UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

IMPACT OF SAMPLE COLLECTION PREPARATION ON METABOLOMIC AND MICROBIOME PROFILES

A THESIS

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

MASTER OF ARTS

By

JACOB JAMES HAFFNER

Norman, Oklahoma

2019

IMPACT OF SAMPLE COLLECTION PREPARATION ON METABOLOMIC AND MICROBIOME PROFILES

A THESIS APPROVED FOR

THE DEPARTMENT OF ANTHROPOLOGY

BY

Dr. Cecil M. Lewis, Jr., Chair

Dr. Courtney Hofman

Dr. Laura-Isobel McCall

Dedicated to Waldo

Acknowledgements

I am immensely grateful for the support from my committee members throughout this project. I thank Dr. Courtney Hofman for her advice, guidance, and reminders to always focus on the impact and implications of research. I thank Dr. Laura-Isobel McCall for allowing me to use her lab and its resources (especially for risking her mass spectrometer), her teachings in mass spectrometry-based metabolomics, and for her counsel. Most of all, I thank Dr. Cecil Lewis for his endless support, patience, and for granting me the opportunity to study at OU and LMAMR. His role in shaping me as a graduate student and researcher cannot be understated.

Additionally, I thank everyone at LMAMR. I thank Dr. Tanvi Honap, Dr. Krithivasan "Krithi" Sankaranarayanan, and Nihan Dagtas for their help with the labwork, aid with bioinformatics processing, and patience throughout this project (especially during its many technical problems). I am in debt to everyone at LMAMR who help make it an incredible lab and work experience. I also thank all the LMAMR graduate students who provided moral support and put up with my many questions: Rita Austin, Robin Singleton, Kristen Rayfield, Sterling Wright, Justin Lund, Dave Jacobson, Abigail Gamble, Christine Woelfel-Monsivais, Samuel Miller, Dr. Allison Mann, and Dr. Nisha Patel.

Lastly, I thank my parents, my sister, and my group of friends in Tulsa for their encouragement throughout this project.

Table of Contents

LIST OF TABLES

List of Figures

Abstract

Anthropological studies of human biology are predominately field-based, and burdened, by the need for well-preserved biological samples. In the emerging application of multi-omics, definitions of "well-preserved" and preservation strategies have had limited study, particularly those data that inform the biology of the human ecology. These human ecological data are a frequent objective of metabolomics and microbiome research. Metabolomics, through exploration of the totality of small molecules known as metabolites, offers a way to directly observe the molecular phenotype. These small molecules offer just as much valuable information to molecular anthropology as DNA and RNA, but there is a surprising lack of inquiry into how these molecules are preserved in samples and how molecular preservation impacts results and interpretations. For most metabolomic studies, the standard sample collection procedure involves snap-freezing the sample within 15 minutes of collection and storing at -80ºC. However, this is often unfeasible for field-based sample collection. Metabolome taphonomy, the study of how metabolome profiles are impacted by environmental processes as well as sample collection and preservation/preparation strategies, is still poorly understood. This thesis considers sample storage, with attention to human gut microbial samples. Consequently, this thesis presents two complementary studies, one with a focus on the metabolome and one focused on the microbiome taxonomic inventory, to determine if the application of the common DNA and RNA preservative RNA*later* provides a valuable method for conserving ecological data from both approaches. Ten human fecal samples previously collected and frozen at -80ºC were homogenized, aliquoted, and subjected to treatment

that simulates different levels of cold storage in the field: 22-25ºC, 4ºC, and -80ºC. To assess the impact of preservation methods on metabolite and bacterial taxonomic profiles, subsets of these aliquots were further treated with different preservation techniques, such as RNA*later*. Metabolomic and bacterial taxonomic profiles were characterized using liquid chromatography-tandem mass spectrometry and 16S rRNA amplicon sequencing, respectively. These results will inform field sample collection and best storage practices for human biological approaches that apply multi-omic studies.

CHAPTER 1

INTRODUCTION

The molecular understanding of human biology has entered a golden era driven by technological and protocol innovations that have allowed for a deeper characterization of the genome, and the molecules driving the phenotype above the genome, identified by the fields of transcriptomics (RNA), proteomics (proteins), and metabolomics (metabolites). However, the quality of these "-omic" big data remains contingent on the biological samples themselves. Common to biological anthropology are samples collected from challenging field sites, whose remote conditions can impact sample preservation. Field-based studies of the human microbiome and metabolome are arguably more sensitive to sample preservation issues than human genome studies because they not only require proper DNA and molecular preservation, but also, unbiased frequency and abundance profiles of the organisms, genes, metabolic pathways, and/or particles the DNA and metabolites represent.

Despite these concerns, there is a surprising lack of inquiry into how small molecules are preserved in samples and how this preservation impacts results. This thesis project addresses preservation concerns with regards to time, storage temperature, and a storage solution called RNA*later*. To address these preservation issues, a two-pronged approach was adopted, one that uses mass spectrometry-based metabolomics and a complementary approach that uses bacterial taxonomic 16S rRNA gene sequencing. Through these approaches, we explore how RNA*later* and storage conditions affect our data generation, whether these sample treatment steps introduce taxonomic and

compositional biases, and how sample preservation techniques might be improved. By addressing such preservation issues, we hope to improve the current methods used to explore molecular anthropology.

RNA*later* and Sample Preservation

The common standard for ensuring sample integrity involves freezing a sampleimmediately after collection (Fiehn 2002; Gorokhova 2005; Reck et al. 2015; van Eijsden et al. 2013). This can include snap freezing, freezing with liquid nitrogen, or placing the samples in a freezer. However, this is not always feasible in certain environments. Remote field sites rarely have access to cold storage technologies, making sample preservation complicated. This issue is especially problematic for RNA-based projects due to the rapid degradation of RNA (Reck et al. 2015). Several sample storage solutions were developed to overcome these in-field sample preservation concerns. One such common reagent, called RNA*later,* is the focus of inquiry for this project.

Ambion Invitrogen RNA*later* Stabilization Storage Solution, referred to hereafter as RNA*later*, is an aqueous storage reagent designed to preserve RNA in tissue samples (Lader 2001). Since its creation in the late 1990s, studies have demonstrated its efficacy at preserving not just RNA, but all nucleic acids (Gorokhova 2005; van Eijsden et al. 2013). Furthermore, RNA*later* has been proven practical at preserving nucleic acids within varying sample material. This includes bone (Cottrell et al. 2015), urine (Cheng et al. 2016), and feces (Reck et al. 2015). While in-house ingredients can vary, RNA*later* is typically comprised of sodium citrate, ethylenediaminetetraacetic acid, ammonium

sulfate, and a buffer (sodium acetate is recommended by the developer) (L. Technologies 2013; Lader 2001). RNA*later* works by penetrating sample cells with the ammonium sulfate salts and forcing a precipitation of nucleic acids and proteins (Lader 2001). This process is commonly known as 'salting out'. The salting out process by RNA*later* deactivates any nucleases found within the cells that would otherwise degrade any present nucleic acids (Gorokhova 2005; Lader 2001; Voigt et al. 2015).

RNA*later* eliminates many in-field sample storage concerns due to its ease of use in the field. According to the manufacturer, collected samples can be placed in certain volumes of RNA*later* (the amount varies depending on sample material) and frozen once cold storage is accessible (Ambion 2014). For example, utilizing RNA*later* for feces requires adding 1 mL of RNA*later* per gram of feces, mixing, and freezing (Reck et al. 2015; Zoetendal et al. 2006). Once placed in RNA*later*, samples can be left at room temperature for 1 week without jeopardizing sample integrity (Ambion 2014; Reck et al. 2015), although standard storage at 4°C, -20°C, or -80°C is eventually necessary to avoid molecular degradation. Samples should be submerged in RNA*later* overnight at 4°C before being transferred to -20°C or -80°C (Ambion 2014). After storage, samples should be blotted using a paper towel and gently rinsed to remove RNA*later* (fecal samples do not receive this step). As stated in the developer manual, samples stored in -20°C or -80°C preserves samples indefinitely (Ambion 2014). However, this thesis questions the adequacy of these storage lengths.

Despite RNA*later*'s frequent usage, there are few studies that have examined the consequences of RNA*later* treatment on data quality (for exceptions see: Choo et al.

2015; Loftfield et al. 2016; Sinha et al. 2016; Wang et al. 2018). Addressed in this thesis are three critical questions: How does RNA*later* impact metabolomic and 16S RNA gene data for functional and microbiome taxonomic characterization, respectively? Does RNA*later* impact the data diversity measures or bias metabolic and microbial data? And lastly, can RNA*later* treatment generate both untargeted metabolomic and 16S rRNA gene data?

Metabolomics and Anthropology

Metabolomics is the study of the total metabolites present and their functional roles within a biological system (Bino et al. 2004; Greaves and Roboz 2014; Patti et al. 2013; Wolfender et al. 2015). The particular definition of metabolite varies, but in this project, metabolites are any small molecule involved in life-sustaining chemical reactions (metabolic reactions, or metabolism) whose molecular weight is under 1500 Daltons (Da) (Viant et al. 2017). Due to the vast numbers of atomic arrangements, metabolites have high levels of structural variability, especially compared to genes and proteins (Fiehn 2002). This extensive assortment of metabolites is generally divided into two categories: endogenous and exogenous. Endogenous metabolites are found naturally in organisms whereas exogenous metabolites are environmentally acquired (Dawes and Ribbons 2003; Wishart 2016). The total sum of all metabolites is known as the metabolome (Fiehn 2002; Patti et al. 2012). Through studying the metabolome, researchers can explore the functional role of metabolites within biological systems (Wolfender et al. 2015). Investigation of these metabolites, and their associated pathways, offers a direct snapshot

of the biological phenotype as formed by interactions between the genotype and environment (Dettmer et al. 2006; Patti et al. 2012). This connection to the phenotype exists because metabolic reactions are fundamental biological processes (DeBerardinis and Thompson 2012). Thus, metabolites and the metabolome represent the ultimate response to genetic and environmental forces (Bino et al. 2004; Nicholson et al. 2011). For example, some human diseases (such as heart disease, diabetes, stroke, and cancer) are caused by genetic, lifestyle, and environmental influences (Rattray et al. 2018; Willett 2002). Studying the metabolome offers a way to thoroughly examine these disease states and their associated molecules from all sources of origin, rather than a single cause (Rappaport and Smith 2010; Rattray et al. 2018; Willett 2002).

Unfortunately, there is no single approach that can completely capture the metabolome (Bino et al. 2004; Dettmer et al. 2006; Wishart et al. 2018). The molecules of interest, experimental conditions, instruments, and data analysis approaches are variably selected by researchers to best answer their research questions. Depending on the project, these conditions can change. One common metabolomics approach involves targeted analysis, which focuses on a group of metabolites related to a specific pathway or metabolite class (Patti et al. 2012). A targeted analysis quantifies a known metabolite or a small number of metabolites (Wolfender et al. 2015). On the other hand, untargeted screenings categorize analytes depending on a change in response to stimuli and focus on measuring as many metabolites as possible (Dunn et al. 2013). Either approach is equally valid at generating metabolomic data and should be selected based on the specific project goals. For this project, an untargeted approach was adopted to investigate sample

treatment effects on all metabolites found in samples. Such untargeted approaches are popular for microbiology and have begun to emerge in molecular anthropology studies (Sankaranarayanan et al. 2015; Velsko 2017).

Metabolomics provides critical information for biological anthropologists. Biological anthropology can be defined as "the study of human biology within the framework of evolution" (Jurmain et al. 2013). With the popularity of genomics, biological anthropologists began studying human biology and evolution through studying biochemical molecules like DNA to explore human genetic origins (Ayala 1995), compare human microbiomes with chimpanzees and gorillas to track changes during human evolution (Moeller et al. 2014), study the relatedness of humans and past hominins like Neanderthals (Ovchinnikov et al. 2000), and much more. These new molecular anthropologists applied biochemical techniques and technologies to answer anthropological questions (Marks 2002). Recently, molecular anthropological work has shifted to studying metabolites. Examples include incorporating targeted MS to study hunter-gatherer diets as a proxy for ancient humans (Turroni et al. 2016), medicinal plant use by Neandertals (Hardy et al. 2012), and detection of possible metabolites associated with longevity in various mammals (Ma et al. 2015). Because metabolomics explores the functional role of metabolites, it represents human biology on a molecular level (DeBerardinis and Thompson 2012). This is because metabolites are integral to all biological processes, including those of health and disease (DeBerardinis and Thompson 2012). Thus, metabolomics has a lot to offer in the exploration of human biology. Despite its value to anthropology, metabolomics's importance has yet to be fully acknowledged

within biological anthropology on the same level as genomics. Some molecular anthropological studies have conducted metabolomic projects (Radini et al. 2016; Sankaranarayanan et al. 2015; Velsko et al. 2017), but these fields remain young, especially with respect to studies that assess the preservation and profile integrity of the metabolome in field-based studies.

Mass Spectrometry-Based Metabolomics

Mass spectrometry (MS) coupled to separation techniques or direct injection (Dettmer et al. 2006) is a common method for metabolomics. Nuclear magnetic resonance is also used in metabolomics (Wang and Bodovitz 2010), but MS is the focus here because it was employed for this project. MS can identify, quantify, and analyze metabolites found in a sample by generating ionized particles and releasing them into the instrument where they are quantified by the detector and subsequently analyzed (Dettmer et al. 2006; Greaves and Roboz 2014). There are varying forms of MS instruments and techniques available, but no current process can detect every metabolite in a sample (Bino et al. 2004; Zamboni, Saghatelian, and Patti 2015). Each technique and instrument present their own biases in metabolite detection (Greaves and Roboz 2014). As a result, MS-based metabolomics projects can be highly varied in their goals, methods, and acquired data (Zamboni, Saghatelian, and Patti 2015).

Sample introduction is a critical part of any MS approach, with chromatographic separation (CS) being a popular method. Generally, CS separates and transfers sample

compounds between two phases to purify the components for identification (Coskun 2016). CS coupled to mass spectrometry is generally divided into gas chromatography-mass spectrometry (GC-MS) or liquid chromatography-mass spectrometry (LC-MS) (Dettmer et al. 2006; Patti et al. 2012; Coskun 2016). CS involves passing samples through a column lined with a special solution (called the stationary phase) via the mobile phase. This mobile phase can be a gas or a liquid, depending on whether gas chromatography or liquid chromatography was employed (Coskun 2016; Greaves and Roboz 2014). Gas chromatography features a gas mobile phase for sample transfer, elutes more nonpolar molecules, and is preferred for thermally stable samples (Greaves and Roboz 2014). On the other hand, liquid chromatography utilizes a liquid as the mobile phase, elutes more polar molecules, is commonly used for thermally volatile samples, and has higher sensitivity for metabolic detection (Coskun 2016; Greaves and Roboz 2014). This greater sensitivity by liquid chromatography is due to its ability to detect molecules greater than 600 Da; gas chromatography is less accurate for molecules above this threshold (Greaves and Roboz 2014). Therefore, liquid chromatography is generally preferred for untargeted metabolomics. Additionally, some MS instruments include a fragmentation step. Such MS instruments utilize tandem mass spectrometry (MS/MS). MS/MS fragments ions and detects molecules a second time, allowing for greater detection and identification of metabolites (Greaves and Roboz 2014). For this project, a MS/MS instrument called the ThermoScientific Q Exactive Plus Hybrid Quadrupole-Orbitrap Mass Spectrometer was used to perform untargeted screenings of samples. This MS/MS instrument was coupled to a ThermoFisher Scientific Vanquish Flex Binary LC

System to perform liquid chromatography. These together allowed LC-MS/MS to be employed for this project. A brief description of the MS/MS instrument and how it works is necessary to understand its role for this project.

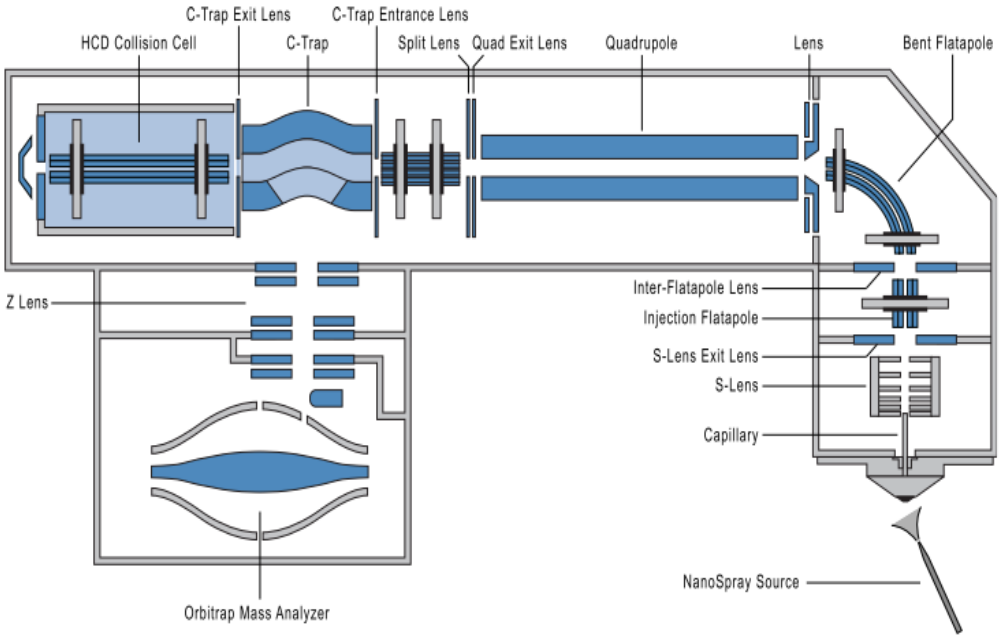


Figure 1. Construction details from the Q Exactive Plus.

This figure was originally published in *Molecular & Cellular Proteomics*. Michalski A., Damoc E., Hauschild J-P., Lange O., Wieghaus A., Makarov A., Nagaraj N., Cox J., Mann M., and Horning S. Mass Spectrometry-based Proteomics Using Q Exactive, a High-performance benchtop Quadrupole Orbitrap Mass Spectrometer. Mol Cell Proteomics. 2011; 10:1-11. © The American Society for Biochemistry and Molecular Biology.

Charged particles travel from the NanoSpray source, through the flatapoles, into the quadrupoles for filtering, collect in the C-trap, fragment in the HCD Collision Cell (for MS/MS), and enter the Orbitrap Mass Analyzer for detection.

For this MS model, ions are initially formed via electrospray ionization (Scheltema et al. 2014). Here, sample molecules move through the CS column from the Vanquish LC System as a liquid and are dissolved by an electrically charged solvent, creating charged ionic droplets (Greaves and Roboz 2014). These nebulized droplets enter the MS/MS instrument and are evaporated to become a gas (Greaves and Roboz 2014; Michalski et al. 2011; Scheltema et al. 2014). Ions then pass through the S-lens. This is a series of rings serving as an ion transfer tube (Michalski et al. 2011). Next, the ions move through the injection flatapole (which selects out ions) into the bent flatapole. The bent flatapole is a series of rods with small gaps between them that allow for droplets to fly out of the flatapole (Scheltema et al. 2014). This design prevents unwanted agents from passing further into the instrument (Michalski et al. 2011). From the bent flatapole, ions are channeled into a chamber of four long cylindrical rods called quadrupoles. These quadrupoles generate an electric field to guide selected ions along while also filtering out unwanted ions (Michalski et al. 2011). This ion filtering will vary according to the experiment. Moving through the quadrupole, ions enter the C-trap and are kept here before moving along. This next step moves ions into the HCD collision cell, where they are fragmented through collision (Michalski et al. 2011; Scheltema et al. 2014). This fragmentation is what defines MS/MS. After the HCD collision cell, ions are sent back to the C-trap before passing through the Z lens. This Z lens leads into the final section of the instrument: the orbitrap mass analyzer (Michalski et al. 2011). The orbitrap is a spindle-shaped metal rod generating an electrical charge. Ions are fired from the C-trap into the orbitrap at high speeds where the electrical force and momentum causes ions to spin and

move across the orbitrap (Greaves and Roboz 2014; Fisher Scientific Inc 2016; Scheltema et al. 2014). Due to the spinning and movement of the ions, a ring of constantly moving ions is formed around the orbitrap. The size of this ring and the ion speed will depend upon the mass-to-charge ratio (*m/z*) of the molecules. The MS instrument detects these rings and subsequently analyzes the fragmented ions (Michalski et al. 2011; Scheltema et al. 2014). Ultimately, the Q Exactive Plus offers highly accurate analyte detection and characterization for metabolomic studies (Michalski et al. 2011; Scheltema et al. 2014).

Despite its many strengths, MS-based metabolomics has problems ranging from database inconsistencies, instrument differences, accurately identifying metabolites, and sample treatment and preparation, to name a few (Johnson and Gonzalez 2012; Matsuda 2016; Patti et al. 2012). These last major weaknesses are highlighted in this thesis: sample preparation and treatment. Prior to separation and injection on an MS instrument, samples must undergo preparation. This step is crucial in extracting analytes, but it results in metabolite losses (Hollywood et al. 2006; Dettmer et al. 2006). Ultimately, sample preparation tends to be where metabolomic experimental errors most frequently occur (Fiehn 2002). These errors commonly include metabolite loss and misidentification. Specific losses will vary depending on the preparation techniques employed, with solid-phase extraction (SPE) and liquid-liquid extraction as the most common sample preparation methods (Dettmer et al. 2006).

Another major problem in MS-based metabolomics is sample treatment. Any procedures following sample collection can create biases in the formation, degradation,

and detection of metabolites. As a result, sample preservation greatly affects metabolomic data. Standard storage methods involve simply snap freezing samples in liquid nitrogen or freeze clamping (Fiehn 2002). However, few studies have explored the effects of sample preparation and treatment on metabolomic experiments. Complications of sample preservation are further compounded by the inability to use RNA*later* (Wang et al. 2018). RNA*later* is known to inhibit metabolomic data due to the components of RNA*later* interfering with the MS instrument and preventing metabolic detection (Loftfield et al. 2016; Sinha et al. 2016a; Sinha et al. 2016b; Wang et al. 2018). Particularly, the ammonium sulfate salts are known to be problematic for MS analysis (Loftfield et al. 2016; Sinha et al. 2016). Only a handful of studies have used RNA*later* for metabolomic projects as a result (Loftfield et al. 2016; Sinha et al. 2016; Wang et al. 2018). This project addresses these preservation concerns with RNA*later* and MS-based metabolomics in efforts to allow usage of RNA*later* for untargeted MS-based metabolomic studies through a SPE protocol. These metabolomics data are paired with an investigation of RNA*later*'s effects on sample integrity through microbiome profile analysis.

The Microbiome and Anthropology

The microbiome is the collective sum of microorganisms (plus their genetic material) living in an environment (Grice and Segre 2012). For humans, the human microbiome is a composite of several microbial communities found in the gut, mouth, reproductive tract, and skin (Grice and Segre 2012; Turnbaugh et al. 2007). Known to

number in the trillions, these microorganisms play a variety of biological roles including digestion, metabolism, and immunity (Blaser and Falkow 2009; Grice and Segre 2012; Turnbaugh et al. 2007).

Human microbiome research elucidates the intersect between genetics, health, environment, and lifestyles (Turnbaugh et al. 2007). Anthropologists have studied the human microbiome with growing interest because the microbiome can shed light on our species' evolutionary history, varying diets, behavior, and diversity (Benezra et al. 2012; Blaser and Falkow 2009; Vuong et al. 2017). Such studies can include investigating differences between human and non-human primate microbiomes (Moeller et al. 2014; Yildirim et al. 2010), examining historical microbiomes (Tito et al. 2012), and comparisons between human hunter-gatherer and industrialized populations (Obregon-Tito et al. 2015; Schnorr et al. 2014). Researchers can utilize microbiome analysis methods through DNA sequencing to explore why some bacterial species are no longer present, why microbial levels of diversity changed, when these changes occurred, and how our relationship with these microbes impacts our biology today. Most common to anthropological studies of the human microbiome are rare and extraordinary samples retrieved from unique environments and cultural practices (Moeller et al. 2014; Obregon-Tito et al. 2015; Sankaranarayanan et al. 2015; Schnorr et al. 2014; Turroni et al. 2016; Yildirim et al. 2010). Anthropologists have a legacy of innovating protocols to facilitate their often complex and challenging sample conditions (Benezra et al. 2012; Kaestle and Horsburgh 2002; Outram 2008; Warinner et al. 2014). This thesis continues in that same spirit, with attention to the microbiome profile via 16S rRNA gene sequencing.

The 16S rRNA Gene

Species identification is a recurring pursuit within biological sciences (Clarridge 2004; Pereira et al. 2010). Early methods for identifying species relied on analyzing physical features and comparing these characteristics to those identified by other researchers (Clarridge 2004; Woese 1987). However, these methods are often challenged by the subjective and varied nature of examining bacteria for physical characteristics and the difficulty in studying unculturable bacteria (Clarridge 2004). In the 1970s and 1980s, the 16S rRNA gene was demonstrated to be more effective at identifying bacterial species than these earlier methods (Fox et al. 1977; Woese 1987; Clarridge 2004). Moreover, 16S rRNA gene sequencing provided a way to taxonomically identify unculturable bacteria (Amann et al. 1995; Pace 1997). Advancements in gene sequencing have further reinforced 16S rRNA gene sequencing as an ideal technique for characterizing bacterial ecologies, such as the gut microbiome (Clarridge 2004; Jovel 2016).

The 16S rRNA gene, also known as 16S rDNA, encodes for a component of prokaryote ribosomes. The 16S rRNA gene is frequently sequenced and studied for taxonomic and phylogenetic purposes within microbiol9ogy because it is ubiquitous amongst bacteria, has highly conserved regions for targeted PCR-based methods and species-specific hypervariable regions for phylogenetic resolution, and is inexpensive and easy to sequence (Clarridge 2004; Fox et al. 1977; Kim and Chun 2014). In particular, the conserved and hypervariable regions make 16S rRNA gene sequences superior to earlier phenotype-based phylogenetic methods (Pace 1997). By aligning different organisms'

16S sequences, researchers can count nucleotide differences as a measure of evolutionary distance between the organisms (Amann et al. 1995; Pace 1997). Thus, researchers can employ 16S rRNA gene sequences to understand how prokaryotic evolution occurred, what species are related to each other, and when species might have diverged through established methods in phylogenetics. While the gene's conserved and hypervariable regions make it an excellent molecular clock for measuring this evolutionary distance (Tsukuda et al. 2017), the 16S rRNA gene sequence is limited at identifying closely-related species or within-species strains (Jovel 2016; Kolbert and Persing 1999), such phylogenetic studies often require more multi-loci or genome studies, which exceed the resource of this thesis. Nevertheless, 16S rRNA gene-based studies remain the most prolific of phylogenetic approaches to microbiology in general, and microbiome specifically, in the last 30 years.

A 16S rRNA gene approach does not replace or diminish the importance of culture-based methods. While phenotypic species identification methods were limited in their ability to characterize unculturable bacteria, it has its advantages for culturable bacteria (Clarridge 2004). Chemotaxonomy, a combination of phylogenetics and culture-based approaches to functional characterization of bacteria, remains a common practice for identifying novel species, and thus, culturing methods remain critical to the study of microbial variation (Clarridge 2004; Prakash et al. 2007). This combination of phenotypic and genotypic information for taxonomic purposes is called polyphasic taxonomy (Prakash et al. 2007; Vandamme et al. 1996). Such polyphasic approaches are

common in microbiology studies and frequently generate taxonomic inventories of bacterial species found in molecular anthropology and microbiome research.

CHAPTER 2

MATERIALS AND METHODS

This research was conducted at the University of Oklahoma Norman Research Campus in the Laboratories of Molecular Anthropology and Microbiome Research (LMAMR).

Sampling

All samples for this project were human fecal samples gathered from villages in Burkina Faso, Africa. A total of 120 individuals from 30 families (four from each family) contributed samples. Informed consent was provided with oversight from the Ministry of Health Ethics Committee located within Centre Muraz, a Burkina Faso national research institute. Participants were equipped with a labeled disposable collection container and a pair of gloves. The collection container and fecal sample were returned to researcher on site. Multiple scoops of sample were placed in collection tubes and these tubes were sealed then placed in a labeled bag. Bags were sealed and placed in ice located on-site. Next, samples were transferred to a -20°C freezer for overnight storage. All processing took place within 15 minutes. Each evening, samples were thawed prior to DNA extraction. DNA was extracted then samples were frozen again. Sampling occurred in this manner over the course of several days. After sample collection, all frozen samples were shipped to LMAMR at Norman, Oklahoma, and stored in a -80°C freezer. Samples were briefly thawed again to extract 2g from each sample for anaerobic culturing. After

the 2g collection, samples were kept frozen in -80°C until treatment for this project. Ten of these 120 total samples were used for this project.

Experimental Design

The design and sample treatment of this project can be grouped into two distinct stages for metabolomics. Stage two was a modification of stage one, but the same samples were used for the entirety of the project.

Ten samples with the largest mass were chosen and set aside for this project. Four grams (g) of each sample was aliquoted, which were the working samples for this project. Of each 4g sample, 1g was removed sequentially as subsamples and stored in -80°C. These 1g subsamples were frozen backups in case more samples were needed. The remaining 3g of each sample was then divided into two separate groups of 1.5g each. One 1.5g group was designated as raw, untreated samples. The second 1.5g group was allocated for treatment with RNA*later*. Each 1.5g group (totaled at 20 separate 1.5g groups, 10 without RNA*later* and 10 with RNA*later*) were then further aliquoted into three distinct 0.5g portions. This created a total of 60 working sample, with six subsamples per sample. Of these six subsamples per sample, a total of three were untreated and the other three were treated with RNA*later*. Thus, 30 total untreated and 30 RNA*later* samples were used for this project.

Following this treatment, the six subsamples from each sample were sorted into temperature groups: -80°C, 4°C, and room temperature (22-25°C). These temperatures were chosen because they are common storage temperatures for metabolomic samples

(Dettmer et al. 2006). Two of the six subsamples were placed in each temperature group, creating two in -80°C, two in 4°C, and the last two in room temperature. These samples were stored in their respective temperature groups for two weeks. Within these temperature groups, one subsample was untreated, and the other was RNA*later*-treated. These steps were repeated for each sample. In the end, the 60 total samples (six each from the ten original samples) had 20 0.5g aliquots in each of the three temperature groups. Half of these 20 were untreated while the other half were RNA*later*-treated. Figure 1 depicts this experimental design.

After two-week storage, RNA*later* samples underwent a RNA*later* cleanup protocol designed to remove RNA*later* components which impact MS analysis. Stage one samples were subjected to the RNA*later* cleanup protocol twice to maximize RNA*later* removal and avoid contaminating the MS instrument. Non-RNA*later* samples did not undergo this protocol and, instead, were prepared for MS immediately after the two-week storage. Following RNA*later* cleanup, samples were prepared for MS analysis then stored in -80°C prior to MS analysis. Remaining sample material was kept at -80°C. This concludes stage one of the project.

Stage two tested the efficacy of undergoing RNA*later* cleanup two times. Of the 30 available RNA*later*-treated 0.5g samples, seven random samples were chosen and set aside for further RNA*later* cleanup. Samples in stage two were treated identically to those from stage one except the stage two samples only underwent one RNA*later* cleanup protocol. Following the single RNA*later* removal, samples underwent the same metabolomic workflow as stage one. Stage one and two were run as separate batches on

the MS instrument. A minimal number of RNA*later* samples (12 total, including blank) were run on the MS instrument to ensure traces of RNA*later* would not negatively affect the MS instrument. Due to time limitations, only RNA*later* samples were run on the MS instrument.

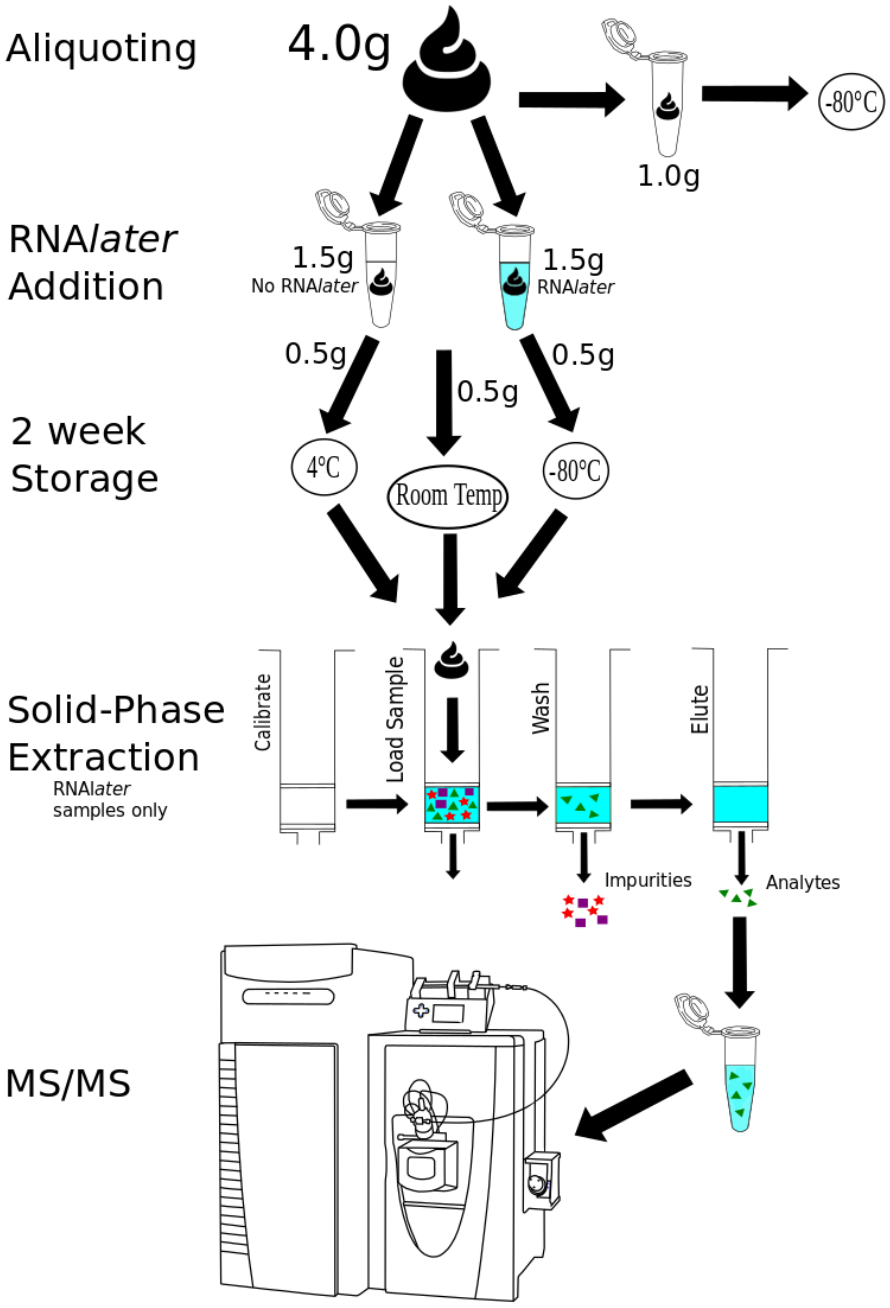


Figure 1: Experimental Design.

Simplified visualization of metabolomics experimental design described previously. For stage one, samples went through SPE again following elution. For stage two, samples went through SPE once. Samples were loaded onto a Vanquisher LC system which injected samples into MS instrument.

Solid-Phase Extraction: RNAlater Cleanup

In this study, the RNA*later* cleanup protocol was optimized from a SPE protocol presented by Cottrell et al. 2015. An Oasis Waters extraction manifold (20 pos, 13x75mm tubes, cat.: WAT200606) connected to a Rocker 400 650mmHg vacuum pump (cat.: 167400–11) were used for SPE. Oasis Waters SPE HLB 1cc Vac Cartridges (10 mg sorbent per cartridge, 30µM particle size, cat.: 186000383) were placed in the extraction manifold. Flow rate was adjusted to approximately 1 drop per second, as per Oasis Waters SPE protocol. Vacuum pressure and flow rate varied depending on the sample buffer, necessitating frequent manual adjustment of the flow rate to match the 1 drop per second ideal. LC-MS grade water and LC-MS grade methanol were always used for RNA*later* cleanup protocol unless otherwise specified. 5mL culture tubes were used to collect flowthrough and were replaced after each cartridge loading.

To prepare RNA*later*-treated samples, RNA*later* was added to create in a 1:1 ratio of RNA*later* to sample. In this sample type, 0.5g of feces were combined with 500µl of RNA*later* was added to each of these samples. Following RNA*later* addition, all samples were mixed with volumes of water spiked with 2µM sulfachloropyridazine as internal standard (IS) to reach a total volume of 5mL. This resulted in 4.5mL of water. For the untreated samples, 5mL of water was used. All samples were then homogenized by sonication to create fecal slurries.

Samples were placed in a FisherScientific Ultrasonic Cleaning Bath (20.8L, cat.: 15-337-435) at maximum power for 10 minutes. Following sonication, samples had organic and supernatant layers. 1000µL of supernatant from each RNA*later*-treated

sample were collected and placed in separate, appropriately labeled 1.5mL collection tubes prior to SPE. 1000µL of supernatant from each of untreated samples were collected, placed in new 1.5mL collection tubes, and set aside. The original samples were then placed in -80°C for storage. From here, only the supernatants of the RNA*later*-treated samples underwent the following steps.

Next, samples were centrifuged at 14,000rpm at 4°C for 10 minutes using an Eppendorf 5242 mini-centrifuge (cat.: 0008643) in a freezer room. Oasis Waters SPE HLB cartridges were rinsed with 3mL methanol and 3mL water to condition and equilibrate the cartridges under vacuum. After all liquid flowed through, 1000µL of samples were each loaded into cartridges under vacuum. Next, 3mL from each of the following were added to every cartridge under vacuum in this order: water, 5% methanol, 50% methanol with 0.1% acetic acid, and 50% methanol. All liquid flowed through the cartridges before adding the subsequent solution. Cartridges were then vacuum-dried to remove any remaining liquid. Next, 1mL methanol was added for elution with new 1mL collection tubes placed to collect eluates, yielding 1000ul of eluate for each sample. For stage one, this RNA*later* cleanup was repeated using the new eluates. New SPE cartridges were put in place and the steps were identical to previous RNA*later* cleanup with the eluates used for sample loading. Stage two moved on to next step of MS treatment.

Following final elution from RNA*later* cleanup, elutes and 1000µL of supernatant from untreated samples were placed in an Eppendorf Vacufuge Vacuum Concentrator

(cat.: 005535). Dessicator function was utilized to dry down samples for MS analysis. After dessication, dried samples were stored in -80°C prior to MS analysis.

1) Add 500µl of RNA*later* to 0.5g aliquots of raw fecal samples. Do not add to non-RNA*later* samples. Add 4L of LC-MS grade water (spiked with IS) to RNA*later* samples. For non-RNA*later* samples, add 4.5L instead. Total volumes should be 5L.
   a) All following reagents and solutions should be LC-MS grade.
2) Homogenize to create slurries. Place in sonicator at maximum power for 10 minutes.
3) Take 1000µl of aqueous supernatant and place in separate tubes. Remaining samples should be stored in -80°C as backup. Set aside supernatants from non-RNA*later* samples. Only RNA*later* supernatants should undergo the following steps.
4) Centrifuge at 14,000rpm at 4°C for 10 minutes.
5) During or after centrifugation, place SPE cartridges into SPE extraction manifold. One cartridge per sample. Place collection tubes (we used culture tubes) in the interior manifold tube rack to collect waste.
6) Rinse each SPE cartridge with 3mL methanol, followed by 3mL water. Cartridges are now conditioned and equilibrated.
   a) Replace waste tubes following each added reagent to cartridge.
7) After centrifugation, load all 1000µL of sample supernatant into cartridges.
8) Add 3mL of methanol to each cartridge.
9) Add 3mL 5% methanol to each cartridge.
10) Add 3mL 50% methanol with 0.1% acetic acid to each cartridge.
11) Add 3mL 50% methanol to each cartridge.
12) Vacuum-dry cartridges for 5 minutes.
13) Place empty 1mL collection tube in interior tube rack. Add 1mL methanol to each cartridge for elution.
14) For second RNA*later* removal, elutes from Step 11 were treated as new samples and protocol was restarted from Step 4.
15) After final elution, place RNA*later*-removed elutes and 1000µL from untreated samples inside Vacufuge. Activate dessicator function to dry down samples for MS analysis.
16) Once liquid is dried, place samples in -80°C until ready for MS analysis.

Table 1. RNA*later* Removal Protocol.

Our optimized RNA*later* cleanup protocol is listed here. The protocol is a normal-phase SPE procedure developed from a similar protocol used by Cottrell et al. 2015. Goal is to isolate and remove the RNA*later* salts from samples without removing metabolites. SPE works by having target molecules bind to a silica sorbent within the cartridge, wash the sorbent, and then elute the target analytes. Chemicals and reagents used for SPE should be selected based off their chemical properties and the properties of the target analytes. For this RNA*later* removal SPE protocol, all solutions and reagents should be LC-MS grade. Only samples treated with RNA*later* should undergo steps 4-14.

Liquid chromatography-tandem mass spectrometry

All LC-MS/MS processing was done using the ThermoFisher Scientific Vanquish Flex Binary LC System (cat.: IQLAAAGABHFAPUMBJC) linked to the ThermoFisher Scientific Q Exactive Plus Hybrid Quadrupole-Orbitrap Mass Spectrometer (cat.: IQLAAEGAAPFALGMBDK). These instrument performed LC and MS/MS, respectively. Samples were removed from -80°C and resuspended using 200µL of 50% LC-MS grade methanol spiked with 0.5µg/mL sulfadimethoxine as an IS. After resuspension, samples were added to a 96-well plate for MS injection. Resuspended samples were injected with an injection volume of 20µL. A Kinetix C18 core-shell column (50x2.1mm, 1.7µM particle size, 100 Å pore size, cat.: 00B-4475-AN) was used for LC. The mobile phase consisted of two solvents: Solvent A as LC-MS grade water with 0.1% formic acid and Solvent B as LC-MS grade acetonitrile with 0.1% formic acid. To avoid RNA*later* components from entering MS instrument, flow was initially directed to waste for 30 seconds and 15 seconds for samples from stages one and two, respectively. This common step prevents contaminating the MS source with unwanted molecules from the mobile phase. After this initial waste redirection, gradient parameters were 5% Solvent B for 1 minute, increase from 5%-100% Solvent B over 8 minutes, remain at 100% Solvent B for 2 minutes, decrease to 5% Solvent B for 30 seconds, and a 1 minute re-equilibration phase at 5% Solvent B. Column temperature and compartment were kept at 40°C and 10°C, respectively, during analysis. For samples from stage one, samples were randomly selected for injection in order to test effects of RNA*later* removal on MS. All stage two samples were injected in order of location on 96-well plate.

Electrospray ionization parameters were: sheath gas flow rate at 35 L/min, auxiliary gas flow rate at 10 L/min, auxiliary gas temperature at 350°C, and sweep gas flow rate at 0 L/min. S-lens RF was at 50V, spray voltage was at 3.8 kV, and the capillary temperature was at 320°C. MS data were acquired in positive mode, with data-dependent acquisition for MS2 data. MS scan ranges were set to 100-1500 m/z. 5 MS/MS scans of the most abundant ion per cycle were recorded. MS1 resolution was set to 35,000 and MS2 resolution was set to 17,500. MS1 and MS2 maximum injection time were both set at 100 ms. MS AGC target was at 1e6 and MS/MS AGC target was at 5e5. 2m/z was used as an isolation window. MS/MS occurred at 2-8 seconds with an exclusion of 10 seconds. Collision energy was increased from 20% to 30% and to 40%.

16S rRNA Gene Amplicon Sequencing

To illustrate bacterial taxonomic profiles, the bacterial 16S rRNA gene V4 hypervariable region was targeted and amplified. 16S amplification and sequencing was performed on all samples following sonication after addition of RNA*later*. As mentioned earlier, a total of 60 samples were utilized in this project. These 60 samples underwent the same 16S procedures, regardless of temperature or RNA*later* treatment. None of the samples used for 16S rRNA gene sequencing underwent the RNA*later* removal protocol.

DNA was extracted using the Qiagen AllPrep PowerViral DNA/RNA Kit (cat:28000-50) with extraction blanks. Extraction protocol followed manufacturer instructions. Final DNA concentration was quantified using the Invitrogen Qubit 2.0 Fluorometer (cat.: Q32866) with the ThermoFisher Scientific Qubit dsDNA Broad Range

assay kit (cat.: Q32850). See supplementary table 1 for these Qubit results. Samples underwent this DNA extraction procedure in batches of ten with one negative blank in each batch, causing the final sample count for 16S rRNA gene sequencing to be 66 total samples (60 samples with 6 negative blanks). These 66 samples all underwent the following steps.

16S copy number quantification was collected via quantitative polymerase chain reaction (qPCR). The Roche FastStart Essential DNA Green MM with SYBR Green I was used (cat.: 06402712001) on a Roche Lightcycler 96 (cat.: 05815916001). 10μM V4 non-Illumina primer stocks of 515f (GTGCCAGCMGCCGCGGTAA) and 806r (GGACTACHVGGGTWTCTAAT) were used as forward and reverse primers, respectively. qPCR negative blanks were included. In-house *Escherichia coli* (*E. coli*) standards (1000x, 100x, 10x 16S copies per μL) were used as positive controls and standards. Initial denaturation was set to 95°C for 10 minutes, with 35 cycles at 95°C for 10 seconds, 52°C for 20 seconds, and 72°C for 30 seconds. Samples were diluted and categorized into two PCR groups according to qPCR results.

16S triplicate PCR was performed with negative and positive controls using the ThermoFisher Scientific Phusion HotStart II High Fidelity DNA Polymerase Enzyme System (cat.: F-549L) on an Analytik Jena Biometra T Professional Trio Thermocycler (cat.: 3408114). 10μM stocks of the universal 515f V4 primers with Illumina adapters were used in all samples. 2.5 μM stocks of 806r V4 barcoded primers with Illumina adapters were similarly employed. These universal reverse primers had unique 12bp GOLAY error-correcting barcodes for multiplexing (Caporaso et al. 2012). Each sample

was given a unique reverse primer and barcode. Supplementary table 3 depicts these sample-to-barcode matches. UV nanopure water used for PCR blanks instead of DNA template. Supplementary table 4 depicts the PCR reaction sheet used, including amount of each reagent. PCR conditions were initial denaturation at 98°C for 30 seconds, 18 (for PCR group 1) or 20 (for PCR group 2) cycles of 98°C for 15 seconds, 52°C for 20 seconds, 72°C for 30 seconds, and final extension at 72°C for 5 minutes.

Sample triplicates were pooled, and the pools were purified using the Qiagen MinElute PCR Purification kit (cat.: 28004) according to the Qiagen protocol. Sample pools were run on a 1% agarose gel and desired fragments (~380 bp) were cut from gel. Excised gel fragments underwent the Qiagen QIAquick Gel Extraction kit (cat.: 28706) following kit instructions. Pools were normalized to 4nM, denatured using 0.5N NaOH, and diluted to a final concentration of 10pM. 15% PhiX control was added to sequencing pool. Final pool was loaded onto an Illumina MiSeq Next Generation Sequencer (cat.: SY-410-1003) using the Illumina MiSeq Reagent 2x250bp v2 Kit (cat.: MS-102-2003) protocol. All 60 working samples, six extraction blanks, two PCR blanks, and two *E. coli* 10X positive controls (totaling 70 samples) were loaded onto the MiSeq as the loading pool. These were the only samples on this MiSeq run.

Data Analysis

Raw MS and MS/MS files were converted to mzXML format using MSConvert (Chambers et al. 2012). MZMine v2.37 was used to identify MS features (Pluskal et al. 2010). MZMine parameters are depicted in supplementary table 5. PCoA plots were

created using a Canberra distance matrix with the ClusterApp online program (http://dorresteinappshub.ucsd.edu:3838/clusterMetaboApp0.9.1/). Molecular networking and library spectral database searches were completed using the Global Natural Products Social Molecular Networking (Wang et al. 2017), also known as GNPS, on the mgf file exported from MZMine. As GNPS only works with MS/MS data, only the MS/MS files were uploaded. GNPS parameters were: precursor and fragment ion mass tolerance: 0.02 Da, minimum cosine score for networking and library matches: 0.7, minimum number of matched MS2 fragment ions for networking and library matches: 4, network topK: 50, maximum connected component size: 100, maximum shift between precursors: 500 Da, analog search: enabled, maximum analog mass difference: 100 Da, precursor window filtering: enabled, 50 Da peak window filtering: enabled, normalization per file: row sum normalization. Results were analyzed by evaluating mirror plot similarity, cosine score, and plausibility of matches.

16S rRNA gene sequences were downloaded from Illumina BaseSpace sequence hub (http://basespace.illumina.com). Raw file outputs gave a unique Sample ID with three numbers (e.g.: Samp254). These were matched to the RCBC number corresponding to the unique reverse barcode for each sample and utilized for demultiplexing. AdapterRemoval v2 (Schubert et al. 2016) was used to trim and merge Read1 and Read2 files with a quality score equal to or greater than 30 phred. Next, these files were collapsed into a single file containing all the merged reads. Quantitative Insights Into Microbial Ecology 1 (QIIME 1) was employed for operational taxonomic unit (OTU) picking (Caporaso et al. 2010). Closed-reference OTU picking was performed using the

EzTaxon Database (Chun et al. 2007) to identify taxonomy and generate a biom file. This OTU picking method was selected because it uses reference sequences for alignment, which allows faster speed, high-quality taxonomic selections and phylogenetic trees, and nonoverlapping amplicons can be compared (Rideout et al. 2014).

Summarizing the biom file revealed all working samples and *E. coli* standards contained at least 12,000 reads. None of our negative controls contained 12,000 reads, so 12,000 was selected as the rarefaction depth. Following rarefaction, phyla and genera level taxonomy of the remaining 60 samples and two *E. coli* standards were processed using Microsoft Excel to calculate the relative frequencies and abundances of bacterial taxa. The top five most abundant phyla and top fifteen genera were identified and plotted using Microsoft Excel.

For alpha diversity analysis, a phylogenetic tree was generated using the EzTaxon Database (Chun et al. 2007). MAFFT (Katoh and Standley 2013) was used to align sequences within a mapping file containing a single representative sequence for each OTU. Next, FastTree (Price, Dehal, and Arkin 2009) was used within QIIME to generate phylogenetic trees based off the aligned OTU sequences. QIIME then generated alpha diversity results from the rarefied biom file and phylogenetic tree. Average abundance of specific taxa was calculated by summing up the total taxa within each sample and dividing the count for each taxon by the total amount. Alpha diversity analyses focused on the top five abundant phyla and top 15 abundant genera across all samples. For beta diversity, QIIME was employed to create unweighted and weighted UniFrac distance matrices. These correspond to presence/absence and abundance of bacterial taxa,

respectively. Next, Principal Coordinate Analysis (PCoA) using these UniFrac distance matrices, as implemented in QIIME, was performed, followed by utilizing Emperor within QIIME to generate three-dimensional PCoA plots. Additional analysis was performed in R version 3.5.3 (R Core Team 2018) using the R package ggplot2 to create boxplots, scatter plots, and two-dimensional PCoA plots (Wickham 2016). Statistical significance of results was determined using the FSA package in R to perform the Kruskal-Wallis test by ranks and Dunn's test on alpha diversity files (Ogle 2017). Dunn's test was employed to validate the findings from significant Kruskal-Wallis tests. Analysis of covariance (ANCOVA) was done in R to test the effects of treatments while controlling for covariates on alpha diversity files. Permutational multivariate analysis of variance (PERMANOVA) was used in QIIME to evaluate statistical significance of beta diversity data. Results from these statistical tests can be found in supplementary table 6. The data from these taxonomic summaries, alpha diversity, and beta diversity analyses were used to inform conclusions about microbiome profiles of our samples.

CHAPTER 3

RESULTS

Metabolome Preservation

While we were able to successfully employ a RNA*later* cleanup protocol to generate untargeted LC-MS/MS data, our RNA*later* cleanup protocol did not appear to preserve metabolomic profiles. The internal standards (sulfachloropyridazine and sulfadimethoxine) were detected, but GNPS spectral library searches identified six total metabolites (not including internal standards) across all samples and stages. Generally, MS-based metabolomics projects have several hundred total identified metabolites, with ranges including 180-860 metabolites depending on the project and its goals (Loftfield et al. 2016; Sankaranarayanan et al. 2015; Turroni et al. 2016; Wang et al. 2018). Since only six metabolites were detected for this project, this indicates metabolomic profiles within samples were not preserved properly. This was true for all samples ran on the MS, irrespective of the number of RNA*later* removals performed.

While the GNPS spectral library search detected six total metabolites, only one compound was a proper match based on mirror plot similarity (Figure 3) and cosine score: urobilinogen. According to the Human Metabolome Database, urobilinogen is a parent compound of the pigment stercobilin, which is known to give feces its color (Wishart et al. 2018). Since urobilinogen is a fecal metabolite, its presence within our samples is expected. However, urobilinogen was only detected in seven samples. Of these seven total samples, five samples had a single RNA*later* removal.

PCoA plots generated using Canberra distance matrices demonstrate that while we were unable to preserve metabolomic profiles, sample treatment still influenced our metabolomic content (Figure 4A). Samples that underwent two RNA*later* cleanups clustered very tightly together, whereas one-time RNA*later* removal samples did not cluster together as strongly. Also, these plots do not show clustering according to storage temperature (Figure 4B). Furthermore, these PCoA plots did not position our negative blank near any of our samples (Figures 4A and 4B). This indicates samples were not completely devoid of metabolites, even though the entire metabolomic profile was not preserved.

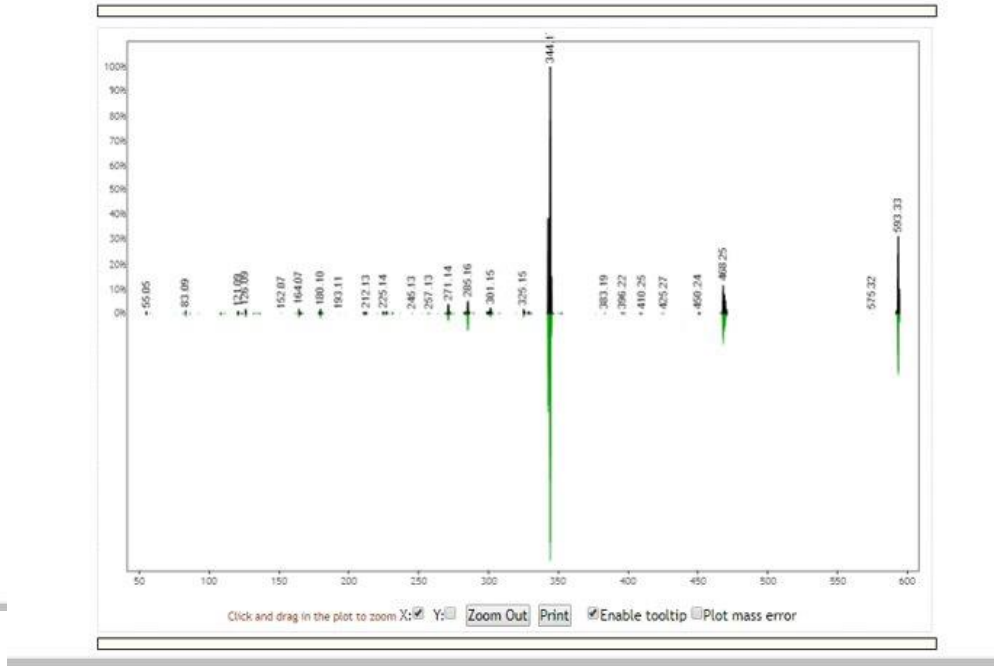


Figure 3. Mirror plot of Urobilinogen from GNPS.

This screenshot was taken from the GNPS website. In GNPS, mirror plots are used to evaluate whether a match from the spectral library database is valid. The bottom bars in green are the raw peaks associated with the matched molecular from the GNPS library. The top black bars are the peaks from the users' submitted data that GNPS believes are a mach. Both sets of peaks are placed on top of each other to easily identify if they match. The size and placement of peaks must be similar for the match to be considered valid. In this case, the peaks from our data and the library data are nearly identical. Therefore, this molecular is a valid match to our data.

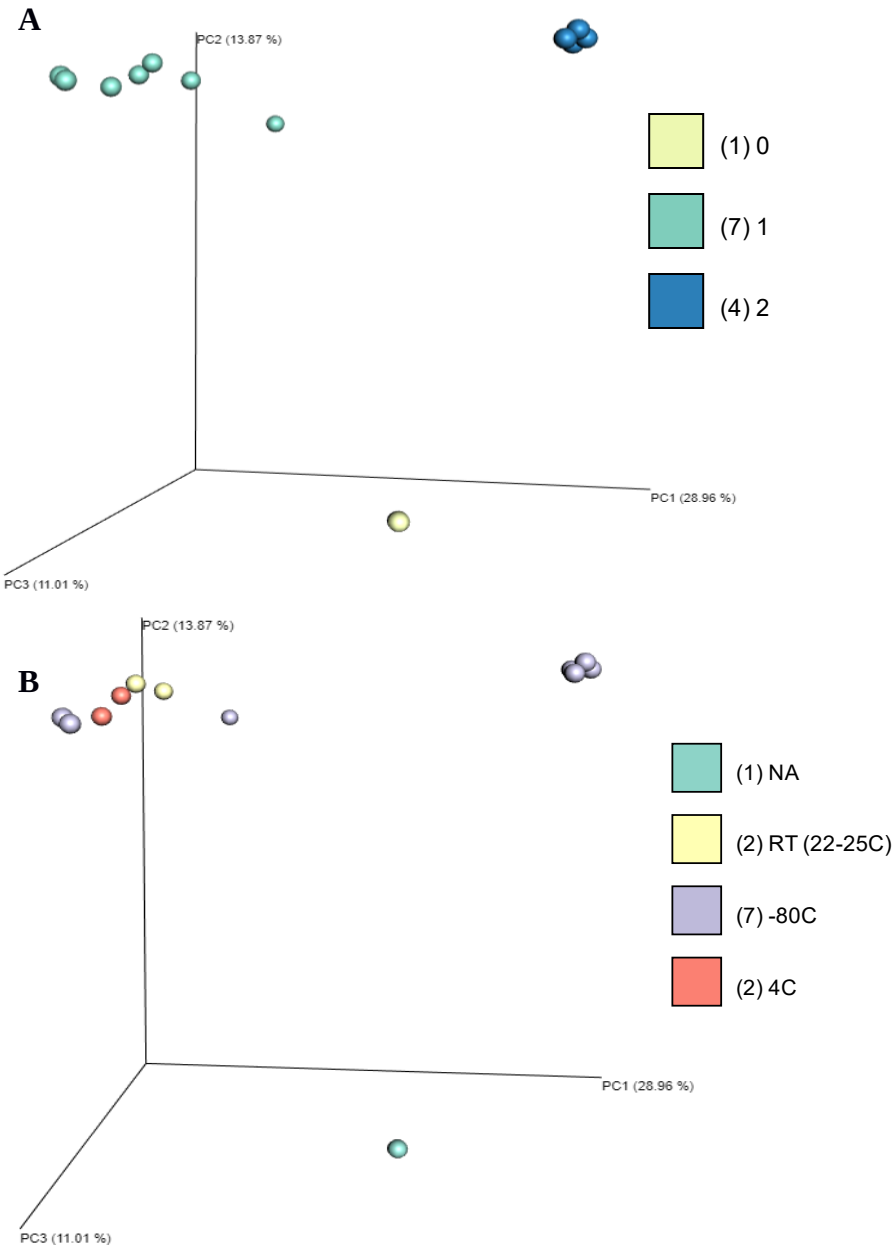


Figure 4. Three-dimensional PCoA plots of metabolomics samples. Generated from Canberra distance matrices.

(**A**) Color-coded number of RNA*later* cleanups by: dark blue samples (2 RNA*later* cleanups), green samples (1 RNA*later* cleanup), and tan (blank, 0 RNA*later* cleanup). Samples that had two RNAlater cleanups clustered together very tightly. On the other hand, samples that underwent a single RNAlater removal did not group together as strongly. These samples clustered in the same general area of the PCoA plot, but did not occupy the same space, unlike the samples that underwent two RNA*later* cleanup protocols. Negative blank located at bottom indicates RNA*later* removal did not deplete all metabolites from samples.

(**B**) Color-coded storage temperature by: green (NA, blank), tan (room temperature, 22-25°C), silver (-80°C), and red (4°C). Samples did not readily cluster due to temperature. The group of clustered -80°C samples were all samples that went through two RNA*later* cleanups, which likely caused the clustering.

Gut Microbiome Profile

DNA from all 60 samples was successfully extracted, amplified, and sequenced. Of the total 12,588,428 reads from the Illumina MiSeq, 11,704,428 reads passed filtering criteria. Approximately 25% of these passing filter reads mapped to our PhiX control according to the MiSeq, resulting in a total 7,765,530 16S reads identified as the samples in this study.

Microbiome taxonomic inventories were first investigated by identifying the most abundant phyla within samples. Samples exhibited high levels of the phyla Firmicutes (76.6% average abundance), Actinobacteria (14.6% average), Bacteroidetes (3.3% average), Proteobacteria (2.5% average), and Euryarchaeota (1.3% average). These specific phyla are the most abundant across all samples (Figure 5). Other identified phyla include Tenericutes, Cyanobacteria, Verrumicrobia, and more (Supplementary table 7A). These remaining phyla are varyingly distributed across samples, as shown by Figure 5.

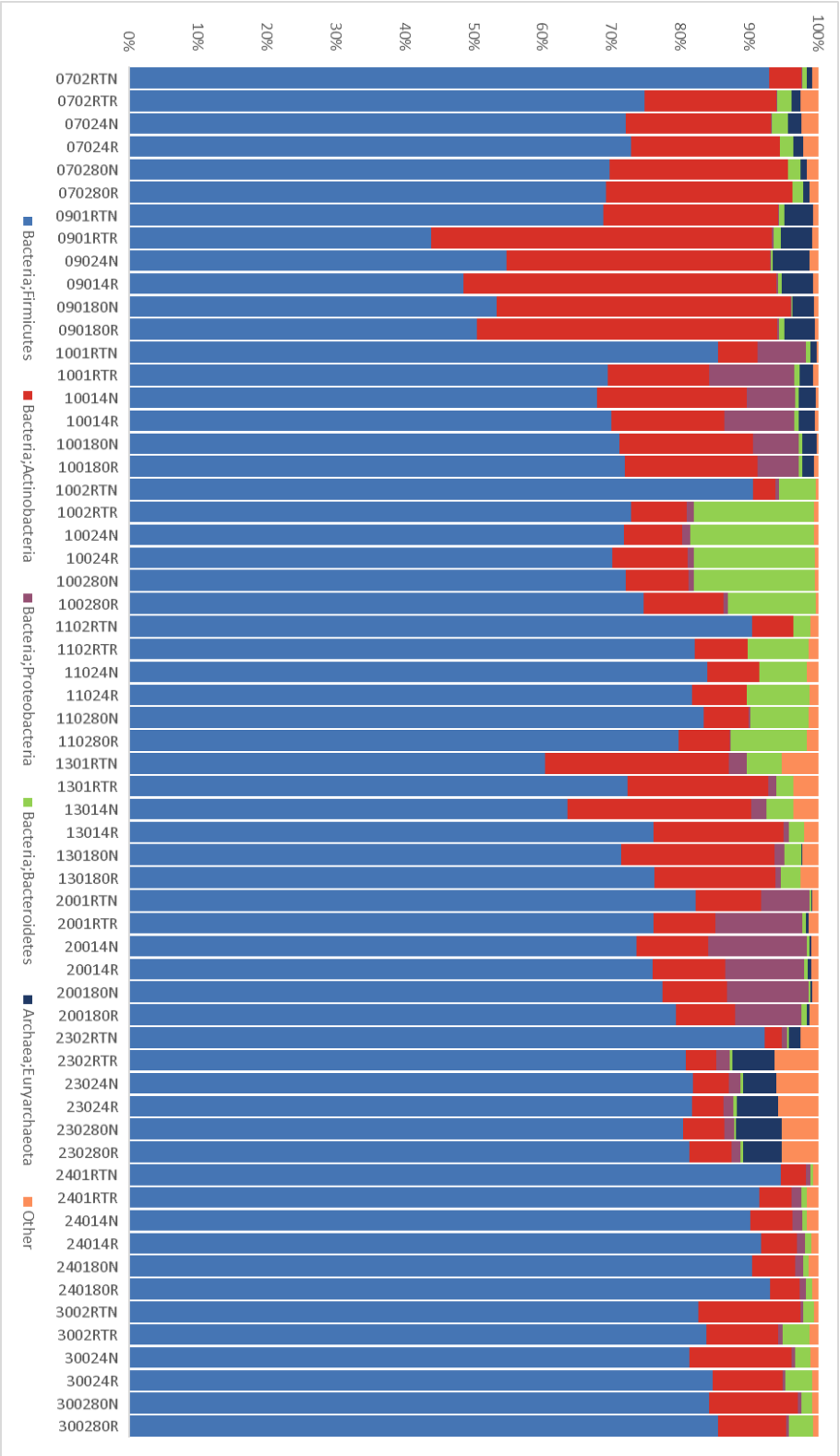


Figure 5. Phylum-level taxonomic summaries. Colored by top five most abundant phyla. Remaining detected phyla were grouped as "Other". Each column represents a sample, sorted by individual. Firmicutes dominates all samples, followed by Actinobacteria, Bacteroidetes, Proteobacteria, and Euryarchaeota. Distribution of phyla varies between samples. Sample names can be broken down into host, storage temperature, and RNA*later*. The first four numbers refer to the family number and individual within that family (e.g.: 0702= family 07, family member 02). These are always individuals. RT/4/80 refer to storage at room temperature (RT), 4°C (4), and -80°C (80), respectively. The final letter, N/R, refers to use of RNA*later*. N means No and R means RNA*later* (Yes). For example, sample 0901RTN is from family 09, family member 02, stored at room temperature, and no RNA*later* treatment.

Next, the top 15 most abundant genera were identified (Figure 6). These top genera include *Blautia* (10.3% average abundance), *Clostridium* (8.3% average), *Collinsella* (6.4% average), *Subdoligranulum* (4.1% average), and *Pediococcus* (3.6% average), to name a few (Supplementary table 7B). Overall, the distribution of genera was highly varied within these samples. For phyla, all samples generally had the same top five abundant phyla with only a small percentage of other phyla (such as Tenericutes or Cyanobacteria). At the genus level, samples usually contained many more genera outside the top 15, showing a greater range of taxa within samples. Of the top 15 most abundant genera, 12 belonged to the phylum Firmicutes. After characterizing the bacterial taxonomic inventories of our samples, the bacterial diversity within samples was then explored.

Alpha diversity analysis of samples indicates the use of RNA*later* affected phylogenetic diversity and microbial richness within samples. RNA*later*-treated samples generally had higher counts of observed bacterial species (Figure 7A) and increased phylogenetic diversity (Figure 7B). When examining samples based on the individual sample donor, the RNA*later* samples exhibited higher microbial richness and phylogenetic diversity compared to non-RNA*later* samples from the same host. The Kruskal-Wallis test by ranks followed by Dunn's test confirmed this observation was statistically significant (both p-values=0.01). Furthermore, alpha diversity analyses also demonstrated that sample taxonomic inventories were largely impacted by the host sample donor. For both microbial richness and phylogenetic diversity, host-based differences were statistically significant (both p-values=0).

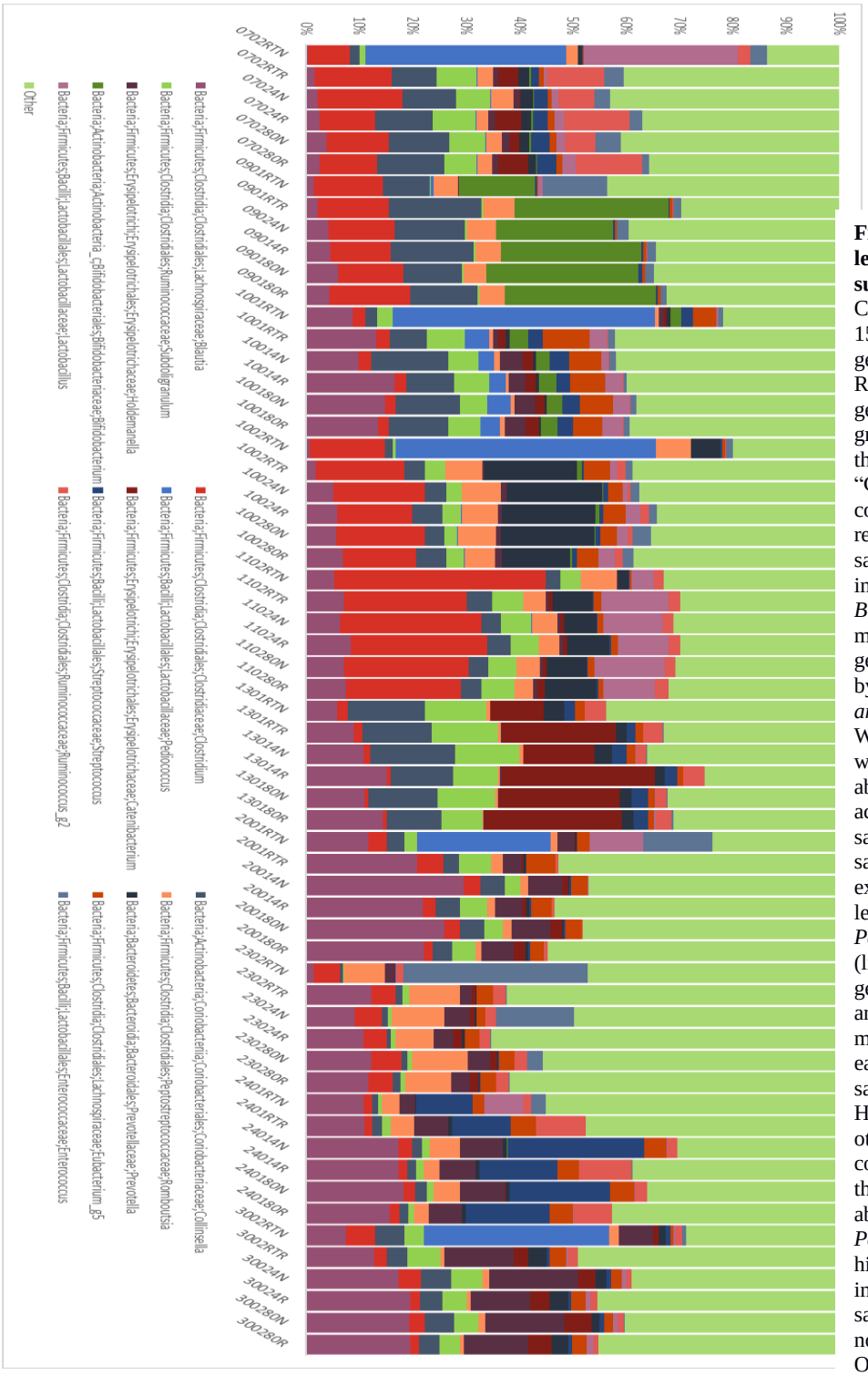
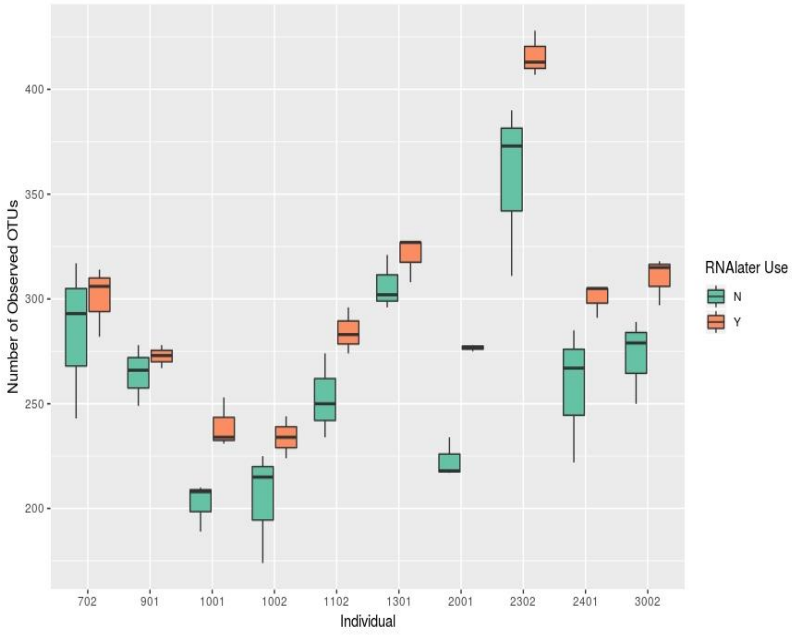


Figure 6. Genus-level taxonomic summaries. Colored by top 15 abundant genera. Remaining genera were grouped under the category "Other". Each column represents a sample, sorted by individual. *Blautia* is the most abundant genus followed by *Clostridium, and Collinsella*. While *Blautia* was the most abundant genus across all samples, five samples exhibited high levels of *Pediococcus* (light blue). This genus was far-and-away the most abundant in each of these samples. However, no other samples contained more than 1% abundance of *Pediococcus.* Its high abundance in these five samples is noteworthy. Other genera are distributed

A



B

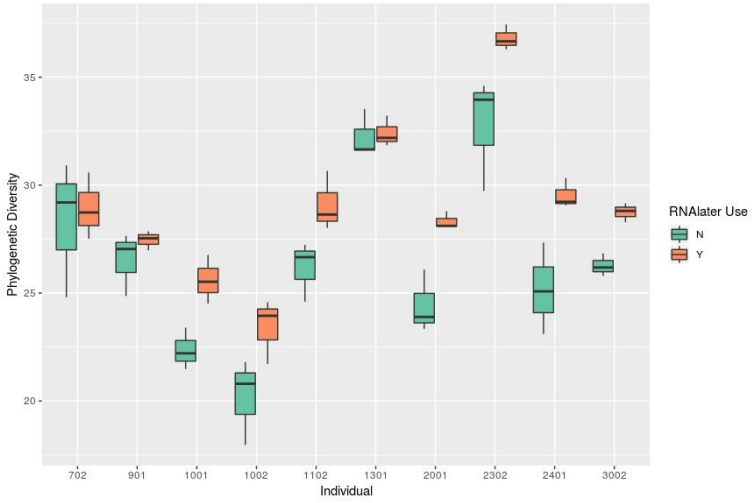


Figure 7. Boxplots of alpha diversity analyses.

Color-coded RNA*later* use by: green (No) and red (Yes). Non-RNA*later* samples tended to show wider ranges of diversity.

(**A**) Microbial richness of samples from all hosts increased in RNA*later* samples. Number of observed species varied dramatically between hosts.

(**B**) Phylogenetic diversity of samples from all hosts generally increased in RNA*later* samples. Number of observed species for each host were very different. Trends exhibited here are like those from Figure 7A.

Despite these observations from RNA*later* and individual sample donors, there were no significant differences in microbial richness or phylogenetic diversity due to storage temperature (p-values=0.62 and 0.53, respectively). RNA*later* samples exhibited increased microbial richness (Figure 8A). When examining the effects of storage temperature in alpha diversity analyses without considering RNA*later use*, all temperatures had relatively equal values for microbial richness and phylogenetic diversity (Figure 8B). However, inspecting both temperature and RNA*later* use together reveals that non-RNA*later* samples stored at room temperature had decreased microbial richness and phylogenetic diversity (Figure 8C).

Beta diversity analyses report that differences between samples were primarily due to host. PCoA plots using both weighted and unweighted UniFrac distances demonstrate strong clustering based on host donor (Figures 9 and 10, respectively). After utilizing PERMANOVA on both weighted and unweighted UniFrac distance matrices (weighted p-value=0.001, unweighted p-value=0.001), these differences were found to be statistically significant. However, seven samples did not cluster due to individual donor based on weighted UniFrac distances (Figure 9B). Instead, these seven samples all did not receive RNA*later* treatment and were kept at room temperature. Interestingly, five of the seven outliers contained the genus *Pediococcus* in high abundance (Figures 6, 10, 11, 12, 13, and Table 2). These five samples had the highest levels of *Pediococcus* abundance from any sample, and it was the most abundant genus in these five samples. PCoA plots using unweighted UniFrac distances do not show these seven outliers (Figures 10A and 10B).

Figures 9 and 10 illustrate samples did not cluster according to RNA*later* use or storage temperature (except for the seven outliers). Nevertheless, PERMANOVA analysis indicates RNA*later* use (weighted p-value=0.001, unweighted p-value=0.002) and storage temperature (weighted p-value=0.001, unweighted p-value=0.001) caused statistically significant differences between samples. Therefore, individual sample donor was the primary force affecting differences between samples, but RNA*later* and storage temperature slightly affected microbiome profiles. These effects were outweighed by the influence of the host.

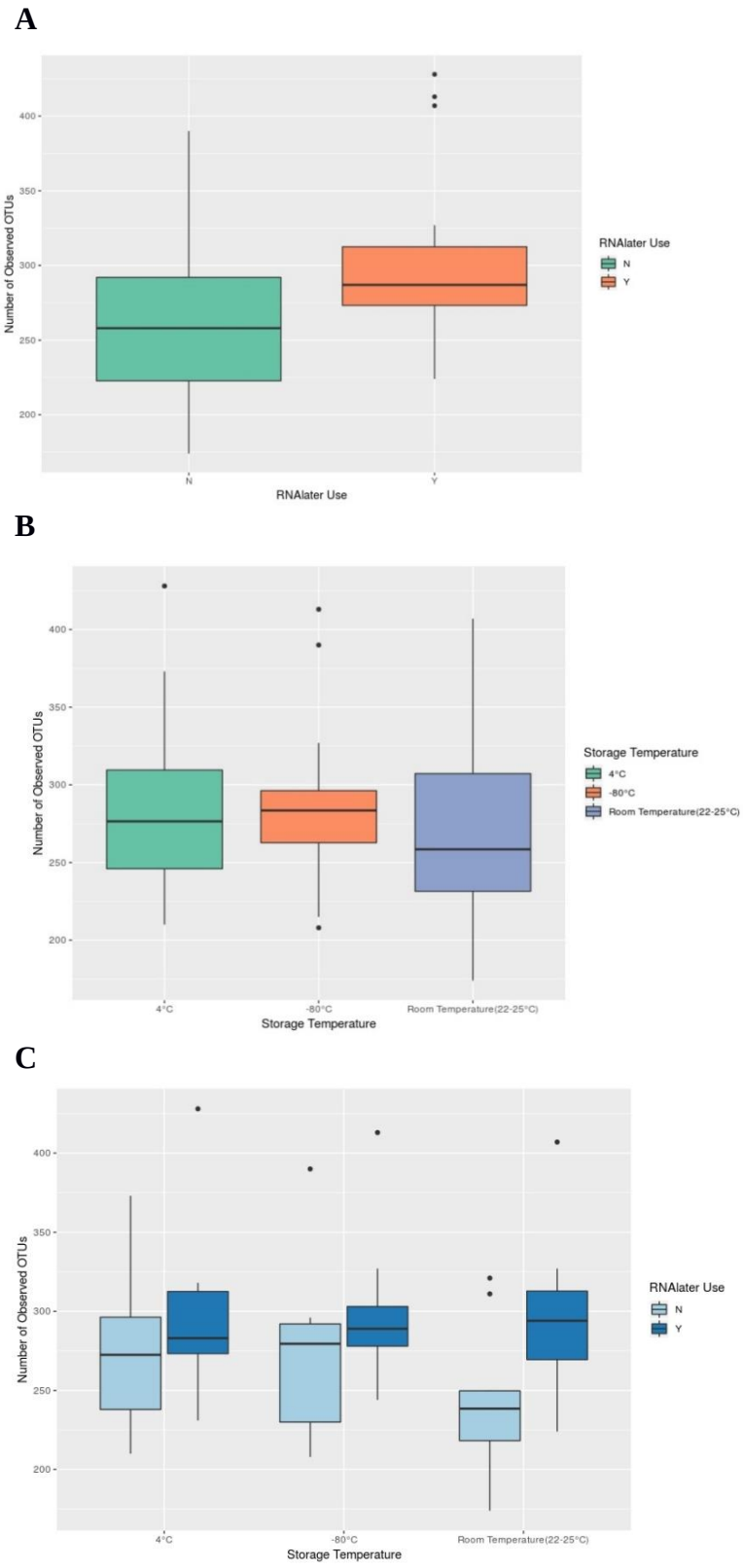


Figure 8. Boxplots of alpha diversity analyses by sample treatment method.

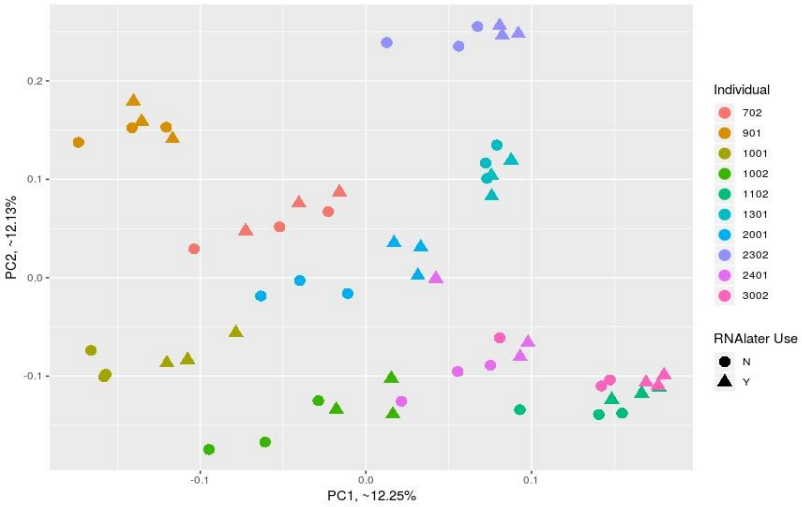
Alpha diversity analyses of storage temperature and RNA*later* use.

(A) Color-coded RNA*later* use by: green (No) and red (Yes). RNA*later*-treated samples had increased microbial richness, although non-RNA*later* samples had a wider range.

(B) Color-coded storage temperatures by: green (4°C), red (-80°C), and blue (room temperature, 22-25°C). All storage temperatures exhibited similar microbial richness.

(**C**) Color-coded RNA*later* use by: light blue (No) and dark blue (Yes). Storage temperature on x-axis. Microbial richness was relatively similar for samples regardless of treatment method. However, RNA*later* samples displayed increased microbial richness regardless of storage temperature. The exceptions are non-RNA*later* samples kept at room temperature, where these samples exhibited reduced microbial richness.

A



B

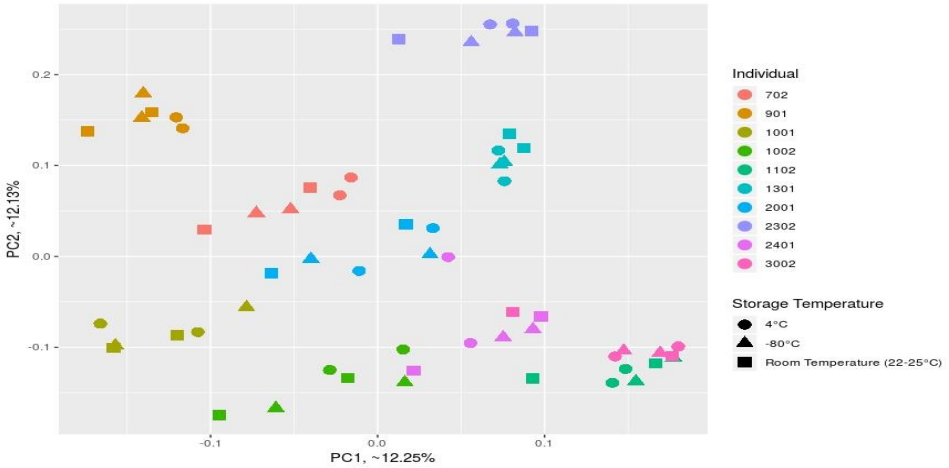


Figure 9. Two-dimensional PCoA plots using unweighted UniFrac distances.

Color-coded individual by: red (702), orange (901), brown (1001), dark green (1002), light green (1102) teal (1301), sky blue (2001), dark blue (2302), purple (2401), and pink (3002). These PCoA plots were generated from beta diversity analyses used to create unweighted UniFrac distance matrices. Differences in presence and absence of taxa were primarily due to host. Samples generally clustered with other samples from their host. Moreover, host clusters were occasionally distinct from each other rather than packing together closely.

(A) Shape-coded RNA*later* use by: circle (No) and triangle (Yes). Differences in presence and absence of taxa were primarily due to host rather than RNA*later* use. Samples did not appear to cluster due to the usage of RNA*later*.

(**B**) Shape-coded storage temperature by: circle (4°C), triangle (-80°C), and square (room temperature, 22-25°C). Unweighted UniFrac plots show samples generally clustered according to host instead of storage temperature. Samples did not cluster based on storage temperature.

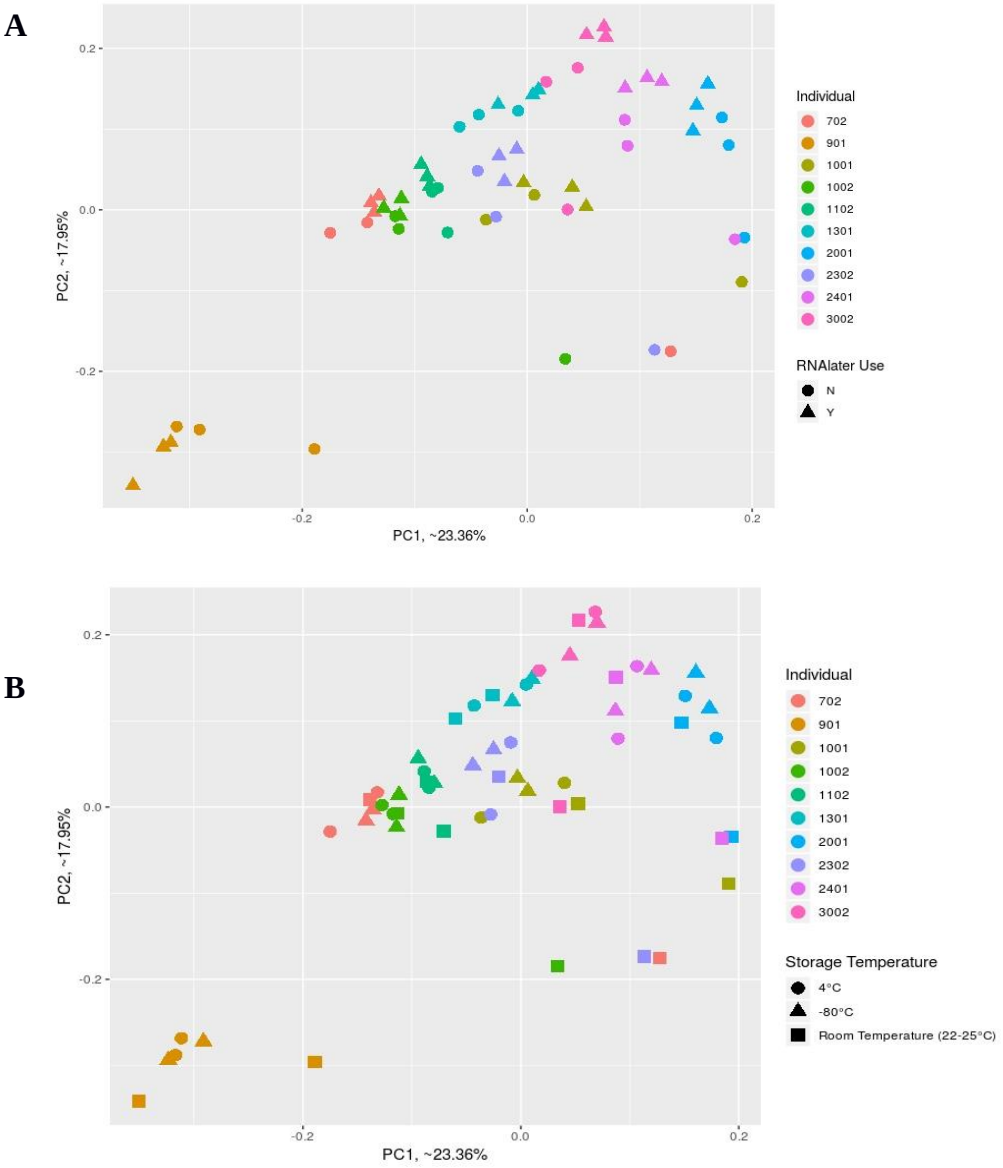


Figure 10. Two-dimensional PCoA plots using weighted UniFrac distances.

Similar to Figure 6 except these plots were created using weighted UniFrac distance matrices. Color-coded individual by: red (702), orange (901), brown (1001), dark green (1002), light green (1102) teal (1301), sky blue (2001), dark blue (2302), purple (2401), and pink (3002). While samples generally clustered based on host differences, almost all the samples were packed tightly together in a large group. There is much more overlap between hosts than in the unweighted PCoA plots.

(A) Shape-coded RNA*later* use by: circle (No) and triangle (Yes).This PCoA plot shows how samples primarily clustered according to host when examining differences between abundance of taxa. RNA*later* use had little effect on clustering. However, within host groups, there appeared to be slight clustering due to RNA*later*.

(**B**) Shape-coded storage temperature by: circle (4°C), triangle (-80°C), and square (room temperature, 22-25°C). Host differences affecting the abundance of bacterial taxa drove sample clustering more than storage temperature. Some clustering due to storage temperature appears to occur within host groups.

Sample	*Pediococcus* Relative Abundance (%)
0702RTN	37.65
1001RTN	49.14
1002RTN	48.78
2001RTN	25.03
2302RTN	0.025
2401RTN	8.33e-5
3002RTN	34.66

Table 2. Seven weighted UniFrac outliers.

Five of these seven outliers showed *Pediococcus* at high abundance, especially compared to other samples. For these each of these five samples, *Pediococcus* was the most abundant genus.

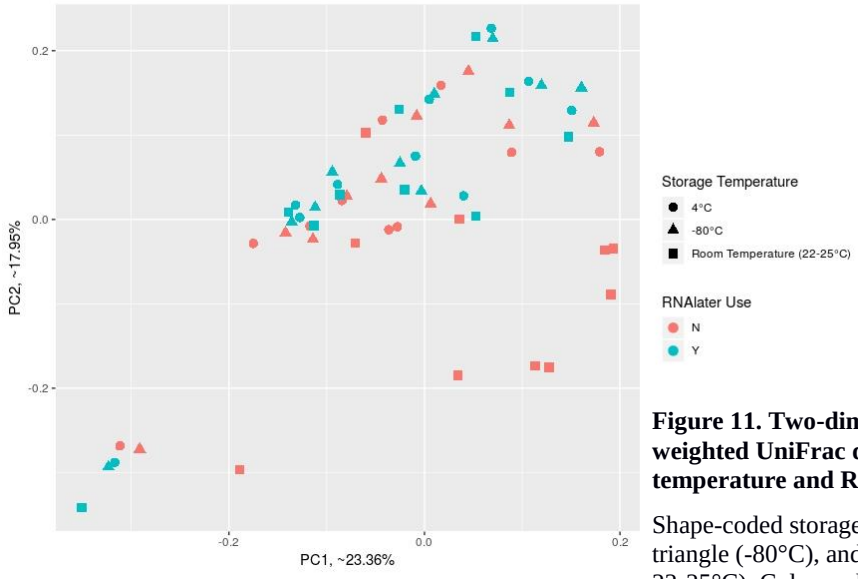


Figure 11. Two-dimensional PCoA plot from weighted UniFrac distances with storage temperature and RNA*later*.

Shape-coded storage temperature: circle (4°C), triangle (-80°C), and square (room temperature, 22-25°C). Color-coded RNA*later* use by: red (No) and blue (Yes). This PCoA plot using weighted UniFrac distance matrices shows samples largely do not cluster according to storage temperature or RNA*later* use. However, this plot highlights the seven outliers located in the lower right half of the plot. These outliers are the only points between 0.0 & 0.2 PC1 and 0.0 & -0.2 PC2. The seven outliers clustered closer to each other than to other samples from their hosts. Five of these outliers had high levels of *Pediococcus*. Clearly the sample treatment methods affected these samples differently than the others.

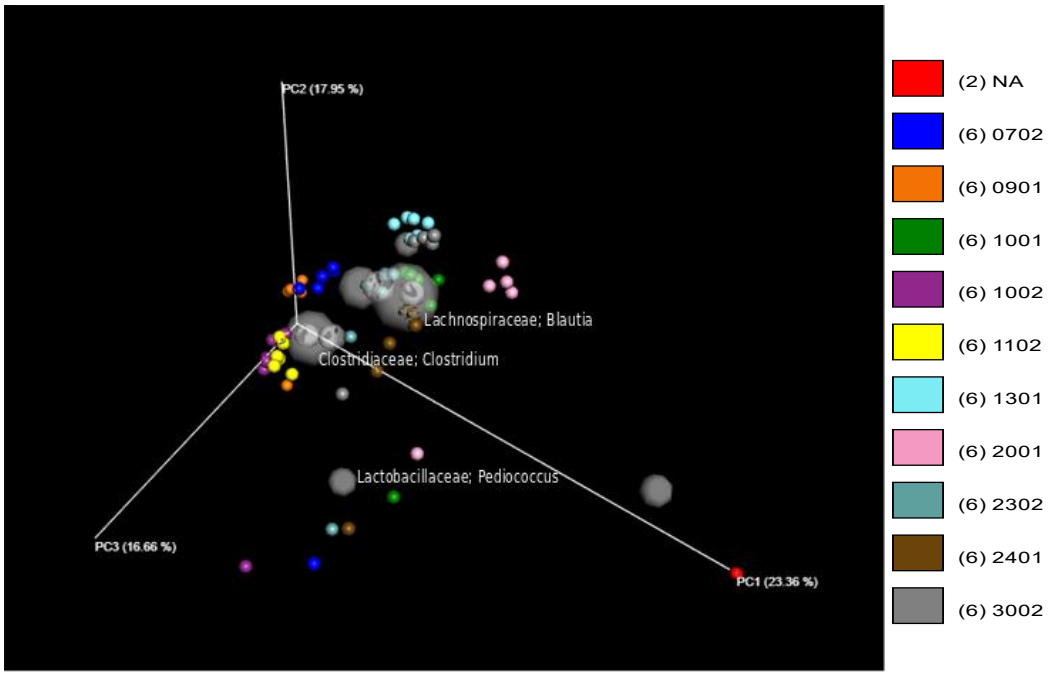


Figure 12. Three-dimensional PCoA biplot from weighted UniFrac with genera added.

Color-coded individual by: red (blank), dark blue (0702), orange (0901), green (1001), purple (1002), yellow (1102), light blue (1301), pink (2001), teal (2302), brown (2401), and grey (3002). Semi-transparent gray clouds correspond to the top ten abundant genera. The locations of these clouds on the PCoA plot correspond to the abundance of each genera within samples. This indicates which genera are driving the clustering of samples. Size of the cloud correlates with the abundance of the specific genus within samples. A small number of genera were labeled for simplicity. *Blautia* and *Clostridium* were the most abundant genera, shown by the size of their clouds. Most of the genera are found close together in the large cluster of samples. Meanwhile, *Pediococcus* was the fifth most abundant genera, but was only found near the seven outliers. These seven outliers are in center of PC axes. These are the pink, green, light blue, dark blue, purple, and grey single samples. Other samples that appear close are due to the captured angle (Figure 11 indicates these seven are dissimilar from other samples). These outliers are closest to the *Pediococcus* cloud, indicating this genus primarily affects their clustering.

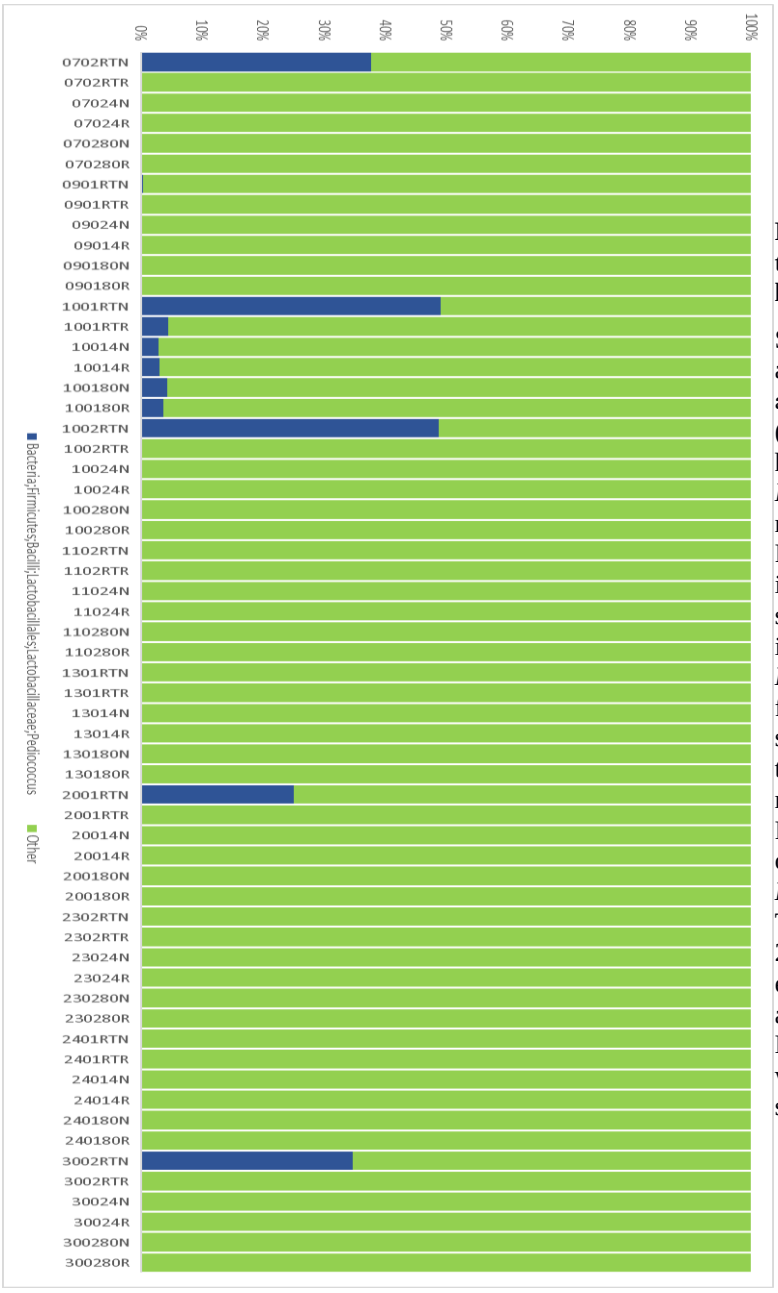


Figure 13. Genus-level taxonomic summaries limited to *Pediococcus*.

Similar to Figure 5 except all non-*Pediococcus* genera are grouped as "Other" (green). This highlights the high abundance of *Pediococcus* in the five room temperature, non-RNA*later* outliers and how it is rarely found in other samples. Interestingly, individual 1001 had <5% *Pediococcus* abundance in five of their six total samples, but these were not to the same degree as the room temperature, non-RNA*later* sample. The five outliers each had 25-50% *Pediococcus* abundance. The remaining two outliers, 2302RTN and 2401RTN, each had less than .03% abundance of *Pediococcus*. Nonetheless, they clustered with high *Pediococcus* samples.

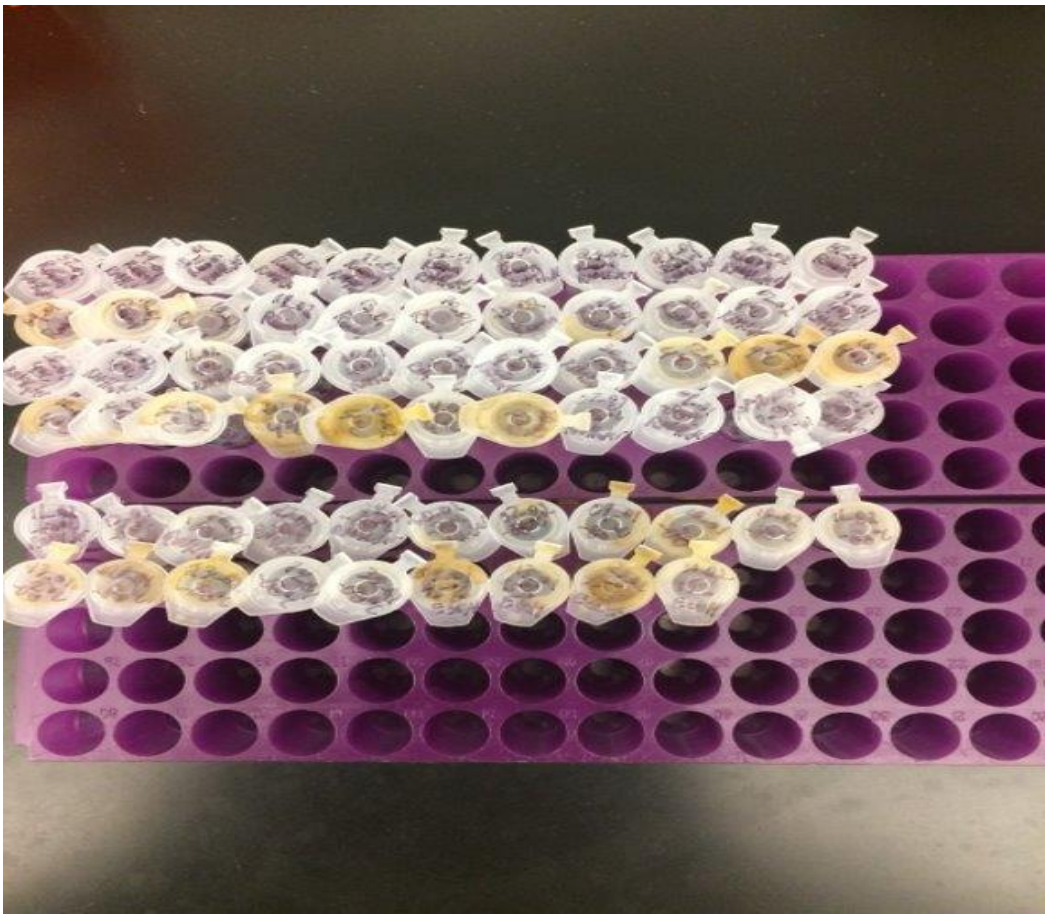


Figure 14. Sample photograph prior to MS injection.
This photograph depicts all 60 samples used for MS analysis. Tubes with color are all samples that were not treated with RNAlater and therefore did not undergo an RNAlater removal protocol. The tubes without any color are samples that underwent the RNAlater removal protocol twice. This image highlights the distinct lack of color samples had following two RNAlater removals. However, only 12 of the RNA*later* samples were ultimately analyzed in the MS.

CHAPTER 4

DISCUSSION

Metabolome Preservation

The goal of this study was to examine how the effects of storage temperature and RNA*later* treatment impact fecal sample integrity through taxonomic classification with 16S rRNA gene sequencing and mass spectrometry-based metabolomics. In the metabolomics approach, we were unable to generate data from 12 samples that were treated with RNA*later* and subsequently underwent a RNA*later* removal process. Only six total metabolites were detected in database searches with only urobilinogen as a confirmed match. RNA*later* removal is the likely culprit influencing MS analysis and sample preservation due to its components commonly interfering with MS analysis. Based on the work done for this project, there are two possible explanations for this lack of metabolomic data.

The first viable answer is that the components of RNA*later* mixing with fecal samples prevented MS analysis. Other researchers have identified this as the primary cause behind metabolomic problems with RNA*later* (Sinha et al. 2016; Wang et al. 2018). However, this explanation is unlikely when considering the detection of urobilinogen and our use of internal standards. While not every tested sample contained urobilinogen, its detection in some samples indicates MS analysis performed properly. Additionally, all samples had been spiked with internal standards (sulfachloropyridazine and sulfadimethoxine) with known masses. Their detection by the MS instrument further demonstrates MS analysis performed normally. With both expected molecules and

internal standards detected within samples, this suggests MS detection performed as expected. This was likely due to the RNA*later* removal having succeeded in removing the negative contents of RNA*later*, which ties into the other potential reason for our results.

The second possible, and more likely, cause for our results is the design of the modified RNA*later* removal protocol. It was intended to isolate RNA*later* components without compromising other analytes, but the protocol was likely too thorough when removing molecules. The components of the removal protocol appear to bias polar molecules due to usage of methanol for elution. Methanol, a polar solvent, would release more polar molecules from the sorbent during elution, potentially causing a low recovery of nonpolar molecules. Moreover, methanol was used previously in the RNA*later* cleanup protocol as part of the washing step. Using methanol as both a washing solvent and an elution solvent likely caused target molecules to wash away rather than bind to the cartridge for subsequent elution. However, urobilinogen is a nonpolar molecule and it was detected within our samples. Therefore, the RNA*later* removal protocol's inclination towards polar compounds does not adequately explain why few metabolites were detected, but it cannot be ruled out as a potential factor.

The SPE protocol changed the physical color of samples, which can represent the loss of metabolites. Samples generally retained their original color after elution in the first RNA*later* removal. For stage one samples, the second SPE treatment caused color loss for all eluates. Some stage two samples kept their color following their single RNA*later* removal (Figure 14). According to the Human Metabolome Database (Wishart

et al. 2018), some metabolites, such as stercobilin and urobilinogen, are associated with feces coloring. The lack of color after RNA*later* removal suggests these molecules, as well as other color-causing compounds, were lost in RNA*later* removal. This observation illustrates the thoroughness of the RNA*later* removal protocol, which likely caused analyte loss. Urobilinogen was found more commonly in samples that underwent a single RNA*later* removal. The second RNA*later* removal largely eliminated urobilinogen from samples, reinforcing the idea that the RNA*later* removal was too effective. Additionally, PCoA plots showed samples clustered strongly based on the number of RNA*later* removals performed (Figures 4A and 4B). Samples that went through the protocol once were grouped together, but samples that underwent the protocol twice occupied the same space as each other. Therefore, undergoing the RNA*later* cleanup protocol resulted in similar metabolomic profiles due to molecule loss, which increased with additional RNA*later* cleanup. Lastly, the fact our MS analysis performed normally demonstrates that our RNA*later* removal process was too effective. Since RNA*later* ordinarily prevents MS analysis from occurring (Loftfield et al. 2016; Sinha et al. 2016), we would have expected our MS analysis to completely fail when using RNA*later* samples. Even through our protocol did not effectively preserve metabolomic profiles, it appeared to remove enough RNA*later* products for proper MS analysis. All in all, there is strong evidence advocating the design of our RNA*later* removal process caused near-total analyte loss within our samples, impacting metabolomic profile preservation. With this conclusion, it is apparent this RNA*later* removal protocol is impractical for MS-based metabolomics.

The overall lack of data is consistent with the results of published literature (Loftfield et al. 2016; Sinha et al. 2016; Wang et al. 2018). These specific projects evaluated different sample storage and treatment methods, including RNA*later*, to determine their impact on fecal microbiome and metabolomic profiles. All these groups were unable to generate untargeted metabolomic data from samples treated with RNA*later* using a MS instrument. However, none of the other projects attempted to remedy the problems RNA*later* causes for untargeted metabolomics. To our knowledge, this thesis is the first project to explore ways of modifying RNA*later* treatment to allow for fecal untargeted MS-based metabolomics. While the RNA*later* cleanup protocol used here was developed from Cottrell et al. 2015, they employed targeted MS analysis on mouse bone fracture callus samples, whereas this project focused on untargeted screenings of the global fecal metabolomic profile. The work by Cottrell et al. 2015 signals RNA*later* treatment can be tweaked to allow for MS analysis, but the RNA*later* cleanup protocol used for this thesis was unable to achieve similar success for fecal untargeted MS-based metabolomics.

While the RNA*later* removal protocol did not preserve metabolomic profiles, this result shows the limits of storage methods for metabolomics. To improve these limits, more research into the effects of RNA*later* treatment on sample integrity and whether this storage method can be improved is necessary. Future research can start by evaluating the modified RNA*later* removal protocol provided in this thesis. While we could not definitively pinpoint the exact cause or causes of why our RNA*later* removal protocol failed, future work can address these questions. This project focused on analyzing

RNA*later*-treated samples but employing MS analysis on non-RNA*later* samples is crucial for understanding how RNA*later* impacts sample preservation. Future work for this project should include MS analyses for samples treated with and without RNA*later*. These non-RNA*later* controls are crucial for future work on this thesis. Next, selected non-RNA*later* samples should also undergo the removal protocol. This would demonstrate how the protocol impacts samples and we could compare how RNA*later* and non-RNA*later* samples differ after undergoing the treatment. By understanding how the RNA*later* removal protocol affects samples on a molecular level, the protocol can be improved. Modifications to the protocol could improve untargeted metabolite yield from samples treated with RNA*later*, possibly enabling RNA*later* usage with metabolomic samples. Utilizing RNA*later* with metabolomic samples would allow researchers to perform more types of molecular analyses on a single sample, providing deeper insights for molecular anthropology, including microbiome studies.

Microbiome Profile Preservation

In this thesis, 16S rRNA gene sequencing was performed to investigate the effects of RNA*later* and different storage temperatures on the bacterial taxonomic inventories of fecal samples. Our most abundant phyla were Firmicutes, Actinobacteria, Bacteroidetes, Proteobacteria, and Euryarcheota, and our most abundant genera included *Blautia*, *Clostridium*, *Collinsella*, *Subdoligranulum*, and *Pediococcus* (Figures 5 and 6). These phyla and genera are expected within microbiome profiles of fecal samples (Choo et al. 2015), indicating our results are consistent with published research. While immediate

freezing without storage solutions is accepted as the gold standard for sample preservation (Choo et al. 2015; Loftfield et al. 2016), our results suggest that the host was the primary factor influencing beta diversity of samples (Figures 9 and 10), with these PCoA plots demonstrating strong clustering when color-coded by host. Samples did not cluster according to RNA*later* use or storage temperature (Figures 9 and 10), but RNA*later* treatment and storage temperature caused samples to be different from each other (these effects were simply outweighed by the host) when examining beta diversity in PCoA plots. Each host had six total samples, but these samples were not identical to each other in the PCoA plots (Figures 9 and 10). Furthermore, samples generally clustered closer to other samples from the same host rather than a different host. Therefore, host differences greatly influenced inter-sample differences in presence/absence of taxa as well as the overall abundance of taxa.

Alpha diversity analyses also indicated that microbial diversity of samples was largely influenced by the host who contributed the sample. Kruskal-Wallis and Dunn's test determined the host had a significant influence on both microbial richness and phylogenetic diversity of samples when considered alone (Supplementary tables 6A and 6B, respectively). ANCOVA tests incorporating both RNA*later* use and storage temperature as covariates reveal that differences in microbial content and abundance between samples were primarily determined by the host (Supplementary Table 6C). RNA*later* use and storage temperature still provided some influence but with less pronounced effects (Supplementary table 6C). PERMANOVA tests further confirmed the host differences to be statistically significant (Supplementary table 6D). Moreover, this

conclusion matches findings from other research where sample treatment influences were overshadowed by the host (Choo et al. 2015; Fouhy et al. 2015; Ribeiro et al. 2018; Sinha et al. 2016; Wang et al. 2018). Ultimately, it seems that the host was the primary force affecting the presence/absence and abundance of bacterial taxa. This holds true when evaluated by itself and when RNA*later* and storage temperature were considered.

Alpha diversity analyses reported usage of RNA*later* by itself had statistically significant effects on both microbial richness and phylogenetic diversity of samples, with Kruskal-Wallis and Dunn's tests determining statistical significance (Supplementary tables 6A and 6B, respectively). Furthermore, ANCOVA tests validate that RNA*later* had significant effects on microbial richness and phylogenetic diversity, even when using storage temperature as a covariate (Supplementary table 6C). For eight of the ten host sample donors, RNA*later*-treated subsamples had higher levels of microbial richness and phylogenetic diversity compared to their non-RNA*later* counterparts (Figures 7A, 7B, and 8A). This conclusion differs from results of another project, where RNA*later* was noted to have less microbial diversity than frozen samples (Dominianni et al. 2014). Our differing results might be caused by the seven *Pediococcus* outliers skewing the data or this thesis's samples having their microbiome profiles altered after being frozen prior to 16S analysis (Bahl et al. 2012). Future work is needed to examine how freeze-thaw cycles impact microbiome profiles, but our data suggests RNA*later* affected the microbiome profile of samples, even when considered with storage temperature, although this effect was less pronounced than that of the host.

Additionally, alpha diversity analyses demonstrated that storage temperature did not have statistically significant effects on microbial richness or phylogenetic diversity when considered as a sole factor. Conversely, Figure 8C indicates that room temperature samples had less microbial richness and phylogenetic diversity compared to the 4°C and -80°C samples. However, only the non-RNA*later* samples kept at room temperature exhibited this decrease (Figure 8C). A Kruskal-Wallis test and Dunn's test both reported the storage temperature differences were not statistically significant (Supplementary tables 6A and 6B, respectively), but both tests only examined storage temperature by itself. Figure 8B supports this statistical finding, as all storage temperatures displayed similar levels of microbial richness when RNA*later* was ignored. It seems that 4°C and -80°C can be effective storage temperatures without RNA*later*, but samples will exhibit decreased microbial richness and phylogenetic diversity (Figure 8C). Thus, samples stored at 4°C and -80°C will be better preserved when treated with RNA*later*. Samples kept at room temperature without RNA*later* show even greater decrease in microbial richness and phylogenetic diversity (Figure 8C). This loss could be due to certain bacterial taxa blooming when stored at room temperature, known as microbial blooming, but it is still unclear (Amir et al. 2017). Despite this, our findings demonstrate storage at 4°C or -80°C have similar preservation effects, but RNA*later* will better preserve microbial richness and phylogenetic diversity at 4°C and -80°C and is critical when storing samples at room temperature.

One noteworthy feature of the weighted PCoA plots (Figures 10, 11, and 12) show seven outliers separating from the main group of samples clustering by host. These

figures had samples clustering by host, but samples usually grouped together in the same region of the plot rather than more varied distribution. In the unweighted PCoA plots, there was no single large group of sample clustering (Figure 9). Samples organized by host, but these host groups were more indiscriminately spread across the plot. These seven outliers from the weighted PCoA plots were all from different hosts, were not treated with RNA*later*, and were stored at room temperature. Examining the taxonomic inventories of these samples reveals that five had high levels of the genus *Pediococcus*, making it the fifth most abundant genus across all samples (Figures 6 and 13). The remaining two outliers both had less than 0.025% relative abundance of *Pediococcus* (Table 2).

Pediococcus is a genus of Gram-positive member of the Lactobacillaceae family, commonly associated with sauerkraut fermentation (Courage 2019; Woese 1987). As a lactic-acid bacterium, *Pediococcus* is frequently found in the human gut microbiome and plays varying roles, such as gluten metabolism (Caminero et al. 2014), probiotics in animal-based diets (David et al. 2013), and reductions of the genus have been linked to cirrhosis (Schnabl and Brenner 2014). However, its prominence in our data is unusual as is not normally a highly abundant genus and is not a focus of fecal microbiome studies.

Weighted PCoA biplots using the abundance of taxonomic genera further indicate that *Pediococcus* drove the clustering of these outliers (Figure 12). While not all seven outliers contained high levels of *Pediococcus*, its remarkable abundance in these room-temperature, non-RNA*later* samples is noteworthy (Figure 13 and Table 2). Moreover, the fact that *Pediococcus* was not detected in the RNA*later*-treated, room temperature

samples from these same hosts hints at a relationship between *Pediococcus* and RNA*later*. Since *Pediococcus* was only found in non-RNA*later* samples, something in RNA*later* inhibited *Pediococcus* growth. To our knowledge, this thesis project is the first to notice the potential connections between *Pediococcus*, high abundance in samples kept at room temperature, and RNA*later*.

Microbial blooming is one possible explanation for the high abundance of *Pediococcus* (Amir et al. 2017). In the Amir et al. 2017 study, Lactobacillaceae were identified as candidate bloomers at room temperature. However, *Pediococcus* was not described in the article. Two-week storage at room temperature for this thesis possibly caused *Pediococcus* to bloom for non-RNA*later* samples, but RNA*later* prevented *Pediococcus* blooming. Species-level identification of bacterial taxa corresponding to *Pediococcus* could possibly explain why this genus was found in such high abundance under these conditions. Regardless, our treatment methods clearly affected these seven sample differently than the rest, with the abundance of *Pediococcus* being a likely cause. Future research can further explore these relationships with the *Pediococcus* genus.

Previous research has indicated that some bacteria, particularly Gram-positive bacteria like Firmicutes (such as *Pediococcus*) and Actinobacteria, will be more abundant in frozen samples than fresh samples (Bahl et al. 2012). This change is caused by alterations to the cellular structure of Gram-positive bacteria. With the samples used in this thesis having undergone multiple freeze-thaw cycles earlier, it is possible this process altered the microbiome profiles of these samples before work for this thesis began, such as causing Gram-positive bacteria to be found in higher abundance than in fresh samples.

This could explain how storage temperature had such minimal negative effects on microbiome profiles. If samples were already impacted by multiple freeze-thaws, their response to treatment methods would be different to samples that were only frozen once. This could also clarify our results contrast with published research. Future work can utilize fresh samples as well as samples that underwent varying levels of freeze-thaws to track how microbiome profiles change due to storage and treatment methods. Tracking the effects of freeze-thaws in this way can shed further light on how current methods, including the gold standard of immediate freezing, might be improved. Moreover, researchers can examine how certain bacterial taxa are better preserved in samples that are fresh, frozen, or treated with storage solutions such as RNA*later*. By understanding exactly how these preservation methods impact sample integrities, researchers can better understand the biological meaning of their results, examine how samples are affected on a molecular level, and ensure high-quality data.

In conclusion, this thesis demonstrates the crucial nature of evaluating how results are affected by sample treatment and molecular preservation. While we were unable to preserve the metabolome profiles of our samples, the preservation of microbiome profiles in our samples indicates that samples treated with RNA*later* will be better preserved than samples without RNA*later*, regardless of sample storage. For samples not treated with RNA*later*, 94°C and -80°C will have similar effects on preservation, but room temperature storage requires RNA*later* to avoid compromising the microbiome profile. Our results suggest researchers sampling at field sites can utilize RNA*later* as an alternative to immediate freezing without dramatically compromising the microbiome

integrities of these samples, but more work is needed as our samples had been frozen and thawed previously. Moreover, our results validate the essential role that sample treatment plays in multi-omics and molecular anthropology projects. Simply by modifying how samples were stored and treated, we were unable to generate data for metabolomics, an entire field of research. Our efforts to counter this issue failed, but they highlight a core problem of sample preservation: choosing how one treats and preserves their samples can cut off access to lines of inquiry. Furthermore, whatever method is chosen will still affect sample integrity. Ideally, a single sample should allow for multiple lines of analysis, but sample storage and treatment processes affect different types of molecules in different ways. Current sample preservation methods require researchers to balance the pros and cons of these methods in order to generate the data they are interested in. By identifying how these methods affect samples molecularly and how these methods can be improved, different molecular analysis can be done on a single sample. This idea of molecular taphonomy, referring to the study of how molecules are preserved in samples, must be explored in greater detail in order to get as much biological information from a sample as possible. By expanding our knowledge of molecular taphonomy, sample preservation methods, and their recurring issues, these improvements can advance sample collection and storage methods for multi-omics and molecular anthropology studies.

References

Ambion | RNA by Life Technologies. 2014. RNAlater® Tissue Collection: RNA Stabilization Solution [Online]. *Ambion Protocols and Manuals*:1-12. URL: https://www.thermofisher.com/order/catalog/product/AM7022.

Amir A., McDonald D., Navas-Molina J.A., Debelius J., Morton J.T., Hyde E., Robbins-Pianka A., and Knight R. 2017. Correcting for Microbial Blooms in Fecal Samples during Room-Temperature Shipping. *American Society for Microbiology* 2(2):1–5.

Ayala, F.J. 1995. The Myth of Eve: Molecular Biology and Human Origins. *Science* 270:1930-1936.

Amann R.I., Ludwig W., and Schleifer K-H. 1995. Phylogenetic Identification and In Situ Detection of Individual Microbial Cells without Cultivation. *Microbiological Reviews* 59(1):143-169. PMID:7535888.

Bahl M.I., Bergström A., and Rask Licht T. 2012. Freezing fecal samples prior to DNA extraction affects the Firmicutes to Bacteroidetes ratio determined by downstream quantitative PCR analysis. *FEMS Microbiology Letters* 329(2):193–197. DOI:10.1111/j.1574-6968.2012.02523.x.

Benezra A., DeStefano J. and Gordon J.I. 2012. Anthropology of microbes. *Proceedings of the National Academy of Sciences* 109(17):6378-6381. DOI: 10.1073/pnas.1200515109.

Bino R.J., Hall R.D., Fiehn O., Kopka J., Saito K., Draper J., Nikolau B.J., Mendes P., Roessner-Turnali U., Beale M.H., Trethewey R.N., Lange B.M., Wurtele E.S., and Summer L.W. 2004. Potential of metabolomics as a functional genomics tool. *Trends in Plant Science* 9(9):418-425. DOI: 10.1016/j.tplants.2004.07.004.

Blaser M.J., and Falkow S. 2009. What are the consequences of the disappearing human microbiota? *Nature Reviews Microbiology* 7(12):887-894. DOI: http://dx.doi.org/10.1038/nrmicro2245.

Caminero A., Herrán A.R., Nistal E., Pérez-Andrés J., Vaquero L., Vivas S., Ruiz de Morales J.M.G., Albillos S.M., and Casqueiro J. 2014. Diversity of the cultivable human gut microbiome involved in gluten metabolism: isolation of microorganisms with potential interest for coeliac disease. *FEMS Microbial Ecology* 88:309-319. DOI: 10.1111/1574-6941.12295.

Caporaso J.G., Kuczynski J., Stombaugh J., Bittinger K., Bushman F.D., Costello E.K., Fierer N., Gonzalez Peña A., Goodrich J.K., Gordon J.I., Huttley G.A., Kelley S.T., Knights D., Koenig J.E., Ley R.E., Lozupone C.A., McDonald D., Muegge B.D., Pirrung M., Reeder J., Sevinsky J.R., Turnbaugh P.J., Walters W.A., Widmann J., Yatsunenko T., Zaneveld J., and Knight R. 2010. QIIME allows analysis of high-throughput community sequencing data. *Nature Methods* 7(5):1–4. DOI:10.1038/nmeth.f.303.QIIME.

Caporaso J.G., Lauber C.L., Walters W.A., Berg-Lyons D., Huntley J., Fierer N., Owens S.M., Betley J., Fraser L., Bauer M., Gormley N., Gilber J.A., Smith G., and Knight R. 2012. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *The ISME Journal* 6:1621-1624. DOI: 10.1038/ismej.2012.8

Chambers M.C., Maclean B., Burke R., Amodei D., Ruderman D.L., Neumann S., Gatto L., Fischer B., Pratt B., Egertson J., Hoff K., Kessner D., Tasman N., Shulman N., Frewen B., Baker T.A., Brusniak M-Y., Paulse C., Creasy D., Flashner L., Kani K., Moulding C., Seymour S.L., Nuwaysir L.M., Lefebvre B., Kuhlmann F., Roark J., Rainer P., Detlev S., Hemenway T., Huhmer A., Langridge J., Connolly B., Chadick T., Holly K., Eckels J., Deutsch E.W., Moritz R.L., Katz J.E., Agus D.B., MacCoss M., Tabb D.L., and Mallick P. 2012. A cross-platform toolkit for mass spectrometry and proteomics. *Nature Biotechnology* 30(10):918–920. DOI:10.1038/nbt.2377.

Cheng Z., Zhou X., Li W., Hu B., Zhang Y., Xu Y., Zhang L., and Jiang H. 2016. Optimization of solid-phase extraction and liquid chromatography-tandem mass spectrometry for simultaneous determination of capilliposide B and its activate metabolite in rat urine and feces: Overcoming nonspecific binding. *Journal of Pharmaceutical and Biomedical Analysis* 131:6-12. DOI: 10.1016/j.jpba.2016.08.003.

Choo J.M., Leong L.E.X., and Rogers G.B. 2015. Sample storage conditions significantly influence faecal microbiome profiles. *Scientific Reports* 5:1–10. DOI:10.1038/srep16350.

Chun J., Lee J.H., Jung Y., Kim M., Kim S., Kwon Kim B., and Woon Lim Y. 2007. EzTaxon: A web-based tool for the identification of prokaryotes based on 16S ribosomal RNA gene sequences. *International Journal of Systematic and Evolutionary Microbiology* 57(10):2259–2261. DOI:10.1099/ijs.0.64915-0.

Clarridge III 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.

Coskun, O. 2016. Separation techniques: Chromatography. *Northern clinics of Istanbul* 3(2):156-160. DOI: 10.14744/nci.2016.32757.

Cottrell J.A., Lin H-N., and O'Connor J.P. 2015. Method for measuring lipid mediators, proteins, and messenger RNAs from a single tissue specimen. *Analytical Biochemistry* 469:34-42. DOI: 10.1016/j.ab.2014.10.004.

Courage K.H. 2019. Cultured: How Ancient Foods Can Feed our Microbiome, First Edition. *Avery Publishing:* New York, New York. Hardcover. ISBN:9781101905289.

David L.A., Maurice C.F., Carmody R.N., Gootenberg D.B., Button J.E., Wolfe B.E., Ling A.V., Devlin A.S., Varma Y., Fischbach M.A., Biddinger S.B., Dutton R.J., Turnbaugh P.J. 2014. Diet rapidly and reproducibly alters the human gut microbiome. *Nature* 505(7484):559-563. DOI:10.1038/nature12820.

Dawes E.A., and Ribbons D.W. 2003. The Endogenous Metabolism of Microorganisms. *Annual Reviw of Microbiology* 16(1):241-264. DOI:10.1146/annurev.mi.16.100162.001325.

DeBerardinis R.J., and Thompson C.B. 2012. Cellular metabolism and disease: what do metabolic outliers teach us? *Cell* 148(6):1132-1144. DOI:10.1016/j.cell.2012.02.032.Cellular

Dettmer K., Aronov P.A., and Hammock B.D. 2006. Mass Spectrometry-Based Metabolomics. *Mass Spectrometry Reviews* 26:51-78. DOI:10.1002/mas.

Dominianni C., Wu J., Hayes R.B., and Ahn J. 2014. Comparison of methods for fecal microbiome biospecimen collection. *BMC Microbiology* 14(1):1–6. DOI:10.1186/1471-2180-14-103.

Dunn W.B., Erban A., Weber R.J.M., Creek D.J., Brown M., Breitling R., Hankemeier T., Goodacre R., Neumann S., Kopka J., and Viant M.R. 2013. Mass appeal: metabolite identification in mass spectrometry-focused untargeted metabolomics. *Metabolomics* 9:44-66. DOI: 10.1007/s11306-012-0434-4.

Fouhy F., Deane J., Rea M.C., O'Sullivan Ó., Ross R.P., O'Callaghan G., Plant B.J., and Stanton C. 2015. The effects of freezing on faecal microbiota as determined using miseq sequencing and culture-based investigations. *PLoS ONE* 10(3):1–12. DOI:10.1371/journal.pone.0119355.

Fox G.E., Magrum L.J., Balch W.E., Wolfe R.S., and Woese C.R. 1977. Classification of methanogenic bacteria by 16S ribosomal RNA characterization. *Proceedings of the National Academy of Sciences* 74(10):4537-4541. DOI: 10.1073/pnas.74.10.4537.

Fiehn, O. 2002. Metabolomics -- the link between genotypes and phenotypes. *Planet Molecular Biology* 48(1-2):155-171. DOI: 10.1007/978-94-010-0448-0_11.

Gorokhova E. 2005. Effects of preservation and storage of microcrustaceans in RNA*later* on RNA and DNA degradation. *Limnology and Oceanography: Methods* 3:143-148. DOI:10.4319/lom.2005.3.143

Greaves J., and Roboz J. 2014. Mass Spectrometry for the Novice. *Taylor & Francis Group, LLC. CRC Press*. International standard book number-13:978-1-4200-9418-3 (Paperback).

Grice E.A., and Segre J.A. 2012. The Human Microbiome: Our Second Genome. *Annual Review of Genomics and Human Genetics* 13:151-170. DOI: 10.1146/annurev-

genom-090711-163814.

Hardy K., Buckley S., Collins M.J., Estalrrich A., Brothwell D., Copeland L., García-Tabernero A., García-Vargas S., de la Rasilla M., Lalueza-Fox C., Huguet R., Bastir M., Santamaría D., Madella M., Wilson J., Fernández Cortés Á., and Rosas A. 2012. Neanderthal medics? Evidence for food, cooking, and medicinial plants entrapped in dental calculus. *Naturwissenschaften* 99:617-626. DOI:10.1007/s00114-012-0942-0.

Hollywood K., Brison D.R., and Goodacre R. 2006. Metabolomics: Current technologies and future trends. *Proteomics* 6(17):4716-4723. DOI: 10.1002/pmic.200600106.

Johnson C.H., and Gonzales F.J. 2012. Challenges and Opportunities of Metabolomics. *Journal of Cellular Physiology* 227(8):2975-2981. DOI: 10.1002/jcp.24002.

Jovel J., Patterson J., Wang W., Hotte N., O'Keefe S., Mitchel T., Perry T., Kao D., Mason A.L., Madsen K.L., and Wong G.K-S. 2016. Characterization of the Gut Microbiome using 16S or Shotgun Metagenomics. *Frontiers in Microbiology* 7(459):1-17. 10.3389/fmicb.2016.00459

Jurmain R., Kilgore L., Trevathan L., and Ciochon R.L. 2013. Introduction to Physical Anthropology, 14th Edition. *Cengage Learning*: Boston, MA. Paperback. ISBN-10:1285061977.

Kaestle F.A., and Horsburgh K.A. 2002. Ancient DNA in Anthropology: Methods, Applications, and Ethics. *Yearbook of Physical Anthropology* 119:92-130. DOI: 10.1002/ajpa.10179.

Katoh K., and Standley D.M. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution* 30(4):772–80. DOI:10.1093/molbev/mst010.

Kim M., and Chun J. 2014. 16S rRNA Gene-Based Identification of *Bacteria* and *Archaea* using the EzTaxon Server. *Methods in Microbiology* 41:61-74. Academic Press. DOI: 10.1016/bs.mim.2014.08.001.

Kolbert C.P., and Persing D.H. 1999. Ribosomal DNA sequencing as a tool for identification of bacterial pathogens. *Current Opinion in Microbiology* 2:299-305. DOI: 10.1016/S1369-5274(99)80052-6.

Life Technologies. 2013. RNA*later* Safety Data Sheet; Product Code.:F7022 [Online]. Published October 29 2013. URL: https://www.thermofisher.com/order/catalog/product/AM7022.

Lader, E. 2001. United States Patent 6,204,375 B1: Methods and Reagents for Preserving RNA in Cell and Tissue Samples. Ambion, Inc., issued March 20, 2001.

Loftfield E., Vogtmann E., Sampson J.N., Moore S.C., Nelson H., Knight R., Chia N., and Sinha R. 2016. Comparison of collection methods for fecal samples for discovery metabolomics in epidemiologic studies. *Cancer Epidemiology Biomarkers and Prevention* 25(11):1483–1490. DOI:10.1158/1055-9965.EPI-16-0409.

Ma S., Yim S.H., Lee S-G., Kim E.B., Lee S-R., Chang K-T., Buffenstein R., Lewis K.N., Park T.J., Miller R.A., Clish C.B., and Gladyshev V.N. 2015. Organization of the Mammalian Metabolome according to Organ Function, Lineage Specialization, and Longevity. *Cell Metabolism* 22:332-343. DOI:http://dx.doi.org/10.1016/j.cmet.2015.07.005

Marks, J. 2002. What is Molecular Anthropology? What Can it Be? *Evolutionary Anthropology* 11:131-135. DOI: 10.1002/evan.10031

Matsuda F. 2016. Technical Challenges in Mass Spectrometry-Based Metabolomics. *Mass Spectrometry* 5(2): 10.5702/massspectrometry.S0052.

Michalski A., Damoc E., Hauschild J-P., Lange O., Wieghaus A., Makarov A., Nagaraj N., Cox J., Mann M., and Horning S. 2011. Mass Spectrometry-based Proteomics Using Q Exactive, a High-performance benchtop Quadrupole Orbitrap Mass Spectrometer. *Molecular & Cellular Proteomics* 10(9):1-11. DOI: 10.1074/mcp.M111.011015.

Moeller A.H., Li Y., Mpoudi Ngole E., Ahuka-Mundeke S., Lonsdorf E.V., Pusey A.E., Peeters M., Hahn B.H., and Ochman H. 2014. Rapid changes in the gut microbiome during human evolution. *Proceedings of the National Academy of Sciences* 111(46):16431-16435. DOI: 10.1073/pnas.1419136111.

Nicholson G., Rantalainen M., Maher A.D., Li J.V., Malmodin D., Ahmadi K.R., Faber J.H., Hallgrímsdóttir I.B., Barrett A., Toft H., Krestyaninova M., Viksna J., Neogi S.G., Dumas M-E., Sarkans U., the MolPAGE Consortium, Silverman B.W., Donnelly P., Nicholson J.K., Allen M., Zondervan K.T., Lindon J.C., Spector T.D., McCarthy M.I., Holmes E., Baunsgaard D., and Holmes C.C. 2011. Human metabolic profiles are stably controlled by genetic and environmental variation. *Molecular Systems Biology* 7:1-11. DOI: 10.1038/msb.2011.57.

Obregon-Tito A.J., Tito R.Y., Metcalf J., Sankaranarayanan K., Clement J.C., Ursell L.K., Zech Xu Z., Van Treuren W., Knight R., Gaffney P.M., Spicer P., Lawson P., Marin-Reyes L., Trujillo-Villarroel O., Foster M., Guija-Poma E., Troncoso-Corzo L., Warinner C., Ozga A.T., and Lewis C.M. 2015. Subsistence strategies in traditional societies distinguish gut microbiomes. *Nature Communications* 6:1-9. DOI: http://dx.doi.org/10.1038/ncomms7505.

Ogle, D.H., Wheeler P., and Dinno A. 2017. *FSA: Fisheries Stock Analysis.* R package version 0.8.22.9000, https://github.com/droglenc/FSA.

Outram A.K. 2008. Introduction to experimental archaeology. *World Archaeology* 40(1):1-6. DOI: 10.1080/00438240801889456.

Ovchinnikov I.V., Götherström A., Romanova G.P., Kharitonov V.M., Lidén K., and Goodwin W. 2000. Molecular analysis of Neanderthal DNA from the north Caucasus. *Nature* 404(6777):490-493. DOI: 10.1038/35006625.

Pace N.R. 1997. A Molecular View of Microbial Diversity and the Biosphere. *Science Magazine* 276:734-740. DOI:10.1126/science.276.5313.734

Patti G.J., Yanes O., and Siuzdak G. 2012. Metabolomics: the apogee of the omic trilogy. *Nature Reviews: Molecular Cell Biology* 13(4):263-269. DOI: 10.1038/nrm3314.

Pereira F., Carnerio J., Matthiesen R., van Asch B., Pinto N., Gusmão L., and Amorim A. 2010. Identification of species by multiplex analysis of variable-length sequences. *Nucleic Acids Research* 38(22):1-17.DOI: 10.1093/nar/gkq865.

Pluskal T., Castillo S., Villar-Briones A., and Orešič M. 2010. MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics* 11. DOI:10.1186/1471-2105-11-395.

Prakash O., Verma M., Sharma P., Kumari K., Singh A., Kumari H., Jit S., Gupta S.K., Khanna M., and Lal R. 2007. Polyphasic approach of bacterial classification -- An overview of recent advances. *Indian Journal of Microbiology* 47(2):98-108. DOI: 10.1007/s12088-007-0022-x.

Price M.N., Dehal P.S., and Arkin A.P. 2009. FastTree: Computing Large Minimum Evolution Trees with Profiles instead of a Distance Matrix. *Molecular Biology and Evolution* 26(7):1641-1650. DOI: doi:10.1093/molbev/msp077.

R Core Team. 2018. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/

Radini A., Buckley S., Estalrrich A., de la Rasilla M., and Hardy K. 2016. Neanderthals, trees and dental calculus: new evidence from El Sidrón. *Antiquity* 90(350):290-301. DOI:10.15184/aqy.2016.21.

Rappaport S.M., and Smith M.T. 2010. Environment and Disease Risks. *Science* 330(6003):460-461. DOI: 10.1126/science.1192603.

Rattray N.J.W., Deziel N.C., Wallach J.D., Khan S.A., Vasiliou V., and Ioannidis J.P.A. 2018. Beyond genomics: understanding exposotypes through metabolomics. *Human Genomics* 12(1):1-14. DOI: 10.1186/s40246-018-0134-x.

Reck M., Tomasch J., Deng Z., Husemann P. Wagner-Döbler I., and On behalf of COMBACTE consortium. 2015. Stool metatranscriptomics: A technical guideline for mRNA stabilisation and isolation. *BMC Genomics* 16(1):1-18. DOI:

10.1186/s12864-015-1694-y.

Ribeiro R.M., de Souza-Basqueira M., Campos de Oliveira L., Salles F.C., Pereira N.B., and Sabino E.C. 2018. An alternative storage method for characterization of the intestinal microbiota through next generation sequencing. *Revista do Instituto de Medicina Tropical de Sao Paulo* 60(May):1–11. DOI:10.1590/S1678-9946201860077.

Rideout J.R., He Y., Navas-Molina J.A., Walters W.A., Ursell L.K., Gibbons S.M., Chase J., McDonald D., Gonzalez A., Robbins-Pianka A., Clemente J.C., Gilbert J.A., Huse S.M., Zhou H-W., Knight R., and Caporaso J.G. 2014. Subsampled open-reference clustering creates consistent, comprehensive OTU definitions and scales to billions of sequences. *PeerJ* 2(545):1-25. DOI: 10.7717/peerj.545

Sankaranarayanan K., Ozga A.T., Warinner C., Tito R.Y., Obregon-Tito A.J., Xu J., Gaffney P.M., Jervis L.L., Cox D., Stephens L., Foster M., Tallbull G., Spicer P., and Lewis C.M. 2015. Gut Microbiome Diversity among Cheyenne and Arapaho Individuals from Western Oklahoma. *Current Biology* 25(24):3161-3169. DOI: 10.1016/j.cub.2015.10.060

Scheltema R.A., Hauschild J-P., Lange O., Hornburg D., Denisov E., Damoc E., Kuehn A., Makarov A., and Mann M. 2014. The Q Exactive HF, a Benchtop Mass Spectrometer with a Pre-filter, High-performance Quadurpole and an Ultra-high-field Orbitrap Analyzer. *Molecular & Cellular Proteomics* 13(12):3698-3708. DOI: 10.1074/mcp.M114.043489.

Schnabl B., and Brenner D.A. 2014. Interactions Between the Intestinal Microbiome and Liver Diseases. *Gastroenterology* 146:1513-1524. DOI:10.1053/j.gastro.2014.01.020

Schnorr S.L., Candela M., Rampelli S., Centanni M., Consolandi C., Basagli G., Turroni S., Biagi E., Peano C., Severgnini M., Fiori J., Gotti R., De Bellis G., Luiselli D., Brigidi P. Mabulla A., Marlowe F., Henry A.G., and Crittenden A.N. 2014. Gut microbiome of the Hadza hunter-gatherers. *Nature Communications* 5(1):1-12. DOI: 10.1038/ncomms4654.

Schubert M., Lindgreen S., and Orlando L. 2016. AdapterRemoval v2: rapid adapter

trimming, identification, and read merging. *BMC Research Notes* 9(1):88. DOI:10.1186/s13104-016-1900-2.

Sinha R., Chen J., Amir A., Vogtmann E., Shi J., Inman K.S., Flores R., Sampson J., Knight R., and Chia N. 2016. Collecting fecal samples for microbiome analyses in epidemiology studies. *Cancer Epidemiology Biomarkers and Prevention* 25(2):407–416. DOI:10.1158/1055-9965.EPI-15-0951.

Sinha R., Vogtmann E., Chen J., Amir A., Shi J., Sampson J., Flores R., Knight R., and Chia N. 2016. "Fecal microbiome in epidemiologic studies " - Response. *Cancer Epidemiology Biomarkers and Prevention* 25(5):870-871. DOI: 10.1158/1055-9965.EPI-16-0161.

ThermoFisher Scientific, Inc. 2016. Thermo Scientific Q Exactive Orbitrap LC-MS/MS System Product Specifications [Online]. URL: https://www.thermofisher.com/order/catalog/product/IQLAAEGAAPFALGMBDK.

Tito R.Y., Knights D., Metcalf J., Obregon-Tito A.J., Cleeland L., Najar F., Roe B., Reinhard K., Sobolik K., Belknap S., Foster M., Spicer P., Knight R., and Lewis Jr. C.M. 2012. Insights from Characterizing Extinct Human Gut Microbiomes. *PLoS ONE* 7(12):1-8. DOI: 10.1371/journal.pone.0051146.

Tsukuda M., Kitahara K., and Miyazaki K. 2017. Comparative RNA function analysis reveals high functional similarity between distantly related bacterial 16S rRNAs. *Scientific Reports* 7(1):1-8. DOI: 10.1038/s41598-017-10214-3.

Turnbaugh P.J., Ley R.E., Hamady M., Fraser-Liggett C., Knight R., and Gordon J.I. The human microbiome project: exploring the microbial part of ourselves in a changing world. *Nature* 449(7164):804-810. DOI:10.1038/nature06244.

Turroni S., Fiori J., Rampelli S., Schnorr S.L., Consolandi C., Barone M., Biagi E., Fanelli F., Mezzullo M., Crittenden A.N., Henry A.G., Brigidi P., and Candela M. 2016. Fecal metabolome of the Hazda hunter-gatherers: a host-microbiome intregrative view. *Scientific Reports* 6:1-9. DOI: 10.1038/srep32826.

Vandamme P., Pot B., Gillis M., De Vos P., Kersters K., and Swings J. 1996. Polyphasic

Taxonomy, a Consensus Approach to Bacterial Systematics. *Microbiological and Molecular Biological Reviews* 60(2):407-438. DOI: 10.1007/s12088-007-0022-x.

van Eijsden, R.G.E., Stassen C., Daenen L., Van Mulders S.E., Bapat P.M., Siewers V., Goossens K.V.Y., Nielsen, J., Delvaux F.R., Van Hummelen P., Devreese B., and Willaert R.G. 2013. A universal fixation method based on quaternary ammonium salts (RNA*later*) for omics-technologies: *Saccharomyces cerevisiae* as a case study. *Biotechnology Letters* 35(6):891-900. DOI: 10.1007/s10529-013-1163-0.

Velsko I.M., Overmyer K.A., Speller C., Klaus L., Collins M.J., Loe L., Frantz L.A.F., Sankaranarayanan K., Lewis Jr C.M., Bautista Rodriguez Martinez J., Chaves E., Coon J.J., Larson G., and Warinner C. 2017. The dental calculus metabolome in modern and historic samples. *Metabolomics* 13(11):134, 1-17. DOI: 10.1007/s11306-017-1270-3.

Viant M.R., Kurland I.J., Jones M.R., and Dunn W.B. 2017. How close are we to complete annotation of metabolomes? *Current Opinion in Chemical Biology* 36:64-69. DOI: 10.1016/j.cbpa.2017.01.001.

Voigt A.Y., Costea P.I., Kultima J.R., Li S.S., Zeller G., Sunagawa S., and Bork P. 2015. Temporal and technical variability of human gut metagenomes. *Genome Biology* 16(1):73-85. DOI: 10.1186/s13059-015-0639-8.

Vuong H.E., Yano J.M., Fung T.C., and Hsiao E.Y. 2017. The Microbiome and Host Behavior. *Annual Review of Neuroscience* 40(1):21-49. DOI: 10.1146/annurev-neuro-072116-031347.

Wang D., and Bodovitz S. 2010. Single cell analysis: the new frontier in 'omics'. *Trends in Biotechnology* 26(6):281-290. DOI: 10.1016/j.tibtech.2010.03.002.

Wang, M., Carver J.J., Phelan V.V., Sanchez L.M., Garg N., Peng Y., Nguyen D.D., Watrous J., Kapono C.A., Luzzatto-Knaan T., Porto C., Bouslimani A., Melnik A. V., Meehan M. J., Liu W.-T., Crüsemann M., Boudreau P.D., Esquenazi E., Sandoval-Calderon M., Kersten R.D., Pace L.A., Quinn R.A., Duncan K.R., Hsu C.-C., Floros D.J., Gavilan R.G., Kleigrewe K., Northen T., Dutton R.J., Parrot D., Carlson E.E., Aigle B., Michelsen C.F., Jelsbak L., Sohlenkamp C., Pevzner P., Edlund A., McLean J., Piel J., Murphy B.T., Gerwick L., Liaw C.-C., Yang Y.-L.,

Humpf H.-U., Maansson M., Keyzers R.A., Sims A.C., Johnson A.R., Sidebottom A.M., Sedio B.E., Klitgaard A., Larson C.B., Boya P C.A., Torres-Mendoza D., Gonzalez D.J., Silva D.B., Marques L.M., Demarque D.P., Pociute E., O'Neill E.C., Briand E., Helfrich E.J.N., Granatosky E.A., Glukhov E., Ryffel F., Houson H., Mohimani H., Kharbush J.J., Zeng Y., Vorholt J. A., Kurita K. L., Charusanti P., McPhail K. L., Nielsen K. F., Vuong L., Elfeki M., Traxler M.F., Engene N., Koyama N., Vining O.B., Baric R., Silva R.R., Mascuch S.J., Tomasi S., Jenkins S., Macherla V., Hoffman T., Agarwal V., Williams P.G., Dai J., Neupane R, Gurr J., Rodríguez A.M.C., Lamsa A., Zhang C., Dorrestein K., Duggan B.M., Almaliti J., Allard P.-M., Phapale P., Nothias L.-F., Alexandrov T., Litaudon M., Wolfender J.-L., Kyle J.E., Metz T.O., Peryea T., Nguyen D.-T., VanLeer D., Shinn P., Jadhav A., Müller R., Waters K. M., Shi W., Liu X., Zhang L., Knight R., Jensen P.R., Palsson B.Ø., Pogliano K., Linington R.G., Gutierrez M., Lopes N.P., Gerwick W.H., Moore B.S., Dorrestein P.C., and Bandeira N. 2017. Sharing and community curation of mass spectrometry data with GNPS. *Nature Biotechnology* 34(8):828–837. DOI:10.1038/nbt.3597.Sharing.

Wang Z., Zolnik C.P., Qiu Y., Usyk M., Wang T., Strickler H.D., Isasi C.R., Kaplan R.C., Kurland I.J., Qi Q., and Burk R.D. 2018. Comparison of Fecal Collection Methods for Microbiome and Metabolomics Studies. *Frontiers in Cellular and Infection Microbiology* 8:1–10. DOI:10.3389/fcimb.2018.00301.

Warinner C., Matias Rodrigues J.F., Vyas R., Trachsel C., Shved N., Grossmann J., Radini A., Hancock Y., Tito R.Y., Fiddyment S., Speller C., Hendy J., Charlton S., Ulrich Luder H., Salazar-García D.C., Eppler E., Seiler R., Hansen L.H., Alfredo Samaniego Castruita J., Barkow-Oesterreicher S., Yik Teoh K., Kelstrup C.D., Olsen J.V., Nanni P., Kawai T., Willerslev E., von Mering C., Lewis Jr. C.M.,, Collins M.J., Gillbert M.T.P., Rühli F., and Cappellini E. 2014. Pathogens and host immunity in the ancient human oral cavity. *Nature Genetics* 46(4):336-344. DOI:10.1038/ng.2906.

Wickham H. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer US.DOI:10.18637/jss.v077.b02.

Willett W.C. 2002. Balancing Life-Style and Genomics Research for Disease Prevention. *Science* 296(5568):695-698.

Wishart D.S. 2016. Emerging applications of metabolomics in drug discovery and precision medicine. *Nature Reviews: Drug Discovery* 15(7):473-484. DOI:

10.1038/nrd.2016.32

Wishart D.S., Feunang Y.D., Marcu A., Guo A.C., Liang K., Vázquez-Fresno R., Sajed T., Johnson D., Li C., Karu N., Sayeeda Z., Lo E., Assempour N., Berjanskii M., Singhal S., Arndt D., Liang Y., Badran H., Grant J., Serra-Cayuela A., Liu Y., Mandal R., Neveu V., Pon A., Knox C., Wilson M., Manach C., and Scalbert A. 2018. HMDB 4.0: The human metabolome database for 2018. *Nucleic Acids Research* 46(D1):D608–D617. DOI:10.1093/nar/gkx1089.

Woese, Carl R. 1987. Bacterial Evolution. *Microbiological Reviews* 52(2):221-271. PMID:2439888.

Wolfender J-L., Marti G., Thomas A., and Bertrand S. 2015. Current approaches and challenges for the metabolite profiling of complex natural extracts. *Journal of Chromatography A* 1382:136-164. DOI: 10.1016/j.chroma.2014.10.091.

Yildrim S., Yeoman C.J., Sipos M., Torralba M., Wilson B.A., Goldberg T.L., Stumpf R.M., Leigh S.R., White B.A., and Nelson K.E. 2010. Characterization of the Fecal Microbiome from Non-Human Wild Primates Reveals Species Specific Microbial Communities. *PLoS ONE* 5(11):1-11. DOI: 10.1371/journal.pone.0013963.

Zamboni N., Saghatelian A., and Patti G.J. 2015. Defining the Metabolome: Size, Flux, and Regulation. *Molecular Cell* 58(4):699-706. DOI: 10.1016/j.molcel.2015.04.021.

Zoetendal E.G., Booijink C.C.G.M., Klassens E.S., Heilig H.G.H.J., Kleerebezem M., Smidt H., and de Vos W.M. 2006. Isolation of RNA from bacterial samples of the human gastrointestinal tract. *Nature Protocols* 1(2):954-959. DOI: 10.1038/nprot.2006.143.

Appendix I: Supplementary Tables

Qubit Quantification Values

Sample Name	DNA Concentration (ng/μL)	RNA Concentration (ng/μL)
0702RTN	1.93	3.26
0901RTN	3.19	11.6
1001RTN	1.93	6.63
1002RTN	2.94	12.8
1102RTN	7.39	13.6
1301RTN	1.71	2.99
2001RTN	9.67	42.3
2302RTN	12.2	24.2
2401RTN	5.18	19.9
3002RTN	5.26	7.81
0702RTR	3.3	14.5
0901RTR	4.09	11.4
1001RTR	4.14	8.42
1002RTR	5.88	14
1102RTR	13.8	13.8
1301RTR	5.25	3.35
2001RTR	10.9	19.9
2302RTR	5.31	6.61
2401RTR	6.2	5.79
3002RTR	18.1	TOO_LOW
07024N	1.38	TOO_LOW
09014N	5.19	14
10014N	2.25	7.89
10024N	2.3	6.47
11024N	11.6	14.6
13014N	1.67	6.38
20014N	9.78	41.9
23024N	10.8	21.8
24014N	5.33	18.9
30024N	8.23	16.1
07024R	3.2	7.21
09014R	5.58	21.9
10014R	7.49	23.2
10024R	4.58	11.4
11024R	13.1	15.4
13014R	3.41	6.5

20014R	12.3	47.3
23024R	20.9	2.93
24014R	12	25.9
30024R	18.5	27
070280N	2.57	5.49
090180N	8.75	32.1
100180N	6.83	22.8
100280N	4.67	17.2
110280N	14.4	21.2
130180N	2.54	16.2
200180N	12.6	63
230280N	13.9	36.6
240180N	4.27	26.8
300280N	6.9	16.3
070280R	2.97	11.5
090180R	8.73	36.4
100180R	5.55	27.2
100280R	4.62	10.6
110280R	13.5	18.2
130180R	4.31	14.3
200180R	16.4	TOO_HIGH
230280R	9.3	28.2
240180R	7.43	30.3
300280R	8.96	30.3

Table S1. Qubit Values.

This table lists values from the Qubit Quantification of all 60 samples. Quantification was done immediately after DNA extraction. Qubit kits had limited detection ranges, so any values that were outside those ranges were designated “TOO_LOW” or “TOO_HIGH”. Samples kept at -80°C tended to have higher concentrations. Moreover, samples treated with RNA*later* generally had higher concentration values than their non-RNA*later* counterparts.

qPCR Reaction Sheet

	35 cycles			
	initial	Denature	Anneal	Elongation
Temp	95 C	95 C	52 C	72 C
Time	600 sec	10 sec	20 sec	30 sec

MasterMix	ul per reaction	# rxn	ul need
H20	5.8	77	446.6
FastStart Essential Green MM	10		770
V4 F (non-barcoded)	0.6		46.2
V4 R (non-barcoded)	0.6		46.2
Total	17		1309

Reaction	17 ul MM
	3 ul sample
	20 ul total

Table S2. Reaction Sheet used for qPCR reactions.

Reaction sheet contains reagents used for master mix, how much of each reagent was used, cycling and amplification parameters, and total number of cycles. Number of reactions (rxns) refers to number of samples (n) at this step (n + 10%). 77 rxns were chosen to ensure there was enough MasterMix.

Sample to Barcode Matches

A

Sample Name	Plate	Plate Position	Golay Barcode	Reverse Complement	Primer#
23_02_RT_N	4	A3	AGACATACCGTA	TACGGTATGTCT	806rcbc290
24_01_RT_N	4	B3	TGTATCTTCACC	GGTGAAGATACA	806rcbc302
23_02_-80_N	4	C3	AGGCACAGTAGG	CCTACTGTGCCT	806rcbc314
23_02_4_R	4	D3	TGTTAAGCAGCA	TGCTGCTTAACA	806rcbc326
09_01_-80_R	4	E3	AAGGGCGCTGAA	TTCAGCGCCCTT	806rcbc338
09_01_-80_N	4	F3	CTCTGCCTAATT	AATTAGGCAGAG	806rcbc350
20_01_-80_N	4	G3	GCATTACTGGAC	GTCCAGTAATGC	806rcbc362
23_02_4_N	4	H3	GAGTACAGTCTA	TAGACTGTACTC	806rcbc374
20_01_-80_R	4	A4	GATCCTCATGCG	CGCATGAGGATC	806rcbc291
20_01_RT_N	4	B4	GACTGACTCGTC	GACGAGTCAGTC	806rcbc303
30_02_4_R	4	C4	CTACTTACATCC	GGATGTAAGTAG	806rcbc315
23_02_-80_R	4	D4	ACGGCGTTATGT	ACATAACGCCGT	806rcbc327
10_02_RT_N	4	E4	GTTTCCGTGGTG	CACCACGGAAAC	806rcbc339
24_01_4_R	4	F4	ATATGACCCAGC	GCTGGGTCATAT	806rcbc351
30_02_4_N	4	G4	TTGGGCCACATA	TATGTGGCCCAA	806rcbc363
20_01_4_N	4	H4	CCTACATGAGAC	GTCTCATGTAGG	806rcbc375
11_02_RT_N	4	A5	ATTATCGTCCCT	AGGGACGATAAT	806rcbc292

20_01_4_R	4	B5	TCGTGGATAGCT	AGCTATCCACGA	806rcbc304
09_01_4_R	4	C5	CTCTTCTGATCA	TGATCAGAAGAG	806rcbc316
30_02_-80_N	4	D5	ACTTTGCTTTGC	GCAAAGCAAAGT	806rcbc328
24_01_-80_R	4	E5	AGGAACCAGACG	CGTCTGGTTCCT	806rcbc340
09_01_RT_N	4	F5	CTCTATTCCACC	GGTGGAATAGAG	806rcbc352
30_02_-80_R	4	G5	CACACAAAGTCA	TGACTTTGTGTG	806rcbc364
11_02_-80_N	4	H5	TCCGTGGTATAG	CTATACCACGGA	806rcbc376

09_02_4_N	4	A6	CCAGACCGCTAT	ATAGCGGTCTGG	806rcbc293
30_02_RT_N	4	B6	GACGCACTAACT	AGTTAGTGCGTC	806rcbc305
11_02_-80_R	4	C6	ATGCTAACCACG	CGTGGTTAGCAT	806rcbc317
10_01_-80_N	4	D6	CAAAGCGGTATT	AATACCGCTTTG	806rcbc329
10_02_-80_N	4	E6	TAATGCCCAGGT	ACCTGGGCATTA	806rcbc341
11_02_4_R	4	F6	ATTGAGTGAGTC	GACTCACTCAAT	806rcbc353
24_01_4_N	4	G6	GCCAAGGATAGG	CCTATCCTTGGC	806rcbc365
30_02_RT_R	4	H6	TCTACGGCACGT	ACGTGCCGTAGA	806rcbc377

11_02_4_N	4	A7	AGCTCTAGAAAC	GTTTCTAGAGCT	806rcbc294
11_02_RT_R	4	B7	GGCGATTTACGT	ACGTAAATCGCC	806rcbc306
10_02_4_R	4	C7	ACCAATCTCGGC	GCCGAGATTGGT	806rcbc318
10_01_-80_R	4	D7	CGAAACTACGTA	TACGTAGTTTCG	806rcbc330
07_02_RT_N	4	E7	TATGAACGTCCG	CGGACGTTCATA	806rcbc342

24_01_-80_N	4	F7	TTATGGTACGGA	TCCGTACCATAA	806rcbc354
10_02_-80_R	4	G7	CGCCACGTGTAT	ATACACGTGGCG	806rcbc366
10_01_4_R	4	H7	ATGCTGCAACAC	GTGTTGCAGCAT	806rcbc378

09_01_RT_R	4	A8	TCCATCGACGTG	CACGTCGATGGA	806rcbc295
10_02_RT_R	4	B8	TAAGGCATCGCT	AGCGATGCCTTA	806rcbc307
20_01_RT_R	4	C8	TATCCAAGCGCA	TGCGCTTGGATA	806rcbc319
10_02_4_N	4	D8	GAGGACCAGCAA	TTGCTGGTCCTC	806rcbc331
23_02_RT_R	4	E8	CCACATTGGGTC	GACCCAATGTGG	806rcbc343
13_01_-80_R	4	F8	GCTAGTTATGGA	TCCATAACTAGC	806rcbc355
07_02_-80_R	4	G8	GCAACCGATTGT	ACAATCGGTTGC	806rcbc367
24_01_RT_R	4	H8	TTCTCATGGAGG	CCTCCATGAGAA	806rcbc379

07_02_-80_N	4	A9	CGATGTGTGGTT	AACCACACATCG	806rcbc296
10_01_RT_N	4	B9	ACCCATACAGCC	GGCTGTATGGGT	806rcbc308
13_01_-80_N	4	C9	GTACTGAAGATC	GATCTTCAGTAC	806rcbc320
10_01_4_N	4	D9	AATAGCATGTCG	CGACATGCTATT	806rcbc332
07_02_4_R	4	E9	TCAGTCAGATGA	TCATCTGACTGA	806rcbc344
13_01_4_R	4	F9	CAGATTAACCAG	CTGGTTAATCTG	806rcbc356
PCRBLK1	4	A1	TCTGAGGTTGCC	GGCAACCTCAGA	806rcbc288
ECOLI10X	4	B1	TCCAACTGCAGA	TCTGCAGTTGGA	806rcbc300
ECOLI10X_2	4	B2	TAAAGACCCGTA	TACGGGTCTTTA	806rcbc301

B

Sample Name	Plate	Plate Position	Golay Barcode	Reverse Complement	Primer#
10_01_RT_R	4	A10	GCGAAGTTGGGA	TCCCAACTTCGC	806rcbc297
07_02_RT_R	4	B10	CGCACTACGCAT	ATGCGTAGTGCG	806rcbc309
13_01_RT_R	4	C10	TCGCCGTGTACA	TGTACACGGCGA	806rcbc321
07_02_4_N	4	D10	CGGAGTAATCCT	AGGATTACTCCG	806rcbc333
13_01_4_N	4	E10	AAGTCACACACA	TGTGTGTGACTT	806rcbc345
13_01_RT_N	4	F10	GGCTGCATACTC	GAGTATGCAGCC	806rcbc357
EB2	4	G10	GTTCCTCCATTA	TAATGGAGGAAC	806rcbc369
EB1	4	H10	GCTATCAAGACA	TGTCTTGATAGC	806rcbc381
EB6	4	A11	GCATTCGGCGTT	AACGCCGAATGC	806rcbc298
EB4	4	B11	CAGTCGTTAAGA	TCTTAACGACTG	806rcbc310
EB3	4	C11	AACTGCGATATG	CATATCGCAGTT	806rcbc322
EB5	4	D11	CTGTGTCCATGG	CCATGGACACAG	806rcbc334
PCRBLK2	4	A2	GATCATTCTCTC	GAGAGAATGATC	806rcbc289
ECOLI10X	4	B1	TCCAACTGCAGA	TCTGCAGTTGGA	806rcbc300
ECOLI10X_2	4	B2	TAAAGACCCGTA	TACGGGTCTTTA	806rcbc301

Table S3. Sample to Barcode Pairings.

These tables contain the samples used and their corresponding reverse barcodes. PCR replicates were done using these unique sample-to-barcode matches. The *E. coli* standards appear twice because the B1 barcode ran low for the last replicates. B2 was used as a result. Following qPCR data analysis, samples were split into two groups based on their CQ values. Groups were amplified by a different number of cycles.

(A) PCR Group 1. All were amplified to 18 cycles.

(B) PCR Group 2. All were amplified to 20 cycles.

PCR Reaction Sheets

A

			Amplify for **18** cycles		
	initial	Denature	Anneal	Elongation	Final
Temp	98 C	98 C	52 C	72 C	72 C
Time	30 s	15 sec	20 sec	30 sec	300 sec

MasterMix	ul per reaction	# rxn	ul need
H20	6.6	61	402.6
Phusion HF Buffer	4		244
Illumina V4 F primer 10uM	1		61
10mM dNTPs	0.4		24.4
Phusion HS II enzyme	0.2		12.2
BSA 2.5 mg/ml	0.8		48.8
Total	13		793

Reaction	13 ul MM
	3 ul sample
	4 ul V4 reverse 2.5 uM
	20 ul total

B

			Amplify for **20** cycles		
	initial	Denature	Anneal	Elongation	Final
Temp	98 C	98 C	52 C	72 C	72 C
Time	30 s	15 sec	20 sec	30 sec	300 sec

MasterMix	ul per reaction	# rxn	ul need
H20	6.6	20	132
Phusion HF Buffer	4		80
Illumina V4 F primer 10uM	1		20
10mM dNTPs	0.4		8
Phusion HS II enzyme	0.2		4
BSA 2.5 mg/ml	0.8		16
Total	13		260

Reaction	13 ul MM
	3 ul sample
	4 ul V4 reverse 2.5 uM
	20 ul total

Table S4. **Reaction Sheets used for PCR.**

These sheets detail the reagents compromising the master mix, how much of each reagent was used, PCR conditions, and number of amplification cycles. Number of rxns still refers to number of samples plus 10%.

(A) PCR Group 1.

(B) PCR Group 2.

MZMine Data Processing Parameters

Mass Detection	MS1 Noise Level	8.00E+04
	MS2 Noise Level	5.00E+03
	Mass Detector	Centroid
Chromatogram Builder	Minimum Time Span (min)	0.05
	Minimum Height	2.40E+05
	m/z tolerance (ppm)	10
Chromatogram Deconvolution	Min peak height	2.40E+05
	Peak duration range (min)	0-2.00
	Baseline level	8.00E+04
	m/z Range (Da)	0.01
	RT range (min)	0.1
Isotope Peaks Grouper	RT tolerance (min)	0.1
	m/z tolerance (ppm)	10
	Monotonic shape	Yes
	Max charge	3
	Representative isotope	Lowest m/z
Alignment	m/z tolerance (ppm)	10
	m/z to RT weight	5 to 1
	RT tolerance (min)	0.1
	Require same charge state	Yes
Row Filtering	RT range (min)	0.2-12
	Keep only peaks with MS2 scar	Yes
	Minimum peaks per row	2
	Minimum peaks per isotope	2
Gap Filling	Intensity tolerance (%)	10
	m/z tolerance (ppm)	10
	RT tolerance (min)	0.2

Table S5. Parameters for MZMine Data Analysis.

Column 1 refers to the different steps of data processing in MZMine. Each step had different parameters and values, as shown here. Columns 2 and 3 correspond to the parameter and the input value, respectively. Retention time is referred to as RT.

Statistical Results

A

Analysis	Evaluated Categories	Test Name	Test statistic	p-value	Statistically Significant
Alpha Diversity	Phylogenetic Diversity & RNA*later* Use	Kruskal-Wallis	chi-squared = 6.9255	0.008498	Yes
	Observed OTUs & RNA*later* Use	Kruskal-Wallis	chi-squared = 6.5077	0.01074	Yes
	Phylogenetic Diversity & Individual	Kruskal-Wallis	chi-squared = 42.638	2.51E-06	Yes
	Observed OTUs & Individual	Kruskal-Wallis	chi-squared = 43.948	1.44E-06	Yes
	Phylogenetic Diversity & Storage Temperature	Kruskal-Wallis	chi-squared = 1.2751	0.5286	No
	Observed OTUs & Storage Temperature	Kruskal-Wallis	chi-squared = 0.95963	0.6189	No

B

Analysis	Evaluated Categories	Test Name	Test statistic	p-value	Statistically Significant
Alpha Diversity	Phylogenetic Diversity & RNA*later* Use	Dunn's test	chi-squared = 6.9255	0.01	Yes
	Observed OTUs & RNAlater Use	Dunn's test	chi-squared = 6.5077	0.01	Yes
	Phylogenetic Diversity & Individual	Dunn's test	chi-squared = 42.638	0	Yes
	Observed OTUs & RNAlater Use	Dunn's test	chi-squared = 43.9485	0	Yes
	Phylogenetic Diversity & Storage Temperature	Dunn's test	chi-squared = 1.2751	0.53	No
	Observed OTUs & Storage Temperature	Dunn's test	chi-squared = 0.9596	0.62	No

C

Analysis	Evaluated Categories	Test Name	Test statistic	p-value	Statistically Significant
Alpha Diversity	Phylogenetic Diversity: Individual & RNA*later* Use	ANCOVA	Indiv F-value = 37.746	Indiv = 2e-16	Yes
			RNA*later* F-value = 46.140	RNA*later* = 3.63e-8	Yes
			Indiv & RNA*later* F-value = 1.554	Indiv & RNA*later* = 0.163	No
	Phylogenetic Diversity: Individual & Storage Temperature	ANCOVA	Indiv F-value = 15.657	Indiv = 4.5e-09	Yes
			StorageTemp F-value = 2.123	StorageTemp = 0.137	No
			Indiv & StorageTemp F-value = 0.405	Indiv & StorageTemp = 0.976	No
	Phylogenetic Diversity: RNA*later* Use & Storage Temperature	ANCOVA	RNA*later* F-value = 6.576	RNA*later* = 0.0132	Yes
			StorageTemp F-value = 0.729	StorageTemp = 0.4869	No
			RNA*later* & StorageTemp F-value = 0.314	RNA*later* & StorageTemp = 0.7317	No
	Observed OTUs: Individual & RNA*later* Use	ANCOVA	Indiv F-value = 37.858	Indiv = 2e-16	Yes
			RNA*later* F-value = 43.308	RNA*later* = 7.22e-8	Yes
			Indiv & RNA*later* F-value = 1.082	Individual & RNA*later* = 0.397	No
	Observed OTUs: Individual & Storage Temperature	ANCOVA	Indiv F-value = 16.209	Indiv = 2.98e-9	Yes
			StorageTemp F-value = 2.206	StorageTemp = 0.128	No
			Indiv & StorageTemp F-value = 0.301	Indiv & StorageTemp = 0.995	No
	Observed OTUs: RNA*later* Use &	ANCOVA	RNA*later* F-value = 6.295	RNA*later* = 0.0151	Yes

	Storage Temperature		StorageTemp F-value = 0.749	StorageTemp = 0.4777	No
			RNA*later* & StorageTemp F-value = 0.627	RNA*later* & Storage Temp = 0.5379	No

D

Analysis	Evaluated Categories	Test Name	Test statistic	p-value	Statistically Significant
Beta Diversity	Unweighted UniFrac Distances of Individual	PERMANOVA	t-value = 12.492386530836116	0.001	Yes
	Unweighted UniFrac Distances of RNA*later* Use	PERMANOVA	t-value = 3.609480950510251	0.002	Yes
	Unweighted UniFrac Distances of Storage Temperature	PERMANOVA	t-value = 2.36318457297859	0.001	Yes
	Weighted UniFrac Distances of Individual	PERMANOVA	t-value = 21.446488362798743	0.001	Yes
	Weighted UniFrac Distances of RNA*later* Use	PERMANOVA	t-value = 8.656130892382821	0.001	Yes
	Weighted UniFrac Distances of Storage Temperature	PERMANOVA	t-value = 5.97720542549215	0.001	Yes

Table S6. Values from various statistical tests. These tables are separated by the test performed. First column refers to the type of analysis the test acted on. Second column is the category/variable considered in these tests. Each statistical test has different considerations so these varied, but primarily focused on the effects of Individual (host), RNA*later* use, and Storage Temperature. Third column is the name of the test. The fourth column contains the test statistic values from these tests. These statistics will be different between tests. The fifth column contains the p-value from each test. The sixth and last column says whether the results were statistically significant. For this project, statistical significance was defined as p-value <0.05.

(A) Results from the Kruskal-Wallis test by ranks, also known as a one-way ANOVA. This nonparametric test evaluates whether significant differences exist in independent samples. This test was run on results

from alpha diversity and found that host and RNA*later* use caused significant differences for both phylogenetic diversity and microbial richness. Storage temperature did not have significant effects.

(**B**) Results from Dunn's test of multiple comparisons. Typically done on a significant result from Kruskal-Wallis to correct for errors in the Kruskal-Wallis test. These results validated the findings of our Kruskal-Wallis tests: host and RNA*later* use had significant effects and storage temperature did not.

(**C**) Results from ANCOVA tests, also known as analysis of covariance. ANCOVA evaluates whether an independent variable has significant effects on a dependent variable while considering the effects of a different independent variable (called the covariate). The test provides results for all variables, so the values for each specific variable are listed. These ANCOVA results indicate if each independent variable had significant effects on the dependent variable, and if these independent variables affected each other. Our results indicate individual and RNA*later* caused significant effects, validating the results from Kruskal-Wallis and Dunn's tests. However, these variables did not have significant effects on each other. Storage temperature did not have significant effects on samples or on other variables. Our conclusions from ANCOVA match those from earlier tests, but further confirm that host and RNA*later* caused significant effects and did not affect each other.

(**D**) PERMANOVA results. Known as Permutational Analysis of Variance, this non-parametric test evaluates significant differences between groups while considering multiple variables. PERMANOVA also considers permutations to ensure accurate results. Our PERMANOVA results indicate that host, RNA*later* use, and storage temperature all had significant effects between samples. The significance of storage temperature contrasts with results from alpha diversity and with PCoA plots, but storage temperature still caused differences between samples. These temperature effects were largely outweighed by the influence of the host. We can conclude that storage temperature and *RNAlater* use will still affect the microbiome profile of samples, but these effects are minimal compared to the host.

Identified Phyla and Genera

A

Phylum	Average Abundance	Number of Samples with >0% Abundance	Number of Samples with >2% Abundance
Firmicutes	0.766368056	60	60
Actinobacteria	0.146322222	60	60
Bacteroidetes	0.032993056	60	24
Proteobacteria	0.0249375	60	14
Euryarchaeota	0.012948611	38	14
Tenericutes	0.009440278	60	10
Cyanobacteria	0.004248611	58	0
Verrucomicrobia	0.002077778	15	4
Spirochaetes	0.000518056	22	0
Streptophyta	8.19444E-05	14	0
Elusimicrobia	3.61111E-05	7	0
Fusobacteria	1.66667E-05	7	0
Lentisphaerae	5.55556E-06	2	0
Synergistetes	2.77778E-06	2	0
Chloroflexi	1.38889E-06	1	0
Planctomycetes	1.38889E-06	1	0

B

Genus	Average Frequency	Number of Samples with >0% Abundance	Number of Samples with >2% Abundance
Blautia	0.10365	60	54
Clostridium	0.083161111	60	47
Collinsella	0.063786111	60	47
Subdoligranulum	0.0414125	60	40
Pediococcus	0.035995833	47	10
Romboutsia	0.035770833	60	37
Holdemanella	0.032661111	57	27
Catenibacterium	0.031088889	52	19
Prevotella	0.030095833	59	21
Bifidobacterium	0.028334722	45	12
Streptococcus	0.026070833	60	20

Lachnospiraceae Eubacterium_g5 (Unknown genus)	0.025148611	60	31
Lactobacillus	0.024423611	60	17
Ruminococcaceae Ruminococcus_g2 (Unknown genus)	0.023888889	60	22
Enterococcus	0.020926389	57	13
Turicibacter	0.01985	60	17
Terrisporobacter	0.0192125	60	17
Mogibacterium	0.014624294	49	11
Escherichia	0.015484722	60	9
Olsenella	0.015338889	54	16
Ruminococcaceae JN713389_g (Unknown genus)	0.015231944	60	11
Intestinibacter	0.015101389	60	17
Faecalibacterium	0.014279167	60	16
Dorea	0.014272222	60	18
Agathobacter	0.012163889	58	9
Bulleidia	0.011606944	56	11
Weissella	0.011316667	52	4
Coriobacteriaceae JN162689_g (Unknown genus)	0.011129167	52	7
Sporobacter	0.010763889	60	5
Lachnospiraceae Ruminococcus_g4 (Unknown genus)	0.010102778	59	11

Table S7. Most Abundant Phyla and Genera.

These data come from the rarefied output files, resulting in 12,000 reads per sample. There were 60 total samples at this point. Positive controls were excluded here. Table is sorted in order of descending abundance. Each row corresponds with a different phyla or genera, as indicated by column one. The second column refers to the average frequency of the specific phyla/genera across all samples. This indicates the percentage of the 12,000 reads per sample that matched to the specific phylum/genus. Columns three and four represent the total number of samples containing the specific phylum/genus with at least 0% and 2% relative abundance, respectively.

(**A**) These are the detected phyla across all samples. A total of 16 phyla were identified. Firmicutes dominates the phyla here, as expected in fecal microbiome profiles. Actinobacteria, Bacteroidetes, Proteobacteria, and Euryarcheota follow Firmicutes, but with significantly less abundance. Eight of these 16 phyla do not contain more than 2% abundance in a single sample, suggesting the distribution of phyla was primarily limited to a handful of phyla.

(**B**) The top 30 abundant genera across all samples. A total of 336 genera were identified, but 30 are presented here. *Blautia* was the most abundant genus, as expected in fecal microbiome profiles. *Clostridium, Collinsella, Subdoligranulum,* and *Pediococcus* were the next most abundant genera. Column four indicates that the number of samples with at least 2% abundance of the specific genus varied more so than phyla. This suggests the overall distribution of genera within samples was highly diverse.

SampleID	Individual	RNAlater	StorageTemp	Cleanups#	Urobilinogen
10_02_RT_1removal	10_02	Y	RT (22-25°C)	1	Yes
13_01_4_1removal	13_01	Y	4°C	1	Yes
13_01_80_1removal	13_01	Y	-80°C	1	No
10_02_80_1removal	10_02	Y	-80°C	1	No
11_02_80_2removal	11_02	Y	-80°C	2	No
10_01_80_2removal	10_01	Y	-80°C	2	Yes
23_02_4_1removal	23_02	Y	4°C	1	Yes
07_02_80_2removal	07_02	Y	-80°C	2	No
20_01_80_2removal	20_01	Y	-80°C	2	Yes
20_01_RT_1removal	20_01	Y	RT (22-25°C)	1	Yes
23_02_80_1removal	23_02	Y	-80°C	1	Yes
blank_1	blank	N	NA	0	No

Table S8. Sample information for metabolomic analysis.

These are the twelve samples that were analyzed on the MS with their corresponding metadata information. First two numbers refer to family and individual (##_##). After the name, storage temperature is RT (room temperature), 4 (4°C), and 80 (-80°C). The final part of the name refers to the number of RNA*later* cleanup protocols the samples went through: 1removal (1 cleanup) and 2removal (2 cleanups). Metabolomic analyses were limited to the individual/host, RNA*later* use, storage temperature, and number of RNA*later* removal protocols performed. The final column lists the samples where GNPS detected urobilinogen.

Sample	Individual	StorageTemp	RNAlater	Reads	Phylogenetic diversity	chao1	Observed OTUs
07024N	702	4	N	68,196.00	30.91952	479.0285714	317
07024R	702	4	Y	86,829.00	30.58945	390.6842105	314
070280N	702	80	N	86,566.00	29.19883	378.0227273	293
070280R	702	80	Y	58,686.00	27.51803	373.7368421	282
0702RTN	702	RT	N	71,189.00	24.80942	363.6756757	243
0702RTR	702	RT	Y	59,957.00	28.73366	422.2777778	306
09014R	901	4	Y	146,735.00	27.86527	339.42	273
090180N	901	80	N	216,028.00	27.04321	368.5294118	266
090180R	901	80	Y	212,199.00	27.5375	371.1395349	278
0901RTN	901	RT	N	137,762.00	24.86659	352.9166667	249
0901RTR	901	RT	Y	107,769.00	26.98019	389.375	267
09014N	901	4	N	148,265.00	27.65255	357.2272727	278
10014N	1001	4	N	78,073.00	22.21157	329.4545455	210
10014R	1001	4	Y	86,884.00	24.51685	336.3333333	231
100180N	1001	80	N	137,098.00	23.40074	281.9655172	208
100180R	1001	80	Y	76,928.00	26.76917	366.3636364	253
1001RTN	1001	RT	N	12,758.00	21.48344	256.0967742	189
1001RTR	1001	RT	Y	59,866.00	25.52583	303.5172414	234
10024N	1002	4	N	80,380.00	21.80402	317.037037	225
10024R	1002	4	Y	71,311.00	24.57321	345.7241379	234
100280N	1002	80	N	99,103.00	20.80152	300.0384615	215
100280R	1002	80	Y	78,486.00	23.94963	309.8780488	244
1002RTN	1002	RT	N	177,581.00	17.96618	280.6363636	174
1002RTR	1002	RT	Y	106,493.00	21.71453	295.0294118	224
11024N	1102	4	N	102,164.00	26.66412	331.8484848	250
11024R	1102	4	Y	133,329.00	28.02333	404.5333333	274
110280N	1102	80	N	158,068.00	27.23012	375.25	274
110280R	1102	80	Y	150,935.00	30.65894	391.0666667	296
1102RTN	1102	RT	N	129,626.00	24.60702	328.8857143	234
1102RTR	1102	RT	Y	101,531.00	28.63765	358.3488372	283
13014N	1301	4	N	45,231.00	31.599	381.3333333	302
13014R	1301	4	Y	148,170.00	31.85311	465.097561	308
130180N	1301	80	N	71,614.00	31.65805	452	296
130180R	1301	80	Y	57,857.00	32.18994	531	327
1301RTN	1301	RT	N	37,263.00	33.5229	420.1666667	321
1301RTR	1301	RT	Y	54,825.00	33.21517	447	327
20014N	2001	4	N	150,426.00	26.08141	366	234
20014R	2001	4	Y	114,140.00	28.1145	439.15625	275
200180N	2001	80	N	178,278.00	23.34557	278.1538462	218
200180R	2001	80	Y	274,649.00	28.79424	388.6756757	278
2001RTN	2001	RT	N	140,656.00	23.89523	299.125	217
2001RTR	2001	RT	Y	83,659.00	28.0913	392.1764706	277
23024N	2302	4	N	182,171.00	33.95504	538.4545455	373
23024R	2302	4	Y	197,924.00	37.44313	558.5	428
230280N	2302	80	N	173,645.00	34.60156	597.4285714	390
230280R	2302	80	Y	112,996.00	36.65942	520.8	413

2302RTN	2302	RT	N	173,188.00	29.73521	410	311
2302RTR	2302	RT	Y	97,993.00	36.2905	516.2	407
24014N	2401	4	N	114,796.00	25.08086	365.1538462	267
24014R	2401	4	Y	135,648.00	30.33615	450.4642857	291
240180N	2401	80	N	80,990.00	27.3397	438.0285714	285
240180R	2401	80	Y	145,291.00	29.22783	429.0277778	305
2401RTN	2401	RT	N	168,097.00	23.10736	353.6666667	222
2401RTR	2401	RT	Y	76,867.00	29.06668	450.5	305
30024N	3002	4	N	159,652.00	25.79397	406.6	279
30024R	3002	4	Y	177,542.00	28.8064	418.125	318
300280N	3002	80	N	138,921.00	26.83494	380.2439024	289
300280R	3002	80	Y	150,606.00	28.27993	398.9354839	297
3002RTN	3002	RT	N	217,085.00	26.18505	331.4285714	250
3002RTR	3002	RT	Y	111,187.00	29.15452	396.1621622	315

Table S9. Sample information for microbiome analysis and results.

All 60 samples that underwent 16S rRNA gene sequencing are listed. Blanks and positive controls were not included in this chart. Sample naming system continues as family and individual, storage temperature, and RNA*later*. Column 5 contains the number of 16S reads mapped to each sample. Columns 6,7, and 8 correspond to phylogenetic diversity, chao1, and number of observed OTUs, respectively. QIIME1 was used to generate these values.

Lane	Read	Cycles	Yield	Projected Yield	Aligned (%)	Error rate (%)	%>=Q30	Cluster PF(%)	Reads PF	Total Reads
Lane 1	Read 1	251	2.93Gbp	2.93Gbp	25.32	1.31	95.12	92.99	11,704,428	12,588,365
	Read 2	12	128.75Mbp	128.75Mbp	0	0	69.71			
	Read 3	251	2.93Gbp	2.93Gbp	25.69	1.27	92.16			
	Non-Index Reads Total	502	5.85Gbp	5.85Gbp	25.5	1.29	93.64			
	Totals	514	5.98Gbp	5.98Gbp	25.5	1.29	93.13			

Table S10. MiSeq Run Summary and Metrics.

Results table was acquired from Illumina BaseSpace SequenceHub for the MiSeq run performed for this thesis. 1 lane of the MiSeq flowcell was used for 3 reads. Reads 1 and 3 ran for 251 cycles each, while read 2 ran for 12 cycles. This was because a 2x250 paired-end run was performed. 5.98 giga base pairs (Gbp, equivalent to 1,000,000 base pairs) were acquired in the run. Of all the total reads, 25.5% aligned to the PhiX positive control to ensure the MiSeq run performs as it should. 12,588,365 total reads were detected by the MiSeq and 11,704,428 of these reads passed the filtering criteria. This equals 92.99% of the total reads passing filter and mapping as 16S reads. 93.13% of the total reads were Q30, meaning there was a 1:1,000 chance that a base was incorrectly identified.