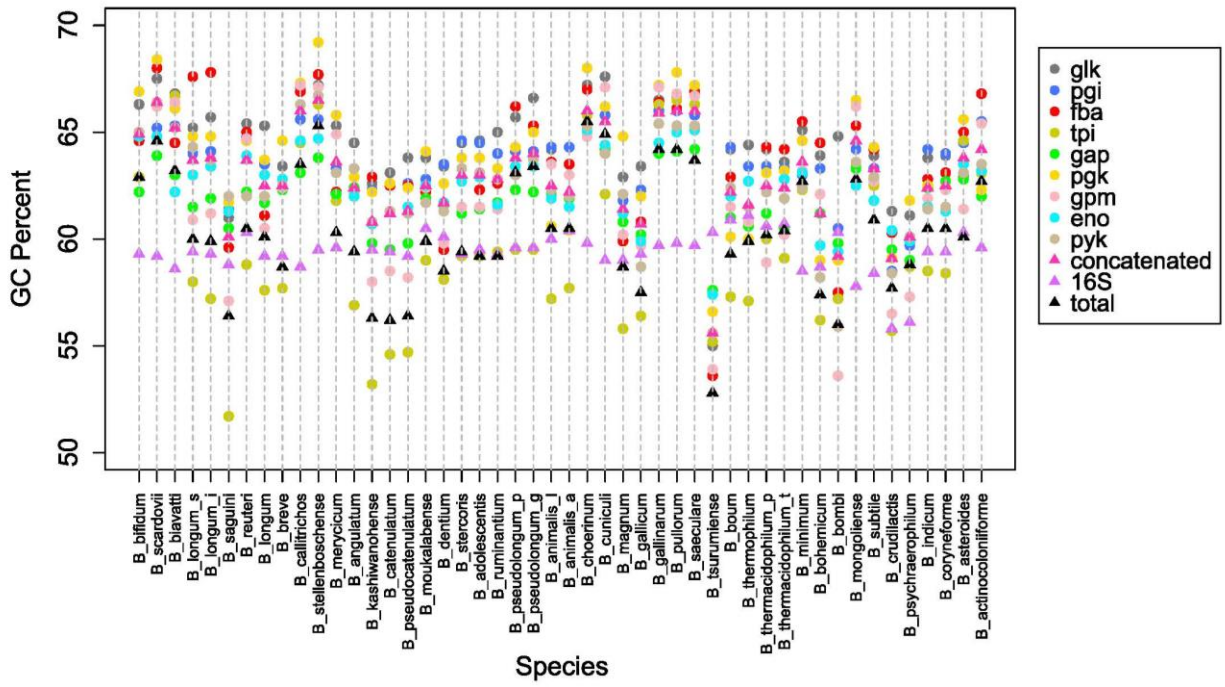


Figure 2.5 | Overall GC content patterns across species. GC percent for each glycolysis gene, 16SrRNA and overall genome, for species listed in Table 2.1.

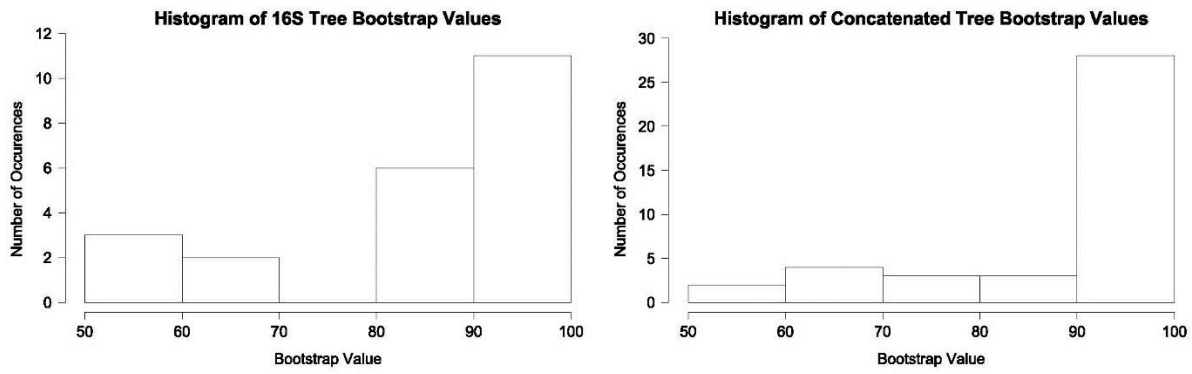


Figure 2.S1 | Histogram of Bootstrap Values. Histograms of bootstrap values from the consensus trees of the 16S phylogenetic tree (right) and the glycolytic proteins concatenated tree (left).

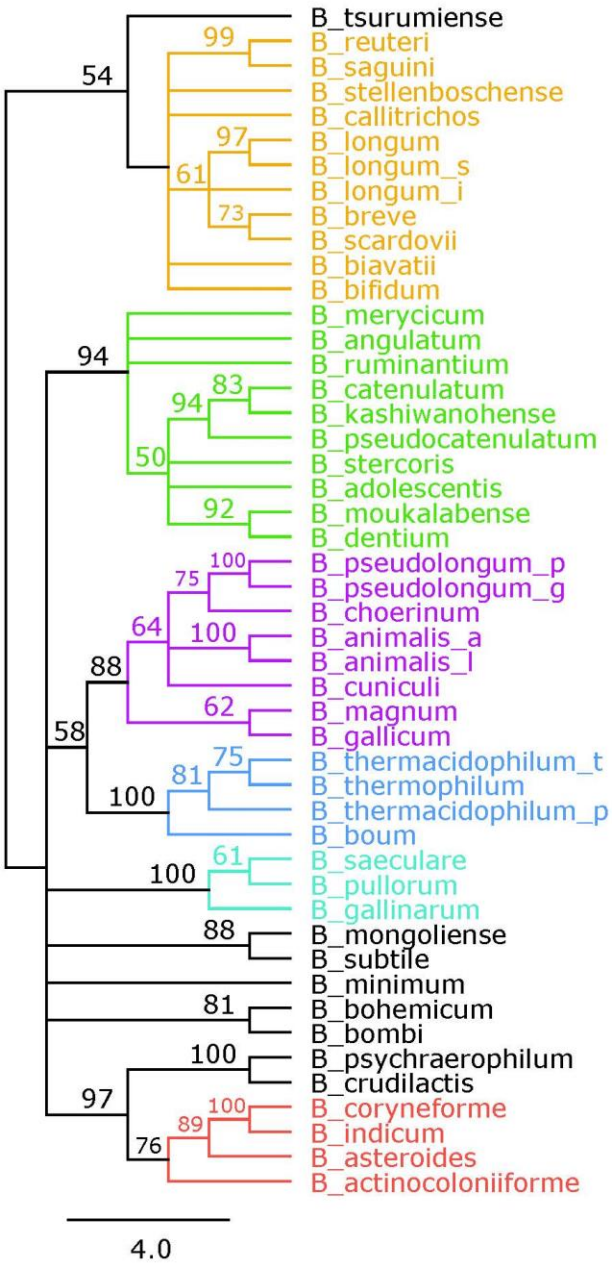


Figure 2.S2 | Pgm Tree. Consensus tree based on alignment of the amino acid sequences of Pgm. Trees were made using RaxML. Bootstrap values are found on each node. Phylogenetic gropus are colored as follows: *Bifidobacterium longum* is orange, *Bifidobacterium adolescentis* is green *Bifidobacterium psdeudolongum* is purple, *Bifidobacterium pullorum* is blue-green, *Bifidobacterium boum* is blue, and *Bifidobacterium asteroides* is red. Species names following the naming convention from Table 2.1.

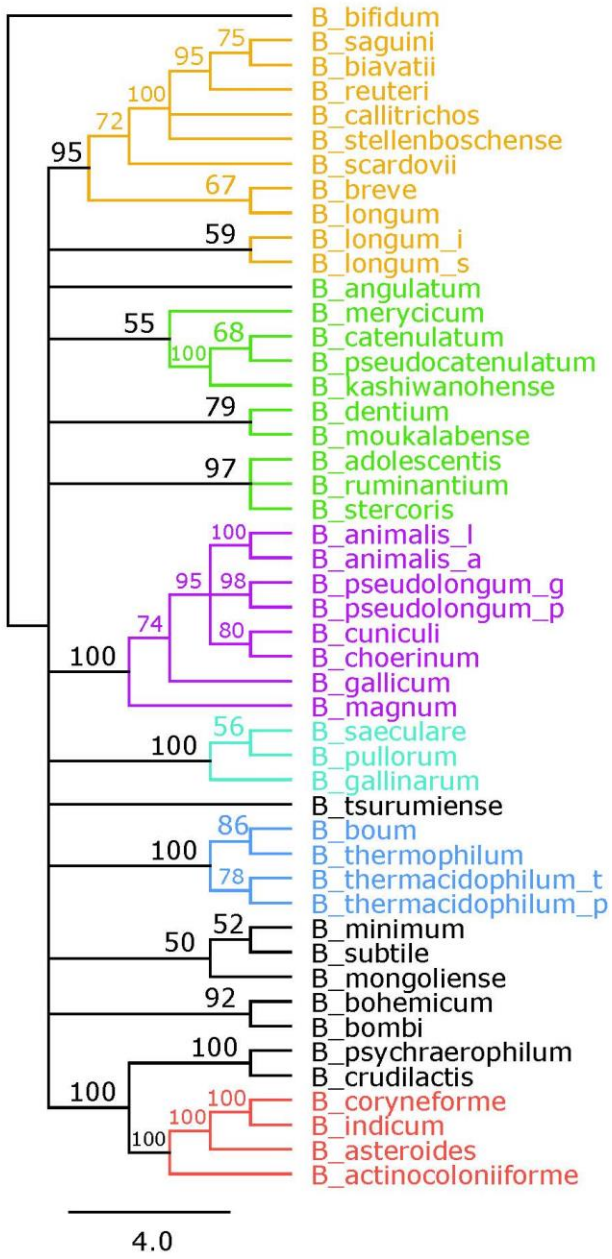


Figure 2.S3 | Pgi Tree. Consensus tree based on alignment of the amino acid sequences of Pgi. Trees were made using RaxML. Bootstrap values are found on each node. Phylogenetic gropus are colored as follows: *Bifidobacterium longum* is orange, *Bifidobacterium adolescentis* is green *Bifidobacterium psdeudolongum* is purple, *Bifidobacterium pullorum* is blue-green, *Bifidobacterium boum* is blue, and *Bifidobacterium asteroides* is red. Species names following the naming convention from Table 2.1.

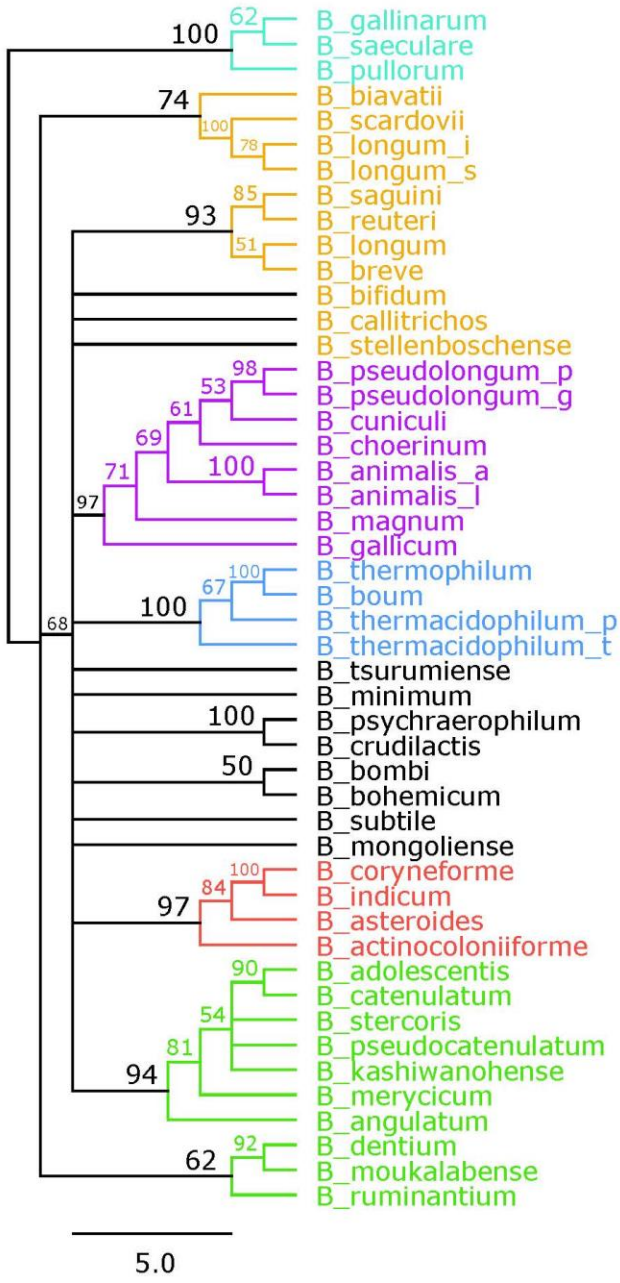


Figure 2.S4 | Fba Tree. Consensus tree based on alignment of the amino acid sequences of Fba. Trees were made using RaxML. Bootstrap values are found on each node. Phylogenetic gropus are colored as follows: *Bifidobacterium longum* is orange, *Bifidobacterium adolescentis* is green *Bifidobacterium psdeudolongum* is purple, *Bifidobacterium pullorum* is blue-green, *Bifidobacterium boum* is blue, and *Bifidobacterium asteroides* is red. Species names following the naming convention from Table 2.1.

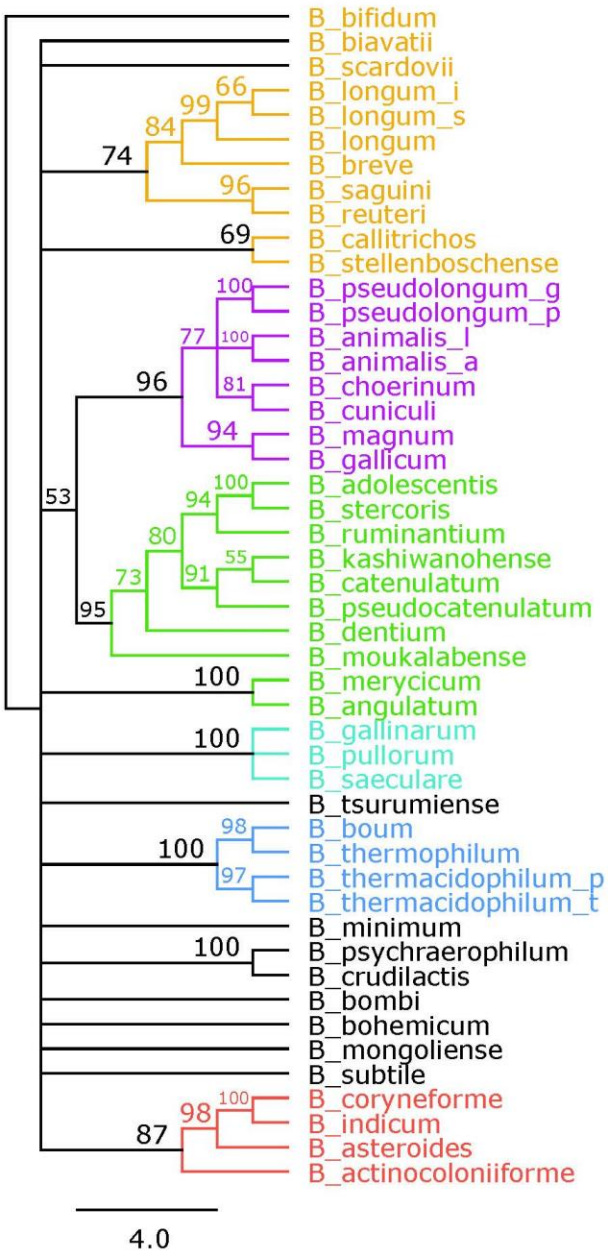


Figure 2.S5 | Tpi Tree. Consensus tree based on alignment of the amino acid sequences of Tpi. Trees were made using RaxML. Bootstrap values are found on each node. Phylogenetic gropus are colored as follows: *Bifidobacterium longum* is orange, *Bifidobacterium adolescentis* is green *Bifidobacterium psdeudolongum* is purple, *Bifidobacterium pullorum* is blue-green, *Bifidobacterium boum* is blue, and *Bifidobacterium asteroides* is red. Species names following the naming convention from Table 2.1.

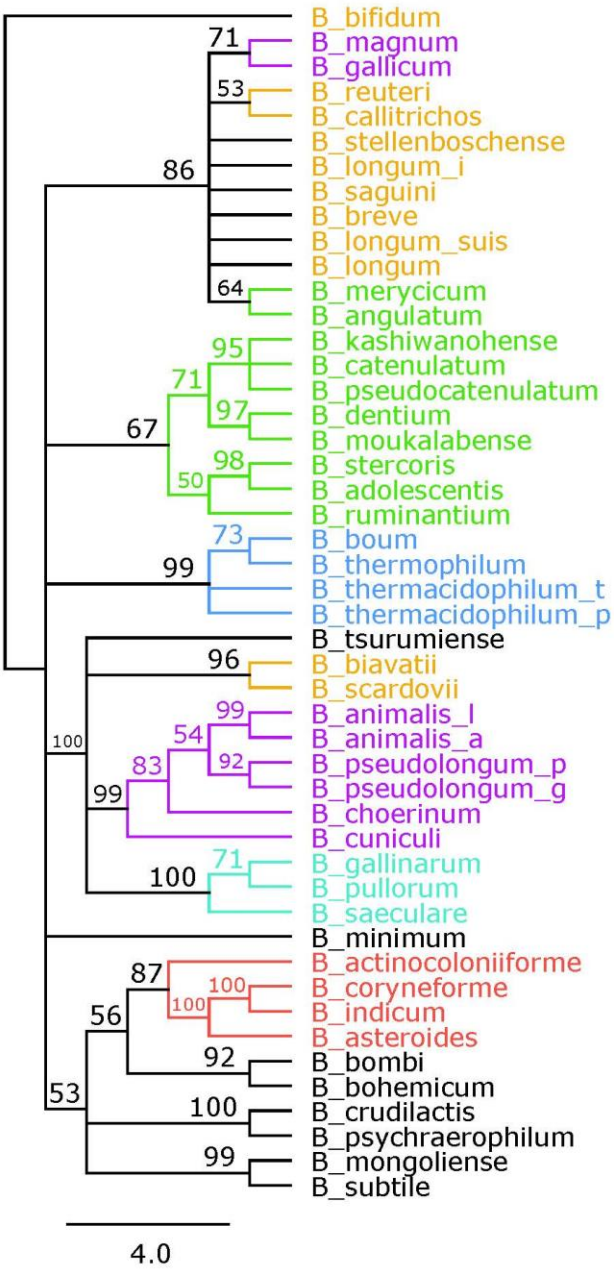


Figure 2.S6 | Gap Tree. Consensus tree based on alignment of the amino acid sequences of Gap. Trees were made using RaxML. Bootstrap values are found on each node. Phylogenetic gropus are colored as follows: *Bifidobacterium longum* is orange, *Bifidobacterium adolescentis* is green *Bifidobacterium psdeudolongum* is purple, *Bifidobacterium pullorum* is blue-green, *Bifidobacterium boum* is blue, and *Bifidobacterium asteroides* is red. Species names following the naming convention from Table 2.1.

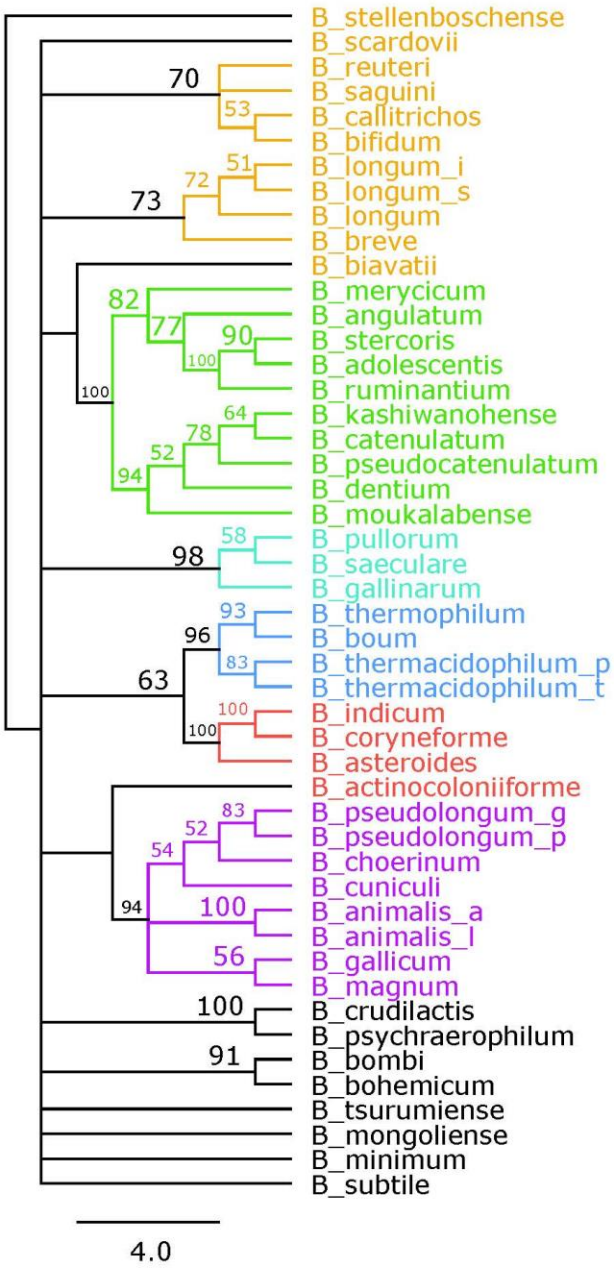


Figure 2.S7 | Pgk Tree. Consensus tree based on alignment of the amino acid sequences of Pgk. Trees were made using RaxML. Bootstrap values are found on each node. Phylogenetic gropus are colored as follows: *Bifidobacterium longum* is orange, *Bifidobacterium adolescentis* is green *Bifidobacterium psdeudolongum* is purple, *Bifidobacterium pullorum* is blue-green, *Bifidobacterium boum* is blue, and *Bifidobacterium asteroides* is red. Species names following the naming convention from Table 2.1.

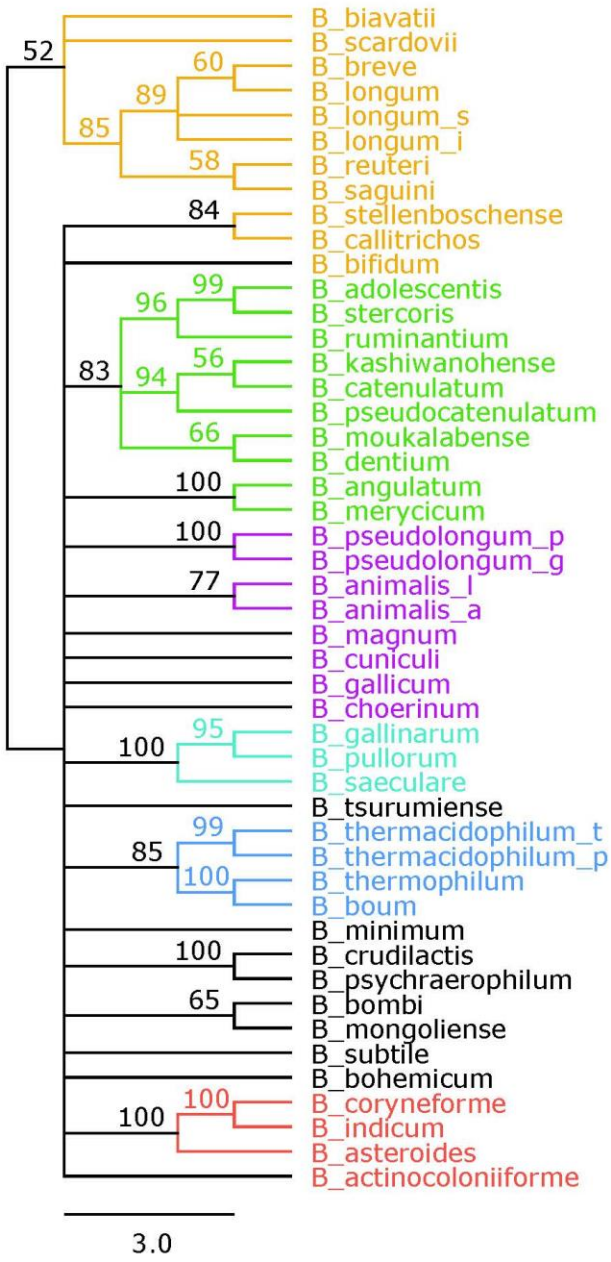


Figure 2.S8 | Gpm Tree. Consensus tree based on alignment of the amino acid sequences of Gpm. Trees were made using RaxML. Bootstrap values are found on each node. Phylogenetic gropus are colored as follows: *Bifidobacterium longum* is orange, *Bifidobacterium adolescentis* is green *Bifidobacterium psdeudolongum* is purple, *Bifidobacterium pullorum* is blue-green, *Bifidobacterium boum* is blue, and *Bifidobacterium asteroides* is red. Species names following the naming convention from Table 2.1.

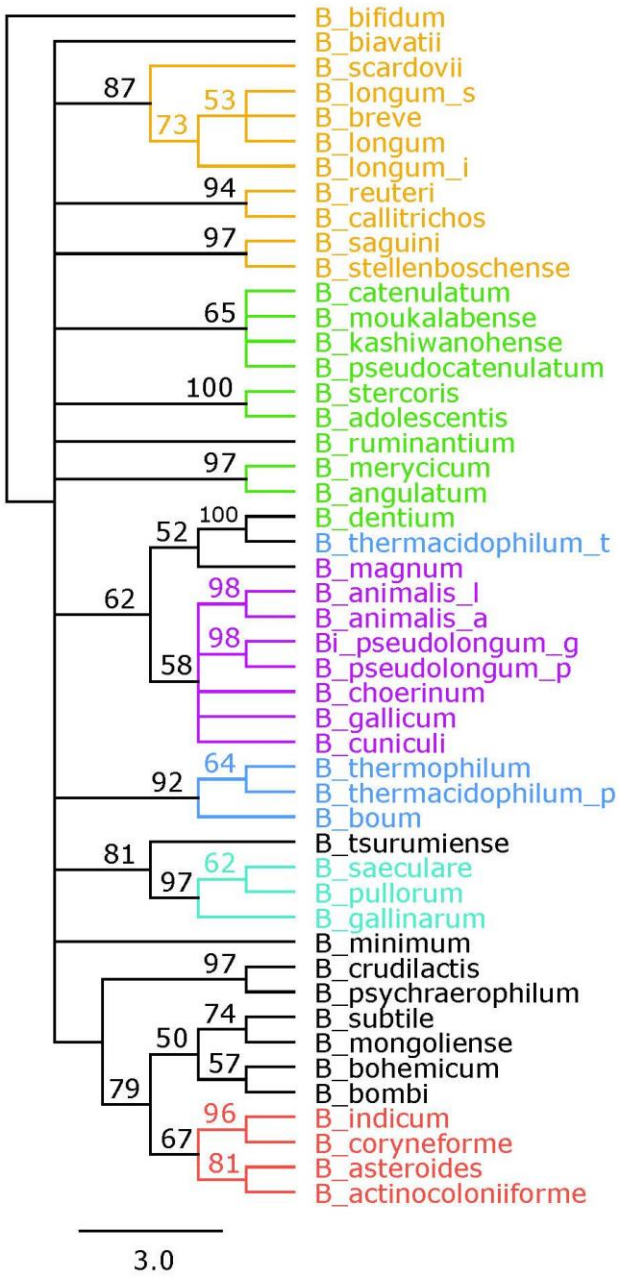


Figure 2.S9 | Eno Tree. Consensus tree based on alignment of the amino acid sequences of Eno. Trees were made using RaxML. Bootstrap values are found on each node. Phylogenetic gropus are colored as follows: *Bifidobacterium longum* is orange, *Bifidobacterium adolescentis* is green *Bifidobacterium psdeudolongum* is purple, *Bifidobacterium pullorum* is blue-green, *Bifidobacterium boum* is blue, and *Bifidobacterium asteroides* is red. Species names following the naming convention from Table 2.1.

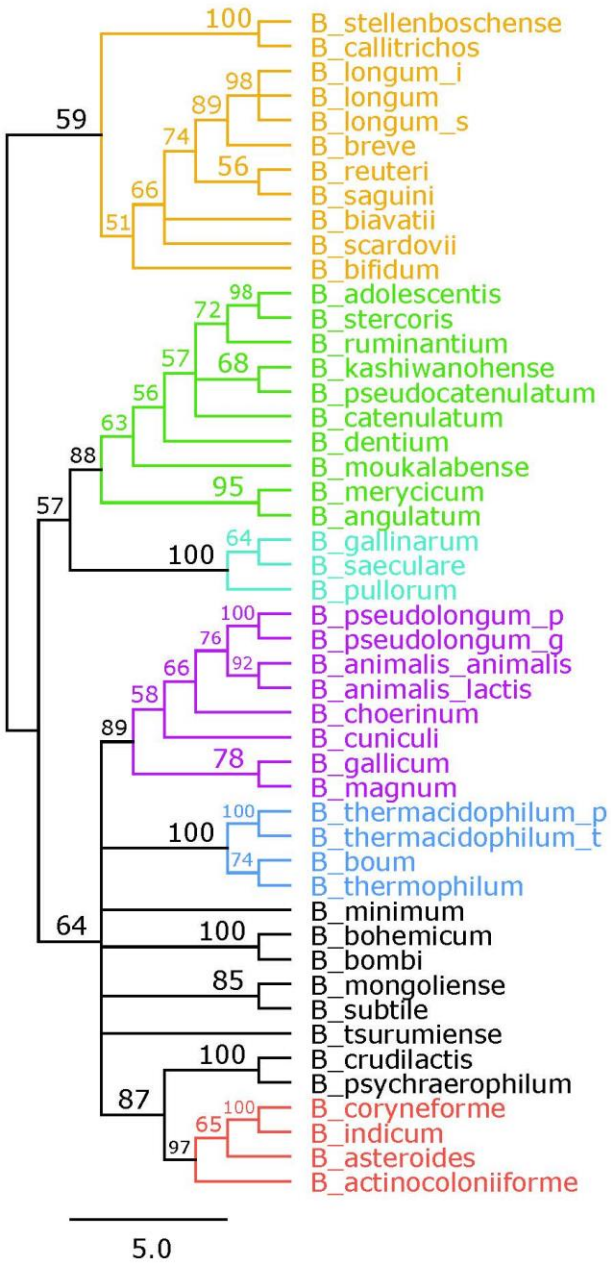


Figure 2.S10 | Pyk Tree. Consensus tree based on alignment of the amino acid sequences of Pyk. Trees were made using RaxML. Bootstrap values are found on each node. Phylogenetic gropus are colored as follows: *Bifidobacterium longum* is orange, *Bifidobacterium adolescentis* is green *Bifidobacterium psdeudolongum* is purple, *Bifidobacterium pullorum* is blue-green, *Bifidobacterium boum* is blue, and *Bifidobacterium asteroides* is red. Species names following the naming convention from Table 2.1.

CHAPTER 3: USING GLYCOLYSIS ENZYME SEQUENCES TO INFORM LACTOBACILLUS PHYLOGENY

3.1. CONTRIBUTION TO WORK

Katelyn Brandt is first author on "Using glycolysis enzyme sequences to inform Lactobacillus phylogeny" published in *Microbial Genomics*. She was responsible for planning and executing experiments. She and Rodolphe Barrangou analyzed data and wrote the manuscript. The following chapter is from Brandt and Barrangou, *Microbial Genomics* 2018;4.

3.2. ABSTRACT

The genus Lactobacillus encompasses a diversity of species that occur widely in nature and encode a plethora of metabolic pathways reflecting their adaptation to various ecological niches, including human, animals, plants, and food products. Accordingly, their functional attributes have been exploited industrially and several strains are commonly formulated as probiotics or starter cultures in the food industry. Although divergent evolutionary processes have yielded the acquisition and evolution of specialized functionalities, all Lactobacillus species share a small set of core metabolic properties, including the glycolysis pathway. Thus, the sequences of glycolytic enzymes afford a means to establish phylogenetic groups with the potential to discern species that are too closely related from a 16S rRNA standpoint. Here, we identified and extracted glycolysis enzyme sequences from 52 species and carried out individual and concatenated phylogenetic analyses. We show that a glycolysis-based phylogenetic tree can robustly segregate lactobacilli into distinct clusters and discern very closely related species. We also compare and contrast evolutionary patterns with genome-wide features and transcriptomic patterns, reflecting genomic drift trends. Overall, results suggest that glycolytic enzymes provide valuable phylogenetic insights and may constitute practical targets for evolutionary studies.

3.3. DATA SUMMARY

RNA sequencing data has been deposited at the National Center for Biotechnology Information, BioProject PRJNA420353.

3.4. INTRODUCTION

Genome adaptation is an important feature for speciation, and evolutionary processes balance various adaptive techniques for optimal growth and survival. At the genome level, adaptation features may include gene synteny conservation, G+C mol% drift, as well as codon bias optimization (1-3). A working balance of these and other forces enable an organism to become uniquely adapted to its niche, and build up competitive advantages in shifting environmental conditions, or overcome predators and competitors. Such unique adaptations are the basis of phylogenetic studies and allow researchers various degrees of discrimination. At the genus and species levels, additions and deletions of genes can be used to define the pan- and core-genome, respectively, and genome architecture can be used to evaluate synteny (4). At the strain level, nucleotide polymorphisms afford the highest resolution opportunities, with the ability to compare and contrast nearly identical isolates and even clonal relatives (5, 6).

For prokaryotic species, various tools and methodologies have been used to compare and contrast genomes, but the challenges are often genus or species-specific, and approaches can vary depending on the desired resolution and encompassed genetic diversity (7). In some cases where within genus diversity is extensive, such as in bifidobacteria and lactobacilli, using canonical housekeeping genes or universal markers (i.e. 16S) has proven difficult or limited (8-11). Also, there has yet to be defined a consistent set of genes to be utilized for MLST studies. Indeed, while universally conserved 16S sequences afford opportunities for metagenomic analyses, their shortcomings and biases are increasingly under scrutiny (12-14).

For some genera, it has become obvious that the 16S resolution limit has been met and a new set of criteria must be established. One such genus is *Lactobacillus.* Belonging to the Lactic Acid Bacteria (LAB) group, this genus is composed of over 150 Gram-positive, low G+C

species (15, 16). Lactobacilli have been used as starter cultures in the food industry for decades, and by humankind for millennia, and as such have been labeled Generally Regarded as Safe (GRAS) and benefit from the Qualified Presumption of Safety (QPS) (17). Food-related studies have led to the assertion that some strains in select species are to be considered probiotic ("live microorganisms which when administered in adequate amounts confer a health benefit on the host") (18), and as such, are now predominantly featured in dairy foods and widely formulated in probiotic dietary supplements (19). Recently, the advent of microbiome studies has revealed that microbial populations are more numerous, diverse, and variable than originally thought (20, 21). With both qualitative and quantitative considerations, associations and sometimes even correlations have been established between members of the microbiome and host health, though the accuracy and precision with which bacteria are identified vary widely and are not universally satisfactory. One such instance concerns the genus *Lactobacillus*, which has been established as an important colonizer of the human gastrointestinal tract (GIT) (22). Additional research is thus needed in this area, as researchers better grasp the role of this genus in health and disease (23-28). Some are already being exploited, such as using lactobacilli as a tool to deliver vaccines (29). Arguably, we are far from exhausting all the possible uses of this functional genus. However, in order to be able to fully utilize the numerous functions of *Lactobacillus*, we must first establish a method that enables us to properly identify and relate the many diverse species within this genus. While 16S sequencing has gotten us this far, it has a limited ability to distinguish between closely related species and represent overall genomic content and reflect genome-wide trends. These shortcomings are certainly not unique to *Lactobacillus*, and with the ever-increasing expansion of our understanding of the microbial world (30), there is a need to identify 16S-independent genomic features that capture diversity on a more granular level. It is

thus imperative that a standard method be developed which allows the proper identification of species. In order to achieve this, we assessed the potential of the widespread glycolysis pathway enzyme sequences to inform phylogeny.

In this paper, we applied a previously used method of phylogenetic analyses using the classical glycolysis enzymes as phylogenetic markers (31) to a diverse set of *Lactobacillus* species in order to establish its effect on a complicated genus. Though previous studies had used glycolysis as an expansion of ribosomal trees (32), we determined how a broad glycolysis-based phylogeny compares to the ribosomal tree. Specifically, previous studies have applied glycolysis-based approaches to Lactic Acid Bacteria in order to define an evolutionary pathway. By adding data from the entirety of the glycolysis and pentose phosphate pathways, Salvetti, et al. were able to apply phenotypic data to explain the branching of the LAB tree, as well as highlight some areas of misclassification in the 16S tree (31). Here, we propose using the entirety of the canonical glycolysis pathway as a replacement phylogenetic marker for the 16S rRNA. Conveniently, glycolysis enzymes, much like the 16S rRNA, are universally present, at least partially, and conserved, and constitute suitable candidates for phylogenetic analyses (33, 34). Here, we demonstrate that this method can assign phylogenetic relationships consistent with what is known from the 16S marker, though at a much higher discriminatory power. Specifically, we compared sequence-based alignment trees of a representative set of lactobacilli using 16S rRNA and glycolysis generated trees. We also analyzed the occurrence and location, expression, and G+C mol% of each glycolysis gene. The location and transcriptional profiles confirm that these genes are conserved and highly transcribed with varying levels of drift.

3.3. IMPACT STATEMENT

Though 16S rRNA-based phylogeny methods have been broadly used, they have a limited ability to precisely ascribe genus-species across the prokaryotic branch of the Tree of Life. In this study, we have shown that using glycolysis enzyme sequences for phylogenetic analyses can be applied to the diverse genus *Lactobacillus*, and is able to consistently unravel phylogenetic groups and precisely ascertain relatedness, even between species nearly identical on the classical ribosomal tree. Because of its universal presence and its greater diversity compared to 16S rRNA sequences, we posit that these sequences could be valuable markers in future phylogenetic and microbiome studies, specifically by providing connections to the other major branches, and enabling increased resolution. This can also be used to help identify unknown and un-culturable species, as the glycolysis enzymes are widespread, variable and allow for greater discriminatory power. Importantly, variability within some of the hypervariable regions within glycolytic sequences can also provide discrimination within a species. Looking forward, expanding this analysis to other genera and phylogenetic branches could open new avenues for evolutionary studies, and investigating the phylogeny, composition and diversity of microbial populations in complex microbiomes.

3.6. METHODS

3.6.1. Genomes

We selected 52 diverse species and subspecies of *Lactobacillus* for analysis, sampled across and throughout the 16S and core- and pan-genome tree (Table 3.1). We ensured this set was representative of this paraphyletic genus, and included species from various niches, as previously established (16). The genomes were mined using Geneious version 9.0.5 (35) to identify the classical glycolysis genes in each species (Fig. 3.S1-2). Four references genomes were used to make a curated database for the glycolysis genes, namely *Lactobacillus acidophilus, Lactobacillus gasseri, Lactobacillus reuteri,* and *Lactobacillus rhamnosus.* The Annotate from Database feature was used to annotate the other genomes. To validate the annotations, especially in the case of multiple hits, a combination of BLAST, GET_HOMOLOGUES, and mRNA-Seq data were used (36, 37). The 16S rRNA sequences were extracted from the genomes and BLAST was used to validate any cases where there were multiple hits. Once annotated and curated, the genes were extracted from the genome. The glycolysis genes were then translated and confirmed by ExPASy (38). For the concatenated tree, the amino acid sequences were joined together in order of their presence in the glycolysis pathway (Fig. 3.S1).

3.6.2. Transcriptional Profiles of Glycolysis Genes

We analyzed RNA transcription profiles from mRNA-Seq analyses for six species (*L. acidophilus, L. amylovorus, L. crispatus, L. delbrueckii_b, L. gasseri, L. helveticus*) with the previously published isolation method, analyses, and mRNA-sequencing (39). Briefly, we used mRNA-Seq data generated by our laboratory to determine the boundaries and quantitative

amounts of RNA transcripts for glycolysis genes as previously described. Samples were grown to mid-log phase and flash-frozen. Single-read RNA-Sequencing was performed on the extracted RNA using an Illumina HiSeq 2500.Data was then quality assessed, trimmed, filtered, and mapped on the reference genomes. Samples were grown to mid-log phase and flash-frozen. Single-read RNA-Sequencing was performed on the extracted RNA using an Illumina HiSeq 2500. Presumably, levels of constitutive transcription reflect biological relevance in the tested conditions and transcript boundaries inform on co-transcribed functional pairs.

3.6.3. Alignments and Trees

Alignments and trees were generated using a previously described methodology (31). Briefly, once curated sequences were extracted, we aligned the sequences using ClustalW (IUB, gap penalty of 15, gap extension of 6.66), MUSCLE (8 iterations), Geneious (global alignment with free end gaps, cost matrix was BLOSUM62 (amino acids) or 65% similarity (nucleotide)), and MAFFT (algorithm was auto, scoring matric was BLOSUM62 and BLOSUM80 (amino acids) or 100PAM or 200PAM (nucleotide), gap penalty of 1.53, offset 0.123), then used trimAl (compareset and automated1) to find a consistent alignment (35, 40-43). Trees were then generated using RaxML (CAT BLOSUM62 (amino acids) or CAT GTR (nucleotide), Bootstrap using rapid hill climbing with random seed 1, replicates were 100). (44). A consensus tree was then established using a 50% threshold level.

3.6.4. R Analyses

Statistical analyses were performed using R version 3.2.2. (45). R was used to create plots, graphs, and quantitative data. Statistical tests used included a two-tailed t-test for

comparing G+C contents. Default settings were used to preform statistical analyses and assess quantitative distributions.

3.7. RESULTS

3.7.1. 16S rRNA Phylogeny

We first generated a 16S rRNA-based tree to use as a reference for our subsequent analyses. A phylogenetic tree based off of the alignment of the 16S rRNA sequences from a representative set of 52 species and sub-species of *Lactobacillus* is depicted in Fig. 3.1. Six phylogenetic groups were identified based on their branching: the *Lactobacillus animalis* group, the *Lactobacillus vaginalis* group, the *Lactobacillus buchneri* group, the *Lactobacillus rhamnosus* group, the *Lactobacillus acidophilus* group, and the *Lactobacillus gasseri* group. These groupings are consistent with historically established relationships, as well as recent core-genome analyses (16, 46). Some of these groups also encompass species that have been historically associated with distinct niches and points of isolation (i.e. mucosal vs. intestinal vs. dairy origins) (16). The groups ranged in size from four to nine genomes with the *L. rhamnosus* group as the smallest and the *L. animalis* group as the largest. The bootstrap values for the 16S tree ranged from 51 to 100. There were 27 nodes that had a bootstrap of 70 or greater (Fig 3.S3). We used these six phylogenetic groups as references for our subsequent analyses, though some species were not assigned to one of these six groups.

3.7.2. Glycolysis Gene Expression

Before using the glycolysis enzymes as phylogenetic markers, we first explored their genetic properties in *Lactobacillus*. Of the 52 *Lactobacillus* species and sub-species selected, 35 species encoded all ten of the classical glycolytic genes present. In contrast, 16 species (encompassing the *L. vaginalis* and *L. buchneri* groups) presented eight of the canonical genes (missing *pfk* and *fba*) (Fig. 3.S2). In such cases, alternative metabolic pathways may be utilized,

such as the pentose phosphate pathway (*Lactobacillus fermentum*) or the phosphoketolase pathway (*L. buchneri*) (47, 48). *Lactobacillus reuteri* uses a mixture of the Embden-Meyerhof pathway and phosphoketolase pathway and thus was the only species with six of the glycolysis genes (Fig. 3.S2) (49).

Next, we characterized the transcripts of glycolysis genes in *Lactobacillus*. Chromosome location and mRNA sequence data were analyzed from six species: *L. acidophilus*, *Lactobacillus amylovorus*, *Lactobacillus crispatus*, *Lactobacillus delbrueckii* subsp. *bulgaricus*, *L. gasseri*, and *Lactobacillus helveticus*. These six species fall into the *L. acidophilus* and *L. gasseri* groups, and all six species contain the complete glycolysis genes complement, allowing for inferences on all of the genes in this study, instead of just a subset. Fig. 3.2 depicts the location of the glycolysis genes on normalized chromosomes for each of these six species. Noteworthy, two operons can be visualized: the *gap*, *pgk*, and *tpi* operon, as well as the *pfk* and *pyk* operon. Furthermore, the operon boundaries are clearly seen in the mRNA coverage data for each of the six species (Fig. 3.3). The remaining five genes have clear start and stop boundaries. Notably, *L. helveticus* has a unique arrangement of the glycolysis genes compared to the other five species, possibly due to the large number of IS elements leading to genome decay, however the operons remain conserved (50). Next, we compared the expression levels of the glycolysis genes to the whole transcriptome. We found that the glycolysis genes are among the most highly expressed genes. Indeed, considering the top 10% of the most highly expressed genes in the cell, nine of the ten glycolysis genes are listed (Fig. 3.4). The only gene absent from the top 10% is *pgm*. Noteworthy, the *gap* gene is consistently among the top three most highly expressed genes in all six species. Such a consistently high transcription level indicates that the *gap* gene is critical to the functionality of the cell and perhaps, as such, less susceptible to changes. This is also

reflected by the conserved location of *gap* in the genome and operon structure amongst the strains studied (Fig. 3.2), potentially indicating uses for *gap* in identification These results demonstrate that glycolysis genes are genomically conserved, organizationally syntenous, and transcriptionally important, showcasing their use as potential phylogenetic markers.

3.7.3. Glycolysis-Based Phylogeny

To create a glycolysis-based phylogeny for the 52 selected *Lactobacillus* species and subspecies, the concatenated amino acid sequences of the glycolysis enzymes were used (Fig. 3.5). The enzymes were concatenated in their order of occurrence in the glycolysis pathway (Fig. 3.S1). For organisms with all enzymes present, this meant ten sequences were concatenated together, whereas only six to eight amino acid sequences were concatenated for the other species (Fig. 3.S2). The six phylogenetic groups identified from the 16S reference tree, namely: *L. animalis, , L. vaginalis, L. buchneri, L. rhamnosus, L. acidophilus,* and *L. gasseri*, were also identified in the concatenated tree and follow the same clustering (colouring) scheme. The bootstrap values for the concatenated tree ranged from 52 to 100. Nodes with bootstrap values equal to or greater than 70 numbered 43, a 59% increase from that of the 16S tree. Overall, the concatenated tree correctly assigned the phylogenetic groups established from the 16S tree. In addition, the concatenated tree better discerned how the phylogenetic groups relate to one another, even within groups. This is supported by the higher bootstrap values (Fig. 3.S3). Trees based off of the individual glycolysis enzymes can be found in Fig. 3.S4-13. The sum of branch lengths for each tree can be found in Table 3.S1. A detailed comparative analysis of various trees structures revealed that overall, there is high congruence in clustering both between and within

the six established groups, though with various levels of discrimination across each protein sequence. Repeatedly, glycolysis-based trees provided more discriminatory power than 16S.

3.7.4. G+C Content Analyses

Next, we looked at the G+C mol% and genomic drift of the glycolysis genes across the various species. Fig. 3.6 shows a notched boxplots comparing the G+C mol% of each sequence set (the 16S sequence, the 10 genes, and the concatenated sequences) in this study, compared to the genome-wide G+C mol%, ranked in increasing order. The G+C mol% of the *pgm* gene is closest to that of the total genome, while the 16s rRNA gene is the farthest and an outlier. The notches are indicative of strong evidence that the medians differ when the notches do not overlap (51). The 16S rRNA gene does not overlap with any other gene. In fact, a two-tailed t-test with a *P* value less than 0.001 (2.2×10^{-16}) revealed that the G+C mol% of the 16S rRNA sequence was statistically distinct from that of the total genome G+C mol%. This indicates that the 16S rRNA gene is not matching the pace of drift of the total genome with regards to G+C mol%. In contrast, all of the glycolysis genes, with the exception of *pfk* and *eno* were not statistically different from the total genome G+C mol% (*P* value greater than 0.01), indicating that G+C mol% drift for glycolysis genes provide insights into the genome-wide G+C mol% drift. This further supports glycolytic sequences as intriguing candidates for both phylogenetic studies, and representatives of genome-wide trends.

The genome sizes in this study ranged from 1.28Mb (*Lactobacillus iners*) to 3.65Mb (*Lactobacillus pentosus*), again reflecting the extensive genomic diversity within this genus. The total G+C mol% ranged from 32.50% (*L. iners*) to 57.00% (*Lactobacillus nasuensis*), which is intriguing given the general assumption that all lactobacilli are low G+C mol% organisms.

Nevertheless, the mean G+C mol% was 40.70%, consistent with *Lactobacillus* being generally perceived as low G+C mol% organisms. Splitting the species into high, medium, and low categories, it becomes apparent that most species are trending towards the lower end of the spectrum, and away from the higher G+C mol% range (Fig. 3.7a). Some of the phylogenetic groups are closely clustered, such as the *L. acidophilus* group, *L. gasseri* group, and the *L. rhamnosus* group, with the exception of *L. delbrueckii_b* (a dairy bacterium) and *L. nasuensis* (an aforementioned outlier in G+C mol%). The *L. animalis* group and *L. buchneri* group are similarly clustered, albeit more loosely. These observations hold true when comparing the G+C mol% of all the individual genes in their respective genomes, perhaps reflecting a consistent and genome-wide pace of drift, rather than variable speeds of drift for each gene (Fig. 3.7b). Again, the 16S sequence has a much higher G+C mol% than most of the other studied genes, with the outlier *L. nasuensis* deviating from the consensus. The G+C mol% of the glycolysis genes within clusters are often times very close, as exemplified by the *L. acidophilus* group.

3.8. DISCUSSION

The genomic and functional attributes of *Lactobacillus* render it a pervasive genus, both in research and in the industry. The benefits and uses of this diverse set of species are well-established and exhaustive, and yet, the list continues to grow. Many *Lactobacillus* strains are now considered to be health-promoting in the form of probiotics and are often found to be a part of a healthy microbiome (26), and are being engineered to promote health host-microbe interactions, and deliver bioactive compounds such as vaccines (52). As microbiome studies expand, we anticipate that the interest in *Lactobacillus* is set to increase, especially given their occurrence in several human-associated microbiomes, encompassing intestinal, vaginal, oral, and skin (21). Many studies have been published discussing the role of *Lactobacillus* in the microbiota, including research into the microbiota changes through disease, enhancing the microbiome as a form of treatment, and how the microbiome reacts to drugs (53-55). The continuously expanding list of uses and studies just illustrates how important it is to accurately identify *Lactobacillus* species. While all species of *Lactobacillus* share some classical features of LAB organisms, notably their ability to produce lactic acid, the similarities between species are relatively few. In fact, even basic characteristics such as niche and isolation source can vary radically. Proper identification is an increasing concern especially when it comes to disease modeling in the human microbiome, as well as the formulation, tracking and efficacy of probiotic strains. Innovative techniques are continuously being developed and often use a combination of 16S rRNA with developing technologies, such as MALDI-TOF (56). However, these tools are not broadly accessible and still rely partially on the sometimes unsatisfactory 16S rRNA. Here, we provide a practical alternative to the classical use of 16S sequencing.

In this paper, we applied the previously proposed methodology of using glycolysis sequences to perform phylogenetic studies (31) in the genus *Lactobacillus*. We demonstrated that this method is a practical and robust approach for *Lactobacillus*. Compared to the traditional 16S rRNA method, this approach was able to consistently identify phylogenetic groupings, with notably high-resolution between closely-related species. While the 16S rRNA-based tree was able to identify the six phylogenetic groups, the concatenated tree was able to add more discrimination both between and within groups, evidenced by the higher bootstrap values in the glycolysis based tree. Our grouping is consistent with a previous study using glycolysis sequences for phylogenetic analysis of *Lactobacillus* species (31). Further analyses based on genomic content revealed clues as to why the glycolysis-based tree was better able to assign species.

First, looking at the organization of the genes in the genomes revealed two conserved operons in *Lactobacillus*, the *gap* operon and the *pfk* operon, with the remaining enzymes showing clear start and stop boundaries. This shared synteny emphasizes the importance of glycolysis gene conservation. Next, we looked at expression level. The glycolysis genes were consistently among the most highly expressed genes in the cell, with the *gap* gene always in the top three most abundant transcripts. These high expression levels indicate a great use and energy expenditure, and thus arguably reflect the biological importance of this gene to the cell. Because of this importance, the glycolysis genes are much less likely to be subjected to loss. The operon structures and expression levels of the glycolysis genes are significant because a main criterion for selecting the 16S rRNA as a phylogenetic marker was its high conservation among species (57). Next, we looked at how the glycolysis genes reflected genomic drift in terms of G+C mol%. First, it would appear that the genus is reaching a stabilizing point in its G+C mol% drift,

though some species with high G+C mol% still have margin for extending the trend (*L. nasuensis, L. zymae, L. fermentum*). Next, we saw that the glycolysis gene G+C mol% was extremely close to that of the genome-wide G+C mol%, while the 16S rRNA was startlingly higher ($P<0.001$), underscoring the fact that the 16S rRNA is by all accounts much different than that of the total genome, whereas the majority of the glycolysis are significantly similar to the total genome G+C mol% (Fig. 3.6). This provides a possible explanation for the reason why the 16S rRNA analyses have been limited at a high-resolution level in *Lactobacillus* and why the glycolysis based tree was able to reach a higher-resolution level. In fact, it has long been noted that 16S rRNA is unable to discriminate between species of lactobacilli due to its high similarity amongst them (58). The individual glycolysis genes are much more similar to the genome as a whole (Fig. 3.6). Additionally, individual glycolysis genes are also able to accurately assign species to groups with a high-resolution (Fig. 3.S4-13). The *gap* gene is of particular note, due to its presence in an operon, consistently high expression, G+C mol% and ability to accurately define species groups. Overall, the glycolysis-based approach was able to provide a higher-resolution phylogeny for *Lactobacillus,* due in part to its conservation, expression, and reflection of genomic drift.

3.9. ACKNOWLEDGEMENTS

The authors thank funding sources for their support and the CRISPR lab for insightful conversations.

3.10. REFERENCES

1. Dandekar T, Snel B, Huynen M, Bork P. Conservation of Gene Order: a fingerprint of proteins that physically interact. Trends in Biochemical Sciences. 1998;23(9):324-8.

2. Karlin S, Campbell AM, Mrazek J. Comparative DNA Analysis Across Diverse Genomes. Annual review of genetics. 1998;32:185-225.

3. Karlin S, Mrázek J, Campbell A, Kaiser D. Characterizations of Highly Expressed Genes of Four Fast-Growing Bacteria. Journal of Bacteriology. 2001;183(17):5025-40.

4. Boekhorst J, Siezen RJ, Zwahlen M-C, Vilanova D, Pridmore RD, Mercenier A, et al. The Complete Genomes of *Lactobacillus plantarum* and *Lactobacillus johnsonii* Reveal Extensive Differences in Chromosome Organization and Gene Content. Microbiology. 2004;150(11):3601-11.

5. Hoffmann M, Zhao S, Pettengill J, Luo Y, Monday SR, Abbott J, et al. Comparative Genomic Analysis and Virulence Differences in Closely Related *Salmonella enterica* Serotype Heidelberg Isolates from Humans, Retail Meats, and Animals. Genome Biology and Evolution. 2014;6(5):1046-68.

6. Morán Losada P, Tümmler B. SNP Synteny Analysis of *Staphylococcus aureus* and *Pseudomonas aeruginosa* Population Genomics. FEMS Microbiology Letters. 2016;363(19):fnw229-fnw.

7. Makarova K, Slesarev A, Wolf Y, Sorokin A, Mirkin B, Koonin E, et al. Comparative Genomics of the Lactic Acid Bacteria. Proceedings of the National Academy of Sciences. 2006;103(42):15611-6.

8. Claesson MJ, van Sinderen D, apos, Toole PW. *Lactobacillus* Phylogenomics – Towards a Reclassification of the Genus. International Journal of Systematic and Evolutionary Microbiology. 2008;58(12):2945-54.

9. Felis GE, Dellaglio F, Mizzi L, Torriani S. Comparative Sequence Analysis of a *recA* Gene Fragment Brings New Evidence for a Change in the Taxonomy of the *Lactobacillus casei* Group. Int J Syst Evol Microbiol. 2001;51(Pt 6):2113-7.

10. Milani C, Turroni F, Duranti S, Lugli GA, Mancabelli L, Ferrario C, et al. Genomics of the Genus *Bifidobacterium* Reveals Species-Specific Adaptation to the Glycan-Rich Gut Environment. Applied and Environmental Microbiology. 2016;82(4):980-91.

11. Milani C, Lugli GA, Turroni F, Mancabelli L, Duranti S, Viappiani A, et al. Evaluation of Bifidobacterial Community Composition in the Human Gut by Means of a Targeted Amplicon Sequencing (ITS) Protocol. FEMS Microbiology Ecology. 2014;90(2):493-503.

12. Baker GC, Smith JJ, Cowan DA. Review and Re-Analysis of Domain-Specific 16S Primers. Journal of Microbiological Methods. 2003;55(3):541-55.

13. Clarridge JE. Impact of 16S rRNA Gene Sequence Analysis for Identification of Bacteria on Clinical Microbiology and Infectious Diseases. Clinical Microbiology Reviews. 2004;17(4):840-62.

14. de la Cuesta-Zuluaga J, Escobar JS. Considerations For Optimizing Microbiome Analysis Using a Marker Gene. Frontiers in nutrition. 2016;3:26.

15. Salvetti E, Torriani S, Felis GE. The Genus *Lactobacillus*: A Taxonomic Update. Probiotics and Antimicrobial Proteins. 2012;4(4):217-26.

16. Sun Z, Harris HMB, 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.

17. Bernardeau M, Vernoux JP, Henri-Dubernet S, Guéguen M. Safety Assessment of Dairy Microorganisms: The *Lactobacillus* genus. International Journal of Food Microbiology. 2008;126(3):278-85.

18. 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.

19. Saxelin M. Probiotic Formulations and Applications, the Current Probiotics Market, and Changes in the Marketplace: A European Perspective. Clinical Infectious Diseases. 2008;46(Supplement 2):S76-S9.

20. Cho I, Blaser MJ. The Human Microbiome: at the interface of health and disease. Nature reviews Genetics. 2012;13(4):260-70.

21. The Human Microbiome Project C. Structure, Function and Diversity of the Healthy Human Microbiome. Nature. 2012;486(7402):207-14.

22. Conlon MA, Bird AR. The Impact of Diet and Lifestyle on Gut Microbiota and Human Health. Nutrients. 2015;7(1):17-44.

23. Li X, Wang N, Yin B, Fang D, Jiang T, Fang S, et al. Effects of *Lactobacillus plantarum* CCFM0236 on Hyperglycemia and Insulin Resisitance in High-fat and Streptozotocin Induced Type 2 Diabetic Mice. Journal of Applied Microbiology. 2016:n/a-n/a.

24. Feng X-B, Jiang J, Li M, Wang G, You J-W, Zuo J. Role of Intestinal Flora Imbalance in Pathogenesis of Pouchitis. Asian Pacific Journal of Tropical Medicine. 2016;9(8):786-90.

25. Hooper LV, Midtvedt T, Gordon JI. How Host-Microbial Interactions Shape the Nutrient Environment of the Mammalian Intestine. Annual Review of Nutrition. 2002;22(1):283-307.

26. Gerritsen J, Smidt H, Rijkers GT, de Vos WM. Intestinal Microbiota in Human Health and Disease: the impact of probiotics. Genes & Nutrition. 2011;6(3):209-40.

27. Shreiner AB, Kao JY, Young VB. The Gut Microbiome in Health and in Disease. Current opinion in gastroenterology. 2015;31(1):69-75.

28. Okai S, Usui F, Yokota S, Hori-i Y, Hasegawa M, Nakamura T, et al. High-Affinity Monoclonal IgA Regulates Gut Microbiota and Prevents Colitis in Mice. Nature Microbiology. 2016;1:16103.

29. O'Flaherty S, Klaenhammer TR. Multivalent Chromosomal Expression of the *Clostridium botulinum* Serotype A Neurotoxin Heavy Chain Antigen and *Bacillus anthracis* Protective Antigen in *Lactobacillus acidophilus*. Applied and Environmental Microbiology. 2016.

30. Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, et al. A New View of the Tree of Life. Nature Microbiology. 2016;1:16048.

31. Brandt K, Barrangou R. Phylogenetic Analysis of the *Bifidobacterium* Genus Using Glycolysis Enzyme Sequences. Frontiers in Microbiology. 2016;7(657).

32. Salvetti E, Fondi M, Fani R, Torriani S, Felis GE. Evolution of lactic acid bacteria in the order Lactobacillales as depicted by analysis of glycolysis and pentose phosphate pathways. Systematic and Applied Microbiology. 2013;36(5):291-305.

33. Fothergill-Gilmore LA. The Evolution of the Glycolytic Pathway. Trends in Biochemical Sciences. 1986;11(1):47-51.

34. Fothergill-Gilmore LA, Michels PAM. Evolution of Glycolysis. Progress in Biophysics and Molecular Biology. 1993;59(2):105-235.

35. 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.

36. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic Local Alignment Search Tool. Journal of molecular biology. 1990;215(3):403-10.

37. Contreras-Moreira B, Vinuesa P. GET_HOMOLOGUES, a Versatile Software Package for Scalable and Robust Microbial Pangenome Analysis. Applied and Environmental Microbiology. 2013;79(24):7696-701.

38. Gasteiger E, Gattiker A, Hoogland C, Ivanyi I, Appel RD, Bairoch A. ExPASy: the proteomics server for in-depth protein knowledge and analysis. Nucleic Acids Research. 2003;31(13):3784-8.

39. Johnson BR, Hymes J, Sanozky-Dawes R, Henriksen ED, Barrangou R, Klaenhammer TR. Conserved S-Layer-Associated Proteins Revealed by Exoproteomic Survey of S-Layer-Forming Lactobacilli. Applied and Environmental Microbiology. 2016;82(1):134-45.

40. Katoh K, Misawa K, Kuma Ki, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Research. 2002;30(14):3059-66.

41. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research. 2004;32(5):1792-7.

42. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, et al. Clustal W and Clustal X Version 2.0. Bioinformatics. 2007;23(21):2947-8.

43. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009;25(15):1972-3.

44. Stamatakis A. RAxML Version 8: A tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies. Bioinformatics. 2014.

45. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2015.

46. Canchaya C, Claesson MJ, Fitzgerald GF, van Sinderen D, apos, Toole PW. Diversity of the Genus Lactobacillus Revealed by Comparative Genomics of Five Species. Microbiology. 2006;152(11):3185-96.

47. Heinl S, Wibberg D, Eikmeyer F, Szczepanowski R, Blom J, Linke B, et al. Insights into the Completely Annotated Genome of *Lactobacillus buchneri* CD034, a Strain Isolated from Stable Grass Silage. Journal of Biotechnology. 2012;161(2):153-66.

48. Cárdenas N, Laiño JE, Delgado S, Jiménez E, Juárez del Valle M, Savoy de Giori G, et al. Relationships Between the Genome and Some Phenotypical Properties of *Lactobacillus fermentum* CECT 5716, a Probiotic Strain Isolated from Human Milk. Applied Microbiology and Biotechnology. 2015;99(10):4343-53.

49. Årsköld E, Lohmeier-Vogel E, Cao R, Roos S, Rådström P, van Niel EWJ. Phosphoketolase Pathway Dominates in Lactobacillus reuteri ATCC 55730 Containing Dual Pathways for Glycolysis. Journal of Bacteriology. 2008;190(1):206-12.

50. Broadbent JR, Hughes JE, Welker DL, Tompkins TA, Steele JL. Complete Genome Sequence for *Lactobacillus helveticus* CNRZ 32, an Industrial Cheese Starter and Cheese Flavor Adjunct. Genome Announcements. 2013;1(4):e00590-13.

51. Chambers JM. Notched Box Plots. Graphical Methods for Data Analysis: Wadsworth International Group; 1983. p. 60-3.

52. Seegers JFML. Lactobacilli as Live Vaccine Delivery Vectors: progress and prospects. Trends in Biotechnology. 2002;20(12):508-15.

53. Bhat M, Arendt BM, Bhat V, Renner EL, Humar A, Allard JP. Implication of the Intestinal Microbiome in Complications of Cirrhosis. World Journal of Hepatology. 2016;8(27):1128-36.

54. Bull-Otterson L, Feng W, Kirpich I, Wang Y, Qin X, Liu Y, et al. Metagenomic Analyses of Alcohol Induced Pathogenic Alterations in the Intestinal Microbiome and the Effect of *Lactobacillus rhamnosus* GG Treatment. PLOS ONE. 2013;8(1):e53028.

55. Shin CM, Kim N, Kim YS, Nam RH, Park JH, Lee DH, et al. Impact of Long-Term Proton Pump Inhibitor Therapy on Gut Microbiota in F344 Rats: Pilot Study. Gut and Liver. 2016;10(6):896-901.

56. Foschi C, Laghi L, Parolin C, Giordani B, Compri M, Cevenini R, et al. Novel approaches for the taxonomic and metabolic characterization of lactobacilli: Integration of 16S rRNA gene sequencing with MALDI-TOF MS and 1H-NMR. PLOS ONE. 2017;12(2):e0172483.

57. Eisen JA. The RecA Protein as a Model Molecule for Molecular Systematic Studies of Bacteria: Comparison of Trees of RecAs and 16S rRNAs from the Same Species. Journal of Molecular Evolution. 1995;41(6):1105-23.

58. Fox GE, Wisotzkey JD, Jurtshuk P. How Close Is Close: 16S rRNA Sequence Identity May Not Be Sufficient To Guarantee Species Identity. International Journal of Systematic and Evolutionary Microbiology. 1992;42(1):166-70.

Table 3.1 | Species and Genome List. Representative set of 52 *Lactobacillus* species and sub-species used in this study. Accession numbers and naming conventions included.

Genus	Species	Subspecies	Strain	Accession Number	Naming Convention	Locus Tag
Lactobacillus	*acidipiscis*		KCTC 13900	NZ_BACS00000000	L_acidipiscis	GSS
Lactobacillus	*acidophilus*		NCFM	NC_006814	L_acidophilus	LAB
Lactobacillus	*algidus*		DSM 15638	NZ_AZDI00000000	L_algidus	FC66
Lactobacillus	*amylolyticus*		DSM 11664	NZ_ADNY00000000	L_amylolyticus	HMPREF0493
Lactobacillus	*amylovorus*		GRL1118	NC_017470	L_amylovorus	LAB52
Lactobacillus	*animalis*		DSM 20602	NZ_AEOF00000000	L_animalis	LACAN
Lactobacillus	*aquaticus*		DSM 21051	NZ_AYZD00000000	L_aquaticus	FC19
Lactobacillus	*brevis*		ATCC 367	NC_008497	L_brevis	LVIS
Lactobacillus	*buchneri*		CD034	NC_018610	L_buchneri	LBUCD034
Lactobacillus	*cacaonum*		DSM 21116	NZ_AYZE00000000	L_cacaonum	FC80
Lactobacillus	*casei*		DSM 20011	NZ_AZCO00000000	L_casei	FC13
Lactobacillus	*coryniformis*	*torquens*	DSM 20004	NZ_AEOS00000000	L_coryniformis_t	EWE
Lactobacillus	*crispatus*		ST1	NC_014106	L_crispatus	LCRIS
Lactobacillus	*curvatus*		CRL 705	NZ_AGBU00000000	L_curvatus	CRL705
Lactobacillus	*delbrueckii*	*bulgaricus*	ATCC BAA-365	NC_008529	L_delbrueckii_b	LBUL
Lactobacillus	*farciminis*		DSM 20184	NZ_AEOT00000000	L_farciminis	LACFC
Lactobacillus	*fermentum*		CECT 5716	NC_017465	L_fermentum	LC40
Lactobacillus	*floricola*		DSM_23037	NZ_AYZL00000000	L_floricola	FC86
Lactobacillus	*gallinarum*		DSM 10532	NZ_BALB00000000	L_gallinarum	JCM2011
Lactobacillus	*gasseri*		ATCC 33323	NC_008530	L_gasseri	LGAS
Lactobacillus	*helveticus*		CNRZ32	NC_021744	L_helveticus	LHE
Lactobacillus	*hilgardii*		DSM 20176	NZ_ACGP00000000	L_hilgardii	HMPREF0519

Table 3.1 | (continued).

Lactobacillus hominis		DSM 23910	NZ_CAKE00000000	L_hominis	BN55
Lactobacillus iners		DSM 13335	NZ_ACLN00000000	L_iners	HMPREF0520
Lactobacillus jensenii		DSM 20557	NZ_AYYU00000000	L_jensenii	FC45
Lactobacillus johnsonii		NCC 533	NC_005362	L_johnsonii	LJ
Lactobacillus kimchicus		JCM_15530	NZ_AZCX00000000	L_kimchicus	FC96
Lactobacillus lindneri		DSM_20690	NZ_JQBT00000000	L_lindneri	IV52
Lactobacillus mali		DSM 20444	NZ_AKKT00000000	L_mali	LMA
Lactobacillus mindensis		DSM_14500	NZ_AZEZ00000000	L_mindensis	FD29
Lactobacillus mucosae		LM1	NZ_CP011013	L_mucosae	LBLM1
Lactobacillus nasuensis		JCM_17158	NZ_AZDJ00000000	L_nasuensis	FD02
Lactobacillus oeni		DSM_19972	NZ_AZEH00000000	L_oeni	FD46
Lactobacillus oris		F0423	NZ_AFTL00000000	L_oris	HMPREF9102
Lactobacillus otakiensis		DSM 19908	NZ_BASH00000000	L_otakiensis	LOT
Lactobacillus parabuchneri		DSM_5707	NZ_AZGK00000000	L_parabuchneri	FC51
Lactobacillus paracasei		N1115	NZ_CP007122	L_paracasei	AF91
Lactobacillus pasteurii		DSM 23907	NZ_CAKD00000000	L_pasteurii	BN53
Lactobacillus pentosus		DSM_20314	NZ_AZCU00000000	L_pentosus	FD24
Lactobacillus plantarum		16	NC_021514	L_plantarum	LP16
Lactobacillus reuteri		DSM 20016	NC_009513	L_reuteri	LREU
Lactobacillus rhamnosus		GG	NC_013198	L_rhamnosus	LGG
Lactobacillus rossiae		DSM_15814	NZ_AZFF00000000	L_rossiae	FD35
Lactobacillus ruminis		ATCC 27782	NC_015975	L_ruminis	LRC
Lactobacillus sakei	*sakei*	DSM 20017	NZ_BALW00000000	L_sakei_s	JCM1157
Lactobacillus salivarius		CECT 5713	NC_017481	L_salivarius	CECT 5713
Lactobacillus sanfranciscensis		TMW 1.1304	NC_015978	L_sanfranciscensis	LSA
Lactobacillus suebicus		DSM 5007	NZ_BACO00000000	L_suebicus	GSK
Lactobacillus sunkii		DSM_19904	NZ_AZEA00000000	L_sunkii	FD17

Table 3.1 | (continued).

Lactobacillus vaginalis	DSM 5837	NZ_ACGV00000000	L_vaginalis	HMPREF0549
Lactobacillus versmoldensis	DSM 14857	NZ_BACR00000000	L_versmoldensis	GSQ
Lactobacillus zymae	DSM_19395	NZ_AZDW00000000	L_zymae	FD38

Table 3.S1 | Sum of Branch Lengths. The sum of branch lengths for each tree can be found in this table. Included are abbreviations for each glycolysis gene and their E.C. number.

Tree	E.C. Number	Sum
16S		146.902
Concatenated		101.7221
Phosphoglucomutase, (*pgm*)	5.4.2.2	111.9324
Glucose-6-phosphate isomerase, (*pgi*)	5.3.1.9	118.2525
6-phosphofructokinase, (*pfk*)	2.7.1.11	69.89766
Fructose-bisphosphate aldolase, (*fba*)	4.1.2.13	86.18618
Triosephosphate isomerase, (*tpi*)	5.3.1.1	152.2155
Glyceraldehyde 3-phosphate dehydrogenase, (*gap*)	1.2.1.12	104.6662
Phosphoglycerate kinase, (*pgk*)	2.7.2.3	109.189
Phosphoglycerate mutase, (*gpm*)	5.4.2.11	243.8897
Enolase, (*eno*)	4.2.1.11	113.4122
Pyruvate kinase, (*pyk*)	2.7.1.40	106.5558

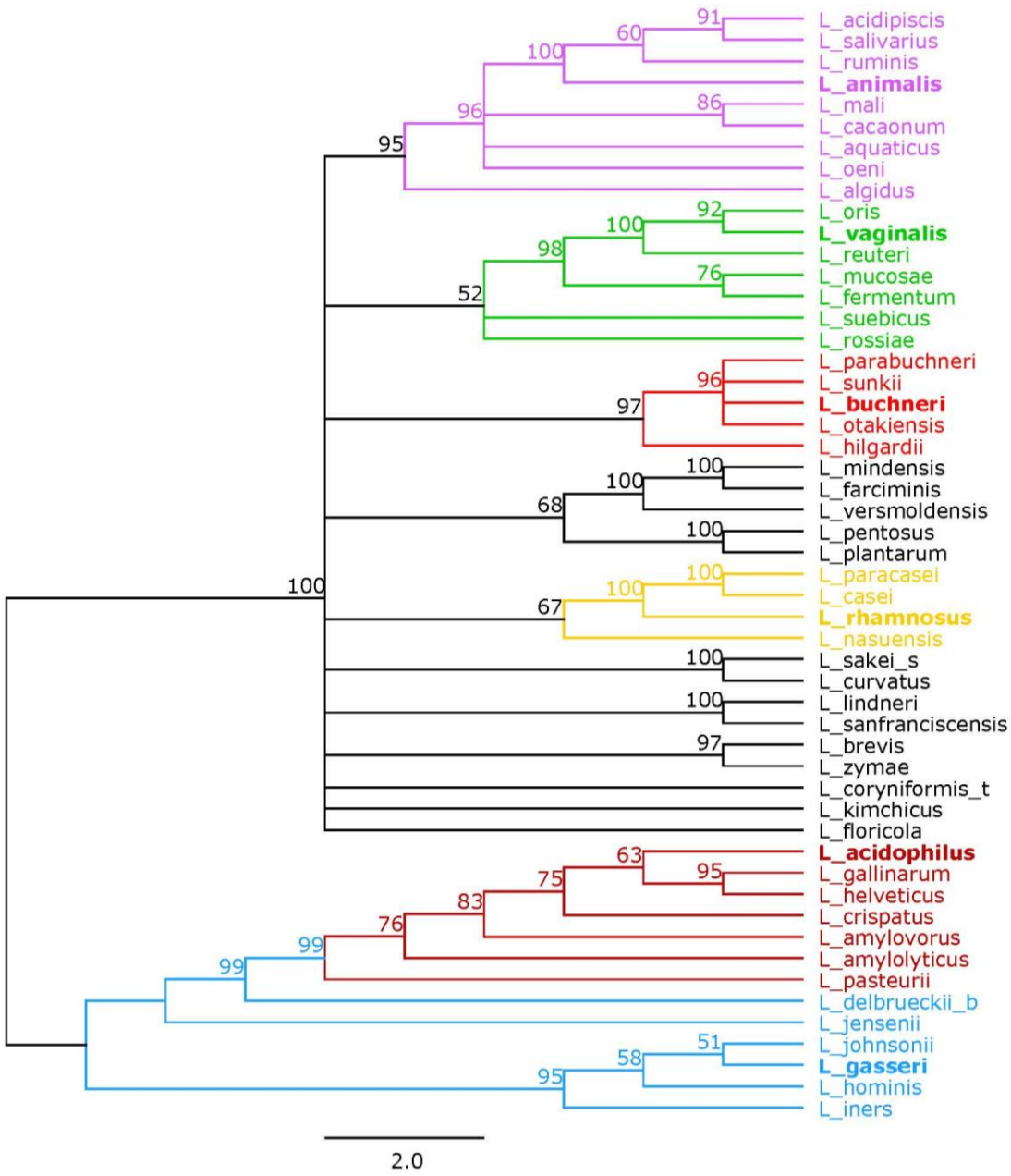


Figure 3.1 | 16S rRNA Tree. Tree based on the alignment of the 16S rRNA sequences using RaxML. Bootsrap values are recorded on the noedes. Groups are coloured as follows: the *L. animalis* group in purple, the *L. vaginalis* group in green, the *L. buchneri* group in red, the *L. rhamnosus* group in yellow, the *L. acidophilus* group in maroon, and the *L. gasseri* group in blue. The representative species in each group is in bold. Species names follow the naming convention shown in Table 3.1.

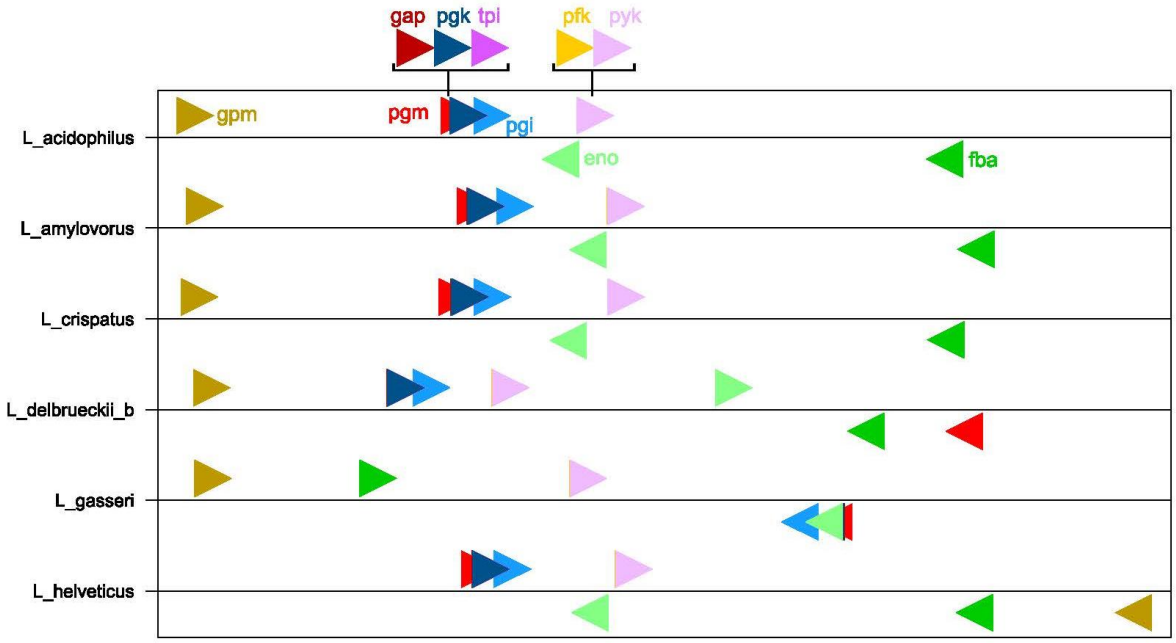


Fig. 3.2 | Genomic Location. Normalized glycolysis gene locations in *L. acidophilus, L. amylovorus, L. crispatus, L. delbrueckii_b, L. gasseri, L. helveticus.* Normalization was calculated by dividing the location on the genome by the total genome size. Right arrows indicate forward direction, left the reverse direction. The genomes are organized in the 5' to 3' direction. Colours are as follows: *pgm* in red, *pgi* in blue, *pfk* in yellow, *fba* in dark green, *tpi* in purple, *gap* in maroon, *pgk* in navy, *gpm* in mustard, *eno* in light green, and *pyk* in lavender.

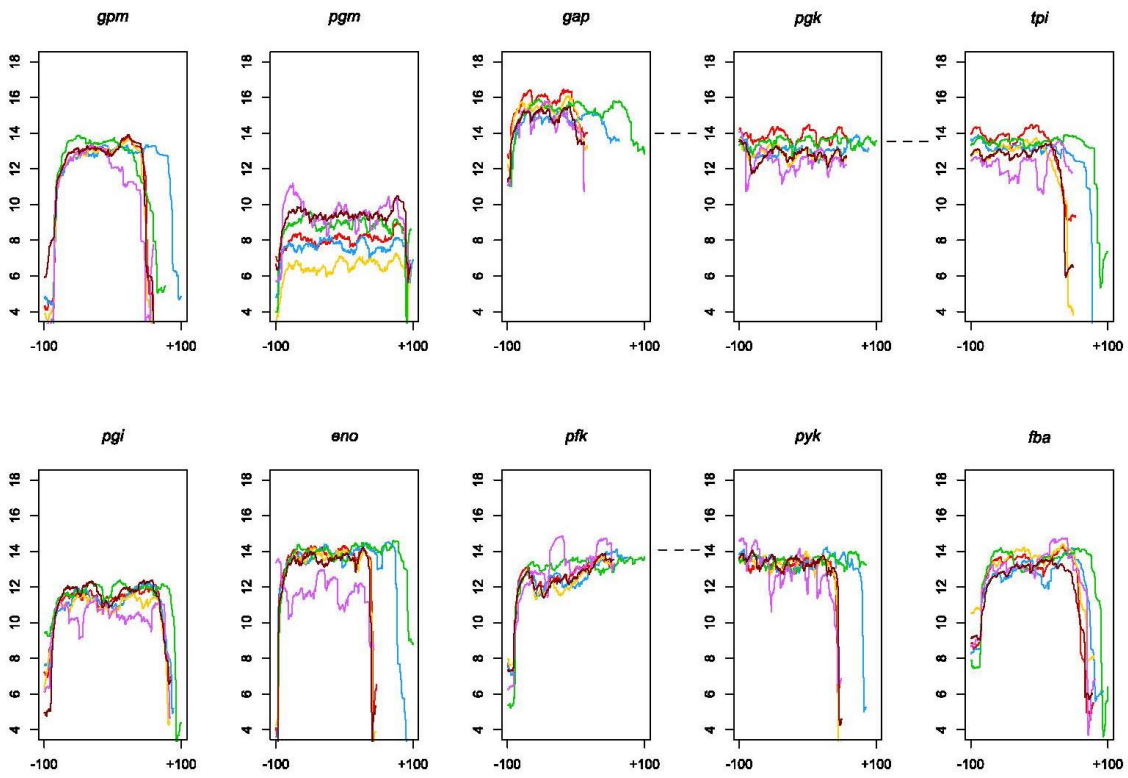


Fig. 3.3 | Glycolysis Genes Transcription. Each plot represents the mRNA-Seq coverage, log2 transformed, for the corresponding glycolysis gene over its length +/-100 represents the number of bases away from the start/end of the gene. The species are as follows: *L. acidophilus* is red, *L. amylovorus* in blue, *L. crispatus* in yellow, *L. delbrueckii_b* in green, *L. gasseri* in purple, and *L. helveticus* in maroon.

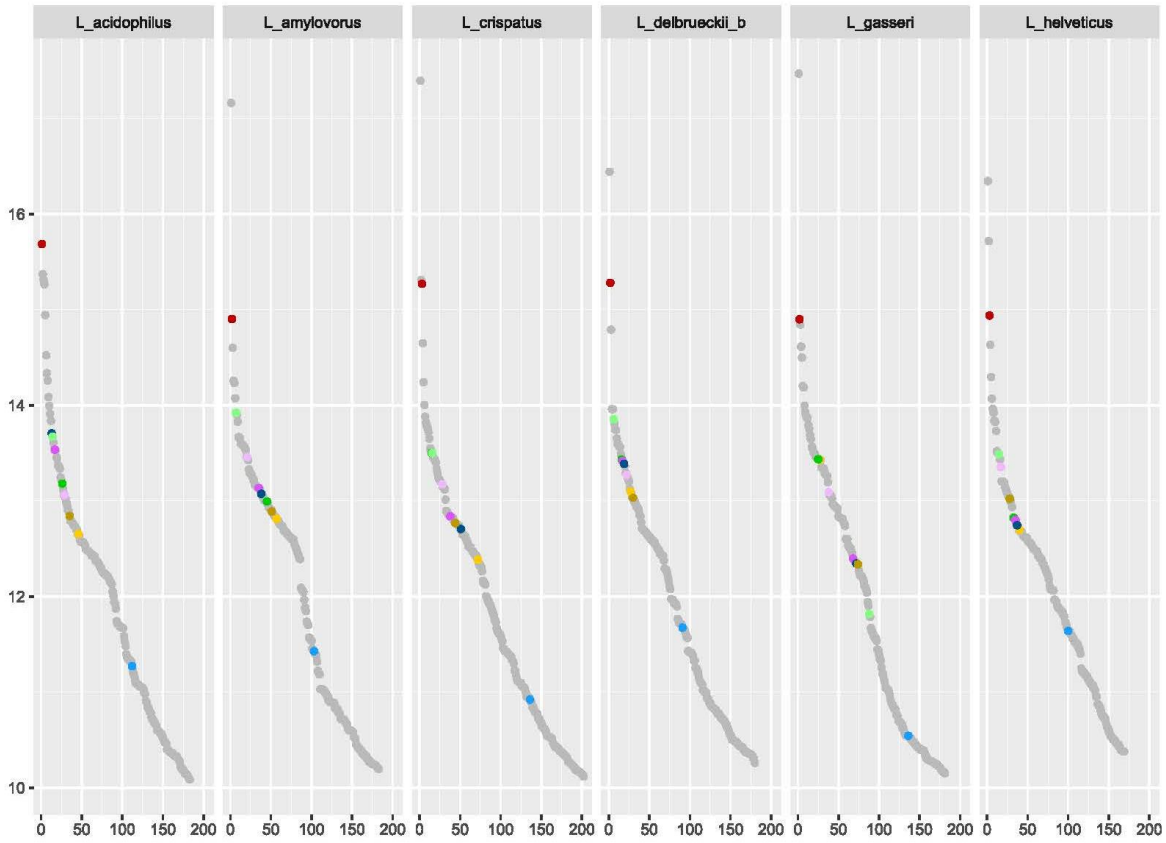


Fig. 3.4 | Ranked Order of mRNA Expression. Top 10% most highly expressed genes in *L. acidophilus, L. amylovorus, L. crispatus, L. delbrueckii_b, L. gasseri, and L. helveticus.* Data is represented as a log2 transformed RPKM. Transcripts are ranked from most abundant to least abundant. Glycolysis genes are coloured as follows: *pgm* in red, *pgi* in blue, *pfk* in yellow, *fba* in dark green, *tpi* in purple, *gap* in maroon, *pgk* in navy, *gpm* in mustard, *eno* in light green, and *pyk* in lavender.

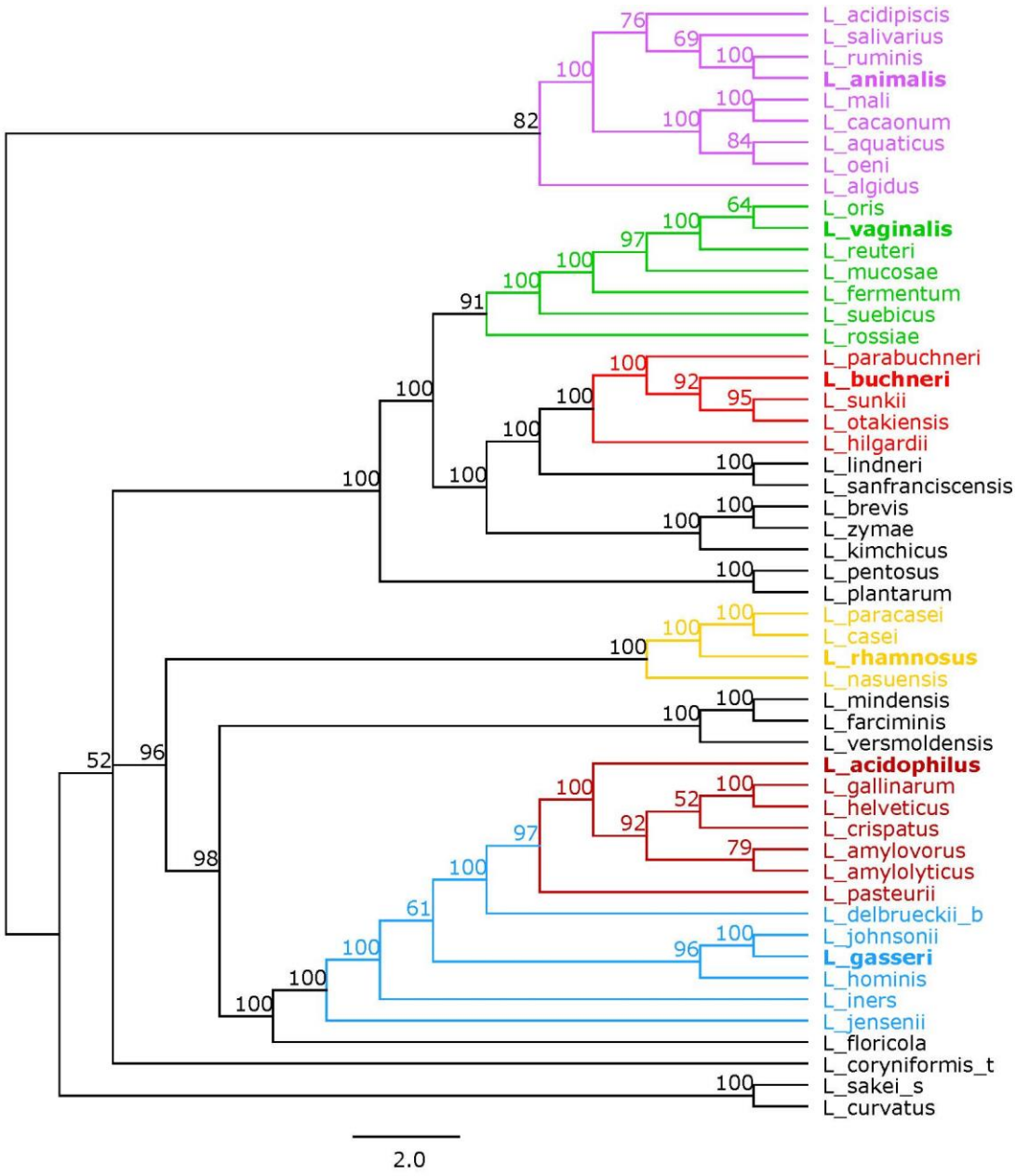


Fig. 3.5 | Concatenated Glycolysis Tree. Tree based off of the alignment of concatenated amino acid sequences of glycolysis enzymes using RaxML. Bootstrap values are recorded on the nodes. Groups are coloured as follows: the *L. animalis* group in purple, the *L. vaginalis* group in green, *L. buchneri* group in red, the *L. rhamnosus* group in yellow, the *L. acidophilus* group in maroon, the *L. gasseri* group in blue. The representative species in each group is in bold. Species name follows the naming convention in Table 3.1.

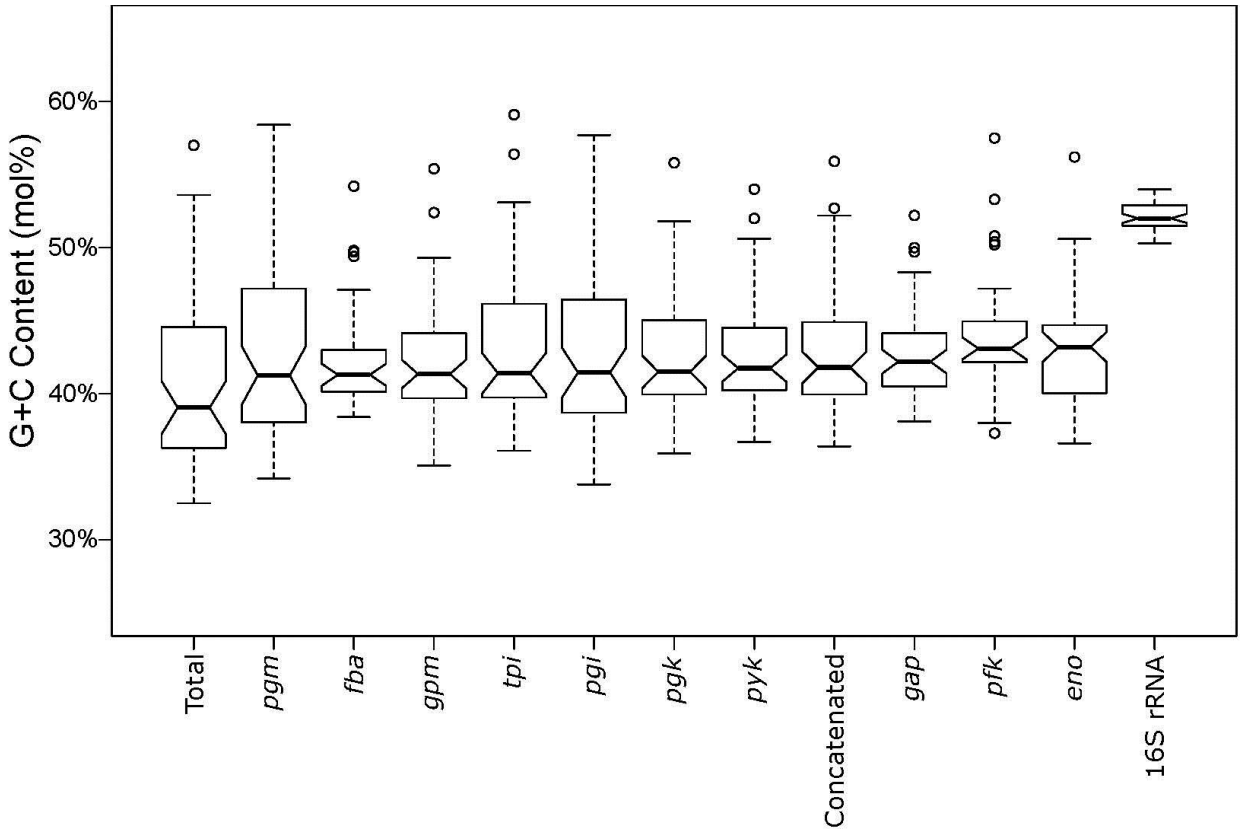


Fig. 3.6 | G+C mol% analysis of *Lactobacillus* Glycolysis Genes. Depicted are notched boxplots of G+C mol% for each glycolysis gene, concatenated genes, 16S rRNA, and total genome. Genes are placed in order of increasing median. If two notches do not overlap, it is an indication of strong evidence for differing medians.

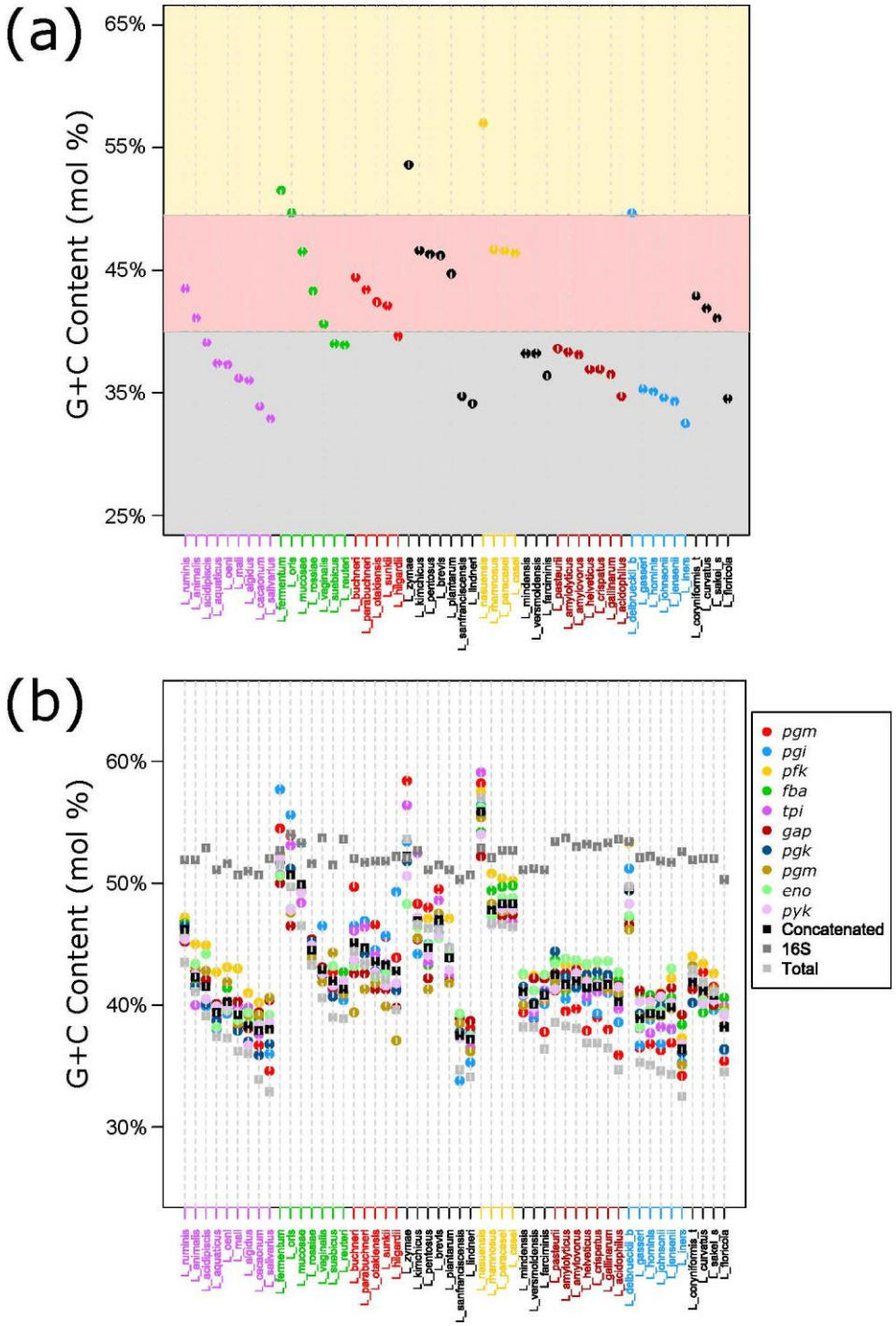


Fig. 3.7 | G+C mol% analysis of *Lactobacillus* Genomes. (a) shows the total G+C mol% for each species. Species are coloured according to their phylogenetic group. (b) shows the G+C mol% of the glycolysis genes, the concatenated glycolysis genes, the 16S, and total G+C mol% for each species. Species are named according to Table 3.1.

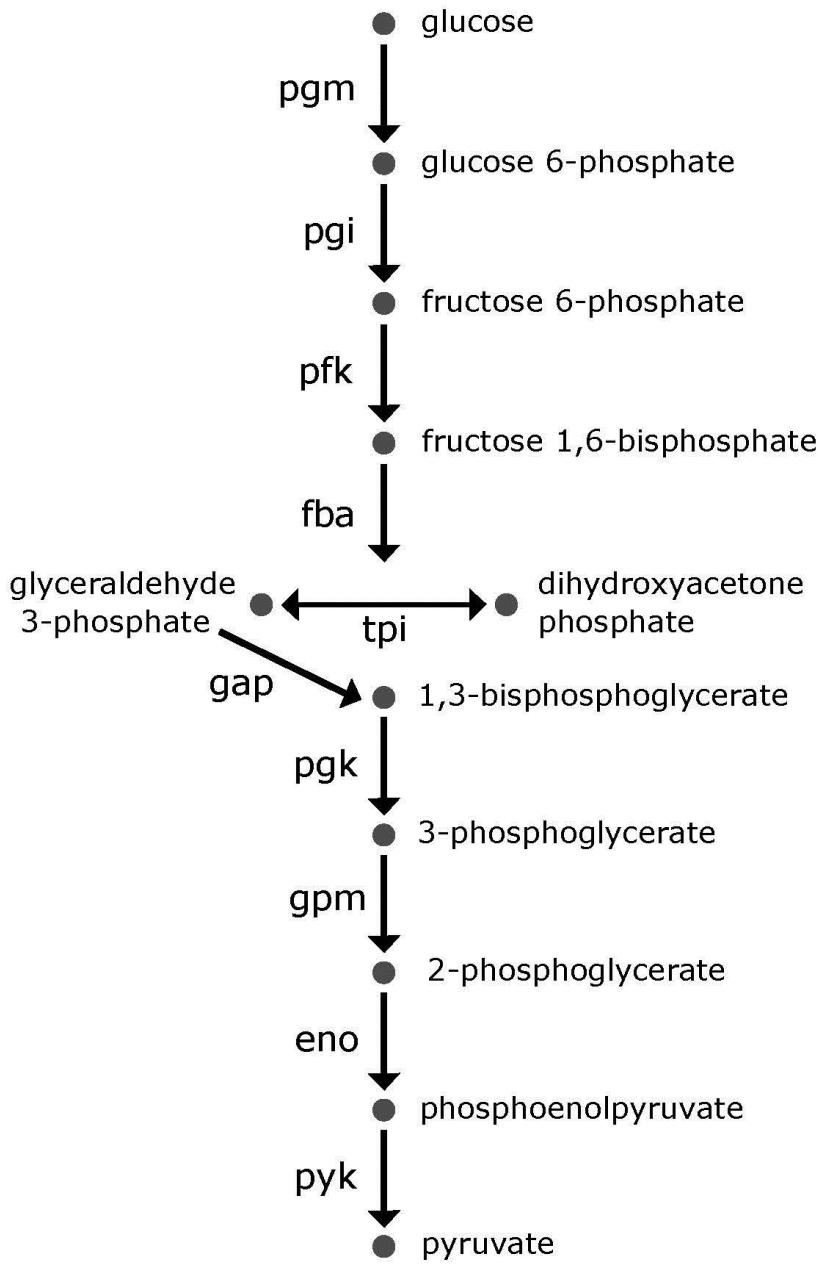


Fig. 3.S1 | Glycolysis Pathway in *Lactobacillus*. Depicted is the glycolysis pathway. Products are to the right of the dots and the enzymes used in this study are to the left of the arrows. Enzyme abbreviations can be found in Table 3.S1.

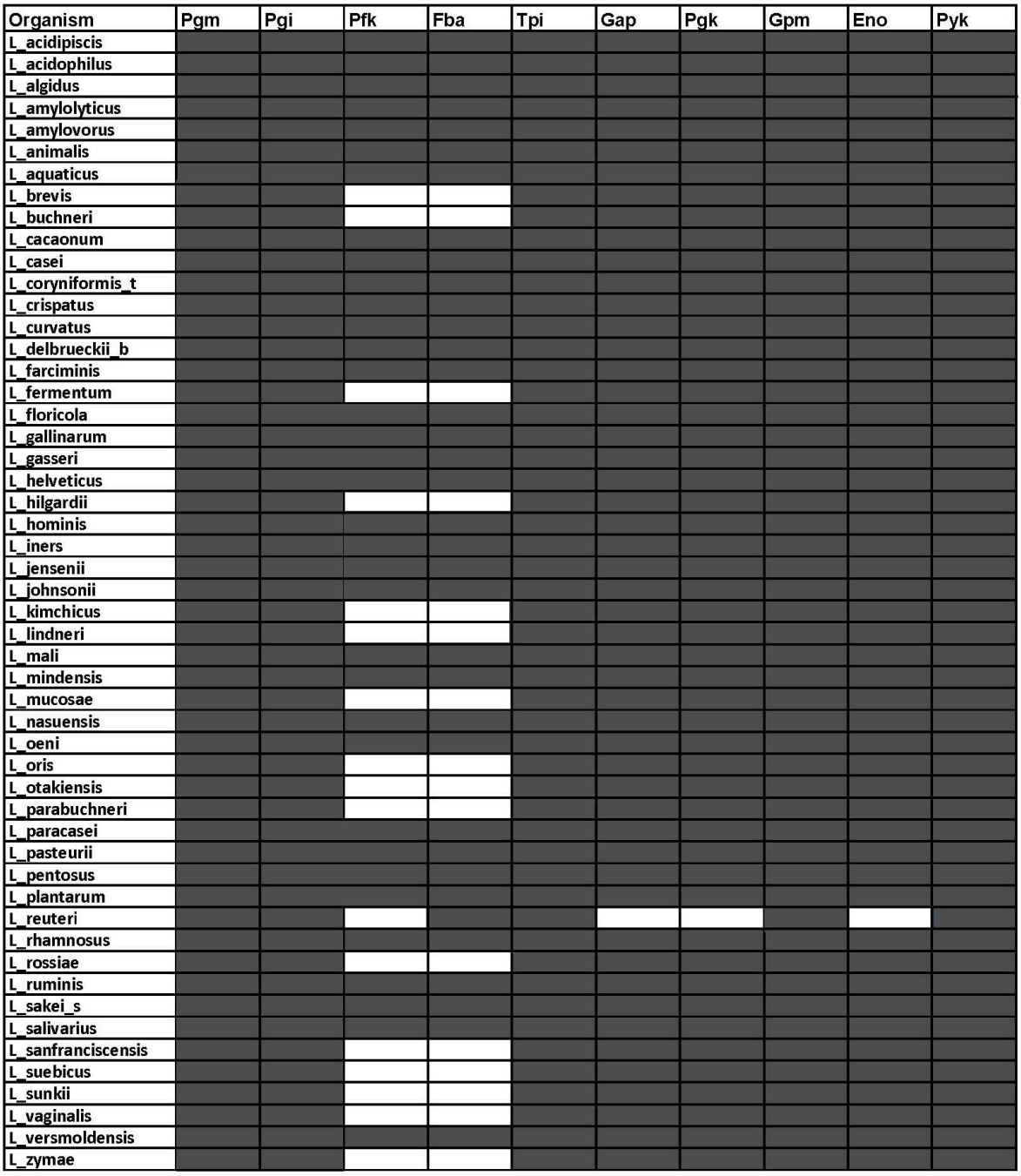


Fig. 3.S2 | Glycolysis Presence in *Lactobacillus*. Shows the presence of each glycolysis gene in each species of this study. Absence is reflected by a white box. Species follow the naming convention from Table 3.1. Gene abbreviations can be found in Table 3.S1.

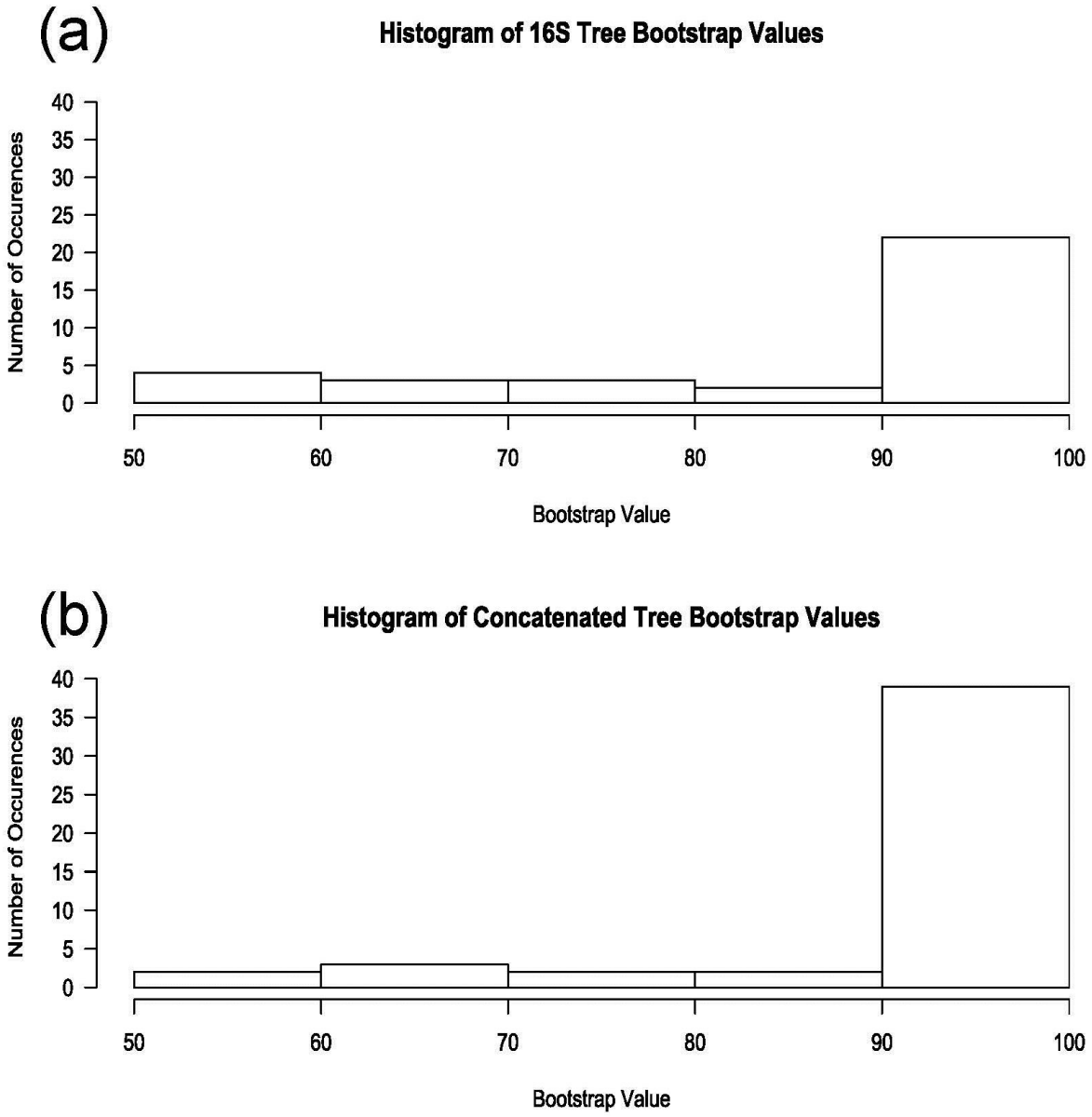


Fig. 3.S3 | Histogram of Bootstrap Values. (a) shows the bootstrap values for the 16S based trees. (b) shows the bootstrap values for the Glycolysis-based Concatenated Tree.

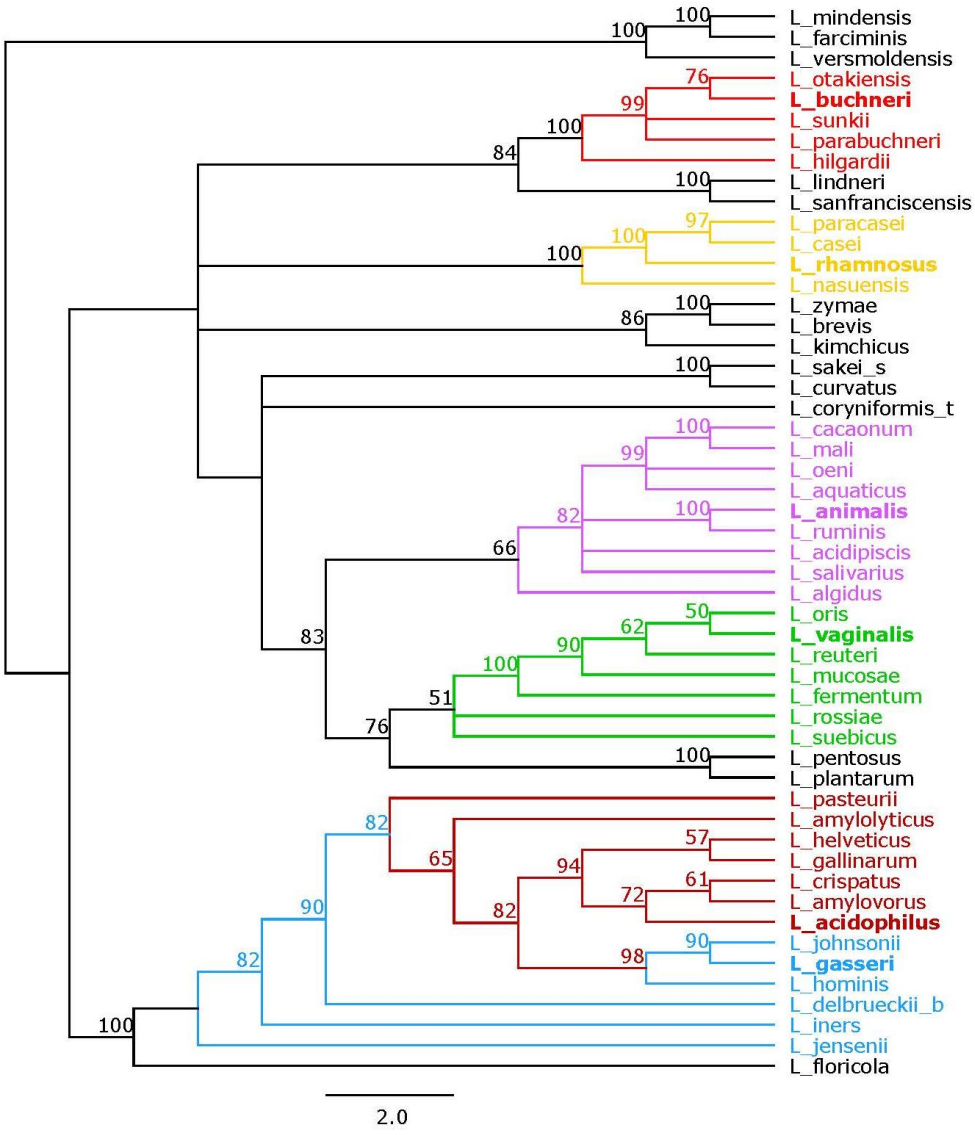


Fig. 3.S4 | Pgm Tree. Tree based off of the alignment of amino acid sequences of Pgm using RaxML. Bootstrap values are recorded on the nodes. Groups are coloured as follows: the *L. animalis* group in purple, the *L. vaginalis* group in green, *L. buchneri* group in red, the *L. rhamnosus* group in yellow, the *L. acidophilus* group in maroon, the *L. gasseri* group in blue. The representative species in each group is in bold. Species name follows the naming convention in Table 3.1.

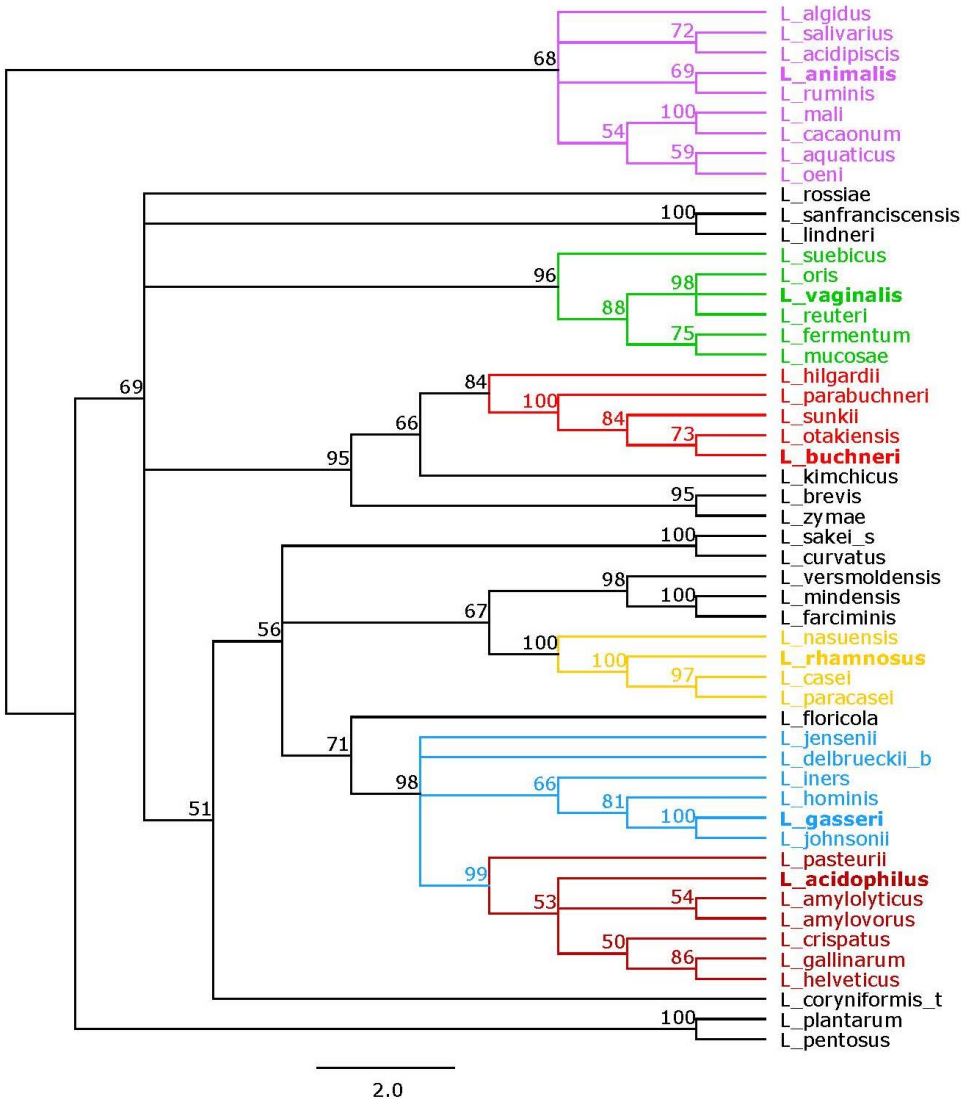


Fig. 3.S5 | Pgi Tree. Tree based off of the alignment of amino acid sequences of Pgi using RaxML. Bootstrap values are recorded on the nodes. Groups are coloured as follows: the *L. animalis* group in purple, the *L. vaginalis* group in green, *L. buchneri* group in red, the *L. rhamnosus* group in yellow, the *L. acidophilus* group in maroon, the *L. gasseri* group in blue. The representative species in each group is in bold. Species name follows the naming convention in Table 3.1.

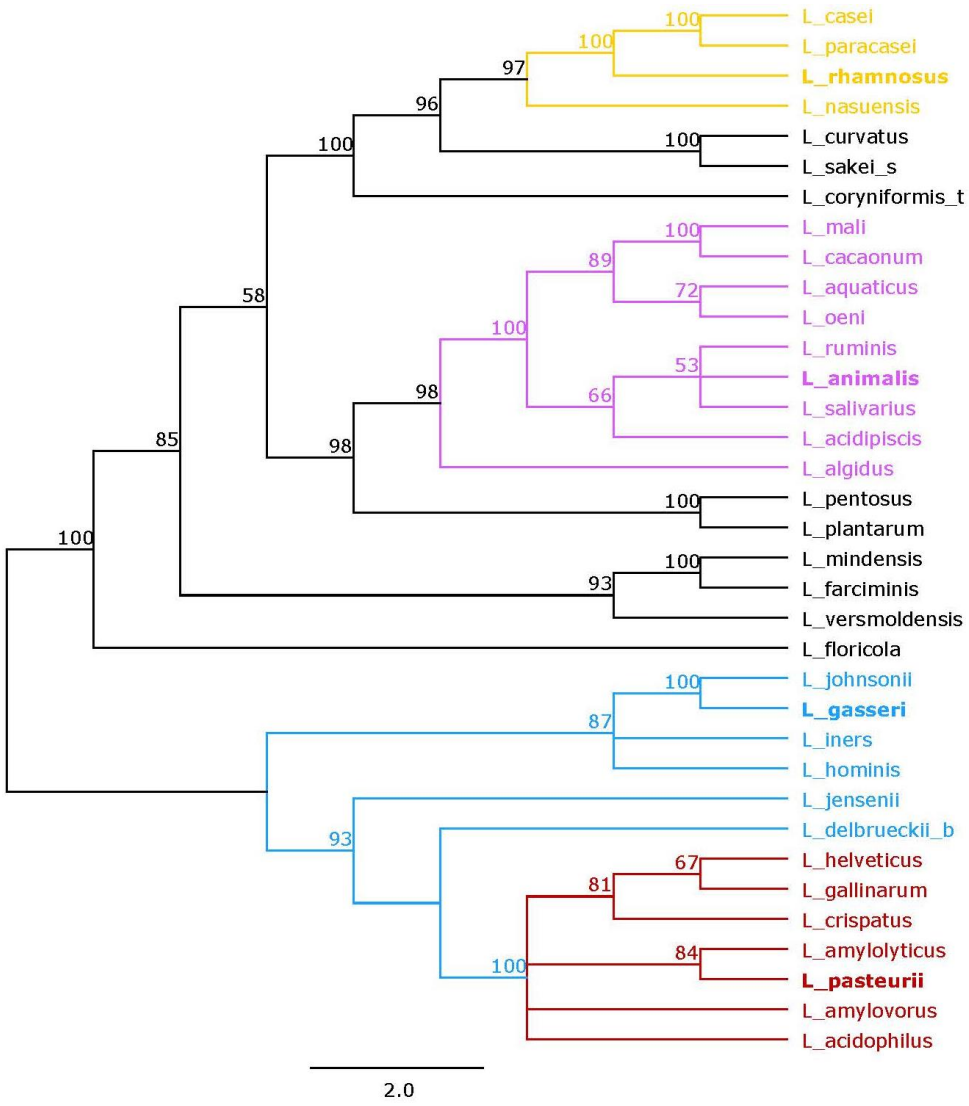


Fig. 3.S6 | Pfk Tree. Tree based off of the alignment of amino acid sequences of Pfk using RaxML. Bootstrap values are recorded on the nodes. Groups are coloured as follows: the *L. animalis* group in purple, the *L. rhamnosus* group in yellow, the *L. acidophilus* group in maroon, the *L. gasseri* group in blue. The representative species in each group is in bold. Species name follows the naming convention in Table 3.1.

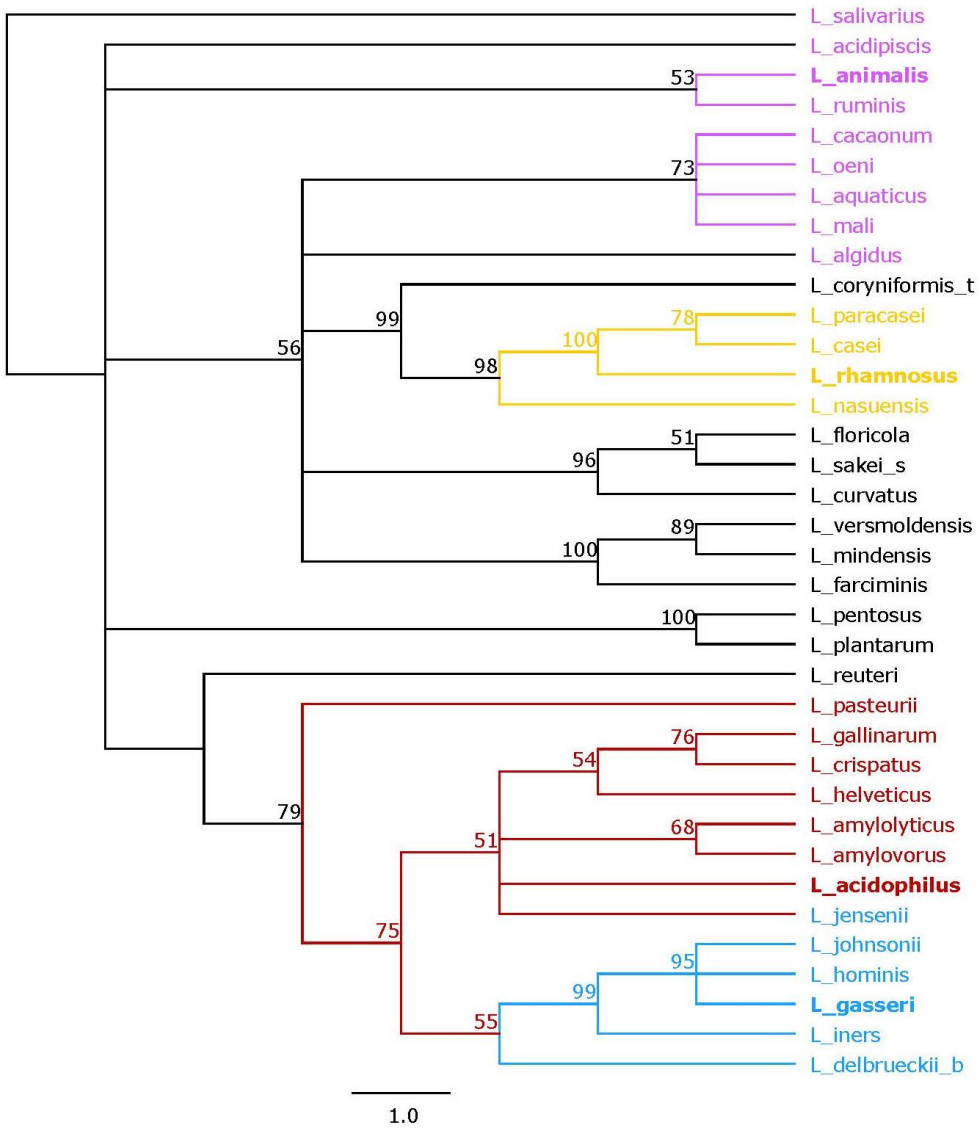


Fig. 3.S7 | Fba Tree. Tree based off of the alignment of amino acid sequences of Fba using RaxML. Bootstrap values are recorded on the nodes. Groups are coloured as follows: the *L. animalis* group in purple, the *L. rhamnosus* group in yellow, the *L. acidophilus* group in maroon, the *L. gasseri* group in blue. The representative species in each group is in bold. Species name follows the naming convention in Table 3.1.

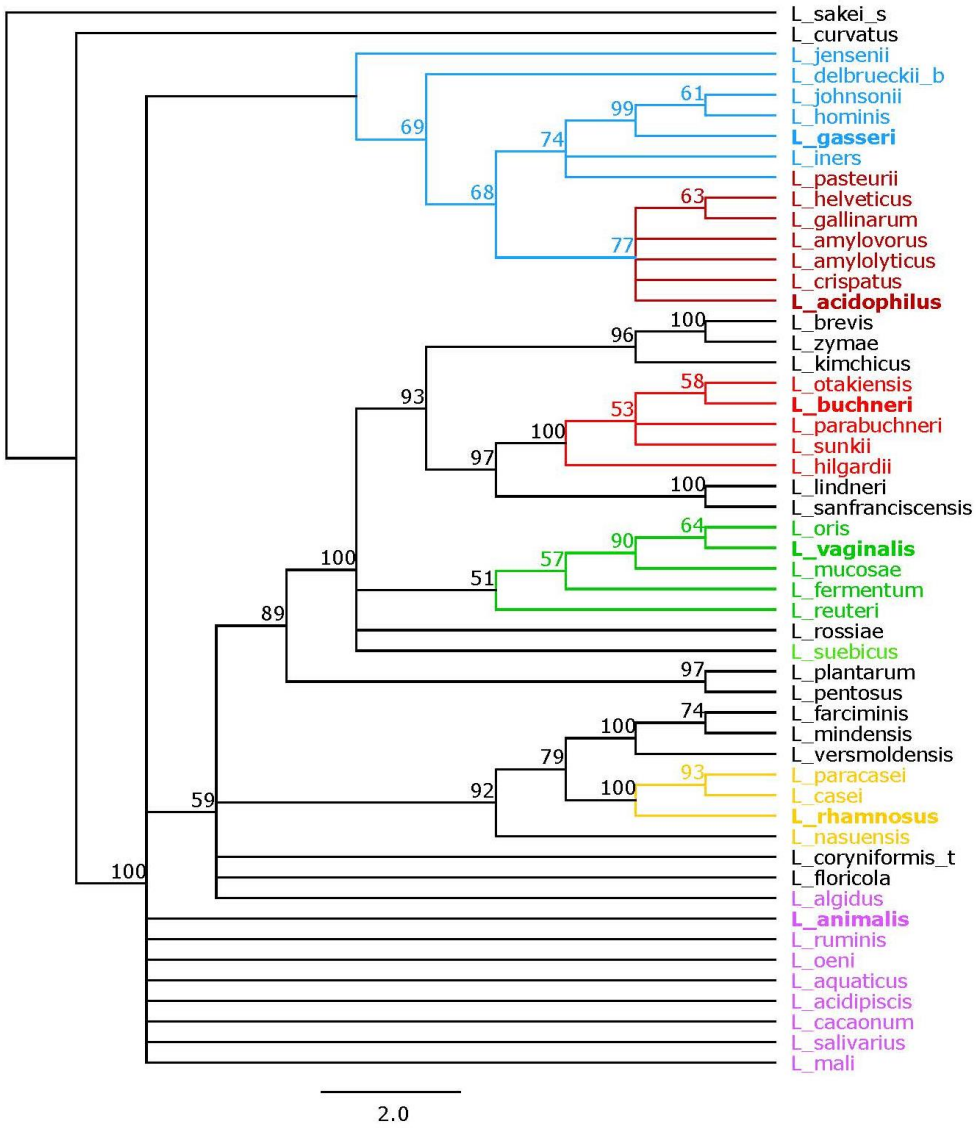


Fig. 3.S8 | Tpi Tree. Tree based off of the alignment of amino acid sequences of Tpi using RaxML. Bootstrap values are recorded on the nodes. Groups are coloured as follows: the *L. animalis* group in purple, the *L. vaginalis* group in green, *L. buchneri* group in red, the *L. rhamnosus* group in yellow, the *L. acidophilus* group in maroon, the *L. gasseri* group in blue. The representative species in each group is in bold. Species name follows the naming convention in Table 3.1.

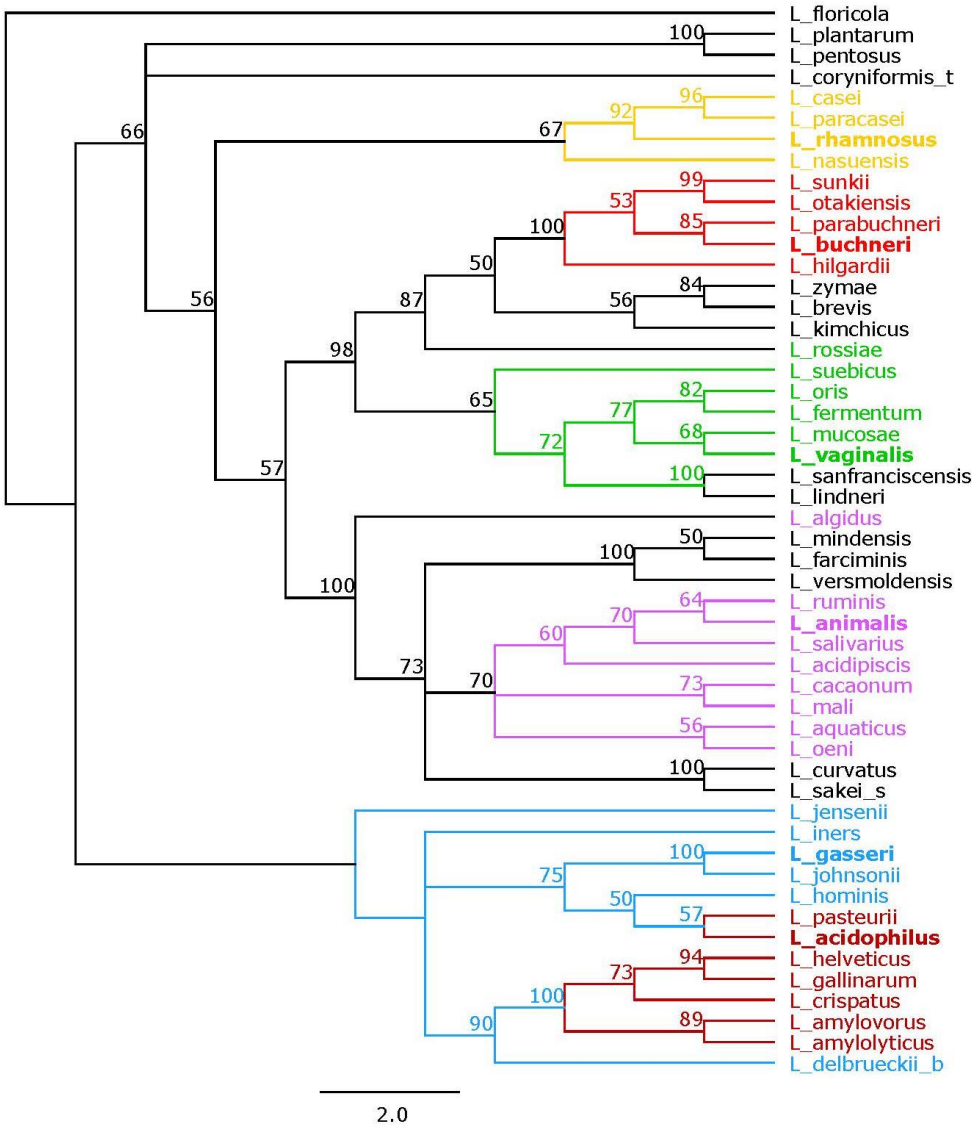


Fig. 3.S9 | Gap Tree. Tree based off of the alignment of amino acid sequences of Gap using RaxML. Bootstrap values are recorded on the nodes. Groups are coloured as follows: the *L. animalis* group in purple, the *L. vaginalis* group in green, *L. buchneri* group in red, the *L. rhamnosus* group in yellow, the *L. acidophilus* group in maroon, the *L. gasseri* group in blue. The representative species in each group is in bold. Species name follows the naming convention in Table 3.1.

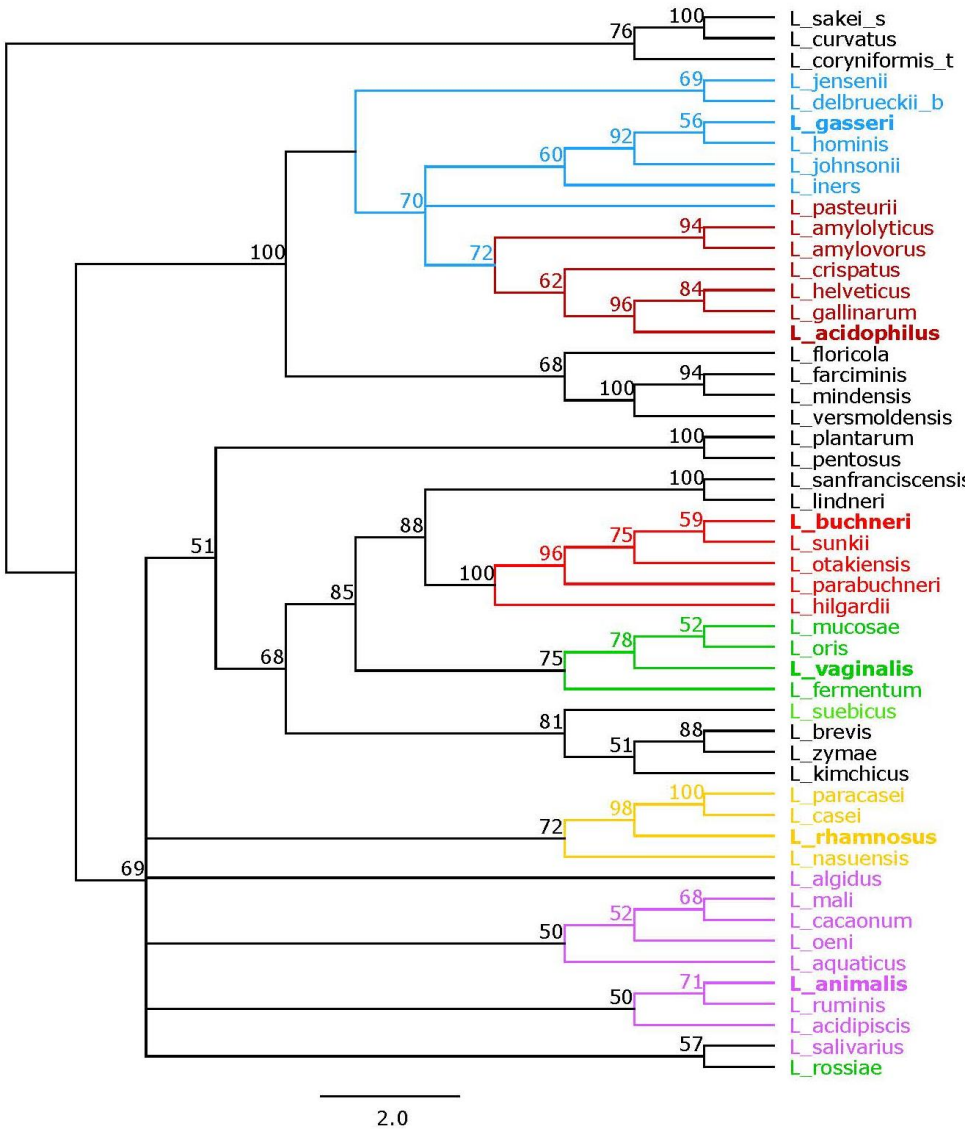


Fig. 3.S10 | Pgk Tree. Tree based off of the alignment of amino acid sequences of Pgk using RaxML. Bootstrap values are recorded on the nodes. Groups are coloured as follows: the *L. animalis* group in purple, the *L. vaginalis* group in green, *L. buchneri* group in red, the *L. rhamnosus* group in yellow, the *L. acidophilus* group in maroon, the *L. gasseri* group in blue. The representative species in each group is in bold. Species name follows the naming convention in Table 3.1.

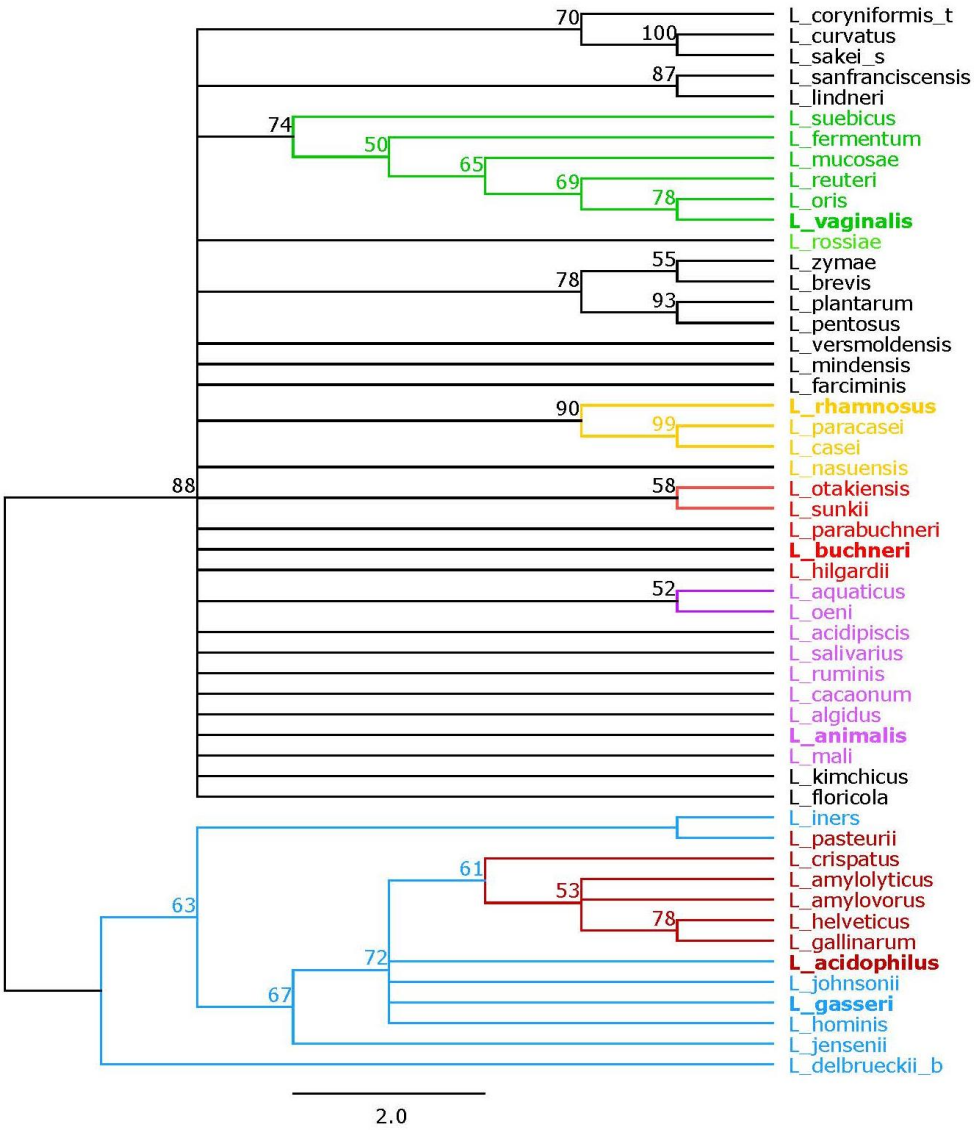


Fig. 3.S11 | Gpm Tree. Tree based off of the alignment of amino acid sequences of Gpm using RaxML. Bootstrap values are recorded on the nodes. Groups are coloured as follows: the *L. animalis* group in purple, the *L. vaginalis* group in green, *L. buchneri* group in red, the *L. rhamnosus* group in yellow, the *L. acidophilus* group in maroon, the *L. gasseri* group in blue. The representative species in each group is in bold. Species name follows the naming convention in Table 3.1.

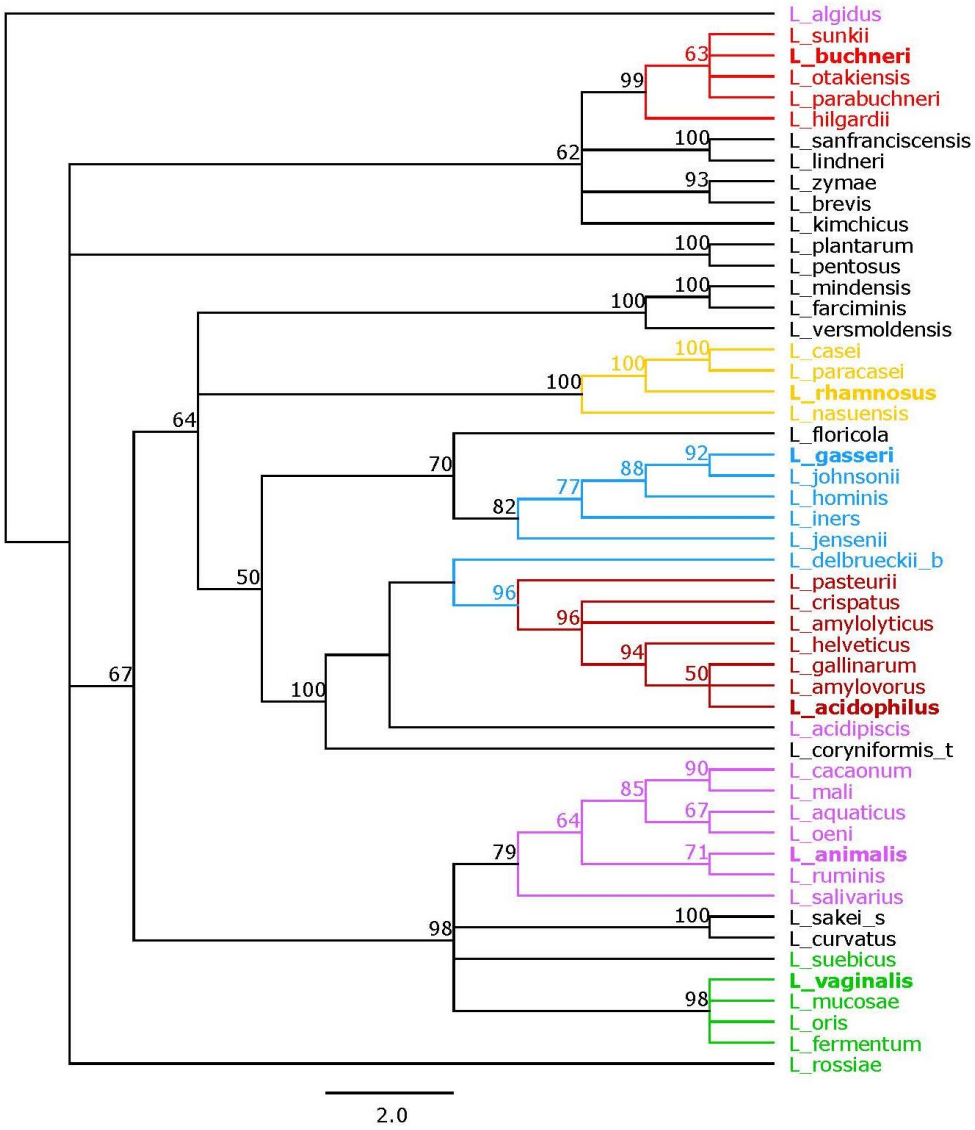


Fig. 3.S12 | Eno Tree. Tree based off of the alignment of amino acid sequences of Eno using RaxML. Bootstrap values are recorded on the nodes. Groups are coloured as follows: the *L. animalis* group in purple, the *L. vaginalis* group in green, *L. buchneri* group in red, the *L. rhamnosus* group in yellow, the *L. acidophilus* group in maroon, the *L. gasseri* group in blue. The representative species in each group is in bold. Species name follows the naming convention in Table 3.1.

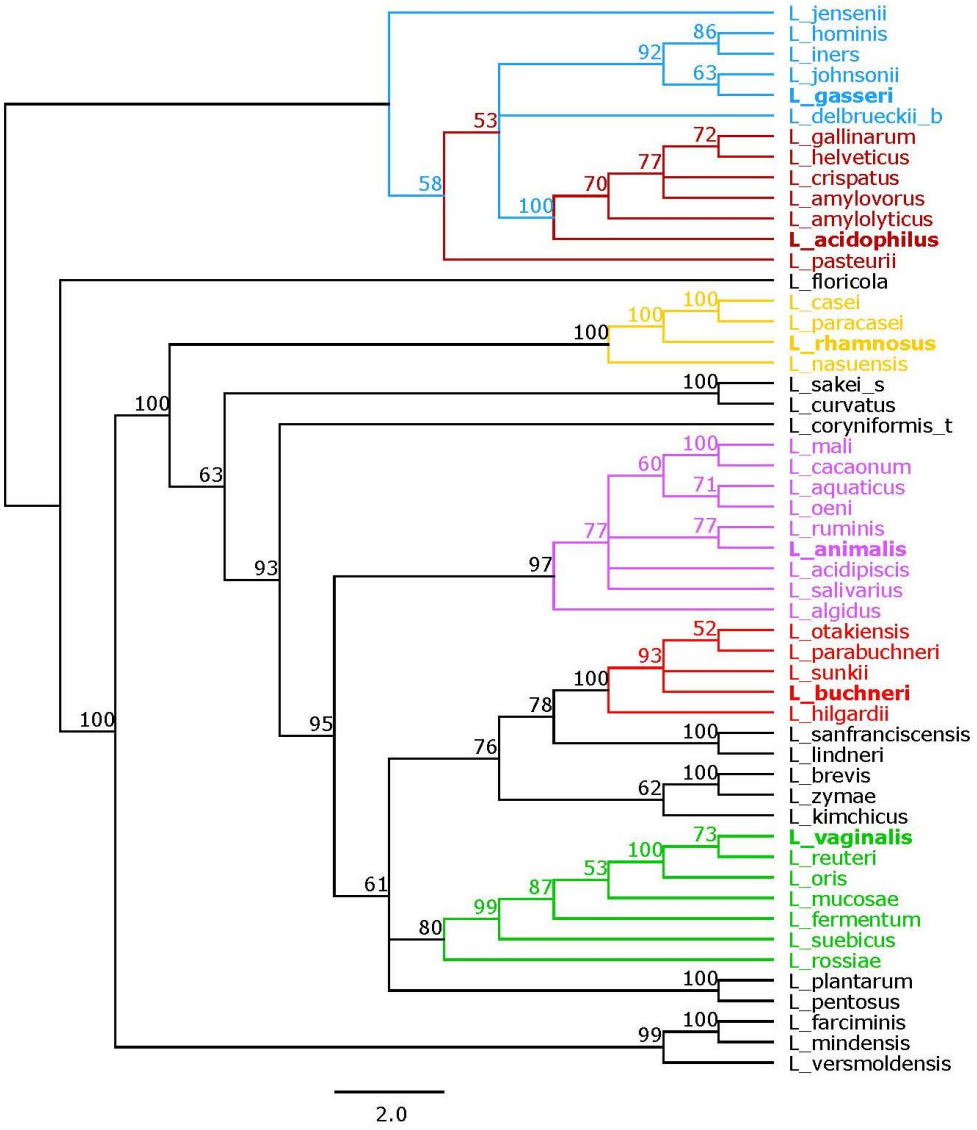


Fig. 3.S13 | Pyk Tree. Tree based off of the alignment of amino acid sequences of Pyk using RaxML. Bootstrap values are recorded on the nodes. Groups are coloured as follows: the *L. animalis* group in purple, the *L. vaginalis* group in green, *L. buchneri* group in red, the *L. rhamnosus* group in yellow, the *L. acidophilus* group in maroon, the *L. gasseri* group in blue. The representative species in each group is in bold. Species name follows the naming convention in Table 3.1.

CHAPTER 4: FUNCTIONAL AND GENOMIC CHARACTERIZATION OF *LACTOBACILLUS FERMENTUM* ATCC 14931

4.1. CONTRIBUTION TO THE WORK

The following is a manuscript in preparation titled "Functional and Genomic Characterization of *Lactobacillus fermentum* ATCC 14931" with the following authorship:

Katelyn Brandt[1,2], Matthew A. Nethery[1,2], Sarah O'Flaherty[2], and Rodolphe Barrangou[1,2]

[1]Functional Genomics Graduate Program, North Carolina State University, Raleigh, NC 27695

[2]Department of Food, Bioprocessing and Nutrition Sciences, North Carolina State University, Raleigh, NC 27695

Katelyn Brandt is first author on this publication. She carried out experiments and analyzed data. Katelyn also wrote the manuscript.

4.2. ABSTRACT

Lactobacillus fermentum, a member of the lactic acid bacteria complex, has recently garnered increased attention due to documented antimicrobial activity and interest in assessing the probiotic potential of select strains that may provide human health benefits. Here, we genomically and functionally characterize the *L. fermentum* type strain ATCC 14931 as a canonical representative of this species. Functionally, we determined carbohydrate utilization, morphology, and stress tolerance. Genomically, we determined the polished whole genome sequence of this type strain and compared it to 37 available genome sequences within this species. Results reveal genetic diversity across nine clades, with variable content encompassing mobile genetic elements, CRISPR-Cas immune systems and genomic islands, as well as numerous genome rearrangement. Interestingly, we determined a high frequency of occurrence of diverse Type I, II, and III CRISPR-Cas systems in 72% of the genomes, with a high level of strain hypervariability.

4.3. INTRODUCTION

Lactobacillus are low-GC, microaerophilic, Gram-positive microorganisms that are members of the lactic acid bacteria (LAB) group (Pot et al., 1994). They are considered ubiquitous in nature and many species and strains benefit from Generally Regarded as Safe (GRAS) or Qualified Presumption of Safety (QPS) status (Bernardeau et al., 2008). They have had a large impact on the food manufacturing, human health, and biotechnology industries. Their ability to spontaneously ferment foods and produce lactic acid has ingratiated lactobacilli into the food manufacturing process, specifically as starter cultures to produce yogurt, cheese, and fermented vegetables (Pfeiler and Klaenhammer, 2007). Several strains of *Lactobacillus* are used as probiotics, meaning they are "live microorganisms which when administered in adequate amounts confer a health benefit on the host" (Hill et al., 2014; Lebeer et al., 2018). Several species are widely studied and utilized, such as *Lactobacillus acidophilus, Lactobacillus gasseri*, and *Lactobacillus delbrueckii.* Additionally, *Lactobacillus* serves as a valuable source of clustered regularly interspaced short palindromic repeats (CRISPR) and associated proteins (Cas), which may be repurposed for a diversity of applications, including the development of genome editing tools (Sun et al., 2015). Recently, there has been an increased interest in assessing the potential of various *Lactobacillus* species and strains for the development of new functional foods, biotechnology tools, and next-generation probiotics. *Lactobacillus fermentum* is one such candidate species being examined for its potential use.

A survey of metagenomic study data using Integrated Microbial Next Generation Sequencing (IMNGS) revealed that the most common metagenomes for *L. fermentum* are fermentation and human gut metagenomes, implying use or effectiveness in food manufacturing and human health. Various studies over the years have looked at the ability of *L. fermentum* to

serve as a potential probiotic or biotechnology tool beyond its current uses in food manufacturing. *L. fermentum* is known for its biofilm formation phenotype and has been studied as a potential biosurfactant in numerous capacities, including for the sterilization of surgical implants (Velraeds et al., 1996; Gan et al., 2002). Some strains of *L. fermentum* have been shown to inhibit pathogens through the production of bacteriocins and antifungal metabolites (Varma et al., 2010; Ghazvini et al., 2016). This, combined with the ability to survive bile salts and lower cholesterol, suggests that some *L. fermentum* strains may have some potential for probiotic applications (Pereira and Gibson, 2002; Pereira et al., 2003). In fact, two *L. fermentum* strains, ME-3 and CECT 5716, have been characterized and both are considered probiotic. *L. fermentum* ME-3 has antioxidant properties as well as having demonstrated antimicrobial capabilities against Gram-negative organisms, such as *Enterococci* and *Staphylococcus aureus* (Mikelsaar and Zilmer, 2009). *L. fermentum* CECT 5716 has the ability to modulate immune responses of host organisms (Diaz-Ropero et al., 2007).

Despite the interest in *L. fermentum*, there have been relatively few studies overall, especially regarding the type strain ATCC 14931 (DSM 20052), with which there is limited knowledge with regards to species genetic and genomic diversity. One study compared five *L. fermentum* strains, but did not include the type strain (Yoo et al., 2017). In order to fully leverage the potential of *L. fermentum*, we should first assess functional and genetic species diversity and identify strains of interest. In this study, we characterized the type strain ATCC 14931 and provided comparative genomic analyses against 37 strains in this species.

4.4. MATERIALS AND METHODS

4.4.1. Growth Conditions

The *L. fermentum* type strain ATCC 14931 was used in functional characterization of carbohydrate utilization, morphology, and stress tolerance. The strain was sourced from DSMZ (20052, Germany). Carbohydrate utilization was tested using the API strip according to the manufacturer's instructions (bioMérieux). For growth curves, overnight cultures were inoculated 1% (v/v) into a 96-well microplate in triplicate. Each well contained 200 µL of SDM (Semi-Defined media (Kimmel and Roberts, 1998)) with 1% carbohydrate. Plates were then sealed with a clear Thermalseal film. Plates were read at 37°C in a Floustar Optima microplate reader (BMG Labtech, Cary, NC) at 600nm for 48 hours.

For scanning electron microscopy, *L. fermentum* ATCC 14931 was grown overnight in MRS (de Man, Rogosa, Sharpe) broth. Cells from 10mL of culture were harvested by centrifugation (20 min, 5,000 rpm), then resuspended in 10mL of 3% glutaraldehyde in 0.1M Na cacodylate buffer (pH 5.5). Suspensions were then filtered using a 0.4 µM pore polycarbonate Nucleopore filter. Filters were washed with 0.1M Na cacodylate buffer (pH 5.5) three times in 30-minute changes. Cells were then dehydrated through a graded series of ethanol to 100% ethanol and were critical-point dried in liquid $C0_2$. Filters were then mounted on stubs with tape and silver paint and were then sputter coated with 50 Å Au/Pd. Samples were viewed using a JEOL JSM-5900LV scanning electron microscope at the Center for Electron Microscopy at North Carolina State University, USA.

Simulated intestinal and gastric juices were prepared as previously described (Charteris et al., 1998; Frece et al., 2005). Overnight cultures were pelleted, washed two times, and resuspended in distilled water. Cells (0.2 mL) were then added to 1 mL of freshly made

simulated gastric or intestinal juice and incubated at 37°C. Viable cell counts were determined by plating on MRS agar at regular intervals.

4.4.2. Genome Sequencing

Long and short reads were generated for *L. fermentum* ATCC 14931. PacBio sequencing was performed by RTL Genomics (Texas, US). DNA was extracted using Qiagen's MagAttract HMW DNA Kit with the following modifications: sample was incubated at 37°C, shaking (900 RPM) overnight with the addition of lysozyme and mutalysin, then eluted with AE. Quality check was performed using dsDNA Broad Range DNA kit on the Qubit Fluorometer 3.0 and Fragment Analyzer by Advanced Analytical Technologies with the High Sensitivity Large Fragment 50KB Analysis kit. Library preparation was performed from SMRTbell Libraries using PacBio Barcoded Adapters for Multiples SMRT sequencing with the following modifications: samples were pooled equimolar, 500 ng per sample of DNA were used, ligation was overnight, and final elution was 12 µL EB. dsDNA High Sensitivity DNA kit on the Qubit Fluorometer 3.0 and Fragment Analyzer using High Sensitivity Large Fragment 50KB Analysis Kit were used to perform library QC. Library preparation for sequencing was performed following PacBio's protocol with a pre-extension time of 120 minutes and final loading of 6pM. Short reads were generated by CoreBiome, Inc. (MN, USA). DNA was extracted using Qiagen's MO Bio PowerFecal for high throughput on QiaCube with bead beating in 0.1mm glass bead plates. Invitrogen's Qiant-iT Picogreen dsDNA Assay was used to quantify DNA. Library preparation was completed using an adapted procedure from Illumina's Nextera Library Prep Kit. Sequencing took place on an Illumina NextSeq using paired-end 2 x 150 reads and Illumina's NextSeq 500/550 High Output V2 kit. Sequence Quality Control was set to filter a Q-Score <20

and length <50; cutadapt (v.1.15) was used to trim adapter sequences. SPAdes (v3.11.0) was used to assemble contigs and QUAST (v4.5) analysis was performed on contigs greater than 1,000 bases. Short and long reads were then combined using Unicycler with default options. Remaining contigs were then hand-curated and joined using primer walking. The genome sequence was annotated using Rapid Annotations Subsystems Technology (RAST) (Aziz et al., 2008). Genomic features can be found in Table 4.1. COG (Clusters of Orthologous Groups) annotations were determined using eggnog-mapper, based on eggnog 4.5 data (Huerta-Cepas et al., 2015; Huerta-Cepas et al., 2017). Antibiotic resistances genes were searched for using ResFinder v 3.1 (Aarestrup et al., 2012).

4.4.3. Comparative Analyses

Thirty-eight *L. fermentum* strains were selected for phylogenetic analyses (Table 4.1). A phylogeny was developed using the glycolysis gene phosphoglucomutase as its basis, following a previously proposed methodology (Brandt and Barrangou, 2018). After extracting the phosphoglucomutase gene sequence, nucleotide sequences were aligned using MUSCLE (maximum iteration was eight) (Edgar, 2004). Trees were then generated using RAxML (CAT GTR, Bootstrap using rapid hill climbing with random seed 1, and 100 replicates) (Stamatakis, 2014). A consensus tree was generated using a 50% threshold. Metadata was added to the cladogram using CLC Genomics (https://www.qiagenbioinformatics.com/).

Seven *L. fermentum* genomes were used for whole genome comparisons (ATCC 14931, LT906621, NZ_AP017973, NZ_CP019030, NZ_CP021790, NC_021235, and NC_017465). A BRIG image was generated using BRIG 0.95, following parameters outlined in the manual

(Alikhan et al., 2011). A MAUVE alignment using all complete genomes was generated using default settings (Darling et al., 2004).

4.4.4. Identification and Annotation of CRISPR-Cas Systems

Potential CRISPR loci were identified in 38 *L. fermentum* strains using the CRISPR recognition tool (CRT) (Bland et al., 2007). Each predicted CRISPR-Cas system was then hand-curated for integrity, content, and assigned a type. Spacer visualization was achieved using CRISPRViz with standard options (Nethery and Barrangou, 2018). mRNA and smRNA were used to analyze transcriptional profiles of the CRISPR loci in ATCC 14931. Cells were grown to mid-log phase and flash-frozen. Total RNA was extracted using Zymo Direct-Zol Miniprep kit (Zymo Research, Irvine, CA) according to a previously described protocol (Theilmann et al., 2017). Library preparation and sequencing were performed by the Roy J. Carver Biotechnology Centre from the University of Illinois (Urbana-Champaign, IL) using an Illumina HiSeq2500. Data was uploaded into Geneious (v. 11.1.5, https://www.geneious.com). Reads were then processed by trimming to an error probability limit of 0.001 and filtered to exclude reads less than 10 nt (smRNA) or a range of 28-150 nt (mRNA). Reads were mapped to the reference genome using Bowtie2 (Langdon, 2015). *trans*-activating-crRNA (tracrRNA) prediction was performed as previously described (Briner et al., 2016). Briefly, we searched for the five modules of tracrRNA and the terminal GC-rich hairpins. Protospacer adjacent motif (PAM) prediction was carried out as previously described (Crawley et al., 2018). Briefly, protospacer hits were determined by BLASTing spacers against publicly available datasets. The flanking regions of positive hits were then used to identify sequence motifs.

4.5. RESULTS

4.5.1. Microbial Phenotypic Characterization

We began by characterizing the growth and survival of the type strain for *L. fermentum*. First, carbohydrate utilization was determined using API strips. Of the 49 sugars tested, *L. fermentum* ATCC 14931 utilized eleven. These sugars were: D-ribose, D-galactose, D-glucose, D-fructose, esculin ferric citrate, D-maltose, D-lactose, D-melibiose, D-saccharose, D-raffinose, and potassium gluconate. Growth curves were performed to confirm growth for galactose, fructose, and D-xylose (Figure 4.1A). These curves showed similar growth for galactose and fructose, and little growth was observed in D-xylose, as indicated by the API strips. We then examined cell morphology, due to the presence of the exopolysaccharide (*eps*) locus in the genome. *L. fermentum* ATCC 14931 colonies were of the ropy phenotype, an indication of EPS formation (data not shown). Further evidence of a potential EPS layer can be visualized in scanning electron microscopy photos (Figure 4.1D, E, and F). Cellular material are shown between clumps of cells (Figure 4.1E) and extending from single cells (Figure 4.1D and F).

Finally, we tested stress tolerance. *L. fermentum* ATCC 14931 was exposed to simulated intestinal juice and simulated gastric juice (Figure 4.1B and C, respectively). After two hours in simulated intestinal juice, there was 49% survival. This continued to decrease until hour six where there was 15% survival and remained steady into hour eight. In simulated gastric juice, *L. fermentum* ATCC 14931 showed 70-75% survival for the first 30 minutes after exposure. By the one-hour point, there was no survival in simulated gastric juice. Growth in media supplemented with oxgall or porcine bile was minimal (data not shown).

4.5.2. Complete Genome Sequence of *L. fermentum* ATCC 14931

A draft genome for *L. fermentum* ATCC 14931 was previously deposited at NCBI in 2009 and updated in 2017 as NZ_ACGI, which contained 74 contigs. We polished and completed the genome sequence and generated a single contig (1.89Mb). The genomic traits for *L. fermentum* ATCC 14931 can be found in Table 4.1. The genome size is 1.89Mb with a GC content of 52.5%. We identified no plasmids in *L. fermentum* ATCC 14931. Next, we annotated the genome using RAST, which identified 1,900 coding sequences and 73 RNAs (15 rRNA and 58 tRNA). We then applied COG identifiers to the genome sequence using EggNOG. Of the 1,900 coding sequences, 1,237 were given a COG designation. The largest COG group was the [S] group (15% of assigned coding sequences), or the unknown function group (Tatusov et al., 2000). Of note, closer examination of the genome revealed several loci of interest, including a putative *eps* locus and one CRISPR-Cas (CRISPR associated) locus. Additionally, there were several annotated transposases and mobile genetic elements (MGE). As the spread of antibiotic resistance is of growing concern, we next analyzed *L. fermentum* ATCC 14931 for any antibiotic resistance genes using ResFinder and found none, which is consistent with the aforementioned GRAS status of this species.

4.5.3. *L. fermentum* Species Genetic Diversity

With a complete genome sequence for the type strain, we next determined how ATCC 14931 compares to other *L. fermentum* strains and carried out comparative genomic analyses. Thirty-seven strains, in addition to ATCC 14931 (Table 4.1), were chosen for comparative analysis using the glycolysis gene phosphoglucomutase (Figure 4.2). Nine clades were identified in the phylogeny. *L. fermentum* ATCC 14391, highlighted by a red asterisk (*), was found to be

a part of a four-member clade that included the strains HFB3 (LJFJ01.1), L930BB (NZ_CBUR), and Lfu21 (NZ_PNBB). Interestingly, HFB3 and Lfu21 were isolated from human fecal samples, while L930BB was isolated from a human colon biopsy (Table 4.1).

Next, we selected six strains to perform whole genome comparisons with *L. fermentum* ATCC 14931. The genomes chosen for further analyses were: LT906621 (IMDO 130101, sourdough), NZ_AP017973 (MTCC 25067, fermented milk), NZ_CP019030 (SNUV175, human vagina), NZ_CP021790 (LAC FRN-92, human oral), NC_021235 (F-6, unknown), and NC_017465 (CECT 5716, human milk). These genomes were chosen as a representative set of the phylogeny generated in Figure 4.2 and are highlighted in red. They all contain a single contig or closed genome and range in size from 1.95Mb to 2.18Mb. GC content for each strain was ~51% (Table 4.1). MTCC 25067 and SNUV175 both carry plasmids. Using these six genomes in addition to ATCC 14931, whole genome analysis was carried out with BRIG (Figure 4.3). From the BRIG analysis, there are several islands in *L. fermentum* ATCC 14931 that do not occur within the other genomes. These islands at approximately 180kbp, 760kbp, and 1550kbp also correlate with GC dips. Further examination of these three islands did not reveal loci of note (Figure 4.S1-3), but several transposases in or around each island were identified (Figure 4.3). There are several smaller GC dips throughout the genome that correlate to either transposases or minor assembly gaps. There were no GC spikes observed. Another island of note is the CRISPR locus of *L. fermentum* ATCC 14931, which is only present in LT906621, annotated at 880kbp. Finally, the GC skew switches around 50kbp and 1090kb. Due to the large presence of transposases, we next used MAUVE to determine gene synteny amongst *L. fermentum* genomes (Figure 4.4). For this analysis, we used all genomes consisting of a single contig/closed genome, in addition to the strains used for the BRIG analysis (Table 4.1). Examination of the MAUVE

alignment showed several small blocks of synteny among the strains, in contrast to the expected large blocks of similarity. These small blocks generated by MAUVE could be combined into larger regions of synteny (outlined in boxes). In addition, there were several rearrangements observed, especially for genomes NZ_CP019030 (SNUV175, human vagina), NZ_CP021790 (LAC FRN-92, human oral), and NZ_CP017151 (NCC2970, unknown) (Figure 4.4). These smaller blocks of synteny and genome rearrangements could be due to the presence of transposons in the genomes.

4.5.4. CRISPR-Cas Immune Systems Diversity

Next, we examined the occurrence and diversity of CRISPR-Cas systems in *L. fermentum* across 38 strains (Figure 4.2). Potential CRISPR loci were identified using the CRISPR recognition tool (CRT) and then hand-curated. Types I, II, and III were all identified in *L. fermentum*. Several loci did not contain the complete *cas* complement due to draft genome sequences or transposons and were thus labelled unknown (Figure 4.2). Of the 38 strains analyzed, 71.8% encoded putative CRISPR-Cas systems. 53.8% of the strains analyzed contained a Type I system, 41.0% a Type II system, and 2.56% a Type III system. This is relatively hypervariable within a species, given the very high relative level of occurrence, and the absence of a single CRISPR-Cas system type that is widely shared across the species is noteworthy. Interestingly, one strain (OKQY01.1), contained a Type I, II, and III system, which is very rare in bacteria. This was the only strain with over 91 spacers in its genome (Figure 4.2).

We then used CRISPRViz to compare the spacer content and, presumably, the history of the strains (Figure 4.5). Type I, II, and III spacers grouped based on CRISPR-Cas systems. As expected, Type I systems encoded for a greater number of spacers than that of the Type II

systems (Toms and Barrangou, 2017). The spacers in *L. fermentum* as a whole were very diverse and we were unable to identify common ancestral spacers for the majority of the strains. Three genomes (NZ_AVAB, NC_010610, and NC_017465) had the most similar spacer arrays, only differing by one or two spacers in any of their Type I loci (Figure 4.5). Interestingly, each of these three genomes belonged to a different clade in the *L. fermentum* phylogeny (Figure 4.2). Of those with Type II systems, the genomes NZ_CP021104, CP020353, NZ_CP011536, NZ_PNBB shared some spacers, but also each had a great deal of unique spacers (Figure 4.5). Specifically, they shared a common ancestry and some newer additions; the main deviation was the large number of additional spacers in NZ_CP011536 (Figure 4.5). Interestingly, these genomes were a part of the same clade, with the exception of NZ_PNBB (Figure 4.2). A few other genomes, such as NZ_JQAU and NZ_PTL, also shared common spacers amongst each other. Even though the spacers varied widely, the repeats in *L. fermentum* did group with high similarity (Figure 4.6).

Next, we characterized the *L. fermentum* ATCC 14931 Type II CRISPR-Cas system. Of the strains used in the BRIG analysis, only IMD0 130101 (LT906621) also coded a Type II system (Figure 4.2). A comparison of the two strains' Type II loci is found in Figure 4.7A. Each strain has the following *cas* genes: *cas9*, *cas1*, *cas2*, and *csn2*. Cas9 is the signature protein for Type II systems and *csn2* is the genetic marker for subtype II-A (Koonin et al., 2017). There were several more spacers in LT906621 (20) than ATCC 14931 (12). The repeat sequences for both systems were the same, only differing in their ancestral repeats, which often acquires SNPs. mRNA-Seq expression was overlaid on ATCC 14931's locus to show active transcription of the *cas* genes (Figure 4.7A).

smRNA-Seq and *in silico* predictions were used to further characterize *L. fermentum* ATCC 14931's CRISPR-Cas system (Figure 4.7). Expression levels for the CRISPR array,

CRISPR RNA (crRNA), leaderRNA (ldrRNA), and tracrRNA were determined as shown in Figure 4.7 (B, C, D, and E, respectively). In the CRISPR locus, the last two crRNAs (most ancestral) were found to be the most highly expressed spacers in the cell. Boundaries were determined for the crRNA, ldrRNA, and tracrRNA. The crRNA was found to consist of a 21 bp section of the CRISPR repeat and a 20 bp section of spacer, which is common in Type II-A CRISPR-Cas systems (Deltcheva et al., 2011; Crawley et al., 2018). The ldrRNA contains a 21 bp portion of repeat and a 20 bp leader. The tracrRNA was found to be 75 bp, which was much shorter than predicted (Figure 4.7E). The structure of the tracrRNA was determined using NUPAK (Figure 4.7G). The tracrRNA sequence modules are colored as previously described (Briner et al., 2014). *L. fermentum* ATCC 14931's tracrRNA consists of all expected modules and contains only a single hairpin. Examining the BLAST results of *L. fermentum*'s Type II spacers, we predicted the PAM of ATCC 14931 to be (C/T)AAA (Figure 4.7F). Finally, a BLASTp comparison between *L. fermentum* ATCC 14931's Cas9 gene sequence, the *Streptococcus thermophilus* (Sth) Cas9 gene sequence, and the *Streptococcus pyogenes* (Spy) Cas9 gene sequence found at most only 32% AA identity between *L. fermentum* ATCC 14931's Cas9 and the other Cas9s. *L. fermentum* ATCC 14931's Cas9 is 1,378 AAs long and its closest relatives are *Lactobacillus gorillae* and *Lactobacillus mucosae,* with 72% and 57% identity, respectively.

4.6. DISCUSSION

In this study, we assess the genetic and phenotypic potential of the *Lactobacillus fermentum* species with focus on the type strain ATCC 14931. Metabolically, *L. fermentum* ATCC 14931 showed the ability to uptake and catabolize a diversity of carbohydrates encompassing D-ribose, D-galactose, D-glucose, D-fructose, esculin ferric citrate, D-maltose, D-lactose, D-melibiose, D-saccharose, D-raffinose, and potassium gluconate. Utilization was then confirmed using growth curves in SDM for galactose, fructose, and D-xylose (Figure 4.1A). The cell morphology of *L. fermentum* ATCC 14931 revealed a ropy phenotype, typical of an EPS-producer. There is a putative *eps* locus in the genome and SEM images appear to visualize an EPS layer on and between cells (Figure 4.1D-F). Close examination of the SEM images revealed strands of cellular material extending from the cells, notably at the extremities (Figure 4.1 D and F). In Figure 4.1F, it appears as if these strands are reaching towards the chain of cells (indicated by a white arrow). EPS has implications in food manufacturing for texture, in human health for biofilm formation, and in biotechnology for pathogen exclusion (Berlanga and Guerrero, 2016; Mende et al., 2016; Sarikaya et al., 2017). In stress challenges (Figure 4.1B and 1C), *L. fermentum* ATCC 14931 showed 15% survival in simulated intestinal juice and no survival after one hour in simulated gastric juice. This contrasts with the well-characterized probiotic *L. acidophilus*, which showed over 80% survival in simulated intestinal juice after five hours and ~20% survival in simulated gastric juice after 50 minutes (Johnson et al., 2013). In addition, there was little to no growth in media containing oxgall and porcine bile. This implies that *L. fermentum* ATCC 14931 would survive poorly in the human GIT and most likely serves only as an allochthonous species introduced to the microbiome through diet. This would potentially pose a major problem for the survival of and colonization of the GIT by *L. fermentum* ATCC 14931.

Improving the previously published genome sequence of *L. fermentum* ATCC 14931 allowed us to set a baseline genomic analysis for the type strain. The GC content (52.50%) is higher than what is typical for the low-GC *Lactobacillus* genus (Brandt and Barrangou, 2018). As lactobacilli are typically considered low-GC organisms, this finding may suggest that *L. fermentum* has seen less genomic drift. It is generally believed that as *Lactobacillus* species become more adapted to their environment, they begin to undergo genome decay (Makarova et al., 2006). Typically, lactobacilli with more than one niche have larger genomes and have undergone less genome decay. This is corroborated by a recent study looking at niche-adaptations in *Lactobacillus*; *L. fermentum*, while included in the study, did not have enough information to assign it a particular niche category (Duar et al., 2017). This could imply that *L. fermentum* is a member of various niches and is still in the process of active adaptation. The portion (15%) of unknown/hypothetical genes certainly implies that there is still much to discover about *L. fermentum* ATCC 14931. We also examined the genome for genes that are relevant for industrial use. As antibiotic resistance genes are raising concerns in both health and biotechnology applications, we examined *L. fermentum* ATCC 14931 for any predicted antibiotic resistance genes and found none.

After examining the genome of *L. fermentum* ATCC 14931, we performed a global phylogeny of *L. fermentum* using 38 genomes (Figure 4.2). This analysis revealed a great deal of diversity among *L. fermentum* strains. Nine clades were identified, with *L. fermentum* ATCC 14931 as a part of a four-member clade, consisting of the strains HFB3 (LJFJ01.1), L930BB (NZ_CBUR), and Lfu21 (NZ_PNBB). Even though *L. fermentum* ATCC 14931 was isolated from fermented beets, its clade members were isolated from human feces/colon biopsies. We would anticipate that similar strains would have similar isolation sources. Since this is not the

case for *L. fermentum* ATCC 14931, this could imply *L. fermentum* enters the human microbiome through food sources and is a transient member (allochthonous), rather than a permanent member of the human microbiome (autochthonous). This fits with data found in IMNGS databases that show *L. fermentum*'s main environments to be food and human gut metagenomes. As transient members, it would also explain why *L. fermentum* does not have a specific niche-adaptation (Duar et al., 2017). This finding also reflects the low survival found under GIT conditions.

Next, we performed whole genome comparisons using BRIG and Mauve with *L. fermentum* ATCC 14931 and other complete genomes. For the BRIG Analysis, six genomes, NC_017465, NC_021235, LT906621, NZ_CP021790, NZ_AP017973, and NZ_CP019030 were chosen to represent the phylogeny from Figure 4.2. Their average genome size and GC% is 2.07Mb and 51.6%, respectively, making *L. fermentum* ATCC 14931 slightly smaller (1.89Mb) and have a slightly higher GC% (52.5%) as compared to the other strains in the analyses. As seen in Figure 4.3, comparing the seven strains via BRIG revealed three genomic islands in *L. fermentum* ATCC 1493 that are absent in the other *L. fermentum* genomes. These islands are identifiable not only based on their absence in the other strains, but by a corresponding decrease in GC content. Further examination revealed that transposases and mobile genetic elements were frequently in and around these islands, which is indicative of acquired genes—potentially through horizontal gene transfer. No other loci of interest were identified (Figure 4.S1-3). Another *L. fermentum* ATCC 14931 island encompassed the CRISPR locus - which is absent in the other genomes - with the exception of LT906621. A continuing examination of GC dips resulted in the identification of several other smaller GC dips in the BRIG alignment, which again correlated mostly with transposases and minor assembly gaps. These loci were often absent

in the other *L. fermentum* genomes. We next analyzed gene synteny using whole genome MAUVE analysis (Figure 4.4). Due to the large number of transposons identified in the BRIG analysis, we elected to include all completed genomes in the MAUVE analysis. Typically, the strains of a species are highly similar, and this manifests as large blocks of co-linearization in the MAUVE alignment. However, our analysis showed only small blocks of similarity and many rearrangements. This is unsurprising given the large number of MGEs discovered in *L. fermentum* ATCC 14931. We were able to show that many of the small blocks identified by MAUVE remained in the same order and could be considered larger blocks of synteny (Figure 4.4). Interestingly, genomes NZ_CP019030, NZ_CP021790, and NZ_CP017151 showed very little commonalities with the other *L. fermentum* genomes. While this could be a reflection of MGE's, it may also imply inaccurate assemblies.

As CRISPR-Cas systems are a valuable reservoir of Cas-based genome editing technologies, we determined the occurrence and diversity of CRISPR systems in the thirty-eight analyzed *L. fermentum* genomes. On a species level, we found that 71.8% of strains encoded a predicted CRISPR system (Figure 4.2). This is higher than *Lactobacillus* in general (62.9%), and bacteria as a whole (46%), suggesting that *L. fermentum* is a potential reservoir for novel CRISPR-based tools (Sun et al., 2015). Type I is the most common system found in *L. fermentum* (53.8%), which reflects the overall dominance and diversity of Type I systems in nature (Makarova et al., 2015). Type I CRISPR-Cas systems have recently been studied for antimicrobial properties, and as such *L. fermentum* could be potentially explored as a programmed antimicrobial in microbiome settings (Gomaa et al., 2014). While Type I is more common than Type II systems, it is the Type II's signature Cas9 programmable endonuclease that is the most popular tool of the CRISPR toolbox (Jinek et al., 2012). 41% of *L. fermentum*

strains contain a predicted Type II system. This is slightly higher than the Type II occurrence rate in lactobacilli (36%) and much higher than the occurrence rate in all bacteria (5%) (Chylinski et al., 2014; Sun et al., 2015). It is of note that one genome (OKQY01.1) was predicted to contain a Type I, II, and III system-- a rare occurrence (Horvath and Barrangou, 2010). Of the strains chosen for whole genome comparisons, all contained a putative Type I system except for *L. fermentum* ATCC 14931, and only *L. fermentum* ATCC 14931 and LT906621 contained a putative Type II system. A global analysis of the spacers found in *L. fermentum* revealed greater diversity than expected (Figure 4.5). Typically, strains of a species share similar spacer history, or, "vaccination records," resulting in the sharing of spacers, especially towards the ancestral ends of loci. In our analysis, we found only a limited number of shared spacers. Of the predicted Type II systems, NZ_CP021104's, CP020353's, NZ_CP011536's, and NZ_PNBB's loci shared common history, specifically in the ancestral spacers. However, there were several deletions or additional spacers in each locus, making the shared spacers a minority. With the exception of NZ_PNBB, these genomes were found in the same clade (Figure 4.2). In contrast, the genomes with the most similar predicted records were NZ_AVAB, NC_010610, and NC_017465. All three putative Type I loci in each strain shared the same vaccination record as the other strains, with the exception of one or two spacers. Intriguingly, these genomes did not share clades (Figure 4.2). Although there was not much congruity in the spacers, the predicted repeats of the *L. fermentum* CRISPR loci did share a high degree of similarity (Figure 4.6).

We then performed an in-depth *in silico* analysis of *L. fermentum* ATCC 14931's putative CRISPR loci and revealed it to be a Type II-A, as evidenced by the *csn2* gene (Figure 4.7A). As the only other strain with a putative Type II CRISPR-Cas system from those genomes selected

for BRIG comparison, *L. fermentum* LT906621 was used to compare CRISPR loci. Both predicted systems were Type II-A, with *L. fermentum* LT906621 coding for a slightly larger CRISPR array. The repeats for each strain were identical, but they shared no common spacers. We also examined the expression levels of *L. fermentum* ATCC 14931's putative CRISPR loci using mRNA and smRNA-seq. mRNA expression levels showed that the *cas* genes are transcribed in *L. fermentum* ATCC 14931. Analysis of expression in the CRISPR array using smRNA-seq revealed that the two most ancestral crRNA are the most highly expressed in *L. fermentum* ATCC 14931's CRISPR locus. This is highly unusual, as the newly acquired crRNA are typically the most highly expressed since they are more recently exposed to infection (Wei et al., 2015; McGinn and Marraffini, 2016). It is possible that there is an internal promoter driving the expression of the ancestral crRNAs, and thus why the expression does not fit canonical expectations. The crRNA, ldrRNA, and tracrRNA had similar sizes as reported previously in lactobacilli (Figure 4.7B-6D) (Crawley et al., 2018). The *in silico* prediction of the tracrRNA was longer than the true boundaries predicted through smRNA-Seq, which has been previously reported (Crawley et al., 2018). This implies that our predictions are conservative compared to what is used *in vivo.* The tracrRNA structure showed the appropriate modules including the lower stem, bulge, upper stem, nexus, and contained a single hairpin (Figure 4.7G). Finally, we predicted *L. fermentum* ATCC 14931's PAM to be (C/T)AAA (Figure 4.7F). It is similar to several predicted PAMs in *L. gasseri* (TAA) (Sanozky-Dawes et al., 2015). Overall, expression for *L. fermentum* ATCC 14931's CRISPR loci fit canonical expectations, with the exception of the highly transcribed ancestral spacers. Despite its similarities to canonical Type II loci, the Cas9 in *L. fermentum* ATCC 14931 is unique, only sharing 32% AA identity with either Sth's or Spy's Cas9—two of the most commonly used Cas9s in genome editing. This is especially

intriguing as the Cas9s of Sth and Spy only share ~32% AA identity with each other. This marks *L. fermentum* ATCC 14931 as a potential new orthogonal Cas9 for tool development.

Overall, this study provides a basis for functional and genetic analysis of *L. fermentum* strains, with characterization of the type strain ATCC 14931. We determined the complete genome sequence of the type strain and carried out comparative genomic analyses revealing high variability within the species, encompassing MGEs and genomic islands. This genetic variability is also illustrated by the occurrence and diversity of hypervariable CRISPR-Cas systems. These observations highlight the value of determining the complete genome sequence of reference and type strains within a species, along with opening new avenues for the study of *Lactobacillus fermentum* strains and related species.

4.7. ACKNOWLEDGEMENTS

We would like to thank the CRISPR lab for insights and support during this project. We also thank Laurel Hedgecock for technical assistance.

4.8. REFERENCES

Aarestrup, F.M., Hasman, H., Vestergaard, M., Zankari, E., Larsen, M.V., Lund, O., et al. (2012). Identification of acquired antimicrobial resistance genes. *Journal of Antimicrobial Chemotherapy* 67(11)**,** 2640-2644. doi: 10.1093/jac/dks261.

Alikhan, N.-F., Petty, N.K., Ben Zakour, N.L., and Beatson, S.A. (2011). BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. *BMC genomics* 12**,** 402-402. doi: 10.1186/1471-2164-12-402.

Aziz, R.K., Bartels, D., Best, A.A., DeJongh, M., Disz, T., Edwards, R.A., et al. (2008). The RAST Server: Rapid Annotations using Subsystems Technology. *BMC Genomics* 9(1)**,** 75. doi: 10.1186/1471-2164-9-75.

Berlanga, M., and Guerrero, R. (2016). Living together in biofilms: the microbial cell factory and its biotechnological implications. *Microb Cell Fact* 15(1)**,** 165. doi: 10.1186/s12934-016-0569-5.

Bernardeau, M., Vernoux, J.P., Henri-Dubernet, S., and Guéguen, M. (2008). Safety assessment of dairy microorganisms: The *Lactobacillus* genus. *International Journal of Food Microbiology* 126(3)**,** 278-285. doi: https://doi.org/10.1016/j.ijfoodmicro.2007.08.015.

Bland, C., Ramsey, T.L., Sabree, F., Lowe, M., Brown, K., Kyrpides, N.C., et al. (2007). CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics* 8**,** 209. doi: 10.1186/1471-2105-8-209.

Brandt, K., and Barrangou, R. (2018). Using glycolysis enzyme sequences to inform Lactobacillus phylogeny. *Microbial Genomics* 4(6)**,** -. doi: doi:10.1099/mgen.0.000187.

Briner, Alexandra E., Donohoue, Paul D., Gomaa, Ahmed A., Selle, K., Slorach, Euan M., Nye, Christopher H., et al. (2014). Guide RNA Functional Modules Direct Cas9 Activity and Orthogonality. *Molecular Cell* 56(2)**,** 333-339. doi: 10.1016/j.molcel.2014.09.019.

Briner, A.E., Henriksen, E.D., and Barrangou, R. (2016). Prediction and Validation of Native and Engineered Cas9 Guide Sequences. *Cold Spring Harbor Protocols* 2016(7)**,** pdb.prot086785. doi: 10.1101/pdb.prot086785.

Charteris, W.P., Kelly, P.M., Morelli, L., and Collins, J.K. (1998). Development and application of an in vitro methodology to determine the transit tolerance of potentially probiotic *Lactobacillus* and *Bifidobacterium* species in the upper human gastrointestinal tract. *J Appl Microbiol* 84(5)**,** 759-768.

Chylinski, K., Makarova, K.S., Charpentier, E., and Koonin, E.V. (2014). Classification and evolution of type II CRISPR-Cas systems. *Nucleic acids research* 42(10)**,** 6091-6105. doi: 10.1093/nar/gku241.

Crawley, A.B., Henriksen, E.D., Stout, E., Brandt, K., and Barrangou, R. (2018). Characterizing the activity of abundant, diverse and active CRISPR-Cas systems in lactobacilli. *Scientific Reports* 8(1)**,** 11544. doi: 10.1038/s41598-018-29746-3.

Darling, A.C.E., Mau, B., Blattner, F.R., and Perna, N.T. (2004). Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome research* 14(7)**,** 1394-1403. doi: 10.1101/gr.2289704.

Deltcheva, E., Chylinski, K., Sharma, C.M., Gonzales, K., Chao, Y., Pirzada, Z.A., et al. (2011). CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature* 471**,** 602. doi: 10.1038/nature09886. https://www.nature.com/articles/nature09886#supplementary-information.

Diaz-Ropero, M.P., Martin, R., Sierra, S., Lara-Villoslada, F., Rodriguez, J.M., Xaus, J., et al. (2007). Two *Lactobacillus strains*, isolated from breast milk, differently modulate the immune response. *J Appl Microbiol* 102(2)**,** 337-343. doi: 10.1111/j.1365-2672.2006.03102.x.

Duar, R.M., Lin, X.B., Zheng, J., Martino, M.E., Grenier, T., Perez-Munoz, M.E., et al. (2017). Lifestyles in transition: evolution and natural history of the genus *Lactobacillus*. *FEMS Microbiol Rev* 41(Supp_1)**,** S27-s48. doi: 10.1093/femsre/fux030.

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* 32(5)**,** 1792-1797. doi: 10.1093/nar/gkh340.

Frece, J., Kos, B., Svetec, I.K., Zgaga, Z., Mrsa, V., and Suskovic, J. (2005). Importance of S-layer proteins in probiotic activity of *Lactobacillus acidophilus* M92. *J Appl Microbiol* 98(2)**,** 285-292. doi: 10.1111/j.1365-2672.2004.02473.x.

Gan, B.S., Kim, J., Reid, G., Cadieux, P., and Howard, J.C. (2002). *Lactobacillus fermentum* RC-14 inhibits *Staphylococcus aureus* infection of surgical implants in rats. *J Infect Dis* 185(9)**,** 1369-1372. doi: 10.1086/340126.

Ghazvini, R.D., Kouhsari, E., Zibafar, E., Hashemi, S.J., Amini, A., and Niknejad, F. (2016). Antifungal Activity and Aflatoxin Degradation of *Bifidobacterium Bifidum* and *Lactobacillus Fermentum* Against Toxigenic *Aspergillus Parasiticus*. *Open Microbiol J* 10**,** 197-201. doi: 10.2174/1874285801610010197.

Gomaa, A.A., Klumpe, H.E., Luo, M.L., Selle, K., Barrangou, R., and Beisel, C.L. (2014). Programmable Removal of Bacterial Strains by Use of Genome-Targeting CRISPR-Cas Systems. *mBio* 5(1)**,** e00928-00913. doi: 10.1128/mBio.00928-13.

Hill, C., Guarner, F., Reid, G., Gibson, G.R., Merenstein, D.J., Pot, B., et al. (2014). The International Scientific Association for Probiotics and Prebiotics consensus statement on the scope and appropriate use of the term probiotic. *Nature Reviews Gastroenterology &Amp; Hepatology* 11**,** 506. doi: 10.1038/nrgastro.2014.66.

Horvath, P., and Barrangou, R. (2010). CRISPR/Cas, the Immune System of Bacteria and Archaea. *Science* 327(5962)**,** 167. doi: 10.1126/science.1179555.

Huerta-Cepas, J., Forslund, K., Coelho, L.P., Bork, P., von Mering, C., Szklarczyk, D., et al. (2017). Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. *Molecular Biology and Evolution* 34(8)**,** 2115-2122. doi: 10.1093/molbev/msx148.

Huerta-Cepas, J., Forslund, K., Sunagawa, S., Bork, P., Szklarczyk, D., Heller, D., et al. (2015). eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Research* 44(D1)**,** D286-D293. doi: 10.1093/nar/gkv1248.

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J.A., and Charpentier, E. (2012). A Programmable Dual-RNA–Guided DNA Endonuclease in Adaptive Bacterial Immunity. *Science* 337(6096)**,** 816. doi: 10.1126/science.1225829.

Johnson, B., Selle, K., O'Flaherty, S., Goh, Y.J., and Klaenhammer, T. (2013). Identification of extracellular surface-layer associated proteins in *Lactobacillus acidophilus* NCFM. *Microbiology (Reading, England)* 159(Pt 11)**,** 2269-2282. doi: 10.1099/mic.0.070755-0.

Kimmel, S.A., and Roberts, R.F. (1998). Development of a growth medium suitable for exopolysaccharide production by *Lactobacillus delbrueckii* ssp. *bulgaricus* RR. *Int J Food Microbiol* 40(1-2)**,** 87-92.

Koonin, E.V., Makarova, K.S., and Zhang, F. (2017). Diversity, classification and evolution of CRISPR-Cas systems. *Curr Opin Microbiol* 37**,** 67-78. doi: 10.1016/j.mib.2017.05.008.

Langdon, W.B. (2015). Performance of genetic programming optimised Bowtie2 on genome comparison and analytic testing (GCAT) benchmarks. *BioData Mining* 8(1)**,** 1. doi: 10.1186/s13040-014-0034-0.

Lebeer, S., Bron, P.A., Marco, M.L., Van Pijkeren, J.-P., O'Connell Motherway, M., Hill, C., et al. (2018). Identification of probiotic effector molecules: present state and future perspectives. *Current Opinion in Biotechnology* 49**,** 217-223. doi: https://doi.org/10.1016/j.copbio.2017.10.007.

Makarova, K., Slesarev, A., Wolf, Y., Sorokin, A., Mirkin, B., Koonin, E., et al. (2006). Comparative genomics of the lactic acid bacteria. *Proceedings of the National Academy of Sciences of the United States of America* 103(42)**,** 15611-15616. doi: 10.1073/pnas.0607117103.

Makarova, K.S., Wolf, Y.I., Alkhnbashi, O.S., Costa, F., Shah, S.A., Saunders, S.J., et al. (2015). An updated evolutionary classification of CRISPR-Cas systems. *Nature reviews. Microbiology* 13(11)**,** 722-736. doi: 10.1038/nrmicro3569.

McGinn, J., and Marraffini, L.A. (2016). CRISPR-Cas Systems Optimize Their Immune Response by Specifying the Site of Spacer Integration. *Molecular cell* 64(3)**,** 616-623. doi: 10.1016/j.molcel.2016.08.038.

Mende, S., Rohm, H., and Jaros, D. (2016). Influence of exopolysaccharides on the structure, texture, stability and sensory properties of yoghurt and related products. *International Dairy Journal* 52**,** 57-71. doi: https://doi.org/10.1016/j.idairyj.2015.08.002.

Mikelsaar, M., and Zilmer, M. (2009). *Lactobacillus fermentum* ME-3 - an antimicrobial and antioxidative probiotic. *Microbial ecology in health and disease* 21(1)**,** 1-27. doi: 10.1080/08910600902815561.

Nethery, M.A., and Barrangou, R. (2018). CRISPR Visualizer: rapid identification and visualization of CRISPR loci via an automated high-throughput processing pipeline. *RNA Biology***,** 1-8. doi: 10.1080/15476286.2018.1493332.

Pereira, D.I., and Gibson, G.R. (2002). Cholesterol assimilation by lactic acid bacteria and bifidobacteria isolated from the human gut. *Appl Environ Microbiol* 68(9)**,** 4689-4693.

Pereira, D.I.A., McCartney, A.L., and Gibson, G.R. (2003). An In Vitro Study of the Probiotic Potential of a Bile-Salt-Hydrolyzing *Lactobacillus fermentum* Strain, and Determination of Its Cholesterol-Lowering Properties. *Applied and Environmental Microbiology* 69(8)**,** 4743. doi: 10.1128/AEM.69.8.4743-4752.2003.

Pfeiler, E.A., and Klaenhammer, T.R. (2007). The genomics of lactic acid bacteria. *Trends in Microbiology* 15(12)**,** 546-553. doi: https://doi.org/10.1016/j.tim.2007.09.010.

Pot, B., Ludwig, W., Kersters, K., and Schleifer, K.-H. (1994). "Taxonomy of Lactic Acid Bacteria," in *Bacteriocins of Lactic Acid Bacteria: Microbiology, Genetics and Applications,* eds. L. De Vuyst & E.J. Vandamme. (Boston, MA: Springer US), 13-90.

Sanozky-Dawes, R., Selle, K., apos, Flaherty, S., Klaenhammer, T., and Barrangou, R. (2015). Occurrence and activity of a type II CRISPR-Cas system in *Lactobacillus gasseri*. *Microbiology* 161(9)**,** 1752-1761. doi: doi:10.1099/mic.0.000129.

Sarikaya, H., Aslim, B., and Yuksekdag, Z. (2017). Assessment of anti-biofilm activity and bifidogenic growth stimulator (BGS) effect of lyophilized exopolysaccharides (l-EPSs) from Lactobacilli strains. *International Journal of Food Properties* 20(2)**,** 362-371. doi: 10.1080/10942912.2016.1160923.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics (Oxford, England)* 30(9)**,** 1312-1313. doi: 10.1093/bioinformatics/btu033.

Sun, Z., Harris, H.M.B., McCann, A., Guo, C., Argimón, S., Zhang, W., et al. (2015). Expanding the biotechnology potential of lactobacilli through comparative genomics of 213 strains

and associated genera. *Nature Communications* 6**,** 8322. doi: 10.1038/ncomms9322 https://www.nature.com/articles/ncomms9322#supplementary-information.

Tatusov, R.L., Galperin, M.Y., Natale, D.A., and Koonin, E.V. (2000). The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic acids research* 28(1)**,** 33-36.

Theilmann, M.C., Goh, Y.J., Nielsen, K.F., Klaenhammer, T.R., Barrangou, R., and Abou Hachem, M. (2017). *Lactobacillus acidophilus* Metabolizes Dietary Plant Glucosides and Externalizes Their Bioactive Phytochemicals. *mBio* 8(6)**,** e01421-01417. doi: 10.1128/mBio.01421-17.

Toms, A., and Barrangou, R. (2017). On the global CRISPR array behavior in class I systems. *Biology Direct* 12(1)**,** 20. doi: 10.1186/s13062-017-0193-2.

Varma, P., Dinesh, K.R., Menon, K.K., and Biswas, R. (2010). *Lactobacillus fermentum* isolated from human colonic mucosal biopsy inhibits the growth and adhesion of enteric and foodborne pathogens. *J Food Sci* 75(9)**,** M546-551. doi: 10.1111/j.1750-3841.2010.01818.x.

Velraeds, M.M., van der Mei, H.C., Reid, G., and Busscher, H.J. (1996). Inhibition of initial adhesion of uropathogenic *Enterococcus faecalis* by biosurfactants from *Lactobacillus* isolates. *Applied and environmental microbiology* 62(6)**,** 1958-1963.

Wei, Y., Chesne, M.T., Terns, R.M., and Terns, M.P. (2015). Sequences spanning the leader-repeat junction mediate CRISPR adaptation to phage in *Streptococcus thermophilus*. *Nucleic acids research* 43(3)**,** 1749-1758. doi: 10.1093/nar/gku1407.

Yoo, D., Cho, S., Kim, H., Bagon, B.B., Oh, J.K., Valeriano, V.D.V., et al. (2017). Complete genome analysis of *Lactobacillus fermentum* SK152 from kimchi reveals genes associated with its antimicrobial activity. *FEMS Microbiology Letters* 364(18). doi: 10.1093/femsle/fnx185.

Table 4.1 | Genome List. Genomic features of 38 *L. fermentum* strains used in this study.

Strain	Sequence Length	GC%	#Sequences	#Plasmids	Accession	Isolation
ATCC 14931	1887974	52.50%	1	0		Fermented beets
MTCC 25067	1954694	51.50%	1	1	AP017973.1	Fermented Milk
VRI-003	1949297	52.10%	1	0	CP020353.1	Commercial Probiotic
IMD0 130101	2089202	51.50%	1	0	LT906621.1	Sourdough
IFO 3956	2098685	51.50%	1	0	NC_010610	Fermented plant material
CECT 5716	2100449	51.50%	1	0	NC_017465	Human milk
F-6	2064620	51.70%	1	0	NC_021235	Unknown
3872	2297851	50.70%	1	1	NZ_CP011536	Milk
NCC2970	1949874	52.20%	1	0	NZ_CP017151	Unknown
47-7	2098685	52.50%	1	0	NZ_CP017712	Unknown
SNUV175	2176678	51.50%	1	3	NZ_CP019030	Human vagina
FTDC 8312	2239921	51.00%	1	0	NZ_CP021104.1	Human feces
LAC FRN-92	2063606	51.80%	1	0	NZ_CP021790.1	Human oral
LfQi6	2098510	52.50%	1	0	NZ_CP025592.1	Human microbiome
HFB3		51.80%	7	0	LJFJ00000000.1	Human gut
28-3-CHN		52.20%	42	0	NZ_ACQG00000000	Human
39		51.60%	55	0	NZ_LBDG00000000	Unknown
L930BB		52.10%	72	0	NZ_CBUR000000000	Unknown
222		52.10%	73	0	NZ_CBZV000000000	Cocoa bean
RI-508		52.20%	74	0	NZ_MKGE00000000.1	Cacao bean fermentation
MD IIE-4657		52.30%	74	0	NZ_PTLW00000000.1	Silage
S6		52.30%	82	0	NZ_FUHZ00000000.1	Unknown
S13		52.30%	85	0	NZ_FUHY00000000.1	Unknown
90 TC-4		51.90%	93	0	NZ_LBDH00000000	Unknown
SHI-2		52.10%	93	0	NZ_NJPQ00000000.1	Human saliva
DSM 20055		52.40%	102	0	NZ_JQAU00000000	Human Saliva
UC0-979C		51.90%	108	0	NZ_LJWZ00000000	Human gastric
279		52.00%	108	0	NZ_PGGI00000000.1	Human feces
103		51.80%	110	0	NZ_PGGE00000000.1	Human cecum
311		51.80%	111	0	NZ_PGGJ00000000.1	Human feces
MTCC 8711		49.70%	116	7	NZ_AVAB00000000	Yogurt
CECT 9269		51.70%	129	0	NZ_OKQY00000000.1	Tocosh
LfU21		51.70%	131	0	NZ_PNBB00000000.1	Human feces
NB-22		51.80%	137	0	NZ_AYHA00000000	Human vagina
NCDC 400		51.60%	138	0	NZ_PDKX00000000.1	Curd
BFE 6620		52.10%	149	0	NZ_NIWV00000000.1	Gari
779_LFER		52.10%	169	0	NZ_JUTH00000000	Unknown
Lf1		52.60%	250	0	NZ_AWXS00000000	Human gut

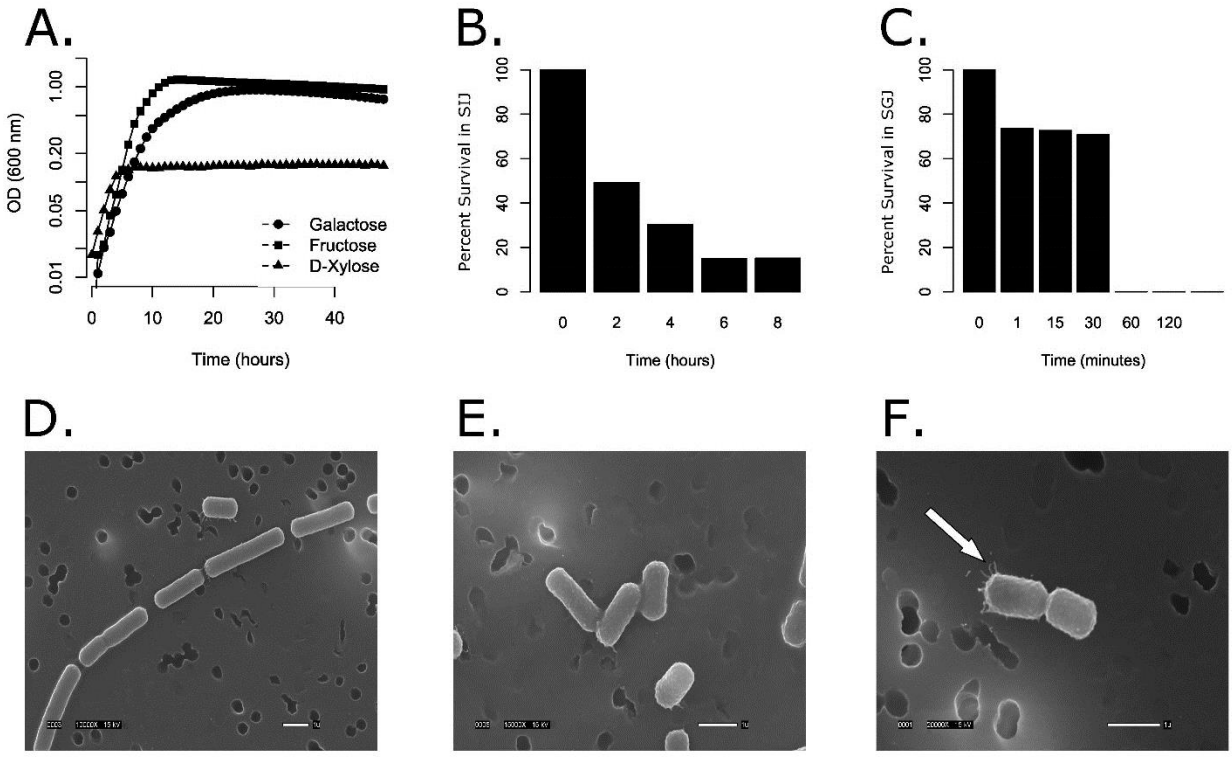


Figure 4.1 | Functional Attributes of ATCC 14931. (A) 48h growth curves of ATCC 14931 on substrates galactose (circle), fructose (square), and D-Xylose (triangle). Data points are the average of two biological replicates. Survival in simulated intestinal (B) or simulated gastric (C) juice. Data is the average of two biological replicates. SEM photos of ATCC 14931 at 15kV and 10000X (E), 15000X (F) or 20000X (G) zoom. A white arrow indicates a feature of interest.

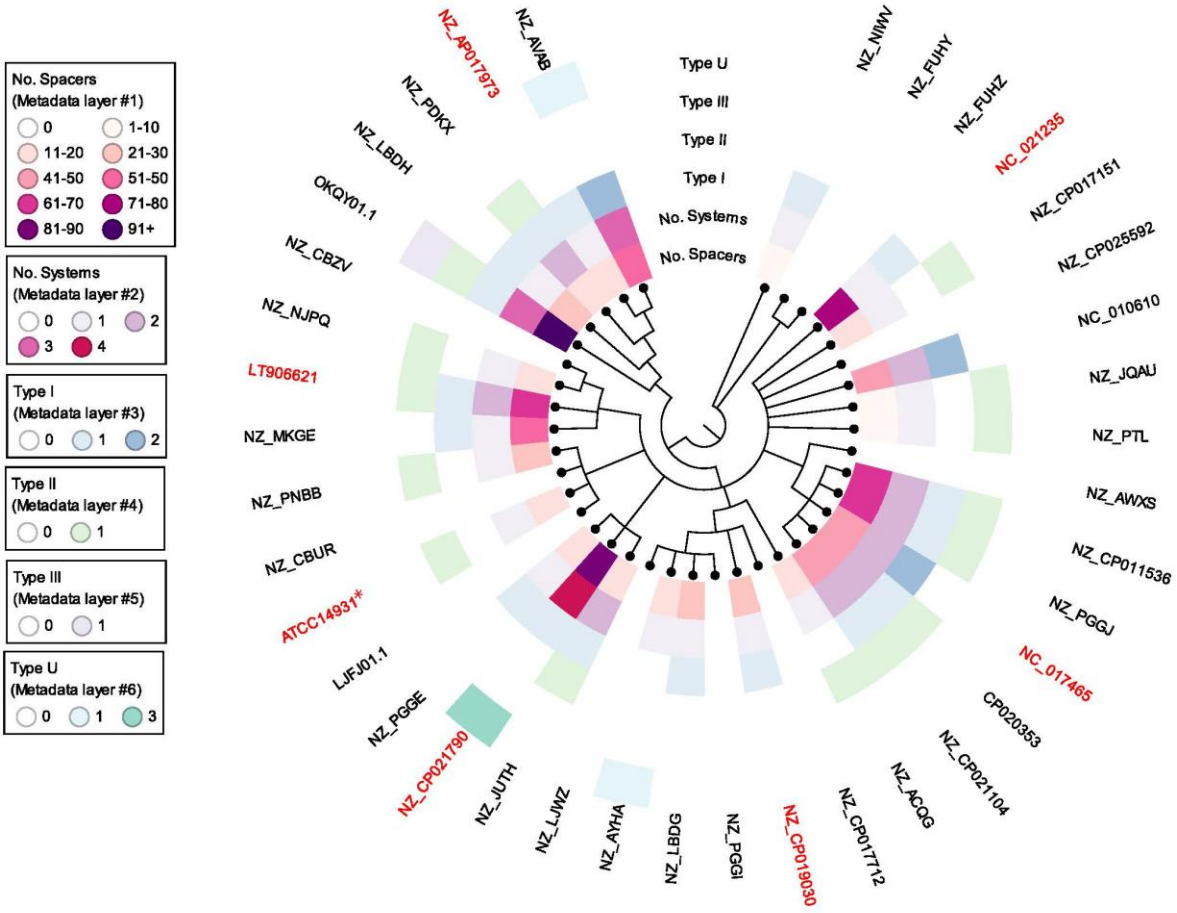


Figure 4.2 | *L. fermentum* Phylogeny. Phylogenetic tree generated of 38 *L. fermentum* strains using RAxML based on the nucleotide alignment of *phosphoglucomutase.* ATCC 14931 is indicated by a red asterisk (*). Genomes in red are used for subsequent BRIG comparisons. Rings refer to CRISPR-Cas analyses performed on the genomes and are (from inside out): number of spacers in the genome, number of total systems in the genome, number of Type I systems in the genome, number of Type II systems in the genome, number of Type III systems in the genome, and number of undefined Types in the genome. Ring legends are in the insets. Strain names can be found in Table 4.1.

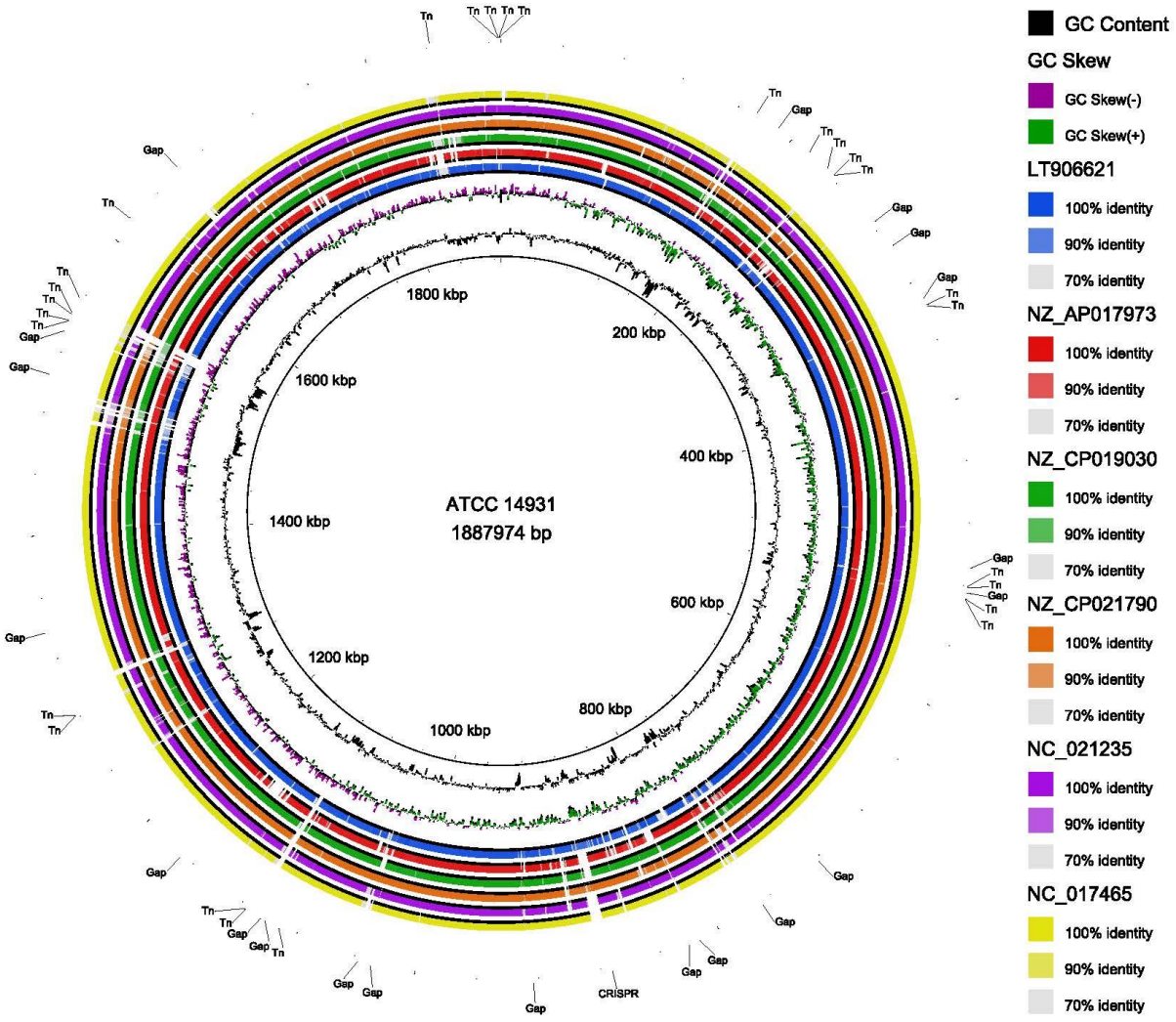


Figure 4.3 | BRIG Analysis. BRIG alignment of seven *L. fermentum* genomes with ATCC 14931 as the reference. The innermost ring denotes genome location. The other rings and color specifications can be found to the right of the ring image. Transposases, CRISPR genes, and minor assembly gaps are annotated outside of the rings. Strain names can be found in Table 4.1.

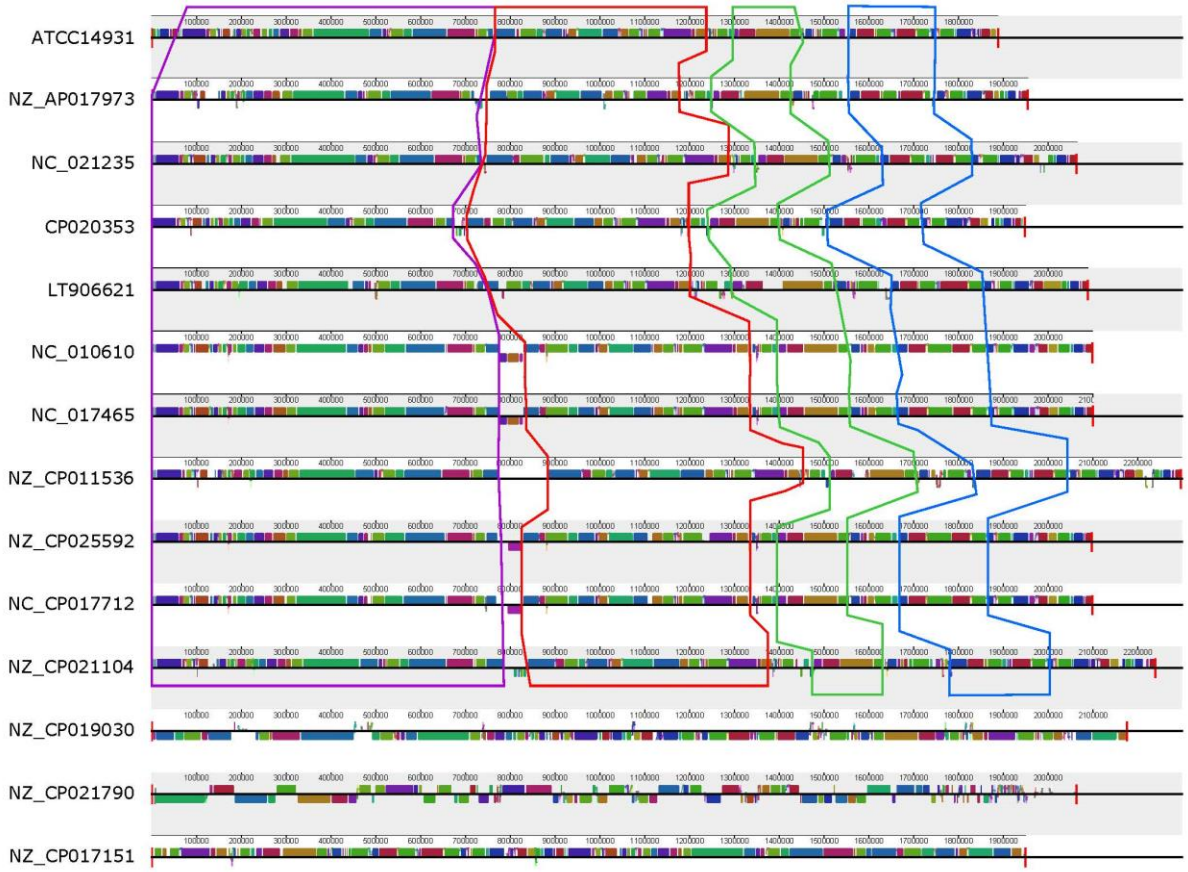


Figure 4.4 | Whole Genome Comparisons. MAUVE alignment of all complete *L. fermentum* genomes with ATCC 14931 set at the reference. Grouped blocks of similarity are boxed. Strain names can be found in Table 4.1.

Figure 4.5 | Global Spacer Visualization. Visualization of CRISPR spacers for 38 *L. fermentum* strains using CRISPRViz. Spacers for putative Type I loci are on the top, with Type III loci in the middle, and Type II loci on the bottom. Ancestral spacers are on the right-hand side of the figure. Strain names can be found in Table 4.1.

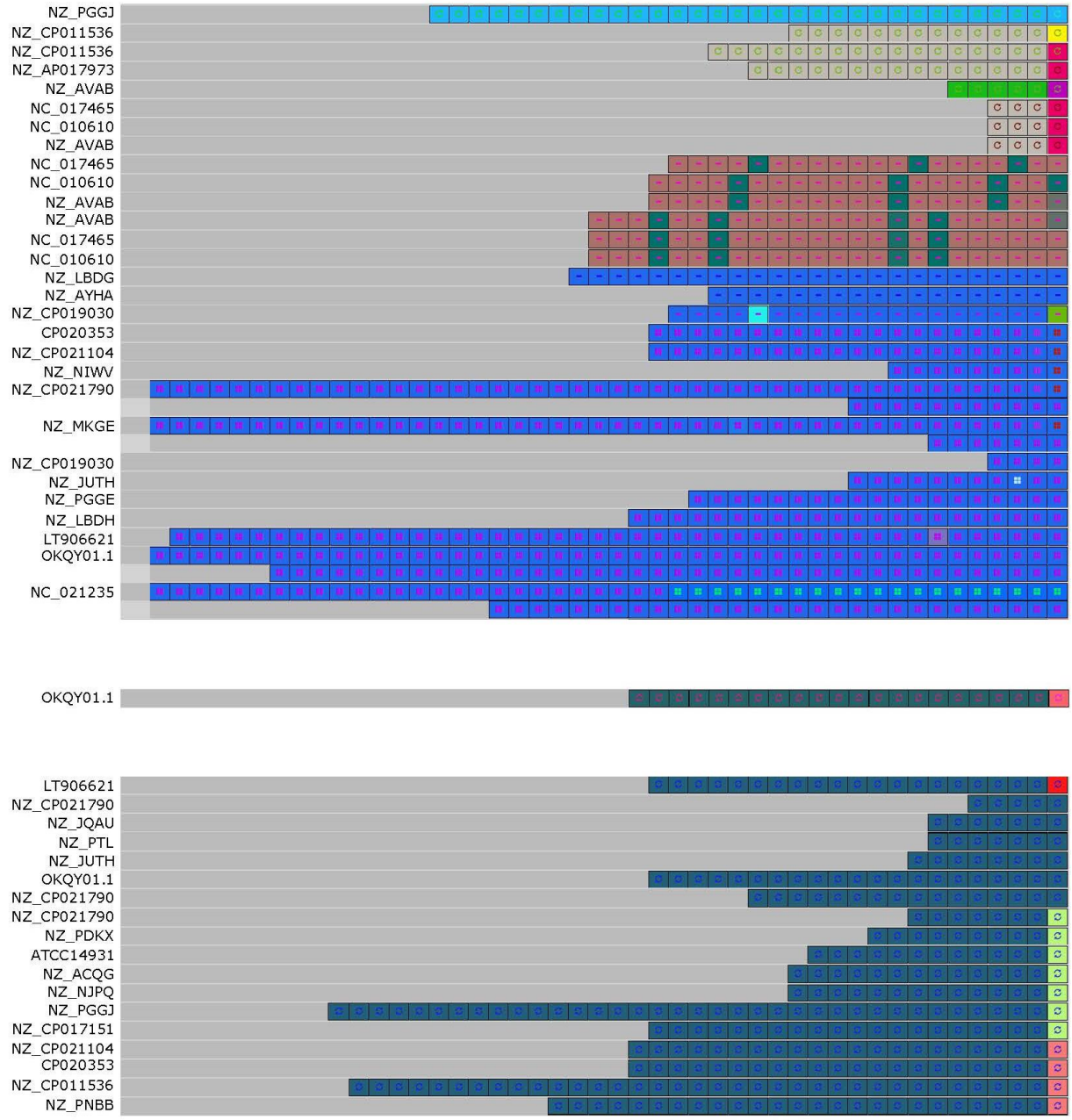


Figure 4.6 | Global Repeat Visualization. Visualization of CRISPR repeats for 38 *L. fermentum* strains using CRISPRViz. Repeats for putative Type I loci are on the top, with Type III loci in the middle, and Type II loci on the bottom. Ancestral repeats are on the right-hand side of the figure. Strain names can be found in Table 4.1.

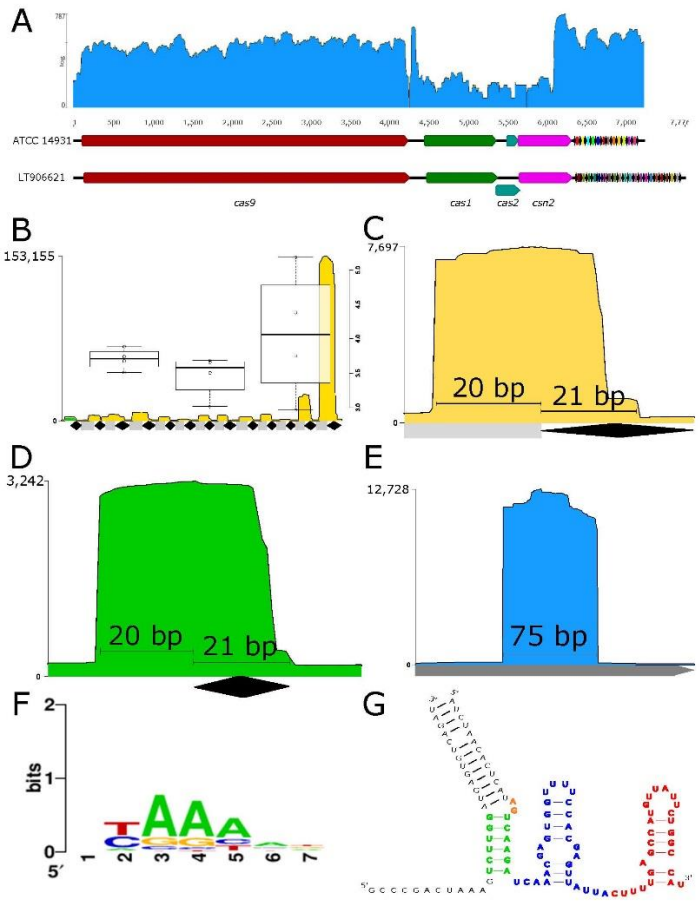
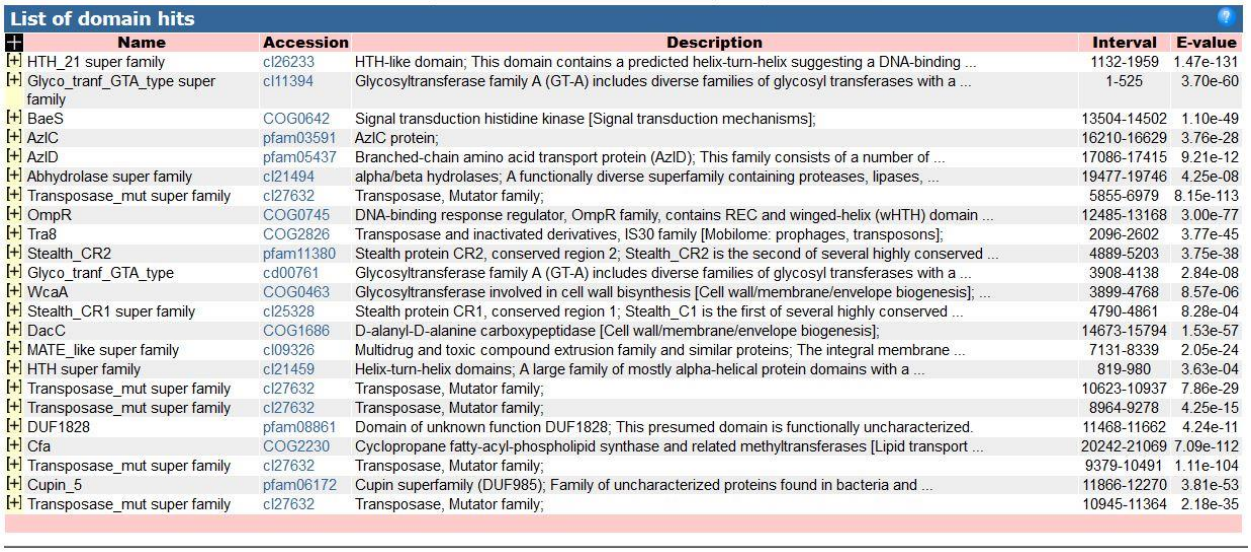


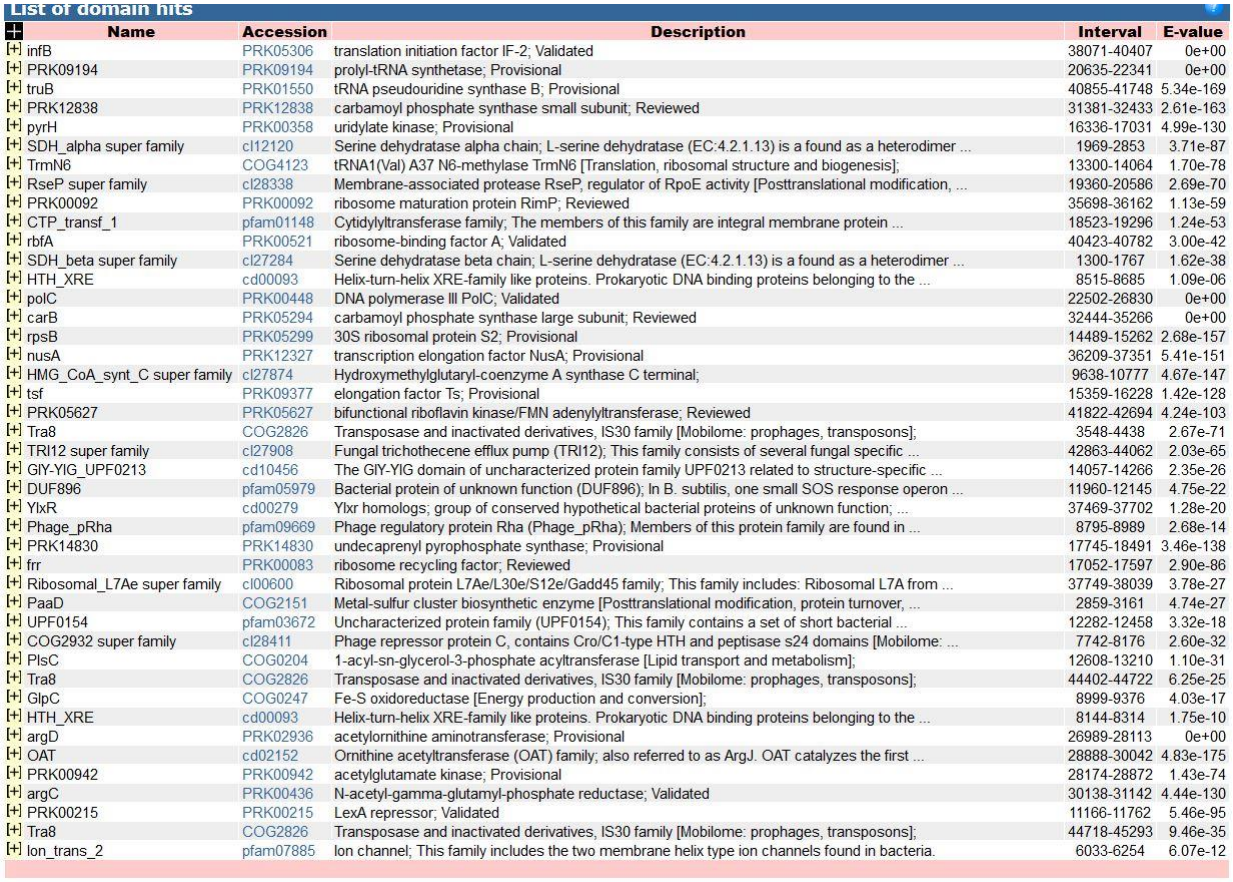
Figure 4.7 | ATCC 14931 CRISPR Expression. (A) Loci comparison of the putative Type II-A CRISPR-Cas loci in ATCC 14931 (top) and LT906621 (bottom). *cas* genes are annotated. The repeat spacer arrays are visualized with the standard diamond/rectangle scheme. mRNA data for ATCC 14931's loci is overlaid. smRNA-seq expression profiles for the (B) repeat spacer array, (C) crRNA (yellow), (D), ldrRNA (green), and (E) tracrRNA (blue), the sequencing coverage is on the y-axis. In (B), the right y-axis shows the log transformed coverage for the boxplots. The leader is highlighted in green. In (C), the spacer (rectangle) and repeat (diamond) boundaries are labelled. In (D), the leader and repeat (diamond) boundaries are labelled. In (E), the predicted boundary is in gray and the actual is labelled. In (F) WebLogo of the proposed PAM sequence for the system. (G) Predicted structure of the tracrRNA with the lower stem (green), bulge (yellow), nexus (blue), and hairpin (red) highlighted.



List of domain hits

Name	Accession	Description	Interval	E-value
HTH_21 super family	cl26233	HTH-like domain; This domain contains a predicted helix-turn-helix suggesting a DNA-binding ...	1132-1959	1.47e-131
Glyco_tranf_GTA_type super family	cl11394	Glycosyltransferase family A (GT-A) includes diverse families of glycosyl transferases with a ...	1-525	3.70e-60
BaeS	COG0642	Signal transduction histidine kinase [Signal transduction mechanisms];	13504-14502	1.10e-49
AzlC	pfam03591	AzlC protein;	16210-16629	3.76e-28
AzlD	pfam05437	Branched-chain amino acid transport protein (AzlD); This family consists of a number of ...	17086-17415	9.21e-12
Abhydrolase super family	cl21494	alpha/beta hydrolases; A functionally diverse superfamily containing proteases, lipases, ...	19477-19746	4.25e-08
Transposase_mut super family	cl27632	Transposase, Mutator family;	5855-6979	8.15e-113
OmpR	COG0745	DNA-binding response regulator, OmpR family, contains REC and winged-helix (wHTH) domain ...	12485-13168	3.00e-77
Tra8	COG2826	Transposase and inactivated derivatives, IS30 family [Mobilome: prophages, transposons];	2096-2602	3.77e-45
Stealth_CR2	pfam11380	Stealth protein CR2, conserved region 2; Stealth_CR2 is the second of several highly conserved ...	4889-5203	3.75e-38
Glyco_tranf_GTA_type	cd00761	Glycosyltransferase family A (GT-A) includes diverse families of glycosyl transferases with a ...	3908-4138	2.84e-08
WcaA	COG0463	Glycosyltransferase involved in cell wall bisynthesis [Cell wall/membrane/envelope biogenesis]; ...	3899-4768	8.57e-06
Stealth_CR1 super family	cl25328	Stealth protein CR1, conserved region 1; Stealth_C1 is the first of several highly conserved ...	4790-4861	8.28e-04
DacC	COG1686	D-alanyl-D-alanine carboxypeptidase [Cell wall/membrane/envelope biogenesis];	14673-15794	1.53e-57
MATE_like super family	cl09326	Multidrug and toxic compound extrusion family and similar proteins; The integral membrane ...	7131-8339	2.05e-24
HTH super family	cl21459	Helix-turn-helix domains; A large family of mostly alpha-helical protein domains with a ...	819-980	3.63e-04
Transposase_mut super family	cl27632	Transposase, Mutator family;	10623-10937	7.86e-29
Transposase_mut super family	cl27632	Transposase, Mutator family;	8964-9278	4.25e-15
DUF1828	pfam08861	Domain of unknown function DUF1828; This presumed domain is functionally uncharacterized.	11468-11662	4.24e-11
Cfa	COG2230	Cyclopropane fatty-acyl-phospholipid synthase and related methyltransferases [Lipid transport ...	20242-21069	7.09e-112
Transposase_mut super family	cl27632	Transposase, Mutator family;	9379-10491	1.11e-104
Cupin_5	pfam06172	Cupin superfamily (DUF985); Family of uncharacterized proteins found in bacteria and ...	11866-12270	3.81e-53
Transposase_mut super family	cl27632	Transposase, Mutator family;	10945-11364	2.18e-35

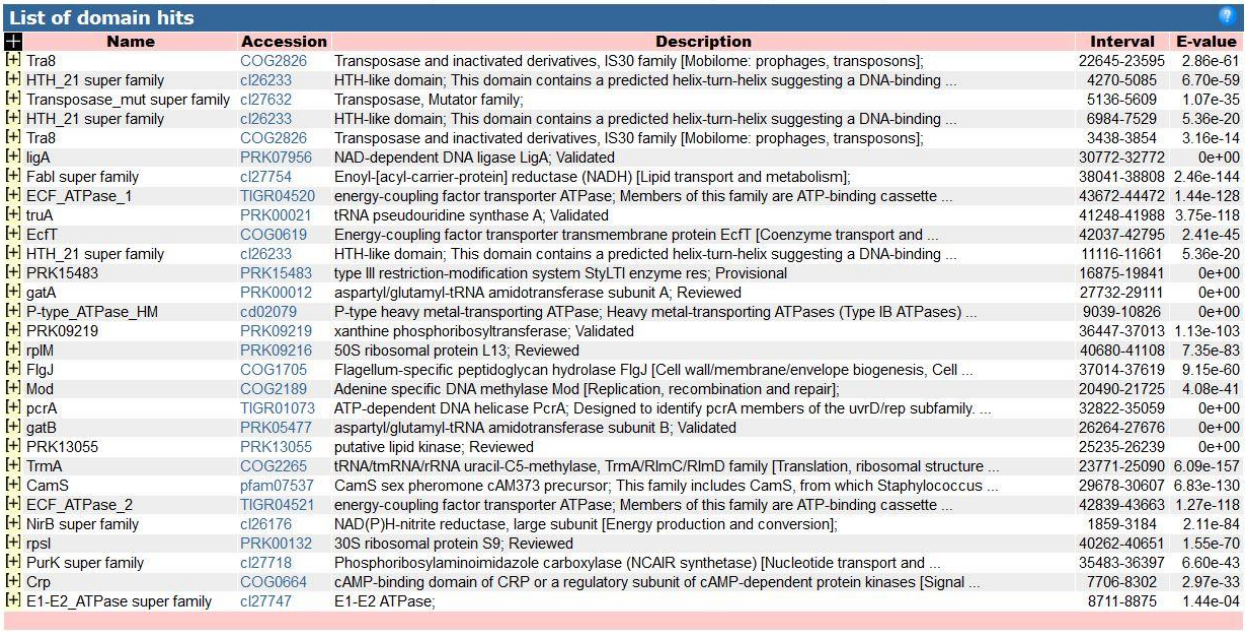
Figure 4.S1 | GC Island at 180kpb. Results of a NCBI Conserved Domain Search of the GC island at 180kpb in *L. fermentum* ATCC 14931.



List of domain hits

Name	Accession	Description	Interval	E-value
infB	PRK05306	translation initiation factor IF-2; Validated	38071-40407	0e+00
PRK09194	PRK09194	prolyl-tRNA synthetase; Provisional	20635-22341	0e+00
truB	PRK01550	tRNA pseudouridine synthase B; Provisional	40855-41748	5.34e-169
PRK12838	PRK12838	carbamoyl phosphate synthase small subunit; Reviewed	31381-32433	2.61e-163
pyrH	PRK00358	uridylate kinase; Provisional	16336-17031	4.99e-130
SDH_alpha super family	cl12120	Serine dehydratase alpha chain; L-serine dehydratase (EC:4.2.1.13) is a found as a heterodimer ...	1969-2853	3.71e-87
TrmN6	COG4123	tRNA1(Val) A37 N6-methylase TrmN6 [Translation, ribosomal structure and biogenesis];	13300-14064	1.70e-78
RseP super family	cl28338	Membrane-associated protease RseP, regulator of RpoE activity [Posttranslational modification, ...	19360-20586	2.69e-70
PRK00092	PRK00092	ribosome maturation protein RimP; Reviewed	35698-36162	1.13e-59
CTP_transf_1	pfam01148	Cytidylyltransferase family; The members of this family are integral membrane protein ...	18523-19296	1.24e-53
rbfA	PRK00521	ribosome-binding factor A; Validated	40423-40782	3.00e-42
SDH_beta super family	cl27284	Serine dehydratase beta chain; L-serine dehydratase (EC:4.2.1.13) is a found as a heterodimer ...	1300-1767	1.62e-38
HTH_XRE	cd00093	Helix-turn-helix XRE-family like proteins. Prokaryotic DNA binding proteins belonging to the ...	8515-8685	1.09e-06
polC	PRK00448	DNA polymerase III PolC; Validated	22502-26830	0e+00
carB	PRK05294	carbamoyl phosphate synthase large subunit; Reviewed	32444-35266	0e+00
rpsB	PRK05299	30S ribosomal protein S2; Provisional	14489-15262	2.68e-157
nusA	PRK12327	transcription elongation factor NusA; Provisional	36209-37351	5.41e-151
HMG_CoA_synt_C super family	cl27874	Hydroxymethylglutaryl-coenzyme A synthase C terminal;	9638-10777	4.67e-147
tsf	PRK09377	elongation factor Ts; Provisional	15359-16228	1.42e-128
PRK05627	PRK05627	bifunctional riboflavin kinase/FMN adenylyltransferase; Reviewed	41822-42694	4.24e-103
Tra8	COG2826	Transposase and inactivated derivatives, IS30 family [Mobilome: prophages, transposons];	3548-4438	2.67e-71
TRI12 super family	cl27908	Fungal trichothecene efflux pump (TRI12); This family consists of several fungal specific ...	42863-44062	2.03e-65
GIY-YIG_UPF0213	cd10456	The GIY-YIG domain of uncharacterized protein family UPF0213 related to structure-specific ...	14057-14266	2.35e-26
DUF896	pfam05979	Bacterial protein of unknown function (DUF896); In B. subtilis, one small SOS response operon ...	11960-12145	4.75e-22
YlxR	cd00279	Ylxr homologs; group of conserved hypothetical bacterial proteins of unknown function; ...	37469-37702	1.28e-20
Phage_pRha	pfam09669	Phage regulatory protein Rha (Phage_pRha); Members of this protein family are found in ...	8795-8989	2.68e-14
PRK14830	PRK14830	undecaprenyl pyrophosphate synthase; Provisional	17745-18491	3.46e-138
frr	PRK00083	ribosome recycling factor; Reviewed	17052-17597	2.90e-86
Ribosomal_L7Ae super family	cl00600	Ribosomal protein L7Ae/L30e/S12e/Gadd45 family; This family includes: Ribosomal L7A from ...	37749-38039	3.78e-27
PaaD	COG2151	Metal-sulfur cluster biosynthetic enzyme [Posttranslational modification, protein turnover, ...	2859-3161	4.74e-27
UPF0154	pfam03672	Uncharacterized protein family (UPF0154); This family contains a set of short bacterial ...	12282-12458	3.32e-18
COG2932 super family	cl28411	Phage repressor protein C, contains Cro/C1-type HTH and peptisase s24 domains [Mobilome: ...	7742-8176	2.60e-32
PlsC	COG0204	1-acyl-sn-glycerol-3-phosphate acyltransferase [Lipid transport and metabolism];	12608-13210	1.10e-31
Tra8	COG2826	Transposase and inactivated derivatives, IS30 family [Mobilome: prophages, transposons];	44402-44722	6.25e-25
GlpC	COG0247	Fe-S oxidoreductase [Energy production and conversion];	8999-9376	4.03e-17
HTH_XRE	cd00093	Helix-turn-helix XRE-family like proteins. Prokaryotic DNA binding proteins belonging to the ...	8144-8314	1.75e-10
argD	PRK02936	acetylornithine aminotransferase; Provisional	26989-28113	0e+00
OAT	cd02152	Ornithine acetyltransferase (OAT) family; also referred to as ArgJ. OAT catalyzes the first ...	28888-30042	4.83e-175
PRK00942	PRK00942	acetylglutamate kinase; Provisional	28174-28872	1.43e-74
argC	PRK00436	N-acetyl-gamma-glutamyl-phosphate reductase; Validated	30138-31142	4.44e-130
PRK00215	PRK00215	LexA repressor; Validated	11166-11762	5.46e-95
Tra8	COG2826	Transposase and inactivated derivatives, IS30 family [Mobilome: prophages, transposons];	44718-45293	9.46e-35
Ion_trans_2	pfam07885	Ion channel; This family includes the two membrane helix type ion channels found in bacteria.	6033-6254	6.07e-12

Figure 4.S2 | GC Island at 760kpb. Results of a NCBI Conserved Domain Search of the GC island at 760kpb in *L. fermentum* ATCC 14931.



List of domain hits

Name	Accession	Description	Interval	E-value
Tra8	COG2826	Transposase and inactivated derivatives, IS30 family [Mobilome: prophages, transposons];	22645-23595	2.86e-61
HTH_21 super family	cl26233	HTH-like domain; This domain contains a predicted helix-turn-helix suggesting a DNA-binding ...	4270-5085	6.70e-59
Transposase_mut super family	cl27632	Transposase, Mutator family;	5136-5609	1.07e-35
HTH_21 super family	cl26233	HTH-like domain; This domain contains a predicted helix-turn-helix suggesting a DNA-binding ...	6984-7529	5.36e-20
Tra8	COG2826	Transposase and inactivated derivatives, IS30 family [Mobilome: prophages, transposons];	3438-3854	3.16e-14
ligA	PRK07956	NAD-dependent DNA ligase LigA; Validated	30772-32772	0e+00
FabI super family	cl27754	Enoyl-[acyl-carrier-protein] reductase (NADH) [Lipid transport and metabolism];	38041-38808	2.46e-144
ECF_ATPase_1	TIGR04520	energy-coupling factor transporter ATPase; Members of this family are ATP-binding cassette ...	43672-44472	1.44e-128
truA	PRK00021	tRNA pseudouridine synthase A; Validated	41248-41988	3.75e-118
EcfT	COG0619	Energy-coupling factor transporter transmembrane protein EcfT [Coenzyme transport and ...	42037-42795	2.41e-45
HTH_21 super family	cl26233	HTH-like domain; This domain contains a predicted helix-turn-helix suggesting a DNA-binding ...	11116-11661	5.36e-20
PRK15483	PRK15483	type III restriction-modification system StyLTI enzyme res; Provisional	16875-19841	0e+00
gatA	PRK00012	aspartyl/glutamyl-tRNA amidotransferase subunit A; Reviewed	27732-29111	0e+00
P-type_ATPase_HM	cd02079	P-type heavy metal-transporting ATPase; Heavy metal-transporting ATPases (Type IB ATPases) ...	9039-10826	0e+00
PRK09219	PRK09219	xanthine phosphoribosyltransferase; Validated	36447-37013	1.13e-103
rplM	PRK09216	50S ribosomal protein L13; Reviewed	40680-41108	7.35e-83
FlgJ	COG1705	Flagellum-specific peptidoglycan hydrolase FlgJ [Cell wall/membrane/envelope biogenesis, Cell ...	37014-37619	9.15e-60
Mod	COG2189	Adenine specific DNA methylase Mod [Replication, recombination and repair];	20490-21725	4.08e-41
pcrA	TIGR01073	ATP-dependent DNA helicase PcrA; Designed to identify pcrA members of the uvrD/rep subfamily. ...	32822-35059	0e+00
gatB	PRK05477	aspartyl/glutamyl-tRNA amidotransferase subunit B; Validated	26264-27676	0e+00
PRK13055	PRK13055	putative lipid kinase; Reviewed	25235-26239	0e+00
TrmA	COG2265	tRNA/tmRNA/rRNA uracil-C5-methylase, TrmA/RlmC/RlmD family [Translation, ribosomal structure ...	23771-25090	6.09e-157
CamS	pfam07537	CamS sex pheromone cAM373 precursor; This family includes CamS, from which Staphylococcus ...	29678-30607	6.83e-130
ECF_ATPase_2	TIGR04521	energy-coupling factor transporter ATPase; Members of this family are ATP-binding cassette ...	42839-43663	1.27e-118
NirB super family	cl26176	NAD(P)H-nitrite reductase, large subunit [Energy production and conversion];	1859-3184	2.11e-84
rpsI	PRK00132	30S ribosomal protein S9; Reviewed	40262-40651	1.55e-70
PurK super family	cl27718	Phosphoribosylaminoimidazole carboxylase (NCAIR synthetase) [Nucleotide transport and ...	35483-36397	6.60e-43
Crp	COG0664	cAMP-binding domain of CRP or a regulatory subunit of cAMP-dependent protein kinases [Signal ...	7706-8302	2.97e-33
E1-E2_ATPase super family	cl27747	E1-E2 ATPase;	8711-8875	1.44e-04

Figure 4.S3 | GC Island at 1550kpb. Results of a NCBI Conserved Domain Search of the GC island at 1550kpb in *L. fermentum* ATCC 14931.

CHAPTER 5: ADAPTIVE RESPONSE TO ITERATIVE PASSAGES OF FIVE *LACTOBACILLUS* SPECIES IN SIMULATED VAGINAL FLUID

5.1. CONTRIBUTION TO THE WORK

The following is a manuscript in preparation titled "Adaptive Response to Iterative Passages of Five *Lactobacillus* Species in Simulated Vaginal Fluid" with the following authorship:

Katelyn Brandt[1,2] and Rodolphe Barrangou[1,2]

[1]Functional Genomics Graduate Program, North Carolina State University, Raleigh, NC 27695

[2]Department of Food, Bioprocessing and Nutrition Sciences, North Carolina State University, Raleigh, NC 27695

Katelyn Brandt is first author on this publication. She carried out experiments and analyzed data. Katelyn Brandt also wrote the manuscript.

5.1. ABSTRACT

Abstract: Microbiome and metagenomic studies have given rise to a new understanding of microbial colonization of various human tissues and their ability to impact our health. One surprisingly under-studied human microbiome, the vaginal microbiome, stands out given its importance for women's health, and is peculiar in terms of its relative bacterial composition simplicity and typical domination by a small number of *Lactobacillus* species. The loss of *Lactobacillus* dominance is associated with disease states such as bacterial vaginosis, and efforts are now underway to understand the ability of *Lactobacillus* species to colonize the vaginal tract and adapt to this dynamic and acidic environment. Here, we investigate how various *Lactobacillus* species often isolated from the vaginal and intestinal cavities genomically and transcriptionally respond to iterative growth in simulated vaginal fluid. We determined the genomes and transcriptomes of *L. acidophilus*, *L. crispatus*, *L. fermentum*, *L. gasseri*, and *L. jensenii* and compared profiles after 50, 100, 500, and 1,000 generations. Overall, we identified relatively few genetic changes consisting of single nucleotide polymorphisms, with more mutations for non-vaginal isolated. Transcriptional profiles were more impacted over time and quantitatively more extensive for non-vaginal species, reflecting a more extensive need to adapt to a more unfamiliar environment.

Importance: With greater understanding of how the microbiome affects the host and impacts human health, recent studies are opening new avenues for the development of probiotics for areas beyond the gastrointestinal tract, notably in women's health. In order to properly develop probiotic formulations, we must first determine the composition and functionalities of the vaginal microbiome, and identify the genetic and functional features that drive adaptation to this niche and its environmental conditions. Thus, we determined the genetic and transcriptional

profile of intestinal and vaginal *Lactobacillus* species over time during extensive culturing in a simulated vaginal fluid to assess the bacterial basis for adaptation to this environment.

5.3. INTRODUCTION

It was originally documented in 1892 that the presence of *Lactobacillus* is the hallmark of a healthy vaginal microbiome [1]. However, it took over a century and the establishment of the Human Microbiome Project (HMP) in 2007 to truly start to understand and extensively study how microbes and microbial genes affect their human hosts, contribute to disease, and impact health [2, 3]. The project has spurred several subsequent microbiome studies, and in 2010, a landmark study shed light on the 1892 observation. In a large cohort, the study showed that the vaginal microbiome of women is not only largely dominated by lactobacilli, but often primarily consists of a single species of *Lactobacillus.* The study identified four state types that were each dominated by a select *Lactobacillus* species: *Lactobacillus crispatus, Lactobacillus iners, Lactobacillus gasseri,* and *Lactobacillus jensenii* [4]. These results were confirmed by subsequent studies, in different populations [5-7]. The HMP revealed that the vaginal microbiome is qualitatively and quantitatively distinct from other sites of the human body, especially with regards to composition and relatively limited diversity [3]. This indicates that different principles shape vaginal microbiome composition and function, and that a specific approach should be considered to understand how these bacteria impact vaginal health and determine how to approach the development of probiotic products for women's health.

A loss of lactobacilli or a dysbiosis in the vaginal microbiome typically correlates with Bacterial Vaginosis (BV) [8]. Treatment of BV results in a cure rate of 80% but a re-occurrence rate of 50% [9]. Additionally, BV is the cause of over 50% of visits by women to health clinics, and as such has been the focus of several studies [9]. Specifically, as the etiology of BV is unknown [10], several groups have been investigating how the microbiome fluctuates to a BV state in order to better manage this condition. Drawing from the success of probiotics in the

gastrointestinal tract, several groups have thought to develop a women's probiotic using *Lactobacillus* with varied results [11]. However, we still have relatively little knowledge and a shallow understanding of how community state types (CSTs) associate, correlate, or possibly cause disease states such as BV. Some correlations have been established and *L. crispatus* is typically considered protective against BV, while *L. iners* is considered an in-between state. [12, 13].There is limited knowledge with regards to which phylogenetic units, let alone species or strains, may provide the most promise for vaginal health.

In order to be an effective probiotic, a species must presumably be able to survive environmental conditions, grow using available resources, and ideally colonize the niche of interest. In this study, we set out to examine how various strains from five distinct *Lactobacillus* species isolated from vaginal and intestinal environments alter their genomes and transcriptomes during long-term serial passages in a simulated vaginal fluid. We specifically selected *L. crispatus, L. gasseri,* and *L. jensenii* strains to represent CSTs associated with a healthy vaginal status [4], as well as *Lactobacillus acidophilus* and *Lactobacillus fermentum,* strains found in the intestinal cavity that have also been occasionally used commercially as potential vaginal probiotics [14, 15].

5.4. MATERIALS AND METHODS

5.4.1. Strain Selection

In order to evaluate the adaptive response to the vaginal environment, five species of *Lactobacillus* were selected: *L. crispatus* JV-V01, *L. gasseri* JV-V03, *L. jensenii* DSM-20557, *L. acidophilus* NCFM, and *L. fermentum* ATCC 14931. These strains represent vaginal species (*L. crispatus, L. gasseri,* and *L. jensenii*), intestinal contaminants (*L. acidophilus*), and an outgroup (*L. fermentum*). Table 5.1 contains information on each strain, encompassing their source of isolation as well as their genomic profiles. *L. acidophilus* and *L. gasseri* both had publically available genomes [16]. The *L. fermentum*'s genome was completed as previously described [Chapter 4]. The *L. jensenii*'s genome, RB055, was sequenced at CoreBiome (St. Paul, MN). Briefly, samples were isolated using MO Bio PowerFecal (Qiagen) for high-throughput on QiaCube (beading beating 0.1mm). DNA quantification was determined using Invitrogen's Qiant-iT Picogreen dsDNA Assay. Library preparation was performed using the Nextera Library Prep kit (Illumina). Sequencing was carried out using Illumina NextSeq 500/550 High Output V2 kit using paired-end reads (2 x 150 reads). Reads were then filtered (Q-score <20; length <50) and adapter sequences were removed using cutadapt (v1.15). Sequences were then assembled using SPAdes (v3.11.0) and QUAST (v4.5). We then annotated RB055 using RAST. The *L. crispatus's* genome was sent to the Roy J. Carver Biotechnology Centre from the University of Illinois (Urbana-Champaign, IL) for DNA extraction and sequencing. Briefly, library preparation for short reads was completed using the Hyper Library construction kit from Kapa Biosystems (Roche). It was then quantified using qPCR before sequencing on a MiSeq flow cell using a MiSeq 500-cycle sequencing kit (v2). Demultiplexing was achieved using bcl2fastq (v2.20) Conversion Software (Illumina). Long reads were generated by converting 600ng of DNA to

barcoded Nanopore libraries suing the Rapid Barcoding kit SQK-RAD004. Sequencing then occurred using a GridION x5 sequencer on SpotON R9.4.1 FLO-MIN106 flow cells. Guppy (v1.8.1) was used for base-calling and Porechops (v0.2.3) was used for demultiplexing and adapter removal. Reads were then assembled using the Unicycler assembler (v0.4.4) and annotated using Prokka (v1.12).

5.4.2. Growth Conditions

Our lab recently developed a simulated vaginal fluid to mimic the vaginal environment and conditions in a laboratory setting (SFV, Table 5.2). Based off of a semi-defined medium, SVF incorporates nutrients and enzymes common to the vaginal environment, such as glycogen, α-amylase, and a pH of 4.5 [17]. Directed evolution was achieved by passaging these strains for 1,000 generations in preconditioned, anaerobic SVF. We assessed adaptation at five generations: 0, 50, 100, 500, and 1,000. These generations were used throughout this study. Generation 0 is defined as the first passage from freezer stock in MRS to growth in SVF and represents the first naïve exposure to a vaginal environment. This serves as our reference point for all other time points analyzed.

Growth medium was preconditioned for 24 hours before inoculation. At inoculation, α-amylase (2mU/mL) was added. Passages were inoculated 1% (vol/vol) from the previous day. Strains were grown for 24 hours anaerobically at 37°C, and passed iteratively for a total of 1,000 generations (6.6 generations/transfer for 151 transfers). After reaching a generation of interest, cultures were stocked during the following inoculation.

For growth curves, MRS and SVF were preconditioned for 24 hours with cysteine (0.05%, wt/vol). Cultures were inoculated from freezer stocks (of the desired generation) into

preconditioned SVF (MRS for generation 0) and grown for 24 hours anaerobically at 37°C. 96-well microplates were inoculated 1% (v/v) from the overnight culture in triplicate. Each well contained 200 µL of SVF or MRS. A clear Thermalseal film was used to then seal the plates. Using a Floustar Optima microplate reader (BMG Labtech, Cary, NC), plates were read at 37°C, 600nm for 48 hours. Growth curves were completed three times. Analyses were carried out between the biological replicates (averages of the technical replicates).

5.4.3. Transcriptome Analysis

Cells were inoculated from freezer stock (of the desired generation) and grown overnight in preconditioned SVF (MRS for generation 0) anaerobically at 37°C. Cells were then passed (1% v/v) and grown to mid-log phase and flash frozen. Zymo Direct-Zol Miniprep kit (Zymo Research, Irvine, CA) was used to extract total RNA, as previously described [18]. mRNA library preparation and sequencing were completed at the Roy J. Carver Biotechnology Centre from the University of Illinois (Urbana-Champaign, IL) using an Illumina HiSeq2500. Data was uploaded into Geneious (v. 11.1.5, https://www.geneious.com). Read processing included trimming (error probability limit of 0.001) and filtering (length extracted 28-150 nt). Reads were then mapped to the reference genome using Bowtie2 [19].

5.4.4. Data Analysis

Statistical analyses were completed using R (v3.5.1). Expression levels were calculated in Geneious (count as partial matches/CDS). Levels were compared using Geneious's built in method (transcripts/median of gene expression rations). Gene significance was determined using the Geneious metrics Differential Expression Absolute Confidence ($\geq|6|$) and the Differential

Expression Log2 Ratio ($\geq|2|$). Significantly expressed genes were hand curated from Geneious and then assigned Clusters of Orthologous Groups (COGs) by eggnog-mapper, based on eggnog 4.5 data [20, 21]. SNP variations were calculated using Geneious and default settings. Generation 0 was used as the reference genome for SNP calling. SNPs were additionally filtered by removing SNPs generated through runs and a coverage cutoff of 20. Chromosomal location of SNPs from non-closed genomes was determined by ordering contigs by size. Mutation rates were determined by total number of SNPs per genome size per generation. Student's t-test ($\alpha=0.05$) was used to determine significance between groups.

5.5. RESULTS AND DISCUSSION

5.5.1. Bacterial Growth in Simulated Vaginal Fluid

In this study, we sought to understand how *Lactobacillus* species adapted to a simulated vaginal environment over time. Directed evolution and adaptation have been studied in bacteria for several species under various conditions through iterative passages [22-25]. Here, we examined how vaginal and non-vaginal strains from five distinct *Lactobacillus* species adapted to iterative passages for 1,000 generations in SVF. Strain selection was guided by several recent studies highlighting their occurrence and potential role in women's health. From the establishment of CSTs, *L. crispatus, L. gasseri,* and *L. jensenii* strains were included as representatives of normal microbiomes. We strategically elected to forego *L. iners* due to its controversial status, potentially linked to its different properties of the other CST species, as well the practical need this species has to grow in blood, rendering it an impossible comparative growth outlier [13, 26]. We chose to include *L. acidophilus* as a representative of an intestinal strain and potential vaginal contaminant to compare and contrast vaginal vs. intestinal isolates. Additionally, *L. acidophilus* is a well-known gut probiotic and has been proposed and even commercialized as a vaginal probiotic [27, 28]. Finally, *L. fermentum,* isolated from beets, was used as an outgroup, as a food isolate which can also be found the human GIT, though this species is also a candidate for women's health applications [29, 30].

To date, the primary concept underlying protection of the vaginal environment by *Lactobacillus* species colonization is competitive exclusion [31], which presumably hinges on the ability of these strains to survive and grow in the vaginal environment. Therefore, we began our analyses by comparing strain growth in MRS (primary defined medium for *Lactobacillus* species growth in a laboratory setting) and SVF [17]. As anticipated, all strains grew well in

MRS and did not showcase altered growth in this medium over the time course of the experiment. As expected, strains grew better in the *Lactobacillus*-optimized medium, MRS, than they did in SVF (Figure 5.1). *L. fermentum* grew the best, while *L. jensenii* grew the least in MRS. In terms of SVF, the vaginal representative strains grew better than their intestinal counterparts, with *L. crispatus* showcasing highest growth, followed by *L. gasseri* and *L. jensenii*. Over time, *L. crispatus* and *L. gasseri* showed relatively little growth change across generations, presumably because they are already adapted to the environmental conditions provided by this simulated vaginal fluid. Towards the latter stages of the experiment, it appears the growth of *L. acidophilus* increases, though the improvement may be relatively limited and growth benefits would likely warrant a more extensive period of time.

5.5.2. Genomic Variations

Having carried out iterative passages over time and sampled the strains at select time points through the 1,000 generations of growth, we next determined their genome sequence to monitor genetic plasticity over time. We did this by mapping draft genome content on the reference genome to assess the occurrence of genetic mutations such as insertions, deletions, duplications and the appearance of single nucleotide variations in a population of cells using deep sequencing. SNPs were called from mRNA-Seq data using Geneious. We then filtered SNPs that were part of runs or had coverage less than 20 reads. Figure 5.2 depicts the number of SNPs over time. Both fixed SNPs (changes observed in 100% of the reads) and partial SNPs (changes observed in over 50% of reads) were mapped. Based on similar studies with *Escherichia coli*, we anticipated adaptation at an early passage stage, with less modifications at the latter stages [32], and our results suggest that indeed the bulk of fixed mutations did occur by

generation 50, though some SNPs appeared thereafter. *L. gasseri, L. acidophilus,* and *L. fermentum* overall had the most amount of SNPs at 50% frequency. *L. gasseri, L. jensenii,* and *L. fermentum* had more SNPs in early generations and either remained steady or dropped in subsequent generations (Figure 5.2A). The *L. acidophilus* SNPs increased over time, in a pattern somewhat consistent with the aforementioned increase in growth in SVF over the course of the experiment. Intriguingly, *L. crispatus* had very little SNPs at 50% frequency (Figure 5.2A). The total SNPs at 50% frequency for *L. crispatus* was significantly different compared to all other species with the exception of *L. jensenii* (Figure 5.2B). For fixed SNPs, the five species distributed in the same two groups, with *L. gasseri, L. acidophilus,* and *L. fermentum* encompassing the high-SNP group, and *L. crispatus* and *L. jensenii* constituting the low group (Figure 5.2C). Within the high group, *L. gasseri* had the most SNPs of the group at generation 50, *L. fermentum* at generations 100 and 500, and *L. acidophilus* at generation 1,000. Of the low group, *L. crispatus* had no fixed SNPs at any time point. *L. crispatus* was significantly different from all species. *L. jensenii* and *L. fermentum* were also significantly different (Figure 5.2D). Mutation rates, defined as number of SNPs per genome size per generation, can be found in Table 5.3. With the exception of *L. crispatus*, mutations rates were similar to those reported for *Lactobacillus rhamnosus*, but slightly higher than those in *E. coli* [33, 34]. None of the strains took on a mutator phenotype [23, 33, 35].

After enumerating the number of SNPs per generation, we next examined how individual SNPs change over time. Figure 5.3 depicts SNPs at 50% frequency (open circles) and 100% frequency (closed circles) at chromosome locations. SNPs are also separated based on the generation they were called. Due to the level of magnification shown on the figure, some SNPs are indistinguishable from others, as in *L. gasseri.* One gene, *rpn family recombination-*

promoting nuclease/putative transposase, had four SNPs occur in subsequent generations, all of which resulted in synonymous changes. Here, it can be seen in most strains that once a SNP is fixed, it typically remains in successive generations. This is most noticeable in *L. gasseri*, whose majority of SNPs fix early and remain fixed. In contrast, *L. fermentum* continuously acquires and loses SNPs, although overall numbers remain fairly steady. This diversity reflects the presence of a mixed microbial population and the rise and fall of various genotypes with no apparent gain of fitness over the course of the experiment. This contrasts with *L. acidophilus*, which increases in total SNPs and especially fixed SNPs in later generations, which is noteworthy and could reflect a functional benefit acquired over time by some members of the mixed population. Overall, mutations occurred throughout the chromosomes, at various locations across species with no common hot-spot, similar to previous studies [33]. It is noteworthy that we did not detect any insertion, deletion, duplication or genomic re-arrangement in any of these five species, perhaps reflecting the overall genetic stability of these strains in these conditions (Figure 5.4). In fact, evolution studies in other LAB species (*L. rhamnosus* and *Lactococcus lactis*) identified large deletions, which was contributed to insertion sequence (IS) elements [33, 35, 36].

5.5.3. Transcriptional Response to Growth in Simulated Vaginal Fluid

After determining the genomic changes over time, we next examined how the transcriptome changed over time to assess whether the regulation of select transcripts reflected adaptation to SVF composition and substrates. We began by comparing expression levels between generation 0 and subsequent generations for each strain (Figure 5.5). Overall, relatively few transcripts showed extensive differential expression. However, there were significant differences in expression for each species, highlighted by colored dots (Figure 5.5). The greatest

number of significant changes were found in *L. gasseri* and *L. fermentum*. *L. jensenii* and *L. crispatus* had the least number of changes (Table 5.4). In fact, there were no significant changes in expression between generation 0 and generation 50 for *L. crispatus*.

After determining which genes were significantly differentially expressed, we investigated the potential functions of these genes. First, we ascribed COG designation for each significantly expressed gene using eggnog. COG definitions can be found in Table 5.S1. Then, we determined the number of genes that were upregulated or downregulated by COG designation (Figure 5.6). After showing no significant changes in generation 50, *L. crispatus* began to show significant changes at tested subsequent generations, with no obvious global trend (Figure 5.6). At generation 100, *L. crispatus* showed the most change in the downregulation of genes of the COG designation replication, recombination and repair (L) and the upregulation of genes with unassigned COGs (NA). Profiles for generation 500 and generation 1,000 both showed downregulation of unassigned COGs (NA) and upregulation of genes with the designation of unknown function (S). Additionally, profiles for generation 500 and generation 1,000 showed up-regulation in amino acid transport and metabolism (E) and nucleotide transport and metabolism (F), respectively. In contrast, *L. gasseri* largely showed downregulation across its significantly expressed genes (Figure 5.6). Additionally, *L. gasseri* had a large number of significantly expressed genes at generation 50 and gradually increased over time. Of the downregulated genes, most were of the designation carbohydrate transport and metabolism (G), function unknown (S), or unassigned a COG (NA) (Figure 5.6). Of the genes upregulated in *L. gasseri*, most were a part of the nucleotide transport and metabolism (F) designation. At generation 1,000, *L. gasseri* showed the most upregulation across all tested species notably for designations function unknown translation, ribosomal structure and biogenesis (J), nucleotide

transport and metabolism (F), and function unknown (S). *L. jensenii* showed the least amount of significant changes and seemed most defined by downregulation over upregulation (Figure 5.6). In general, most downregulation occurred at generations 500 and 1,000 in carbohydrate transport and metabolism (G). Overall, *L. acidophilus* showed relatively more changes in expression than *L. jensenii*, with no obvious functional patterns (Figure 5.6). The most commonly downregulated genes comprised an unassigned COG function (NA). In terms of upregulation, generation 50 and generation 1,000 stand out for upregulation of nucleotide transport and metabolism (F). *L. fermentum* had the most significant changes at generations 500 and 1,000 (Figure 5.6). Downregulation was common in both, specifically in amino acid transport and metabolism (E) and carbohydrate transport and metabolism (G). Upregulation occurred the most in generation 500 and occurred primarily in translation, ribosomal structure in biogenesis (J), and function unknown (S). Overall, downregulation was common, which perhaps reflects the rich composition of MRS, in which more substrates are available. Across species, the COG designation (G) carbohydrate transport and metabolism was often downregulated and (F) nucleotide transport and metabolism was upregulated. Additionally, changes occurred often in the unknown function (S) and unassigned (NA) groups. These changes have been noted in other studies as well [17]. In contrast to genetic changes, which primarily occurred early on (most SNPs detected by generation), it appears most transcriptional impact was detected at later generations, suggesting a more significant impact over time on the transcriptome than the genome (Table 5.4).

5.5.4. Selected Transcription

Finally, we highlighted one locus from each species (Figure 5.7). These genes were selected from Figure 5.5 based on their consistent significance, and have been marked by an

asterisk (*). Figure 5.7 depicts a select locus highly impacted for each species and the corresponding transcriptional profile at each analyzed generation. Each gene represented a different COG: *L. crispatus*-unassigned, *L. gasseri*-(G), *L. jensenii*-(O), *L. acidophilus*-(S), and *L. fermentum*-(E) and (G). *L. crispatus* and *L. fermentum* loci both had decreasing transcription as generations passed. *L. gasseri* and *L. jensenii* had genes that downregulated early and remained around the same level. For *L. crispatus,* a hypothetical gene was selected. This gene was one of three hypothetical genes that were significantly downregulated at generations 500 and 1,000, as compared to generation 0. It was not assigned a COG, and this group was more downregulated in generations 500 and 1,000. A domain search of the hypothetical gene depicted returned no known domains. From *L. gasseri,* a *PTS fructose transporter subunit II ABC* locus is depicted (Figure 5.7). It is one of two genes that were consistently downregulated in all generations compared to generation 0 and was easily distinct in two-way plots (Figure 5.5) (the other *pfkb*). Its assigned COG group (G) was downregulated in all generations (Figure 5.6). A clpE-like protein locus is depicted for *L. jensenii* (Figure 5.7). It was one of the few genes that showed significant expression between generations 0 and 50, and remained significantly expressed in subsequent generations (Figure 5.5). Its expression profile for generations 50, 100, 500, and 1,000 was quite consistent. The *clpE-like* gene was assigned the COG (O) for post-translational modification, protein turnover, and chaperones. Though a small group, it was consistently downregulated across generations, while a few of this group were upregulated in generation 1,000 (Figure 5.6). *L. fermentum* had many genes significantly expressed across generations and a few were expressed in subsequent generations (Figure 5.5). One such gene was *gluconate permease,* which was downregulated in all generations compared to 0; its expression profile is shown in Figure 5.7. It was assigned both an (E) and (G) COG designation. Each group

was downregulated in generation 50 and 100. However, in generations 500 and 1,000, genes of each group were both down- and upregulated (Figure 5.6).

The locus in *L. acidophilus* was perhaps the most interesting. This was the only highlighted loci that also had an acquired nonsynonymous SNP (Figure 5.7), reflecting a convergence between the transcriptional response and genetic mutation pattern over the course of the experiment. This was the only gene to show subsequent SNPs in the same location and a correlated significant change in expression. The SNP was a change from an A to a C (transversion) and caused an amino acid change from a glutamic acid to alanine. Intriguingly, this SNP was acquired in generation 500 and remained in generation 1,000. In correlation, there was no significant change of expression until generation 500. It was significantly downregulated in generation 500 and generation 1,000 but was not significantly expressed in generation 50 and generation 100 (Figure 5.5). The locus (LBA1020) is annotated as a mucus binding protein. It had no homologs (threshold of 40% similarity) in any of the other *Lactobacillus* strains in this study. Additionally, no other works reference this locus. Despite this, it is highly interesting because mucin is added to SVF, indicating a direct adaption.

5.6. CONCLUSION

In conclusion, we established how vaginal and non-vaginal strains adapt to a simulated vaginal environment. Previous directed evolution experiments in LAB showed niche-specific and stress adaptations after 1,000 generations, with defined phenotypes [33, 35], however we did not see strong changes in phenotype. The only correlation between passage, genomic change, and transcription change was the mucin binding locus in *L. acidophilus*, as mucin is a component of SVF. As this locus has not been fully characterized, it is a potential feature of interest for future studies. Additionally, we did not see the previously reported large chromosome changes or mutator phenotypes [33, 35]. We did see greater growth in SVF and less mutations in our vaginal strains, specifically *L. crispatus*, as compared to the non-vaginal strains.

Overall, vaginal strains showed better growth in SVF, fewer genetic alterations and modest transcriptional changes compared to non-vaginal strains. Additionally, while performing relatively poorly at initial exposure to SVF, non-vaginal strains did not greatly improve over time in a simulated vaginal environment. They had modest growth gains, relatively few genetic changes primarily consisting of SNPs, and limited transcriptional responses, reflecting media composition differences. This implies that the differences between vaginal strains and non-vaginal strains is functionally, genetically, and transcriptionally significant. These findings open new avenues to investigate and characterize members of the vaginal microbiome and inform the genetic and functional assessment of strains of potential value for enhancing women's health, with preference for vaginal over intestinal isolates for candidate probiotics.

5.7. ACKNOWLEDGEMENTS

We would like to thank the CRISPR lab for insights and support for this project. We also thank Dr. Courtney Klotz during figure development.

5.8. REFERENCES

1. Vaneechoutte, M. (2017) The human vaginal microbial community. Research in Microbiology.

2. Turnbaugh, P.J. et al. (2007) The Human Microbiome Project. Nature 449, 804.

3. Lloyd-Price, J. et al. (2017) Strains, functions and dynamics in the expanded Human Microbiome Project. Nature 550 (7674), 61-66.

4. Ravel, J. et al. (2011) Vaginal microbiome of reproductive-age women. Proceedings of the National Academy of Sciences 108 (Supplement 1), 4680-4687.

5. Ravel, J. et al. (2013) Daily temporal dynamics of vaginal microbiota before, during and after episodes of bacterial vaginosis. Microbiome 1 (1), 29.

6. Callahan, B.J. et al. (2017) Replication and refinement of a vaginal microbial signature of preterm birth in two racially distinct cohorts of US women. Proceedings of the National Academy of Sciences 114 (37), 9966-9971.

7. Ma, Z. and Li, L. (2017) Quantifying the human vaginal community state types (CSTs) with the species specificity index. PeerJ 5, e3366.

8. (June 04, 2015) Bacterial Vaginosis https://www.cdc.gov/std/tg2015/bv.htm, (accessed November 27, 2017).

9. Huang, B. et al. (2014) The Changing Landscape of the Vaginal Microbiome. Clinics in laboratory medicine 34 (4), 747-761.

10. Onderdonk, A.B. et al. (2016) The Human Microbiome during Bacterial Vaginosis. Clinical Microbiology Reviews 29 (2), 223-238.

11. Kovachev, S. (2018) Defence factors of vaginal lactobacilli. Critical Reviews in Microbiology 44 (1), 31-39.

12. Madhivanan, P. et al. (2014) Characterization of culturable vaginal *Lactobacillus* species among women with and without bacterial vaginosis from the United States and India: a cross-sectional study. Journal of Medical Microbiology 63 (7), 931-935.

13. Petrova, M.I. et al. (2017) *Lactobacillus iners*: Friend or Foe? Trends in Microbiology 25 (3), 182-191.

14. Zarate, G. and Nader-Macias, M. (2006) Influence of probiotic vaginal lactobacilli on in vitro adhesion of urogenital pathogens to vaginal epithelial cells. Letters in Applied Microbiology 43 (2), 174-180.

15. Reid, G. et al. (2001) Oral probiotics can resolve urogenital infections. FEMS Immunology & Medical Microbiology 30 (1), 49-52.

16. Altermann, E. et al. (2005) Complete genome sequence of the probiotic lactic acid bacterium Lactobacillus acidophilus NCFM. Proc Natl Acad Sci U S A 102 (11), 3906-12.

17. Pan, M., Characterixation of *Lactobacillus* Strains for Vaginal and Intestinal Applications, Food Science, Nutrition and Bioprocessing, North Carolina State University, Raleigh, NC, 2019.

18. Theilmann, M.C. et al. (2017) *Lactobacillus acidophilus* Metabolizes Dietary Plant Glucosides and Externalizes Their Bioactive Phytochemicals. mBio 8 (6), e01421-17.

19. Langdon, W.B. (2015) Performance of genetic programming optimised Bowtie2 on genome comparison and analytic testing (GCAT) benchmarks. BioData Mining 8 (1), 1.

20. Huerta-Cepas, J. et al. (2017) Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. Molecular Biology and Evolution 34 (8), 2115-2122.

21. Huerta-Cepas, J. et al. (2015) eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Nucleic Acids Research 44 (D1), D286-D293.

22. Barrick, J.E. et al. (2009) Genome evolution and adaptation in a long-term experiment with *Escherichia coli*. Nature 461, 1243.

23. Elena, S.F. and Lenski, R.E. (2003) Evolution experiments with microorganisms: the dynamics and genetic bases of adaptation. Nature Reviews Genetics 4 (6), 457-469.

24. Finkel, S.E. and Kolter, R. (1999) Evolution of microbial diversity during prolonged starvation. Proceedings of the National Academy of Sciences 96 (7), 4023.

25. Jerison, E.R. and Desai, M.M. (2015) Genomic investigations of evolutionary dynamics and epistasis in microbial evolution experiments. Current Opinion in Genetics & Development 35, 33-39.

26. Macklaim, J.M. et al. (2011) At the crossroads of vaginal health and disease, the genome sequence of *Lactobacillus iners* AB-1. Proceedings of the National Academy of Sciences 108 (Supplement 1), 4688.

27. Reid, G. (2000) In vitro testing of *Lactobacillus acidophilus* NCFM^{TM} as a possible probiotic for the urogenital tract. International Dairy Journal 10 (5), 415-419.

28. Borges, S. et al. (2014) The role of lactobacilli and probiotics in maintaining vaginal health. Archives of Gynecology and Obstetrics 289 (3), 479-489.

29. Deidda, F. et al. (2016) In Vitro Activity of *Lactobacillus fermentum* LF5 Against Different *Candida* Species and *Gardnerella vaginalis*: A New Perspective to Approach Mixed Vaginal Infections? Journal of Clinical Gastroenterology 50, S168-S170.

30. do Carmo, M.S. et al. (2016) *Lactobacillus fermentum* ATCC 23271 Displays In vitro Inhibitory Activities against *Candida* spp. Front Microbiol 7, 1722.

31. Ojala, T. et al. (2014) Comparative genomics of *Lactobacillus crispatus* suggests novel mechanisms for the competitive exclusion of *Gardnerella vaginalis*. BMC Genomics 15 (1), 1070.

32. Lenski, R.E. et al. (1991) Long-Term Experimental Evolution in *Escherichia coli.* I. Adaptation and Divergence During 2,000 Generations. The American Naturalist 138 (6), 1315-1341.

33. Douillard, F.P. et al. (2016) Polymorphisms, Chromosomal Rearrangements, and Mutator Phenotype Development during Experimental Evolution of *Lactobacillus rhamnosus* GG. Applied and Environmental Microbiology 82 (13), 3783.

34. Drake, J.W. et al. (1998) Rates of spontaneous mutation. Genetics 148 (4), 1667-1686.

35. Bachmann, H. et al. (2012) Microbial domestication signatures of *Lactococcus lactis* can be reproduced by experimental evolution. Genome Research 22 (1), 115-124.

36. Schneider, D. et al. (2000) Long-term experimental evolution in *Escherichia coli.* IX. Characterization of insertion sequence-mediated mutations and rearrangements. Genetics 156 (2), 477-488.

Table 5.1 | Strains Selection. Selection of *Lactobacillus* strains used in this study.

Species	Strain	Genome	Origin	Representation	Code	Color	Ref
Lactobacillus crispatus	JV-V01	RB287	Vaginal Flora	Vaginal Species	*Lcr*	Purple	This Study
Lactobacillus gasseri	JV-V03	NZ_ACGO	Urogenital	Vaginal Species	*Lga*	Light Green	
Lactobacillus jensenii	DSM20557	RB055	Vaginal Discharge	Vaginal Species	*Lje*	Pink	This Study
Lactobacillus acidophilus	NCFM	NC006814	Milk	Intestinal Contaminant	*Lac*	Light Blue	[16]
Lactobacillus fermentum	ATCC 14931	ATCC 14931	Fermented Beets	Outgroup	*Lfe*	Orange	Chapter 4

Table 5.2 | Simulated Vaginal Fluid Composition [17]. Components of vaginal media used in this study.

Component	**Final Concentration (w/v)**
Tween 80	0.1%
Ammonium citrate	0.2%
Sodium acetate	0.5%
$MgSO_4$-$7H_2O$	0.01%
$MnSO_4$-H_2O	0.005%
K_2HPO_4	0.2%
No. 3 Protease Peptone	0.3%
Urea	0.05%
Glucoses	1%
Glycogen	1%
Lactic Acid	88mM
Mucin	0.025%
Albumin	0.4%
Vitamin Solution	1X
α-Amylase	2mU/mL

Table 5.3 | Mutation Rate. Mutation rate for fixed mutations at generation 1,000.

Species	**# Mutations**	**Genome Size (bp)**	**Mutation Rate**
L. crispatus	0	2318358	0
L. gasseri	6	2011855	2.98×10^{-09}
L. jensenii	3	1610429	1.86×10^{-09}
L. acidophilus	13	1993560	6.52×10^{-09}
L. fermentum	10	1887974	5.30×10^{-09}

Table 5.4 | Significantly Expressed Genes. Number of significantly expressed genes compared to generation 0.

Species	**Generation 50**	**Generation 100**	**Generation 500**	**Generation 1,000**
L. crispatus	0	32	78	67
L. gasseri	83	98	117	191
L. jensenii	5	8	33	59
L. acidophilus	60	53	76	83
L. fermentum	8	67	234	137

Table 5.S1 | COG Description. COG symbols and their descriptions.

COG Symbol	Description
NA	Unassigned
Z	Cytoskeleton
Y	Nuclear structure
W	Extracellular structures
V	Defense mechanisms
U	Intracellular trafficking, secretion, and vesicular transport
T	Signal transduction mechanisms
S	Function unknown
R	General function prediction only
Q	Secondary metabolites biosynthesis, transport, and catabolism
P	Inorganic ion transport and metabolism
O	Post-translational modification, protein turnover, and chaperones
N	Cell motility
M	Cell wall/membrane/envelope biogenesis
L	Replication, recombination and repair
K	Transcription
J	Translation, ribosomal structure, and biogenesis
I	Lipid transport and metabolism
H	Coenzyme transport and metabolism
G	Carbohydrate transport and metabolism
F	Nucleotide transport and metabolism
E	Amino acid transport and metabolism
D	Cell cycle control, cell division, chromosome partitioning
D	Energy production and conversion
B	Chromatin structure and dynamics
A	RNA processing and modification

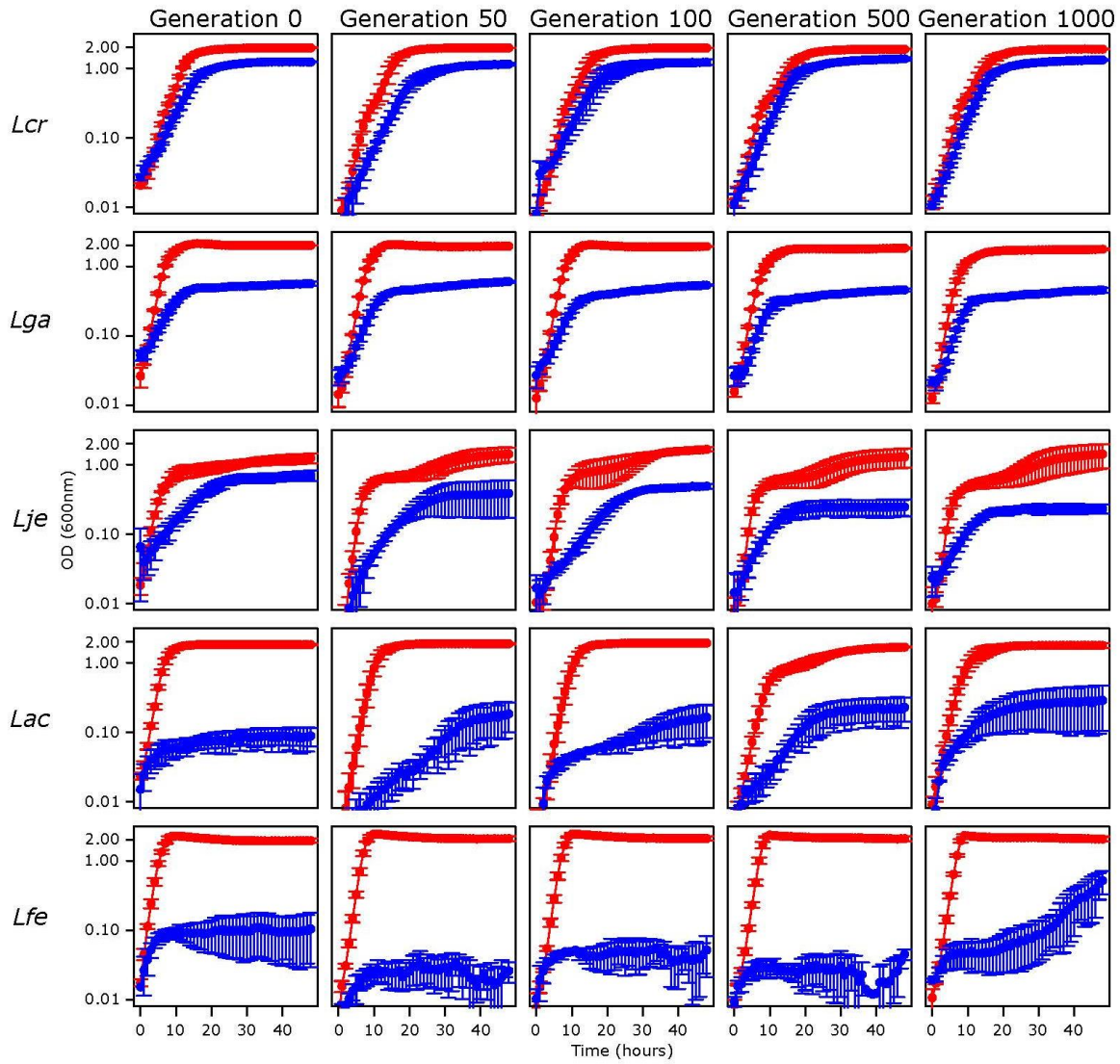


Figure 5.1 | Growth in SVF and MRS. 48-hour growth curves performed in SVF (blue) and MRS (red). Standard error bars are over three biological replicates. Species are organized in rows and are labelled on the left. Generations are organized by columns and are labeled at the top. Species are labeled according to Table 5.1.

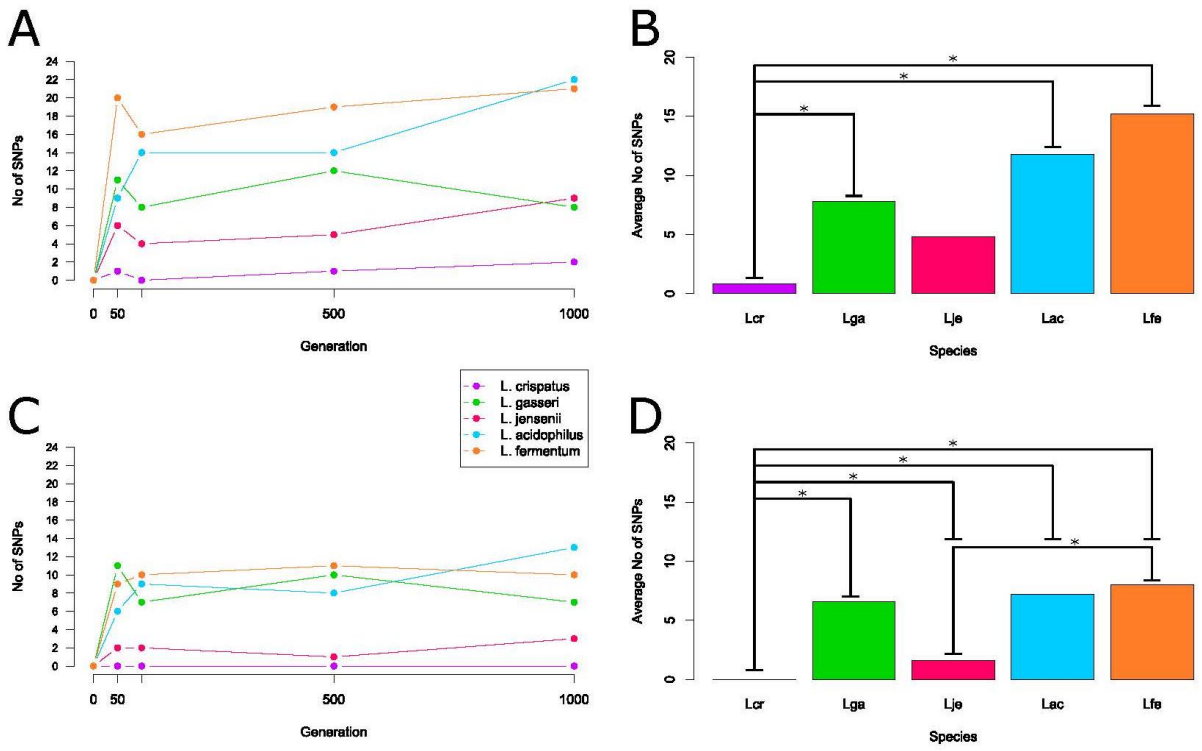


Figure 5.2 | SNP Count. Total number of sites with (A) over 50% frequency or (C) 100% frequency. Number of SNPs is on the y-axis and generation is on the x-axis. Average number of SNPs with (B) over 50% frequency or (D) 100% frequency. Significance determined by a Student's t-test ($p<0.05$) is indicated by an asterisk (*). Species are labelled and colored according to Table 5.1.

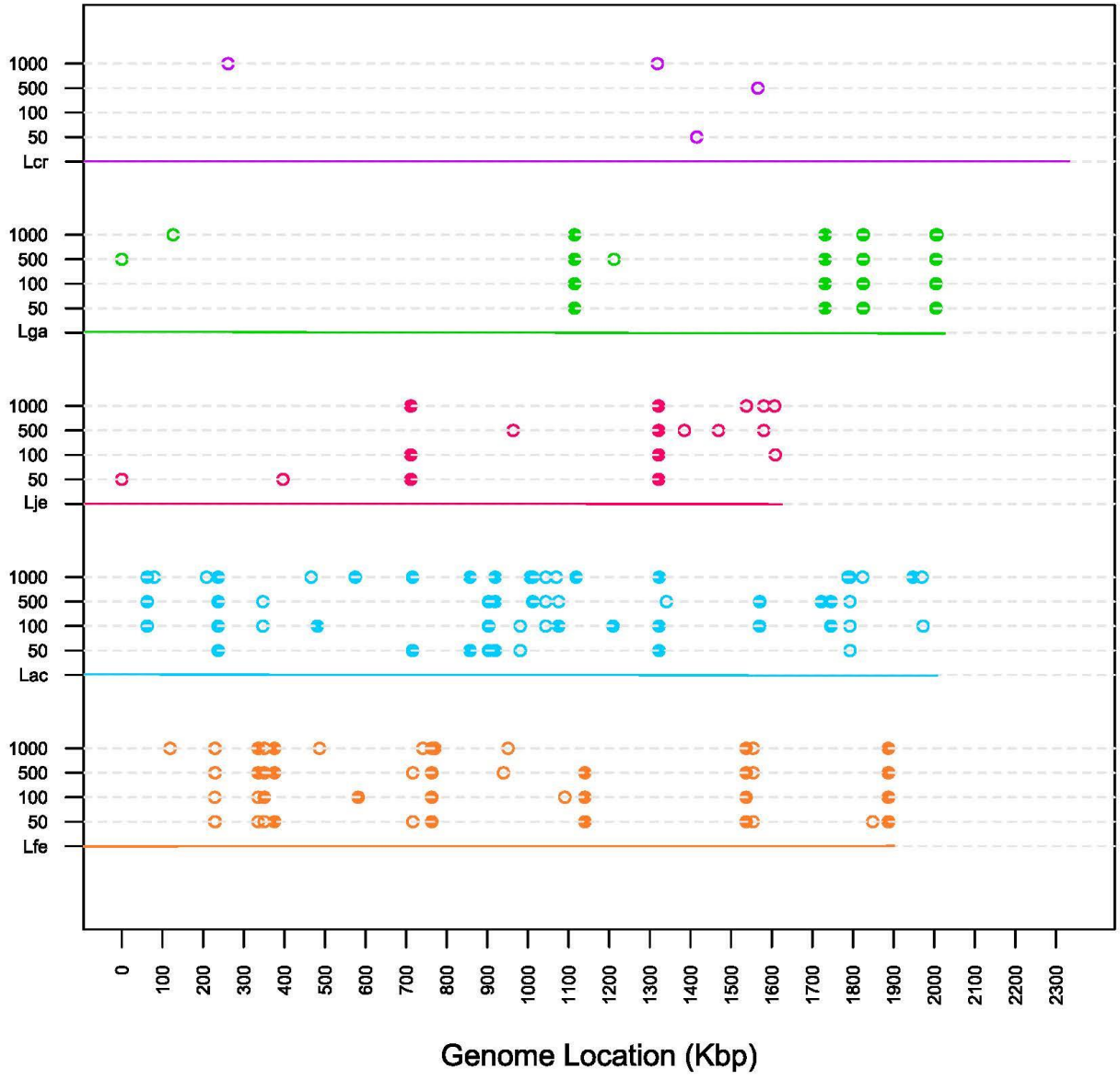


Figure 5.3 | SNPs by Location. SNPs are recorded by chromosome location and generation in which they occur. Chromosome is represented by a colored line and species label is on the y-axis. Genome location is recorded on the x-axis. SNPs with over 50% frequency are represented by an open circle. SNPs with 100% frequency are represented by a closed circle. Species are labeled and colored according to Table 5.1.

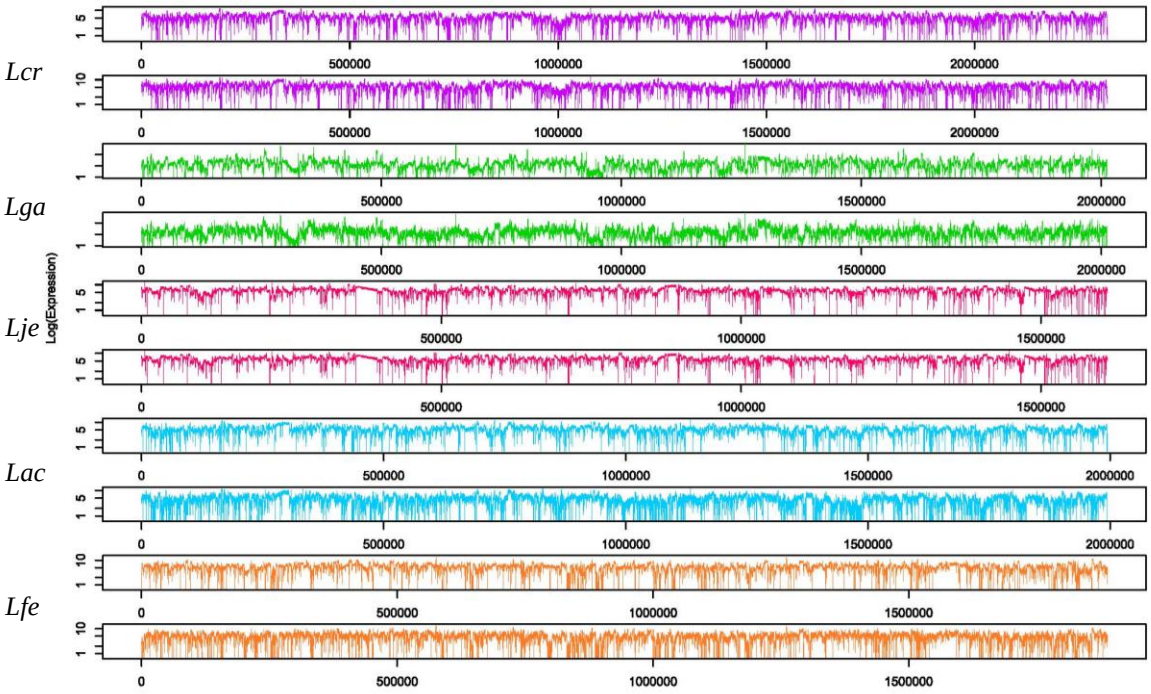


Figure 5.4 | Whole Genome Transcription. mRNA coverage over the entire genome. The chromosome location (bp) is on the x-axis. The log coverage is on the y-axis. Species are labelled and colored according to Table 5.1. For each species, the top graph is generation 0, the bottom graph is generation 1,000.

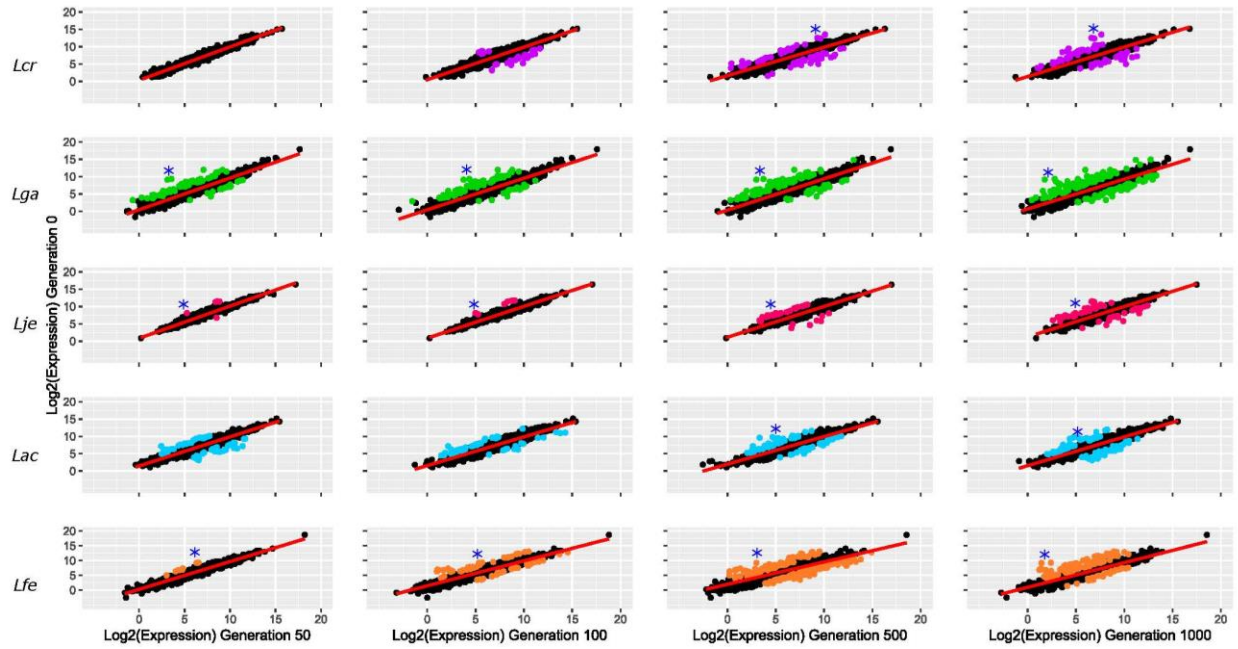


Figure 5.5 | Two-way Expression. Two-way plots comparing expression between generation 0 and remaining generations. Generation 0 is along the y-axis. The x-axis denotes the compared generation. The red line is a fitted linear model. Significantly expressed genes are colored. A blue asterisk (*) denotes genes that are analyzed in Figure 5.7. Species are ordered by row and are labelled on the left. Species are labeled and colored according to Table 5.1.

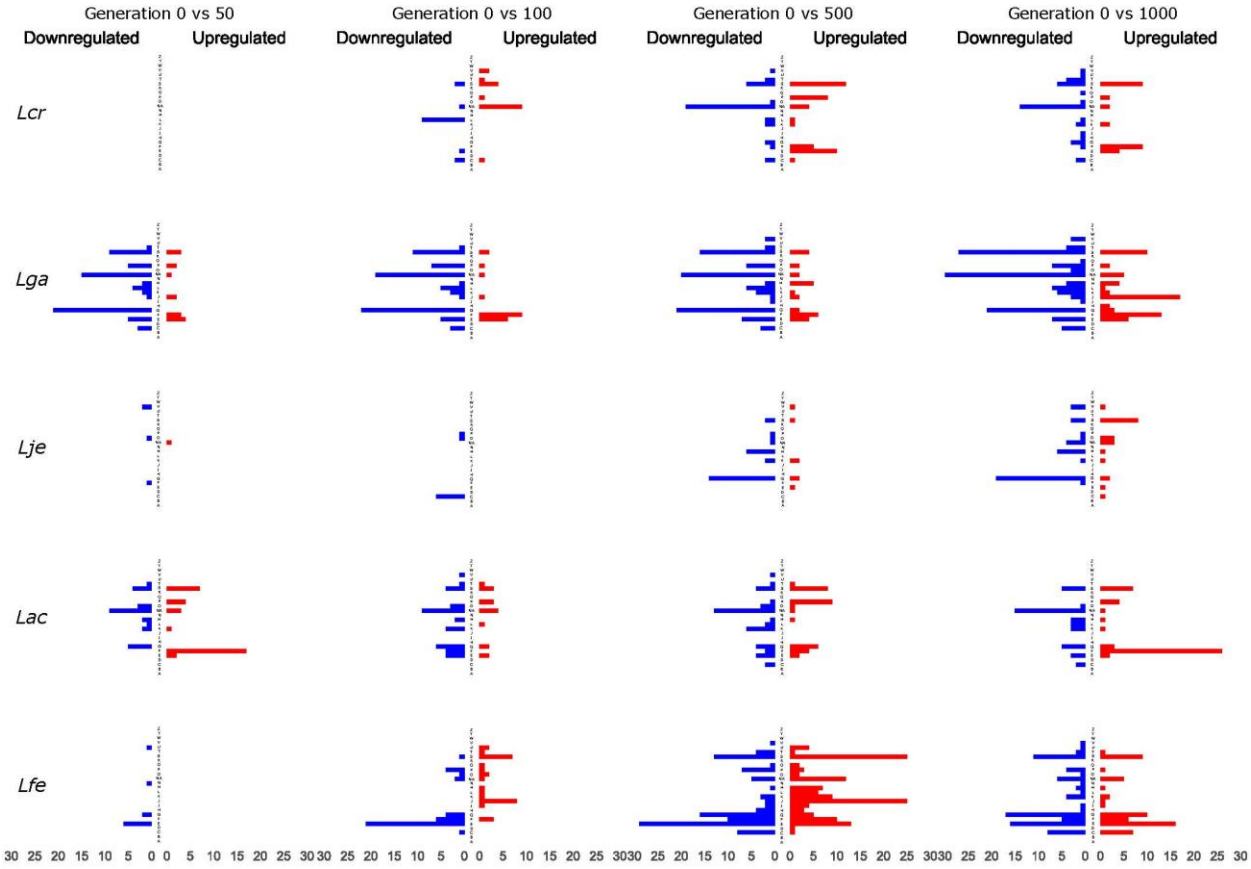


Figure 5.6 | COG Regulation. Counts of upregulated and downregulated significantly expressed genes by COG designation. COGs definitions are in Table 5.S1. Species are organized in rows and are labelled on the left. Generations are organized by columns and are labeled at top. Species are labeled according to Table 5.1.

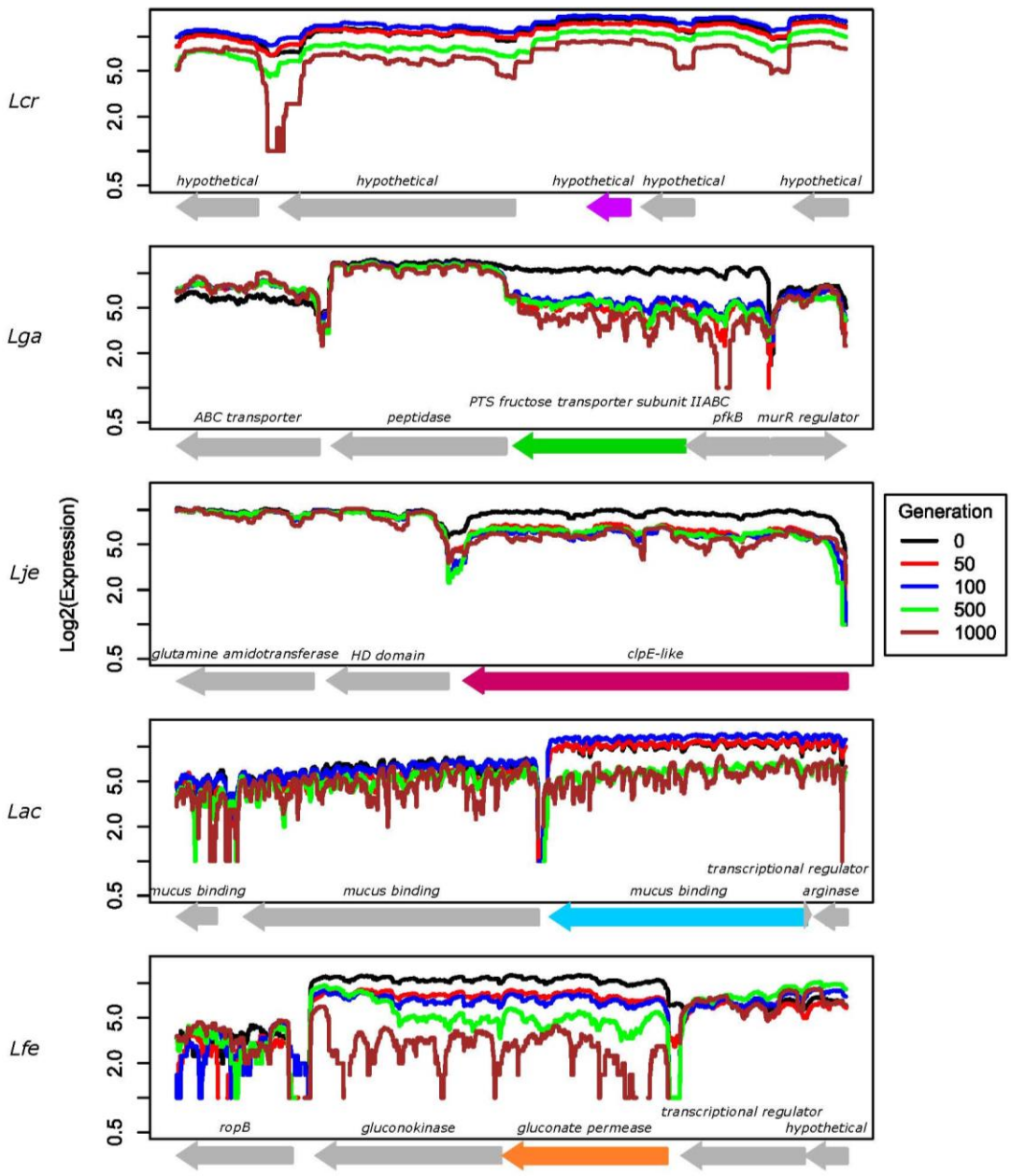


Figure 5.7 | Locus Transcription. Expression of select locus for each species. Locus of interest is colored according to Table 5.1. Two loci upstream and downstream of the locus are shown and colored in gray (*L. jensenii* was at the end of a contig, so only loci upstream are shown). Expression is on the y-axis; position is on the x-axis. Arrows denote gene direction. Species are organized in rows and are labelled on the left. Species are labeled according to Table 5.1.

CHAPTER 6: FUTURE DIRECTIONS

6.1. CONCLUDING THOUGHTS

6.1.1. Further Research in Lactic Acid Bacteria Phylogenetic Analyses

The phylogenetic method described in Chapter 2 and applied in Chapter 3 was developed to allow for easy identification and phylogenetic assignment of lactic acid bacteria (LAB) using the universal glycolysis pathway. The glycolytic approach was applied to all sequenced species in *Bifidobacterium* in Chapter 2. However, Chapter 3 applied the methodology to a subset of *Lactobacillus*. The approach should then be applied to all known and sequenced *Lactobacillus* species. The glycolytic approach provides a robust phylogeny in less time and uses less computational power as compared to core genome analyses. This is invaluable in a genus for which taxonomy is often subject to debate. The methodology should be extended to other LAB genera, such as *Lactococcus, Pediococcus, Streptococcus,* and others. Once validated in other genera, it would then be desirable to perform a complex analysis across LAB using the glycolytic gene-based method.

While using the entire glycolysis pathway has advantages, it should be determined whether the process can be streamlined. Chapters 2 and 3 analyzed the ability of the glycolysis enzymes to determine phylogeny on an individual basis, as well as a whole pathway. No single enzyme performed as well as the whole pathway, but some did provide a basis for strong phylogenetic branches, such as glyceraldehyde 3-phosphate and pyruvate kinase. It should be determined whether a subset of genes (perhaps three or four) would provide the necessary information for robust analyses, thus further refining and enhancing the process.

Finally, future analyses should focus on applying the glycolysis method at a strain level. Chapter 4 briefly explored this for a phylogenetic tree of *Lactobacillus fermentum* species using the phosphoglucomutase enzyme. This approach will need to be validated in several other

species to determine its universal potential and appeal. As strains can be relatively similar, care will have to be taken to determine if the full pathway, a subset of enzymes, or a single enzyme is the best approach. At a strain level, using nucleotide sequences over amino acid sequences would be preferred as synonymous variations would be captured and add an extra layer of granularity. When testing the validity of using a subset or single enzyme, care must be taken in selection. Several species lack some glycolytic genes, most notably *Lactobacillus reuteri*. Despite the care needed, there is no reason that this approach should not be applied outside of *Bifidobacterium* and *Lactobacillus*, nor why it should not be applied at a strain level.

6.1.2. Further Research in Species Characterization

Consumer interest has shifted in the last few years. Rising concerns from antibiotic use and a push for "natural" products opens the door for the development of microbial products as probiotics, functional foods, and biotherapeutics. In order to do this, scientists must investigate and characterize new species of interest. In Chapter 4, we examined the type strain of *Lactobacillus fermentum* and investigated the species' genomic diversity. This provides a basis to identify functional attributes of interest. In order to determine the suitability of *L. fermentum* as a valuable industrial species, future studies should perform *in vivo* characterization of some important functional features, such as adhesion for probiotics, CRISPR-Cas systems for biotechnology, and exopolysaccharide (EPS) production for biotherapeutics. More importantly, due to the high level of genome variability determined in Chapter 4, strain level characterization will be needed. Beyond *L. fermentum*, more species and strains of *Lactobacillus* and LAB should be investigated. The long history of use and presence of LAB in relevant environments reviewed in Chapter 1 implies the rich reservoir for potential tool development in LAB. Future work

should expand the promising, well-characterized LAB toolbox in order to diversify the molecular toolbox, industrial applications, and fields of use.

6.1.3. Future Work in the Vaginal Microbiome

In Chapter 5, we investigated how *Lactobacillus* species adapt and grow in a simulated vaginal environment. Future work should establish how *Lactobacillus* species interact with one another within such an environment. As the vaginal microbiome is commonly dominated by a single species, these studies should encompass a diverse mixed population, to determine the rules that guide dominance and determine competitiveness. Beyond the ability to compete, it would be interesting to determine whether there are any synergistic effects between select *Lactobacillus* species. Finally, studies should be completed to examine the loss of competition, or to determine when a species is no longer fit to dominate. This would mimic dysbiosis and could possibly shed light on how bacterial vaginosis forms or re-occurs.

Beyond microbial interaction, studies need to focus on host interaction as well. Some studies have correlated host behavior with changes in the vaginal microbiome, but there is not a genetic component involved as of yet. This is very interesting due to the established ethnic imbalance among the Community State Types. In order to better understand the genetic underpinnings of the vaginal microbiome, the Human Microbiome Project should be further utilized. Recent work has included the survey of metagenomics data, which should be analyzed for potential functional genes of interest. This work and the work mentioned above is especially important in determining which genes and species to utilize in commercial formulations. While *Lactobacillus crispatus* has the strongest association with host health, it may not be the healthiest option for everyone. Without knowing how *L. crispatus* will grow in the presence of other

lactobacilli, nor how it comes into dominance, it is not possible to state that *L. crispatus* should be universally used. Single-species studies are essential to ascertaining some of these answers, but to answer the problem in full, community studies will need to be undertaken.

Lastly, a human vaginal model must be developed. This will be essential in establishing community relationships, validating mode of action, and the development of probiotics and treatments. One option would be an animal model that creates similar environmental conditions, such as cell structure and carbohydrate availability. It may then be possible to modify the microbiome so that it mimics humans. The further development and advancement of three-dimensional models may be an alternative route. This is appealing as there is no known mammal that matches the human vaginal microbiome in biology or microbiome.

6.2. CONTRIBUTION TO THE FIELD

In this work, we have applied genomic techniques to help expand the LAB toolbox in industry. We hope our findings will inform product development and guide the design of future studies. As our understanding of how microbes affect and interact with their environments deepens, we need to apply our knowledge of genomics, as performed here, to fully leverage the opportunities provided and technologies available.

The glycolytic phylogeny methodology has several benefits to industry. At the time of writing, there was no accepted universal method in LAB, and certainly not for *Lactobacillus*. Traditional 16S rRNA provided an unsatisfactory phylogeny. Core-genome methods provide strong phylogenies, but are heavy computationally. And while studies have attempted similar approaches using different enzymes, there has not been a set group. The glycolytic approach provides a uniform, quick, and robust method of phylogeny and taxonomic assessment. This is vital for the assignment of new species and the identification of new isolates, which are particular assets in microbiome research. This method may also be applied to current products in quality control, such as probiotics. Finally, these findings can aid in the selection and development of species and strains for product development.

The analysis of *L. fermentum* was performed due to the increased interest in assessing the potential of this species as a next-generation probiotic or a potential biotherapeutic. Before either could be safely applied, a more in-depth understanding of how the species functions and its mode of action needs to be determined. Vital to species characterization is the type strain, in this case being ATCC 14931. The type strain sets the benchmark on which other strains are compared. Here, we provided some of the first studies on the type strain *L. fermentum* ATCC 14931. Mainly, we performed genome analyses that compared ATCC 14931 to other *L. fermentum*

strains. At the time of writing, we had no knowledge of any other studies that attempted this. Our results showed that *L. fermentum* is highly variable. This means that each strain must be individually assessed and that claims cannot be generalized to the species as a whole. We do find a potential untapped reservoir of putative CRISPR systems that could be used as genome editing tools. There is variety in the species as a whole and the type strain's Type II system is distinct from those used in industry, allowing for an expansion of the CRISPR toolbox.

The human vaginal microbiome has been the subject of numerous microbiome and metagenomic studies. While this has provided insights into the structure of the environment, it has not been able to reveal how it functions. There is interest in harnessing the natural healthy microbiome to combat issues such as dysbiosis—similar to probiotics modulating the gut microbiome. In order to do this, functional assays must be performed, but first, proper species selection is needed. Our studies help guide this selection process. Work looking into the potential development of vaginal probiotics often includes species that are not dominant members of the vaginal microbiome. Our study shows that vaginal species are perhaps a better alternative. We have shown that vaginal species are better suited to their environment over non-vaginal species, underscoring the importance of isolation source in product development. Additionally, by performing a long-term experiment, we showed that adaptation to vaginal environment is not trivial for non-vaginal species—human adapted or not. This shows that the engineering of a non-vaginal species may not be an easy feat. This study will aid in new development of potential vaginal probiotics. It also lays the foundation for the evaluation of how species grow and come to dominate the vaginal environment.

APPENDICES

APPENDIX A: REPRINT OF "PHYLOGENETIC ANALYSIS OF THE *BIFIDOBACTERIUM* GENUS USING GLYCOLYSIS ENZYME SEQUENCES"

frontiers in Microbiology

METHODS
published: 09 May 2016
doi: 10.3389/fmicb.2016.00657

Phylogenetic Analysis of the *Bifidobacterium* Genus Using Glycolysis Enzyme Sequences

Katelyn Brandt[1,2] *and Rodolphe Barrangou*[1,2]*

[1] *Functional Genomics Graduate Program, North Carolina State University, Raleigh, NC, USA,* [2] *Department of Food, Bioprocessing and Nutrition Sciences, North Carolina State University, Raleigh, NC, USA*

Bifidobacteria are important members of the human gastrointestinal tract that promote the establishment of a healthy microbial consortium in the gut of infants. Recent studies have established that the *Bifidobacterium* genus is a polymorphic phylogenetic clade, which encompasses a diversity of species and subspecies that encode a broad range of proteins implicated in complex and non-digestible carbohydrate uptake and catabolism, ranging from human breast milk oligosaccharides, to plant fibers. Recent genomic studies have created a need to properly place *Bifidobacterium* species in a phylogenetic tree. Current approaches, based on core-genome analyses come at the cost of intensive sequencing and demanding analytical processes. Here, we propose a typing method based on sequences of glycolysis genes and the proteins they encode, to provide insights into diversity, typing, and phylogeny in this complex and broad genus. We show that glycolysis genes occur broadly in these genomes, to encode the machinery necessary for the biochemical spine of the cell, and provide a robust phylogenetic marker. Furthermore, glycolytic sequences-based trees are congruent with both the classical 16S rRNA phylogeny, and core genome-based strain clustering. Furthermore, these glycolysis markers can also be used to provide insights into the adaptive evolution of this genus, especially with regards to trends toward a high GC content. This streamlined method may open new avenues for phylogenetic studies on a broad scale, given the widespread occurrence of the glycolysis pathway in bacteria, and the diversity of the sequences they encode.

Keywords: ***Bifidobacterium*, glycolysis, phylogeny, probiotic, evolution**

OPEN ACCESS

Edited by:
Francesca Turroni,
University College of Cork, Ireland

Reviewed by:
Abelardo Margolles,
Consejo Superior de Investigaciones Cientificas, Spain
Gabriele Andrea Lugli,
University of Parma, Italy

***Correspondence:**
Rodolphe Barrangou
rbarran@ncsu.edu

Specialty section:
This article was submitted to Microbial Symbioses, a section of the journal Frontiers in Microbiology

Received: *29 February 2016*
Accepted: *20 April 2016*
Published: *09 May 2016*

Citation:
Brandt K. and Barrangou R (2016) Phylogenetic Analysis of the Bifidobacterium Genus Using Glycolysis Enzyme Sequences. Front. Microbiol. 7:657. doi: 10.3389/fmicb.2016.00657

INTRODUCTION

Bifidobacterium species are an important component of the human gastrointestinal tract (GIT) microbiome, and exert critical functional roles, especially during the establishment of gut microbial composition early in life. Consequently, they are the subject of extensive microbiological and genetics studies, to investigate their probiotic phenotypes, and genotypes, respectively. Actually, many studies are investigating the genetic basis for their health-promoting functionalities, both in industry and academia. This genus is often found in the GIT of animals (Ventura et al., 2014), and is the predominant phylogenetic group early in human life (Turroni et al., 2012a). Indeed, a mounting body of evidence has established vertical transmission between the mother and infants (Milani et al., 2015), notably through the selective nurture of bifidobacteria

through diverse non-digestible human-milk oligosaccharides (HMOs) that are a critical component of breast milk (Sela, 2011). These HMOs selectively drive the colonization of the infantile GIT by species that encode prebiotic transporters and hydrolases (Turroni et al., 2012b). Recently, a dichotomy has been established between healthy term babies with a normal gut microbiome, and preterm infants whom have not been colonized by *Bifidobacterium* species (Arboleya et al., 2015). Several studies have implicated the expansive carbohydrate uptake and catabolism gene repertoire of bifidobacteria as the key driver of adaptation of this genus to the infant diet (Milani et al., 2014). In fact, several species of bifidobacteria have shown unique genome composition adaptation trajectories in their carbohydrate utilization machinery, rendering them competitive in this environment (Pokusaeva et al., 2011; Ventura et al., 2012).

To better understand how these organisms have emerged as potent early-life colonizers, there has been a surge in genome sequencing in recent years. At the time of writing, 47 established species and subspecies have been sequenced (Milani et al., 2016), providing a wealth of genomic information, which serves as a valuable tool for understanding the species and strain diversity within this polymorphic genus, as well as unraveling the key elements that drive health-promoting and colonization phenotypes in humans. However, given the democratization of sequencing technologies in general, and genome and microbiome sequencing in particular, it is imperative that tools and methods be available to analyze this high-throughput data, and specifically allow experimentalists to parse out the complex phylogeny of this broad genus. Indeed, basic questions being addressed regarding the occurrence, diversity and functions of various *Bifidobacterium* species in the human GIT will require the ability to accurately and consistently assign phylogeny.

Fundamentally, as new sequences become available, it is important to know where to place strains on the phylogenetic tree of *Bifidobacterium.* Whereas the affordability, accessibility and ability to generate high-throughput data have become somewhat straightforward, a key challenge lies in the analysis of these sequences, regarding assembly, comparative analyses and phylogenetic assignments. Historically, 16S rRNA sequences have been used across the phylogenetics field for classification and sequence tree-based assignments, but there are growing concerns about the adequacy and sustainability of this method (Fox et al., 1992), notably with regards to the availability of proper references (Clarridge, 2004), and the actual levels of conservation of sequences targeted by "universal" primers (Baker et al., 2003). Because of this, new approaches have been suggested, ranging from multi-locus approaches, using housekeeping genes (Eisen, 1995), to core-genome analyses (Medini et al., 2005). For *Bifidobacterium*, efforts have been focused on creating a phylogeny based on whole and/or conserved genomic sequences, namely the pan-genome and the core-genome, respectively (Lukjancenko et al., 2011; Lugli et al., 2014). While the core-genome is arguably comprehensive, core-genome assembly is time consuming and computationally intense. Alternative methods need to be developed, to allow rapid and convenient phylogenetic screening of new and potentially unknown sequences. Preferably, such a method would provide high resolution, low-throughput, robust, accurate, and affordable information.

Notwithstanding phenotypic diversity between organisms that have specialized metabolic pathway combinations, and the corresponding genomic complement, there are core biochemical pathways and processes that are broadly distributed across the Tree of Life. Noteworthy, glycolysis is a fundamental process for most cells, and may be construed as the biochemical backbone of most, if not all, living organisms. Indeed, this process allows for the genesis of energy through the catabolism of simple carbohydrates. This pathway is, at least partially, present in all genomes (Fothergill-Gilmore and Michels, 1993) and consequently constitutes a promising biochemical, and thus genetic, marker for phylogenetic studies. Because these genes are important, they are typically members of the house-keeping genomic set, and are widely dispersed across the Tree of Life. However, they are likely subject to less selective pressure than other phylogenetic markers (i.e., ribosomal sequences), and thus afford a more diverse set of sequences to encompass a broad range of assorted sequences (Fothergill-Gilmore, 1986). Therefore, we set out to assess the potential of glycolytic genes, and the sequences of the proteins they encode, for bifidobacteria phylogenetic studies. In particular, we determined the occurrence and diversity of these glycolytic enzyme genes in the genomes of bifidobacteria, and compared and contrasted sequence alignment-based trees with one another, and to those derived from alternative sequences, notably the core-genome, and the 16S rRNA-based reference tree. Our results show how the glycolysis protein sequences can be used as suitable markers to create a phylogeny of *Bifidobacterium* that is as accurate as the core-genome based phylogeny, but much less computationally demanding. We also explore how basic features of the genetic sequences of glycolysis can reveal trends and patterns of evolution among the different *Bifidobacterium* species and the genus as a whole.

MATERIALS AND METHODS

Genetic Sequences Sampling and Reference Genomes

We used sequences derived from a total of 48 *Bifidobacterium* genomes from distinct species and subspecies, as listed in **Table 1**. *Bifidobacterium stercoris* was included in this analysis, as a separate species, but it was recently renamed as a strain of *Bifidobacterium adolescentis* (Killer et al., 2013). Our results (see below) show that *B. stercoris* is always a close neighbor of *B. adolescentis*, consistent with the newest findings. These genomes were mined for the presence of glycolytic enzymes using Geneious version 9.0.5(Kearse et al., 2012). We selectively elected to pursue a scheme based on canonical glycolysis genes, as to generate a broadly applicable method. Nevertheless, the classical glycolysis genes do not universally occur in bacterial genomes. Furthermore, some organisms do carry alternative pathways, such as the bifid shunt in *bifidobacterium,* which could prove valuable, but are not widely distributed. The nine canonical glycolysis enzymes from bifidobacteria (de Vries and

Stouthamer, 1967) were found in each genome. Four reference species (*Bifidobacterium longum* subsp. *longum*, *B. adolescentis*, *Bifidobacterium animalis sub. lactis*, and *Bifidobacterium breve*) were used to make a database of the nine genes. The Annotate from Database feature was used (with 40% nucleotide sequence similarity cut-off) to identify glycolytic orthologs in the other genomes. As all genomes had been previously annotated, we confirmed the original annotation to the database annotation

TABLE 1 | Species and genome list.

Genus	Species	Subspecies	Strain	Accession number	Naming convention	Locus tag
Bifidobacterium	*actinocoloniiforme*		DSM 22766	NZ_CP011786	B_actinocoloniiforme	AB656
Bifidobacterium	*adolescentis*		ATCC 15703	NC_008618	B_adolescentis	BAD
Bifidobacterium	*angulatum*		LMG 11039	NZ_JGYL00000000	B_angulatum	BIANG
Bifidobacterium	*animalis*	*animalis*	ATCC 22527	NC_017834	B_animalis_a	BANAN
Bifidobacterium	*animalis*	*lactis*	DSM 10140	NC_012815	B_animalis_l	BALAT
Bifidobacterium	*asteroides*		PRL 2011	NC_018720	B_asteroides	BAST
Bifidobacterium	*biavatii*		DSM 23969	NZ_JDUU00000000	B_biavatti	OU23
Bifidobacterium	*bifidum*		LMG 13200	NZ_JSEB00000000	B_bifidum	LMG13200
Bifidobacterium	*bohemicum*		DSM 22767	NZ_JDUS00000000	B_bohemicum	OU21
Bifidobacterium	*bombi*		DSM 19703	NZ_JDTS00000000	B_bombi	OT95
Bifidobacterium	*boum*		LMG 10736	NZ_JGYQ00000000	B_boum	BBOU
Bifidobacterium	*breve*		UCC 2003	NC_020517	B_breve	Bbr
Bifidobacterium	*callitrichos*		DSM 23973	NZ_JGYS00000000	B_callitrichos	BCAL
Bifidobacterium	*catenulatum*		JCM 1194	NZ_AP012325	B_catenulatum	BBCT
Bifidobacterium	*choerinum*		LMG 10510	NZ_JGYU00000000	B_choerinum	BCHO
Bifidobacterium	*coryneforme*		LMG 18911	NZ_CP007287	B_coryneforme	BCOR
Bifidobacterium	*crudilactis*		LMG 23609	NZ_JHAL00000000	B_crudilactis	DB51
Bifidobacterium	*cuniculi*		LMG 10738	NZ_JGYV00000000	B_cuniculi	BCUN
Bifidobacterium	*dentium*		Bd1	NC_013714	B_dentium	BDP
Bifidobacterium	*gallicum*		DSM 20093	NZ_ABXB00000000	B_gallicum	BIFGAL
Bifidobacterium	*gallinarum*		LMG 11586	NZ_JGYX00000000	B_gallinarum	BIGA
Bifidobacterium	*indicum*		LMG 11587	NC_CP006018	B_indicum	BINDI
Bifidobacterium	*kashiwanohense*		JCM 15439	NZ_AP012327	B_kashiwanohense	BBKW
Bifidobacterium	*longum*	*longum*	NCC 2705	NC_004307	B_longum	BL
Bifidobacterium	*longum*	*infantis*	ATCC 15697	NC_011593	B_longum_i	Blon
Bifidobacterium	*longum*	*suis*	LMG 21814	NZ_JGZA00000000	B_longum_s	BLSS
Bifidobacterium	*magnum*		LMG 11591	NZ_JGZB00000000	B_magnum	BMAGN
Bifidobacterium	*merycicum*		LMG 11341	NZ_JGZC00000000	B_merycicum	BMERY
Bifidobacterium	*minimum*		LMG 11592	NZ_JGZD00000000	B_minimum	BMIN
Bifidobacterium	*mongoliense*		DSM 21395	NZ_JGZE00000000	B_mongoliense	BMON
Bifidobacterium	*moukalabense*		DSM 27321	NZ_AZMV00000000	B_moukalabense	BMOU
Bifidobacterium	*pseudocatenulatum*		JCM 1200	NZ_AP012330	B_pseudocatenulatum	BBPC
Bifidobacterium	*pseudolongum*	*globosum*	LMG 11569	NZ_JGZG00000000	B_pseudolongum_g	BPSG
Bifidobacterium	*pseudolongum*	*pseudolongum*	LMG 11571	NZ_JGZH00000000	B_pseudolongum_p	BPSP
Bifidobacterium	*psychraerophilum*		LMG 21775	NZ_JGZI00000000	B_psychraerophilum	BPSY
Bifidobacterium	*pullorum*		LMG 21816	NZ_JGZJ00000000	B_pullorum	BPULL
Bifidobacterium	*reuteri*		DSM 23975	NZ_JGZK00000000	B_reuteri	BREU
Bifidobacterium	*ruminantium*		LMG 21811	NZ_JGZL00000000	B_ruminantium	BRUM
Bifidobacterium	*saeculare*		LMG 14934	NZ_JGZM00000000	B_saeculare	BSAE
Bifidobacterium	*saguini*		DSM 23967	NZ_JGZN00000000	B_saguini	BISA
Bifidobacterium	*scardovii*		LMG 21589	NZ_JGZO00000000	B_scardovii	BSCA
Bifidobacterium	*stellenboschense*		DSM 23968	NZ_JGZP00000000	B_stellenboschense	BSTEL
Bifidobacterium	*stercoris*		DSM 24849	NZ_JGZQ00000000	B_stercoris	BSTER
Bifidobacterium	*subtile*		LMG 11597	NZ_JGZR00000000	B_subtile	BISU
Bifidobacterium	*thermacidophilum*	*porcinum*	LMG 21689	NZ_JGZS00000000	B_thermacidophilum_p	BPORC
Bifidobacterium	*thermacidophilum*	*thermacidophilum*	LMG 21395	NZ_JGZT00000000	B_thermacidophilum_t	THER5
Bifidobacterium	*thermophilum*		JCM 7027	–	B_thermophilum	BTHER
Bifidobacterium	*tsurumiense*		JCM 13495	NZ_JGZU00000000	B_tsurumiense	BITS

List of the 48 species and subspecies used in this study. Accession numbers and naming conventions included.

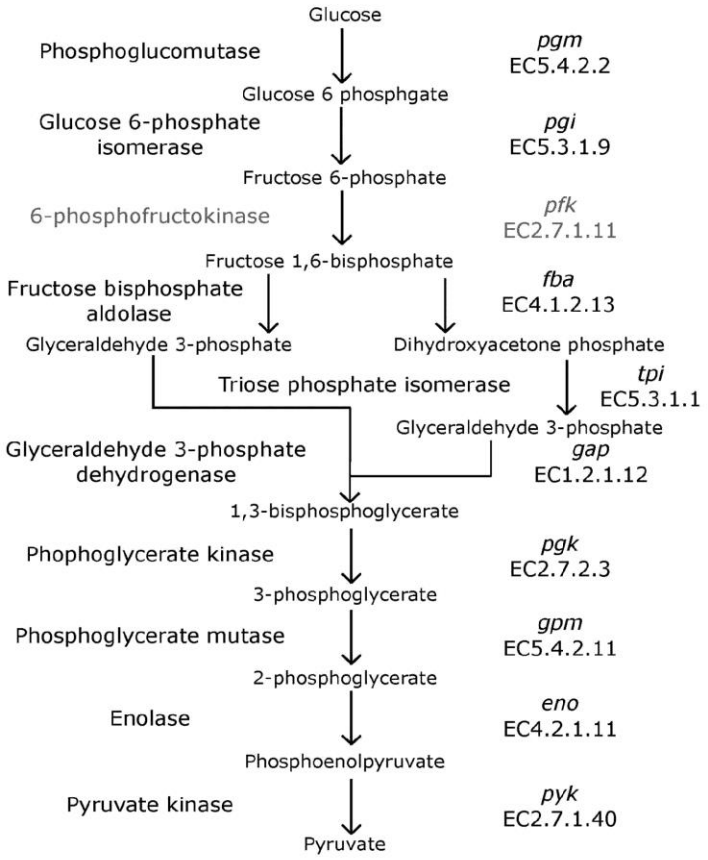


FIGURE 1 | Glycolysis pathway. Traditional biochemical pathway of glycolysis. Enzyme names are listed to left of arrows, and gene names and EC numbers are shown on the right. 6-phosphofructokinase is faded to represent its absence in *Bifidobacterium*.

manually to validate this method of mining. In cases where multiple hits were obtained, BLAST (Altschul et al., 1990) analyses were carried out to select the correct homolog. Translated sequences were confirmed using ExPasy (Gasteiger et al., 2003). For the 16S rRNA analysis, the 16S rRNA sequences were extracted manually from each genome. In case of multiple hits, BLAST analyses were carried out to select the right sequences. For increased robustness, the glycolysis enzyme sequences were concatenated in order of occurrence in the glycolysis pathway (Lang et al., 2013).

Genesis of Sequence Alignment-based Trees

Five different alignments were made for each tree using Geneious version 9.0.5. ClustalW (Larkin et al., 2007) was used, with the BLOSUM scoring matrix, and settings of gap creation at −10 cost, and gap extension at −0.1 cost per element. For the 16S rRNA alignment, ClustalW was set so that the cost matrix was IUB, with a gap opening penalty of 15, and gap extension cost of 6.66. MUSCLE (Edgar, 2004) was used with the setting of eight maximum number of iterations for the amino acid sequences and the 16S rRNA alignments. The Geneious Pairwise Alignment was set so that the alignment type was global alignment with free end gaps and the cost matrix was BLOSUM62 for the amino acid sequences. For the 16S rRNA gene analysis, the alignment type was global alignment with free end gaps and a cost matrix of 65% similarity (5.0/−4.0). MAFFT (Katoh et al., 2002) was used twice, for both the amino acid sequences and the 16S rRNA sequences. For the amino acid sequences the first alignment had an algorithm setting of auto, a scoring matrix of BLOSUM62, a gap open penalty of 1.53, and an offset value of 0.123. The second

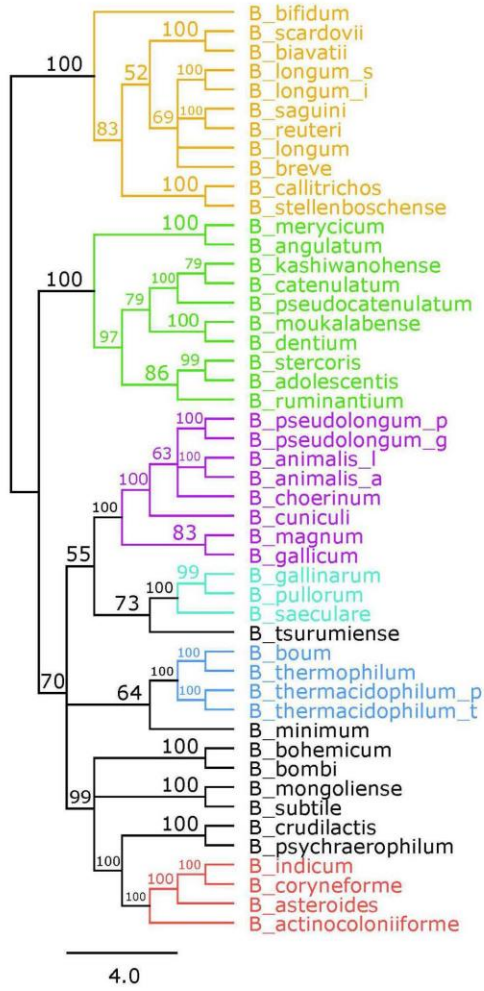


FIGURE 2 | Glycolytic proteins concatenated tree. Consensus tree based on alignment of the concatenated amino acid sequences of the glycolysis pathway found in *Bifidobacterium*. Trees were made using RaxML. Bootstrap values are found on each node. Phylogenetic groups are colored as follows: *Bifidobacterium longum* is orange, *Bifidobacterium adolescentis* is green, *Bifidobacterium psdeudolongum* is purple, *Bifidobacterium pollorum* is blue-green, *Bifidobacterium boum* is blue, and *Bifidobacterium asteroides* is red. Species names follow the naming convention from **Table 1**.

alignment had an algorithm setting of auto, a scoring matrix of BLOSUM80, a gap open penalty of 1.53, and an offset value of 0.123. For the first 16S rRNA alignment, the algorithm was set to auto, the scoring matrix was set to 100 PAM/k = 2, the gap open penalty was set to 1.53, and the offset value was set to 0.123. The second alignment for the 16S rRNA was set so that the algorithm was auto, the scoring matrix was 200 PAM/k = 2, the gap open penalty was 1.53, and the offset value was 0.123. trimAl (Capella-Gutiérrez et al., 2009) was used to select a consistent alignment between the five alignments. The parameters were compareset and automated1. Using Geneious, trees were made from the respective consistent alignments. The trees were generated using RaxML version 7.2.8 (Stamatakis, 2006b, 2014). For the protein based trees the parameters were set so that the model was CAT (Lartillot and Philippe, 2004) BLOSUM62, the algorithm was Bootstrap using rapid hill climbing with random seed 1, and the number of bootstrap replicates was 100 (Stamatakis, 2006a). For the 16S rRNA tree, the nucleotide model was GTR CAT, the algorithm was Bootstrap using rapid hill climbing with random seed 1, and the number of bootstrap replicates was 100. A consensus tree was then built using the consensus builder in Geneious, at a 50% support threshold. The consensus tree was used in all further analyses. The sums of branch lengths for each tree were found by adding the branch lengths together in Mega6 (Tamura et al., 2013).

TABLE 2 | Sum of branch lengths for each tree.

Gene	E. C. number	Sum
Phosphoglucomutase (pgm,1)	5.4.2.2	125.03
Glucose-6-phosphate isomerase (pgi,2)	5.3.1.9	153.43
Fructose bisphosphate aldolase (fba, 4)	4.1.2.13	151.76
Triose phosphate isomerase (tpi, 5)	5.3.1.1	170.61
Glyceraldehyde 3-phosphate dehydrogenase (gap, 6)	1.2.1.12	103.07
Phsophoglycerate kinase (pgk, 7)	2.7.2.3	132.41
Phosphoglycerate mutase (gpm, 8)	5.4.2.11	174.7
Enolase (eno, 9)	4.2.1.11	145.06
Pyruvate kinase (pyk, 10)	2.7.1.40	107.56
Concatenated	–	99.56
16S rRNA	–	204.99

Sum of branch lengths for each tree. EC number for each enzyme is also listed.

Statistical Analyses

All statistical analyses were carried out using R version 3.2.2 (R Core Team, 2015). This software was also used to generate plots, graphs and display quantitative data throughout the manuscript.

RESULTS

Glycolytic Enzyme Sequence-based Phylogeny

Bifidobacteria contain nine of the 10 traditional enzymes (**Figure 1**) commonly found in the glycolysis pathway (de Vries and Stouthamer, 1967). Phylogenetic analyses were carried out using the amino acid sequences of the proteins encoded by the aforementioned glycolysis genes. A comprehensive tree based on sequence alignment of the concatenated sequences of the glycolytic enzymes found in *Bifidobacterium*

is shown in **Figure 2**. Six separate phylogenetic groups were identified, as previously established from the core-genome (Milani et al., 2016). These groups are: the *B. longum* group (orange), the *B. adolescentis* group (green), the *Bifidobacterium pseudolongum* group (purple), the *Bifidobacterium pollurom* group (blue-green), the *Bifidobacterium boum* group (blue), and the *Bifidobacterium asteroides* group (red; Bottacini et al., 2014). The number of individuals in each group varied between 3 and 11, with the *B. longum* group being the most diverse. *Bifidobacterium angulatum* and *Bifidobacterium merycicum* were moved to the *B. adolescentis* group due to a high bootstrap value in the concatenated tree. The concatenated tree has bootstrap values that range from 52 to 100. We observe a total of 34 bootstrap values of 70 and above (Supplementary Figure S1). Trees based on sequence alignments of the individual enzymes of glycolysis can be found in Supplementary Figures S2–S10. Interestingly, all of the individual trees resolved the phylogenetic groups found in the core-genome with only the Gap and Eno trees providing alternative locations for a few branches, notably *Bifidobacterium magnum*, *Bifidobacterium gallicum*, and *Bifidobacterium thermacidophilum sub. thermacidophilum*. **Table 2** shows the sum of branch lengths for each tree. The 16S rRNA tree has the largest sum at 204.99, while the concatenated tree had the smallest sum at 99.56. The consistent clustering into these six phylogenetic trees illustrates how robust and valuable the glycolytic sequences are with regards to phylogenetic information. It also shows that this method is congruent with the core-genome.

16S rRNA-based Reference Phylogeny

A reference phylogeny was generated using the 16S rRNA sequences of each of the 48 species and sub-species included in this study (**Figure 3**). The six phylogenetic groups are identified and colored the same as in the concatenated tree. We elected to assign the *B. angulatum* and *B. merycicum* from the *B. longum* group to the *B. adolescentis* group, consistent with the concatenated tree. Noteworthy, the tree has bootstrap values that range from 51 to 100, with 17 nodes at values of 70 and above, which is half the amount found in the concatenated tree (Supplementary Figure S1). With regards to size, we point out that the concatenated tree is based on overall sequences ranging between 3,205 amino acids and 3,479 amino acids, which quantitatively compares as approximately twice the amount to the 16S rRNA ~1,600 nt range, in terms of input-information amounts.

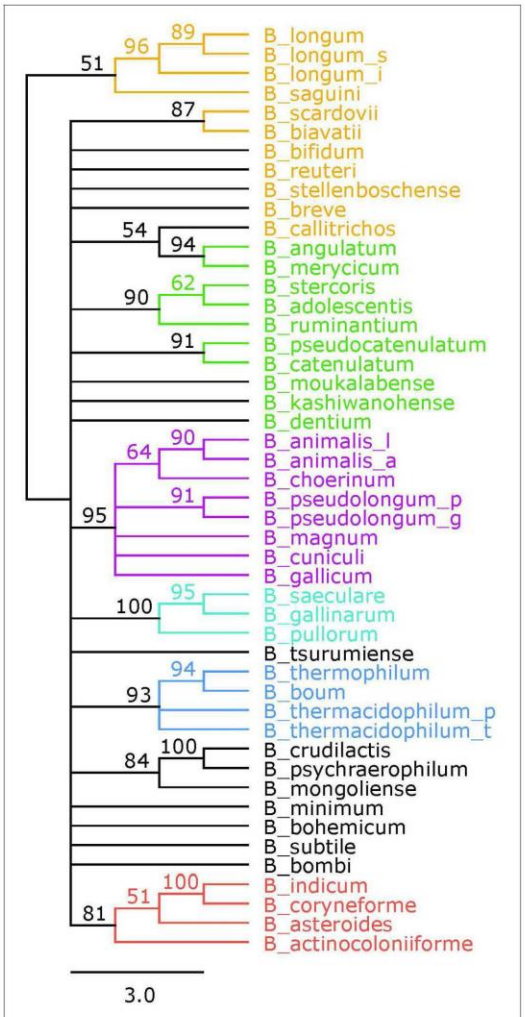


FIGURE 3 | 16S rRNA phylogenetic tree. Consensus tree based on alignment of the 16S rRNA sequences. Trees were made using RaxML. Bootstrap values are found on each node. Phylogenetic groups are colored as follows: *B. longum* is orange, *B. adolescentis* is green, *B. psdeudolongum* is purple, *B. pollorum* is blue-green, *B. boum* is blue, and *B. asteroides* is red. Species names follow the naming convention from **Table 1**.

Genome-Wide Analyses

The overall genome sizes in this study ranged from 1.73 Mb for *Bifidobacterium indicum* to 3.26 Mb for *Bifidobacterium biavatii*, with an average of 2.28 Mb and a median of 2.17 Mb. The GC content ranged from 52.8% for *Bifidobacterium tsurumiense* to 65.5% for *Bifidobacterium choerinum*, with an average of 60.4% and a median of 60.2%. This substantiates the perception that bifidobacteria are generally categorized as high-GC content organisms, at the genome-wide level (Ventura et al., 2007). However, a thorough analysis of GC content across the phylogenetic groups revealed that even among these high-GC organisms there are three distinct subsets of high, medium, and low-GC bifidobacteria (**Figure 4A**). Most of the species fall in the upper medium-GC range, with the low-GC range being the least populated. There are some noteworthy groupings between the phylogenetic groups, specifically the *B. pullorum* and the *B. boum* groups, for which the entire groups are packed tightly

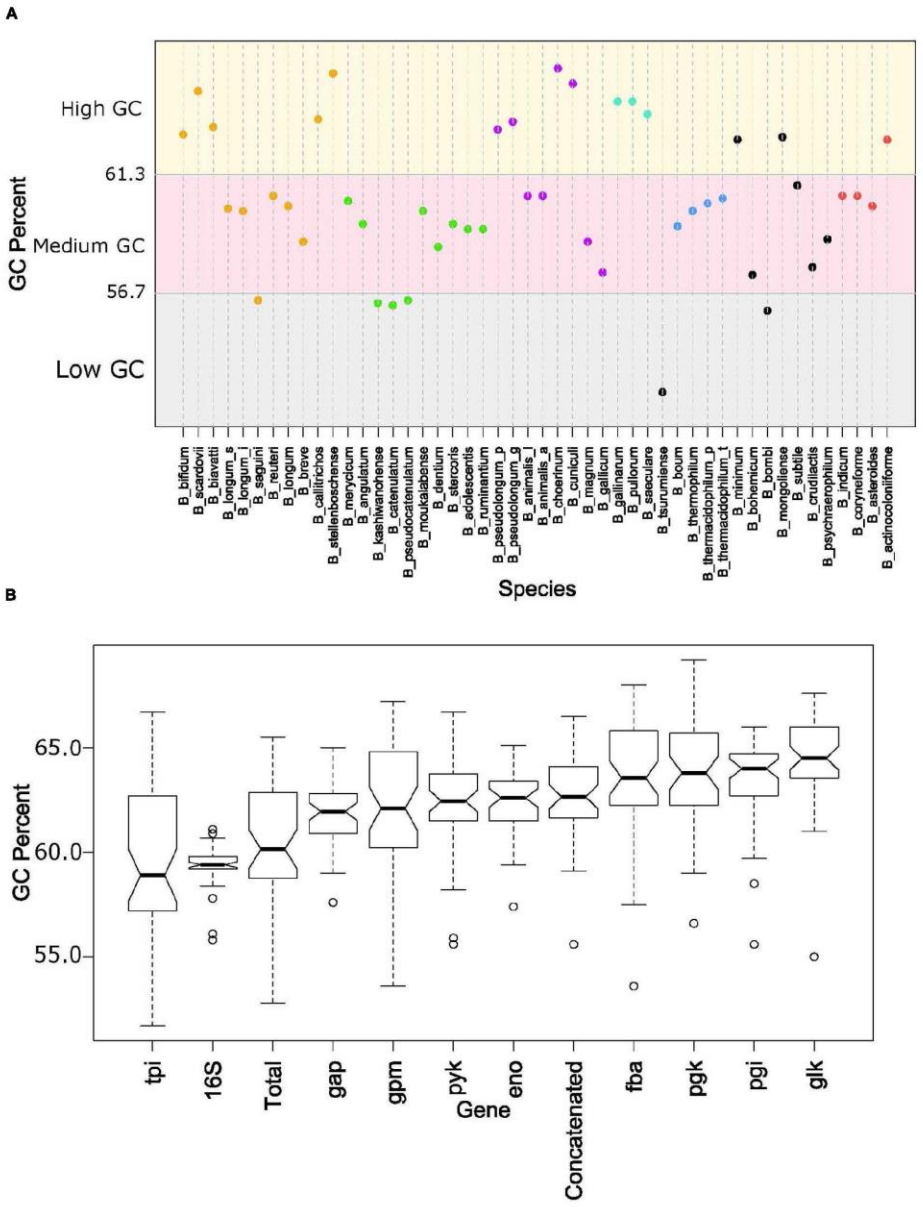


FIGURE 4 | GC content by species and glycolytic genes. (A) Shows the total GC content of each species organized by the glycolytic concatenated tree. Spectrum is split into three groups: low GC from 0.52 to 0.567 (gray), Medium GC from 0.567 to 0.613 (pink), and High GC from 0.613 to 0.66 (yellow). Phylogenetic groups are colored as follows: *B. longum* is orange, *B. adolescentis* is green, *B. psdeudolongum* is purple, *B. pollorum* is blue-green, *B. boum* is blue, and *B. asteroides* red. Species names following the naming convention from **Table 1**. **(B)** contains notched boxplots of the GC values of each gene and total GC. Boxes are ranked in order of median. Notches that do not overlap are indicative of strong evidence of difference between two medians.

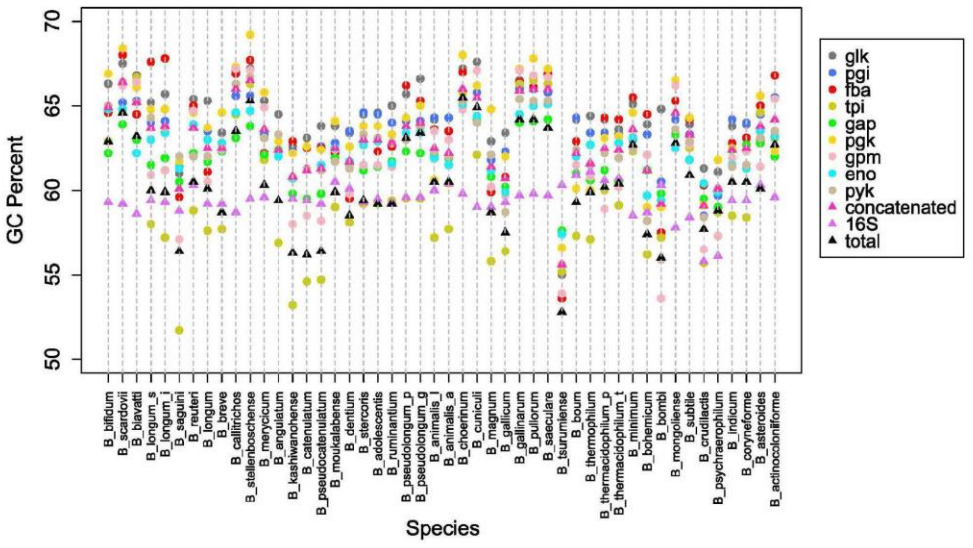


FIGURE 5 | Overall GC content patterns across species. GC percent for each glycolysis gene, 16SrRNA and overall genome, for species listed in **Table 1**.

in the high GC region and the medium GC region, respectively. All of the other groups, except the *B. longum* group, span two of these subsets. For the *B. longum* group, *Bifidobacterium saguini* lies just at the border between the low and medium GC subsets. This group has the largest spread, consistent with being the most diverse in the concatenated and 16S rRNA trees.

Next we looked at how the GC content varied across the trees. **Figure 4B** shows boxplots of the GC content of each tree and the total GC content. Except for the 16S rRNA and *tpi* trees, all other trees had median GC values with strong evidence of being higher than the median total GC content (Chambers, 1983). Looking on an individual basis, over half of the genomes have 16S rRNA and *tpi* GC values below their total GC, while the other genes are either above or close to their total GC (**Figure 5**). Again, the *B. pullorum* and *B. boum* groups are tightly packed in regards to their GC spread amongst their glycolysis genes, 16S rRNA, and total GC. In contrast, the *B. longum* group has the largest spread, a parallel to its higher diversity in the phylogenetic trees.

DISCUSSION

Bifidobacterium is a diverse genus of human intestinal beneficial microbes that provide health-promoting functionalities, as illustrated by their broad use as probiotics in foods and dietary supplements (Turroni et al., 2014). Recently, extensive genomic analyses of diverse species, subspecies and phylogenetic groups have provided insights into their adaptation to the human gut, notably with regards to their ability to colonize the intestinal cavity in general, and utilize non-digestible carbohydrates in particular (Milani et al., 2016). Studies investigating the use of human breast milk oligosaccharides illustrate the important contribution of these probiotics in establishing the human gut microbiome at the early stages of life (Sela, 2011). Yet, these studies also reveal that there are many distinct and diverse *Bifidobacterium* species and phylogenetic groups that colonize the human GIT, perhaps with idiosyncratic genomic attributes, and their corresponding functionalities (Chaplin et al., 2015). These organisms have specifically adapted to their environment to competitively utilize available nutrients (Sánchez et al., 2013). In the human gut, these consist of non-digestible complex oligosaccharides that are not adsorbed, nor broken down in the upper GIT. Whereas, plant-based fibers are important in the adult diet, HMOs are important components of the infant diet. Furthermore, *Bifidobacterium* have even been successful in helping each other through cross-feeding (Turroni et al., 2015). Thus, we addressed the need to establish practical means to allocate phylogeny with minimalistic information based on sequences that encode glycolysis, the biochemical spine of most cells.

Here, we have shown that a multigene approach using glycolysis sequences can be used to uncover genomic trends and to make an accurate phylogenetic tree, based on a relatively small amount of information. The concatenated glycolysis tree in **Figure 2** is congruent with both the 16S rRNA tree and the established core-genome-based tree (Milani et al., 2016). The only notable exception is the placement of *B. merycicum* and *B. angulatum*. However, the relocation was between two

neighboring phylogenetic groups in the concatenated and core-genome based trees. The glycolysis pathway is perhaps as, if not more, robust and accurate than the 16S rRNA tree. Compared to the 16S rRNA, the bootstrap values of the concatenated tree were higher on average. This leads to more confidence in the placement of species and the identification of phylogenetic groups, which in comparison, can appear arbitrarily located on the 16S rRNA. The concatenated tree is able to identify groups as well as the core-genome based tree. In fact, all of the phylogenetic groups from the core-genome were consistently found across the glycolytic pathway based trees. However, the glycolysis-based trees have the advantage of being much less labor intensive than the core-genome approach. This allows for accurate phylogenetic mapping of new strains or species, possibly encompassing unknown species, in less time and with less data than a core-genome. This approach is high resolution, low throughput, affordable, and accurate. Part of the success of this approach is the universality of glycolysis. Glycolysis is the biochemical backbone of the cell, and as such all organisms have at least some part of the glycolysis pathway represented (Fothergill-Gilmore and Michels, 1993). Even though these are slower-evolving genes, the changes that are made are enough to make an accurate phylogeny (Fothergill-Gilmore, 1986), evidenced from the congruence between our trees and the core-genome based tree. Even though the glycolysis enzymes are considered "slow evolvers," our data shows they are evolving at different rates amongst themselves. This can be explained by the fact that the glycolysis pathway is adapted by organisms to best fit their own unique environment and requirements (Bar-Even et al., 2012), as seen here in the *Bifidobacterium* and their bifid shunt (Sela et al., 2010). Some of the genes have specialized secondary functions, such as enolase acting as a cell surface receptor in *Bifidobacterium* (Candela et al., 2009). All of this makes the glycolysis pathway an excellent phylogenetic marker candidate. The various rates in evolution and moonlighting abilities also allow for further applications in recognizing adaptive trends.

The functional diversity of bifidobacteria is underpinned by multi-dimensional variety in their genomes, including overall content, organization, sequence diversity, and others. In extreme cases, even a two-fold difference in genome size can be observed. Despite being generally perceived as high GC organisms, they vary enough to have distinct relative classes of high, middle, and low-GC, amongst themselves (**Figure 4A**). Yet, there are non-random patterns and phenomena that drive these differences. The phylogenetic groups are clustered in specific regions of the GC continuum. Some groups are more tightly packed than others. A general trend that is observed across the genus is an evolutionary movement toward a high(er) GC content. The higher end of the spectrum is more densely populated then the lower end of the spectrum, indicative of an upward trend. This is reflected by the increased GC content in the individual glycolysis genes, when compared to the total GC content. Of the glycolysis genes, only one, *tpi*, does not show strong evidence for being different from the genome-wide (total) GC content. Critically, all of the other genes are above the total GC content. When we combine the overall genomic data with the GC-content groupings and trends discovered using glycolysis as phylogenetic markers, we posit the hypothesis that, over time, the GC content within the genomes of bifidobacteria increases, as to deviate further away from the 50% value, as the organisms adapt, and their genomes evolve accordingly.

Because of the broad occurrence of the glycolysis pathway in the Tree of Life, it is a suitable candidate marker to use in phylogenetic studies, likely beyond its application in bifidobacteria. In addition to being conserved genes that capture genetic diversity, glycolysis genes are consistently amongst the most highly expressed in not only *Bifidobacterium* (Turroni et al., 2015), but other organisms as well (Barrangou et al., 2006). This reflects both the importance of these sequences genetically (as illustrated by GC content drift), and functionally (as illustrated by their propensity for high levels of constitutive transcription). Because of this, it may be possible to correlate transcriptional data to phylogenetic studies on a broader scale. From here, it could be feasible to assign species and map data to known references using transcriptomic, genomic, or meta-data. Indeed, as the democratization of metagenomic technologies continues, and the need to assign phylogenetic information to partial genomic information increases, we propose that this method be used to provide insights into the phylogeny of un-assigned contigs. Overall, this approach allows for accurate phylogenetic mapping, congruent with a core-genome and more robust than the 16S rRNA phylogenetic approach, as well as inference on genomic adaptation, using either genomic, transcriptomic, or meta-data in a timely fashion and with minimal computation.

AUTHOR CONTRIBUTIONS

KB and RB designed and carried out experiments, interpreted results, and wrote the manuscript.

ACKNOWLEDGMENT

We would like to thank the Dr. Todd Klaenhammer lab and the CRISPR lab for providing insights and support during this project.

FUNDING

This study was supported by startup funds from North Carolina State University. KB is a recipient of a NIEHS training grant.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fmicb.2016.00657

REFERENCES

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410. doi: 10.1016/s0022-2836(05)80360-2

Arboleya, S., Sánchez, B., Milani, C., Duranti, S., Solís, G., Fernández, N., et al. (2015). Intestinal microbiota development in preterm neonates and effect of perinatal antibiotics. *J. Pediatr.* 166, 538–544. doi: 10.1016/j.jpeds.2014.09.041

Baker, G. C., Smith, J. J., and Cowan, D. A. (2003). Review and re-analysis of domain-specific 16S primers. *J. Microbiol. Methods* 55, 541–555. doi: 10.1016/j.mimet.2003.08.009

Bar-Even, A., Flamholz, A., Noor, E., and Milo, R. (2012). Rethinking glycolysis: on the biochemical logic of metabolic pathways. *Nat. Chem. Biol.* 8, 509–517. doi: 10.1038/nchembio.971

Barrangou, R., Azcarate-Peril, M. A., Duong, T., Conners, S. B., Kelly, R. M., and Klaenhammer, T. R. (2006). Global analysis of carbohydrate utilization by *Lactobacillus acidophilus* using cDNA microarrays. *Proc. Natl. Acad. Sci. U.S.A.* 103, 3816–3821. doi: 10.1073/pnas.0511287103

Bottacini, F., Ventura, M., van Sinderen, D., and O'Connell Motherway, M. (2014). Diversity, ecology and intestinal function of bifidobacteria. *Microb. Cell Fact.* 13(Suppl. 1), S4–S4. doi: 10.1186/1475-2859-13-S1-S4

Candela, M., Biagi, E., Centanni, M., Turroni, S., Vici, M., Musiani, F., et al. (2009). Bifidobacterial enolase, a cell surface receptor for human plasminogen involved in the interaction with the host. *Microbiology* 155, 3294–3303. doi: 10.1099/mic.0.028795-0

Capella-Gutiérrez, S., Silla-Martínez, J. M., and Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25, 1972–1973. doi: 10.1093/bioinformatics/btp348

Chambers, J. M. (1983). "Notched box plots," in *Graphical Methods for Data Analysis* (Belmont, CA: Wadsworth International Group), 60–63.

Chaplin, A. V., Efimov, B. A., Smeianov, V. V., Kafarskaia, L. I., Pikina, A. P., and Shkoporov, A. N. (2015). Intraspecies genomic diversity and long-term persistence of *Bifidobacterium longum*. *PLoS ONE* 10:e0135658. doi: 10.1371/journal.pone.0135658

Clarridge, J. E. (2004). Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. *Clin. Microbiol. Rev.* 17, 840–862. doi: 10.1128/CMR.17.4.840-862.2004

de Vries, W., and Stouthamer, A. H. (1967). Pathway of glucose fermentation in relation to the taxonomy of bifidobacteria. *J. Bacteriol.* 93, 574–576.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32, 1792–1797. doi: 10.1093/nar/gkh340

Eisen, J. A. (1995). The RecA protein as a model molecule for molecular systematic studies of bacteria: comparison of trees of RecAs and 16S rRNAs from the same species. *J. Mol. Evol.* 41, 1105–1123. doi: 10.1007/BF00173192

Fothergill-Gilmore, L. A. (1986). The evolution of the glycolytic pathway. *Trends Biochem. Sci.* 11, 47–51. doi: 10.1016/0968-0004(86)90233-1

Fothergill-Gilmore, L. A., and Michels, P. A. M. (1993). Evolution of glycolysis. *Prog. Biophys. Mol. Biol.* 59, 105–135. doi: 10.1016/0079-6107(93)90001-Z

Fox, G. E., Wisotzkey, J. D., and Jurtshuk, P. (1992). How close is close: 16S rRNA sequence identity may not be sufficient to guarantee species identity. *Int. J. Syst. Evol. Microbiol.* 42, 166–170. doi: 10.1099/00207713-42-1-166

Gasteiger, E., Gattiker, A., Hoogland, C., Ivanyi, I., Appel, R. D., and Bairoch, A. (2003). ExPASy: the proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res.* 31, 3784–3788. doi: 10.1093/nar/gkg563

Katoh, K., Misawa, K., Kuma, K. I, and Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* 30, 3059–3066. doi: 10.1093/nar/gkf436

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., et al. (2012). Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 28, 1647–1649. doi: 10.1093/bioinformatics/bts199

Killer, J., Sedláèek, I., Rada, V., Havlík, J., and Kopeènı, J. (2013). Reclassification of *Bifidobacterium* stercoris Kim et al. 2010 as a later heterotypic synonym of *Bifidobacterium adolescentis*. *Int. J. Syst. Evol. Microbiol.* 63, 4350–4353. doi: 10.1099/ijs.0.054957-0

Lang, J. M., Darling, A. E., and Eisen, J. A. (2013). Phylogeny of bacterial and archaeal genomes using conserved genes: supertrees and supermatrices. *PLoS ONE* 8:e62510. doi: 10.1371/journal.pone.0062510

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., et al. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics* 23, 2947–2948. doi: 10.1093/bioinformatics/btm404

Lartillot, N., and Philippe, H. (2004). A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol. Biol. Evol.* 21, 1095–1109. doi: 10.1093/molbev/msh112

Lugli, G. A., Milani, C., Turroni, F., Duranti, S., Ferrario, C., Viappiani, A., et al. (2014). Investigation of the evolutionary development of the genus *Bifidobacterium* by comparative genomics. *Appl. Environ. Microbiol.* 80, 6383–6394. doi: 10.1128/aem.02004-14

Lukjancenko, O., Ussery, D. W., and Wassenaar, T. M. (2011). Comparative genomics of *Bifidobacterium*, *Lactobacillus* and related probiotic genera. *Microb. Ecol.* 63, 651–673. doi: 10.1007/s00248-011-9948-y

Medini, D., Donati, C., Tettelin, H., Masignani, V., and Rappuoli, R. (2005). The microbial pan-genome. *Curr. Opin. Genet. Dev.* 15, 589–594. doi: 10.1016/j.gde.2005.09.006

Milani, C., Lugli, G. A., Duranti, S., Turroni, F., Bottacini, F., Mangifesta, M., et al. (2014). Genomic encyclopedia of type strains of the genus *Bifidobacterium*. *Appl. Environ. Microbiol.* 80, 6290–6302. doi: 10.1128/aem.02308-14

Milani, C., Mancabelli, L., Lugli, G. A., Duranti, S., Turroni, F., Ferrario, C., et al. (2015). Exploring vertical transmission of bifidobacteria from mother to child. *Appl. Environ. Microbiol.* 81, 7078–7087. doi: 10.1128/aem.02037-15

Milani, C., Turroni, F., Duranti, S., Lugli, G. A., Mancabelli, L., Ferrario, C., et al. (2016). Genomics of the genus *Bifidobacterium* reveals species-specific adaptation to the glycan-rich gut environment. *Appl. Environ. Microbiol.* 82, 980–991. doi: 10.1128/aem.03500-15

Pokusaeva, K., Fitzgerald, G. F., and Sinderen, D. (2011). Carbohydrate metabolism in bifidobacteria. *Genes Nutr.* 6, 285–306. doi: 10.1007/s12263-010-0206-6

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Sánchez, B., Ruiz, L., Gueimonde, M., Ruas-Madiedo, P., and Margolles, A. (2013). Adaptation of bifidobacteria to the gastrointestinal tract and functional consequences. *Pharmacol. Res.* 69, 127–136. doi: 10.1016/j.phrs.2012.11.004

Sela, D. A. (2011). Bifidobacterial utilization of human milk oligosaccharides. *Int. J. Food Microbiol.* 149, 58–64. doi: 10.1016/j.ijfoodmicro.2011.01.025

Sela, D. A., Price, N. P. J., and Mills, D. (2010). "Metabolism of bifidobacteria," in *Bifidobacteria: Genomics and Molecular Aspects*, eds B. Mayo and D. van Sinderen (Norwich: Caister Academic Press).

Stamatakis, A. (2006a). "Phylogenetic models of rate heterogeneity: a high performance computing perspective," in *Proceedings 20th IEEE International Parallel & Distributed Processing Symposium* (Rhodes Island: IEEE).

Stamatakis, A. (2006b). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22, 2688–2690. doi: 10.1093/bioinformatics/btl446

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30, 1312–1313. doi: 10.1093/bioinformatics/btu033

Tamura, K., Stecher, G., Peterson, D., Filipski, A., and Kumar, S. (2013). MEGA6: molecular evolutionary genetics analysis version 6.0. *Mol. Biol. Evol.* 30, 2725–2729. doi: 10.1093/molbev/mst197

Turroni, F., Duranti, S., Bottacini, F., Guglielmetti, S., Van Sinderen, D., and Ventura, M. (2014). *Bifidobacterium bifidum* as an example of a specialized human gut commensal. *Front. Microbiol.* 5:437. doi: 10.3389/fmicb.2014.00437

Turroni, F., Özcan, E., Milani, C., Mancabelli, L., Viappiani, A., van Sinderen, D., et al. (2015). Glycan cross-feeding activities between bifidobacteria under in vitro conditions. *Front. Microbiol.* 6:1030. doi: 10.3389/fmicb.2015.01030

Turroni, F., Peano, C., Pass, D. A., Foroni, E., Severgnini, M., Claesson, M. J., et al. (2012a). Diversity of bifidobacteria within the infant gut microbiota. *PLoS ONE* 7:e36957. doi: 10.1371/journal.pone.0036957

Turroni, F., Strati, F., Foroni, E., Serafini, F., Duranti, S., van Sinderen, D., et al. (2012b). Analysis of predicted carbohydrate transport systems encoded by *Bifidobacterium bifidum* PRL2010. *Appl. Environ. Microbiol.* 78, 5002–5012. doi: 10.1128/AEM.00629-12

Ventura, M., Canchaya, C., Tauch, A., Chandra, G., Fitzgerald, G. F., Chater, K. F., et al. (2007). Genomics of Actinobacteria: tracing the evolutionary

history of an ancient phylum. *Microbiol. Mol. Biol. Rev.* 71, 495–548. doi: 10.1128/MMBR.00005-07

Ventura, M., Turroni, F., Lugli, G. A., and van Sinderen, D. (2014). Bifidobacteria and humans: our special friends, from ecological to genomics perspectives. *J. Sci. Food Agric.* 94, 163–168. doi: 10.1002/jsfa.6356

Ventura, M., Turroni, F., Motherway, M. O. C., MacSharry, J., and van Sinderen, D. (2012). Host–microbe interactions that facilitate gut colonization by commensal bifidobacteria. *Trends Microbiol.* 20, 467–476. doi: 10.1016/j.tim.2012.07.002

Conflict of Interest Statement: 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.

APPENDIX B: REPRINT OF "USING GLYCOLYSIS ENZYME SEQUENCES TO INFORM *LACTOBACILLUS* PHYLOGENY"

MICROBIAL GENOMICS

RESEARCH ARTICLE

Brandt and Barrangou, *Microbial Genomics* 2018;4

DOI 10.1099/mgen.0.000187

MICROBIOLOGY SOCIETY

OPEN DATA OPEN MICROBIOLOGY

Using glycolysis enzyme sequences to inform *Lactobacillus* phylogeny

Katelyn Brandt[1,2] and Rodolphe Barrangou[1,2,*]

Abstract

The genus *Lactobacillus* encompasses a diversity of species that occur widely in nature and encode a plethora of metabolic pathways reflecting their adaptation to various ecological niches, including humans, animals, plants and food products. Accordingly, their functional attributes have been exploited industrially and several strains are commonly formulated as probiotics or starter cultures in the food industry. Although divergent evolutionary processes have yielded the acquisition and evolution of specialized functionalities, all *Lactobacillus* species share a small set of core metabolic properties, including the glycolysis pathway. Thus, the sequences of glycolytic enzymes afford a means to establish phylogenetic groups with the potential to discern species that are too closely related from a 16S rRNA standpoint. Here, we identified and extracted glycolysis enzyme sequences from 52 species, and carried out individual and concatenated phylogenetic analyses. We show that a glycolysis-based phylogenetic tree can robustly segregate lactobacilli into distinct clusters and discern very closely related species. We also compare and contrast evolutionary patterns with genome-wide features and transcriptomic patterns, reflecting genomic drift trends. Overall, results suggest that glycolytic enzymes provide valuable phylogenetic insights and may constitute practical targets for evolutionary studies.

DATA SUMMARY

RNA sequencing data has been deposited at the National Center for Biotechnology Information, BioProject PRJNA420353.

INTRODUCTION

Genome adaptation is an important feature for speciation, and evolutionary processes balance various adaptive techniques for optimal growth and survival. At the genome level, adaptation features may include gene synteny conservation, G+C mol% drift, as well as codon bias optimization [1–3]. A working balance of these and other forces enable an organism to become uniquely adapted to its niche, and build up competitive advantages in shifting environmental conditions, or overcome predators and competitors. Such unique adaptations are the basis of phylogenetic studies and allow researchers various degrees of discrimination. At the genus and species levels, additions and deletions of genes can be used to define the pan- and core-genome and genome architecture can be used to evaluate synteny [4]. At the strain level, nucleotide polymorphisms afford the highest resolution opportunities, with the ability to compare and contrast nearly identical isolates and even clonal relatives [5, 6].

For prokaryotic species, various tools and methodologies have been used to compare and contrast genomes, but the challenges are often genus- or species-specific, and approaches can vary depending on the desired resolution and encompassed genetic diversity [7]. In some cases where within genus diversity is extensive, such as in bifidobacteria and lactobacilli, using canonical housekeeping genes or universal markers (i.e. 16S rRNA) has proven difficult or limited [8–11]. Also, there has yet to be defined a consistent set of genes to be utilized for multilocus sequence typing studies. Indeed, while universally conserved 16S rRNA sequences afford opportunities for metagenomic analyses, their shortcomings and biases are increasingly under scrutiny [12–14].

Received 13 March 2018; Accepted 7 May 2018

Author affiliations: [1]Genomic Sciences Graduate Program, North Carolina State University, Raleigh, NC 27695, USA; [2]Department of Food, Bioprocessing and Nutrition Sciences, North Carolina State University, Raleigh, NC 27695, USA.

***Correspondence:** Rodolphe Barrangou, rbarran@ncsu.edu

Keywords: *Lactobacillus*; phylogeny; glycolysis; evolution.

Abbreviations: *eno*, enolase-encoding gene; *fba*, fructose bisphosphate aldolase-encoding gene; *gap*, glyceraldehyde 3-phosphate dehydrogenase-encoding gene; *gpm*, phosphoglycerate mutase-encoding gene; LAB, lactic acid bacteria; mRNA-Seq, mRNA sequencing; *pfk*, 6-phosphofructokinase-encoding gene; *pgi*, glucose-6-phosphate isomerase-encoding gene; *pgk*, phosphoglycerate kinase-encoding gene; *pgm*, phosphoglucomutase-encoding gene; *pyk*, pyruvate kinase-encoding gene; *tpi*, triosephosphate isomerase-encoding gene.

Data statement: All supporting data, code and protocols have been provided within the article or through supplementary data files. One supplementary table and thirteen supplementary figures are available with the online version of this article.

For some genera, it has become obvious that the 16S rRNA resolution limit has been met and a new set of criteria must be established. One such genus is *Lactobacillus*. Belonging to the lactic acid bacteria (LAB) group, this genus is composed of over 150 Gram-positive, low G+C species [15, 16]. Lactobacilli have been used as starter cultures in the food industry for decades, and by humankind for millennia, and as such have been labelled generally regarded as safe (GRAS) and benefit from the qualified presumption of safety (QPS) [17]. Food-related studies have led to the assertion that some strains in select species are to be considered probiotic ('live microorganisms which when administered in adequate amounts confer a health benefit on the host') [18] and, as such, are now predominantly featured in dairy foods and widely formulated in probiotic dietary supplements [19]. Recently, the advent of microbiome studies has revealed that microbial populations are more numerous, diverse and variable than originally thought [20, 21]. With both qualitative and quantitative considerations, associations and sometimes even correlations have been established between members of the microbiome and host health, though the accuracy and precision with which bacteria are identified vary widely and are not universally satisfactory. One such instance concerns the genus *Lactobacillus*, which has been established as an important colonizer of the human gastrointestinal tract [22]. Additional research is thus needed in this area, as researchers better grasp the role of this genus in health and disease [23–28]. Some lactobacilli are already being exploited, for example, as a tool to deliver vaccines [29]. Arguably, we are far from exhausting all the possible uses of this functional genus. However, in order to be able to fully utilize the numerous functions of *Lactobacillus*, we must first establish a method that enables us to properly identify and relate the many diverse species within this genus. While 16S rRNA sequencing has gotten us this far, it has a limited ability to distinguish between closely related species and represent overall genomic content and reflect genome-wide trends. These shortcomings are certainly not unique to *Lactobacillus,* and with the ever-increasing expansion of our understanding of the microbial world [30], there is a need to identify 16S rRNA-independent genomic features that capture diversity on a more granular level. Thus, it is imperative that a standard method be developed that allows the proper identification of species. In order to achieve this, we assessed the potential of the widespread glycolysis pathway enzyme sequences to inform phylogeny.

In this paper, we applied a previously described method of phylogenetic analysis using the classical glycolysis enzymes as phylogenetic markers [31] to a diverse set of *Lactobacillus* species in order to establish its effect on a complicated genus. Though previous studies had used glycolysis as an expansion of ribosomal trees [32], we determined how a broad glycolysis-based phylogeny compares to the ribosomal tree. Specifically, previous studies have applied glycolysis-based approaches to LAB in order to define an evolutionary pathway. By adding data from the entirety of the glycolysis and pentose phosphate pathways, Salvetti *et al.* [32] were able to apply phenotypic data to explain the branching of the LAB tree, as well as highlight some areas of misclassification in the 16S rRNA tree [32]. Here, we propose using the entirety of the canonical glycolysis pathway as a replacement phylogenetic marker for the 16S rRNA. Conveniently, the glycolysis pathway, much like the 16S rRNA, is universally present, at least partially, conserved, and constitutes a set of suitable candidates for phylogenetic analyses [33, 34]. Here, we demonstrate that this method can assign phylogenetic relationships consistent with what is known from the 16S rRNA marker, though at a much higher discriminatory power. Specifically, we compared sequence-based alignment trees of a representative set of lactobacilli using 16S rRNA- and glycolysis-based approaches. We also analysed the occurrence and location, expression, and G+C mol% of each glycolysis gene. The location and transcriptional profiles confirm that these genes are conserved and highly transcribed with varying levels of drift.

IMPACT STATEMENT

Though 16S rRNA-based phylogeny methods have been broadly used, they have a limited ability to precisely ascribe genus species across the prokaryotic branch of the tree of life. In this study, we have shown that using glycolysis enzyme sequences for phylogenetic analyses can be applied to the diverse genus *Lactobacillus*, and is able to consistently unravel phylogenetic groups and precisely ascertain relatedness, even between species nearly identical on the classical ribosomal tree. Because of their universal presence and greater diversity compared to 16S rRNA sequences, we posit that these sequences could be valuable markers in future phylogenetic and microbiome studies, specifically by providing connections to the other major branches, and enabling increased resolution. This can also be used to help identify unknown and un-culturable species, as the glycolysis enzymes are widespread, variable and allow for greater discriminatory power. Importantly, variability within some of the hypervariable regions within glycolytic sequences can also provide discrimination within a species. Looking forward, expanding this analysis to other genera and phylogenetic branches could open new avenues for evolutionary studies, and for investigating the phylogeny, composition and diversity of microbial populations in complex microbiomes.

METHODS

Genomes

We selected 52 diverse species and subspecies of *Lactobacillus* for analysis, sampled across and throughout the 16S rRNA and core- and pan-genome tree (Table 1). We

Table 1. Species and genomes list

This shows the representative set of 52 *Lactobacillus* species and sub-species used in this study. Accession numbers and naming conventions are included.

Genus	Species	Subspecies	Strain	Accession no.	Naming convention	Locus tag
Lactobacillus	*acidipiscis*		KCTC 13900	NZ_BACS00000000	L_acidipiscis	GSS
Lactobacillus	*acidophilus*		NCFM	NC_006814	L_acidophilus	LBA
Lactobacillus	*algidus*		DSM 15638	NZ_AZDI00000000	L_algidus	FC66
Lactobacillus	*amylolyticus*		DSM 11664	NZ_ADNY00000000	L_amylolyticus	HMPREF0493
Lactobacillus	*amylovorus*		GRL1118	NC_017470	L_amylovorus	LAB52
Lactobacillus	*animalis*		DSM 20602	NZ_AEOF00000000	L_animalis	LACAN
Lactobacillus	*aquaticus*		DSM 21051	NZ_AYZD00000000	L_aquaticus	FC19
Lactobacillus	*brevis*		ATCC 367	NC_008497	L_brevis	LVIS
Lactobacillus	*buchneri*		CD034	NC_018610	L_buchneri	LBUCD034
Lactobacillus	*cacaonum*		DSM 21116	NZ_AYZE00000000	L_cacaonum	FC80
Lactobacillus	*casei*		DSM 20011	NZ_AZCO00000000	L_casei	FC13
Lactobacillus	*coryniformis*	*torquens*	DSM 20004	NZ_AEOS00000000	L_coryniformis_t	EWE
Lactobacillus	*crispatus*		ST1	NC_014106	L_crispatus	LCRIS
Lactobacillus	*curvatus*		CRL 705	NZ_AGBU00000000	L_curvatus	CRL705
Lactobacillus	*delbrueckii*	*bulgaricus*	ATCC BAA-365	NC_008529	L_delbrueckii_b	LBUL
Lactobacillus	*farciminis*		DSM 20184	NZ_AEOT00000000	L_farciminis	LACPC
Lactobacillus	*fermentum*		CECT 5716	NC_017465	L_fermentum	LC40
Lactobacillus	*floricola*		DSM 23037	NZ_AYZL00000000	L_floricola	FC86
Lactobacillus	*gallinarum*		DSM 10532	NZ_BALB00000000	L_gallinarum	JCM2011
Lactobacillus	*gasseri*		ATCC 33323	NC_008530	L_gasseri	LGAS
Lactobacillus	*helveticus*		CNRZ32	NC_021744	L_helveticus	LHE
Lactobacillus	*hilgardii*		DSM 20176	NZ_ACGP00000000	L_hilgardii	HMPREF0519
Lactobacillus	*hominis*		DSM 23910	NZ_CAKE00000000	L_hominis	BN55
Lactobacillus	*iners*		DSM 13335	NZ_ACLN00000000	L_iners	HMPREF0520
Lactobacillus	*jensenii*		DSM 20557	NZ_AYYU00000000	L_jensenii	FC45
Lactobacillus	*johnsonii*		NCC 533	NC_005362	L_johnsonii	LJ
Lactobacillus	*kimchicus*		JCM_15530	NZ_AZCX00000000	L_kimchicus	FC96
Lactobacillus	*lindneri*		DSM 20690	NZ_JQBT00000000	L_lindneri	IV52
Lactobacillus	*mali*		DSM 20444	NZ_AKKT00000000	L_mali	LMA
Lactobacillus	*mindensis*		DSM 14500	NZ_AZEZ00000000	L_mindensis	FD29
Lactobacillus	*mucosae*		LM1	NZ_CP011013	L_mucosae	LBLM1
Lactobacillus	*nasuensis*		JCM_17158	NZ_AZDJ00000000	L_nasuensis	FD02
Lactobacillus	*oeni*		DSM 19972	NZ_AZEH00000000	L_oeni	FD46
Lactobacillus	*oris*		F0423	NZ_AFTL00000000	L_oris	HMPREF9102
Lactobacillus	*otakiensis*		DSM 19908	NZ_BASH00000000	L_otakiensis	LOT
Lactobacillus	*parabuchneri*		DSM 5707	NZ_AZGK00000000	L_parabuchneri	FC51
Lactobacillus	*paracasei*		N1115	NZ_CP007122	L_paracasei	AF91
Lactobacillus	*pasteurii*		DSM 23907	NZ_CAKD00000000	L_pasteurii	BN53
Lactobacillus	*pentosus*		DSM 20314	NZ_AZCU00000000	L_pentosus	FD24
Lactobacillus	*plantarum*		16	NC_021514	L_plantarum	LP16
Lactobacillus	*reuteri*		DSM 20016	NC_009513	L_reuteri	LREU
Lactobacillus	*rhamnosus*		GG	NC_013198	L_rhamnosus	LGG
Lactobacillus	*rossiae*		DSM 15814	NZ_AZFF00000000	L_rossiae	FD35
Lactobacillus	*ruminis*		ATCC 27782	NC_015975	L_ruminis	LRC
Lactobacillus	*sakei*	*sakei*	DSM 20017	NZ_BALW00000000	L_sakei_s	JCM1157
Lactobacillus	*salivarius*		CECT 5713	NC_017481	L_salivarius	CECT 5713
Lactobacillus	*sanfranciscensis*		TMW 1.1304	NC_015978	L_sanfranciscensis	LSA
Lactobacillus	*suebicus*		DSM 5007	NZ_BACO00000000	L_suebicus	GSK
Lactobacillus	*sunkii*		DSM 19904	NZ_AZEA00000000	L_sunkii	FD17
Lactobacillus	*vaginalis*		DSM 5837	NZ_ACGV00000000	L_vaginalis	HMPREF0549

Table 1. cont.

Genus	Species	Subspecies	Strain	Accession no.	Naming convention	Locus tag
Lactobacillus	*versmoldensis*		DSM 14857	NZ_BACR00000000	L_versmoldensis	GSQ
Lactobacillus	*zymae*		DSM 19395	NZ_AZDW00000000	L_zymae	FD38

ensured this set was representative of this paraphyletic genus and included species from various niches, as previously established [16]. The genomes were mined using Geneious version 9.0.5 [35] to identify the classical glycolysis genes in each species (Figs S1 and S2, available with the online version of this article). Four reference genomes were

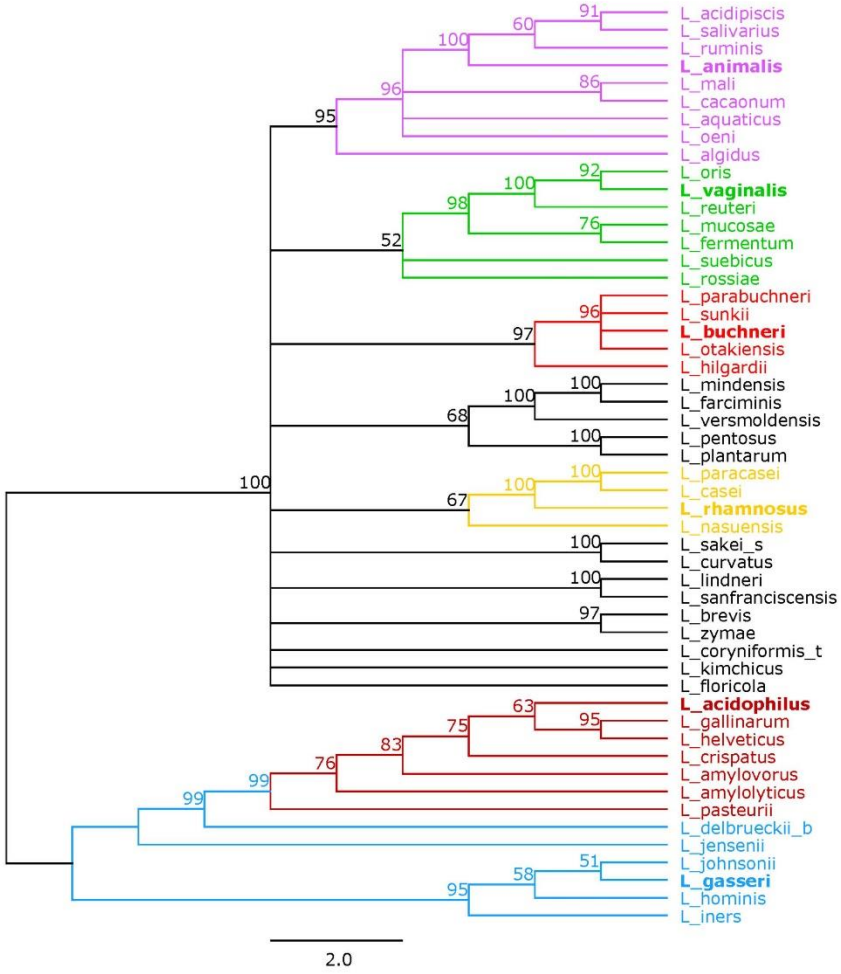


Fig. 1. 16S rRNA tree. Tree based on the alignment of the 16S rRNA sequences using RaxML. Bootstrap values are recorded on the nodes. Groups are coloured as follows: the *L. animalis* group in purple, the *L. vaginalis* group in green, the *L. buchneri* group in red, the *L. rhamnosus* group in yellow, the *L. acidophilus* group in maroon, and the *L. gasseri* group in blue. The representative species in each group is in bold. Species names follow the naming convention shown in Table 1.

used to make a curated database for the glycolysis genes, namely *Lactobacillus acidophilus, Lactobacillus gasseri, Lactobacillus reuteri* and *Lactobacillus rhamnosus*. The Annotate from Database feature was used to annotate the other genomes. To validate the glycolysis annotations, especially in the case of multiple hits, a combination of BLAST, GET_HOMOLOGUES and mRNA-Seq (mRNA sequencing) data was used [36, 37]. The 16S rRNA sequences were extracted from the genomes and BLAST was used to validate any cases where there were multiple hits. Once annotated and curated, the genes were extracted from the genome. The glycolysis genes were then translated and confirmed by ExPASy [38]. For the concatenated tree, the amino acid sequences were joined together in order of their presence in the glycolysis pathway (Fig. S1).

Transcriptional profiles of glycolysis genes

We analysed RNA transcription profiles from mRNA-Seq data for six species (*L. acidophilus, Lactobacillus amylovorus, Lactobacillus crispatus, Lactobacillus delbrueckii* subsp. *bulgaricus, L. gasseri, and Lactobacillus helveticus*) with the previously published isolation method, mRNA sequencing and analyses [39]. Briefly, we used mRNA-Seq data generated in our laboratory to determine the boundaries and quantitative amounts of RNA transcripts for glycolysis genes as previously described. Samples were grown to mid-log phase and flash-frozen. Single-read RNA sequencing was performed on the extracted RNA using an Illumina HiSeq 2500. Data was then quality assessed, trimmed, filtered and mapped on the reference genomes. Presumably, levels of constitutive transcription reflect biological relevance in the tested conditions and transcript boundaries inform on co-transcribed functional pairs.

Alignments and trees

Alignments and trees were generated using a previously described methodology [31]. Briefly, once curated sequences were extracted, we aligned the sequences using CLUSTALW (IUB, gap penalty of 15, gap extension of 6.66), MUSCLE (eight iterations), Geneious [global alignment with free end gaps, cost matrix was BLOSUM62 (amino acids) or 65 % similarity (nucleotide)] and MAFFT [algorithm was auto, scoring matrix was BLOSUM62 and BLOSUM80 (amino acids) or 100PAM and 200PAM (nucleotide), gap penalty of 1.53, offset 0.123], then used trimAl (compareset and automated1) to find a consistent alignment [35, 40–43]. Trees were then generated using RaxML [CAT BLOSUM62 (amino acids) or CAT GTR (nucleotide), Bootstrap using rapid hill climbing with random seed 1, replicates were 100] [44]. A consensus tree was then established using a 50 % threshold level.

R analyses

Statistical analyses were performed using R version 3.2.2. [45]. R was used to create plots, graphs and quantitative data. Statistical tests used included a two-tailed *t*-test for comparing G+C contents. Default settings were used to

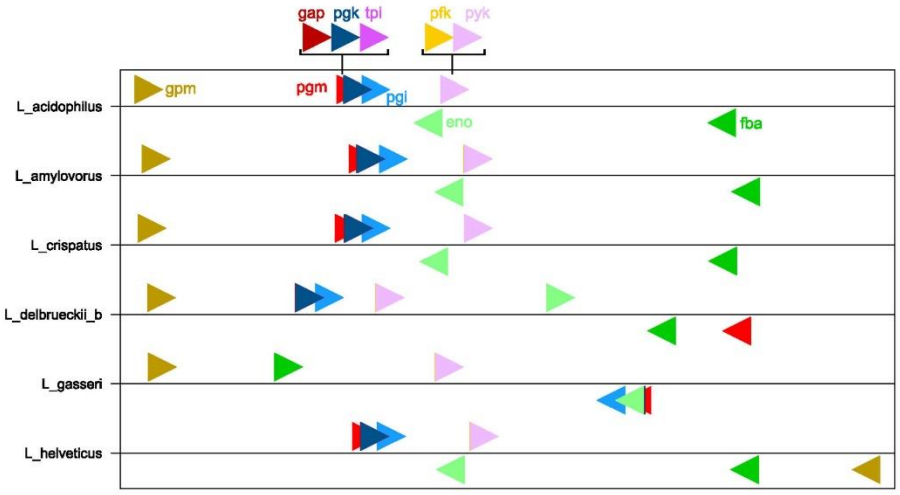


Fig. 2. Genomic location. Normalized glycolysis gene locations in *L. acidophilus, L. amylovorus, L. crispatus, L. delbrueckii* subsp. *bulgaricus, L. gasseri* and *L. helveticus*. Normalization was calculated by dividing the location on the genome by the total genome size. Right arrows indicate forward direction, left reverse direction. The genomes are organized in the 5′ to 3′ direction. Colours are as follows: *pgm* in red, *pgi* in blue, *pfk* in yellow, *fba* in dark green, *tpi* in purple, *gap* in maroon, *pgk* in navy, *gpm* in mustard, *eno* in light green and *pyk* in lavender.

preform statistical analyses and assess quantitative distributions.

RESULTS

16S rRNA phylogeny

We first generated a 16S rRNA-based tree to use as a reference for our subsequent analyses. A phylogenetic tree based on the alignment of the 16S rRNA sequences from a representative set of 52 species and sub-species of *Lactobacillus* is depicted in Fig. 1. Six phylogenetic groups were identified based on their branching: the *Lactobacillus animalis* group, the *Lactobacillus vaginalis* group, the *Lactobacillus buchneri* group, the *L. rhamnosus* group, the *L. acidophilus* group and the *L. gasseri* group. These groupings are consistent with historically established relationships, as well as recent core-genome analyses [16, 46]. Some of these groups also encompass species that have been historically associated with distinct niches and points of isolation (i.e. mucosal vs intestinal vs dairy origins) [16]. The groups ranged in size from four to nine genomes with the *L. rhamnosus* group as the smallest and the *L. animalis* group as the largest. The bootstrap values for the 16S rRNA tree ranged from 51 to 100. There were 27 nodes that had a bootstrap of 70 or greater (Fig. S3). We used these six phylogenetic groups as references for our subsequent analyses, though some species were not assigned to one of these six groups.

Glycolysis gene expression

Before using the glycolysis enzymes as phylogenetic markers, we first explored their genetic properties in *Lactobacillus*. Of the 52 *Lactobacillus* species and sub-species selected, 35 species encoded all ten of the classical glycolytic genes. In contrast, 16 species (encompassing the *L. vaginalis* and *L. buchneri* groups) presented eight of the canonical genes (missing *pfk* and *fba*) (Fig. S2). In such cases, alternative metabolic pathways may be utilized, such as the pentose phosphate pathway (*Lactobacillus fermentum*) or the phosphoketolase pathway (*L. buchneri*) [47, 48]. *L. reuteri* uses a mixture of the Embden–Meyerhof pathway and phosphoketolase pathway and, thus, was the only species with six of the glycolysis genes (Fig. S2) [49].

Next, we characterized the transcripts of glycolysis genes in *Lactobacillus*. Chromosome location and mRNA sequence data were analysed from six species: *L. acidophilus*, *L. amylovorus*, *L. crispatus*, *L. delbrueckii* subsp. *bulgaricus*, *L. gasseri* and *L. helveticus*. These six species fall into the *L. acidophilus* and *L. gasseri* groups, and all six species contain the complete complement of glycolysis genes, allowing

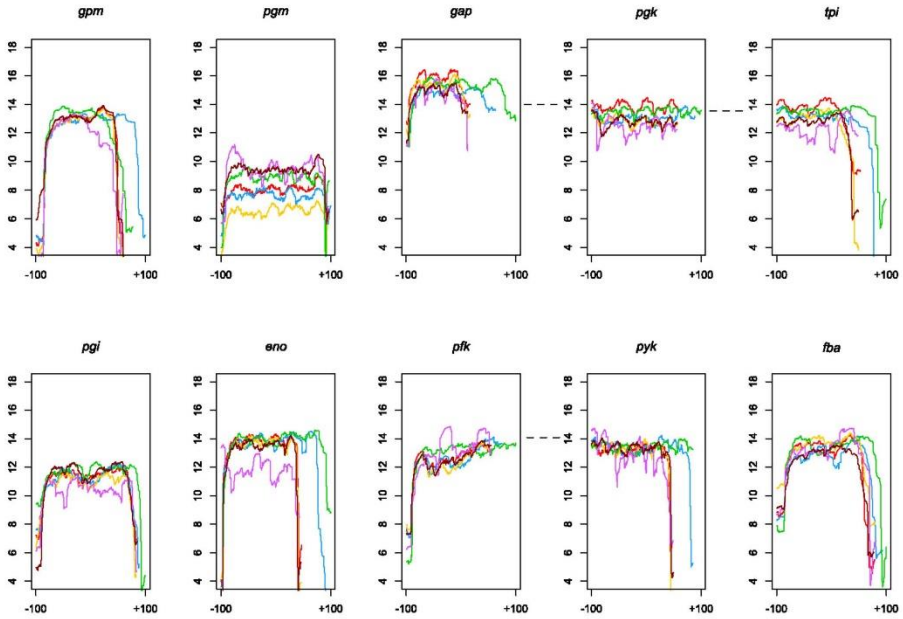


Fig. 3. Glycolysis genes transcription. Each plot represents the mRNA-Seq coverage, log2 transformed, for the corresponding glycolysis gene over its length; ±100 represents the number of bases away from the start/end of the gene. The species are plotted as follows: *L. acidophilus* is red, *L. amylovorus* in blue, *L. crispatus* in yellow, *L. delbrueckii* subsp. *bulgaricus* in green, *L. gasseri* in purple and *L. helveticus* in maroon.

for inferences on all of the genes in this study, instead of just a subset. Fig. 2 depicts the location of the glycolysis genes on normalized chromosomes for each of these six species. It is noteworthy that two operons can be visualized: the *gap*, *pgk* and *tpi* operon, as well as the *pfk* and *pyk* operon. Furthermore, the operon boundaries are clearly seen in the mRNA coverage data for each of the six species (Fig. 3). The remaining five genes have clear start and stop boundaries. Notably, *L. helveticus* has a unique arrangement of the glycolysis genes compared to the other five species, possibly due to the large number of IS elements leading to genome decay; however, the operons remain conserved [50]. Next, we compared the expression levels of the glycolysis genes to the whole transcriptome. We found that the glycolysis genes are among the most highly expressed genes. Indeed, considering the top 10 % of the most highly expressed genes in the cell, nine of the ten glycolysis genes are listed (Fig. 4). The only gene absent from the top 10 % is *pgm*. Strikingly, the *gap* gene is consistently among the top three most highly expressed genes in all six species. Such a consistently high transcription level indicates that the *gap* gene is critical to the functionality of the cell and perhaps, as such, less susceptible to changes. This is also reflected by the conserved location of *gap* in the genome and operon structure amongst the strains studied (Fig. 2), potentially indicating uses for *gap* in identification. These results demonstrate that glycolysis genes are genomically conserved, organizationally syntenous and transcriptionally important, showcasing their use as potential phylogenetic markers.

Glycolysis-based phylogeny

To create a glycolysis-based phylogeny for the 52 selected *Lactobacillus* species and subspecies, the concatenated amino acid sequences of the glycolysis enzymes were used (Fig. 5). The enzymes were concatenated in their order of occurrence in the glycolysis pathway (Fig. S1). For organisms with all enzymes present, this meant ten sequences were concatenated together, whereas only six to eight amino acid sequences were concatenated for the other species (Fig. S2). The six phylogenetic groups identified from the 16S rRNA reference tree, namely *L. animalis, L. vaginalis, L. buchneri, L. rhamnosus, L. acidophilus* and *L. gasseri*, were also identified in the concatenated tree and follow the same clustering (colouring) scheme. The bootstrap values

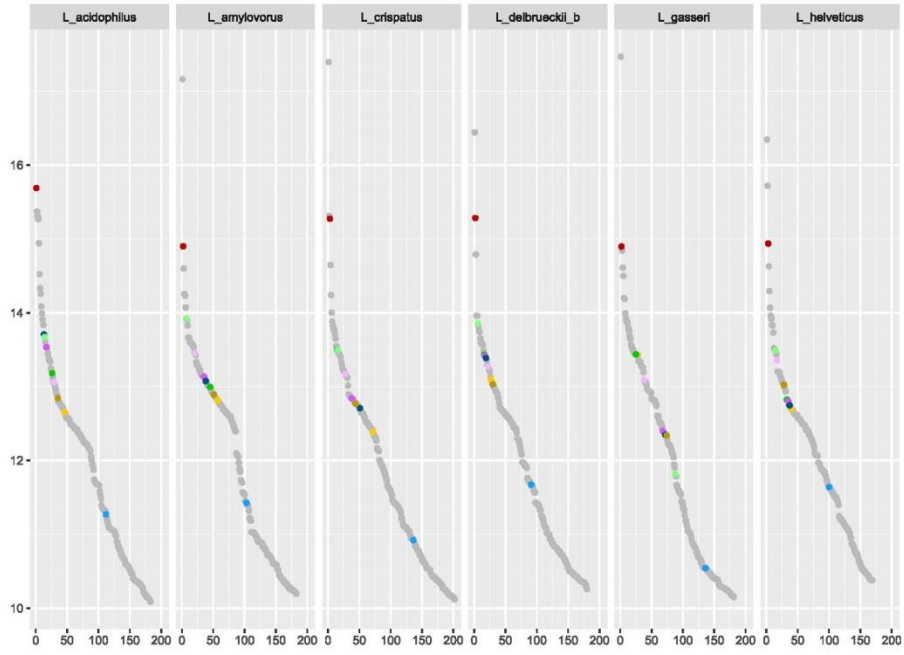


Fig. 4. Ranked order of mRNA expression. Top 10 % most highly expressed genes in *L. acidophilus, L. amylovorus, L. crispatus, L. delbrueckii* subsp. *bulgaricus, L. gasseri* and *L. helveticus*. Data is represented as a log2 transformed RPKM (Reads Per Kilobase of transcript, per Million mapped reads). Transcripts are ranked from most abundant to least abundant. Glycolysis genes are coloured as follows: *pgm* in red, *pgi* in blue, *pfk* in yellow, *fba* in dark green, *tpi* in purple, *gap* in maroon, *pgk* in navy, *gpm* in mustard, *eno* in light green and *pyk* in lavender.

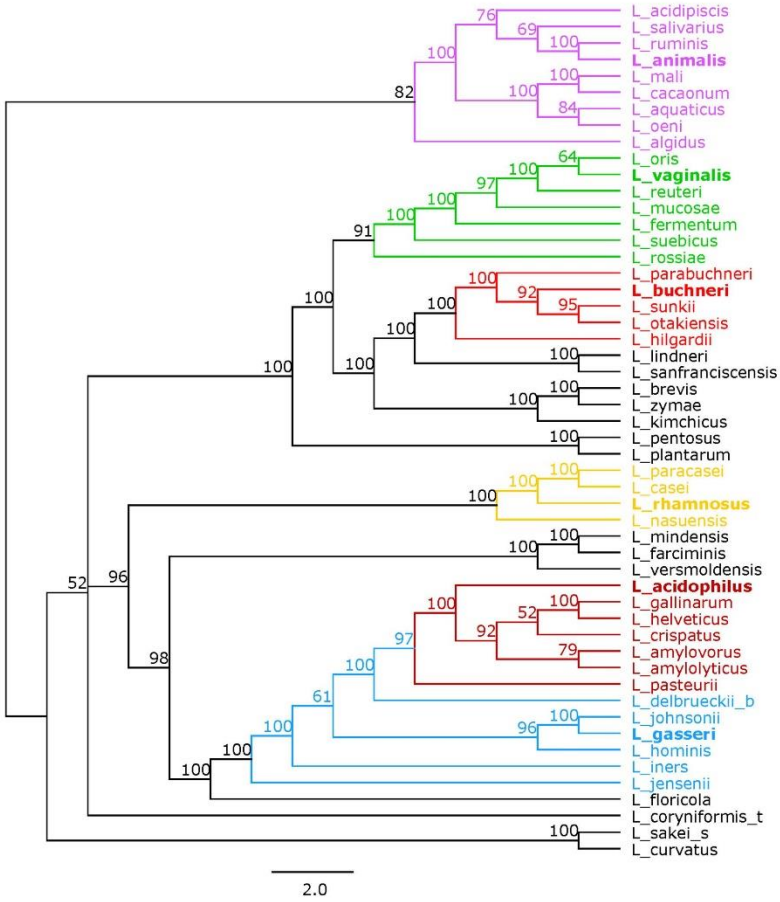


Fig. 5. Concatenated glycolysis tree. Tree based on the alignment of concatenated amino acid sequences of glycolysis enzymes using RaxML. Bootstrap values are recorded on the nodes. Groups are coloured as follows: the *L. animalis* group in purple, the *L. vaginalis* group in green, *L. buchneri* group in red, the *L. rhamnosus* group in yellow, the *L. acidophilus* group in maroon, and the *L. gasseri* group in blue. The representative species in each group is in bold. Species name follows the naming convention shown in Table 1.

for the concatenated tree ranged from 52 to 100. Nodes with bootstrap values equal to or greater than 70 numbered 43, a 59 % increase from that of the 16S rRNA tree. Overall, the concatenated tree correctly assigned the phylogenetic groups established from the 16S rRNA tree. In addition, the concatenated tree better discerned how the phylogenetic groups relate to one another, even within groups. This is supported by the higher bootstrap values (Fig. S3). Trees based on the individual glycolysis enzymes can be found in Figs S4–S13. The sum of branch lengths for each tree can be found in Table S1. A detailed comparative analysis of various trees structures revealed that overall there is high congruence in clustering both between and within the six established groups, though with various levels of discrimination across each protein sequence. Repeatedly, glycolysis-based trees provided more discriminatory power than the 16S rRNA tree.

G+C content analyses

Next, we looked at the G+C mol% and genomic drift of the glycolysis genes across the various species. Fig. 6 shows notched boxplots comparing the G+C mol% of each sequence set (the 16S rRNA sequence, the 10 genes and the concatenated sequences) in this study, compared to the genome-wide G+C mol%, ranked in increasing order. The

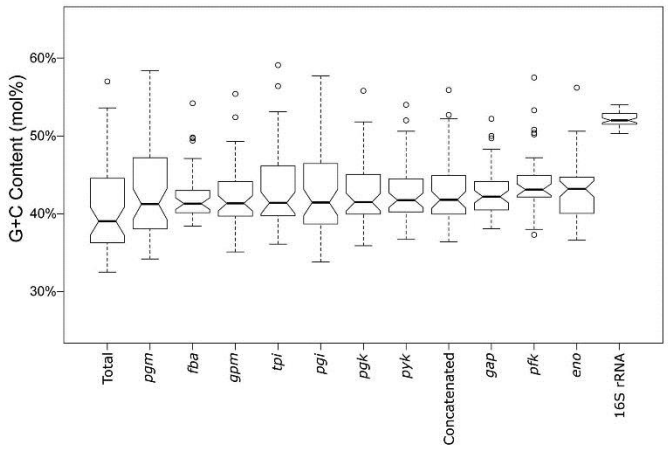


Fig. 6. G+C mol% analysis of *Lactobacillus* glycolysis genes. Depicted are notched boxplots of G+C mol% for each glycolysis gene, concatenated genes, 16S rRNA and total genome. Genes are placed in order of increasing median. If two notches do not overlap, it is an indication of strong evidence for differing medians.

G+C mol% of the *pgm* gene is closest to that of the total genome, while the 16S rRNA gene is the farthest. The notches are indicative of strong evidence that the medians differ when the notches do not overlap [51]. The 16S rRNA gene does not overlap with any other gene. In fact, a two-tailed *t*-test with a *P* value less than 0.001 (2.2×10^{-16}) revealed that the G+C mol% of the 16S rRNA sequence was statistically distinct from that of the total genome G+C mol %. This indicates that the 16S rRNA gene is not matching the pace of drift of the total genome with regards to G+C mol%. In contrast, all of the glycolysis genes, with the exception of *pfk* and *eno*, were not statistically different from the total genome G+C mol% (*P* value greater than 0.01), indicating that G+C mol% drift for glycolysis genes provide insights into the genome-wide G+C mol% drift. This further supports glycolytic sequences as intriguing candidates for both phylogenetic studies and representatives of genome-wide trends.

The genome sizes in this study ranged from 1.28 Mb (*Lactobacillus iners*) to 3.65 Mb (*Lactobacillus pentosus*), again reflecting the extensive genomic diversity within this genus. The total G+C mol% ranged from 32.50 % (*L. iners*) to 57.00 % (*Lactobacillus nasuensis*), which is intriguing given the general assumption that all lactobacilli are low G+C mol % organisms. Nevertheless, the mean G+C mol% was 40.70 %, consistent with *Lactobacillus* being generally perceived as low G+C mol% organisms. Splitting the species into high, medium and low categories, it becomes apparent that most species are trending towards the lower end of the spectrum, and away from the higher G+C mol% range (Fig. 7a). Some of the phylogenetic groups are closely clustered, such as the *L. acidophilus* group, *L. gasseri* group and the *L. rhamnosus* group, with the exception of *L. delbrueckii* subsp. *bulgaricus* (a dairy bacterium) and *L. nasuensis* (an aforementioned ex in G+C mol%). The *L. animalis* group and *L. buchneri* group are similarly clustered, albeit more loosely. These observations hold true when comparing the G+C mol% of all the individual genes in their respective genomes, perhaps reflecting a consistent and genome-wide pace of drift, rather than variable speeds of drift for each gene (Fig. 7b). Again, the 16S rRNA sequence has a much higher G+C mol% than most of the other studied genes, with the outlier *L. nasuensis* deviating from the consensus. The G+C mol% of the glycolysis genes within clusters are often times very close, as exemplified by the *L. acidophilus* group.

DISCUSSION

The genomic and functional attributes of *Lactobacillus* render it a pervasive genus, both in research and in industry. The benefits and uses of this diverse set of species are well-established and exhaustive, and yet, the list continues to grow. Many *Lactobacillus* strains are now considered to be health-promoting in the form of probiotics and are often found to be a part of a healthy microbiome [26]. They are also being engineered to promote healthy host–microbe interactions and deliver bioactive compounds such as vaccines [52]. As microbiome studies expand, we anticipate that the interest in *Lactobacillus* is set to increase, especially given their occurrence in several human-associated microbiomes, encompassing intestinal, vaginal, oral and skin communities [21]. Many studies have been published discussing the role of *Lactobacillus* in the microbiota, including research into the microbiota changes through

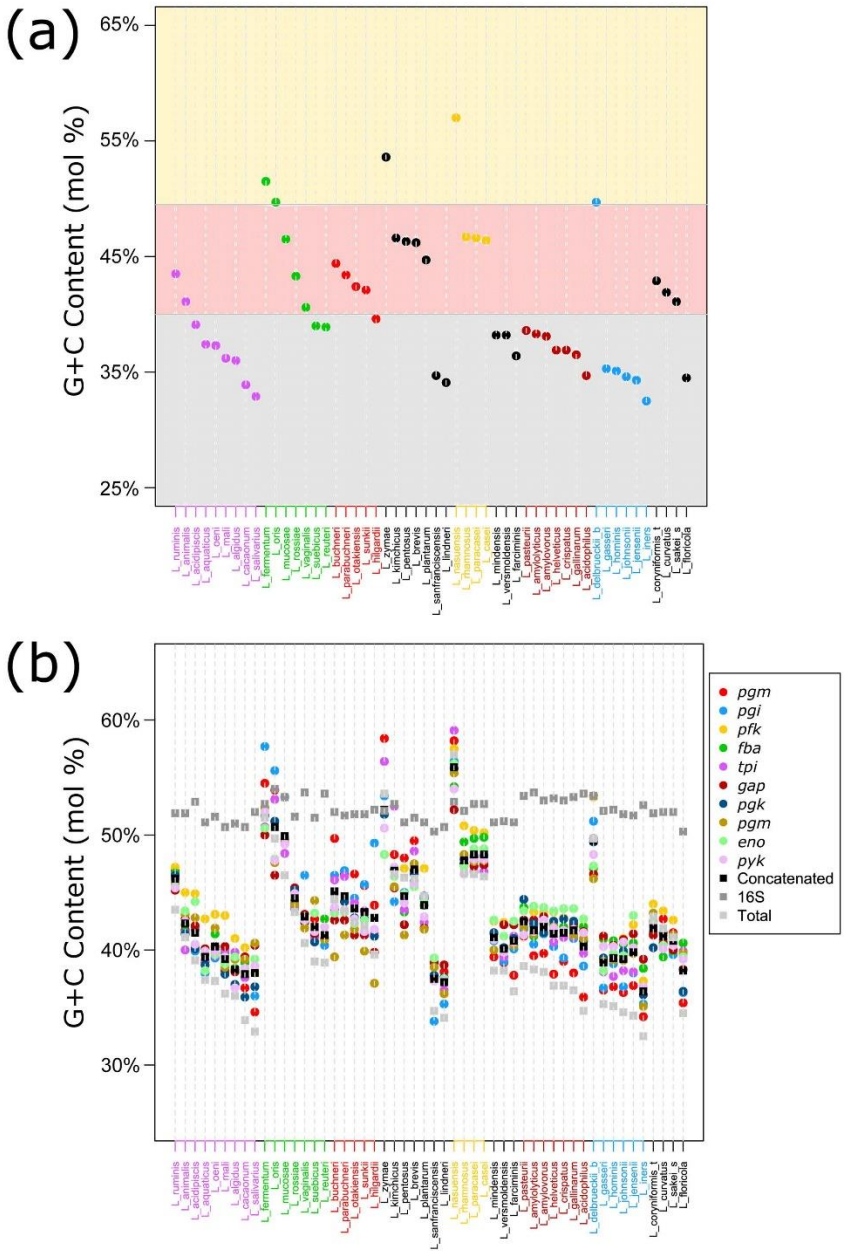


Fig. 7. G+C mol% analysis of *Lactobacillus* genomes. (a) shows the total G+C mol% for each species. Species are coloured according to their phylogenetic group. (b) shows the G+C mol% of the glycolysis genes, the concatenated glycolysis genes, the 16S rRNA and total G+C mol% for each species. Species are named according to Table 1.

disease, enhancing the microbiome as a form of treatment, and how the microbiome reacts to drugs [53–55]. The continuously expanding list of uses and studies just illustrates how important it is to accurately identify *Lactobacillus* species. While all species of *Lactobacillus* share some classical features of LAB organisms, notably their ability to produce lactic acid, the similarities between species are relatively few. In fact, even basic characteristics such as niche and isolation source can vary radically. Proper identification is an increasing concern especially when it comes to disease modelling in the human microbiome, as well as the formulation, tracking and efficacy of probiotic strains. Innovative techniques are continuously being developed and often use a combination of 16S rRNA with developing technologies, such as MALDI-TOF [56]. However, these tools are not broadly accessible and still rely partially on the sometimes unsatisfactory 16S rRNA. Here, we provide a practical alternative to the classical use of 16S rRNA sequencing.

In this paper, we applied the previously proposed methodology of using glycolysis sequences to perform phylogenetic studies [31] in the genus *Lactobacillus*. We demonstrated that this method is a practical and robust approach for *Lactobacillus*. Compared to the traditional 16S rRNA method, this approach was able to consistently identify phylogenetic groupings, with notably high-resolution between closely related species. While the 16S rRNA-based tree was able to identify the six phylogenetic groups, the concatenated tree was able to add more discrimination both between and within groups, evidenced by the higher bootstrap values in the glycolysis-based tree. Our grouping is consistent with a previous study using glycolysis sequences for phylogenetic analysis of *Lactobacillus* species [32]. Further analyses based on genomic content revealed clues as to why the glycolysis-based tree was better able to assign species.

First, looking at the organization of the genes in the genomes revealed two conserved operons in *Lactobacillus*, the *gap* operon and the *pfk* operon, with the remaining enzymes showing clear start and stop boundaries. This shared synteny emphasizes the importance of glycolysis gene conservation. Next, we looked at expression level. The glycolysis genes were consistently among the most highly expressed genes in the cell, with the *gap* gene always in the top three most abundant transcripts. These high expression levels indicate a great use and energy expenditure and, thus, arguably reflect the biological importance of this gene to the cell. Because of this importance, the glycolysis genes are much less likely to be subjected to loss. The operon structures and expression levels of the glycolysis genes are significant because a main criterion for selecting the 16S rRNA as a phylogenetic marker was its high conservation among species [57]. Next, we looked at how the glycolysis genes reflected genomic drift in terms of G+C mol%. First, it would appear that the genus is reaching a stabilizing point in its G+C mol% drift, though some species with high G+C mol% still have margin for extending the trend (*L. nasuensis, Lactobacillus zymae,* and *L. fermentum*). Next, we saw that the glycolysis gene G+C mol% was extremely close to that of the genome-wide G+C mol%, while the 16S rRNA was startlingly higher ($P<0.001$), underscoring the fact that the 16S rRNA is by all accounts much different than that of the total genome, whereas the majority of the glycolysis genes are significantly similar to the total genome G+C mol % (Fig. 6). This provides a possible explanation for the reason why the 16S rRNA analyses have been limited at a high-resolution level in *Lactobacillus* and why the glycolysis-based tree was able to reach a higher-resolution level. In fact, it has long been noted that 16S rRNA is unable to discriminate between species of lactobacilli due to its high similarity amongst them [58]. The individual glycolysis genes are much more similar to the genome as a whole (Fig. 6). Additionally, individual glycolysis genes are also able to accurately assign species to groups with a high resolution (Figs S4–S13). The *gap* gene is of particular note, due to its presence in an operon, consistently high expression, G+C mol% and ability to accurately define species groups. Overall, the glycolysis-based approach was able to provide a highe-resolution phylogeny for *Lactobacillus,* due in part to its conservation, expression and reflection of genomic drift.

Funding information
This study was supported by start-up funds from North Carolina State University (Raleigh, USA). K.B. is a recipient of a National Institute of Environmental Health Sciences (NIEHS) training grant.

Acknowledgements
The authors thank the funding sources for their support and the CRISPR lab for insightful conversations.

Conflicts of interest
The authors declare that there are no conflicts of interest.

References

1. Dandekar T, Snel B, Huynen M, Bork P. Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem Sci* 1998;23:324–328.
2. Karlin S, Campbell AM, Mrázek J. Comparative DNA analysis across diverse genomes. *Annu Rev Genet* 1998;32:185–225.
3. Karlin S, Mrázek J, Campbell A, Kaiser D. Characterizations of highly expressed genes of four fast-growing bacteria. *J Bacteriol* 2001;183:5025–5040.
4. Boekhorst J, Siezen RJ, Zwahlen MC, Vilanova D, Pridmore RD *et al.* The complete genomes of *Lactobacillus plantarum* and *Lactobacillus johnsonii* reveal extensive differences in chromosome organization and gene content. *Microbiology* 2004;150:3601–3611.
5. Hoffmann M, Zhao S, Pettengill J, Luo Y, Monday SR *et al.* Comparative genomic analysis and virulence differences in closely related *Salmonella enterica* serotype eidelberg isolates from humans, retail meats, and animals. *Genome Biol Evol* 2014;6:1046–1068.
6. Losada PM, Tümmler B. SNP synteny analysis of *Staphylococcus aureus* and *Pseudomonas aeruginosa* population genomics. *FEMS Microbiol Lett* 2016;363:fnw229-fnw.
7. Makarova K, Slesarev A, Wolf Y, Sorokin A, Mirkin B *et al.* Comparative genomics of the lactic acid bacteria. *Proc Natl Acad Sci USA* 2006;103:15611–15616.
8. Claesson MJ, van Sinderen D, O'Toole PW. *Lactobacillus* phylogenomics-towards a reclassification of the genus. *Int J Syst Evol Microbiol* 2008;58:2945–2954.
9. Felis GE, Dellaglio F, Mizzi L, Torriani S. Comparative sequence analysis of a *recA* gene fragment brings new evidence for a

change in the taxonomy of the *Lactobacillus casei* group. *Int J Syst Evol Microbiol* 2001;51:2113–2117.

10. **Milani C, Turroni F, Duranti S, Lugli GA, Mancabelli L *et al.*** Genomics of the genus *Bifidobacterium* reveals species-specific adaptation to the glycan-rich gut environment. *Appl Environ Microbiol* 2016;82:980–991.
11. **Milani C, Lugli GA, Turroni F, Mancabelli L, Duranti S *et al.*** Evaluation of bifidobacterial community composition in the human gut by means of a targeted amplicon sequencing (ITS) protocol. *FEMS Microbiol Ecol* 2014;90:493–503.
12. **Baker GC, Smith JJ, Cowan DA.** Review and re-analysis of domain-specific 16S primers. *J Microbiol Methods* 2003;55:541–555.
13. **Clarridge JE.** Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. *Clin Microbiol Rev* 2004;17:840–862.
14. **de La Cuesta-Zuluaga J, Escobar JS.** Considerations for optimizing microbiome analysis using a marker gene. *Front Nutr* 2016;3:26.
15. **Salvetti E, Torriani S, Felis GE.** The genus *Lactobacillus*: a taxonomic update. *Probiotics Antimicrob Proteins* 2012;4:217–226.
16. **Sun Z, Harris HM, McCann A, Guo C, Argimón S *et al.*** Expanding the biotechnology potential of lactobacilli through comparative genomics of 213 strains and associated genera. *Nat Commun* 2015;6:8322.
17. **Bernardeau M, Vernoux JP, Henri-Dubernet S, Guéguen M.** Safety assessment of dairy microorganisms: the *Lactobacillus* genus. *Int J Food Microbiol* 2008;126:278–285.
18. **Hill C, Guarner F, Reid G, Gibson GR, Merenstein DJ *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;11:506–514.
19. **Saxelin M.** Probiotic formulations and applications, the current probiotics market, and changes in the marketplace: a European perspective. *Clin Infect Dis* 2008;46:S76–S79.
20. **Cho I, Blaser MJ.** The human microbiome: at the interface of health and disease. *Nat Rev Genet* 2012;13:260–270.
21. **Human Microbiome Project Consortium.** Structure, function and diversity of the healthy human microbiome. *Nature* 2012;486:207–214.
22. **Conlon M, Bird A.** The impact of diet and lifestyle on gut microbiota and human health. *Nutrients* 2015;7:17–44.
23. **Li X, Wang N, Yin B, Fang D, Jiang T *et al.*** Effects of *Lactobacillus plantarum* CCFM0236 on hyperglycaemia and insulin resistance in high-fat and streptozotocin-induced type 2 diabetic mice. *J Appl Microbiol* 2016;121:1727–1736.
24. **Feng XB, Jiang J, Li M, Wang G, You JW *et al.*** Role of intestinal flora imbalance in pathogenesis of pouchitis. *Asian Pac J Trop Med* 2016;9:786–790.
25. **Hooper LV, Midtvedt T, Gordon JI.** How host-microbial interactions shape the nutrient environment of the mammalian intestine. *Annu Rev Nutr* 2002;22:283–307.
26. **Gerritsen J, Smidt H, Rijkers GT, de Vos WM.** Intestinal microbiota in human health and disease: the impact of probiotics. *Genes Nutr* 2011;6:209–240.
27. **Shreiner AB, Kao JY, Young VB.** The gut microbiome in health and in disease. *Curr Opin Gastroenterol* 2015;31:69–75.
28. **Okai S, Usui F, Yokota S, Hori-I Y, Hasegawa M *et al.*** High-affinity monoclonal IgA regulates gut microbiota and prevents colitis in mice. *Nat Microbiol* 2016;1:16103.
29. **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;82:6091–6101.
30. **Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ *et al.*** A new view of the tree of life. *Nat Microbiol* 2016;1:16048.
31. **Brandt K, Barrangou R.** Phylogenetic analysis of the *Bifidobacterium* genus using glycolysis enzyme sequences. *Front Microbiol* 2016;7:00657.
32. **Salvetti E, Fondi M, Fani R, Torriani S, Felis GE.** Evolution of lactic acid bacteria in the order *Lactobacillales* as depicted by analysis of glycolysis and pentose phosphate pathways. *Syst Appl Microbiol* 2013;36:291–305.
33. **Fothergill-Gilmore LA.** The evolution of the glycolytic pathway. *Trends Biochem Sci* 1986;11:47–51.
34. **Fothergill-Gilmore LA, Michels PA.** Evolution of glycolysis. *Prog Biophys Mol Biol* 1993;59:105–235.
35. **Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M *et al.*** Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 2012;28:1647–1649.
36. **Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ.** Basic local alignment search tool. *J Mol Biol* 1990;215:403–410.
37. **Contreras-Moreira B, Vinuesa P.** GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis. *Appl Environ Microbiol* 2013;79:7696–7701.
38. **Gasteiger E, Gattiker A, Hoogland C, Ivanyi I, Appel RD *et al.*** ExPASy: the proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res* 2003;31:3784–3788.
39. **Johnson BR, Hymes J, Sanozky-Dawes R, Henriksen ED, Barrangou R *et al.*** Conserved S-layer-associated proteins revealed by exoproteomic survey of S-layer-forming lactobacilli. *Appl Environ Microbiol* 2016;82:134–145.
40. **Katoh K, Misawa K, Kuma K, Miyata T.** MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* 2002;30:3059–3066.
41. **Edgar RC.** MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 2004;32:1792–1797.
42. **Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA *et al.*** Clustal W and Clustal X version 2.0. *Bioinformatics* 2007;23:2947–2948.
43. **Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T.** trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 2009;25:1972–1973.
44. **Stamatakis A.** RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 2014;30:1312–1313.
45. **Core Team R.** *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2015.
46. **Canchaya C, Claesson MJ, Fitzgerald GF, van Sinderen D, O'Toole PW.** Diversity of the genus *Lactobacillus* revealed by comparative genomics of five species. *Microbiology* 2006;152:3185–3196.
47. **Heinl S, Wibberg D, Eikmeyer F, Szczepanowski R, Blom J *et al.*** Insights into the completely annotated genome of *Lactobacillus buchneri* CD034, a strain isolated from stable grass silage. *J Biotechnol* 2012;161:153–166.
48. **Cárdenas N, Laiño JE, Delgado S, Jiménez E, Juárez del Valle M *et al.*** Relationships between the genome and some phenotypical properties of *Lactobacillus fermentum* CECT 5716, a probiotic strain isolated from human milk. *Appl Microbiol Biotechnol* 2015;99:4343–4353.
49. **Arsköld E, Lohmeier-Vogel E, Cao R, Roos S, Rådström P *et al.*** Phosphoketolase pathway dominates in *Lactobacillus reuteri* ATCC 55730 containing dual pathways for glycolysis. *J Bacteriol* 2008;190:206–212.
50. **Broadbent JR, Hughes JE, Welker DL, Tompkins TA, Steele JL.** Complete genome sequence for *Lactobacillus helveticus* CNRZ 32, an industrial cheese starter and cheese flavor adjunct. *Genome Announc* 2013;1:e00590-13.

51. **Chambers JM.** Notched box plots. *GraphicalMethods for Data Analysis*. Belmont, CA: Wadsworth International Group; 1983. pp. 60–63.
52. **Seegers JF.** Lactobacilli as live vaccine delivery vectors: progress and prospects. *Trends Biotechnol* 2002;20:508–515.
53. **Bhat M, Arendt BM, Bhat V, Renner EL, Humar A *et al.*** Implication of the intestinal microbiome in complications of cirrhosis. *World J Hepatol* 2016;8:1128–1136.
54. **Bull-Otterson L, Feng W, Kirpich I, Wang Y, Qin X *et al.*** Metagenomic analyses of alcohol induced pathogenic alterations in the intestinal microbiome and the effect of *Lactobacillus rhamnosus* GG treatment. *PLoS One* 2013;8:e53028.
55. **Shin CM, Kim N, Kim YS, Nam RH, Park JH *et al.*** Impact of Long-term proton pump inhibitor therapy on gut microbiota in F344 rats: pilot study. *Gut Liver* 2016;10:896–901.
56. **Foschi C, Laghi L, Parolin C, Giordani B, Compri M *et al.*** Novel approaches for the taxonomic and metabolic characterization of lactobacilli: integration of 16S rRNA gene sequencing with MALDI-TOF MS and 1H-NMR. *PLoS One* 2017;12:e0172483.
57. **Eisen JA.** The RecA protein as a model molecule for molecular systematic studies of bacteria: comparison of trees of RecAs and 16S rRNAs from the same species. *J Mol Evol* 1995;41:1105–1123.
58. **Fox GE, Wisotzkey JD, Jurtshuk P.** How close is close: 16S rRNA sequence identity may not be sufficient to guarantee species identity. *Int J Syst Bacteriol* 1992;42:166–170.

APPENDIX C: CHARACTERIZING THE ACTIVITY OF ABUNDANT, DIVERSE AND ACTIVE CRISPR-CAS SYSTEMS IN LACTOBACILLI

C.1. CONTRIBUTION TO WORK

The following is a reprint of Crawley et al. “Characterizing the activity of abundant, diverse and active CRISPR-Cas systems in lactobacilli”, published in *Scientific Reports* 8: 11544. Katelyn Brandt is an author on this publication. She helped carry out experiments and analyze data for this manuscript. She was involved in the editing process.

www.nature.com/scientificreports

SCIENTIFIC REPORTS

OPEN

Characterizing the activity of abundant, diverse and active CRISPR-Cas systems in lactobacilli

Received: 5 April 2018
Accepted: 9 July 2018
Published online: 01 August 2018

Alexandra B. Crawley[1], Emily D. Henriksen[2], Emily Stout[2], Katelyn Brandt[1] & Rodolphe Barrangou[1,2]

CRISPR-Cas systems provide immunity against phages and plasmids in bacteria and archaea. Despite the popularity of CRISPR-Cas9 based genome editing, few endogenous systems have been characterized to date. Here, we sampled 1,262 publically available lactobacilli genomes found them to be enriched with CRISPR-Cas adaptive immunity. While CRISPR-Cas is ubiquitous in some *Lactobacillus* species, CRISPR-Cas content varies at the strain level in most *Lactobacillus* species. We identified that Type II is the most abundant type across the genus, with II-A being the most dominant sub-type. We found that many Type II-A systems are actively transcribed, and encode spacers that efficiently provide resistance against plasmid uptake. Analysis of various CRISPR transcripts revealed that guide sequences are highly diverse in terms of crRNA and tracrRNA length and structure. Interference assays revealed highly diverse target PAM sequences. Lastly, we show that these systems can be readily repurposed for self-targeting by expressing an engineered single guide RNA. Our results reveal that Type II-A systems in lactobacilli are naturally active in their native host in terms of expression and efficiently targeting invasive and genomic DNA. Together, these systems increase the possible Cas9 targeting space and provide multiplexing potential in native hosts and heterologous genome editing purpose.

CRISPR-Cas (Clustered regularly interspaced short palindromic repeats and CRISPR associated genes) systems have been shown to protect bacteria and archaea from invasive mobile genetic elements (MGEs)[1–3]. These systems are identified by a genetic locus with a CRISPR repeat-spacer array and *cas* genes[4,5]. During *adaptation*, the first stage of CRISPR immunity, foreign DNA sequences from MGEs are copied and pasted iteratively into the array as unique spacer sequences flanked by conserved repeats[4,6–9]. The second stage of CRISPR immunity, *expression*, leads to the biogenesis of individual small crRNAs (CRISPR RNAs), that each contain a single partial spacer and partial repeat; these RNAs act as a guide molecule to direct the Cas proteins to a complementary foreign nucleic acid target[2,10,11]. Some specific subtypes of CRISPR-Cas systems, including Type II-A, require a second RNA molecule, called the tracrRNA (trans-acting CRISPR RNA), to generate the individual crRNAs capable of guiding the signature Cas9 endonuclease[11–13]. The final stage of CRISPR immunity, *interference*, is the targeting and cleavage of foreign DNA when it is reintroduced into the cell[4,11,14]. Cas proteins are able to distinguish self from non-self targets through the occurrence of a PAM (protospacer adjacent motif) on the foreign target that is not present when then spacer is stored in the repeat-spacer array[7,15–17].

CRISPR is fairly common in bacteria, occurring in just under half of all bacterial species sequenced to date in publically available databases[18,19]. Though the stages of CRISPR-Cas immunity are universal, there are two main classes of systems that can be further broken down into six types and 23 subtypes that utilize different Cas proteins and crRNA structures[19–21]. Though Type II-A systems can only be identified in 5% of bacteria, they are arguably the most used, since the molecular machinery from this subtype can be repurposed to generate Cas9-based genome editing tools[4,5,22–24]. Despite being relatively rare, Type II-A systems are known to occur preferentially in firmicutes, like lactic acid bacteria, occurring in almost 30% of all lactobacilli[19,25,26].

Interestingly, the majority of our knowledge of CRISPR activity in their native host has been limited to a few model systems, namely *Streptococcus pyogenes* (Type II-A)[23], *Streptococcus thermophilus* (two Type II-As, one Type I-E, and one III-A)[1,6,11,13–15,27,28], *Sulfolobus islandicus* (Type III-B)[29,30], *Pseudomonas aeruginosa* (Type IE and IIA)[31,32], and *Escherichia coli* (Type I-E)[2,8,33]. Unfortunately, some CRISPR systems in *E. coli* and other

[1]North Carolina State University Functional Genomics, Raleigh, NC, 27695, USA. [2]North Carolina State University Department of Food, Bioprocessing and Nutrition Sciences, Raleigh, NC, 27695, USA. Correspondence and requests for materials should be addressed to R.B. (email: rbarran@ncsu.edu)

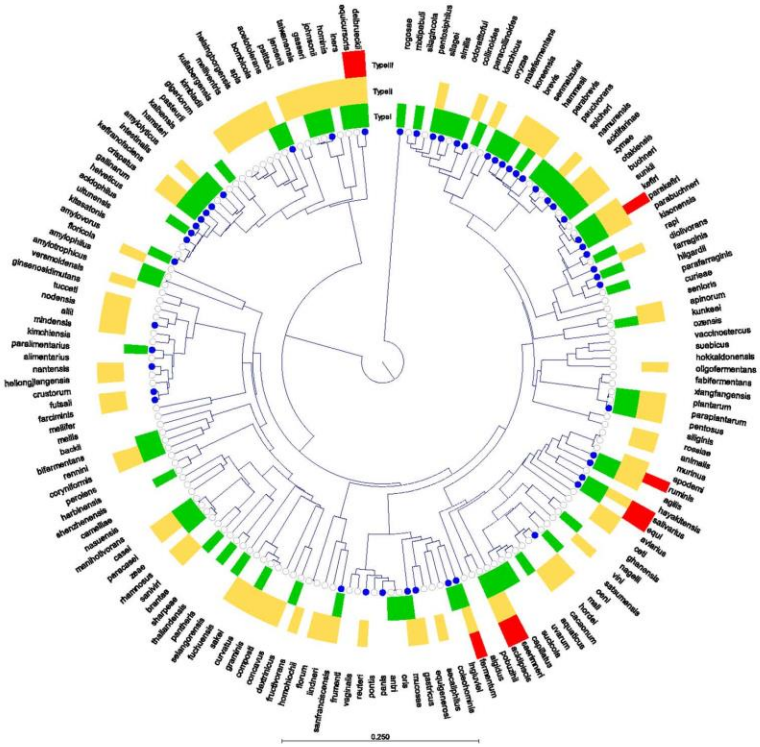


Figure 1. Occurrence of CRISPR-Cas systems in lactobacilli. The core genome of lactobacilli was identified by Sun *et al.* 2015. This tree displays the phylogenetic relationship of one representative genome from each of the 171 lactobacilli species used in this study based on the core genome. The metadata rings display the presence of CRISPR-Cas systems in any strain from that species. Type I systems are shaded in green, Type II in yellow, and Type III in red. The nodes are colored blue if a Type V-U putative Cas protein was identified in that species. A lack of color demonstrates a lack of CRISPR-Cas systems. The species is listed in the outer ring.

organisms do not appear to be natively active and most work must be performed *in vitro* or with heterologous CRISPR machinery, leaving our knowledge of native activity in the original somewhat shallow.

With relatively little known about the native activity of many different endogenous systems, we first identified a large selection of uncharacterized CRISPR-Cas systems. To fully characterize the Type II systems, we then predicted all system components for each system, including the PAM, tracrRNA, and crRNA. Next, we determined CRISPR interference to assess whether each individual system was active through investigating acquisition, expression, and interference. Finally, we used one model system, *Lactobacillus gasseri*, to investigate a novel species of tracrRNA to develop biotechnological CRISPR-Cas9 based genetic engineering tools using the native CRISPR components.

Results

Lactobacilli encode complete, diverse, and active CRISPR-Cas systems. Despite the growing popularity of CRISPR-Cas, only a handful of systems have been characterized to date. We set out to understand the native variability in occurrence and activity of CRISPR using endogenous systems occurring in lactobacilli, as it has been published that they are enriched in CRISPR-Cas systems 6-fold compared to the canonical rate of occurrence for bacteria (5% of all bacteria vs. 30% of all lactobacilli)[19,25]. Our *in silico* searches of 1,262 strains of lactobacilli, accounting for 171 different *Lactobacillus* species and closely related lactic acid bacteria, confirmed diversity across both classes of systems, focusing on Types I, II and III (Figs 1, 2, 3 Panel A, Table S1). We were unable to detect Type IV, V or VI CRISPR-Cas systems in lactobacilli, though several V-U proteins were detected in our genomes (Table S1). Noteworthy, these results are consistent with previous studies documenting that Types I, II and III are most dominant and widespread in nature, though the size of the Type I arrays are smaller than the reported average array size for this type[34]. As these V-U systems are still putatively uncharacterized, we have not included them in determining the rate of occurrence of CRISPR-Cas systems in lactobacilli[20,35].

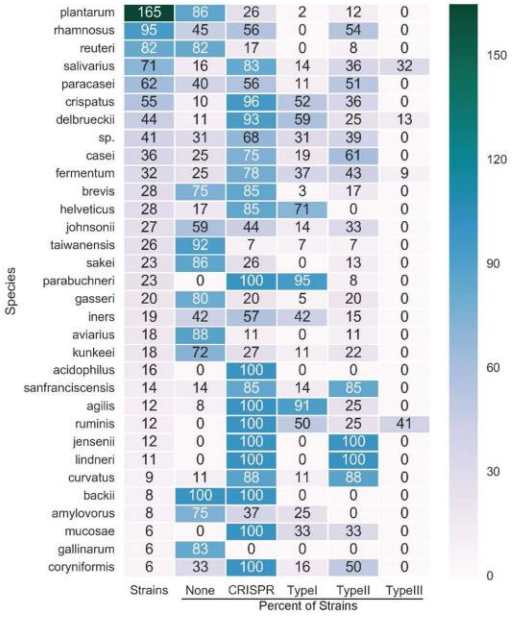


Figure 2. Strain-specific distribution in lactobacilli. For species where there were at least 6 representative genomes, the rate of occurrence of CRISPR repeats and complete systems is displayed. The number of strains investigated is in the first column. The remaining columns list the percent of strains containing: No *cas* genes, CRISPR repeats, Type I systems, Type II systems, or Type III systems.

We detected CRISPR repeats in 59.7% (753 of 1,262) of lactobacilli genomes and most often detected a single CRISPR-Cas locus in a genome (Fig. 2). Two strains of lactobacilli contained Type I, II and III systems in the same genome: *Lactobacillus fermentum* (strains NB-22 and MTCC 8711), *Lactobacillus equicursoris* (strain 66c) (Fig. 1). Multiple systems were often detected in the same genome; occasionally this corresponded to a single subtype with two distinct sets of *cas* gene and CRISPR arrays, but most often corresponded to a distinct Type I-E and II-A system in the same genome. The subtype I-E was the predominant Type I system identified in lactobacilli, accounting for 210 of the 268 Type I systems identified (Figs 1, 2). Likewise, the II-A subtype was the predominant Type II system, accounting for 290 of the 393 Type II systems identified. CRISPR-Cas systems are ubiquitous in 14 of the 171 species (*Lactobacillus parabuchneri*, *Lactobacillus jensenii*, *Lactobacillus ruminis*, *Lactobacillus agilis*, *Lactobacillus linderi*, *Lactobacillus mucosae*, *Lactobacillus pentosus*, *Lactobacillus farciminis*, *Lactobacillus kefiranofaciens*, *Lactobacillus animalis*, *Lactobacillus kefiri*, *Lactobacillus buchneri*, *Lactobacillus parakefiri*, *Lactobacillus equicursoris*) analyzed here and are rarely found in eight species (*Lactobacillus plantarum*, *Lactobacillus reuteri*, *Lactobacillus taiwanensis*, *Lactobacillus sakei*, *Lactobacillus gasseri*, *Lactobacillus avarius*, *Lactobacillus gallinarum*, *Lactobacillus paralimentarius*). There are three species that always contain CRISPR repeats but are always devoid of *cas* genes (*Lactobacillus acidophilus*, *Lactobacillus backii*, *Lactobacillus crustorum*); and conversely, one species, *Lactobacillus paracollinoides*, that always contains *cas* genes, but never contains CRISPR repeats.

The most notable CRISPR trend in lactobacilli is the enrichment of Type II systems, expanding the known Cas9 space to novel proteins, including short II-A Cas9s, long II-A Cas9s, and II-C Cas9s (Fig. 3). The Cas9s from lactobacilli contain an entire clade of Cas9s that is divergent from the canonical Cas9s, mainly *S. pyogenes* (Spy), *S. thermophilus* CRISPRs 1 and 3 (Sth1, Sth3, respectively), *Staphylococcus aureus* (Sau), and *Neisseria meningitides* (Nme) (Fig. 3 Panel B)[1,12,14,36,37]. Though the lactobacilli Cas9 proteins contain the same motifs as the canonical Cas9s, they are highly dissimilar, sharing sometimes as low as 40% similarity at the protein coding level with Spy, Sth1, Sth3, Sau or Nme. Even within the clade of lactobacilli-specific Cas9s, there is great diversity in protein sequences, sometimes as low as 60% similarity to other lactobacilli Cas9s.

In CRISPR biology, *cas1* is currently considered the universal gene as it is found in most CRISPR-Cas systems and drives the acquisition stage of immunity[9,19]. Despite *cas1* being the universal *cas* gene, *cas2* was the most conserved gene amongst all *cas* genes identified (Fig. 3 Panel D). In addition to *cas* conservation and divergence, we observed evidence of maintenance and activity in the CRISPR arrays. The arrays contained between 2 and 135 spacers, with the median array containing 20 spacers. On average, the Type I systems contained the longest CRISPR arrays (27 spacers Type I, 19.5 spacers Type II, 9 spacers undefined) (Table S1). When arrays are inactivated, they can accumulate mutations in repeats and show evidence of degeneration through inconsistent

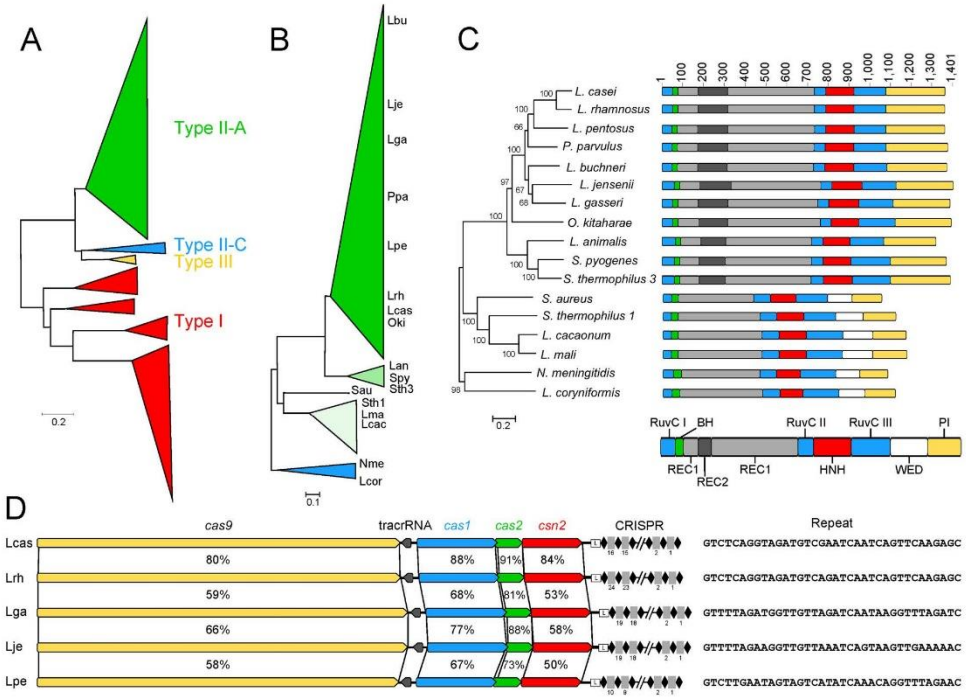


Figure 3. Diversity of CRISPR-Cas systems in lactobacilli. (**A**) The total diversity of CRISPR-Cas systems in lactobacilli was determined through the phylogenetic distribution of the Cas1 proteins. The ML tree is rooted on the Type I to Type II split. (**B**) Diversity of Type II systems was determined through alignment of the Cas9 protein. This tree is rooted on the outgroup II-C. (**C**) Cas9 protein domains were mapped from known protein crystal structures. The long II-A Cas9s – any Cas9 longer than 1250 amino acids – was mapped to the Streptococcus pyogenes Cas9. The short II-A and II-C Cas9s – less than 1,250 amino acids – were mapped to the Staphylococcus aureus Cas9. (**D**) Comparisons of entire CRISPR loci revealed high amounts of diversity in all Cas proteins (yellow arrow – Cas9, blue arrow – Cas1, green arrow - Cas2, red arrow - Csn2), tracrRNA sequence (dark grey arrow), leader sequence (box L), array length (black diamonds - repeats, grey rectangles - spacers), and CRISPR repeat sequence. The percent identities for the Cas proteins compare the protein above and below the percentage.

length of repeats and spacers[14,26]; in contrast, CRISPR repeats in lactobacilli remain intact in terms of length and sequence across the entire array suggesting they are still actively maintained and functional.

crRNA biogenesis and active transcription of CRISPR RNAs. Expression is the second stage of CRISPR interference. To determine the activity of CRISPR expression in lactobacilli, we investigated the crRNA transcripts via small RNA-Sequencing. We were able to determine that crRNAs were some of the most highly transcribed small RNAs in cells, even reaching 199,539 transcripts of a single crRNA in *Lactobacillus pentosus* (2.5%, in 8,000,000 total reads), making that crRNA the 4th most highly expressed small RNA in the cell (Figs 4, S1, S2). When visualizing the crRNA transcripts, we found it very striking to observe the sharp boundaries of processed crRNAs; this demonstrates the cleavage of pre-crRNAs to individual crRNAs is precise and consistent. As seen with other organisms, the length of processed crRNAs was conserved within an array but differed between systems. Interestingly, the spacer portion of the crRNA was consistently 20 nucleotides long in all Type II-A crRNAs (Figs 4, S1, S2). Interestingly, the repeat portion of the crRNA was unique to each CRISPR system, ranging from 13 nucleotides in *Oenococcus kitaharae* to 25 nucleotides in *L. casei*. The II-C crRNAs in *Lactobacillus coryniformis* were comprised of 17 nucleotides in the spacer portion and 22 nucleotides in the repeat portion.

We observed an interesting trend in the expression pattern of the first repeat in the CRISPR array. The 5′ end of the leader RNA, ldrRNA, as we propose to name it, contains 20 nucleotides of the promoter-like leader sequence (Figs 4, S3). The length of the leader transcribed in the ldrRNA is the same length of spacer sequence transcribed in the downstream crRNAs and the length of the repeat transcribed in the ldrRNA is also the same length of repeat transcribed in crRNAs. This RNA was first seen in *S. thermophilus* by Wei *et al.*[28], but this finding has not been investigated in other organisms.

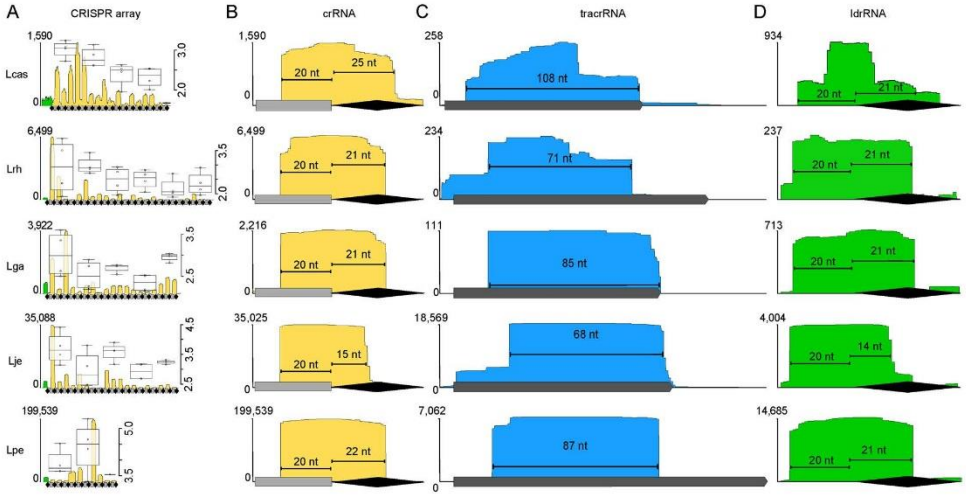


Figure 4. Expression of CRISPR transcripts. (**A**) Expression profile of the entire CRISPR-Cas array reveals the transcription levels of the ldrRNA (green) and crRNAs (yellow) for *L. casei* (Lcas), *L. rhamnosus* (Lrh), *L. gasseri* (Lga), *L. jensenii* (Lje), and *L. pentosus* (Lpe). The left y-axis shows the sequencing coverage depth at each position; the right y-axis shows the log transformed coverage depth for the box plots. Over laid box plots show the distribution of transcription level in four crRNA increments; the transcript level for each individual crRNA is marked by open circles in the box plots length (black diamonds - repeats, grey rectangles - spacers). (**B**) The boundaries of a highly expressed crRNA is shown for each organism. The length of the spacer portions of the crRNA (grey rectangle), is strictly conserved at 20 nt for all crRNAs. The length of the repeat portion of the crRNAs (black diamond), varies between organisms. (**C**) The tracrRNA transcriptional profile reveals the boundaries of tracrRNA processing. The gray bar on the x-axis is the *in silico* prediction for the tracrRNA. (**D**) The transcriptional profile for the ldrRNA for each organism closely matches the crRNA transcriptional profile.

The tracrRNAs were predicted *in silico* according to Briner *et al.*[12], looking for the 5 modules found in the tracrRNA: upper stem, bulge, lower stem, nexus and ending with one to three terminal hairpins; one of which being a GC-rich transcription terminating hairpin (Fig. 4). The expression boundaries of the tracrRNAs are clearly defined, further demonstrating the expression stage of CRISPR-Cas immunity is active. We found that our predictions for the tracrRNAs were often too conservative and the transcription terminating hairpins are often not a part of the final tracrRNAs (Figs 4, S4). As a consequence, in lactobacilli, there is most often only a single terminal hairpin, though two or three were typically predicted (Figs 5, S4, S5). The RNaseIII processing sites are best determined via boundary mapping, as they are often unpredictable[38,39]. All but two of the tracrRNAs we looked at contained the bulged stem loop nexus typical of and unique to lactobacilli. Among the tracrRNAs investigated here, five groups are completely unique and likely orthogonal to other systems known to date based on the predicted structures of the sgRNAs, the Cas9 sequences, and their predicted PAM targets.

Interference stage is active against foreign DNA. The final stage of CRISPR interference is sequence-specific targeting and cleavage of complementary foreign DNA upon introduction to the cell. To determine whether the CRISPR systems were active, we first needed to determine what sequences these systems natively target. The protospacers corresponding to the spacer sequences already stored in CRISPR arrays revealed these systems provide immunity against phages, plasmids, and prophages (Fig. 5, Table S2). In particular, *L. jensenii* is under strong predatory pressure from phage LV-1 as 10 different spacers target separate sequences on the same phage (Table S2). Beyond predatory phages, many spacers targeted prophage and mobile elements such as transposons, suggesting that beyond immunity, CRISPR-Cas systems might be active in maintaining genome homeostasis and helping control the flow of horizontal gene transfer.

The PAM sequences were predicted using the flanking regions of the protospacers. To test whether Cas9 was able to recognize these predicted PAMs, we cloned a native spacer sequence from each endogenous array into a plasmid and included the predicted PAM and tested several mutated variants. The plasmid interference assay was able to determine whether Cas9 is able to recognize the PAM sequence provided, and also demonstrated that the system was active through the ability of Cas9 to target and cleave the foreign DNA and preventing the uptake of plasmid DNA (Figs 6, S6).

We were able to demonstrate that five different CRISPR-Cas systems have endogenous interference activity, with a range of interference efficiencies. One phenomenon we observed was flexibility in PAM targeting by Cas9, which was seen most prominently in *L. gasseri* (Figs 6, S6). The PAM 5′-cTAAC-3′ performed perfect interference and did not have any escapees, while the PAM 5′-ccAAC-3′ allowed one log of transformants to survive and

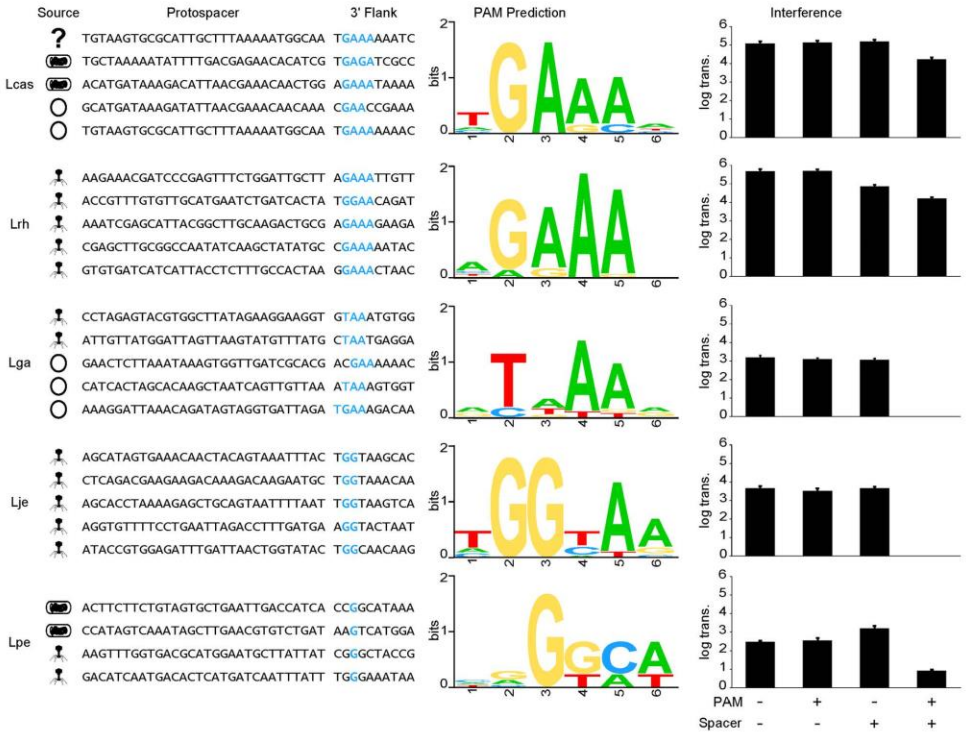


Figure 6. PAM prediction and validation. Representative spacers from each organism are displayed with their protospacer targets. The source of each protospacer sequence was determined to be either phage (phage), plasmid (circle), chromosomal origin (bacterial cell), or unknown metagenomics origin (?). Ten nucleotides from the 3′ flank for each protospacer was used to predict the PAM sequence (blue text) for each Cas9. All of the 3′ flanks for each protospacer were aligned to generate a Weblogo for each PAM prediction. Finally, plasmid interference assays were used to test the ability of each Cas9 to recognize and cleave plasmid DNA. Constructs containing PAMs and protospacers homologous to native spacers in the host genome were used to determine plasmid interference. The log number of transformants is plotted to show the efficiency of the native CRISPR system in eliminating each construct; error bars are based on three independent replicates. The most effective PAM for each organism is shown: Lcas 5′-tGAAAA-3′, Lrh 5′-aGAAA-3′, Lga 5′-cTAACc-3′, Lje 5′-tGGc-3′, and Lpe 5′-gTTAAT-3′.

The biological function of the ldrRNA is unknown[28], and this is the first broad investigation into the expression patterns of the ldrRNA in multiple II-A systems. We hypothesize this RNA might provide a ruler-anchor mechanism for determining how crRNAs are processed due to its strict size conservation that matches the crRNAs processing boundaries. Additionally, the similarity between the ldrRNA and crRNA may suggest an alternative role for the ldrRNA priming Cas9 for adaptation or crRNA loading. We observed crRNAs expression across the array consistent with previous reports of expression trends in that expression is highest at the 5′ end of array[28,50]. The first crRNAs may be more stable because they are transcribed first, making them more available for tracrRNA-binding and protection by Cas9 and providing immunity against the most recently seen MGEs (Figs 4, S1).

Lactobacilli Cas9 are known to utilize tracrRNAs with unique sequences and structures; the diversity in these RNAs suggest each individual RNA is likely not compatible with Cas9s from other systems[12,36]. The lock-and-key specificity of tracrRNAs with Cas9s opens the door for multiplexing potential and concurrent use of different systems simultaneously for genome editing. Additionally, through understanding the native processing sites on tracrRNAs and crRNAs, minimal guide sequences can be used to develop single guide RNAs (sgRNAs) from these sequences (Figs 5, S5).

We were able to demonstrate that five different CRISPR-Cas systems have endogenous interference activity, with a range of interference efficiencies. The differences in PAM sequences suggest there is an entire spectrum of endogenous PAMs that can be used with different Cas9s. The range in ability to target and cleave could be a result of an imperfect PAM, differences in crRNA expression activity, differences in the background ability of the

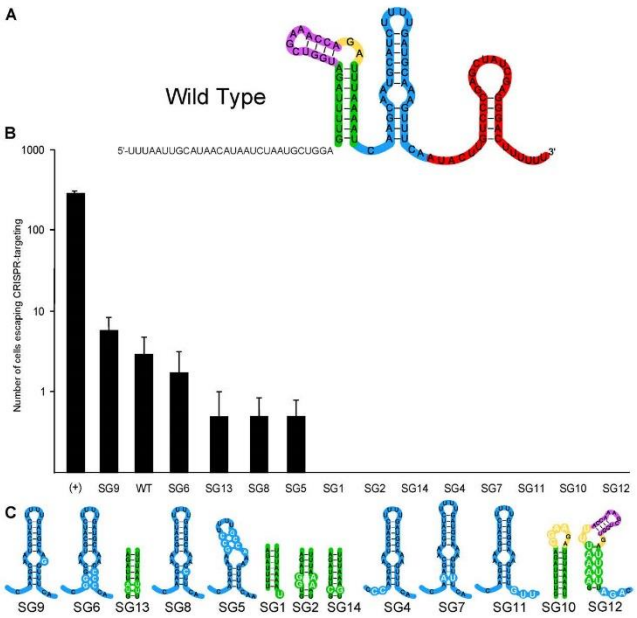


Figure 7. Self-targeting assays. (**A**) An Lga single guide RNA was designed to target the chromosome. (**B**) The ability of each guide to target and cleave the chromosome determines the transformation efficiency of each guide. Error bars are based on three independent replicates. (**C**) Mutations were made to particular modules in the nexus (blue module), the lower stem (green module), the bulge (yellow module) and the upper stem (purple module). Each construct is named SG for Single Guide. The wild type guide is called WT.

organisms to take-up plasmid DNA, or true differences in the targeting activity of each Cas9. With this assay we cannot compare Cas9 activity between organisms due to the design of the experiments meant to characterize endogenous activity of systems; however, interference levels within an organism can be compared to determine activity of each Cas9 with different PAMs.

This PAM flexibility may allow Cas9 to recognize sequences on rapidly mutating phages while also providing circumstantial evidence for a native mechanism of primed acquisition[7,15,27]. If Cas9 is able to flexibly target minor PAM mutants, it may bind the target long enough to acquire a new spacer from the invader. When defining what the true PAM is for each Cas9, it is important to consider there may be a difference between the sequences that allow for spacer acquisition and sequences that permit Cas9 recognition and binding[7,15,17,27,51]. When predicting the PAM using native protospacer sequences, we infer the sequence that Cas9 recognizes during the adaptation stage of CRISPR-Cas immunity; this prediction is likely more stringent than the total recognition space during the interference stage. When determining the PAM through depletion assays, broader flexibility is seen in PAM sequences which may be an evolutionary advantage during immunity as phages and other MGEs are known to rapidly evolve[7,15,52].

It should be noted that the constructs with imperfect PAMs did occasionally show interference; this is likely due to PAM flexibility and the ability of Cas9 to promiscuously, though less effectively, recognize non-canonical or sub-ideal PAMs (Figs 6, S6). There were also instances where the predicted PAM was likely not the optimal PAM as many colonies were able to escape CRISPR targeting (Fig. S6). Escapees can point to several issues with CRISPR activity. The Cas9 protein may not be fully active and cannot fully eliminate all targets. Another possibility includes potential biases inherent to target sequences that affect the ability of Cas9 to interfere.

Once all components required for Cas9 functionality had been determined, we chose one system to develop into single guide RNA targeting technology. Interestingly, the ability to increase guide efficiency through mutagenesis seen here is contradictory to the Spy sgRNA data presented in Briner *et al.*[12], and may be specific to lactobacilli or *L. gasseri*. This is the first investigation of perturbations allowable to double stemmed nexus tracrRNAs; modulation of Cas9 activity through mutations to the double stemmed nexus may be a function unique to these structures. Additionally, this is the first experiment to express an engineered single guide RNA and achieve self-targeting with an endogenous Cas9; previous approaches have relied on heterologous Cas machinery and engineered repeat-spacer arrays. This research opens the door to perform genome editing or targeted killing in bacteria containing native Cas9s with engineered sgRNAs.

Overall, here we present evidence of activity in the expression and interference phases of CRISPR immunity and circumstantial evidence for active acquisition in lactobacilli. Through investigation of the genetic diversity

of CRISPR-Cas systems in hosts were they are naturally enriched, we found five potentially orthogonal systems that contain divergent Cas1s, Cas9s, ldrRNAs, CRISPR repeats, tracrRNAs, and PAMs. Insights into the transcriptional boundaries of the crRNAs and tracrRNAs during the expression stage, allowed us to successfully design a single guide RNA in *L. gasseri* that is able to mimic the native crRNA:tracrRNA duplex and have potentially designed guides that Cas9 can utilize better than the wildtype guides. We explored the native targets of CRISPR-Cas spacers to determine not only what predators attack lactobacilli, but were also able to infer what PAM sequence each Cas9 likely targets. Through plasmid interference assays, we confirmed the relative efficiency of each PAM and noticed a trend of flexible PAM targeting that may have implications both for the bacterial adaptive immunity and for genome editing applications of Cas9. In the literature it has been suggested that most CRISPR-Cas systems are not active or have low targeting activity against DNA[37,53], but this does not appear to be the case with lactobacilli. The diversity of spacer sequences suggests lactobacilli live in a competitive environment under high phage pressure; likely due to the constant threat from invading DNA, CRISPR-Cas systems in lactobacilli need to be constitutively active and ready for defense.

The popularity of CRISPR-Cas systems exploded when Cas9 was first used as a genome-editing tool[54]. Through characterization of all three stages of CRISPR-Cas interference in Type II systems, we were able to develop the basic information necessary to develop potential new genome-editing tools that can be used both natively in bacteria and heterologously in eukaryotic systems. The systems we investigated here cluster into five consistent phylogenetic groups based on Cas1 sequence, Cas9 sequence, ldrRNA and crRNA sequence and length, tracrRNA sequence and structure, and PAM recognition sequence. Future studies will likely show these separate phylogenetic groups are orthogonal systems that contain independent machinery not capable of cross-talk and can be used to multiplex systems for genome editing. By characterizing the native functions of CRISPR-Cas machinery in their hosts, we are able to expand the Cas9 toolbox. The tools created from these systems will be capable of targeting a broader range of sequences due to novel PAM sequences, enabling more precise targeting, and can be used concurrently to multiplex with different Cas9s due to novel sgRNAs.

Methods

***In silico* analyses.** 1,262 *Lactobacillus* genomes were downloaded from NCBI (Table S1). CRISPR-Cas content was detected using the CRISPRdisco pipeline[55]. The core genome tree was generated using the proteins identified by Sun *et al.*, 2015 and aligned used the CLC Genomics ® Workbench. The tree was generated with 100 bootstrap replications in CLC Genomics. The metadata was added to the tree with the results of our CRISPR-Cas annotations.

Protein sequences for the universal Cas1 protein and Type II signature protein, Cas9, were aligned using MUSCLE[56]. Neighbor-joining trees with 100 bootstrap replications were generated using MEGA6[57]; the Cas1 tree was rooted on the Type I-Type II CRISPR-Cas system split, while the Cas9 tree was rooted on the Type II-C branch containing *Neisseria meningitidis* and *Lactobacillus coryniformis*. The highly investigated Cas9 proteins from *Streptococcus pyogenes*, *Streptococcus thermophilus*, *Staphylococcus aureus*, and *N. meningitidis* were included in the analysis to demonstrate the diversity in the Cas9 dataset. A smaller subset of Cas9s from all the systems identified were selected for further characterization based diversity throughout the Cas9 space and uniqueness within the group.

Using the alignment of the Cas9 proteins, the protein motifs as identified by Nishimasu *et al.*[58] for Spy Cas9 and Ran *et al.*[37] for Sau Cas9 were mapped onto the selected subset of Cas9 proteins.

To identify native protospacer targets encoded by the CRISPR arrays, spacers were BLASTed against publically available data including the nr/nt, HTGS, WGS, and SRA databases (Table S2). Positive hits were defined as covering at least 80% of the spacer length with 90% or higher sequence identity. The 10 nucleotide flanking regions on the 5′ and 3′ ends of the protospacer sequences were aligned by hand and submitted to WebLogo[59] for sequence motif identification.

Plasmid generation with inserts. Interference plasmids were generated to test activity of CRISPR-Cas systems using native machinery *in vivo*. A protospacer sequence was selected for each organism by selecting a spacer that exhibited a highly expressed crRNA. PAM mutants were designed to test flexibility and spacing of the Cas recognition machinery. Double stranded inserts were generated by annealing extended oligos containing the protospacer, PAM, and *BamH*I/*Sac*I or *Hind*III/*Spe*I restriction sites. Plasmids were heat shocked into chemically competent *Escherichia coli* D10 or GM1829 cells and plated on selective media containing erythromycin and IPTG/Xgal (Thermo-Fischer). Positive clones were grown in overnight shaking cultures and plasmids were extracted using the QIAGEN Spin MiniPrep kit. The PAM and protospacer sequences were confirmed via Sanger sequencing at the NC State Genomic Science Lab (Raleigh, NC). Plasmids were quantified using a NanoDrop 2000c. Oligos used to generate these plasmids can be found in Table S3.

Plasmid interference assay. Transformations were optimized for *Lactobacillus casei*, *Lactobacillus rhamnosus*, *Lactobacillus gasseri*, *Lactobacillus jensenii*, and *Lactobacillus pentosus*. Overnight cultures were inoculated into 100 mL of Man-de Rossa-Sharpe (MRS) broth with or without 2% glycine at an OD of 0.05 at 600 nm. Cultures were grown to OD 0.50, with some species receiving ampicillin at a final concentration of 10 ug/mL. Cells were pelleted by centrifugation at 5,000 × g for 15 minutes. Some cultures received a lithium acetate [7 mM phosphate buffer, pH 7.4, 600 mM sucrose, 100 mM lithium acetate, 10 mM dithiothreitol] incubation for 30 minutes and spinning at 4,500 × g for 15 minutes. Pellets were resuspended in 50 mL of 3.5X Sucrose Magnesium Electroporation Buffer (SMEB) buffer containing 7 nM phosphate buffer, pH 7.4, 952 mM sucrose, 3.5 mM $MgCl_2$. The cultures were centrifuged at 4,000 × g for 15 minutes, resuspended in 25 mL 3.5X SMEB, centrifuged at 4,000 × g for 20 minutes, and resuspended in a final 1 mL of 3.5X SMEB. 100 uL of competent cells were added to 400 ng of plasmid and pipetted into a pre-chilled 2 mm gap electroporation cuvette. The cultures were

electroporated at a constant voltage of 2.5 kV. Post electroporation, the cells were immediately added to 900 uL of pre-warmed MRS with 1%v/v recovery buffer [2 M sucrose, 20 mM $CaCl_2$, 200 mM $mgCl_2$] and recovered overnight. Cells were plated on MRS agar containing erythromycin and grown anaerobically for two to five days. Colonies were counted to determine interference capabilities of Cas9 with the different PAM variants. Standard error was calculated based on three replications.

RNA-Seq. Cultures were grown to mid-log phase, harvested by centrifugation, and lysed via bead-beating in Trizol (Life Technologies, Carlsbad, CA) with 0.5 mm glass beads (MO BIO Laboratories, Carlsbad, CA). RNA was purified from the lysate using the Direct-zol RNA Miniprep Kit with in column DNase digestion (Zymo Research, Irvine, CA). Total RNA was submitted to the University of Illinois Roy J. Carver Biotechnology Center High-Throughput Sequencing and Genotyping Unit, and smRNA libraries were prepared with the NextFlex Small RNA-Seq Library Prep kit V2 (Bio Scientific, Austin, TX) for size-selected fragments 17 to 200 nt in length. The libraries were sequenced in a single lane of Illumina HiSeq. 2500 with a read length of 180 nt. Data was received de-multiplexed and uploaded into Geneious® for adapter removal followed by quality trimming to an error probability limit of 0.001, filtering to exclude reads <15 nt, and mapping to the reference genome for each species using Bowtie2[60]. Box plots were generated with the statistical program R.

Self-targeting assay. Synthetic single guide RNAs (sgRNA) were designed for *L. gasseri* based on the RNA-Seq confirmed boundaries for the tracrRNA and crRNAs. A protospacer sequence flanked by the PAM 5′-cTAAC-3′ in the FruK was selected as the target for a chromosomal self-targeting assay. The corresponding spacer sequence was designed in the guide RNA. A highly expressed promoter for the *tuf* gene was cloned in front of the sgRNA. Using the transformation protocol for *L. gasseri* in the plasmid interference assays, plasmids containing the promoter and single guide were transformed into the cells. Overnight recovered cells were plated on minimal MRS containing 10% fructose, 3 ug/ml erythromycin, and bromocresol purple to assess the ability of the transformants to still metabolize fructose.

Data availability. The BioProject ID for this experiment is PRJNA400806. The raw small RNA data can be reached using the following SRA Accession Numbers: SRR5997381-SRR5997390.

References

1. Barrangou, R. *et al.* CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315**, 1709–1712 (2007).
2. Brouns, S. J. *et al.* Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science* **321**, 960–964, https://doi.org/10.1126/science.1159689 (2008).
3. Marraffini, L. A. & Sontheimer, E. J. CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. *Science* **322**, 1843–1845, https://doi.org/10.1126/science.1165771 (2008).
4. Hille, F. & Charpentier, E. CRISPR-Cas: biology, mechanisms and relevance. *Philos Trans R Soc Lond B Biol Sci* **371**, https://doi.org/10.1098/rstb.2015.0496 (2016).
5. Barrangou, R. The roles of CRISPR-Cas systems in adaptive immunity and beyond. *Curr Opin Immunol* **32**, 36–41, https://doi.org/10.1016/j.coi.2014.12.008 (2015).
6. Wei, Y., Terns, R. M. & Terns, M. P. Cas9 function and host genome sampling in Type II-A CRISPR-Cas adaptation. *Genes Dev* **29**, 356–361, https://doi.org/10.1101/gad.257550.114 (2015).
7. Paez-Espino, D. *et al.* Strong bias in the bacterial CRISPR elements that confer immunity to phage. *Nat Commun* **4**, 1430, https://doi.org/10.1038/ncomms2440 (2013).
8. Arslan, Z., Hermanns, V., Wurm, R., Wagner, R. & Pul, U. Detection and characterization of spacer integration intermediates in type I-E CRISPR-Cas system. *Nucleic Acids Res* **42**, 7884–7893, https://doi.org/10.1093/nar/gku510 (2014).
9. Nuñez, J. K. *et al.* Cas1–Cas2 complex formation mediates spacer acquisition during CRISPR–Cas adaptive immunity. *Nat Struct Mol Biol* **21**, 528–534, https://doi.org/10.1038/nsmb.2820, http://www.nature.com/nsmb/journal/v21/n6/abs/nsmb.2820.html#supplementary-information (2014).
10. Gasiunas, G., Barrangou, R., Horvath, P. & Siksnys, V. Cas9–crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. *Proc Natl Acad Sci USA* **109**, E2579–2586, https://doi.org/10.1073/pnas.1208507109 (2012).
11. Karvelis, T. *et al.* crRNA and tracrRNA guide Cas9-mediated DNA interference in Streptococcus thermophilus. *RNA biology* **10**, 841–851 (2013).
12. Briner, A. E. *et al.* Guide RNA functional modules direct Cas9 activity and orthogonality. *Molecular cell* **56**, 333–339 (2014).
13. Sapranauskas, R. *et al.* The Streptococcus thermophilus CRISPR/Cas system provides immunity in Escherichia coli. *Nucleic acids research* **39**, 9275–9282 (2011).
14. Horvath, P. *et al.* Diversity, activity, and evolution of CRISPR loci in Streptococcus thermophilus. *J Bacteriol* **190**, 1401–1412, https://doi.org/10.1128/JB.01415-07 (2008).
15. Deveau, H. *et al.* Phage response to CRISPR-encoded resistance in Streptococcus thermophilus. *Journal of bacteriology* **190**, 1390–1400 (2008).
16. Marraffini, L. A. & Sontheimer, E. J. Self versus non-self discrimination during CRISPR RNA-directed immunity. *Nature* **463**, 568–571, https://doi.org/10.1038/nature08703 (2010).
17. Mojica, F. J., Diez-Villasenor, C., Garcia-Martinez, J. & Almendros, C. Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiology* **155**, 733–740, https://doi.org/10.1099/mic.0.023960-0 (2009).
18. Grissa, I., Vergnaud, G. & Pourcel, C. The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. *BMC Bioinformatics* **8**, 172, https://doi.org/10.1186/1471-2105-8-172 (2007).
19. Makarova, K. S. *et al.* An updated evolutionary classification of CRISPR-Cas systems. *Nat Rev Microbiol* **13**, 722–736, https://doi.org/10.1038/nrmicro3569 (2015).
20. Shmakov, S. *et al.* Diversity and evolution of class 2 CRISPR-Cas systems. *Nat Rev Microbiol* **15**, 169–182, https://doi.org/10.1038/nrmicro.2016.184 (2017).
21. Burstein, D. *et al.* New CRISPR-Cas systems from uncultivated microbes. *Nature* **542**, 237–241, https://doi.org/10.1038/nature21059 (2017).
22. Jinek, M. *et al.* A programmable dual-RNA–guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816–821 (2012).
23. Cong, L. *et al.* Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819–823, https://doi.org/10.1126/science.1231143 (2013).
24. Mali, P. *et al.* RNA-guided human genome engineering via Cas9. *Science* **339**, 823–826, https://doi.org/10.1126/science.1232033 (2013).

25. Sun, Z. *et al.* Expanding the biotechnology potential of lactobacilli through comparative genomics of 213 strains and associated genera. *Nat Commun* **6**, 8322, https://doi.org/10.1038/ncomms9322 (2015).
26. Horvath, P. *et al.* Comparative analysis of CRISPR loci in lactic acid bacteria genomes. *Int J Food Microbiol* **131**, 62–70, https://doi.org/10.1016/j.ijfoodmicro.2008.05.030 (2009).
27. Paez-Espino, D. *et al.* CRISPR immunity drives rapid phage genome evolution in Streptococcus thermophilus. *MBio* **6**, https://doi.org/10.1128/mBio.00262-15 (2015).
28. Wei, Y., Chesne, M. T., Terns, R. M. & Terns, M. P. Sequences spanning the leader-repeat junction mediate CRISPR adaptation to phage in Streptococcus thermophilus. *Nucleic Acids Res* **43**, 1749–1758, https://doi.org/10.1093/nar/gku1407 (2015).
29. Held, N. L., Herrera, A. & Whitaker, R. J. Reassortment of CRISPR repeat-spacer loci in Sulfolobus islandicus. *Environ Microbiol* **15**, 3065–3076, https://doi.org/10.1111/1462-2920.12146 (2013).
30. Deng, L., Garrett, R. A., Shah, S. A., Peng, X. & She, Q. A novel interference mechanism by a type IIIB CRISPR-Cmr module in Sulfolobus. *Mol Microbiol* **87**, 1088–1099, https://doi.org/10.1111/mmi.12152 (2013).
31. Haurwitz, R. E., Jinek, M., Wiedenheft, B., Zhou, K. & Doudna, J. A. Sequence- and structure-specific RNA processing by a CRISPR endonuclease. *Science* **329**, 1355–1358, https://doi.org/10.1126/science.1192272 (2010).
32. Wiedenheft, B. *et al.* Structural basis for DNase activity of a conserved protein implicated in CRISPR-mediated genome defense. *Structure* **17**, 904–912, https://doi.org/10.1016/j.str.2009.03.019 (2009).
33. Ivancic-Bace, I., Cass, S. D., Wearne, S. J. & Bolt, E. L. Different genome stability proteins underpin primed and naive adaptation in E. coli CRISPR-Cas immunity. *Nucleic Acids Res* **43**, 10821–10830, https://doi.org/10.1093/nar/gkv1213 (2015).
34. Toms, A. & Barrangou, R. On the global CRISPR array behavior in class I systems. *Biol Direct* **12**, 20, https://doi.org/10.1186/s13062-017-0193-2 (2017).
35. Smargon, A. A. *et al.* Cas13b Is a Type VI-B CRISPR-Associated RNA-Guided RNase Differentially Regulated by Accessory Proteins Csx27 and Csx28. *Mol Cell* **65**, 618–630 e617, https://doi.org/10.1016/j.molcel.2016.12.023 (2017).
36. Esvelt, K. M. *et al.* Orthogonal Cas9 proteins for RNA-guided gene regulation and editing. *Nat Meth* **10**, 1116–1121, https://doi.org/10.1038/nmeth.2681, http://www.nature.com/nmeth/journal/v10/n11/abs/nmeth.2681.html#supplementary-information (2013).
37. Ran, F. A. *et al. In vivo* genome editing using Staphylococcus aureus Cas9. *Nature* **520**, 186–191, https://doi.org/10.1038/nature14299, http://www.nature.com/nature/journal/v520/n7546/abs/nature14299.html#supplementary-information (2015).
38. Deltcheva, E. *et al.* CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature* **471**, 602–607 (2011).
39. Pertzev, A. V. & Nicholson, A. W. Characterization of RNA sequence determinants and antideterminants of processing reactivity for a minimal substrate of Escherichia coli ribonuclease III. *Nucleic Acids Res* **34**, 3708–3721, https://doi.org/10.1093/nar/gkl459 (2006).
40. Sanozky-Dawes, R., Selle, K., O'Flaherty, S., Klaenhammer, T. & Barrangou, R. Occurrence and activity of a type II CRISPR-Cas system in Lactobacillus gasseri. *Microbiology* **161**, 1752–1761, https://doi.org/10.1099/mic.0.000129 (2015).
41. Chaudhary, K., Chattopadhyay, A. & Pratap, D. Anti-CRISPR proteins: Counterattack of phages on bacterial defense (CRISPR/Cas) system. *J Cell Physiol*, https://doi.org/10.1002/jcp.25877 (2017).
42. Hynes, A. P. *et al.* An anti-CRISPR from a virulent streptococcal phage inhibits Streptococcus pyogenes Cas9. *Nat Microbiol*, https://doi.org/10.1038/s41564-017-0004-7 (2017).
43. Pawluk, A. *et al.* Naturally Occurring Off-Switches for CRISPR-Cas9. *Cell* **167**, 1829–1838 e1829, https://doi.org/10.1016/j.cell.2016.11.017 (2016).
44. Li, M., Wang, R. & Xiang, H. Haloarcula hispanica CRISPR authenticates PAM of a target sequence to prime discriminative adaptation. *Nucleic Acids Res* **42**, 7226–7235, https://doi.org/10.1093/nar/gku389 (2014).
45. Li, M., Wang, R., Zhao, D. & Xiang, H. Adaptation of the Haloarcula hispanica CRISPR-Cas system to a purified virus strictly requires a priming process. *Nucleic Acids Res* **42**, 2483–2492, https://doi.org/10.1093/nar/gkt1154 (2014).
46. Levy, A. *et al.* CRISPR adaptation biases explain preference for acquisition of foreign DNA. *Nature* **520**, 505–510, https://doi.org/10.1038/nature14302 (2015).
47. Fagerlund, R. D. *et al.* Spacer capture and integration by a type I-F Cas1-Cas2-3 CRISPR adaptation complex. *Proc Natl Acad Sci USA* **114**, E5122–E5128, https://doi.org/10.1073/pnas.1618421114 (2017).
48. Staals, R. H. *et al.* Interference-driven spacer acquisition is dominant over naive and primed adaptation in a native CRISPR-Cas system. *Nat Commun* **7**, 12853, https://doi.org/10.1038/ncomms12853 (2016).
49. Fineran, P. C. *et al.* Degenerate target sites mediate rapid primed CRISPR adaptation. *Proc Natl Acad Sci USA* **111**, E1629–1638, https://doi.org/10.1073/pnas.1400071111 (2014).
50. McGinn, J. & Marraffini, L. A. CRISPR-Cas Systems Optimize Their Immune Response by Specifying the Site of Spacer Integration. *Mol Cell* **64**, 616–623, https://doi.org/10.1016/j.molcel.2016.08.038 (2016).
51. Leenay, R. T. & Beisel, C. L. Deciphering, Communicating, and Engineering the CRISPR PAM. *J Mol Biol* **429**, 177–191, https://doi.org/10.1016/j.jmb.2016.11.024 (2017).
52. Leenay, R. T. *et al.* Identifying and Visualizing Functional PAM Diversity across CRISPR-Cas Systems. *Mol Cell* **62**, 137–147, https://doi.org/10.1016/j.molcel.2016.02.031 (2016).
53. Kleinstiver, B. P. *et al.* Engineered CRISPR-Cas9 nucleases with altered PAM specificities. *Nature* **523**, 481–485, https://doi.org/10.1038/nature14592 (2015).
54. Barrangou, R. & Doudna, J. A. Applications of CRISPR technologies in research and beyond. *Nat Biotechnol* **34**, 933–941, https://doi.org/10.1038/nbt.3659 (2016).
55. Crawley, A. B., Henriksen, J. R. & Barrangou, R. CRISPRdisco: an automated pipeline for the discovery and analysis of CRISPR-Cas systems. *The CRISPR Journal* **1**, epub ahead of print. (2018).
56. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**, 1792–1797, https://doi.org/10.1093/nar/gkh340 (2004).
57. Tamura, K., Stecher, G., Peterson, D., Filipski, A. & Kumar, S. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Mol Biol Evol* **30**, 2725–2729, https://doi.org/10.1093/molbev/mst197 (2013).
58. Nishimasu, H. *et al.* Crystal structure of Cas9 in complex with guide RNA and target DNA. *Cell* **156**, 935–949 (2014).
59. Crooks, G. E., Hon, G., Chandonia, J. M. & Brenner, S. E. WebLogo: a sequence logo generator. *Genome Res* **14**, 1188–1190, https://doi.org/10.1101/gr.849004 (2004).
60. Langdon, W. B. Performance of genetic programming optimised Bowtie2 on genome comparison and analytic testing (GCAT) benchmarks. *BioData Min* **8**, 1, https://doi.org/10.1186/s13040-014-0034-0 (2015).

Acknowledgements

We would like to thank all members of the Barrangou and Klaenhammer lab for guidance on transformation protocols for lactobacilli, especially Dr. Yong Jun Goh, Dr. Sarah O'Flaherty, Rosemary Sanozky-Dawes, and Evelyn Durmaz. Additionally, we thank Anna Townsend for technical assistance. This was work was funded through the Ag Foundation and NC State University internal funding. This work was sponsored by the NC AgFoundation.

Author Contributions

R.B., A.B.C., E.D.H., and E.S. were involved in planning and design of the experiments. R.B., A.B.C., E.D.H., E.S. and K.B. analyzed data. A.B.C., E.D.H., E.S., and K.B. performed experiments. A.B.C. and R.B. wrote the manuscript.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-018-29746-3.

Competing Interests: R.B. and A.B.C. are inventors on several CRISPR-related patents. R.B. is a co-founder and investor in Locus Biosciences and Intellia Therapeutics and an investor in Caribou Biosciences.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

APPENDIX D: APPLICATIONS OF CRISPR TECHNOLOGIES ACROSS THE FOOD SUPPLY CHAIN

D.1. CONTRIBUTION TO THE WORK

Katelyn Brandt is first author on "Applications of CRISPR technologies across the food supply chain" published in *Annual Reviews: Food Science and Technology*. She and Rodolphe Barrangou wrote the manuscript and prepared the figures. The following chapter is from Brandt and Barrangou *Annual Review of Food Science and Technology* 10(1), 133-150.

D.2. REPRINT OF "APPLICATIONS OF CRISPR TECHNOLOGIES ACROSS THE FOOD SUPPLY CHAIN"

ANNUAL REVIEWS

Annual Review of Food Science and Technology

Applications of CRISPR Technologies Across the Food Supply Chain

Katelyn Brandt[1,2] and Rodolphe Barrangou[1,2]

[1]Genomic Sciences, Bioinformatics Research Center, North Carolina State University, Raleigh, North Carolina 27695, USA; email: rbarran@ncsu.edu

[2]Department of Food, Bioprocessing & Nutrition Sciences, North Carolina State University, Raleigh, North Carolina 27695, USA

Annu. Rev. Food Sci. Technol. 2019. 10:133–50

The *Annual Review of Food Science and Technology* is online at food.annualreviews.org

https://doi.org/10.1146/annurev-food-032818-121204

Keywords

CRISPR-Cas, genome editing, crop development, food supply chain, livestock

Abstract

The food industry faces a 2050 deadline for the advancement and expansion of the food supply chain to support the world's growing population. Improvements are needed across crops, livestock, and microbes to achieve this goal. Since 2005, researchers have been attempting to make the necessary strides to reach this milestone, but attempts have fallen short. With the introduction of clustered regularly interspaced short palindromic repeats (CRISPRs) and CRISPR-associated (Cas) proteins, the food production field is now able to achieve some of its most exciting advancements since the Green Revolution. This review introduces the concept of applying CRISPR-Cas technology as a genome-editing tool for use in the food supply chain, focusing on its implementation to date in crop, livestock, and microbe production, advancement of products to market, and regulatory and societal hurdles that need to be overcome.

INTRODUCTION

With a continuously expanding world population, the food supply chain faces some of its greatest challenges since the Green Revolution. It has been predicted that total crop production must double from the 2005 levels by 2050 to meet the needs of the world population (Tilman et al. 2011). It has been ten years since this prediction was made and crop production is still struggling to meet this goal. Models have estimated that a 2.4% annual increase in crop yield is necessary to reach the 2050 milestone. Unfortunately, a 2013 report revealed that none of the top produced crops in the world are anywhere close to this number (Ray et al. 2013). There are many contributing factors to this lack of increase in crop production. One factor is the amount of arable land available for crops. It has been determined that clearing more land is not the most desirable option. Although increasing arable land may not be viable, recent projections from the United States Department of Agriculture (USDA) show that the number of acres planted will remain stable (USDA 2018b). Moving past arable land, many crops across the globe struggle to thrive because of disease and stress conditions (Ma et al. 2018). Additional strains to the food supply chain are being anticipated. As nations become more affluent, increased demand for livestock is likely. This not only means that livestock production must increase but consequently that more crops will be needed to support the corresponding increase in feed consumption. Increases in demands for cleaner energy also mean that more crops will be outsourced for biofuels (USDA 2018b). Overall, changes must be made to meet global needs. Because of the hurdles listed above, most research is being focused on improving crops to increase yield. Many techniques enable such improvements, but most researchers are turning to the promise of genome editing.

Genome editing can be achieved through various methodologies, but the three most common are zinc finger nucleases (ZFNs), transcription activator–like effector nucleases (TALENs), and clustered regularly interspaced short palindromic repeats (CRISPRs). CRISPR has emerged as the clear leader among the three technologies. It was first fully characterized as an adaptive immune response in bacteria in 2007, when it was established as a phage-resistance mechanism in yogurt cultures (Barrangou et al. 2007). Shortly thereafter, researchers elucidated the CRISPR-Cas (CRISPR-associated) mechanism and repackaged this molecular machinery as a genome-editing tool, and the CRISPR craze was born (Pennisi 2013). Many factors have contributed to CRISPR's unprecedented level of popularity across scientific disciplines, not the least of which has been the ease of access—or democratization—of CRISPR. Addgene, a nonprofit plasmid repository, has made CRISPR readily accessible for $65 per plasmid. This universal access has opened up research avenues to companies, academics, and nonprofits alike, and already 3,400 laboratories have received CRISPR shipments (LaManna & Barrangou 2018). The popularity of CRISPR lies not only in its accessibility but most importantly in its success. Shown to be easy to use, highly specific, and programmable, CRISPR has become the tool of choice for many researchers. CRISPR efficacy has been demonstrated in many model organisms and industrial workhorses used in biotechnology, medicine, and agriculture. Although still relatively new to the agriculture field, CRISPR has shown great promise in its ability to overcome many of the hurdles facing the industry, specifically, how to improve crop yield to avoid the upcoming food gap crisis. Perhaps it is fitting that the most promising tool in agriculture was first discovered in the food industry.

In this review, we examine CRISPR as it applies to the food supply chain. We begin with CRISPR biology to examine what has enabled its successful transition from an adaptive immune response to a genome-editing tool. We also discuss unique applications and approaches, beyond the basic uses of genome editing, that can be further taken advantage of. We then detail the successful application of CRISPR in crops, livestock, and microbes. CRISPR has presented unique opportunities to researchers in these areas and has easily outpaced incumbent technologies. However,

the fast pace of adoption has prevented regulatory agencies from keeping up with the CRISPR craze. We discuss impacts CRISPR will likely have on the regulatory process and how societal concerns may affect future paths.

CRISPR BIOLOGY

The knowledge of CRISPR biology has rapidly expanded since 1987, when it was first noted as unusual repetitive sequences in bacterial genomes (Ishino et al. 1987). In the early stages of CRISPR discovery, three main types were identified: Type I, II, and III. Today the different CRISPR systems are divided into 2 classes, 6 types, and 23 subtypes, with additional putative types to be validated (Koonin et al. 2017). Class 1 is characterized by a multiunit effector complex. This class consists of Types I, III, and IV. Class 2 is characterized by single protein effectors. This class consists of Type II, V, and VI (Makarova et al. 2011a,b). The single protein effector of Class 2 has garnered a great deal of attention, specifically in terms of genome engineering. The Type II signature protein Cas9 is the most widely popular and most used effector protein of the CRISPR-Cas family and as such is the main focus of this review. Regardless of the class, all CRISPR systems are DNA encoded, RNA mediated, and nucleic acid targeting (Barrangou et al. 2007, Brouns et al. 2008, Hale et al. 2009, Marraffini & Sontheimer 2008). Each CRISPR-Cas locus consists of the *cas* genes and the CRISPR array. The *cas* genes are specific to each CRISPR type, whereas the CRISPR array consists of the leader, which drives transcription of the array, and the repeats and spacers (Horvath & Barrangou 2010). **Figure 1*a*** depicts a canonical Type II CRISPR locus. Spacers are derived from invading mobile genetic elements (MGEs) and act as the memory for the adaptive immune response. CRISPR biology occurs in three stages: adaptation, expression, and interference.

Adaptation

Adaptation has two main steps: the acquisition of new spacers and the integration of spacers into the genome. **Figure 1*b*** depicts the canonical Type II adaptation stage. This stage is directed by the nearly universal Cas1–Cas2 proteins (Type VI is the exception) (Makarova et al. 2015). Acquisition occurs when foreign DNA enters the cell and is interrogated by Cas. The system copies a small piece of the foreign DNA and incorporates it into the host's genome. DNA from an invading source is termed a protospacer, whereas incorporated DNA is termed a spacer (Deveau et al. 2008). Spacers are added iteratively, with new spacers consistently being added at the leader end of the repeat–spacer array (Barrangou et al. 2013). For the Type II system, a protospacer is identified via the presence of a signature known as the protospacer adjacent motif (PAM; 2–7-nt PAM sequence) (Horvath et al. 2008). This acts as a recognition signal for the target sequence. It is also useful in distinguishing self from nonself, as the PAM is not incorporated into the CRISPR locus (Marraffini & Sontheimer 2010). This is important for subsequent infections. Additionally, as the spacers are added consecutively and consistently, the array becomes a recorded history of infection events for the organism (Andersson & Banfield 2008, Tyson & Banfield 2008). This can then be used to identify related strains and their history.

Expression

The next stage is expression or biogenesis of CRISPR RNA (crRNA). Transcription begins at the leader end of the repeat–spacer array. The leader sequence, which is immediately adjacent to the array, contains a promoter sequence for the expression of crRNA (Carte et al. 2014).

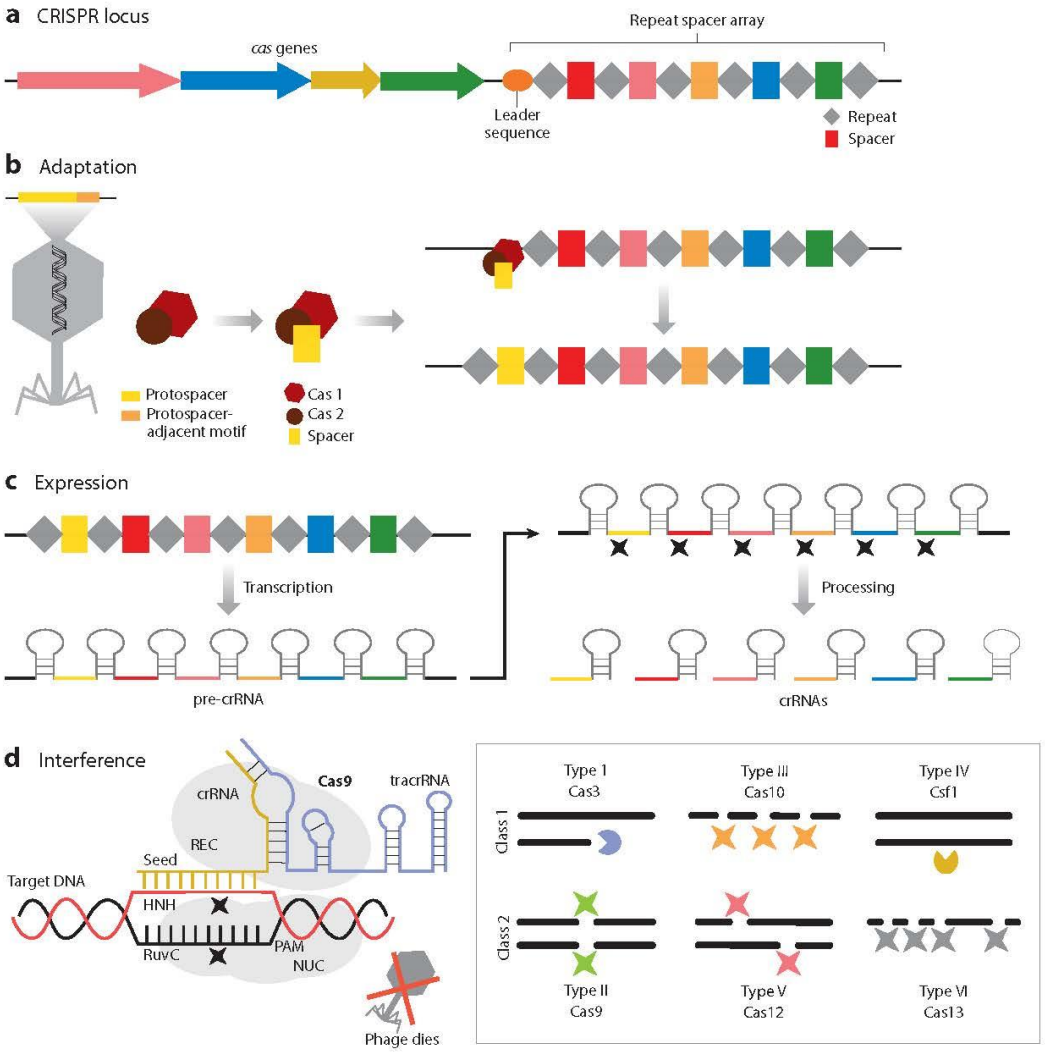


Figure 1

CRISPR (clustered regularly interspaced short palindromic repeat) biology. (*a*) CRISPR locus containing *cas* genes, a leader sequence, and a repeat–spacer array. (*b*) CRISPR adaptation involves the excision of a protospacer from invading DNA and its incorporation as a spacer into the host's repeat–spacer array. (*c*) CRISPR expression involves the transcription and processing of crRNAs (CRISPR RNAs). (*d*) CRISPR interference is achieved when Cas9 bound to the crRNA::tracrRNA complex recognizes the PAM in phage DNA, leading to double-strand break and death of the phage. The inset depicts the different cleavage types of each CRISPR system.

crRNA abundance depends on proximity to the leader, with the most proximal being the most abundant (Zoephel & Randau 2013). Having the newest spacers incorporated after the leader makes sense, as it is more likely for a host to be infected by a more recent phage as opposed to an earlier exposure. Transcription produces one long precrRNA (precursor crRNA) that is processed into mature crRNAs by cleaving within the repeat region (**Figure 1*c***) (Deltcheva et al. 2011). A crRNA is composed of the latter end of a spacer and the beginning of the repeat sequence. Each has a defined length for the system (Crawley et al. 2018). Therefore, it is the crRNA that provides the target for the CRISPR system. In Type II systems, the tracrRNA (*trans*-activating crRNA) is essential to the processing of precrRNA and binding to the effector protein Cas9 (Deltcheva et al. 2011). Cas9 recognizes bound crRNA–tracrRNA complexes and undergoes processing. From here, Cas9 remains bound to the mature crRNA::tracrRNA for the final stage of CRISPR biology.

Interference

The process of targeting nucleic acids is termed interference. This stage is directed by the effector protein or complex. The effector complex is formed, recognizes foreign nucleic acids, and generates nucleic acid cleavage. In Type IIs (**Figure 1*d***), the only required components are the Cas9 bound to the crRNA:tracrRNA duplex (Barrangou 2015). Much like during adaptation, DNA is scanned for a PAM site (Sternberg et al. 2014). Cas9 binds to the noncomplementary strand of the PAM and then checks for sequence similarity between the spacer in the crRNA and the target DNA. If no mismatches are identified in the seed region, then Cas9 enacts an exact double-strand break through the use of two nickases (**Figure 1*d***) (Sternberg et al. 2014). Targeted DNA includes phages, plasmids, and chromosomal DNA. Target and cleavage are characteristic of each CRISPR type (**Figure 1**, inset) (Barrangou & Horvath 2017). Type I uses an exonuclease that cleaves and degrades one strand of DNA. Type IIIs target mRNA and cleave in a ruler mechanism. Type Vs use nickases like Type IIs do; however, the result is staggered nicking of the DNA strand. Type VIs introduce a cut in mRNA and then nonspecifically continue to process its target. Finally, the exact cleavage mechanism of Type IV remains unknown.

CRISPR AS A GENOME-EDITING TOOL

Although CRISPR is natively an immune defense system, its basic biology has made it easily adaptable as a genome-editing tool. Although most of the CRISPR systems may be co-opted in this manner, it is the Type II system that has gained the most popularity. One of the earliest systems discovered, the Type II system attracted researchers because of its single effector protein Cas9. Unlike the other known systems at the time, Cas9 was the only protein needed to identify, target, and cleave target DNA sequences. Not only was Cas9 the only protein needed, it was discovered in 2012 that Cas9 was programmable. By providing a synthesized single-guide RNA (sgRNA), researchers could provide Cas9 with specific targets of interest. The sgRNA mimics the native crRNA::tracrRNA, and the only change needed to alter the target is to the spacer region (Jinek et al. 2012). The ability to deliver only a single protein and RNA made exogenous use of the Type II system remarkably simple. **Figure 2** depicts how Cas9 can be used as a genome-editing tool. Programmable sgRNAs provide an edge to CRISPR over other genetic editing techniques. One advantage is that sgRNA synthesis is provided by numerous companies at a relatively low cost (Wang 2015). Changing a target is no longer a cumbersome process. Additionally, it is possible to multiplex sgRNAs, enabling simultaneous editing of targets in a single cell (Cong et al. 2013, Mali et al. 2013). The only requirement for Cas9 targeting outside of the delivered components is the

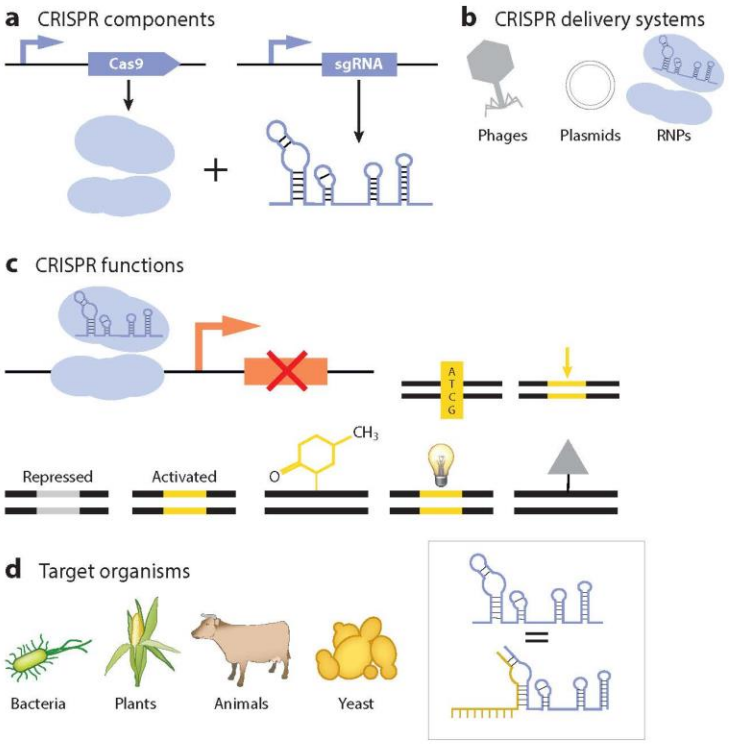


Figure 2

CRISPR (clustered regularly interspaced short palindromic repeats) editing. (*a*) CRISPR components [Cas9 and single-guide RNA (sgRNA)] needed for editing in target organisms. (*b*) CRISPR delivery can be achieved through phages, plasmids, or ribonucleoproteins (RNPs). (*c*) CRISPR can be utilized to cause knockouts, single nucleotide polymorphisms, knockins, repression, activation, epigenetic modification, imaging, and recruitment. (*d*) CRISPR effectiveness has been demonstrated in various target organisms such as bacteria, plants, animals, and yeast. The inset shows the replacement of the crRNA::tracrRNA complex with a single oligonucleotide sgRNA for genome editing.

presence of a PAM in the target region. Once all the components are designed, several delivery options are available: plasmid, phage, or ribonucleoprotein (RNP) delivery. Once delivered, Cas9 technology can be used in numerous ways. The canonical use is to introduce a blunt double-strand break, leaving the native machinery responsible for repairing the breaks (Barrangou & Doudna 2016). One method a cell could use is nonhomologous end-joining (NHEJ). NHEJ fixes double-strand breaks but often at the expense of an added or removed base, creating an INDEL (insertion–deletion) mutation in the target gene. Homologous recombination is an alternative pathway available to the cell. Cells use their own copy of the targeted gene to replace the cut copy. Alternatively, researchers can provide a template to introduce their own desired changes. This can lead to the deletion of a gene, introduction of a gene, or repair of a damaged copy of a gene (Gaj et al. 2013).

Beyond the canonical cleavage event, there are many alternative uses for CRISPR. Although still using the same cleavage scheme, Cas9 can be applied in several unique ways. For instance, self-targeting in bacteria can be used as selection markers or to find rare mutations in a population

(Selle et al. 2015). It is also possible to use CRISPR to alter microbiota composition by targeting undesirable species (Gomaa et al. 2014). Researchers have also altered Cas9 itself to allow for further applications. Nickase Cas9 (nCas9) introduces a single-strand nick as opposed to a double-strand break and was developed for more controlled knockout applications to reduce the risk of off-target effects (Doudna & Charpentier 2014, Tsai et al. 2014). Deactivated Cas9 (dCas9) does not introduce a break because of the inactivation of the two nickase domains (RuvC and HNH) but uses Cas9's homing and binding abilities for a variety of uses (Barrangou & Doudna 2016). dCas9 becomes a transcriptional regulator when combined with activator or repressor domains (Gilbert et al. 2014, Larson et al. 2013); when used in this manner the technology is termed CRISPR activation (CRISPRa) or CRISPR interference (CRISPRi), respectively. Along the same lines, dCas9 can be used for imaging, epigenetic modification, recruitment, and more (Chen et al. 2013, Hilton et al. 2015). Currently, many efforts are underway to fuse dCas9 to various effector domains (e.g., acetyl- or methyl-transferases for epigenetics; deaminases for base editing; or recombinases, gyrases, or helicases to alter DNA locally).

Despite its many advantages, the technology still has limits. Protein size, protein efficiency, and PAM presence can all affect aspects of the overall impact of genetic editing, including delivery, off-target effects, and lack of viable targets (Barrangou & Doudna 2016). As most research applications currently use either SpyCas9 or SauCas9, new Cas9s are being mined and characterized to overcome current difficulties. These new Cas9s—including the new nanoCas9s—differ in efficiency, PAM sequences, and size (Crawley et al. 2018). A second approach is the mining of other CRISPR systems. Cas12 (formerly Cpf1) of Type V is gaining in popularity with its staggered cut double-strand break. Cas3 of Type I, which is an exonuclease, is being utilized as an antimicrobial. With more types being added, it is expected that we are not yet at the saturation level of CRISPR applications.

CRISPR IN AGRICULTURE

The Plant Toolbox

Facing the looming food crisis, plant scientists have been working to improve crops to feed the world's population in the future. Specifically, they have been attempting to breed and develop crops with increased yield, disease and pest resistance, and stress tolerance. Unfortunately, this has been an extremely slow, cumbersome, and expensive process. Traditional breeding in plants for these traits may take 7 to 12 years (Acquaah 2009), and only one trait can be improved at a time. Although results can be achieved with traditional methods, this timeline does not allow for necessary measurable gains by the 2050 deadline. Researchers have thus begun looking at genome-editing technologies as an alternative to traditional breeding. The main technologies utilized in plant genome editing to date have been ZFNs, TALENs, and, most recently, CRISPR. A detailed review comparing the three technologies in plants was completed by Bortesi & Fischer (2015). Of the three, CRISPR is the leading approach for the reasons highlighted above: programmability, efficiency, and specificity, as well as accessibility, ease of use, and multiplexing. Both ZFNs and TALENs are proteins and require new proteins to be designed for each target, a lengthy and costly process compared with the CRISPR sgRNA design, which enables easy and quick reprogramming (Khan et al. 2018). The ability to multiplex with multiple sgRNAs gives CRISPR a further edge over ZFNs and TALENs, as this allows for faster development of desired traits. Additionally, because new sgRNAs are easily and cheaply designed, it is possible to develop high-throughput screens with sgRNAs (S.J. Liu et al. 2017). An additional benefit of CRISPR is that it is not limited by DNA methylation states as ZFNs and TALENs are (Hsu et al. 2013). One

concern with CRISPR is the presence of off-target effects, especially with highly repetitive genomes such as those found in plants; however, researchers have already begun to investigate methods to overcome this, such as using better guide-design tools and chemically altered Cas9s or sgRNAs (Armin et al. 2017). These reasons are why, functionally, CRISPR should be a promising enabler for plant genome editing. Practically, CRISPR has already shown great promise. Multiple groups have determined that CRISPR is effective within the first generation. Other studies have shown homozygous germline alterations that remain stable in later generations (Bortesi & Fischer 2015). Compellingly, innovative delivery and use of CRISPR could allow for nontransgenic crops and non-GMO crops (see below).

Implementation in Crops

CRISPR was first shown to be a viable tool for plant genome engineering in 2013. Jiang et al. (2013) showed proof of concept in *Arabidopsis*, tobacco, sorghum, and rice. Proof of concept was also demonstrated in wheat (Shan et al. 2013, Upadhyay et al. 2013). Beyond ability, groups have already begun using CRISPR to generate crops with the desired traits discussed above. Stress tolerance has been addressed by DuPont Pioneer (now Corteva Agriscience) with its engineered drought-resistant maize (Shi et al. 2017). Cassava has been developed that has increased protection from cassava brown streak disease, which is in the area of disease-resistant crops. And to increase yields, flowering times in soybeans have been altered (Yupeng et al. 2018).

Traditional fruits and vegetables have also been enhanced by CRISPR. Tomatoes have been edited by CRISPR (Brooks et al. 2014). Both tomatoes and potatoes have had multiple targets and traits engineered. Pathenocarpy has been a specific target among the many in tomatoes, an industry-relevant issue specifically with regard to heat-stress conditions (Karkute et al. 2017). Powdery mildew disease resistance is another industry concern that has been addressed (Nekrasov et al. 2017). Potatoes have been edited for a waxy phenotype (Andersson et al. 2017). Fruit engineering, still in the early stages, has begun in strawberries and apples (Gomez et al. 2018, Nishitani et al. 2016).

Path to Market

Along with researchers, companies have also joined the CRISPR journey. Most of the international agricultural businesses have already begun incorporating CRISPR into their pipelines. Last year Syngenta announced its acquisition of CRISPR IP (intellectual property) to begin its use in several of their crops. This list includes many of the crops in which editing has already been successful, such as corn, wheat, and rice, but also new plants such as sunflowers (Maurer 2017). Recently, a group was able to increase grain yield in rice (Miao et al. 2018). This leaves the path open for companies to target similar genes to overcome food shortages. Another group has developed mushrooms with decreased browning that the USDA has decided will not be regulated (Waltz 2016). This sets up the framework for companies' regulatory groups as they develop new products (see further discussion below). One crop is already available in the market—nonbrowning apples (Waltz 2018). Even more companies have begun making commitments to release crops in the future. DuPont is one such company; it announced the release of waxy corn by 2020 (Bomgardner 2017) and is building up an IP portfolio that is open in principle to the agricultural sector, even including competitors such as Bayer and BASF.

CRISPR IN THE ANIMAL KINGDOM

The Livestock Toolbox

In addition to crops, researchers have also been addressing livestock improvement as well. Unlike crop engineering, livestock engineering has been relatively limited in scope and scale. A recent review goes into detail about the history of genetic editing in livestock (Telugu et al. 2017). In short, traditional manipulation techniques such as selective breeding, random trans-genesis, and stem cells were not sufficiently effective to produce reliable results. The advent of genetic editing tools, such as CRISPR, has begun to change this in the animal kingdom. Impressively, groups have already established knockdown and knockin capabilities in livestock. Easy delivery, specificity, and reliable methodology have once again allowed CRISPR to outshine other technologies.

Commercial Implementation

CRISPR implementation in the animal kingdom has occurred mainly in three species: cattle, pigs, and chicken. These account for most of the husbandry business. Cattle have been the focus of many genetic editing approaches, most notably using TALENs for hornless cows (Carlson et al. 2016). However, CRISPR repurposed beyond the canonical knockout is leaving its mark on the field. Using nCas9—an altered Cas9 that cleaves only a single DNA strand—increased resistance to tuberculosis has been introduced into cattle (Gao et al. 2017). Most recently a CRISPR knockin was used to produce only male offspring (Rosenblum 2018). The rationale for this is that males grow faster and bigger than females. In another species, researchers have created lean pigs. Leaner pigs are at a lower risk of mortality, as pigs struggling to regulate temperature and fat deposits have a detrimental effect on pig production. Editing pigs to be leaner will increase pig production and save farmers money on swine going to market (Zheng et al. 2017). Porcine reproductive and respiratory syndrome virus (PRRSV)-resistant pigs have been developed and will soon become commercialized (Van Eenennaam 2018). Additionally, porcine endogenous retroviruses (PERVs) have been removed from pigs, which has a potential impact on the human health industry (Yang et al. 2015). Last, chickens, or more specifically chicken eggs, have been edited. A knockout in chickens has removed a protein from egg whites that is known to cause some allergic reactions (Oishi et al. 2016). The next big target in the animal kingdom will be aquaculture. However, genetic engineering in fish is already facing many regulatory hurdles (Ledford 2015). It has yet to be determined what regulations will be enacted and how this may affect regulation of other genetically engineered livestock. To date, proof of concept has been demonstrated in zebrafish, catfish, and salmon (Edvardsen et al. 2014, J. Liu et al. 2017, Khalil et al. 2017).

CRISPR IN MICROBES

Back to the Future

Ironically, with its success in multicellular organisms, researchers have now come full circle and are looking at new ways to implement CRISPR in microbes. CRISPR has many useful advantages in the microbial industry, both natural and engineered. The CRISPR locus itself allows for strain typing, which is a method that fingerprints each strain, allowing for identification in proprietary blends in food fermentation and probiotic products (Barrangou & Horvath 2012). CRISPR naturally provides immunity against bacteriophages. However, this can be harnessed to specifically vaccinate organisms from bacterial viruses widely encountered in large fermentations to decrease

waste in the food-manufacturing process (Selle & Barrangou 2015). From a genetic engineering standpoint, all the advantages outlined above are applicable to microbes (**Figure 2**). Engineering can be exogenous or endogenous depending on the presence and functionality of a CRISPR-Cas system and the intended use. Finally, it is possible to implement CRISPR for the targeted killing of microbes. This may take the form of self-targeting, in which CRISPR acts as a programmable and specific antimicrobial agent. Self-targeting would be used as a selection marker for the screening of specific alterations or rare natural mutations that occur in mixed bacterial populations (Selle et al. 2015). An antimicrobial agent would be used in mixed populations to remove undesirable organisms from culture blends (Beisel et al. 2014, Gomaa et al. 2014).

Industrial Applications

Even though microbes are the original source of CRISPR, applications in microbiology are still being newly discovered and implemented. One main use to date in bacteria has been typing, specifically of industrially relevant lactobacilli and streptococci as well as foodborne pathogens such as *Escherichia coli* and *Salmonella* (Barrangou & Horvath 2012). In these applications, researchers use the sequence variability within repeat–spacer arrays to distinguish strains. Strains can be completely identical in the rest of their genomes but still see differences in the CRISPR systems. This allows companies to specifically track their strains. Other work has focused on CRISPR's genetic engineering feasibility. CRISPR has been combined with recombineering and used as a selection marker in the probiotic *Lactobacillus reuteri* (Oh & van Pijkeren 2014). Other bacteria of interest in which CRISPR has been effective include *Bacillus subtilis*, *Clostridium*, *Corynebacterium glutamicum*, and *E. coli* (Altenbuchner 2016, Cleto et al. 2016, Jiang et al. 2015, Nagaraju et al. 2016). CRISPR applications in other microbes such as yeasts and fungi have also begun. Proof of concept has been shown in *Candida albicans*, *Saccharomyces cerevisiae*, and *Ustilago maydis* (Biot-Pelletier & Martin 2016, Schuster et al. 2016, Vyas et al. 2015). One study utilized wine yeasts and targeted urea production to limit the potential for carcinogen production in the fermentation process (Vigentini et al. 2017). From a genomic standpoint, fungi have been notoriously difficult to work with and manipulate. But recent proof of concept studies in *Aspergillus*, *Myceliophthora thermophila*, and *Penicillium chrysogenum* show promise for the future (Nødvig et al. 2015, Pohl et al. 2016, Q. Liu et al. 2017).

Company Stakes

A clear indicator of the growing interest in CRISPR in the microbial field is the increasing investment by companies. Several companies have been begun incorporating CRISPR to enhance their microbial products. These companies are using various systems and techniques to improve food products, study the food-chain microbiome, and utilize canonical genome engineering as an alternative to antibiotics. DuPont has been using CRISPR for typing and phage protection in dairy strains for more than a decade (Barrangou et al. 2007). AgBiome recently launched a spin-off company called LifeEdit, whose purpose is to study microbes, specifically soil microbes, and their CRISPR systems to improve agricultural business (Teater 2018). Novozymes is using CRISPR to edit fungi (Nødvig et al. 2018). Locus Biosciences is taking a unique approach by repurposing CRISPR as an antimicrobial agent with the Type I exonuclease Cas3 against *Clostridium difficile*, *Pseudomonas aeruginosa*, and *Enterobacteriaceae* (Teater 2018). These are just a few of the many companies currently investing in CRISPR in the microbiological arena.

WELCOME TO THE CRISPR WORLD

With the promise of new research avenues, faster product development, and solutions to global issues, both industry and academia are embracing CRISPR technology. Studies have shown that CRISPR is effective in plants, animals, yeast, fungi, and bacteria (**Figure 3**). It is no longer a matter of if CRISPR products will be developed but when. Some products, such as nonbrowning apples and faster-growing salmon, have already been released in certain markets (Van Eenennaam 2018, Waltz 2018). Other products, e.g., CRISPR-modified pork and crops such as corn and soybeans to name a few, have begun commercial development research and within two years will be ready. Because of their many advantages, CRISPR technologies have allowed agriculture to generate next-generation products much faster than government regulatory policies can be put in place. Because

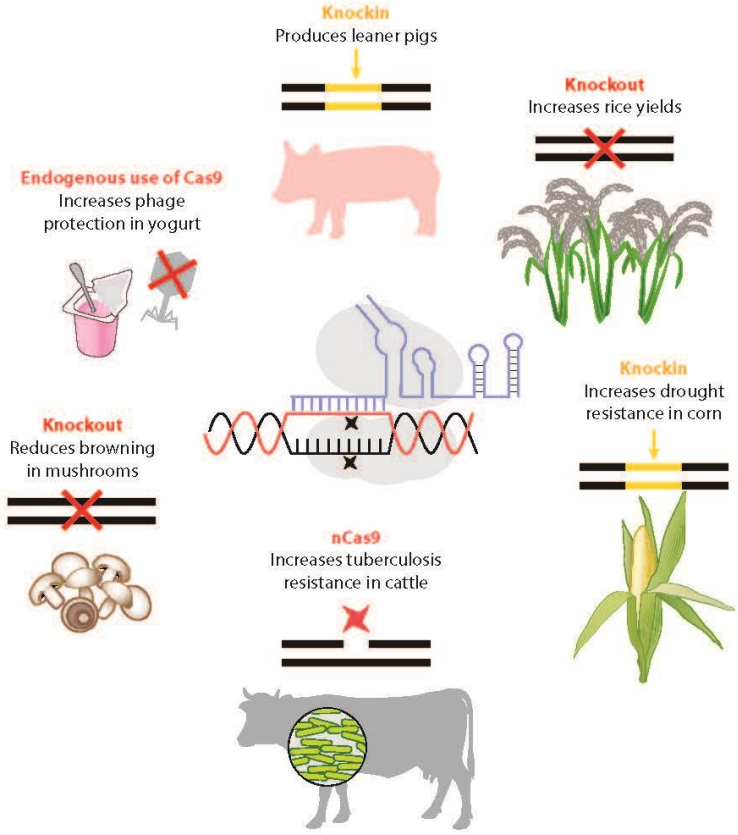


Figure 3

The CRISPR (clustered regularly interspaced short palindromic repeats) farm. CRISPR-Cas9 technology has been utilized in many ways to edit multiple target organisms. Starting from the top of the figure and moving clockwise, there has been a knockin to produce leaner pigs, a knockout to increase rice yields, a knockin to increase drought resistance in corn, nCas9 to increase tuberculosis resistance in cattle, a knockout to reduce browning in mushrooms, and, finally, endogenous use of Cas9 to increase phage protection in yogurt.

of the many CRISPR products entering, or close to entering, the marketplace, a heavy emphasis has been placed on regulation in the CRISPR food arena (Van Eenennaam 2018, Waltz 2018, Wolt et al. 2016). Integral to the conversation is CRISPR's status, or nonstatus, as a GMO. As it currently stands, GMOs are most often defined on the basis of an alteration to an organism that results in the presence of foreign DNA from another species, yielding a transgenic organism. Defined as such, CRISPR technically can readily create non-GMO products. This can be achieved through either CRISPR delivery or novel CRISPR applications. In terms of delivery, the use of RNPs allows DNA-free use of CRISPR editing by solely providing a Cas9 protein loaded with a guide RNA, which is used to cut DNA and prompt the natural repair mechanisms to generate mutations naturally. In terms of application, utilizing CRISPR as a screening tool to identify natural rare mutants in the population allows for new product development without any editing at all. It has become abundantly clear that regulatory policies currently cannot keep up with the CRISPR revolution.

In response to the CRISPR craze, both the United States and the European Union are set to make policy decisions on gene-edited crops and animals this year considering the new technological advances. There are two general approaches for GMO regulation (Araki & Ishii 2015). The first regulation route is product based. In this situation, the end product is under scrutiny. If no foreign DNA is present in the final product, it is not subjected to regulation. The USDA announced earlier this year that this was the route it would use. It stated that if the crop could have been produced naturally, the USDA will not regulate the product, independent of methodology (USDA 2018a). This opens the door for any genetically engineered crop, whether it was developed through CRISPR, ZFNs, TALENs, or other means. The second regulatory route is process based, which is commonly used in the EU. In this situation, the methodology of how a new product is developed is under scrutiny. This approach has also been adopted by the FDA in regard to genetically engineered animals (Van Eenennaam 2018). Methodology-based regulation is considered more conservative than product-based regulation. An alternative regulatory system has recently been proposed in Norway. This process would engage a tiered system. It is hoped that this will make understanding regulations easier and allow products to be regulated on the basis of how much was changed instead of a blanket regulation. This looks to be particularly promising for aquaculture (Fletcher 2017). Recently, the European Court of Justice ruled that edited crops would fall under the existing GMO framework, creating a regulatory burden for commercialization and a challenging route for global companies to manage product approval and regulation differently across regions (Stokstad 2018). Overall, companies favor the first regulation route in which the decision is product based, as illustrated by the recent USDA ruling on nonbrowning white button mushrooms and apples and flax with enhanced omega-3 oil (Waltz 2018). This allows easier entry for agriculture businesses to commercialize crops into the marketplace as millions of dollars are trimmed from the process, and the commercialization window is reduced by years. Some countries have begun incorporating a more blended approach, especially when agricultural trade among countries with different regulatory processes is considered (Araki & Ishii 2015, Ramessar et al. 2008).

As CRISPR has entered the world stage, warnings against potential pitfalls about launching CRISPR agricultural products have arisen. Many interested parties are concerned about a potential fallout reminiscent of the public resistance to GMOs. To avoid this, calls have been made to engage the public in the discussion. A recent study in China has shown that a minority supports the notion of CRISPR crops (Cui & Shoemaker 2018). The authors cite that most were hesitant to support CRISPR-edited foods because of a general lack of understanding of the methodology itself as well as a desire to better understand the risks involved. Interestingly, support is greater for human genetic editing for life-ending diseases than for CRISPR crops. A separate study also revealed modest support for CRISPR when applied to life-threatening conditions (McCaughey et al. 2016). It appears to several researchers that consumers wish to better understand the

risk–benefit analysis for CRISPR foods (Cui & Shoemaker 2018, Funk & Kennedy 2016, Ishii & Araki 2016). It has also been determined that consumers are more interested in benefits they can directly measure (such as increased omega-3s) than those that are indirectly measured (such as benefits to the farmers that may lower costs) (Lucht 2015). Complicating the situation is a lack of trust in scientists when it comes to genetically engineered foods (Cui & Shoemaker 2018, Ishii & Araki 2016, Lucht 2015), despite several studies and reports by esteemed academic organizations that these products do not pose a safety threat to human health. Opinion on CRISPR-edited food is also affected by the media, which many feel is negative (Malyska et al. 2016). To combat these issues, the idea of labeling CRISPR-edited foods has been raised. Many cite the need to increase public trust as a reason to accept labeling. It has been shown that when GMO foods are labeled, consumers tend to trust that food company more (Davis 2018). A study in Switzerland has also revealed that when GMO products are labeled and offered alongside non-GMO products, they are more likely to be purchased than when GMO products are offered alone (Lucht 2015). Overall the trends reveal that consumers desire and likely need a selection of product options and a greater understanding of where their food comes from. Those against the labeling of CRISPR foods cite economic issues involved with adding the labels. There is also fear that this will cause consumers to not purchase the products (McFadden 2017). Another concern is the feasibility of creating labels. As the products could be produced through natural means and not genetic editing intervention, there is concern that it will be impossible to validate labels (Araki & Ishii 2015).

Although there are many differing and conflicting opinions on how to best handle and regulate CRISPR products in the food supply chain, it is clear that the public's opinion will play a major role. With that in mind, it is imperative that researchers, companies, and governments work together to educate and disseminate information to the public so they can make informed decisions on their food choices.

CONCLUSIONS

We are living in a CRISPR world and the future is now. From farm to fork, CRISPR is influencing the food supply chain at every level: from starter cultures to crop and livestock improvement. CRISPR is just one example of how microbes are constantly shaping our world. Advances through microbiome and metagenomic studies are beginning to reveal just how much potential lies ahead. This is especially true in the food industry, with bacteria and yeasts widely used in fermentations and manufacturing of many foods and beverages. This also includes the ability to exploit several CRISPR-based technologies in microbiomes that span the farm (e.g., soil microbiome, livestock microbiome, feed microbiome), the manufacturing facilities (e.g., fermentation tanks, processing lines, food safety control points, packaging environments), and the consumer (e.g., oral and gut microbiomes), as along with CRISPR, microbes are farm to fork. Research is addressing how soil microorganisms affect crops and their yield. The microbiomes of livestock affect their health and growth. Finally, human gut health is largely impacted by its microbiome(s), which is often influenced by diet. It is clear that we are far from exhausting the capabilities of CRISPR in the food supply chain and have much further to go in the utilization of CRISPR-based technologies for the manufacturing of healthier and more sustainable food products.

DISCLOSURE STATEMENT

R.B. is an inventor on several patents related to various uses of CRISPR-based technologies and a shareholder of Dow–DuPont, Caribou Biosciences, Intellia Therapeutics, Locus Biosciences, and Inari, companies with business interests in CRISPR-based applications.

ACKNOWLEDGMENTS

K.B. and R.B. acknowledge support from NC State University and the NC Agricultural Foundation. The authors also acknowledge graphical assistance from Sarah Brandt.

LITERATURE CITED

Acquaah G. 2009. *Principles of Plant Genetics and Breeding*. Hoboken, NJ: Wiley-Blackwell

Altenbuchner J. 2016. Editing of the *Bacillus subtilis* genome by the CRISPR-Cas9 System. *Appl. Environ. Microbiol.* 82:5421–27

Andersson AF, Banfield JF. 2008. Virus population dynamics and acquired virus resistance in natural microbial communities. *Science* 320:1047–50

Andersson M, Turesson H, Nicolia A, Fält A-S, Samuelsson M, Hofvander P. 2017. Efficient targeted multiallelic mutagenesis in tetraploid potato (*Solanum tuberosum*) by transient CRISPR-Cas9 expression in protoplasts. *Plant Cell Rep.* 36:117–28

Araki M, Ishii T. 2015. Towards social acceptance of plant breeding by genome editing. *Trends Plant Sci.* 20:145–49

Armin S, Felix W, Jacqueline B, Holger P, David E. 2017. Towards CRISPR/Cas crops: bringing together genomics and genome editing. *New Phytol.* 216:682–98

Barrangou R. 2015. Diversity of CRISPR-Cas immune systems and molecular machines. *Genome Biol.* 16:247

Barrangou R, Coûté-Monvoisin A-C, Stahl B, Chavichvily I, Damange F, et al. 2013. Genomic impact of CRISPR immunization against bacteriophages. *Biochem. Soc. Trans.* 41:1383–91

Barrangou R, Doudna JA. 2016. Applications of CRISPR technologies in research and beyond. *Nat. Biotechnol.* 34:933–41

Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, et al. 2007. CRISPR provides acquired resistance against viruses in prokaryotes. *Science* 315:1709–12

Barrangou R, Horvath P. 2012. CRISPR: new horizons in phage resistance and strain identification. *Annu. Rev. Food Sci. Technol.* 3:143–62

Barrangou R, Horvath P. 2017. A decade of discovery: CRISPR functions and applications. *Nat. Microbiol.* 2:17092

Beisel CL, Gomaa AA, Barrangou R. 2014. A CRISPR design for next-generation antimicrobials. *Genome Biol.* 15:516

Biot-Pelletier D, Martin VJJ. 2016. Seamless site-directed mutagenesis of the *Saccharomyces cerevisiae* genome using CRISPR-Cas9. *J. Biol. Eng.* 10:6

Bomgardner MM. 2017. A new toolbox for better crops. *Chem. Eng. News* 95:30–34

Bortesi L, Fischer R. 2015. The CRISPR/Cas9 system for plant genome editing and beyond. *Biotechnol. Adv.* 33:41–52

Brooks C, Nekrasov V, Lippman ZB, Van Eck J. 2014. Efficient gene editing in tomato in the first generation using the clustered regularly interspaced short palindromic repeats/CRISPR-associated9 system. *Plant Physiol.* 166:1292–97

Brouns SJJ, Jore MM, Lundgren M, Westra ER, Slijkhuis RJH, et al. 2008. Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science* 321:960–64

Carlson DF, Lancto CA, Zang B, Kim E-S, Walton M, et al. 2016. Production of hornless dairy cattle from genome-edited cell lines. *Nat. Biotechnol.* 34:479–81

Carte J, Christopher RT, Smith JT, Olson S, Barrangou R, et al. 2014. The three major types of CRISPR-Cas systems function independently in CRISPR RNA biogenesis in *Streptococcus thermophilus*. *Mol. Microbiol.* 93:98–112

Chen B, Gilbert LA, Cimini BA, Schnitzbauer J, Zhang W, et al. 2013. Dynamic imaging of genomic loci in living human cells by an optimized CRISPR/Cas system. *Cell* 155:1479–91

Cleto S, Jensen JVK, Wendisch VF, Lu TK. 2016. *Corynebacterium glutamicum* metabolic engineering with CRISPR interference (CRISPRi). *ACS Synth. Biol.* 5:375–85

Cong L, Ran FA, Cox D, Lin S, Barretto R, et al. 2013. Multiplex genome engineering using CRISPR/Cas systems. *Science* 339(6121):819–23

Crawley AB, Henriksen ED, Stout E, Brandt K, Barrangou R. 2018. Characterizing the activity of abundant, diverse and active CRISPR-Cas systems in lactobacilli. *Sci. Rep.* 8:11544

Cui K, Shoemaker SP. 2018. Public perception of genetically-modified (GM) food: a Nationwide Chinese Consumer Study. *npj Sci. Food* 2:10

Davis V. 2018. GMO labeling makes public more likely to trust food companies. *Science* 362

Deltcheva E, Chylinski K, Sharma CM, Gonzales K, Chao Y, et al. 2011. CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature* 471:602–7

Deveau H, Barrangou R, Garneau JE, Labonté J, Fremaux C, et al. 2008. Phage response to CRISPR-encoded resistance in *Streptococcus thermophilus*. *J. Bacteriol.* 190:1390–400

Doudna JA, Charpentier E. 2014. The new frontier of genome engineering with CRISPR-Cas9. *Science* 346(6213):1258096

Edvardsen RB, Leininger S, Kleppe L, Skaftnesmo KO, Wargelius A. 2014. Targeted mutagenesis in Atlantic salmon (*Salmo salar* L.) using the CRISPR/Cas9 system induces complete knockout individuals in the F0 generation. *PLOS ONE* 9:e108622

Fletcher R. 2017. Norwegian legislative debate offers hope for gene editing in aquaculture. *The Fish Site*, Dec. 6. **https://thefishsite.com/articles/norwegian-legislative-debate-offers-hope-for-gene-editing-in-aquaculture**

Funk C, Kennedy B. 2016. *The New Food Fights: U.S. Public Divides Over Food Science*. Washington, DC: Pew Research Center

Gaj T, Gersbach CA, Barbas CF. 2013. ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering. *Trends Biotechnol.* 31:397–405

Gao Y, Wu H, Wang Y, Liu X, Chen L, et al. 2017. Single Cas9 nickase induced generation of NRAMP1 knockin cattle with reduced off-target effects. *Genome Biol.* 18:13

Gilbert LA, Horlbeck MA, Adamson B, Villalta JE, Chen Y, et al. 2014. Genome-scale CRISPR-mediated control of gene repression and activation. *Cell* 159:647–61

Gomaa AA, Klumpe HE, Luo ML, Selle K, Barrangou R, Beisel CL. 2014. Programmable removal of bacterial strains by use of genome-targeting CRISPR-Cas systems. *mBio* 5:e00928-13

Gomez MA, Lin ZD, Moll T, Luebbert C, Chauhan RD, et al. 2018. Simultaneous CRISPR/Cas9-mediated editing of cassava eIF4E isoforms nCBP-1 and nCBP-2 reduces cassava brown streak disease symptom severity and incidence. *Plant Biotechnol. J.* In press

Hale CR, Zhao P, Olson S, Duff MO, Graveley BR, et al. 2009. RNA-guided RNA cleavage by a CRISPR RNA-Cas protein complex. *Cell* 139:945–56

Hilton IB, D'Ippolito AM, Vockley CM, Thakore PI, Crawford GE, et al. 2015. Epigenome editing by a CRISPR-Cas9-based acetyltransferase activates genes from promoters and enhancers. *Nat. Biotechnol.* 33:510

Horvath P, Barrangou R. 2010. CRISPR/Cas, the immune system of Bacteria and Archaea. *Science* 327:167–70

Horvath P, Romero DA, Coûté-Monvoisin A-C, Richards M, Deveau H, et al. 2008. Diversity, activity, and evolution of CRISPR loci in *Streptococcus thermophilus*. *J. Bacteriol.* 190:1401–12

Hsu PD, Scott DA, Weinstein JA, Ran FA, Konermann S, et al. 2013. DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat. Biotechnol.* 31:827–32

Ishii T, Araki M. 2016. Consumer acceptance of food crops developed by genome editing. *Plant Cell Rep.* 35:1507–18

Ishino Y, Shinagawa H, Makino K, Amemura M, Nakata A. 1987. Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in *Escherichia coli*, and identification of the gene product. *J. Bacteriol.* 169:5429–33

Jiang W, Zhou H, Bi H, Fromm M, Yang B, Weeks DP. 2013. Demonstration of CRISPR/Cas9/sgRNA-mediated targeted gene modification in *Arabidopsis*, tobacco, sorghum and rice. *Nucleic Acids Res.* 41:e188

Jiang Y, Chen B, Duan C, Sun B, Yang J, Yang S. 2015. Multigene editing in the *Escherichia coli* genome via the CRISPR-Cas9 system. *Appl. Environ. Microbiol.* 81:2506–14

Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E. 2012. A programmable dual-RNA–guided DNA endonuclease in adaptive bacterial immunity. *Science* 337(6096):816–21

Karkute SG, Singh AK, Gupta OP, Singh PM, Singh B. 2017. CRISPR/Cas9 mediated genome engineering for improvement of horticultural crops. *Front. Plant Sci.* 8:1635

Khalil K, Elayat M, Khalifa E, Daghash S, Elaswad A, et al. 2017. Generation of myostatin gene-edited channel catfish (*Ictalurus punctatus*) via zygote injection of CRISPR/Cas9 system. *Sci. Rep.* 7:7301

Khan MHU, Khan SU, Muhammad A, Hu L, Yang Y, Fan C. 2018. Induced mutation and epigenetics modification in plants for crop improvement by targeting CRISPR/Cas9 technology. *J. Cell. Physiol.* 233:4578–94

Koonin EV, Makarova KS, Zhang F. 2017. Diversity, classification and evolution of CRISPR-Cas systems. *Curr. Opin. Microbiol.* 37:67–78

LaManna CM, Barrangou R. 2018. Enabling the rise of a CRISPR world. *CRISPR J.* 1:205–8

Larson MH, Gilbert LA, Wang X, Lim WA, Weissman JS, Qi LS. 2013. CRISPR interference (CRISPRi) for sequence-specific control of gene expression. *Nat. Protoc.* 8:2180

Ledford H. 2015. Salmon is first transgenic animal to win US approval for food. *Nature.* **http://doi.org/10.1038/nature.2015.18838**

Liu J, Zhou Y, Qi X, Chen J, Chen W, et al. 2017. CRISPR/Cas9 in zebrafish: an efficient combination for human genetic diseases modeling. *Hum. Genet.* 136:1–12

Liu Q, Gao R, Li J, Lin L, Zhao J, et al. 2017. Development of a genome-editing CRISPR/Cas9 system in thermophilic fungal *Myceliophthora* species and its application to hyper-cellulase production strain engineering. *Biotechnol. Biofuels* 10:1

Liu SJ, Horlbeck MA, Cho SW, Birk HS, Malatesta M, et al. 2017. CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells. *Science* 355(6320):aah7111

Lucht JM. 2015. Public acceptance of plant biotechnology and GM crops. *Viruses* 7:4254–81

Ma X, Mau M, Sharbel TF. 2018. Genome editing for global food security. *Trends Biotechnol.* 36:123–27

Makarova KS, Aravind L, Wolf YI, Koonin EV. 2011a. Unification of Cas protein families and a simple scenario for the origin and evolution of CRISPR-Cas systems. *Biol. Direct* 6:38

Makarova KS, Haft DH, Barrangou R, Brouns SJJ, Charpentier E, et al. 2011b. Evolution and classification of the CRISPR–Cas systems. *Nat. Rev. Microbiol.* 9:467–77

Makarova KS, Wolf YI, Alkhnbashi OS, Costa F, Shah SA, et al. 2015. An updated evolutionary classification of CRISPR–Cas systems. *Nat. Rev. Microbiol.* 13:722–36

Mali P, Yang L, Esvelt KM, Aach J, Guell M, et al. 2013. RNA-guided human genome engineering via Cas9. *Science* 339:823–26

Malyska A, Bolla R, Twardowski T. 2016. The role of public opinion in shaping trajectories of agricultural biotechnology. *Trends Biotechnol.* 34:530–34

Marraffini LA, Sontheimer EJ. 2008. CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. *Science* 322:1843–45

Marraffini LA, Sontheimer EJ. 2010. Self versus non-self discrimination during CRISPR RNA-directed immunity. *Nature* 463:568

Maurer A. 2017. *Syngenta acquires non-exclusive license to use CRISPR-Cas9 gene-editing for agriculture applications.* Press Release, Novemb. 3. **https://www.ncbiotech.org/news/syngenta-acquires-non-exclusive-license-use-crispr-cas9-gene-editing-agriculture-applications**

McCaughey T, Sanfilippo PG, Gooden GE, Budden DM, Fan L, et al. 2016. A global social media survey of attitudes to human genome editing. *Cell Stem Cell* 18:569–72

McFadden BR. 2017. The unknowns and possible implications of mandatory labeling. *Trends Biotechnol.* 35:1–3

Miao C, Xiao L, Hua K, Zou C, Zhao Y, et al. 2018. Mutations in a subfamily of abscisic acid receptor genes promote rice growth and productivity. *PNAS* 115(23):6058–63

Nagaraju S, Davies NK, Walker DJF, Köpke M, Simpson SD. 2016. Genome editing of *Clostridium autoethanogenum* using CRISPR/Cas9. *Biotechnol. Biofuels* 9:219

Nekrasov V, Wang C, Win J, Lanz C, Weigel D, Kamoun S. 2017. Rapid generation of a transgene-free powdery mildew resistant tomato by genome deletion. *Sci. Rep.* 7:482

Nishitani C, Hirai N, Komori S, Wada M, Okada K, et al. 2016. Efficient genome editing in apple using a CRISPR/Cas9 system. *Sci. Rep.* 6:31481

Nødvig CS, Hoof JB, Kogle ME, Jarczynska ZD, Lehmbeck J, et al. 2018. Efficient oligo nucleotide mediated CRISPR-Cas9 gene editing in Aspergilli. *Fungal Genet. Biol.* 115:78–89

Nødvig CS, Nielsen JB, Kogle ME, Mortensen UH. 2015. A CRISPR-Cas9 system for genetic engineering of filamentous fungi. *PLOS ONE* 10:e0133085

Oh J-H, van Pijkeren J-P. 2014. CRISPR–Cas9-assisted recombineering in *Lactobacillus reuteri*. *Nucleic Acids Res.* 42:e131

Oishi I, Yoshii K, Miyahara D, Kagami H, Tagami T. 2016. Targeted mutagenesis in chicken using CRISPR/Cas9 system. *Sci. Rep.* 6:23980

Pennisi E. 2013. The CRISPR craze. *Science* 341:833–36

Pohl C, Kiel JAKW, Driessen AJM, Bovenberg RAL, Nygård Y. 2016. CRISPR/Cas9 based genome editing of *Penicillium chrysogenum*. *ACS Synth. Biol.* 5:754–64

Ramessar K, Capell T, Twyman RM, Quemada H, Christou P. 2008. Trace and traceability—a call for regulatory harmony. *Nat. Biotechnol.* 26:975–78

Ray DK, Mueller ND, West PC, Foley JA. 2013. Yield trends are insufficient to double global crop production by 2050. *PLOS ONE* 8:e66428

Rosenblum A. 2018. Meet the woman using CRISPR to breed all-male "terminator cattle." *MIT Technol. Rev.* **https://www.technologyreview.com/s/609699/meet-the-woman-using-crispr-to-breed-all-male-terminator-cattle/**

Schuster M, Schweizer G, Reissmann S, Kahmann R. 2016. Genome editing in *Ustilago maydis* using the CRISPR–Cas system. *Fungal Genet. Biol.* 89:3–9

Selle K, Barrangou R. 2015. CRISPR-based technologies and the future of food science. *J. Food Sci.* 80:R2367–72

Selle K, Klaenhammer TR, Barrangou R. 2015. CRISPR-based screening of genomic island excision events in bacteria. *PNAS* 112:8076–81

Shan Q, Wang Y, Li J, Zhang Y, Chen K, et al. 2013. Targeted genome modification of crop plants using a CRISPR-Cas system. *Nat. Biotechnol.* 31:686–88

Shi J, Guo H, Wang H, Lafitte HR, Archibald RL, et al. 2017. ARGOS8 variants generated by CRISPR-Cas9 improve maize grain yield under field drought stress conditions. *Plant Biotechnol. J.* 15:207–16

Sternberg SH, Redding S, Jinek M, Greene EC, Doudna JA. 2014. DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. *Nature* 507:62–67

Stokstad E. 2018. European court ruling raises hurdles for CRISPR crops. *Science.* **https://doi.org/10.1126/science.aau8986**

Teater B. 2018. *Five company success stories highlighted at CED Conference.* Press Release, Febr. 18. **https://www.ncbiotech.org/news/five-company-success-stories-highlighted-ced-conference**

Telugu BP, Park KE, Park CH. 2017. Genome editing and genetic engineering in livestock for advancing agricultural and biomedical applications. *Mamm. Genome* 28:338–47

Tilman D, Balzer C, Hill J, Befort BL. 2011. Global food demand and the sustainable intensification of agriculture. *PNAS* 108:20260–64

Tsai SQ, Wyvekens N, Khayter C, Foden JA, Thapar V, et al. 2014. Dimeric CRISPR RNA-guided FokI nucleases for highly specific genome editing. *Nat. Biotechnol.* 32:569

Tyson GW, Banfield JF. 2008. Rapidly evolving CRISPRs implicated in acquired resistance of microorganisms to viruses. *Environ. Microbiol.* 10:200–07

Upadhyay SK, Kumar J, Alok A, Tuli R. 2013. RNA-guided genome editing for target gene mutations in wheat. *Genes Genomes Genet.* 3:2233–38

USDA (US Dep. Agric.). 2018a. *Secretary Perdue issues USDA statement on plant breeding innovation.* Press Release 0070.18, March 28. **https://www.usda.gov/media/press-releases/2018/03/28/secretary-perdue-issues-usda-statement-plant-breeding-innovation**

USDA (US Dep. Agric.). 2018b. *Agricultural projections to 2027.* Rep. OCE-2018–1, Interag. Agric. Proj. Comm., Washington, DC

Van Eenennaam AL. 2018. The importance of a novel product risk-based trigger for gene-editing regulation in food animal species. *CRISPR J.* 1:101–06

Vigentini I, Gebbia M, Belotti A, Foschino R, Roth FP. 2017. CRISPR/Cas9 system as a valuable genome editing tool for wine yeasts with application to decrease urea production. *Front. Microbiol.* 8:2194

Vyas VK, Barrasa MI, Fink GR. 2015. A *Candida albicans* CRISPR system permits genetic engineering of essential genes and gene families. *Sci. Adv.* 1(3):e1500248

Waltz E. 2016. Gene-edited CRISPR mushroom escapes US regulation. *Nature* 532:293

Waltz E. 2018. With a free pass, CRISPR-edited plants reach market in record time. *Nat. Biotechnol.* 36:6–7

Wang B. 2015. Disruptive CRISPR gene therapy is 150 times cheaper than zinc fingers and CRISPR is faster and more precise. *Next Big Future.* **https://www.nextbigfuture.com/2015/06/disruptive-crispr-gene-therapy-is-150.html**

Wolt JD, Yang B, Wang K, Spalding MH. 2016. Regulatory aspects of genome-edited crops. *In Vitro Cell. Dev. Biol. Plant* 52:349–53

Yang L, Güell M, Niu D, George H, Lesha E, et al. 2015. Genome-wide inactivation of porcine endogenous retroviruses (PERVs). *Science* 350(6264):1101–4

Yupeng C, Li C, Xiujie L, Chen G, Shi S, et al. 2018. CRISPR/Cas9-mediated targeted mutagenesis of GmFT2a delays flowering time in soya bean. *Plant Biotechnol. J.* 16:176–85

Zheng Q, Lin J, Huang J, Zhang H, Zhang R, et al. 2017. Reconstitution of UCP1 using CRISPR/Cas9 in the white adipose tissue of pigs decreases fat deposition and improves thermogenic capacity. *PNAS* 114:E9474–82

Zoephel J, Randau L. 2013. RNA-Seq analyses reveal CRISPR RNA processing and regulation patterns. *Biochem. Soc. Trans.* 41:1459–63