WHOLE GENOME SEQUENCING

AND COMPARISON OF POTENTIALLY PROBIOTIC

LACTOBACILLUS STRAINS

By

AARON ABRAHAM

Bachelor of Science in Biology

The University of Tulsa

Tulsa, OK

2016

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
MASTER OF SCIENCE
May, 2018

WHOLE GENOME SEQUENCING

AND COMPARISON OF POTENTIALLY PROBIOTIC *LACTOBACILLUS* STRAINS

Thesis Approved:

Dr. Gerwald Koehler

Thesis Adviser

Dr. Rashmi Kaul

Dr. Randall Davis

ACKNOWLEDGEMENTS

I would first like to thank Dr. Gerwald Koehler for taking me on in his lab. At the time I came to him he was unsure he had sufficient space in his lab, but he made it work. Throughout my two years, Dr. Koehler has always kept an open door policy which I appreciate very much. Any time I had questions or needed to discuss something, he would stop what he was doing to help me. Thank you to Dr. Senait Assefa, whose guidance through the various lab techniques and equipment was invaluable. She was always there to help me navigate the often intricate steps for my project. A thank you as well to Dr. Rashmi Kaul and Dr. Randall Davis for being on my committee. Dr. Kaul was my first point of contact when applying to the Master's program, and throughout the program she has always made herself available for any questions I had, whether about school or career opportunities. Similarly, Dr. Davis has always given me great advice when I have gone to him. A general thank you to the OSU-CHS family for creating an open and comfortable environment, and one that fosters collaboration. Thank you to my girlfriend Lauren, who was always there to motivate me and keep me focused on the task at hand. Lastly, thank you to my parents, Abraham and Jyoti, for the support they have given me through the years. I would not be here were it not for all of their love and sacrifices.

Acknowledgements are the view of the author and are not endorsed by Oklahoma State University or the committee members.

Name: AARON ABRAHAM

Date of Degree: MAY, 2018

Title of Study: WHOLE GENOME SEQUENCING AND COMPARISON OF POTENTIALLY PROBIOTIC *LACTOBACILLUS* STRAINS

Major Field: BIOMEDICAL SCIENCES

Abstract: *Lactobacillus* is a well-known and ubiquitous genus of bacteria, many species of which are probiotic in nature. While many lactobacilli have been characterized as being probiotic, the cellular mechanisms and genes exerting beneficial effects on the host have not been fully elucidated. The primary objective of this project is to characterize a potential new species of *Lactobacillus* isolated from prairie voles, with a specific emphasis on probiotic potential. This was assessed by comparative analysis with genomes of known probiotic species of *Lactobacillus* to determine the presence of genes related to probiosis. This work is based on generating whole genome draft sequences of the vole *Lactobacillus* strains. The Nextera XT DNA Library Prep protocol was used in conjunction with the Illumina MiSeq system to sequence the samples. The Qiagen CLC Genomics Workbench was employed for *de novo* assembly of paired sequence reads into sequence contigs. The contig numbers obtained for the *Lactobacillus* strain PV017, PV019, PV025, PV034, PV037, and PV039 genomes were 53, 42, 54, 109, 164, 90, , respectively. Rapid Annotation using Subsystem Technology (RAST) was used to annotate the draft genomes and identify protein-encoding genes related to potential probiotic characteristics such as adhesion, D-alanylation of lipoteichoic acid, and bile hydrolysis. Psi-Blast queries were performed to determine the closest protein matches to those found in the six strains. Across all categories, PV034 had more distinctive results, indicating that is the most unique of the six strains. Potential directions for future studies include the improvement of the genome assemblies by using long-read sequencing across repetitive regions and functional characterization of probiotic candidate genes *in vitro* and in prairie voles to confirm probiotic effects.

TABLE OF CONTENTS

Chapter Page

I. INTRODUCTION 1

II. REVIEW OF LITERATURE...4

Next-Generation Sequencing Platforms 4
Genes Related to Probiotic Efficacy 5
D-Alanylation of Lipoteichoic Acids...6
Bile Hydrolysis 7
Adhesion 8
Origin of Strains..9

III. MATERIALS AND METHODS...11

Strains and Culture Conditions 11
Genomic DNA Isolation 11
Sequencing Library Construction 13
Tagmentation of Input DNA...13
PCR Amplification and PCR Amplicon Clean-up...................................14
Library Normalization 14
Library Quantification, Dilution, and Sequencing...................................15

Chapter Page

IV. RESULTS 20

Extracted Genomic DNA Quality 20
Sequencing Results – Assembly and Annotation 22
RAST Subsystem Assignments 23
Potentially Probiotic Genes Identified by RAST 30
D-Alanylation of Lipoteichoic Acid 31
Bile Hydrolysis 31
Adhesion 32
Presence of Phages 32

V. CONCLUSION 34

REFERENCES 38

LIST OF TABLES

Table Page

1 List of *Lactobacillus* Strains Used in This Study 17
2 List of Nextera XT Kit v2 index 1 (i7) and index 2 (i5) Used for DNA Sequencing Library Construction..18
3 DNA Concentration Quantified with Qubit® 2 Fluorometer 21
5 RAST Genome Information...23
6 RAST Subsystem Breakdown for the Six *Lactobacillus* Strains.............................30
7 Putative phage-derived sequences as detected by PHASTER analysis 33

LIST OF FIGURES

Figure | Page

1 Standard Curve Showing Threshold Cycle as Function of DNA Concentration......19
2 PV017 Gene Subsystem Assignments......24
3 PV019 Gene Subsystem Assignments......25
4 PV025 Gene Subsystem Assignments......26
5 PV034 Gene Subsystem Assignments......27
6 PV037 Gene Subsystem Assignments......28
7 PV039 Gene Subsystem Assignments......29

CHAPTER I

INTRODUCTION

For many centuries, humans across many different cultures have consumed foods that contain lactic acid bacteria. Generally, these are fermented products such as yogurt, kefir, sauerkraut, and beer. Most, if not all, rely on bacteria of the genus *Lactobacillus* for the fermentation process [1]. Beyond their use in food production, lactobacilli are ubiquitous throughout nature, including the human body. Lactobacilli comprise a large part of the commensal microbial community found within us; collectively, these microorganisms are known as the microbiota [2]. This complex community of microbes can be found in many places such as the oral cavity [3], GI tract [3, 4], and female genitourinary tract [5, 6] of the human body; for this thesis, their role within the GI tract is most relevant.

There is a growing body of research that suggests the microbiota in these various locations have myriad effects on the host, with significant contributions to the normal healthy functioning of the gastrointestinal tract. It is involved in many processes including proper digestion and absorption of certain nutrients, pathogen inhibition, promoting host immunity, maintaining the integrity of the epithelial lining, and regulation of host fat storage [7, 8]. This closely intertwined symbiosis did not come to be over night. Thanks to the insights of new technologies, the co-evolution of bacteria of the GI

microbiota and humans can be traced back thousands of years, showing that they have had millennia to develop such a symbiotic relationship [9, 10]. The GI microbiota as a whole is incredibly well adapted to our bodies, complementing and altering our own physiology by carrying out functions that we have not needed to evolve on our own [2, 8]. These high levels of adaptation do pose a challenge in regards to research, as isolates from one organism will not necessarily thrive if transplanted into another. As such, researchers should be careful when drawing conclusions about the impact on human health from experiments done with bacterial strains from non-human organisms.

Probiotics are an important group of bacteria that often compliment the functions of the resident microbial community. The key difference is that probiotics tend to exist transiently within a host instead of permanently. According to the World Health Organization, probiotics are live microorganisms that, when administered in adequate amounts, confer a health benefit to the host [11]. These health benefits come through various processes, only some of which have been elucidated (mentioned above). There are many reasons there is so much interest in probiotics. Whether it is the rise of drug-resistant "super bugs" (combined with the lack of new antibiotics) or the growing threat of a widespread pandemic, governments and researchers are exploring new avenues of treatment for such issues, and probiotics are a prime candidate [4].

While there is general consensus that probiotics have significant potential in the treatment of various diseases, there first needs to be a much deeper understanding of the cellular mechanisms underlying their positive health effects. As research continues in this area, new relationships and mechanisms are constantly being discovered [7]. There are many lactobacilli species that have been characterized as probiotics. For example, *L. johnsonii*

strains have been shown to exhibit characteristics that are beneficial to humans [14], and *L. rhamnosus* GG has been studied extensively and been shown to ameliorate nosocomial diarrhea in children [15].

Despite the ubiquity of these bacteria, only for a fraction of their existence have we been aware of them, and an even smaller fraction of time has seen us begin to appreciate these incredible organisms. Only in the last one or two decades have researchers been able to explore the microbiota and specifically probiotics and their complex relationship with their host. The advent of new technologies, primarily those related to whole genome sequencing and genomic analysis, has given researchers the tools they need to begin to elucidate the complex mechanisms these probiotic bacteria utilize to confer their benefits upon the host [16]. The aim of this study was to identify potentially probiotic genes within six different genomes of strains of *Lactobacillus* isolated from prairie voles [14]. Whole genome sequencing was carried out using the Illumina MiSeq platform; genomes were assembled using the CLC Genomics Workbench, and analyzed using the Rapid Annotation using Subsystem Technology (RAST) platform.

CHAPTER II

REVIEW OF LITERATURE

Next-Generation Sequencing Platforms

Ever since the significance of DNA became widely understood, scientists have grappled with the task of sequencing genomes efficiently. Arguably the biggest undertaking in this regard in recent years was the Human Genome Project, an international and multi-institutional collaboration with one goal: to sequence the human genome. This project cost $2.7 billion over the course of 13 years [17]. Researchers utilized Sanger sequencing, which is considered a first-generation technology. As impressive as this was, it also highlighted the need for better technologies to meet the inevitable increase in demand for genomic sequencing [17]. This demand has led to the development of several platforms that deliver vastly larger volumes of data in much less time [18]. Furthermore, the costs have been drastically reduced; recently, Illumina was one of the first companies to break the sub-$1000 mark for complete sequencing of a human genome, though that cost is not yet commonplace [18, 19]. In the current landscape, there are three main platforms that are widely used: Ion Torrent's PGM, the Illumina MiSeq, and the Pacific Biosciences RS; collectively, these are considered second-generation

platforms [20]. The Illumina platforms utilize sequencing by synthesis, a proprietary process by which single nucleotides can be identified as they are incorporated into a growing DNA strand [21]. Each nucleotide has a unique fluorescent label, allowing for true base by base sequencing. One of the biggest advantages of this technique is it eliminates any issues with handling single nucleotide repeat sequences [22]. Furthermore, the Illumina MiSeq offers the highest throughput per run (meaning it generates the most genomic data) and importantly lower error rates when compared to the Ion Torrent and Pacific Biosciences platforms [21, 23].

Genes Related to Probiotic Efficacy

The main factors when considering probiotic candidacy are host specificity, health benefits, and safety [24]. Host specificity generally refers to adaptations that allow the bacteria to better survive and persist in the host. For example, in the GI tract, a probiotic would need to be able to survive the extremely low pH of the stomach; furthermore, resistance to digestive molecules and enzymes would be imperative as well. Probiotics tend to exist transiently in the GI tract, but the ability to attach to the epithelial lining helps them better persist in the GI tract, as well as increase their ability to inhibit pathogens, a common health benefit of probiotics. As researchers continue to form an understanding of the host-probiotic paradigm, it is important that an exploration of the underlying genes is conducted. Currently this area is lacking, but discussed below are some of the genes that have been identified as being involved with probiotic mechanisms.

D-alanylation of Lipoteichoic Acids

Lactobacillus is a gram-positive genus, meaning its constituents have a thick layer of peptidoglycan, which is composed of a complex mix of proteins, polysaccharides, and teichoic acids [25]. Given the number of molecules to be found in this layer, there can be significant inter-species and inter-strain variation in its composition, which in turn leads to unique properties seen in each strain and/or species [26]. Among the teichoic acids, lipoteichoic acid (LTA) has been linked to multiple probiotic characteristics [26]. D-alanyl ester substitutions are commonly observed in Lactobacilli species, and this incorporation has been shown to affect acid tolerance, adhesion, and resistance to antimicrobial compounds, all of which are important factors for probiotics [27]. These alterations require four proteins which are coded for by the *dlt*ABCD operon: the protein encoded by *dlt*A is Dcl, a carrier protein ligase responsible for ATP-dependent activation of the D-alanine subunit; Dcp is encoded by *dlt*C and is a D-alanine carrier protein [28]. *dlt*B and *dlt*D encode proteins whose roles cannot be stated with complete confidence, though there are a few studies that indicate where *dltD* is involved. Studies have reported that *dltD* is involved in the formation of the D-alanyl: Dcp complex, and the *dlt*B protein is predicted to be involved with membrane transport of the complex [28, 29, 30].

Studies in *Bacillus subtilis* and other Gram-positive bacteria have shown that targeted inactivation of any of these four genes results in significant phenotypic changes stemming from a lack of D-alanyl ester substitutions in LTA [29, 31]. These changes include modulations to autolysin activity, altered resistance to antimicrobial peptides, and modified adhesion and immunomodulation [32, 33, 34]. A study conducted in *Listeria monocytogenes* showed that a *dlt*A knockout mutant had reduced adherence to various

cell lines [33]. Similarly, a study conducted by Vélez, et. al. showed that *dlt*D knockouts of *Lactobacillus rhamnosus* GG (ATCC 53103), a known probiotic strain, have a decreased survival in gastric juices, increased rates of autolysis, and increased sensitivity to human beta-defensin-2; interestingly, however, there were no significant alterations in adhesion to human epithelial cells or immunomodulation in *in vitro* experiments [29]. Although the four genes of the operon encode unique proteins, many studies conducted with various Gram-positive bacteria have illustrated that the end result is the same: a decrease or complete absence of D-alanyl esters in LTA [29].

Bile Hydrolysis

The GI tract is a very harsh environment, full of digestive enzymes and highly acidic gastric juices. Primary bile acids are key digestive compounds that are synthesized in the liver from cholesterol and then conjugated with glycine or taurine via an amide bond; this step is key as it increases the solubility and stability of the final molecule [35]. The ability to hydrolyze bile salts is often a key characteristic included in the criteria for selection of probiotic strains [35]. Bile salt hydrolases (BSHs) belong to the choloylglycine hydrolase family of enzymes, and have been found in multiple bacterial genera, namely *Lactobacillus* [36, 37, 38] and *Bifidobacterium* [39, 40], both of which have many probiotic species. BSHs act by removing the glycine or taurine, resulting in a decrease in bile activity. The overall importance of BSHs is not quite clear. One study showed that five Lactobacilli strains with varying levels of BSH activity all colonized the mouse GI tract equally well [36], while another showed that *Lactobacillus amylovorus* with reduced BSH activity had decreased growth rates in the presence of bile salts [41].

Furthermore, *bsh* gene mutations in *L. monocytogenes* and *L. plantarum* resulted in increased sensitivity to bile salts [38, 42].

The effects of BSHs on the host physiology are also unclear. Many studies have shown that administration of probiotics can significantly lower cholesterol levels in pigs [43, 44]. One hypothesis suggests that these effects can be partially attributed to BSHs, either due to increasing demand for cholesterol for *de novo* bile synthesis (because deconjugated bile salts are excreted at a higher rate) or because lower levels of conjugated bile salts results in less absorption of dietary cholesterol in the GI tract [45, 46].

By deconjugating bile salts, BSHs could also impair the normal host handling of lipids, whether through micelle formation, absorption, or emulsification [47]. In this regard, BSH activity has been associated with growth defects in chickens [48]. Together, these studies highlight the fact that although there is still not a clear consensus on the exact role, mechanisms, or effects of BSHs in the host, they are clearly deserving of further exploration.

Adhesion

A major factor when selecting probiotic bacteria is their ability to adhere to the GI epithelia, as it likely promotes other desirable traits such as inhibiting the adherence of pathogens [49], host immunomodulation, and increased residence time in the GI tract [26, 50]. By doing this, probiotics help strengthen the resident microbiota which are an important component of the GI mucosal barrier.

Adherence is a multifaceted process that has not yet been fully understood. The mechanisms of adherence are still broadly unknown, though studies have pointed to the involvement of many different proteins. The one most relevant for this thesis is fibronectin binding protein. Fibronectin is one of many extracellular-matrix (ECM) proteins found within the protective mucosal layer of the GI tract. It is vital for proper development in vertebrates, with key roles in cell adhesion, migration, growth and differentiation [51]. Given its prevalence in the mucosal layer, it offers a good binding site for bacteria trying to prolong their journey through the GI tract. A study involving 18 strains of *L. acidophilus* and *L. casei* showed that all were able to bind to fibronectin, as well as collagen type IV and fibrin, two other ECM proteins found in the mucosal layer [52]. *L. agilis* has also been shown *in vitro* to localize in areas where fibronectin is present [53]. Pathogenic bacteria also target fibronectin in an attempt to anchor themselves in the GI tract. *Streptococcus pneumoniae*, *S. aureus*, and *E. coli* have all been shown to bind fibronectin well [54, 55].

Origin of Strains

Lactobacillus is a genus of Gram-positive bacteria of the family *Lactobacillaceae*, order Lactobacillales, class Bacilli, phylum Firmicutes. The genus was first characterized by Martinus Beijernick in the early 1900s to describe bacteria he isolated from various fermented foods [56]. The genus is part of the broader group of lactic acid bacteria (LAB), which are characterized by the production of lactic acid as their primary or sole end product of carbohydrate digestion [56]. They are Gram-positive, non-spore forming, non-motile, and require a nutritionally rich environment for growth.

The bacterial strains used in this study were isolated from the GI tract of prairie voles; they were from either the cecum or colon of the animals [14]. In order to isolate lactobacilli, luminal contents of the vole intestines were plated on Lactobacilli MRS (de Mann, Rogosa and Sharpe Medium) agar for enrichment of lactobacilli. This resulted in 30 strains for further analysis. These strains were tested *in vitro* for probiotic activity such as acid and bile resistance, pathogen inhibition, adhesion, and antibiotic susceptibility [14]. Based on the results of these tests, six strains were picked for further genetic analysis in the present study.

CHAPTER III

MATERIALS AND METHODS

Strains and culture conditions

The *Lactobacillus* strains used in this study are shown in Table 1. The procedure and protocols described in previous research were followed for routine maintenance and culturing [14]. Briefly, all bacterial strains were grown from glycerol stocks (stored at -80°C) in de Mann, Rogosa and Sharpe (MRS) agar and broth medium under anaerobic or aerobic conditions. Stocks from individual strains were used to inoculate MRS agar plates that subsequently were incubated under anaerobic growth conditions at 37°C for 48 h. A GasPak™ 100 container and the EZ Anaerobe Pouch system (BD Diagnostics, Sparks, MD) were used to generate anaerobic growth conditions.

Genomic DNA Isolation

For genomic DNA extraction, 3 to 4 isolated colonies from each of the seven strains (Table 1) were inoculated into 50 mL centrifuge tubes containing 45 mL of fresh MRS broth and incubated overnight (~14 h) at 37°C without shaking. Following overnight incubation, bacterial cells were harvested by centrifugation at 4°C for 10 min at 4500×*g*.

Genomic DNA was isolated using a ZR Fungal/Bacterial DNA MiniPrep kit (Zymo Research, Irvine CA, USA) following the manufacturer's instructions. In brief, the harvested cells were lysed in 750µl of ZR lysis buffer and bashed in ZR bashing tubes (Zymo Research, Irvine CA, USA) using a Mini Beadbeater 96 (Biospec Products, Bartlesville, Oklahoma, USA). After disruption the homogenates were processed for genomic DNA isolation following the manufacturer's instructions (Zymo Research, Irvine CA, USA). Depending on the pellet size, in some strains, duplicate or triplicate columns were used to increase DNA yield without exceeding column (Zymo-Spin™ IV, and Zymo-Spin™ IIC, Zymo Research) capacities. Accordingly, purified DNA extracts from each column were eluted, quantitated and visualized separately. After isolation, 2µL of the purified *Lactobacillus* DNA was used for DNA quantification. DNA concentrations were determined by 260 nm and 280 nm readings using a Take 3 Micro-volume plate in a BioTek Synergy 2 Multimode Microplate Reader (BioTek Instruments, Inc. Winooski, Vermont). To assess DNA sample purity, 260nm/280nm ratios were calculated for each sample. The integrity of the extracted genomic DNA was evaluated by agarose (1% w/v) gel electrophoresis in 1× Tris-Acetate-EDTA buffer. A high molecular weight DNA band (>10 kb) was indicative of good-quality genomic DNA. TriDye™ 2-log DNA Ladder (0.1 – 10.0 kb; New England BioLabs Inc., Ipswich, Massachusetts) was used for size estimation.

To accurately ascertain double-stranded (ds) DNA concentrations, the isolated DNAs were quantified with a Qubit® 2 Fluorometer (Invitrogen, Life Technologies, Carlsbad, CA) in conjunction with the Qubit® dsDNA HS (high sensitivity) Assay Kit (Invitrogen, Thermo Fisher Scientific, Carlsbad , CA). Subsequently, DNA samples were diluted with

molecular grade water to the optimal concentration of 1 ng in 5µl for sequencing library construction.

Sequencing Library Construction

Genomic DNA libraries were prepared using the Nextera XT DNA Library Preparation Kit (Illumina San Diego, CA) following the manufacturer's instructions with the following exceptions: for fragmenting and tagging, and amplification steps, capped 0.2 ml PCR tubes were used instead of a 96-well TCY plate NTA (Nextera XT Tagment XT Tagment Amplicon Plate). Rather than the 96-well version of the protocol (CAA, Clean Amplified Plate; SGP; StoraGe Plate; LNP, Library Normalization Plate, low binding Eppendorf tubes (1.5 ml) were used. The main reasons for replacing 96 well plates with tubes was the low sample number and the unavailability of a magnetic stand for 96 well plates required for cleaning (PCR products), washing (beads) and normalizing (library) steps. The available magnetic stand was DynaMag™-2 (Life Technologies, Eugene, Oregon) designed for holding Eppendorf tubes (1.5 mL).

Tagmentation of Input DNA

As per the manufacturer's instructions, 1 ng of the input DNA was tagmented (fragmented and tagged with adapter sequences) by the transposome (Nextera XT transposome simultaneously fragments the input DNA and adds adapter sequences to the ends) on a PTC 200 DNA Engine thermocycler (Bio-Rad, Hercules, CA) with pre-heated lid and incubated at 55°C for 5 min. The reaction was held at 10°C and immediately neutralized with NT (Neutralize Tagment Buffer) followed by a 5 min incubation period at room temperature.

PCR Amplification and PCR Amplicon Clean-Up

In this step, the transposome-tagmented DNA was amplified by PCR via a limited PCR cycle approach (adapter sequence uses a limited PCR reaction to amplify the inserted DNA). Briefly, 15 µl of the Nextera PCR Master Mix (NPM), and 5 µl of each of the index 1 (i7) and index 2 (i5) primers (see Table Primers) were added into each individual tube containing transposome-tagmented DNA and then amplified by PCR to amplify the adapter-flanked DNA fragments. Limited cycle PCR conditions consisted of 1 cycle of 72°C for 3 min and 95°C for 30 s; 12 cycles of 95°C for 10 s, 55°C for 30s,72°C for 30s, and then 1 cycle of 72°C for 5 min followed by a hold at 4°C. Following amplification, the reaction was cleaned-up and purified by adding 90 µl (per reaction tube) of Agencourt AMPure XP magnetic beads (Agencourt, Beckman Coulter, Inc., Indianapolis, IN). The magnetic beads were washed twice with freshly prepared 80% (v/v) ethanol and re-suspended in 52.5 µl of the resuspension buffer (Illumina, San Diego, CA). After further mixing and incubating processes the supernatants were cleared and transferred (50 µl) into clean low binding Eppendorf tubes. The eluted supernatants were stored at -20° up to 24 hrs.

Library normalization

The Nextera XT DNA Library preparation kit includes bead-based sample normalization steps prior to cluster generation and sequencing. As a result, 20 µl of the eluted supernatants were purified using a mixture of library normalization beads (LNB1) and library normalization additives (LNA1) following manufacturer's instructions (Illumina, San Diego, CA). Forty five µl of the combined mixture of LNB1/LNA1 (45µl) and 20 µl of the libraries were incubated on a shaker (at 1200 rpm) for 30 min for library

normalization. Following normalizations the tubes were placed on a DynaMag™-2 (Life Technologies) for 2 min or until the supernatants had cleared. For washing the beads, 45 µl of LNW1 (Library Normalization Wash 1) was added to each sample tube and further incubated on a shaker (at 1200 rpm) for 5 min. Thereafter, the tubes again were placed on the DynaMag™-2 magnetic stand for 2 min or until the supernatants had cleared. The washing step was repeated one more time. Finally, each normalized library was eluted with 30µl of 0.1 N NaOH and incubated on a shaker (at 1200 rpm) for 5 min. Following this incubation, 30 µl of library normalization storage buffer I (LNS1) were added to each tube. To ensure all samples in the tubes were completely re-suspended, eluted samples were incubated for 5 min. Samples were well mixed by gentle pipetting up and down to re-suspend the beads with another short period of incubation (5 min shaking at 1200 rpm). Following incubation, the tubes were placed on the DynaMag™-2 magnetic stand for 2 min or until the supernatants had cleared. From each sample, 30µl of the clarified supernatant was transferred to a clean Eppendorf tube and stored at -20° up to 48 hrs.

Library quantification, dilution, and sequencing

Because the Nextera XT DNA Library Preparation Kit uses bead-based sample normalization, library quantification was not essential, but we chose to quantify the normalized DNA to confirm the final library concentration. A pooled library was prepared by pooling equal volumes of the individual normalized libraries (5 µl of the normalized libraries from each of the 6 strains) from each tube, and then quantified by quantitative PCR using a KAPA Library Quantification Kit (qPCR assay) for Illumina platforms (Kappa Biosystems, Boston, MA). DNA library concentrations for the NGS

was then generated from the standard curve (Figure 1) and adjusted to 1.4 nM [57]. Subsequently, the quantified library (1.4 nM) was diluted with ice cold Hybridization Buffer (Illumina, Hayward, CA, USA) to a final concentration of 4 pM. Also, Illumina PhiX control (PhiX control v3) library (10 nM) was denatured in 0.2N NaOH (fresh) and diluted to a final concentration of 4 pM with ice cold Hybridization Buffer before being loaded onto the V2 MiSeq Illumina reagent cartridge -500 cycles-PE (M85888397-500V2) sequencing Kit on the MiSeq platform (Illumina, Hayward, CA, USA). While genomic sequencing did not rely on 16S rRNA gene amplicon sequencing, in the current study we also conducted to assess the role of estrogen on microbiome composition and diversity using the 16S rRNA primers covering the V4 hyper variable region. The approach was designed to integrate the sequence of the specific Illumina multiplexing sequencing primers and dual-index-paired-end approaches. With this approach since it was possible to run up to 384 samples on a single flow cell (single run) we decided to pool libraries from both the 16S rRNA and genomic libraries generated by 16S rRNA primers and Nextera XT technologies, respectively, and loaded on to the same reagent cartridge and run for sequencing. Although sequencing of a whole genome library does not necessarily require the addition of PhiX control library, for low diversity samples (such as low complexity amplicon pools from 16S rRNA) adding as little as 5% PhiX DNA (Illumina Technical Support Note) provides balanced signals at each cycle to improve the overall run and data quality. Accordingly, a 10% PhiX control library was added to the mix. For a 10 % PhiX run, 900 µl of 4 pM library (450 µl each from 16S r RNA and genomic libraries) and 100 µl of 4 pM of PhiX were combined in a low binding Eppendorf tube and 600 µl of the combined library/PhiX solution was then loaded into

well number 17 on the reagent cartridge for clustering and sequencing on a MiSeq instrument. In addition, 3 µl (at 100 µM concentration) from each of the read 1 sequencing primer for V4 region (read 1 forward for V4 region), index primer for V4 region and read 2 sequencing primer for V4 region (read 2 reverse primer for V4 region) was spiked into wells 12, 13 and 14, respectively. Notably, the combined library/PhiX solution loaded was 4.0 pM overall with 3.6 pM library concentration, 0.4 pM PhiX concentration, and 0.000515N NaOH concentrations. Samples were sequenced using 2 x 250 bp paired-end reads.

Table 1. List of *Lactobacillus* strains used in this study, as isolated by Assefa, *et al* [14]

Species	Strains	Type	Character	Origin
***Lactobacillus* spp.**	PV017	Probiotic strain		Cecum
***Lactobacillus* spp.**	PV018	Probiotic strain		Cecum
***Lactobacillus* spp.**	PV019	Probiotic strain		Cecum
***Lactobacillus* spp.**	PV025	Probiotic strain	*	Colon
***Lactobacillus* spp.**	PV034	Probiotic strain	*	Cecum
***Lactobacillus* spp.**	PV037	Probiotic strain	*	Colon
***Lactobacillus* spp.**	PV039	Probiotic strain		colon

PV, Prairie vole isolates; * Hydrogen peroxide producer

Table 2 List of Nextera XT Kit v2 index 1 (i7) and index 2 (i5) used for DNA Sequencing Library Construction

Strain	i7_ index ID	i7_Index sequence	i5_index ID	i5_index sequence	
PV017	N701	TAAGGCGA	S502	CTCTCTAT	
PV019	N702	CGTACTAG	S502	CTCTCTAT	
PV025	N703	AGGCAGAA	S502	CTCTCTAT	
PV034	N704	TCCTGAGC	S502	CTCTCTAT	
PV037	N705	GGACTCCT	S502	CTCTCTAT	
PV039	N706	TAGGCATG	S502	CTCTCTAT	

PV, Prairie vole isolates

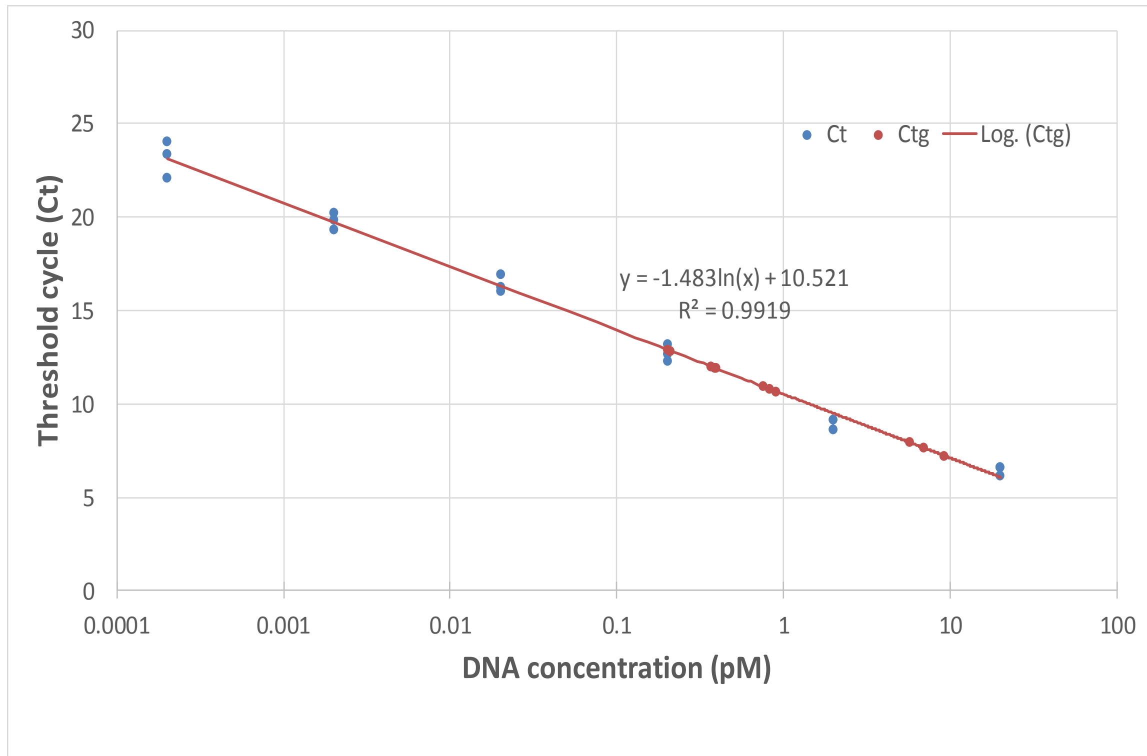

Fig. 1. A standard curve showing threshold cycle (Ct) on the y-axis and the 452 bp KAPA Illumina DNA standard concentrations on the x-axis. Slope (-1.48ln), y-intercept (10.521) and correlation (0.9919) values depicted here were calculated using Excel software to provide information about the performance of the real time reaction. The DNA Library concentration for the NGS was generated from the standard curve by the Applied Biosystems 7500 real time PCR system software. Both standard (0.0002 to 20 pM) and genomic library dilutions (1:1; 1:10;1:100; 1:1000; 1:2000: 1:4000) were assayed in triplicate in a 10 µl reaction containing 4µl of the template DNA, and 6 µl of KAPA SYBR®FASTqPCR Master Mix (20x) using the manufacturer's standard protocols. The cycling conditions used were 95°C for 5 min and 35 cycles at 95°C for 30 s and 60°C for 45 sec. Ct and Ctg; threshold cycle from the standard DNA and genomic DNA Library, respectively. Actual values (average of 3 replicates) are shown below.

Standard Conc. (pM)	0.0002	0.002	0.02	0.2	2	20
Ct	23.81	19.81	16.42	12.74	8.87	6.53
g Conc (pM)	0.209	0.382	0.827	7.32		
Ctg	12.87	11.95	10.81	7.6		

CHAPTER IV

RESULTS

Extracted Genomic DNA Quality

Agarose gel electrophoresis was used for the DNA samples to determine the integrity of the extracted DNA. A band above 10 kb was indicative of good quality; the six strains picked had such bands present (data not shown). Before sequencing library construction was possible, the extracted DNA must be diluted to the appropriate concentration. For the Nextera XT DNA Prep Kit, the manufacturer's optimal concentration is 0.2 ng/µl. The DNA extract concentrations were determined using the Qubit® 2 Fluorometer, and based on those values the samples were diluted to reach the optimal concentration (Table 1).

Table 3: DNA concentration quantified with Qubit® 2 Fluorometer

Strain ID	Assay Conc. ng/ml	Stock Conc. ng/ml	Assay Type	Dilution Factor	Stock Conc. ng/µl	Input DNA Conc. for genomics sequencing	Input DNA (µl) required for 500 µl @ 0.2 ng/µl	Amount H20 added to bring up to 500 µl	Total Volume (µl) @ 0.2 ng/µl
PV017	321	6.42E+04	dsDNA HS	200	64.20	0.2 ng/µl	1.56	498.44	500
PV019	369	7.38E+04	dsDNA HS	200	73.80	0.2 ng/µl	1.36	498.65	500
PV025	352	7.04E+04	dsDNA HS	200	70.40	0.2 ng/µl	1.42	498.58	500
PV034	298	5.96E+04	dsDNA HS	200	59.60	0.2 ng/µl	1.68	498.32	500
PV037	425	8.50E+04	dsDNA HS	200	85.00	0.2 ng/µl	1.18	498.82	500
PV039	343	6.86E+04	dsDNA HS	200	68.50	0.2 ng/µl	1.46	498.54	500

Note: the recommended input DNA at 0.2 ng/µl was prepared in large volume (500 µl) to minimize pipetting errors. ds, double stranded; HS, High Sensitivity; Conc., Concentration.

Sequencing Results – Assembly and Annotation

The six sequencing libraries were sequenced on single MiSeq run and the resulting FastQ files for each sample were assembled in CLC Genomics Workbench using default parameters for paired Illumina sequence reads. The resulting contigs for the PV017, PV019, PV025, PV034, PV037, and PV039 samples were used for annotation using the Rapid Annotation using Subsystem Technology (RAST) online service. Unsurprisingly, all six genomes were relatively similar in genome length. The smallest genome was PV017 at 1,511,831 base pairs and the largest was PV025 at 1,660,462 base pairs; the

remaining genome lengths are found below in Table 1. Another important value to take note of is the number of contigs in each genome. A lower number of contigs is indicative of a more complete assembly; a perfectly assembled genome would be a single contig. The use of short-read sequencing tends to yield a larger number of contigs than long-read sequencing. However, these values are generally not beyond an acceptable range. PV017, PV019, and PV025 had the lowest number of contigs, with 52, 42, and 54, respectively. PV034 and PV037 saw a spike in the contig number, with 109 and 164, respectively; PV039 had 90 (Table 1).

Strain ID	PV017	PV019	PV025	PV034	PV037	PV039
RAST Genome ID	6666666.282908	6666666.282909	6666666.282910	6666666.282911	6666666.282912	6666666.282913
ID	AAgenome1	AAgenome2	AAgenome3	AAgenome4	AAgenome5	AAgenome6
Size	1,511,831	1,512,924	1,660,462	1,540,138	1,655,982	1,664,654
GC Content	33.8	33.8	33.7	34.3	33.8	33.7
N50	61215	75369	48258	26296	14908	26692
L50	10	8	9	17	35	16
Number of Contigs with PEGs[1] [total]	52 [61]	42 [47]	54 [61]	109 [115]	164 [179]	90 [99]
Number of Subsystems	237	238	242	251	242	243
Number of Coding Sequences	1435	1438	1606	1480	1579	1603
Number of RNAs	53	54	53	43	45	55

Table 4: RAST Genome Information

An important metric of assembly quality is N50. When the contigs are arranged based on size, it is the length of the contig C such that 50% of the bases are contained in the contigs of size C or larger. A larger N50 value is beneficial because it signifies a more complete genomic assembly. PV017 and PV019 have the largest N50 values by far, at 61215 and 75369 respectively (Table 1). PV025 sees the first significant decrease to 48258; from there, PV034, PV037, and PV039 have N50s of 26296, 14908, and 26692, respectively (Table 1).

RAST Subsystem Assignments

The genome size for PV017 is 1,511,831 base pairs. The GC content is 33.8%, and RAST identified 1435 coding sequences and 237 subsystems (Table 1); the breakdown of the subsystems is shown below in Figure 1 and Table 2. These subsystem assignments include 47% of the genes identified by RAST (Figure 1).

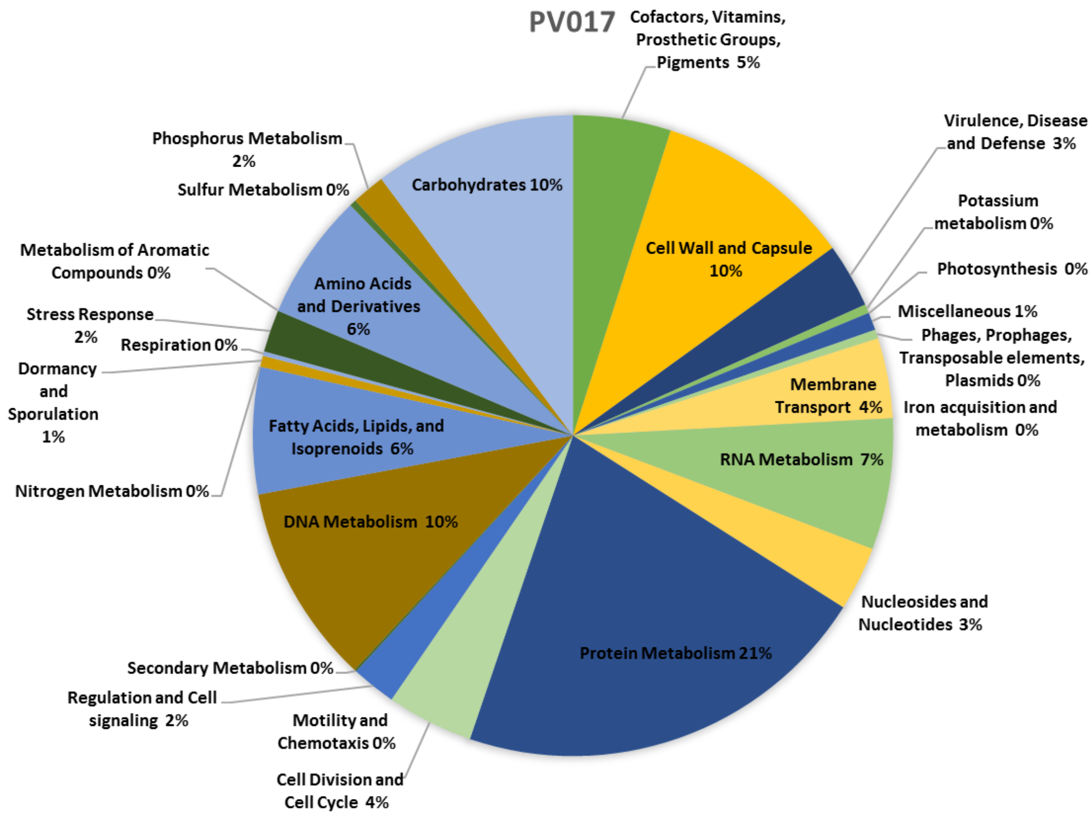

Figure 1: PV017 Gene Subsystem Assignments

The reported genome size for PV019 is 1,512,924 base pairs. RAST identified 1438 coding sequences and 238 subsystems (Table 1); this represents 47% of the discovered features of the genome (Figure 2). The breakdown of the subsystems is shown in Figure 2 and Table 2.

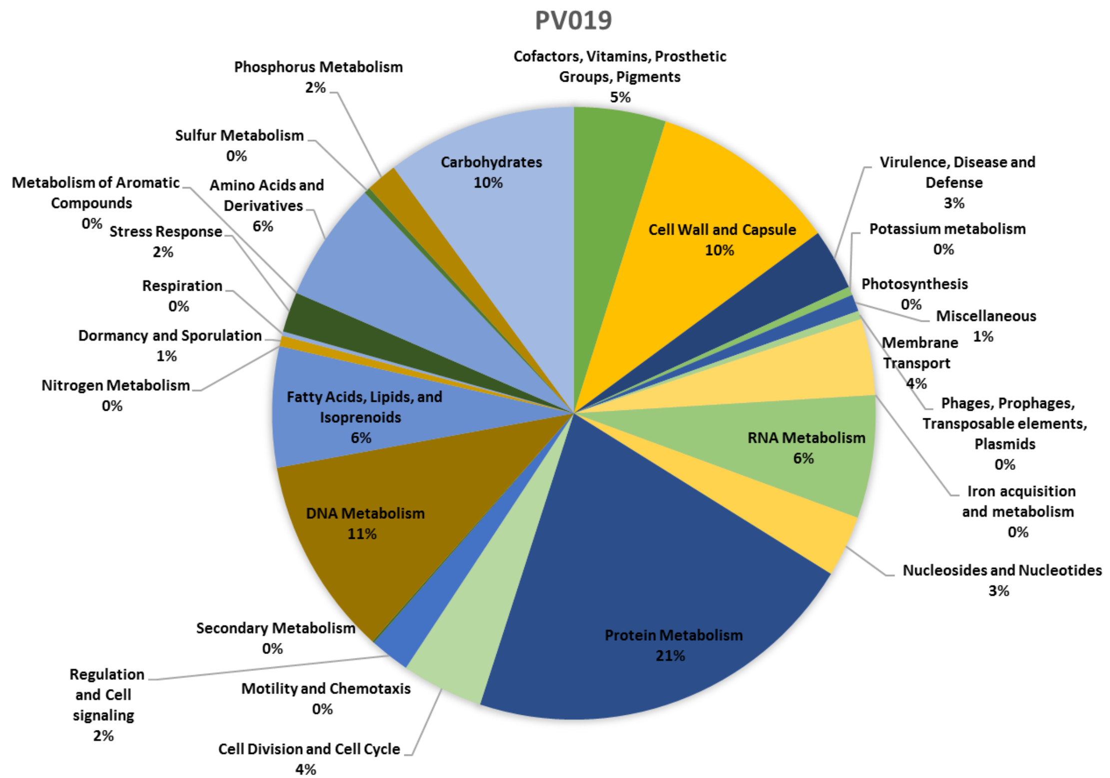

Figure 2: PV019 Gene Subsystem Assignments

Strain PV025 has a 1,660,462 base pair genome. This is the second largest genome, reporting 1606 coding regions and 242 subsystems (Table 1). The subsystem assignments accounts for 43% of the genome as identified by RAST (Figure 3 and Table 2).

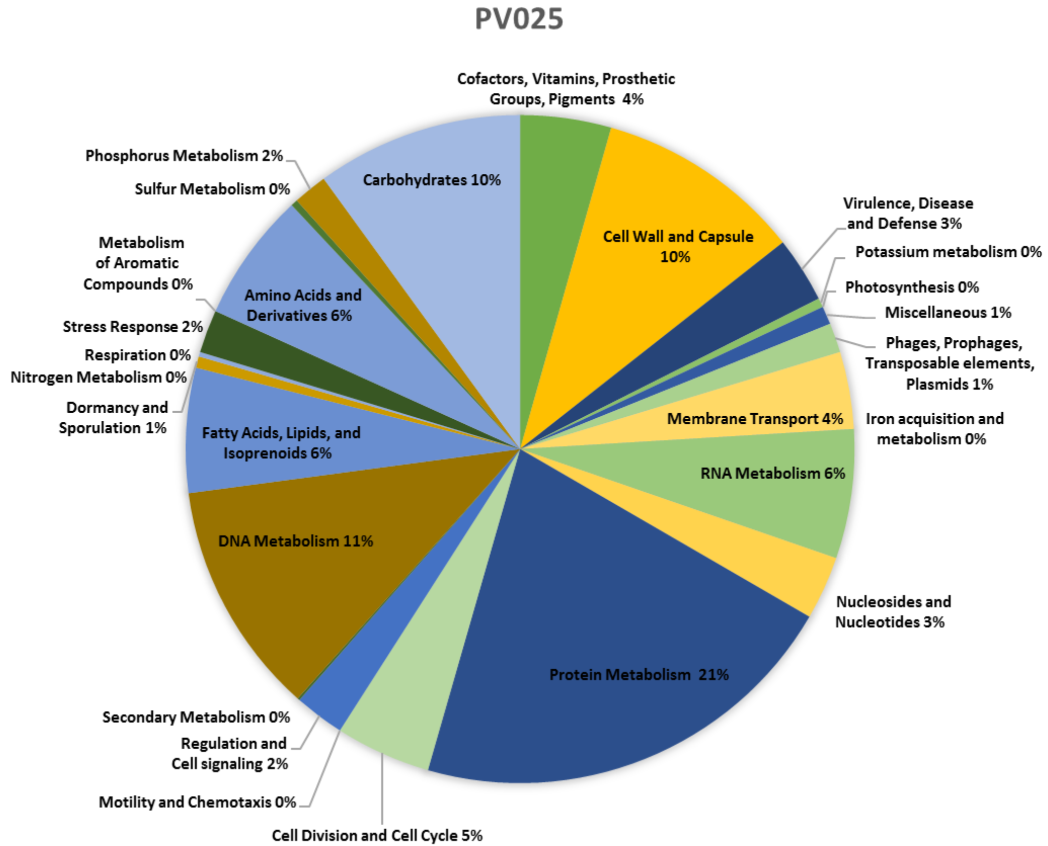

Figure 3: PV025 Gene Subsystem Assignments

The genome of PV034 is 1,540,138 base pairs in size, and it has 34.3% GC content – the highest of the six strains (Table 1). In line with this, RAST identified 252 subsystems (the breakdown can be found in Figure 4 and Table 2), which is also the highest. Interestingly, however, there are only 1480 coding sequences identified, which is on the low end for the six strains (Table 1).

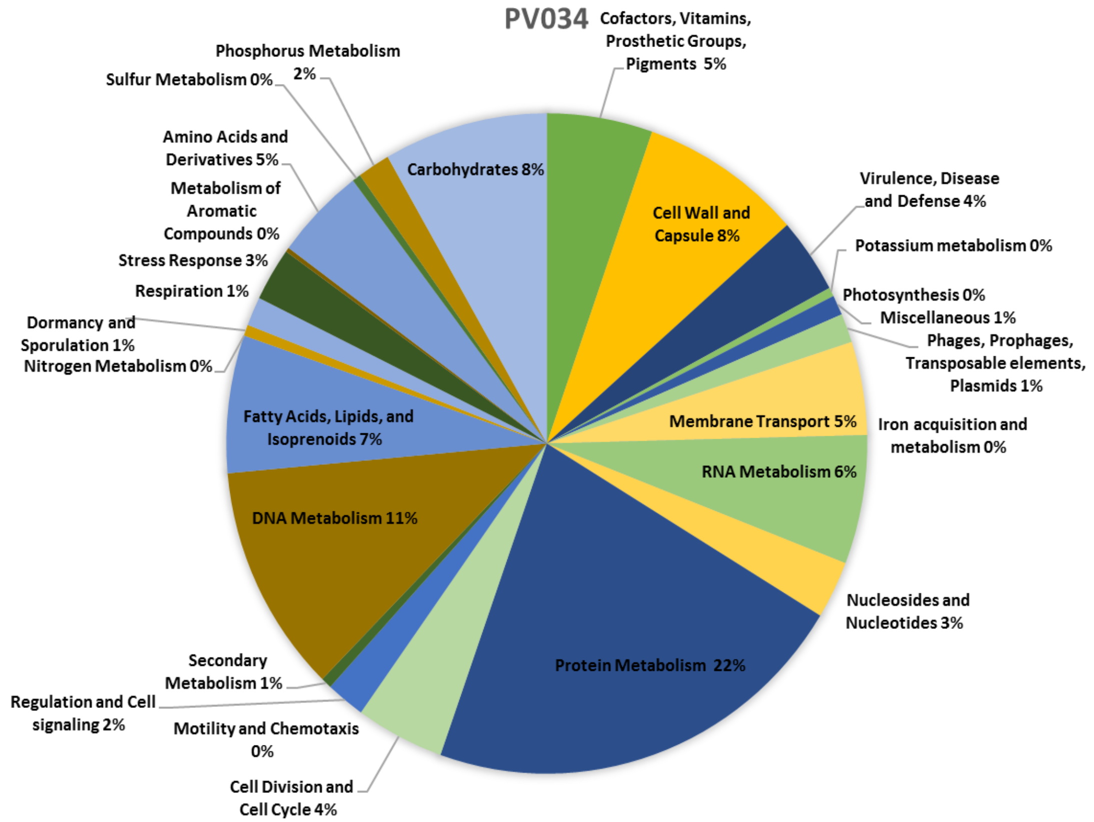

Figure 4: PV034 Gene Subsystem Assignments

The PV037 genome is 1,655,982 base pairs in length, with 33.8% GC content and 1579 coding sequences (Table 1). RAST identified 242 subsystems; 44% of the identified genes were assigned to these subsystems (Figure 5 and Table 2).

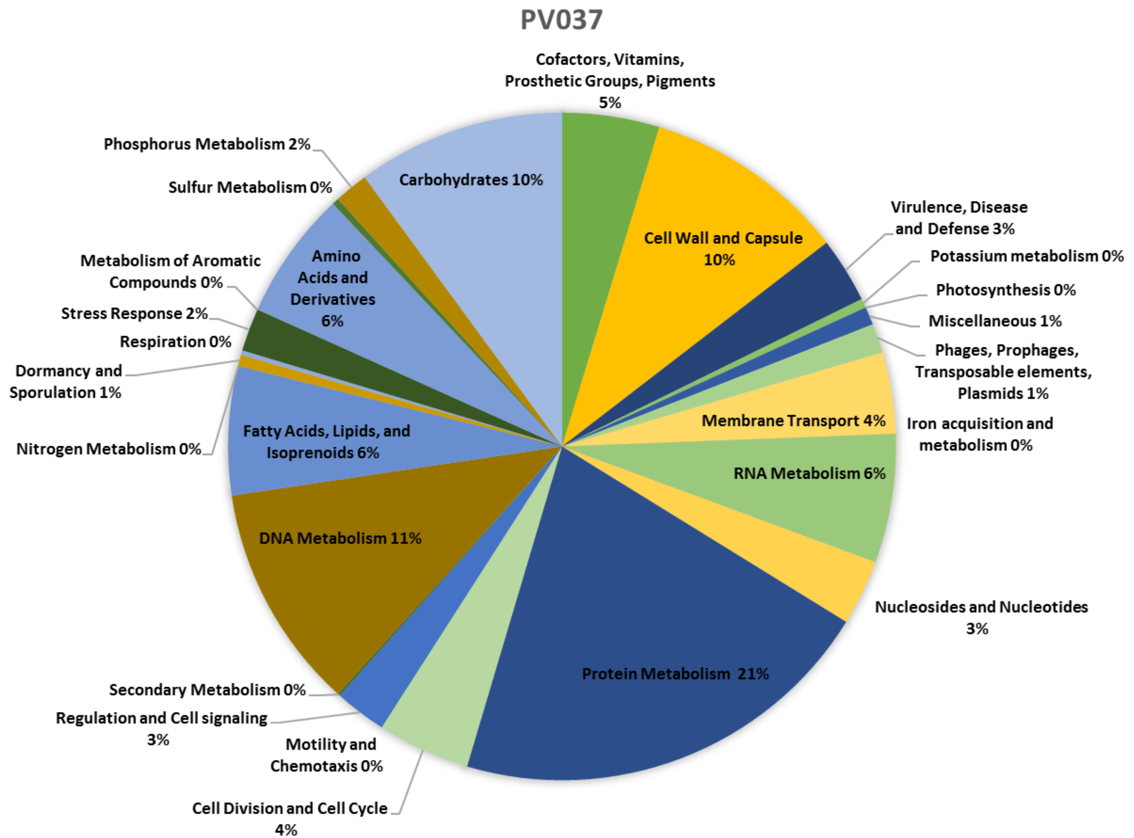

Figure 5: PV037 Gene Subsystem Assignments

At 1,664,654 base pairs, the genome of PV039 is the largest of the six strains. It has 33.7% GC content and 1603 coding sequences. RAST identified 243 subsystems which include 43% of the genes identified. The breakdown of the subsystems can be seen in Figure 6 and Table 2.

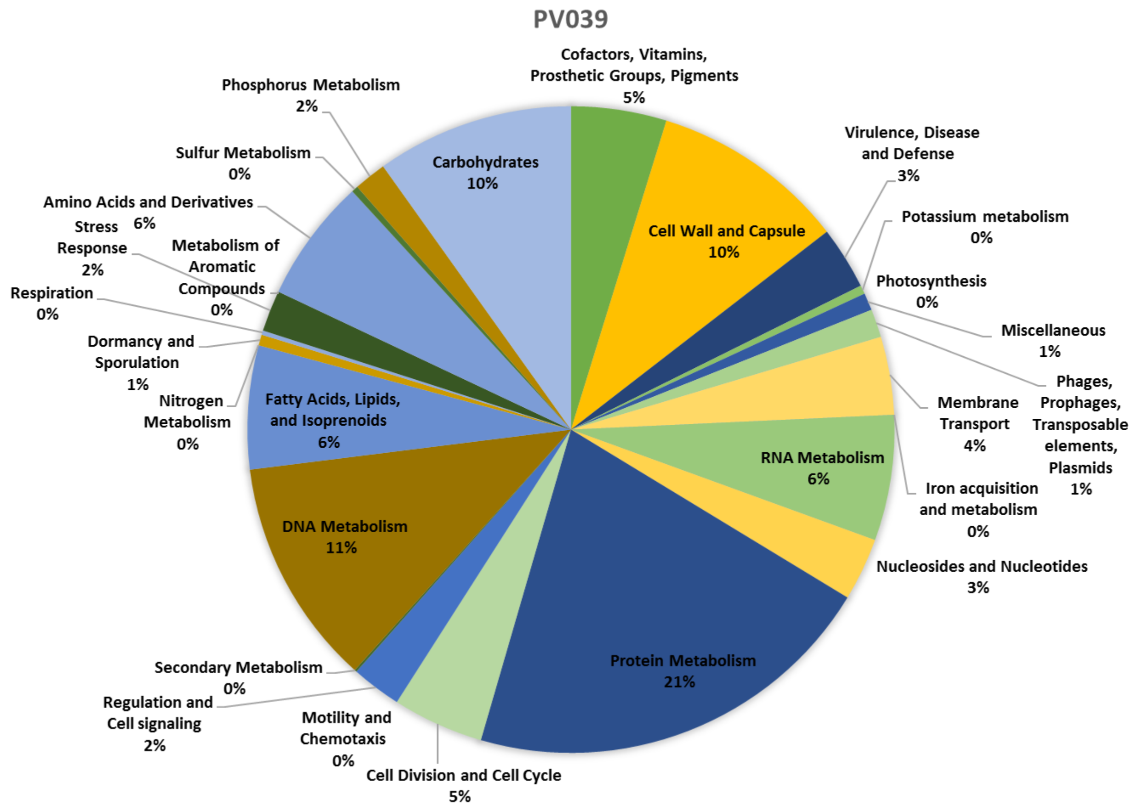

Figure 6: PV039 Gene Subsystem Assignments

	PV017	PV019	PV025	PV034	PV037	PV039
Cofactors, Vitamins, Prosthetic Groups, Pigments	44	44	40	49	43	44
Cell Wall and Capsule	90	90	91	75	90	90
Virulence, Disease and Defense	29	29	29	34	29	29
Potassium metabolism	4	4	4	4	4	4
Photosynthesis	0	0	0	0	0	0
Miscellaneous	8	8	8	9	8	8
Phages, Prophages, Transposable elements, Plasmids	4	4	13	13	13	13
Membrane Transport	36	36	34	42	36	36
Iron acquisition and metabolism	0	0	0	0	0	0
RNA Metabolism	59	58	57	58	57	58
Nucleosides and Nucleotides	29	29	28	26	29	29
Protein Metabolism	189	190	192	199	190	192
Cell Division and Cell Cycle	39	39	42	41	41	42
Motility and Chemotaxis	0	0	0	0	0	0
Regulation and Cell signaling	20	19	22	18	23	23
Secondary Metabolism	1	1	1	5	1	1
DNA Metabolism	90	94	103	103	100	105
Fatty Acids, Lipids, and Isoprenoids	57	57	55	62	57	57
Nitrogen Metabolism	0	0	0	0	0	0
Dormancy and Sporulation	5	5	5	5	5	5
Respiration	2	2	2	13	2	2
Stress Response	19	19	19	24	19	19
Metabolism of Aromatic Compounds	0	0	0	2	0	0
Amino Acids and Derivatives	57	57	57	42	57	57
Sulfur Metabolism	3	3	3	4	3	3
Phosphorus Metabolism	15	15	15	15	15	15
Carbohydrates	91	91	91	76	92	91

Table 5: RAST Subsystem Breakdown for Six Strains

Potentially Probiotic Genes Identified by RAST

The genomes of the six strains were further explored using Psi-Blast as well as PHASTER. Psi-Blast runs selected proteins against the NCBI database of submitted genomes and proteins and compiles matches. The proteins searched were those related to D-alanylation of LTA, bile hydrolysis, and adhesion. For all proteins in each category, PV034 had unique matches not shared by any of the other six strains. Notably, the remaining strains all had the same matches.

D-Alanylation of Lipoteichoic Acid

RAST only identified the *dltB*, *dltD,* and *dltR* genes in all six strains; the absence of the other three genes of the operon, *dltA, B,* and *C*, is peculiar. It is worth exploring this more, as it may be the result of a limitation or error of RAST. In PV034, the DltB protein, putatively labeled as a 406 amino acid D-alanyl transfer protein. A Psi-Blast search identified a 405 amino acid protein in *L. hominis* as the most similar protein with 85% sequence identity. For PV017, PV019, PV025, PV037, and PV039, the DltB protein is the same 406 amino acid protein identified in PV034, but the best match is a 407 amino acid protein from *L. johnsonii*, with 81% sequence identity. The DltD protein in PV034 was characterized as a 429 amino acid protein; at 72% similarity, the best match is a 440 amino acid protein in *L. gasseri*. In the remaining five strains, DltD is much larger at 755 amino acids. The best match at 62% is a 757 amino acid protein in *L. hominis*. In all strains, the DltR regulatory protein is a 222 amino acid protein. In PV034, the best match is a 221 amino acid protein in *L. crispatus*. The sequence identity is 79%. In the other strains, the best match is also 221 amino acids but from *L. gasseri*, and sequence identity is 74%.

Bile Hydrolysis

While there are multiple relevant proteins involved in bile resistance, this paper focuses on choloylglycine hydrolase, a protein involved in the inactivation of primary bile salts. In all strains, RAST identified the protein as having 326 amino acids. The best match in PV034 is a 325 amino acid linear amide C-N hydrolase found in *L. sp. Marseille* –

P3519, with 62% sequence identity. In the remaining strains the best match at 61% similarity is a 309 amino acid protein found in *L. reuteri*.

Adhesion

Fibronectin binding protein was focused on in this thesis. This protein is the only one that yielded the same results for all strains, except for sequence identity, which again was different for PV034. The protein in the six strains was characterized at 564 amino acids in length. The best match for all strains was a 563 amino acid protein from *L. hominis*; in PV034, the sequence identity is 82%, while the remaining strains share 79% identity.

Presence of Phages

Phage Search Tool Enhanced Release (PHASTER; [58]) analysis was used to identify prophage genes or genomes in the *Lactobacillus* genome sequences. Table 4 lists phage-related sequences detected by PHASTER. No phages of were detected in the draft genomes of PV017 and PV019. Intact prophages were detected in the genomes sequences of PV025, PV037, and PV039. PV025 and PV039 harbored additional incomplete phage-related sequences. Two regions in the PV034 genome showed putative remnants of two phages, one of which scored relatively high in the PHASTER analysis (for details see Table 4.)

Strain ID	Region #	Completeness (score) [a]	Most common phage (number of similar proteins) [b]	Genome contig: position [c]	Region Length [kb]	GC content %	Number of proteins [d]
PV017	0						
PV019	0						
PV025	1	Incomplete (10)	PHAGE_Lactob_JCL1032_NC_019456(2)	AAgenome3_S121contig_18_consensus: 15805-23907	8.1	32.1	21
PV025	2	Intact (130)	PHAGE_Lactob_JCL1032_NC_019456(13)	AAgenome3_S121contig_31_consensus: 9353-41160	31.8	34.6	29
PV034	1	Questionable (90)	PHAGE_Lactob_KC5a_NC_007924(23)	AAgenome4_S122contig_24_consensus: 252-23677	23.4	34.6	30
PV034	2	Incomplete (20)	PHAGE_Strept_phiARI0460_1_NC_031913 (8)	AAgenome4_S122contig_41_consensus: 876-16799	15.9	32.8	26
PV037	1	Intact (100)	PHAGE_Lactob_JCL1032_NC_019456(12)	AAgenome5_S123contig_88_consensus: 1202-22989	21.7	34.5	17
PV039	1	Incomplete (10)	PHAGE_Lactob_JCL1032_NC_019456(2)	AAgenome6_S124: 15813-23915	8.1	32.2	21
PV039	2	Intact (130)	PHAGE_Lactob_JCL1032_NC_019456(13)	AAgenome6_S124contig_2_consensus: 7346-39153	31.8	34.3	29

Table 6: Putative phage-derived sequences as detected by PHASTER analysis

CHAPTER V

CONCLUSION

When selecting and studying bacteria for their probiotic potential, it is important to consider not just probiotic factors but also host adaptation factors. Without the ability to survive long enough in a host, a probiotic bacteria cannot exert its positive effects. Host adaptation factors were the primary focus of this study, specifically those related to survival and persistence through the GI tract. All strains have genes related to adhesion, bile resistance, and D-alanylation of LTA, all of which have been shown to be contribute to probiotic potential.

Genomic annotation by RAST showed that all six strains are quite similar. When comparing protein families, all strains except for PV034 showed nearly identical protein presence [Appendix, Figure A1], as well as no variance between the six strains in the presence of the genes and proteins relevant to this study. It is unsurprising to see such high similarity, as these strains have likely had significant time to become highly adapted to their prairie vole host. The RAST subsystem breakdown in Table 2 shows that there is very little variance across all subsystems for five strains. Interestingly, PV034 is quite unique. As can be seen in Figure A1**,** PV034 has substantially lower protein family similarity as well as subsystem assignments [Results, Table 2]. Additionally, PV034 reported unique best match proteins in all cases except in fibronectin binding protein, but

even there the percent identity was slightly higher for PV034. Given this clear distinction, PV034 certainly warrants further attention in future studies.

NGS platforms have greatly reduced the complexity of the task of whole genome sequencing. That said, there is still much room for improvement. For example, while the Illumina MiSeq platform was used in part because of its ability to handle single nucleotide repeat sequences, the genomic assembly quality is still hindered by the short read lengths. Genomes often have repeat sequences, or multiple copies of a gene or set of genes. The sequencer will not be able to identify these elements as unique from one another if the read lengths are shorter than the repeat region. If this occurs, extra contigs will be formed, thus creating a less contiguous assembly. While there are platforms that can better resolve repeat sequences, these long read sequencers also have their own pitfalls. Longer read lengths are more difficult for the sequencer to handle, resulting in less read depth in each base throughout the sequenced fragment. Combining the two techniques is possible, but the cost and complexity of this can be prohibitive for many researchers. In addition to the sequencers themselves, the assembly and annotation software are similarly limited, adding in additional inaccuracies to final genomic assemblies and data. For example, PV017, PV019, and PV025 all have a low number of contigs, while PV034, PV037, and PV039 are much higher although still acceptable. However, this illustrates that none of the assemblies are complete or closed, so there is likely misplaced or missing genomic data. Furthermore, RAST was only able to assign

approximately 50% of discovered genes to subsystems, and such annotation tools may assign genes incorrectly. Although these steps all have flaws, it should not be misunderstood that the current landscape of whole genome sequencing and analysis is vastly better than before the development of NGS platforms, which are continuously being improved.

It is important to note that many of the mechanisms for probiotic action are putative. While new technologies have allowed significant insights into the world of probiotics, mutant analysis studies, which are perhaps the best way to connect genotype to phenotype, are still fairly limited in probiotic Lactobacilli strains [26]. Studies involving gene-knockout mutants offer a more direct path to establishing causal relationships between specific genes and probiotics' positive health effects. In this regard, this study is meant to serve as a catalog, albeit not an exhaustive one, of genes of interest for future studies involving mutant strains. Going forward, there are many other genes and elements worth exploring. For example, the presence of phage elements in all strains but PV017 and PV019 have some amount of phage presence, whether questionable or fully intact. Phages are often considered sources of novel genetic diversity [59]. In this way, it is possible that infection by a phage can imbue a bacteria with very unique and advantageous characteristics, making it a better probiotic candidate. As such, these phage elements identified should be explored further to establish what genes they are carrying.

In regards to native genes and functions, the three focused on in this study are multi-faceted processes that have not been fully understood but certainly involve many proteins. For example, there are many other extracellular matrix proteins in addition to fibronectin present in the mucosal layer of the GI tract that bacteria can target. Similarly,

bile salt hydrolases are not the only proteins relevant to bile resistance. Multidrug efflux pumps have been indicated in this process as well [26, 60], and all six strains have these efflux pumps.

To contextualize this study, it is important to understand that researchers have only uncovered a small fraction of the intricate interplay of the microbiota and probiotics with their hosts. There is not a definitive consensus on what genes or functions are most important for probiotic potential. The six strains studied all have genes that the literature has implicated in probiosis; likely, there are many more that have yet to be identified as important. The results of this study certainly implicate these strains as good probiotic candidates in the prairie vole model. Future studies can use the methods and genes identified here to further characterize these strains and establish more firm connections between the genes and any probiotic effects they may have.

REFERENCES

1. Turpin, W., *et al.*, *Lactobacilli as multifaceted probiotics with poorly disclosed molecular mechanisms.* Int J Food Microbiol, 2010. **143**(3): p. 87-102.
2. Backhed, F., *et al.*, *The gut microbiota as an environmental factor that regulates fat storage.* Proc Natl Acad Sci U S A, 2004. **101**(44): p. 15718-23.
3. Ahrne, S., *et al.*, *The normal Lactobacillus flora of healthy human rectal and oral mucosa.* J Appl Microbiol, 1998. **85**(1): p. 88-94.
4. Reid, G., S.O. Kim, and G.A. Kohler, *Selecting, testing and understanding probiotic microorganisms.* FEMS Immunol Med Microbiol, 2006. **46**(2): p. 149-57.
5. Antonio, M.A., S.E. Hawes, and S.L. Hillier, *The identification of vaginal Lactobacillus species and the demographic and microbiologic characteristics of women colonized by these species.* J Infect Dis, 1999. **180**(6): p. 1950-6.
6. Gustafsson, R.J., *et al.*, *The Lactobacillus flora in vagina and rectum of fertile and postmenopausal healthy Swedish women.* BMC Womens Health, 2011. **11**(1): p. 17.
7. Palmer, C., *et al.*, *Development of the human infant intestinal microbiota.* PLoS Biol, 2007. **5**(7): p. e177.
8. Backhed, F., *et al.*, *Host-bacterial mutualism in the human intestine.* Science, 2005. **307**(5717): p. 1915-20.

9. Lee, Y.K. and S.K. Mazmanian, *Has the microbiota played a critical role in the evolution of the adaptive immune system?* Science, 2010. **330**(6012): p. 1768-73.

10. Ozen, M. and E.C. Dinleyici, *The history of probiotics: the untold story.* Benef Microbes, 2015. **6**(2): p. 159-65.

11. FAO/WHO (2001) Evaluation of health and nutritional properties of powder milk and live lactic acid bacteria. Food and Agriculture Organization of the United Nations and World Health Organization expert consultation report.

14. Assefa, S., *et al.*, *Lactobacilli with probiotic potential in the prairie vole (Microtus ochrogaster).* Gut Pathogens, 2015. **7**(1): p. 35.

15. *Szajewska, Hanna, et al.* Efficacy of Lactobacillus GG in prevention of nosocomial diarrhea in infants. *The Journal of Pediatrics.* ***138****(3): p. 361 – 365*.

16. Johnson, B.R. & Klaenhammer, T.R. *Impact of genomics on the field of probiotic research: historical perspectives to modern paradigms.* Antonie van Leeuwenhoek, 2014. **106**(1): p. 141-156.

17. "Human Genome Project Completion: Frequently Asked Questions." *National Human Genome Research Institute (NHGRI)*, www.genome.gov/11006943/human-genome-project-completion-frequently-asked-questions/

18. Barba, M., H. Czosnek, and A. Hadidi, *Historical perspective, development and applications of next-generation sequencing in plant virology.* Viruses, 2014. **6**(1): p. 106-36.

19. Bahassi el, M. and P.J. Stambrook, *Next-generation sequencing technologies:*

breaking the sound barrier of human genetics. Mutagenesis, 2014. **29**(5): p. 303-10.

20. Hayden, E.C., *Is the $1,000 Genome for Real?* Nature, 2014.
21. Quail, M.A., *et al.*, *A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers.* BMC Genomics, 2012. **13**(1): p. 341.
22. "Sequencing by Synthesis (SBS) Technology." *Sequencing Quality Scores*, www.illumina.com/science/technology/next-generation-sequencing/sequencing-technology.html.
23. Loman, N.J., *et al.*, *Performance comparison of benchtop high-throughput sequencing platforms.* Nat Biotechnol, 2012. **30**(5): p. 434-9.
24. Tuomola E, C.R., Playne M, Isolauri E, Salminen S, *Quality assurance criteria for probiotic bacteria.* American Journal of Clinical Nutrition, 2001. **73**: p. 393-398.
25. Delcour J, F.T., Deghorain M, Palumbo E, Hols P, *The biosynthesis and functionality of the cell-wall of lactic acid bacteria.* Antonie Van Leeuwenhoek, 1999. **76**(1-4): p. 159-184.
26. Lebeer, S., J. Vanderleyden, and S.C. De Keersmaecker, *Genes and molecules of lactobacilli supporting probiotic action.* Microbiol Mol Biol Rev, 2008. **72**(4): p. 728-64.
27. Jens Walter, D.M.L., Mohammed Alqumber, Christoph Rockel, Corinna Hermann, Markus Pfitzenmaier, Gerald W. Tannock, *d-Alanyl ester depletion of teichoic acids in Lactobacillus reuteri 100-23 results in impaired colonization of*

the mouse gastrointestinal tract. Environmental Microbiology, 2007. **9**(7): p. 1750-1760.

28. Neuhaus, F.C. and J. Baddiley, *A continuum of anionic charge: structures and functions of D-alanyl-teichoic acids in gram-positive bacteria.* Microbiol Mol Biol Rev, 2003. **67**(4): p. 686-723.

29. Perea Velez, M., *et al.*, *Functional analysis of D-alanylation of lipoteichoic acid in the probiotic strain Lactobacillus rhamnosus GG.* Appl Environ Microbiol, 2007. **73**(11): p. 3595-604.

30. Debabov, D.V., M.Y. Kiriukhin, and F.C. Neuhaus, *Biosynthesis of lipoteichoic acid in Lactobacillus rhamnosus: role of DltD in D-alanylation.* J Bacteriol, 2000. **182**(10): p. 2855-64.

31. Perego M, G.P., Minutello A, Strauch MA, Leopold K, Fischer W, *Incorporation of D-alanine into lipoteichoic acid and wall teichoic acid in Bacillus subtilis. Identification of genes and regulation.* Journal of Biological Chemistry, 1995. **270**(26): p. 15598-15606.

32. Steen, A., *et al.*, *Autolysis of Lactococcus lactis is increased upon D-alanine depletion of peptidoglycan and lipoteichoic acids.* J Bacteriol, 2005. **187**(1): p. 114-24.

33. Abachin E, P.C., Pellegrini E, Milohanic E, Fiedler F, Berche P, Trieu-Cuot P, *Formation of D-alanyl-lipoteichoic acid is required for adhesion and virulence of Listeria monocytogenes.* Molecular Microbiology, 2002. **43**(1): p. 1-14.

34. Morath, S., A. Geyer, and T. Hartung, *Structure-function relationship of cytokine induction by lipoteichoic acid from Staphylococcus aureus.* J Exp Med, 2001.

193(3): p. 393-7.

35. Begley, M., C. Hill, and C.G. Gahan, *Bile salt hydrolase activity in probiotics.* Appl Environ Microbiol, 2006. **72**(3): p. 1729-38.

36. Bateup, J.M., *et al*., *Comparison of Lactobacillus strains with respect to bile salt hydrolase activity, colonization of the gastrointestinal tract, and growth rate of the murine host.* Appl Environ Microbiol, 1995. **61**(3): p. 1147-9.

37. Christiaens, H., *et al*., *Cloning and expression of a conjugated bile acid hydrolase gene from Lactobacillus plantarum by using a direct plate assay.* Appl Environ Microbiol, 1992. **58**(12): p. 3792-8.

38. Verstraete, I.D.S.L.V.H.M.V.W.H.C.W., *Significance of bile salt hydrolytic activities of lactobacilli.* Journal of Applied Microbiology, 1995. **79**(3): p. 292-301.

39. Grill, J.P., *et al*., *Bifidobacteria and probiotic effects: action of Bifidobacterium species on conjugated bile salts.* Curr Microbiol, 1995. **31**(1): p. 23-7.

40. Kim, G.B., *et al*., *Cloning and characterization of the bile salt hydrolase genes (bsh) from Bifidobacterium bifidum strains.* Appl Environ Microbiol, 2004. **70**(9): p. 5603-12.

41. Schneider, J.P.G.C.C.J.M.A.F., *Isolation and characterization of a Lactobacillus amylovorus mutant depleted in conjugated bile salt hydrolase activity: relation between activity and bile salt resistance.* Journal of Applied Microbiology, 2001. **89**(4): p. 553-563.

42. Begley, M., *et al*., *Contribution of three bile-associated loci, bsh, pva, and btlB, to gastrointestinal persistence and bile tolerance of Listeria monocytogenes.*

Infect Immun, 2005. **73**(2): p. 894-904.

43. Smet, I.D., P.D. Boever, and W. Verstraete, *Cholesterol lowering in pigs through enhanced bacterial bile salt hydrolase activity.* British Journal of Nutrition, 1998. **79**(2): p. 185-194.

44. M du Toita, C.M.A.P.F., L.M.T Dicksb, U Schillingera, P Haberera, B Warliesc, F Ahrensc, W.H Holzapfela, *Characterisation and selection of probiotic lactobacilli for a preliminary minipig feeding trial and their effect on serum cholesterol levels, faeces pH and faeces moisture content.* International Journal of Food Microbiology, 1998. **40**(1-2): p. 93-104.

45. Klaver, F.A. and R. van der Meer, *The assumed assimilation of cholesterol by Lactobacilli and Bifidobacterium bifidum is due to their bile salt-deconjugating activity.* Appl Environ Microbiol, 1993. **59**(4): p. 1120-4.

46. M.T. Liong, N.P.S., *Bile salt deconjugation ability, bile salt hydrolase activity and cholesterol co-precipitation ability of lactobacilli strains.* International Dairy Journal, 2004. **15**(4): p. 391-398.

47. Smet, I.D., *et al.*, *In Vitro Study of Bile Salt Hydrolase (BSH) Activity of BSH Isogenic Lactobacillus plantarum 80 Strains and Estimation of Cholesterol Lowering through Enhanced BSH Activity.* Microbial Ecology in Health and Disease, 1994. **7**(6): p. 315-329.

48. Feighner, S.D. and M.P. Dashkevicz, *Subtherapeutic levels of antibiotics in poultry feeds and their effects on weight gain, feed efficiency, and bacterial cholyltaurine hydrolase activity.* Appl Environ Microbiol, 1987. **53**(2): p. 331-6.

49. Tuomola, E.M., A.C. Ouwehand, and S.J. Salminen, *The effect of probiotic*

bacteria on the adhesion of pathogens to human intestinal mucus. FEMS Immunology & Medical Microbiology, 1999. **26**(2): p. 137-142.

50. Buck, B. L., *et al*. *Functional Analysis of Putative Adhesion Factors in Lactobacillus Acidophilus NCFM.* Applied and Environmental Microbiology, 2005. **71**(12): p. 8344–8351.

51. Roumen Pankov, K.M.Y., *Fibronectin at a glance.* Journal of Cell Science, 2002. **115**: p. 3861-3863.

52. Holzapfel, U.S.C.G.W.H., *In vitro adherence and other properties of lactobacilli used in probiotic yoghurt-like products.* International Dairy Journal, 2005. **15**(12): p. 1289-1297.

53. Kapczynski, D.R., R.J. Meinersmann, and M.D. Lee, *Adherence of Lactobacillus to intestinal 407 cells in culture correlates with fibronectin binding.* Curr Microbiol, 2000. **41**(2): p. 136-41.

54. van der Flier, M., *et al*., *Adherence of Streptococcus pneumoniae to immobilized fibronectin.* Infect Immun, 1995. **63**(11): p. 4317-22.

55. Wadström T., *et al*., *Binding of Microbial Pathogens to Connective Tissue Fibronectin: An Early Step in Localized and Invasive Infections*. The Pathogenesis of Bacterial Infections, 1985, p: 193-207.

56. Sonomoto, K., Atsushi Yokota, *Lactic Acid Bacteria and Bifidobacteria: Current Progress in Advanced Research*. 2011: Caister Academic Press.

Speziale, P., *et al*., *Fibronectin binding to a Streptococcus pyogenes strain.* J Bacteriol, 1984. **157**(2): p. 420-7.

57. Kozich JJ. Westcott SL. Baxter NT. Highlander SK and. Schloss PD.

Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. Applied and Environmental Microbiology, 2013. **79**(17): p. 5112-5120

58. Arndt, D., Grant, J., Marcu, A., Sajed, T., Pon, A., Liang, Y., Wishart, D.S. (2016) PHASTER: a better, faster version of the PHAST phage search tool. Nucleic Acids Res., 2016 May 3.

    Zhou,Y., Liang,Y., Lynch,K.H., Dennis,J.J. and Wishart,D.S. (2011) PHAST: a fast phage search tool. Nucleic Acids Res., 39, W347–W352.

59. Wilson BAS, Abigail A; Whitt, Dixie D.; Winkler, Malcolm E. *Mechanisms of Genetic Modification and Exchange. Bacterial Pathogenesis: A Molecular Approach* 2011. p. 116-124

60. Begley M, Gahan CGM and Hill C. *The interaction between bacteria and bile*. FEMS Microbiology Reviews, 2005. **29**: p. 625-651.

VITA

Aaron Abraham

Candidate for the Degree of

Master of Science

Thesis: WHOLE GENOME SEQUENCING AND COMPARISON OF POTENTIALLY PROBIOTIC *LACTOBACILLUS* STRAINS

Major Field: Biomedical Sciences

Biographical:

Education:

Completed the requirements for the Master of Science in Biomedical Sciences at Oklahoma State University Center for Health Sciences, Tulsa, Oklahoma in May 2018.

Completed the requirements for the Bachelor of Science in Biology at The University of Tulsa, Tulsa, Oklahoma, in 2016.