The effect of sample processing methodology on observed metagenomic and metatranscriptomic microbiome profiles from healthy human stool.

by

Molly Pratt

A Thesis submitted to the Faculty of Graduate Studies of
The University of Manitoba
in partial fulfillment of the requirements of the degree of

Master of Science

Department of Medical Microbiology and Infectious Diseases
University of Manitoba
Winnipeg
Treaty No. 1 Territory

Abstract

Disease-associated changes to the gastrointestinal (GI) microbiome have been detected in a variety of chronic human illnesses. Currently, GI microorganisms are thought to play a major role in systemic health and homeostasis through interactions with the immune system of their host. In recent years, the available technology for capturing and characterizing the GI microbiome has expanded greatly, resulting in a rapid expansion of the field. In particular, culture-independent technologies allowing for direct analysis of microbial genes, transcripts, proteins, and other metabolites (collectively referred to as meta-omics) from stool samples show potential for personalized disease detection and monitoring through non-invasive screening. However, due to the high degree of variability inherent in microbiome profiles, establishing a consensus for microbial signatures or biomarkers of disease across studies is difficult. Differences in study methodology can further complicate comparisons, since the laboratory protocols for sequence-based data capture and analysis are varied, and there are several commercial kits and reagents available for the storage and isolation of microbial DNA and RNA from stool for downstream microbiome profiling. Research has shown that the choice of nucleic acid extraction kit and storage method can affect resulting nucleic acid quality. In the current methodological study, a single stool sample from a healthy donor is divided and processed using multiple commercially available methods for nucleic acid stabilization and isolation in order to evaluate the effect of processing on observed sequence-based microbiome profiles. The research herein demonstrates that the experimental methodology used to stabilize and isolate nucleic acids from human stool samples can significantly impact the ability to capture GI microbiome diversity from metagenomic and metatranscriptomic data. Notably, GI microbiome characteristics commonly used as health markers in disease research are differentially impacted by the specific combination of preservative reagent and nucleic acid extraction kit. This study additionally describes important considerations for future meta-omic microbiome research, including the choice of bacterial lysis approach, bioinformatic analysis methodology, and potentially detrimental vendor mismatching. Ultimately, the current research supports the use of informed experimental design and expands current understanding of how biological sample integrity and experimental bias may affect observed meta-omic microbiome profiles.

Acknowledgements

Firstly I would like to thank my supervisor, Dr. Gary Van Domselaar. Thank you for providing me with opportunities, connections, and most of all, your time—a scarce and precious resource in the time of COVID-19. I am grateful to you, and to the entire Bioinformatics group, for fostering such a positive (virtual) environment during a global pandemic. I would also like to say thank you to my supervisory committee members, Drs. Morag Graham, Charles Bernstein, and Denice Bay, for their expertise, guidance, and encouragement throughout the research process.

A very special thanks must be given to Christine Bonner and the other superstars in the Genomics Core. Your kindness, expertise, and commitment to this project have had an invaluable impact on its quality. Thank you also to Allison Poppel and the VADA team, as well as Dr. Will Hsiao's lab, for providing me with opportunities to expand my data science skills.

I would like to thank everyone in the Bioinformatics lab for their support and friendship. Special thanks to my fellow graduate students, Jill, Aaron, Nelson, and Jeremiah, for always checking in and supporting each other in various ways. Thank you to the entire MMID team and, in particular, to Angela Nelson, who makes everything run seamlessly and is also the kindest soul. I would also like to thank Drs. Natalie Knox and Jessica Forbes for their advice and mentorship, both prior to and during graduate studies.

Finally, thank you to my family and friends for their unwavering encouragement and support throughout the entire process. To Warren and Zach, thank you for truly being the best neighbours and lending me your space in times of need. To Noel, thank you for always being there for me.

Table of Contents

List of Tables

List of Figures

Abbreviations

bp	Base Pairs
CD4+	Cluster of Differentiation 4
CFU	Colony Forming Unit(s)
CLR	Centre-log Ratio
CRA	Colorectal Adenoma
CRC	Colorectal Cancer
cDNA	Complementary DNA
DNA	Deoxyribonucleic Acid
ENS	Enteric Nervous System
F/B Ratio	Firmicutes:Bacteroidetes Ratio
FDR	False Discovery Rate
GALT	Gut-Associated Lymphoid Tissue
gDNA	Genomic DNA
GI	Gastrointestinal
HMP/iHMP	Human Microbiome Project / Integrated HMP
HTS	High-Throughput Sequencing
IBD	Inflammatory Bowel Disease
IHMS	International Human Microbiome Standards Project
LSU	Large subunit of the ribosomal protein.
mg	Milligram
µl	Microlitre
mm	Millimetre
mRNA	Messenger Ribonucleic Acid
NAS	DNA/RNA Shield
ng	Nanogram
NGS	Next Generation Sequencing
NOD	Nucleotide-binding oligomerization domain
NP	No Preservative
PCA	Principal Components Analysis
PCR	Polymerase Chain Reaction
pM	Picomolar
RNA	Ribonucleic Acid
RNAP	RNAprotect
RNA-Seq	RNA Sequencing
rRNA	Ribosomal RNA
SOP	Standard Operating Procedure
SSU	Small subunit of the ribosomal protein.
WGS	Whole-Genome Sequencing

Contribution of Authors

The following thesis chapter includes content from published literature reviews authored in part by Molly Pratt. This content is provided in the introductory chapter in order to provide support for the rationale of the current study and to contextualize its findings within the broader field of meta-omic microbiome research.

Molly Pratt is the first author and was responsible for the original draft preparation of the review manuscript "Colorectal Cancer Screening in Inflammatory Bowel Diseases—Can Characterization of GI Microbiome Signatures Enhance Neoplasia Detection?", which was published in Gastroenterology as Pratt et al., 2022. Jessica D. Forbes, Natalie C. Knox, Gary Van Domselaar, and Charles N. Bernstein additionally contributed to this manuscript in various roles related to conceptualization, review & editing, and supervision. Two tables from this manuscript are included in the current thesis as supplementary tables in order to support the background information given in several introductory subsections. Namely, that stool is an informative medium for studying changes in the GI microbiome via meta-omics, and that meta-omics research can be applied to discover consensus signatures of disease. Molly Pratt was primarily responsible for reviewing and summarizing the information in these tables.

Molly Pratt is the first author and was primarily responsible for the original draft preparation of the review manuscript "Microbiome-Mediated Immune Signaling in Inflammatory Bowel Disease and Colorectal Cancer: Support From Meta-omics Data", which was published in Frontiers in Cell and Developmental Biology as Pratt et al., 2021. Jessica D. Forbes, Natalie C. Knox, Charles N. Bernstein, and Gary Van Domselaar additionally participated in developing the idea for the article and participated in the final editing. Relevant introductory information (including one figure) regarding the development and rise of meta-omics technologies, as well as key findings regarding disease-associated microbially-mediated metabolic disruption, from this manuscript were incorporated into the current thesis in order to

support the rationale and hypotheses of this study. As well, one table from Pratt et al., 2021 was included in the current thesis as a supplemental table because the information summarized in this table provides context to results from the current study. Molly Pratt was primarily responsible for conducting the literature review as well as summarizing the information in the figure and the table.

1 INTRODUCTION

1.1 The Gastrointestinal Microbiome

1.1.1 Microbial Communities in the GI Tract

The community of microorganisms that inhabit the gastrointestinal (GI) tract, including archaea, fungi, bacteria, and viruses that have co-evolved alongside their hosts, are referred to as the GI microbiota. The microbiota, along with their collective genetic material and metabolic activity, contribute to the diverse micro-ecosystem that is the GI microbiome, frequently called the gut microbiome. Though commonly referred to as though it were a single, uniform community, the GI microbiome is truly a diverse and dynamic collection of microbial niches that are heavily shaped by regional anatomical and biochemical landscapes. The GI tract itself is a multi-organ system; heterogeneous both along its length as well as cross-sectionally. Correspondingly, bacterial density in the GI tract changes dramatically, moving from the stomach (10^1–10^3 CFU/ml) to the distal colon (10^{10}–10^{12} CFU/ml) (O'Hara and Shanahan, 2006). The viability or presence of microorganisms in regions of the GI tract is influenced by biochemical gradients, including mucosal thickness, pH, and oxygen availability. Particular GI regions and their bacterial inhabitants are clearly summarized in (Martinez-Guryn et al., 2019). Within the lower GI tract, where the majority of microbial biomass is located, a distinction can be further made between the mucosal-associated and luminal communities (Eckburg et al., 2005). Thus, the gut microbiome is more aptly described as a collection of microbial communities that possess unique biodiversities, in terms of both community membership and functionality.

The bacterial component of the GI microbiota is the focus of most microbiome research to date, though explorations of other microbial organisms within the gut, such as archaea, fungi, and viruses, are becoming increasingly common (Hoffmann et al., 2013; Mar Rodríguez et al., 2015; Hannigan et al.,

2018; Guzzo et al., 2022). Consequently, much of what we currently know about the GI microbiome relates to the composition and function of GI bacteria. The development of an individual's GI microbiome begins during or shortly after birth, and its composition is impacted by a variety of factors, including genetics, birth mode, infant diet, and antibiotic use (Kim et al., 2019). Following weaning, individual GI microbiomes tend to become more diverse and may be altered by major changes in diet, exercise, and other lifestyle factors such as smoking status, pregnancy, acute or chronic illness, and environmental exposures (Yatsunenko et al., 2012; Lee et al., 2022). Absent of major lifestyle changes, the healthy adult GI microbiota reaches a relatively stable equilibrium state wherein inter-individual variation exceeds intra-individual variation over short to intermediate time periods (24 hours– 6 months) (Human Microbiome Project Consortium, 2012b; Mehta et al., 2018; Lee et al., 2022). Individual GI microbiomes may look similar in individuals that are genetically related, or that share common diets, the combination of which leads to broader ethnic similarities (Hillman et al., 2017) and has implications for disease cohort selection and properly matched healthy cohort controls.

1.1.2 Association with Human Health & Disease

As previously discussed, GI microbiota composition can vary among healthy populations according to geography, age, and other factors. Establishing a consensus for healthy GI communities is additionally complicated due to their relative diversity (compared to other body sites, *i.e.*, vaginal communities), and a large degree of inter-individual variation at lower taxonomic levels (Human Microbiome Project Consortium, 2012b). However, some broad hallmarks of healthy stool communities have been established in North American and European populations: At the phylum level, healthy GI communities are dominated by Firmicutes and Bacteroidetes, followed by Actinobacteria and Proteobacteria at lower abundances (Qin et al., 2010; Human Microbiome Project Consortium, 2012b;

Lee et al., 2022). Specific markers of health that are used as primary outcomes in the current study are discussed in Section 1.3. It is important to note that our understanding of what constitutes a so-called healthy microbiome will expand as researchers explore different approaches for community characterization. For example, characterization of GI microbial gene expression has revealed functional "shifts" during disease that cannot be captured via taxonomic analysis alone (Schirmer et al., 2018; Becattini et al., 2021).

Microbial GI community alterations have been associated with a variety of chronic illnesses, including type 2 diabetes mellitus (Qin et al., 2012), obesity (Magne et al., 2020), inflammatory bowel diseases (IBDs; Forbes et al., 2018), and colorectal cancer (CRC; Feng et al., 2015; Yachida et al., 2019), as well as several extraintestinal diseases such as multiple sclerosis and rheumatoid arthritis (Forbes et al., 2018). Currently, GI microorganisms are thought to play a major role in systemic bodily health through interactions with the immune system of their host, beginning in infancy (Patrick et al., 2020).

Much of what is known about the role of the gut microbiome in systemic host health comes from germ-free and gnotobiotic animal studies. For example, compared to their colonized counterparts, germ-free animals are more susceptible to infection and have severely compromised enteric mucosal immunity (Shanahan, 2002). Gnotobiotic mice, colonized with microorganisms from human fecal material, have been used to elucidate disease mechanisms, including drivers of tumorigenesis in CRC (Wong et al., 2017). Discoveries from studies such as these have fueled the now common conception of the GI microbiome as a symbiotic organ due to its extensive metabolic contributions to human health (O'Hara and Shanahan, 2006).

Importantly, the Enteric Nervous System (ENS) and Gut-Associated Lymphoid Tissue (GALT) of the host are co-localized with the dense microbial communities in the lower GI tract. There is evidence that the microbial community in the GI tract interacts with these host tissues through complex signalling

to affect changes in the local environment, such as modulating immune responses and inflammation (Thaiss et al., 2016; Martinez-Guryn et al., 2019; Cullen et al., 2020) (*see section 1.4.1*). Therefore, the microorganisms are also shaping their physiological environment. Disease-associated changes to the GI microbial community, often referred to as a microbial "imbalance," are considered to be dysbiotic; the symbiotic microbial organ is not functioning properly to maintain host health.

1.1.3 Dysbiosis: Cause or Symptom?

The concept of microbial dysbiosis is somewhat ill-defined. The criteria for classifying dysbiotic microbial communities are typically based on symptoms, disease manifestation, or deviation from a healthy comparator, rather than inherent properties of the underlying community (Walker, 2017; Shapiro et al., 2022). However, recent characterization of GI microbiomes from healthy individuals has enabled researchers to identify some hallmarks of GI dysbiosis. Lower microbial diversity and reduced abundance (or loss) of particular microorganisms or metabolites that are considered to be beneficial are emerging patterns of GI dysbioses (Lloyd-Price et al., 2019; De Vos et al., 2022). Likewise, particular IBD-associated bacteria may act as pathobionts since there is evidence that bacteria or their metabolic products can induce or exacerbate colitis in murine models (Ohkusa et al., 2003; Yang et al., 2020; Federici et al., 2022). Although GI dysbioses are associated with a wide array of human illnesses, the precise role of the GI microbiota in many diseases (i.e., whether dysbiosis is a trigger or symptom) remains to be discovered. This question is currently driving much of microbiome disease research, particularly for chronic GI-associated illnesses including IBD (Lloyd-Price et al., 2019) and CRC (Ternes et al., 2021).

Difficulties with the temporal relationship between dysbiosis and disease exist because changes in the microbiome could occur years before symptom onset. There are also many confounding factors: Dysbiosis may be influenced by genetic susceptibility, diet, antibiotic usage, smoking status, and other

environmental exposures, which could have individual as well as cumulative effects on the gut microenvironment, including epithelial barrier integrity and inflammation. For example, comparison of dysbiotic microbiome profiles from several immune-mediated inflammatory disease cohorts (Forbes et al., 2018) reveals that the abundance of certain GI taxa (*e.g. Actinomyces* and *Eggerthella* spp.) are consistently overabundant in all disease cohorts compared to healthy controls. However, the relative abundance of many taxa is significantly different between disease cohorts, suggesting unique etiologies. In order to achieve increased resolution with respect to changes to the microbial GI environment, it may be important to distinguish between taxonomic and functional dysbiosis. In microbiome disease research, focus is shifting toward the latter (Franzosa et al., 2019) because it is difficult to relate taxonomic relative abundance changes alone to complex and multifactorial disease etiologies. Within the GI microenvironment, multiple taxa may perform similar ecological functions (*i.e.,* anaerobic fermentation, short-chain fatty acid production, etc.), resulting in community-wide, gene-level redundancy, hence the difficulty defining a consensus "healthy" microbiota composition.

1.2 Meta-omics for Microbiome Research

1.2.1 The Rise of Culture-Independent Microbial Profiling Techniques

The myriad of life forms in the GI tract possess variable cell structures and growth conditions, as well as a wide range of genomic characteristics, including genome size, complexity, and genomic GC content. This makes capturing the community as a whole without bias extremely challenging. Many laboratory methods have been developed for capturing particular organisms, and thus are not optimized for broad community characterization. As such, despite a resurgence of complex culture technique development, the majority of GI bacterial species have never been cultured *in vitro* (Lagier et al., 2018).

In recent decades, the available technology for capturing and characterizing the GI microbiome has expanded greatly, resulting in fewer barriers due to cost and computational requirements, driving a rapid expansion of the field (Cullen et al., 2020). In particular, the advent of Next Generation Sequencing (NGS) techniques has allowed for direct analysis of microbial genomes without the need for culturing. Two common NGS techniques for studying microbial communities are taxonomic marker gene sequencing, or metataxonomics—for example, the (highly conserved between bacteria and archaea) 16S rRNA gene is sequenced as a representative of taxon abundance—and shotgun metagenomics, which uses whole-genome sequencing (WGS) to characterize the entire genomic content of a given community. These techniques, in particular, have allowed researchers to investigate the previously underestimated biodiversity of the GI microbiome. Culture-independent technologies allowing for direct untargeted analysis of microbial genes, transcripts, proteins, and other metabolites from diverse microbial communities are collectively termed meta-omics. Metatranscriptomics, metaproteomics, and metabolomics are other meta-omics techniques that aim to capture information about microbial communities at varying resolution (Figure 1). The application of multiple meta-omics techniques to samples from the same cohort yields correlated functional profiles (Lloyd-Price et al., 2019), supporting the utility of microbial sequence data for downstream prediction of functional activity. As a result of these new technologies, the healthy human gut microbiome has been well characterized (Méndez-García et al., 2018; Forster et al., 2019).

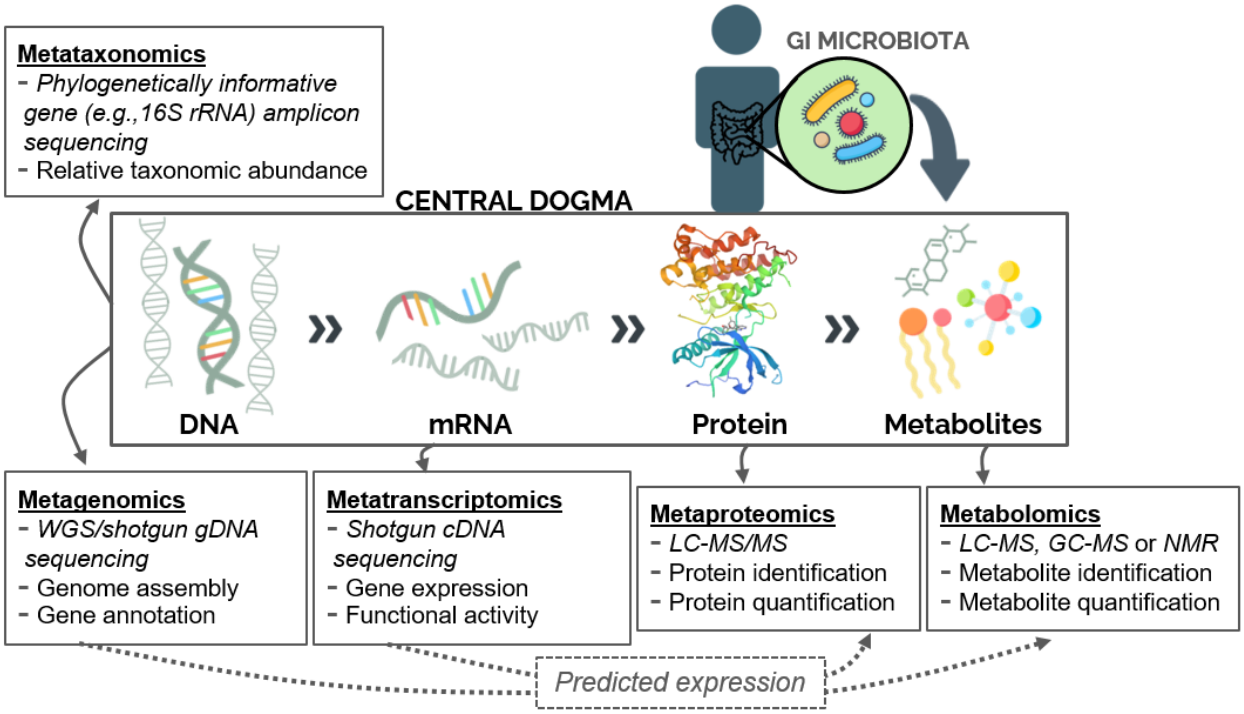


Figure 1. Meta-omics techniques for studying the human gut microbiome. According to the central dogma of molecular biology, information about biological systems increases in resolution moving from DNA to RNA to proteins and metabolites. Microbial communities can be characterized based on their collective gene content (metataxonomics or metagenomics), gene transcripts (metatranscriptomics), protein pool (metaproteomics), or metabolite pool (metabolomics). Additionally, sequence data can be used to predict downstream expression of proteins or metabolites (dashed arrows). Data analysis and interpretations (not shown) are accomplished with a variety of bioinformatics tools.
ABBR: WGS: Whole genome sequencing; LC-MS/MS: Liquid chromatography with tandem mass spectrometry; GC-MS/MS: Gas chromatography with tandem mass spectrometry; NMR: Nuclear magnetic resonance. (Image from Pratt et al., 2021).

1.2.2 Sequence-based Characterization: Metagenomics and Metatranscriptomics

Metagenomics and metatranscriptomics are two sequence-based approaches that can be used for microbiome profiling (*i.e.*, answering the popular questions "who is there" and "what are they doing") and also for downstream prediction of functional activity, according to the central dogma of molecular biology (Figure 1). Metagenomics describes the genetic content of a microbial community within a sample (typically stool or intestinal biopsy in the case of the GI microbiome) from microbial DNA, whereas metatranscriptomics uses reverse transcription to evaluate gene expression patterns from microbial messenger RNA (mRNA). Both techniques involve shotgun sequencing of nucleic acids isolated from a biological or environmental source and thus produce similar data structures in the form of sequence reads. Despite an increase in the number of metagenomic investigations, the mechanism of dysbiosis in many chronic diseases remains unclear, and it can be challenging to identify pathogenic agents from metagenomic abundance studies alone due to inter-individual community differences (Hold et al., 2014; Forbes et al., 2018; Knox et al., 2019b). Thus, there is a need for increased resolution in microbiome characterization in order to uncover mechanisms of disease. Metatranscriptomics presents an opportunity for sequence-based characterization that can inform on microbial gene expression and therefore functional activity.

Many tools and approaches have been developed for metagenomic microbiome characterization since its inception. However, not all of these approaches for data capture and analysis are directly applicable to metatranscriptomics, despite similar underlying data structures. Sequence-based approaches begin with the isolation of nucleic acids from biological samples. The isolation of microbial mRNA for metatranscriptomics is particularly complex for several reasons. First, RNA, and mRNA in particular, is easily degraded and thus requires considerate storage and laboratory preparation (Deutscher, 2006). Additionally, microbial RNA isolated from biological samples is dominated by ribosomal RNAs (95% or

more; Zoetendal et al., 2006), which, while useful for taxonomic investigation, cannot inform on microbial functional activity. Therefore, an additional processing step that depletes bacterial rRNA is required for downstream functional profiling (Macklaim and Gloor, 2018). However, unlike eukaryotic mRNAs, bacterial transcripts lack a distinguishing poly-A tail, further complicating unbiased mRNA capture. Fortunately, several effective approaches for prokaryotic mRNA enrichment have been developed (Petrova et al., 2017). Once bacterial mRNAs are successfully isolated, there are additional considerations for metatranscriptomic sequence library preparation because mRNA libraries are inherently less diverse than metagenomic libraries. Thus, development and optimization of products and workflows for metatranscriptomic data capture and analysis is ongoing (Bashiardes et al., 2016; Macklaim and Gloor, 2018; Zhang et al., 2021a).

1.2.3 Contemporary Meta-omic Investigations

When applied individually or in concert, meta-omics can be used to profile and predict disease-specific, location-specific, or longitudinal changes in microbial communities, and many such studies have provided valuable insights into human-associated microbial community dynamics (Human Microbiome Project Consortium, 2012b; Gevers et al., 2014; Hall et al., 2017; Schirmer et al., 2018). In particular, several large meta-omic cohort studies have been conducted that have greatly expanded our understanding of the role of the GI microbiome in human health and disease.

In addition to characterization, comprehensive studies such as the Human Microbiome Project (HMP), launched by the National Institutes of Health in 2007, have pioneered microbiome reference genome development and the integration of multiple meta-omics techniques in human cohorts (Human Microbiome Project Consortium, 2012a; The Integrative HMP (iHMP) Research Network Consortium, 2014; Proctor et al., 2019). The microbiome analysis pipelines used in the current study were in fact

developed for HMP research (Beghini et al., 2021). The first iteration of HMP research revealed that the taxonomic composition of the microbiome was not a strong correlate with host phenotype on its own. An early application of metatranscriptomics by HMP researchers revealed species-specific biases in transcriptional activity during disease (Schirmer et al., 2018). Thus, the second phase of the HMP integrated multiple meta-omic approaches, along with concordant host physiological measurements, to explore the interplay of host-microbiome interactions in three disease cohorts.

Discoveries from the HMP and other large-scale GI microbiome studies such as the European-based MetaHIT (Qin et al., 2010) are foundational to ongoing research. However, gold-standards for laboratory protocols and analysis workflows have yet to be established for GI microbiome research. For example, the laboratory protocols for metatranscriptomics analysis are quite varied (Cardona et al., 2012; Franzosa et al., 2014; Reck et al., 2015; Mehta et al., 2018; Lloyd-Price et al., 2019); thus, independent validation is needed to ensure consistency and quality of downstream sequence data.

1.3 Measuring the Microbiome: Indicators of Health & Disease

1.3.1 Stool as a Proxy for the Luminal GI Microbiome

Stool metagenomes represent the luminal GI microbiome rather than the mucosal-associated microbiome, which is captured through more invasive endoscopic methods (Eckburg et al., 2005). However, stool has several advantages for microbiome research: It is readily available, easy to collect, and contains a high proportion of microbial content relative to intestinal biopsies. Collection of stool is noninvasive, which may encourage individuals to participate in research cohorts or biobanks as compared with the use of more invasive sampling techniques (*e.g.,* endoscopic biopsy), especially when access to healthcare services is reduced, such as the current pandemic. Despite lacking site-specificity, under particular disease conditions, luminal microbiome profiles commonly reflect those of inflamed or

cancerous tissues compared to adjacent healthy tissues (Gevers et al., 2014; Zeller et al., 2014; Feng et al., 2015; Zhang et al., 2019), suggesting that changing selective pressures in the GI tract during disease may broadly shape luminal microbial community composition. Many meta-omic studies have detected significant differences in fecal microbiomes between CRC cases and healthy control samples (Appendix 1, Table S1), for example, supporting the role of stool as an informative medium for studying changes in the GI microbiome. The following sections will discuss common metrics that are used to evaluate stool microbiome profiles.

1.3.2 The Firmicutes:Bacteroidetes Ratio

The Firmicutes:Bacteroidetes (F/B) ratio is a quantitative metric based on the relative abundance of the two dominant GI phyla: Firmicutes and Bacteroidetes (Human Microbiome Project Consortium, 2012b; Abu-Ali et al., 2018). A broad range for healthy populations is reported between 0–8 (Magne et al., 2020). However, the relationship between these two phyla appears to be heavily dependent on the study population in question (Stojanov et al., 2020). For example, the F/B ratio differs according to age; it has been reported that infants and elderly individuals have significantly lower F/B ratios compared to adults (Mariat et al., 2009).

An "increased" F/B ratio, or expansion of Firmicutes at the expense of Bacteroidetes (compared to healthy controls), has been frequently cited in literature as a hallmark of obesity, and the F/B ratio in general has been proposed as a biomarker of health in that field and others. However, in the context of modern meta-omics it has recently been suggested to shift toward biomarkers that more accurately describe the metabolic activity of GI bacteria, rather than broad taxonomic measures, which may be differentially captured across methodologies (Stojanov et al., 2020). With this in mind, researchers may choose to establish alternate, though related, functional measures of community composition, such as the

ratio of facultative anaerobes to obligate anaerobes assessed in the iHMP (Lloyd-Price et al., 2019). This metric can be loosely associated with the F/B ratio based on *in vitro* growth conditions of Firmicutes and Bacteroidetes (Malele et al., 2018). Although its viability as a disease biomarker is questionable, the F/B ratio remains an informative taxonomic measure. Since it has been shown to be influenced by methodology in 16S rRNA metataxonomic studies (Magne et al., 2020), we have included it as a quantitative outcome for comparison of major GI phyla capture in the current methodological study.

1.3.3 Alpha Diversity

Perhaps the most dominant measure in microbiome research, alpha diversity (or within-sample diversity, a term adapted from the field of ecology; Whittaker 1960) describes the heterogeneity of a biological sample in terms of a particular microbial feature (for example, species, gene, or transcript). Two components of alpha diversity are directly adopted from ecological metrics: richness and evenness. The former is typically described using a count of the unique features within a dataset, and the latter describes the way that features are distributed within a sample. For example, a lower evenness score (approaching zero) indicates dominance of a particular feature, whereas a higher evenness score (approaching one) indicates a balanced population (Pielou, 1966). The Shannon Index (Shannon, 1948) and other equations (*e.g.*, Chao Index, Simpson Index) have been developed to describe these components in a single diversity metric. Regarding human-associated microbial communities, alpha diversity depends largely on body site. The GI microbiome is among the most diverse human-associated communities (Human Microbiome Project Consortium, 2012b) and for several GI microbiome model diseases, lower alpha diversity is a hallmark of dysbiosis (Forbes et al., 2018; Lloyd-Price et al., 2019).

1.3.4 Beta Diversity

Another prominent approach for comparing microbiome profiles is assessing the between-sample, or beta, diversity (Whittaker, 1960). Beta diversity describes how similar microbiome feature profiles are to each other, often quantified via some kind of distance or dissimilarity matrix. There are a variety of matrices in the literature that take unique approaches to measure dissimilarity, so care must be taken to select an appropriate matrix based on the nature of the data (Gloor et al., 2017). Additional information about dissimilarity matrices for microbiome meta-omics is given in section 1.5.2.

Because there are many features (genes, species, *etc*.) in microbiome data, quantitative between-sample comparison often makes use of dimension reduction techniques. Dissimilarity matrices are calculated to summarize the feature profiles, and these dissimilarities are commonly represented by linear distances, using ordination with principal component analysis (PCA). A PCA plot will quantify the proportion of overall variance in the data that is captured by each axis, or principal component, allowing for a high-level overview of the similarity between samples; those that have highly similar feature profiles will be located proximally on the PCA plot and vice versa. Qualitative differences between sample profiles can also be illustrated using heatmaps or other visualization techniques.

1.3.5 Differential Detection: Microbial Biomarkers

Microbial features that are found to be significantly enriched in either healthy or disease cohorts via differential abundance analysis are sometimes proposed as biomarkers. For example, a recent study by Federici et al. (2022) reported a particular clade of *Klebsiella pneumoniae* found to be enriched in IBD patients compared to healthy controls in multiple geographically distinct cohorts. The authors subsequently characterized the bacterial strains within the enriched clade and used them to develop a

mouse model of disease and to design targeted phage therapeutics, eventually culminating in a Phase I clinical trial (ClinicalTrials.gov: NCT04737876).

In contrast to the in-depth exploration performed by Federici and colleagues, many disease biomarkers proposed from *in situ* association studies are not investigated beyond their initial discovery. Furthermore, due to the variability of *in situ* discovery approaches, it is important to validate potential biomarkers in external cohorts. Differential abundance testing can be challenging because it typically requires large sample sizes in order to observe small effect sizes and has many confounders (Zhang et al., 2021b). Meta-analyses that incorporate data from multiple cohorts can reveal consensus signatures of disease, for example, in CRC fecal microbiomes (Appendix 1, Table S2). Thus, caution should be exercised when interpreting results from small individual biomarker studies.

1.4 The Value of Meta-omic Microbiome Research

1.4.1 The Role of Microbial Metabolites in Host Health

The intestinal epithelium is a crucial interface for host microbiome interactions. In a healthy gut, the host's immune system must be able to recognize and tolerate commensal organisms while retaining its ability to defend against pathogens or commensal invasion. For example, microbial taxa that are considered protective stimulate CD4+ T regulatory cell proliferation and maintenance of gut immune homeostasis (Atarashi et al., 2011; Knox et al., 2019a), whereas pathogenic organisms are recognized by Toll-like receptors and NOD-like receptors on CD4+ T cells, resulting in a coordinated adaptive immune response (Himmel et al., 2008). Similarly, the gut microbiota responds to host immune activation and local inflammation by altering gene expression and metabolite production (Thaiss et al., 2016; Becattini et al., 2021; Wilmes et al., 2022). Gathering a better understanding of this complex, bi-directional signaling is the basis for untargeted microbiome functional characterization.

A wealth of *in vitro* and *in vivo* research has shown that the GI microbiome plays an important role in the metabolism or modulation of biological effector molecules such as amino acids, polyamines, bile acids, fatty acids, B vitamins, and sphingolipids, which can directly or indirectly impact homeostasis, inflammation, and/or tumorigenesis (Pratt et al., 2021; De Vos et al., 2022; Wilmes et al., 2022). Additionally, given the impact of the early-life environment on individual microbiome development and subsequent health outcomes (Nielsen et al., 2020; Boutin et al., 2021), the role of commensal bacteria in immune development and sensitization is an active area of research. Studies using meta-omics to investigate gut microbiome functionality, specifically in IBD patients, have observed species-specific differences in metabolic activity corresponding to IBD disease state (Rehman et al., 2010; Schirmer et al., 2018) or therapy response (Ananthakrishnan et al., 2017), supporting the utility of meta-omics for investigating microbial metabolites.

1.4.2 Meta-omic Alterations in Model Diseases

Much of the meta-omic GI microbiome disease research to date has been conducted with IBD or CRC cohorts, and as such these are considered to be model diseases in this field. The gut microbiome is thought to be directly implicated in the etiopathogenesis of both IBD and CRC (Sellon et al., 1998; Peloquin and Nguyen, 2013). Although dysbioses have been characterized in both diseases, the cause and effect relationship between inflammation/tumorigenesis and dysbiosis remains unclear. Current theories hypothesize that alterations in the normal gut microbiome, caused by some environmental exposure, for example, antibiotic use (Ungaro et al., 2014), can trigger an inflammatory immune response that persists in the genetically susceptible host (Yang and Jobin, 2017; Kaplan et al., 2019; Knox et al., 2019a; Szamosi et al., 2020).

Meta-omics studies have been instrumental in revealing that microbial dysbiosis in IBD and CRC goes beyond taxonomic imbalance. While there are discrepancies regarding the differential expression of specific proteins or metabolites between cohorts, the research collectively paints a picture of systemic dysregulation of multiple microbe-mediated compounds in disease. For example, amino acid and fatty acid metabolism are commonly dysregulated in IBD or CRC compared to healthy controls (Appendix 1, Table S3), although the pattern of dysregulation is inconsistent. Other biochemical classes and pathways significantly dysregulated in a subset of studies include bile acids, vitamins B3 and B5, and sphingolipids. Interestingly, a similar pattern of metabolic dysregulation was apparent in the metabolome of a healthy CRC high-risk population (based on heritage and diet) compared to a healthy low-risk population (Ocvirk et al., 2020; Appendix 1, Table S3). This suggests that dysbiotic microbiomes are present before disease manifestation and contribute to CRC development. These shared patterns of metabolic dysregulation, captured through meta-omics, could indicate a shared etiology between IBD and CRC.

1.4.3 Microbiome-based Therapeutics & Diagnostics

Due to its suspected involvement in the etiopathogenesis of model diseases IBD and CRC, there is considerable interest in evaluating individual GI microbiomes within a clinical setting to aid screening and diagnosis using a personalized approach (Knox et al., 2019b; Pratt et al., 2022). As such, microbiome-based therapies aimed at ameliorating dysbiosis, including fecal microbiota transplantation and pre- or probiotics, have been investigated in IBD with promising results in some, but not all, patients (Matsuoka and Kanai, 2015).

Microbial signatures have advantages over other currently available noninvasive tests for early CRC and adenoma detection, for example. Changes in microbiome structure may be detected years before a bleeding lesion is detectable by currently available non-invasive tests. Further research may determine

whether there are different microbial signatures based on degree of dysplasia (*i.e.*, adenoma and flat low-grade dysplasia lesions vs high-grade dysplasia and cancer lesions). Alternatively, microbial signatures that identify any dysplasia could trigger colonoscopy rather than pursuing a paradigm of colonoscopy at set time intervals, even though those time intervals may not be appropriate for all persons. In a multi-omics study that differentiated patients with low or high-grade dysplasia from those with variable stages of CRC, Yachida et al. reported observable shifts in the microbiome-associated metagenome and metabolome from very early stages of cancer development (Yachida et al., 2019), illustrating that early changes in microbiome structure can be detected with an appropriate study design. Establishing consensus biomarkers from meta-omic stool profiles may aid non-invasive screening efforts and early CRC detection (Pratt et al., 2022).

However, many of the techniques used in meta-omic biomarker discovery require a substantial computational effort and corresponding technical expertise, making them difficult to implement routinely in a clinical setting. In addition, due to variation in study and cohort design, as well as a lack of detailed methodology in published research, the results of many contemporary meta-omics studies are difficult to reproduce. Thus, clinical meta-omics is currently hindered by a lack of consensus approaches and reproducibility challenges in microbiome research.

1.5 Challenges & Obstacles to Meta-omic Microbiome Research

1.5.1 Ground truths & Mock Communities for Microbiome Research

Similar to macro-ecological studies, the ground truth composition of an individual's GI microbiome cannot be determined. Rather, community compositions can only be observed, and observations may be heavily dependent on the tool or method used to make them. Moreover, the composition of a given individual's GI microbiome is variable and can change gradually over time with

aging, or more rapidly, in response to lifestyle change or injury (illness), further complicating the idea of "ground truth" comparison. Single-kingdom explorations capturing only bacteria or only fungi, for example, cannot comprehensively describe the complex communities and ecosystems within the GI tract. In the absence of a ground truth, microbiome researchers often employ "mock communities" in order to measure the suitability of a given technique for capturing mixed microbial taxa.

Mock communities may be assembled from whole-cell mixed microbial cultures (Tourlousse et al., 2021), balanced or unbalanced mixes of pre-extracted and quantified microbial genomic DNA (Forbes et al., 2018), or even microbial genomic sequence reads assembled into a synthetic mock-community *in situ* (Rajan et al., 2019). While the latter approach is typically used as ground truth for comparing taxonomic classifiers, the former two can be used as controls; assessing the efficacy of nucleic acid extraction in the case of whole-cell communities, and base-calling error or sequencing bias in the case of community DNA co-sequenced alongside biological samples (Tourlousse et al., 2022). These synthetic communities are typically designed to include a range of genomic GC contents and a mix of cell wall structures (*i.e.,* Gram-positive and Gram-negative) in order to simulate the complexity of microbial communities in biological samples (Tourlousse et al., 2021).

Despite these considerations, a disadvantage of whole-cell mock communities for fecal microbiome research is differential cell lysis due to the physical matrix containing the bacterial communities; bacterial cells contained within the matrix of a stool sample are more difficult to lyse than bacterial cells in solution, thus hindering direct comparison of nucleic acid extraction efficacy with stool samples (Mandal et al., 2020). Mock community DNA from non-GI-associated species may be spiked into stool in order to address the issue of matrix and quantify the absolute extraction error (Costea et al., 2017). However, this method could obscure the resulting community profile relative abundance values and thus is not practical outside of controlled error estimation/benchmarking (during pilot studies, for

example). Recently developed whole-stool reference materials, created from homogenized and pooled human stool standardized across multiple batches (Zymo Research, Mandal et al., 2020), are promising research aids.

1.5.2 Bioinformatics Analyses: An Excess of Choice

There are countless bioinformatics pipelines that share a common goal: to characterize microbial communities from shotgun sequence data. Many computational tools are designed and built by research teams for a particular application and subsequently published as open-source software, with varying levels of accompanying documentation. Some tools are optimized for speed and memory usage, employing k-mer-based approaches (where k-mers are unique subsequences of length *k*) to rapidly map sequences to indexed databases (*e.g.*, Kraken, CLARK; Wood and Salzberg, 2014; Ounit et al., 2015). Other tools aim to minimize false-positive assignments through highly stringent database curation (*e.g.*, MetaPhlAn; Beghini et al., 2021). Regarding taxonomic classification, an assortment of tools exist for both assembly-based (*e.g.*, MEGAHIT; Li et al., 2015) and assembly-free direct sequence mapping (*e.g.*, MEGAN; Huson et al., 2007), each with unique requirements or considerations. For example, some tools provide confidence scores with their assignments (*e.g.*, CLARK, PhymmBL) (Brady and Salzberg, 2011), while others do not (*e.g.*, MetaPhlAn). Therefore, the choice of analysis tool is not arbitrary and decision making during software selection should be informed by a basic level of computational knowledge, as well as a clear understanding of the ability of a given tool to capture the particular microbiome outcomes that one wishes to measure.

It has been demonstrated that significant differences in microbial community 16S rRNA gene profiles from stool and other environments can be introduced during data pre-processing, and the choice of differential abundance software can likewise produce dramatically different results from the same input

dataset (Szamosi et al., 2020; Nearing et al., 2022). Other studies have evaluated the combined effect of sequence depth and reference database on metagenomic data and found that differences in database composition and size affect observed taxonomic profiles, with larger or more comprehensive databases typically providing more classifications per sample (Rajan et al., 2019; Tamames et al., 2019). The strategy used for taxonomic assignment can also have a great impact on taxonomic profiles; for example, k-mer-based assignment approaches rely on matches to sequence substrings and thus, have increased risk of identifying species that are proximally related to community taxa (and thus an increased false positive rate as the taxonomic resolution increases)(Rajan et al., 2019).

Another important consideration for microbiome sequence data analysis is the nature of the data itself. Meta-omic data are typically noisy, sparse (zero-inflated), high-dimensional, and are not normally distributed. Although many tools created for genomic microbiome analysis have been adapted from the related fields of ecology and RNA-Seq, the suitability of these methods directly applied to sequence data analysis has been re-evaluated (McMurdie and Holmes, 2014; Gloor et al., 2017). Due to the nature of NGS approaches, genomic sequence data are inherently compositional; the sum of the data is dependent on the capacity of the sequencing instrument, and the observed abundance of each feature is dependent on that of all other features. This is in contrast to ecological data, which is accurately represented by counts, and has implications for data normalization and ordination (Gloor et al., 2017). Microbiome researchers should therefore be aware of compositional data assumptions and of analysis methodologies that satisfy compositional requirements. For example, the practice of rarefying microbiome data, once widely used, is now considered inappropriate—due to the exclusion of data, which results in inflated false positives during comparisons—and alternative normalization approaches informed by statistical theory (McMurdie & Holmes, 2014) or log-ratio transformations of the data in lieu of normalization (Gloor et al., 2017) are favoured. Additionally, the commonly used distance/dissimilarity measures UniFrac, Bray-Curtis, and

Jensen-Shannon do not account for the compositional nature of high-throughput sequencing (HTS) data and may produce spurious results if data compositionality is not addressed. This high degree of variability between metagenomic analysis approaches often makes the comparison of results between studies challenging, if not impossible.

Meta-analyses that incorporate published microbiome sequence data from multiple studies are valuable to mask inter-study heterogeneity or inconsistencies. Meta-analyses can help overcome challenges associated with the low signal-to-noise ratios often seen in small metagenomics studies. They are especially valuable in explorations of the microbiome aimed at biomarker discovery. Increased sample size, and the use of multiple independent cohorts for disease classification model development and evaluation, contribute to an overall lower risk of bias in disease prediction models developed from meta-analyses compared with individual studies (Moons et al., 2014). Regardless of the application, public microbiome sequence data should be assessed for potential sources of bias originating prior to sequence generation (see below) prior to any direct comparison and where possible, public data should be processed and analyzed using the same approach (software + database) as the data to which it is compared.

1.5.3 Standard Laboratory Approaches for Microbiome Capture from Stool Samples

Sequence-based meta-omics approaches rely on information captured from microbial nucleic acids, DNA and RNA. Importantly, nucleic acid integrity within stool samples can be affected by temperature, time, the presence of nucleases, and degradation during sample processing (Reck et al., 2015; Szóstak et al., 2022). Due to the impracticality of freezing stool samples immediately upon collection and maintaining a cold chain during transport, chaotropic reagents are often used to stabilize nucleic acids during long periods of storage and prevent excessive degradation. This is accomplished by inactivating nucleases (found within the sample as well as its surrounding environment) via denaturation with

guanidinium thiocyanate or other chaotrope (Shen, 2019). Common stabilization reagents used in stool microbiome research include commercial offerings such as OMNIgene® GUT (DNA Genotek, Bethlehem, PA, USA), RNAlater™ (Thermo Fisher Scientific, Waltham, MA, USA; Costea et al., 2017), RNAprotect (Qiagen, Germantown, MD, USA; Reck et al., 2015), and DNA/RNA Shield (Zymo Research, Irvine, CA, USA; Koorakula et al., 2022); other studies use 100% ethanol (Franzosa et al., 2014; Lloyd-Price et al., 2019), guanidinium thiocyanate (Dore et al., 2015), or other in-house buffers to stabilize nucleic acids in stool. Commercial offerings, including those listed above, typically use guanidinium thiocyanate within a mixture of other reagents, which may not be disclosed (see Appendix 4). It is worth noting that although RNALater™ is recommended by some researchers for metagenomic sample stabilization due to its ability to preserve nucleic acid integrity during room temperature storage (Zoetendal et al., 2006; Costea et al., 2017), it has been shown to introduce a transcriptional bias that is maintained throughout storage (Reck et al., 2015), and as such may not be ideal for metatranscriptomic profiling.

Isolation of nucleic acids from mixed microbial communities in stool samples is additionally challenging because many existing microbiology approaches are optimized for a particular organism or cell properties (ie., Gram-positive or Gram-negative cell wall structure). Mechanical cell lysis, accomplished by homogenization of the sample in the presence of small (millimetre-range) inert beads, has been shown to be effective for lysing mixed cell communities and is a common approach to microbiome nucleic acid isolation from stool (Santiago et al., 2014; Tourlousse et al., 2021). However, many commercial kits make use of alternative lysis methods such as chemical (using high pH, salt, or detergents to disrupt cell membranes), enzymatic (using lysozyme and/or Proteinase K to digest cell wall components), heat-based, or a combination of approaches. In metagenomic research, the Gram-positive

to Gram-negative ratio has been shown to increase with bead-beating time (Szóstak et al., 2022), indicating that observed taxonomic profiles are influenced by nucleic acid extraction methodology.

Regarding the isolation of nucleic acids from cell lysates, various techniques have been developed. Solid-phase nucleic acid purification is a common approach, used in many commercial kits (Shen, 2019; Table 1). Solid-phase purification is typically achieved using a spin column containing an inert matrix (*e.g.,* silica) that will bind to the nucleic acids under particular pH and salt buffer conditions and centrifugation (Shen, 2019). In recent years, microbiome researchers, including iHMP investigators, have made use of newer affinity-based nucleic acid extraction via magnetic beads (Table 1). The use of magnetic beads for binding nucleic acids in solution allows for high-throughput automation on instruments such as the chemagic™ (Perkin Elmer, Waltham, MA, USA), which employs metal rods that are magnetized and de-magnetized for rapid separation and resuspension of nucleic acids throughout the protocol.

Commercially developed nucleic acid extraction kits for metagenomics differ in their cell lysis methods and isolation approach, and many methodological studies that compare isolation protocols have been conducted (Santiago et al., 2014; Reck et al., 2015; Costea et al., 2017; Lim et al., 2018; Tourlousse et al., 2021; Koorakula et al., 2022; Table 1). Several large-scale projects, including the International Human Microbiome Standards (IHMS) Project (Costea et al., 2017) and the HMP (Human Microbiome Project Consortium, 2012a), have attempted to develop standard protocols for stool metagenomics, some of which are listed in Table 1. However, recommended protocols can quickly become obsolete. Commercial vendors frequently alter the names and components of their kits, or may discontinue certain offerings without notice. For example, the frequently cited rRNA depletion method from Epicentre, RiboZero, was discontinued in 2018 and later rebranded as an “improved” Ribo-Zero Plus (Illumina), which uses enzymatic rRNA depletion rather than the Streptavidin bead-based pulldown method

employed in the previous version. During writing (2022), Illumina released another updated “Ribo-Zero Plus Microbiome” product, specifically marketed for microbiome research and bacterial metatranscriptomics. Although this product was not available in time to be incorporated in this study, it could be valuable for future explorations. Additionally, the popular MoBio PowerFecal/PowerMicrobiome DNA/RNA Isolation kits were rebranded in 2016 as Qiagen QIAamp PowerFecal/PowerMicrobiome DNA/RNA kits (Kit B, Table 1). Likewise, the Qiagen QIAamp DNA Stool Mini Kit, which was evaluated by (Costea et al., 2017; Tourlousse et al., 2021) and found to be highly accurate for taxonomic profiling of human fecal samples (Protocol Q, Table 1), was discontinued by the manufacturer and replaced with the QIAamp Fast DNA Stool Mini Kit (Kit C, Table 1), a product with a unique catalogue number yet almost identical name. As such, recommendations for particular products are frequently outdated, hindering workflow reproducibility. Furthermore, there is evidence that different sample types may require unique approaches (Koorakula et al., 2022). Thus, despite the suggestions provided by the IHMS and HMP, there appears to be no consensus approach to sequence-based stool meta-omics in the literature, and methodological validation research is ongoing (Lim et al., 2018; Tourlousse et al., 2021).

Table 1. Validated Protocols for Metagenomic DNA Extraction from Stool.

Protocol	Kit Name (Vendor)	Details of Lysis Approach	Isolation Approach	Evaluated in: (Refs)
Kit A, current study	Quick-DNA Fecal/Soil Microbe Kit (Zymo Research, Irvine, CA, USA)	10 minutes of bead-beating in ZR BashingBead Lysis Tubes (0.6 mL of 0.1 mm and 0.5 mm silica-based beads; Zymo Research) on a vortex genie adaptor (MO BIO Laboratories Inc. Carlsbad, CA, USA) in the presence of Genomic Lysis Buffer (Zymo Research)	Solid-phase extraction (spin column)	Tourlousse et al., 2021
"Protocol Q", IHMS	Qiagen QIAamp DNA Stool kit (Qiagen, Germantown, MD, USA*; discontinued)*	Bead-beating with Fastprep™ instrument (MP Biomedicals, Irvine, CA, USA) for 8 minutes, 15 s in the presence of 0.3g 0.1mm zirconia beads	Solid-phase extraction (spin column)	Costea et al., 2017; Tourlousse et al., 2021
Kit C, current study	QIAamp Fast DNA Stool Mini Kit (Qiagen)	Thermal lysis in InhibitEX buffer at 70C for 10 minutes in the presence of proteinase K	Solid-phase extraction (spin column)	
Kit B, current study	QIAamp PowerFecal Pro / RNeasy PowerMicrobiome (Qiagen) (*formerly* MoBio PowerFecal DNA/RNA Isolation)	15 minutes of bead-beating in PowerBead Pro Tubes (Ceramic 1.4 mm, Ceramic 2.8 mm, Glass 0.5 mm, Metal 2.38 mm, and Garnet 0.70 mm; Qiagen) on a vortex Genie adaptor (Mo Bio Laboratories Inc.)	Solid-phase extraction (spin column)	Reck et al., 2015; Koorakula et al., 2022
"Protocol H", IHMS	In-house SOP	Bead-beating with Bead Beater™ (BioSpec Products, Bartlesville, OK, USA) cycles of 5 min medium speed, rest for 10 min, 5 min medium speed again (total 10 minutes) in the presence of 750mg of 0.1mm glass beads	Organic extraction (solvent precipitation)	Costea et al., 2017
"Protocol N", Tourlousse et al., 2021	ISOSPIN Fecal DNA kit (Nippon Gene Co., Tokyo, Japan)	Bead-beating with Fastprep™ instrument (MP Biomedicals), 3 x 60s, bead size and material unknown.	Solid-phase extraction (spin column)	Tourlousse et al., 2021
iHMP protocol - DNA and RNA isolation for metagenomics and metatranscriptomics	Chemagic MSM I with the Chemagic DNA Blood Kit-96 (Perkin Elmer, Waltham, MA, USA)	Automated on Chemagic MSM I (Perkin Elmer) in the presence of magnetic beads and Lysis Buffer B (guanidinium chloride). Unknown time.	Magnetic bead-based purification	Lloyd-Price et al., 2019

1.5.4 Methodological Variation among Methodological Studies

A major drawback of current reporting in GI microbiome research is a lack of properly detailed methodologies. Many investigators provide the name of the kit used for nucleic acid isolation but do not go into detail regarding particulars, which can also affect the study outcomes. For example, genomic library input amounts and intensity of mechanical lysis approaches (time/velocity) are often left out of publications, obscuring key aspects of data generation. This issue persists even in projects that aim to provide frameworks for future research. For example, protocols from the iHMP (available at ibdmdb.org) provide a broad overview of nucleic acid isolation methods (i.e., combined chemical and mechanical lysis with magnetic bead-based purification) without specifying the type and size of beads used for mechanical lysis, nor the time or velocity of bead-beating, both of which can affect resulting profiles (Costea et al., 2017). Other prominent studies in the field, such as (Schirmer et al., 2018), do not provide any detail regarding the method of nucleic acid isolation, instead focusing on sequencing approach and data analysis in their Methods. These gaps in methodology hinder the reproducibility of microbiome research.

As a consequence of the vast amount of choice presented to researchers for both meta-omic nucleic acid extraction kit and sequence data analysis tool, many methodological comparison studies have been conducted. However, key differences in the approaches used hinder straightforward comparisons across such studies. Comparisons are often made between commercial DNA or RNA extraction kits (Lim et al., 2018), sometimes with the inclusion of in-house developed reagents, standard operating procedures (SOPs), or modifications (Reck et al., 2015; Tourlousse et al., 2021). One study evaluating the nucleic acid extraction approaches used by MetaHIT and the HMP found that each is differently suited for capturing particular microbial taxa (Wesolowska-Andersen et al., 2014). Nucleic acid extraction may be evaluated in parallel with storage methods such as stabilization reagents (chaotropes), temperature, and time (Reck et al., 2015), or with downstream library preparation methods (Costea et al., 2017). In contrast,

a single extraction approach may be used to facilitate the comparison of sequence data capture and analysis methods, as in (Alberti et al., 2014; Rajan et al., 2019). Importantly each methodological comparison is conducted under particular conditions, which are highly controlled in order to isolate the effect of the parameters being tested. A consequence of this is that the results may not be broadly applicable to "real-world" scenarios wherein researchers have limited control over one or more of these parameters.

Furthermore, although the results from these methodological studies are undoubtedly informative and useful for microbiome research, incorporating their findings into the "big picture" of microbiome research is not straightforward. Individual studies often have different outcomes by which they measure the success or suitability of a given approach. For example, studies may focus on the quantity (yield) and size (fragmentation) of extracted nucleic acids as primary outcomes (Cardona et al., 2012; Reck et al., 2015). Alternatively, outcomes may result from data analysis such as within-sample taxonomic richness (Cardona et al., 2012), beta diversity based on community composition at varying taxonomic levels (Costea et al., 2017; Lim et al., 2018), or agreement with the ground truth when a mock community is used (Tourlousse et al., 2021). As such, it is important to remember that what is deemed "best" by any individual study depends entirely on the outcomes used to measure success.

Despite these inconsistencies, there is consensus within the field of GI microbiome research regarding certain parameters involved in stool sample storage and processing. For example, it is known from metataxonomic research that storage conditions, namely temperature and use of a stabilizing reagent, as well as nucleic acid extraction approach, can alter 16S rRNA microbiome profiles (Cardona et al., 2012; Santiago et al., 2014). The effect of temperature on nucleic acid quality has also been evaluated in WGS metagenomics and metatranscriptomics (Franzosa et al., 2014; Reck et al., 2015; Koorakula et al., 2022). The gold standard used for comparison in these studies is flash-frozen stool, and the results are generally in agreement that stool samples collected for research should be frozen at -20 °C or lower as soon as

possible following collection, and that freeze-thaw cycles should be minimized where possible to prevent excessive nucleic acid degradation (Cardona et al., 2012; Wu et al., 2019). However, the application of this standard approach may be impractical for studies employing home-collected stool, due to the auto-defrost cycling that occurs in most residential freezers. Therefore, temperature variation during sample storage may be out of the investigators control and a stabilizing agent may be required. Additionally, mechanical cell lysis via bead-beating has been shown to improve recovery of hard-to-lyse Gram-positive bacterial rRNAs from stool stored with (Lim et al., 2018) or without (Santiago et al., 2014) a stabilizing buffer and subsequently increase observed microbial diversity, compared to non-mechanical (chemical, enzymatic, or heat-based) lysis approaches.

1.6 Hypotheses & Objectives

Despite recommendations from the IHMS and HMP, there is no widely-accepted standard approach to stool metagenomics and metatranscriptomics. Many researchers have demonstrated that nucleic acid stabilization and extraction can have a profound effect on 16S rRNA microbiome community profiles (Cardona et al., 2012; Santiago et al., 2014; Lim et al., 2018). However, to our knowledge, the combined effect of commercial stabilizing reagents and extraction kits on taxonomic and functional microbiome outcomes has not been explored for WGS-based approaches. Based on these findings, as well as findings from primary metatranscriptomic microbiome characterizations (Schirmer et al., 2018; Lloyd-Price et al., 2019), it was hypothesized that i) stool sample storage and processing significantly impacts observed stool microbiome profiles and ii) that metatranscriptomics will provide added evidence, relative to metagenomics alone, into GI functional activity.

The experimental objectives of this research were to evaluate several different laboratory approaches, using commercially available reagents, for metagenomic and metatranscriptomic stool microbiome profiling. Also, to describe the impact of these methods on the resulting observed microbiome diversity

profiles, in order to identify microbiome profile outcomes that are associated with laboratory methodology and inform future meta-omics endeavors in our laboratory. Specifically, the aims of this project were to i) divide a single human stool sample into three different nucleic acid preservant conditions, ii) extract DNA and RNA from stool using multiple commercial kits (from different vendors or with differing approaches), and iii) generate and compare taxonomic and functional profiles based on the nucleic acid sequences from each experimental group. The profiles are compared according to common indicators of microbial community health taken from modern microbiome research in order to evaluate how study results may be biased according to methodology.

2 METHODS

2.1 Stool Sample Collection & Storage

A single human stool sample from a healthy individual was procured from BioIVT (Westbury, NY, USA). The fecal sample was collected and processed by the vendor as follows:
Feces collected from one individual (40 g total) was subsequently divided into 3 equal aliquots: 1) 1 x 14 g (No Processing) LOT# HMN358862A; 2) Feces Homogenate in 1:1 RNAprotect (cat# 76104, Qiagen, Germantown, MD, USA; Appendix 4) per g of tissue, LOT# HMN358862B; 3) Feces Homogenate in 1:1 DNA/RNA Shield (cat# R1100-50, Zymo Research, Irvine, CA, USA; Appendix 4) per g of tissue, LOT# HMN358862C.

The aliquots were shipped from the vendor site on dry ice via FedEx to the laboratory at 1015 Arlington St. in Winnipeg, MB, Canada, whereupon they were stored at -80 ℃ immediately upon receipt. The aliquots underwent a single freeze/thaw cycle to subdivide them into smaller ~200 mg (single-time use) sub-aliquots for subsequent nucleic acid extraction.

2.2 Nucleic Acid Extraction from Stool Samples

Genomic DNA and total RNA were extracted separately from the distinctly preserved aliquots of stool [aliquots 1), 2) or 3) as above, denoted as no preservative (NP), RNAprotect (RNAP), and DNA/RNA Shield (NAS) in this study]. Templates were prepared (200 mg per tube) in parallel to represent technical replicates. In order to evaluate the effect of different extraction approaches on microbiome profiling, multiple commercial kits with key differences in their lysis approach (details below) were used to isolate DNA or RNA.

2.2.1 Phase I - DNA Extraction Approaches for Metagenomics

The following kits were used for genomic DNA extraction from stool aliquots:

A) Quick-DNA Fecal/Soil Microbe Kit (cat # D6010, Zymo Research). This kit has been used for microbiome explorations in our laboratory previously, so an existing optimized version of the manufacturer's instructions were followed ["NML-revised Zymo protocol" in Appendix 2], which specifies 10 minutes of bead-beating on a vortex Genie adaptor (cat # 13000-V1-24, MO BIO Laboratories Inc. Carlsbad, CA, USA) at maximum speed. This kit uses proprietary ZR BashingBead Lysis Tubes, which contain 0.1 mm and 0.5 mm silica-based beads, as well as proprietary Genomic Lysis Buffer (containing guanidinium thiocyanate and glycerol) for a combined chemical and mechanical approach to bacterial cell lysis. Six stool replicates per preservation method (total 18 stool samples), as well as two blank samples were processed with this kit.

B) QIAamp PowerFecal Pro Kit (cat # 51804, Qiagen). The manufacturer's protocol was followed with the following specifications: 15 minutes of bead-beating on a vortex Genie adaptor (MO BIO Laboratories Inc.) at maximum speed. This kit uses a combination of chemical and mechanical cell lysis via proprietary solution CD1 (containing sodium thiocyanate) and PowerBead Pro Tubes containing a mixture of differently sized beads (Ceramic 1.4 mm, Ceramic 2.8 mm, Glass 0.5 mm, Metal 2.38 mm, and Garnet 0.70 mm). Six stool replicates per preservation method (N=18), as well as two blank samples were initially processed with this kit. Due to unexpected inconsistency in terms of DNA yield, an additional three replicates per preservation method (N=9) and one additional blank sample were subsequently processed, for a total of 27 stool samples and 3 blank samples.

C) QIAamp Fast DNA Stool Mini Kit (cat # 51604, Qiagen). The manufacturer's instructions for pathogen detection were followed with the following specifications: thermal lysis was carried out by incubation in InhibitEX buffer at 70 ℃ for 10 minutes. This kit requires incubation at a high temperature and Proteinase

K addition for a combined thermal and enzymatic lysis approach. Six stool replicates per preservation method (total 18 stool samples), as well as two blank samples were processed with this kit.

At least 2 blank (mock) extractions were performed per kit in parallel with the stool extractions and quantified to assess contamination. A suitable positive extraction control mimicking the complexity of the stool matrix was not available at the time of DNA extraction. DNA concentration (ng/µl) was measured for all samples using the Qubit High Sensitivity (HS) 1xdsDNA Assay at 2 µl (Thermo Fisher Scientific, Waltham, MA, USA), which quantitates DNA concentrations ranging from 0.005 - 120 ng/µl using a fluorescence-based quantification. Samples that were below the limit of detection were recorded as such, while samples above the limit of detection were measured again with the 1xdsDNA Assay at 1 µl sample volume, and when necessary additionally assessed using the Qubit dsDNA Broad Range (BR) Assay (Thermo Fisher Scientific), which quantitates DNA concentrations from 4–4,000 ng/µl.

2.2.2 Phase II - RNA Extraction Approaches for Metatranscriptomics

The following kits were used for total RNA extraction from 200 mg stool aliquots:

A) Quick-RNA Fecal/Soil Microbe Microprep (cat # R2040, Zymo Research). The manufacturer's instructions were followed with the following specifications: 5 minutes of bead-beating on a vortex Genie adaptor (MO BIO Laboratories Inc.) at maximum speed, included DNase I (Qiagen) treatment, and following a 2-minute incubation, elution of RNA into 10 µl of RNase-Free water (Zymo Research). This kit makes use of proprietary ZR BashingBead Lysis Tubes, which contain 0.1 mm and 0.5 mm silica-based beads, as well as S/F RNA Lysis Buffer (containing sodium iodide) for a combined chemical and mechanical approach to bacterial cell lysis. Six stool replicates per preservation method (total 18 stool samples), as well as two blank samples, and two ZymoBIOMICS™ Fecal Reference with TruMatrix™ Technology (cat# D6323, Zymo Research) samples were processed with this kit. In order to optimize

RNA input for downstream library preparation, some stool samples were pooled together, for a total of 11 stool samples (representing 3-4 technical replicates per group).

B) RNeasy PowerMicrobiome Kit (cat # 26000-50, Qiagen). The manufacturer's instructions were followed with the following specifications: the optional phenol-based lysis step was not performed, 3 minutes of bead-beating on a vortex Genie adaptor (MO BIO Laboratories Inc.) at maximum speed, 70% ethanol used in lieu of solution PM4 to prevent co-purification of small RNAs (5s RNAs, tRNAs and degraded RNAs), we elected to include one DNase I (Qiagen) treatment, and following a 2-minute incubation, elution of RNA into 50 µl of RNase-Free water (Qiagen). This kit uses proprietary PowerBead Tubes containing glass 0.1 mm beads and solution PM1 for a combined chemical and mechanical approach to bacterial cell lysis. Six stool replicates per preservation method (total 18 stool samples), as well as two blank samples, and two ZymoBIOMICS™ Fecal Reference with TruMatrix™ Technology (cat# D6323, Zymo Research) samples were processed with this kit. In order to optimize RNA input for downstream library preparation, some stool samples were pooled together, for a total of 11 stool samples (representing 3-4 technical replicates per group).

Two blank (mock) RNA extractions were performed per kit, in parallel with stool extractions, and quantified to assess contamination. Two aliquots of microbial reference material composed of stool from healthy donors, ZymoBIOMICS™ Fecal Reference with TruMatrix™ Technology (cat# D6323, Zymo Research), were co-extracted per kit (100 µl, RNeasy PowerMicrobiome; 250 µl, Quick-RNA Fecal/Soil Microbe) as a positive RNA extraction control and quantified. RNA concentration (ng/µl) was measured employing 1 µl of RNA sample using the Qubit High Sensitivity (HS) RNA Assay (Thermo Fisher Scientific), which captures RNA concentrations ranging from 0.2 - 200 ng/µl. Samples that were below the limit of detection were recorded as such and excluded from further processing, while samples above

the limit of detection were diluted 1:10 (Qiagen) or 1:2 (Zymo) in Nuclease-free water and reassessed with the same HS RNA assay.

2.3 Illumina Library Preparations

Genomic DNA isolated from stool samples was normalized to 50 ng in 100 µl prior to library preparation. Libraries were prepared from stool samples and extraction blanks according to a revised version of the Nextera XT DNA Library Prep Kit (Illumina, San Diego, United States) [NML 1/4th Volume Nextera XT protocol, A.2]. Additional libraries were prepared in parallel from two mock community DNA standards containing either a balanced (cat # D6305, Zymo Research) or log (cat # D6311, Zymo Research) distribution of pre-quantified microbial DNA. A total of 72 libraries were evenly pooled, and the combined pool was size-selected at a range of 600–1000 bp using a BluePippin (Sage Science, Beverly, MA, US). Library size distribution was determined by electrophoresis on a 2200 TapeStation (Agilent Technologies Inc. Santa Clara, CA, USA) using a Genomic DNA ScreenTape® and DNA concentration (ng/µl) was measured using the Qubit High Sensitivity (HS) 1xdsDNA Assay at 1 µl (Thermo Fisher Scientific). The final pooled library had an average peak size of 830 bp and a concentration of 4.29 ng/µl.

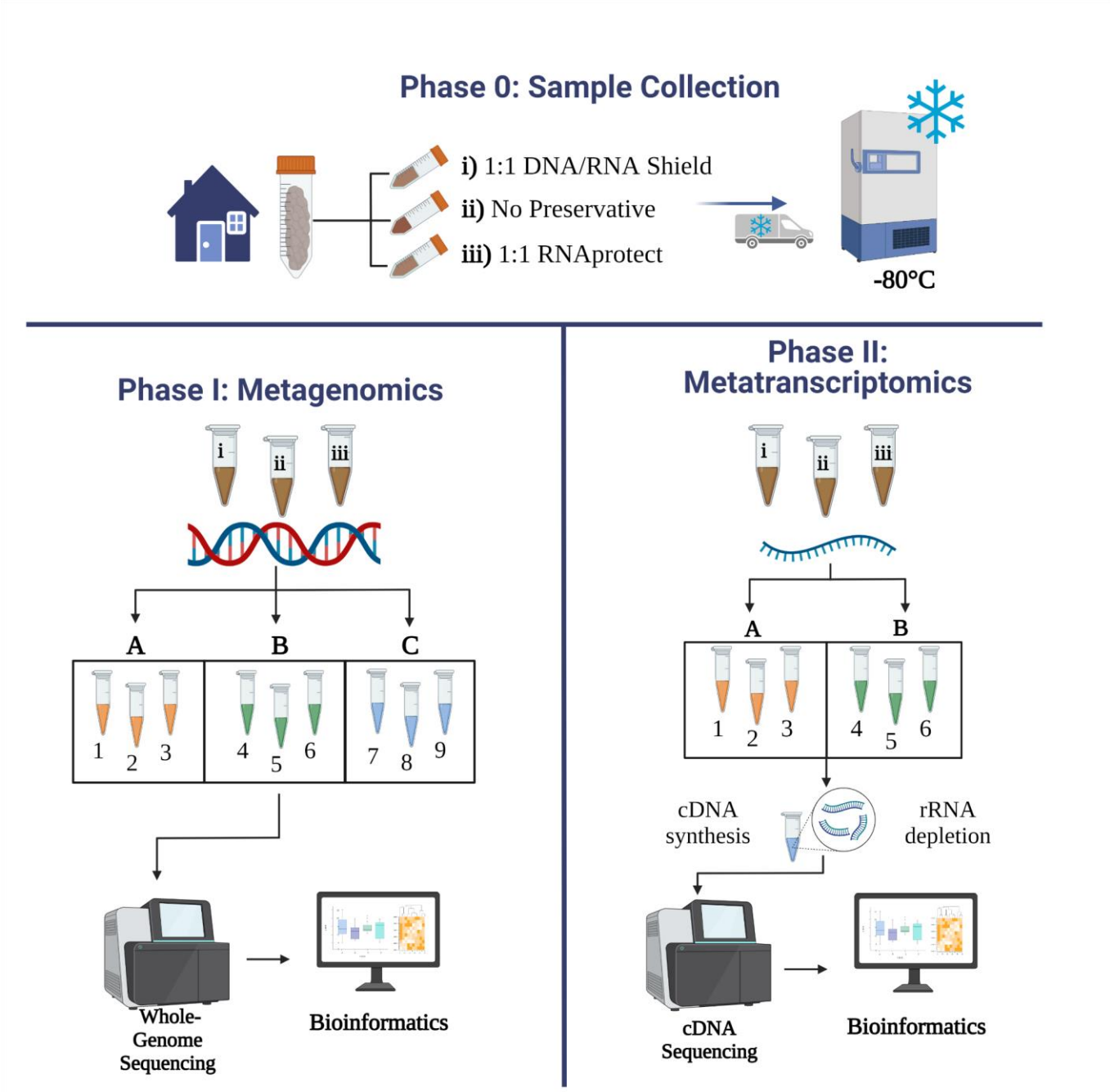


Figure 2. Graphical Methods. The experimental design of the current study is divided into three phases. In Phase 0, a single stool sample is collected from a healthy donor. Metagenomic (Phase I) and metatranscriptomic (Phase II) profiling of preserved stool aliquots (i-iii) using commercial nucleic acid extraction approaches (A-C). Observed microbiome profiles from experimental groups (1–9, Phase I or 1–6, Phase II) are compared against one another to evaluate changes to profiles that can be associated with methodology. Image created in BioRender.

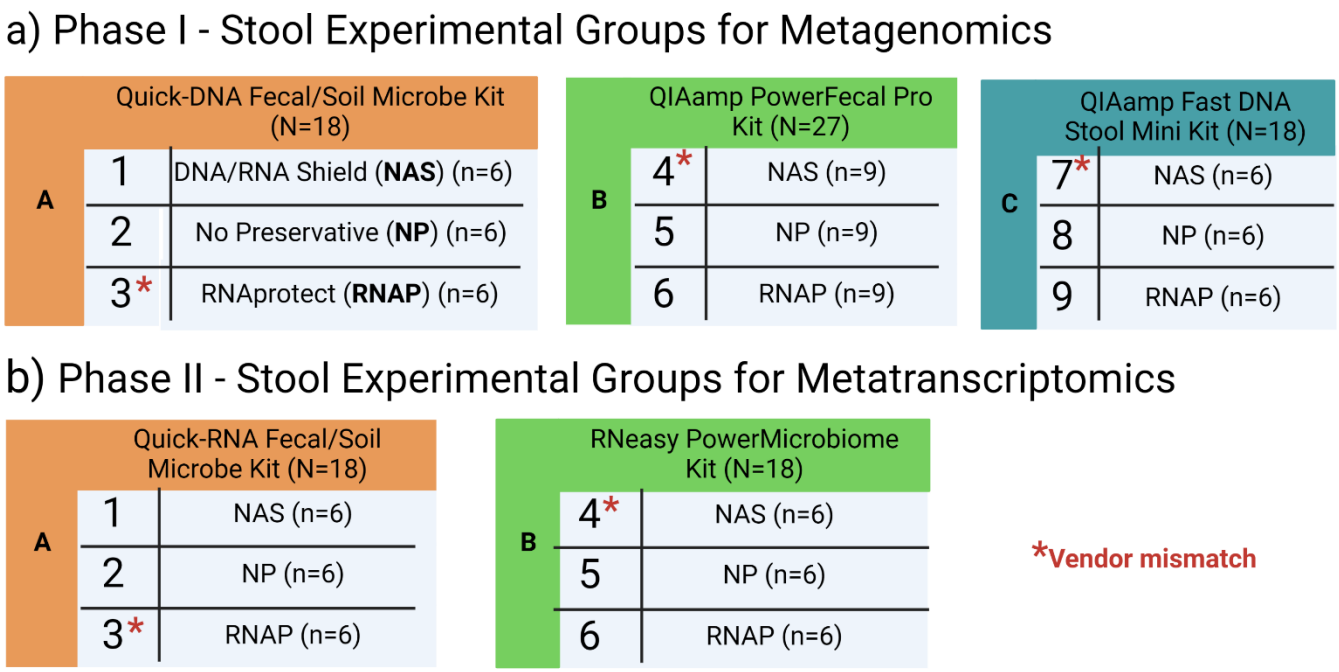


Figure 3. Experimental Groupings Key for Stool Samples According to Nucleic Acid Preservation Method and Extraction Kit. Metagenomic data was generated from 9 experimental groups in Phase I (a) and metatranscriptomic data was generated from 6 experimental groups in Phase II (b). Technical replicates (extraction) for each extraction kit (A-C) and Group (1-9) are indicated with *n* values. Groupings employing a preservation reagent from one vendor and an extraction kit from the other vendor are considered a vendor mismatch and are indicated with an asterisk. Image created in BioRender.

Total RNA isolated from stool samples and positive controls was normalized to 200 ng in 11 µl prior to library preparation. Libraries were prepared from stool and control samples according to the Illumina Stranded Total RNA Prep, Ligation with Ribo-Zero Plus (cat # 20040525, Illumina) protocol, which includes rRNA depletion and reverse transcription into cDNA. In order to prevent elevated levels of index hopping on the NextSeq2000 (Illumina), which uses patterned flow cells, a unique dual indexing scheme (cat # 20025019, Integrated DNA Technologies, Inc. Coralville, IA, US) was used. A total of 22 cDNA libraries were evenly pooled and small fragments were removed using a 0.8x AMPure bead (Beckman Coulter Life Sciences, Indianapolis, IN, US) clean-up. The final pooled library had an average peak size of 307 bp and a concentration of 9.70 ng/µl.

2.4 High-throughput Sequencing of Libraries by Illumina Platforms

The pooled metagenomic library was sequenced on an Illumina NextSeq550 platform (Illumina) at a loading concentration of 1.3 pM, producing an average of 3.2 million paired-end (2x151 bp) reads per sample. Sequence data were uploaded to the Integrated Rapid Infectious Disease Analysis (IRIDA) platform (Matthews et al., 2018) under project #1714 for storage and management. The pooled metatranscriptomic library was sequenced on an Illumina NextSeq2000 platform (Illumina) at a loading concentration of 650 pM, producing an average of 108 million paired-end (2x151 bp) reads per sample. Sequence data were uploaded to IRIDA under project #1999.

2.5 Quality Control of Raw Sequence Reads by KneadData

Raw DNA sequence reads from IRIDA were processed using the KneadData v0.7.6 pipeline from the bioBakery suite of tools (Beghini et al., 2021), which uses Trimmomatic v0.33 (Bolger et al., 2014) to trim and filter raw reads based on quality scores and Bowtie 2 v2.3.5.1 (Langmead and Salzberg, 2012) to filter out reads mapping to the human genome contaminant database included with KneadData.

MultiQC v1.8 (Ewels et al., 2016) was used to generate quality summary reports. Stool samples retaining fewer than 1 million reads following quality control (QC) were excluded from further analysis in order to minimize false-negative assignments due to reduced sequence depth. Detailed analysis code is provided in Appendix 3.

Raw cDNA reads from microbial RNA were also processed using KneadData v0.7.6 (Beghini et al., 2021). Additional bacterial and archaeal 16S and 23S rRNA contaminant databases were downloaded from SortMeRNA (Kopylova et al., 2012). These representative databases contain a subset of sequences from the SILVA rRNA SSU and LSU nonredundant reference databases (Quast et al., 2013). MultiQC (Ewels et al., 2016) was used to generate quality summary reports. Detailed analysis code is provided in Appendix 3.

2.6 Taxonomic Read Mapping by MetaPhlAn

Paired-end metagenomic and metatranscriptomic reads passing QC were concatenated into a single .fastq file per sample and a community taxonomic profile was generated for each sample using the MetaPhlAn v3.0.13 pipeline from bioBakery (Beghini et al., 2021). Briefly, taxonomy is assigned to sequences based on conserved clade-specific markers and relative abundances of each taxa are estimated based on normalized weighted read counts (Segata et al., 2012). Taxonomic classification of sequences from mock community DNA standards (Zymo Research) was used to evaluate optimal parameters for detecting low abundance taxa from metagenomic sequences without introducing false positive assignments (Appendix 1, Figure S1). Extraction blanks were evaluated for contamination based on their community taxonomic profiles and then excluded from further analysis. Optimal parameters for taxonomic profiling of metatranscriptomic sequences were determined in a pilot study and used to evaluate the ZymoBIOMICS™ Fecal Reference with TruMatrix™ Technology (Zymo Research) control sample (Appendix 1, Figure S3) as well as the experimental stool samples. The resulting taxonomic

relative abundance profiles were used to calculate and visualize microbial diversity. Detailed analysis code is provided in Appendix 3.

2.7 Functional Read Mapping by HUMAnN

Community functional profiles were generated using the HUMAnN v3.0.0 pipeline from bioBakery. HUMAnN uses a tiered search strategy to profile microbial metabolic pathways from metagenomic or metatranscriptomic sequence data and provides tables of gene families, pathway abundance, and pathway coverage (Beghini et al., 2021). Detailed analysis code is provided in Appendix 3.

2.8 Statistical Analysis of Microbiome Profiles in RStudio

All analyses detailed below were performed in RStudio v2021.09.0 using R version 4.1.1 (R Core Team, 2021). Detailed code for all analyses are available in Appendix 3.

For beta diversity analyses, feature count tables were processed as follows in RStudio: pseudocounts were imputed to replace zeros in each data set using a Bayesian-multiplicative replacement method via the cmultRepl function included in the zCompositions v1.3.4 R package (Palarea-Albaladejo and Martín-Fernández, 2015). Data were then centre log-ratio (CLR) transformed using the codaSeq.clr function from the CoDaSeq v0.99.6 R package (Gloor and Reid, 2016). Euclidean distances were calculated from the CLR matrix using the vegdist function from the vegan v.2.5.7 R package (Oksanen et al., 2020), and principal components analysis was carried out using the pco function from the ecodist v2.0.7 R package (Goslee and Urban, 2007). Analysis of similarities was conducted using the anosim function from the vegan R package. Indicator species analysis was carried out using the multipatt function from the indicspecies v1.7.12 R package. Confidence intervals for beta diversity PCA coordinates were

calculated and drawn using the stat_ellipse function from the ggplot2 v3.3.5 R package (Wickham, 2016) at a level of 95%.

Multivariable association analysis to detect associations between functional meta-omic features and experimental groupings was carried out using the MaAsLin2 v1.8.0 R package (Mallick et al., 2021). Data were assessed for heteroskedasticity using a Bartlett test of homogeneity of variances and one-way analyses of group means were conducted without assuming equal variances (Welch's ANOVA) via the stats v.4.1.1 R package. All pairwise two-tailed T-tests were performed using Bonferroni correction via the stats v.4.1.1 R package.

2.9 Data Visualization

Results Figures 4, 6-11, 15-19, and 23 were created using the ggplot2 v3.3.5 R package in RStudio. Detailed R code for these figures is provided in A.3.2. Taxonomic heatmaps in Figures 5 and 14 were generated using hclust2 v 1.0.0 (Beghini et al., 2021) during taxonomic profiling (Jupyter Notebook). Functional pathway representations in Figures 13 and 21 were generated using HUMAnN utility scripts. Figures 12 and 20, depicting functional pathway associations were generated by MaAsLin. Figures that have been created or modified using BioRender are indicated as such.

2.10 Comparison with Published Microbiome Data

Published taxonomic profiles from healthy human stool generated with MetaPhlAn were downloaded from (Abu-Ali et al., 2018) and used for comparison against the taxonomic stool profiles generated in this study (Appendix 1, Figure S2). Additionally, paired-end metatranscriptomic sequence data from 6 samples was downloaded from the ZymoBIOMICS™ Fecal Reference Database (https://www.fecalreferencedb.com/). This raw sequence data was processed using KneadData and

MetaPhlAn software as above and the resulting taxonomic profiles were used to evaluate the success of the Fecal Reference extraction carried out in this study (Appendix 1, Figure S3).

2.11 Publication of Sequence Data

Metagenomic and metatranscriptomic sequence data from all stool samples, as well as mock community DNA sequences, have been deposited with links to BioProject accession number PRJNA892575 in the NCBI BioProject Database. https://www.ncbi.nlm.nih.gov/bioproject/PRJNA892575

Individual SRA BioSample accession numbers for the 87 sequenced samples are SAMN31389111–SAMN31389197 (SRA #: SRR21986380 -SRR21986444, Phase I; SRR22116534 - SRR22116555, Phase II).

3 RESULTS

3.1 Nucleic Acid Yield Varies by Experimental Approach

Nucleic acids were isolated from stool sample replicates according to the experimental groupings in Figure 3. Reagents for nucleic acid stabilization and extraction were sourced from two vendors, Qiagen and Zymo Research. For the purposes of this study, we define a vendor mismatch as any experimental grouping that uses a stabilizing reagent from one vendor and an extraction protocol from another (Figure 3). Following nucleic acid extraction, DNA or RNA concentrations in each sample were quantified using the Qubit (Thermo Fisher Scientific). Among the three DNA extraction kits evaluated, Kit A (Quick-DNA Fecal/Soil Microbe Kit, Zymo Research) yielded the highest consistent DNA concentrations across all preservant conditions: replicates from groups 1–3 all yielded DNA concentrations above 25 ng/µl (Figure 4a). Conversely, Kit B (QIAamp PowerFecal Pro Kit, Qiagen) yielded the most variable DNA concentrations, ranging from 0.1 (NP, Group 5) to over 200 ng/µl (NP, Group 5). In particular, DNA yield from NP stool processed with Kit B (Group 5) was statistically distinct from that of stool processed with the same kit and preserved in NAS (Group 4, Figure 4a, $p < 0.05$), and also from all groups processed with Kit C (Groups 7–9, $p < 0.05$). Kit C (QIAamp Fast DNA Stool Mini Kit, Qiagen) yielded consistent DNA concentrations across all preservant conditions (Groups 7–9), although the yield was lower compared to samples processed with Kit A (all Kit C samples had DNA concentrations below 15 ng/µl).

Regarding RNA extraction, stool processed with Kit B (RNeasy PowerMicrobiomeKit, Qiagen) and preserved in NP or RNAP (Groups 5 and 6) had significantly higher RNA yield compared to all other groupings (Figure 4b, $p < 0.05$). Stool from groups 1–3 processed with Kit A (Quick-RNA Fecal/Soil Microbe Kit, Zymo Research), did not have significant differences in terms of RNA yield. The data in Figure 4 illustrates how nucleic acid yield can be impacted by the combination of preservation and

extraction approach. Interestingly, stool that was preserved in NAS from Zymo Research and extracted with a Qiagen bead-beating kit yielded lower nucleic acid concentrations compared to other groupings from the same kit (DNA: Group 4 vs. 5, $p < 0.05$; RNA: Group 4 vs. 5 and 6, $p < 0.05$), possibly indicating a detrimental vendor mismatch. However, this reduced yield did not noticeably affect downstream taxonomic profiles.

The following sections will present results from Phase I and Phase II of metagenomic and metatranscriptomic data analysis, respectively, in order to address the first research hypothesis. In the final Results section, metagenomic and metatranscriptomic data are directly compared in order to address the second hypothesis.

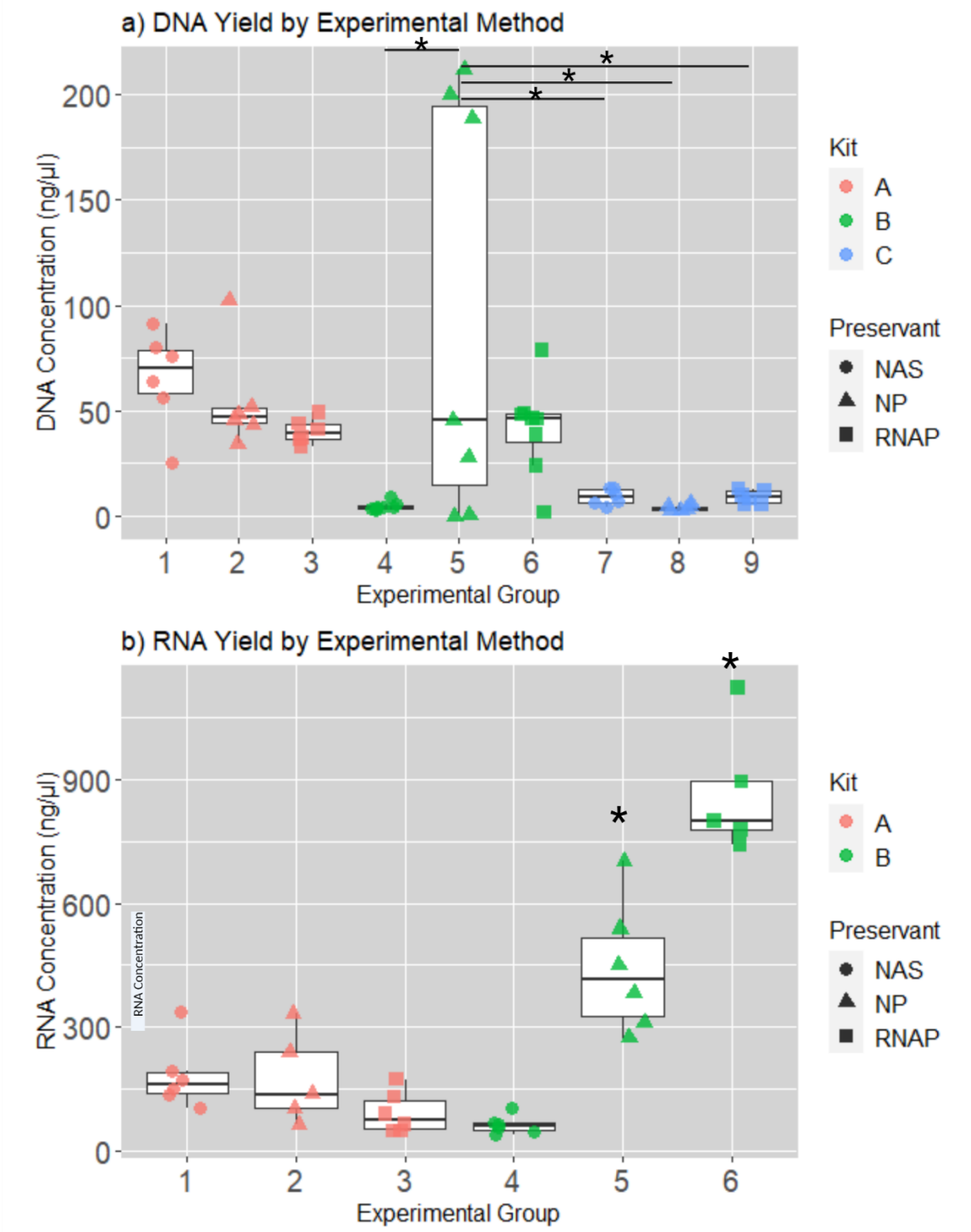


Figure 4 . Nucleic Acid Yield from Stool by Experimental Method. Nucleic acid concentration (ng/µl) by experimental group for Phase I DNA (a) and Phase II RNA (b). DNA concentrations from Group 5 samples are distinct from several other groups. RNA concentrations from Group 5 and 6 samples are distinct from all other groups. *$p < 0.05$ by pairwise two-tailed T-tests with Bonferroni correction. Image modified with BioRender.

3.2 PHASE I Results - Metagenomics

3.2.1 Taxonomic Profiles from DNA Cluster According to Extraction Method and Preservant

For Phases I and II of this study, the bioBakery suite of analysis tools (Beghini et al., 2021) was chosen for several reasons: firstly, these tools are compatible with compositional data structures and analysis requirements (Gloor et al., 2017); as well, bioBakery tools were designed for human microbiome research and have been used with DNA and RNA sequence data in many prominent microbiome studies within the past decade (Human Microbiome Project Consortium, 2012b; Franzosa et al., 2014; Gevers et al., 2014; Abu-Ali et al., 2018; Mehta et al., 2018; Schirmer et al., 2018; Lloyd-Price et al., 2019; De Filippis et al., 2021); furthermore, the software and supporting documentation for bioBakery tools are regularly maintained and updated according to new discoveries in the field (Beghini et al., 2021); and, finally, the associated ChocoPhlAn clade marker gene database enables highly accurate and stringent taxonomic and functional profiling.

Following quality trimming and contaminant filtering of metagenomic reads with KneadData, taxonomic relative abundance profiles produced by MetaPhlAn were evaluated for differences at the phylum and species level. Clustering of samples according to average linkage (computed with Euclidean distances) of species-level taxonomic profiles is illustrated in Figure 5. The heatmap shows approximate clustering of stool samples based on the kit used to extract the metagenomic DNA. Imperfect sub-groupings are also observed according to stool preservative, where NP samples cluster together, and NAS or RNAP samples form separate, mixed clusters. Of note is one stool sample (CQFMRNAP12), which is clustered with several blank samples and appears to contain no taxonomic profile. Conversely, one blank sample (CQFMB2) is clustered with several stool samples and contains a rich taxonomic profile. The two samples in question are labelled in orange in Figure 5. Due to the shorthand labelling convention used for sample tubes in the laboratory, it is highly likely that these two samples (“C12” and “C2”, respectively)

were mistakenly swapped or mislabeled. Therefore, both of these samples were excluded from further analysis. There were 5 remaining stool replicates and 1 remaining blank sample in the associated experimental group (Group 9).

Metagenomic extraction blanks were found to contain negligible contamination: one blank sample, processed with Kit C, was assigned a taxonomic profile consisting of 3 bacterial species. However, the total sequence count for this sample (~131 k reads) was far below the analysis cutoff of 1 million and as such, this contamination was considered to be minor and acceptable. The taxonomic profiles of MOCK samples (MOCK1 and MOCKLOG2), generated from microbial community DNA standards (Zymo Research), were used to assess optimal parameters for accurate taxonomic profiling with MetaPhlAn (Appendix 1, Figure S1). The default parameters set by MetaPhlAn were determined to provide accurate taxonomic profiles for the MOCK samples with only a single false positive assignment in the balanced community at the species level (species-level FDR = 0.048; Appendix 1, Figure S1). The taxonomic assignment of this single taxon was accurate at the genus level, and as such this was considered to be acceptable (genus-level FDR = 0).

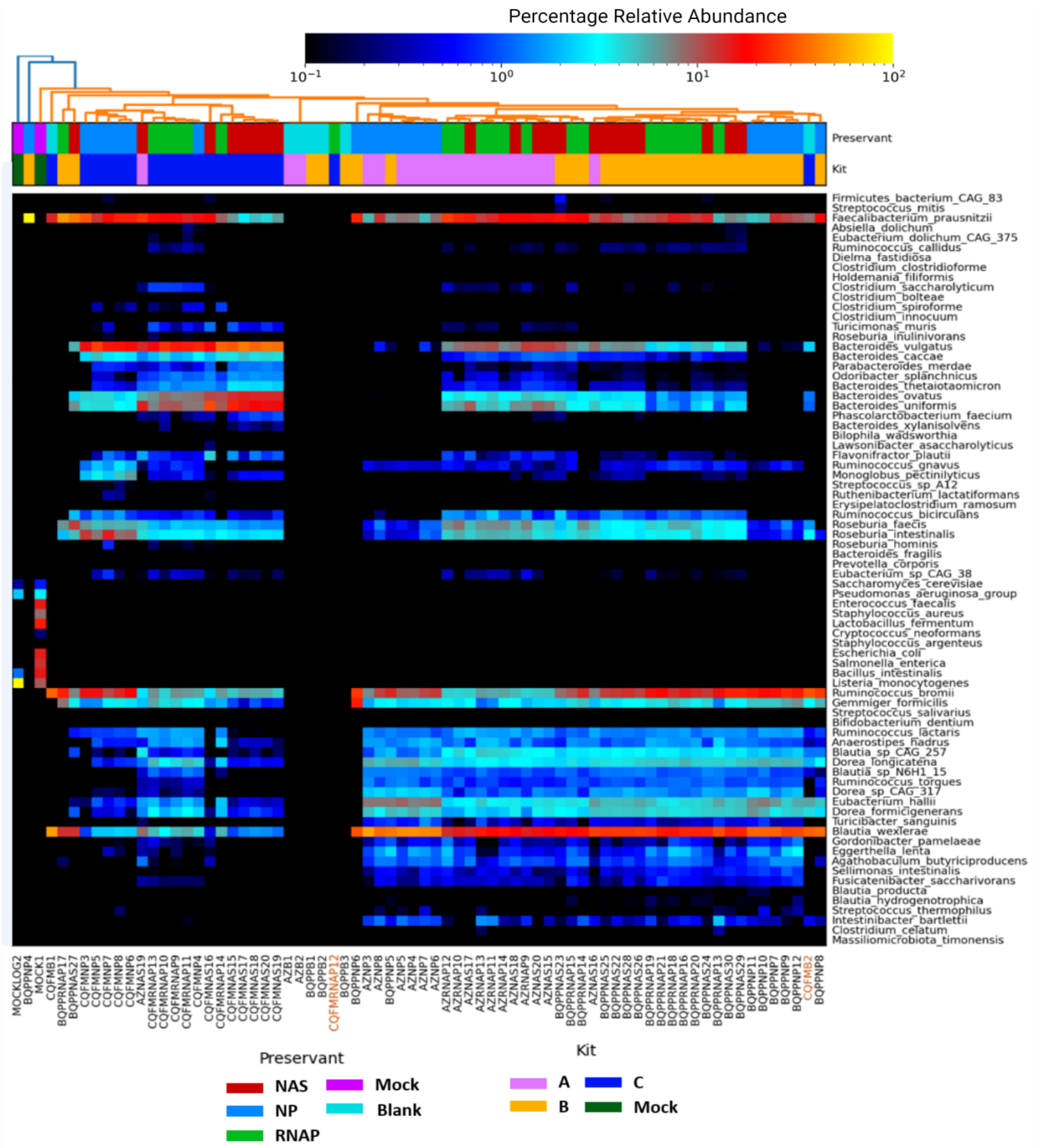


Figure 5. Species Abundance Heatmap of Metagenomic Data. Sample names (bottom) and metadata (top) are arranged along the x-axis, which is clustered according to average linkage with Euclidean distances. The names of two samples that were removed from analysis are indicated in orange. Assigned species names are arranged along the y-axis. Heatmap colouring corresponds to relative abundance.

3.2.2 Recovery of Major GI Phylum Bacteroidetes Depends on Experimental Approach

Major differences between experimental approaches can be observed at the phylum level in taxonomic profiles of metagenomic data from stool samples (Figure 6). Recovery of the Gram-negative phylum Bacteroidetes is lower than expected (reported healthy relative abundance is between 10-80%; Abu-Ali et al., 2018) when stool samples are not preserved (NP) and a bead-beating kit is used (Groups 2 and 5; mean relative abundance = 0.34% and 0.15%, respectively) but not when a thermal lysis kit is used (Group 8; mean relative abundance = 35.7%, Figure 6). Conversely, Bacteroidetes are relatively more abundant (mean = 51.2% vs. 17.2%), at the expense of Firmicutes and Actinobacteria, in samples processed with the thermal lysis approach (Kit C) regardless of preservant (Groups 7–9) as compared to preserved samples processed with Kits A or B (Groups 1, 3, 4, and 6). This is likely due to the thermal lysis approach being less vigorous and lysing fewer Gram-positive cells, compared to bead-beating, which has been shown to be critical for the detection of Gram-positive organisms in stool samples (Santiago et al., 2014). Although, it is worth noting that Gram-positive Firmicutes were still detected with the thermal lysis method at relative abundances between 23.9–64.0%, which is within the expected range of 20–90% (Abu-Ali et al., 2018; Appendix 1, Figure S2).

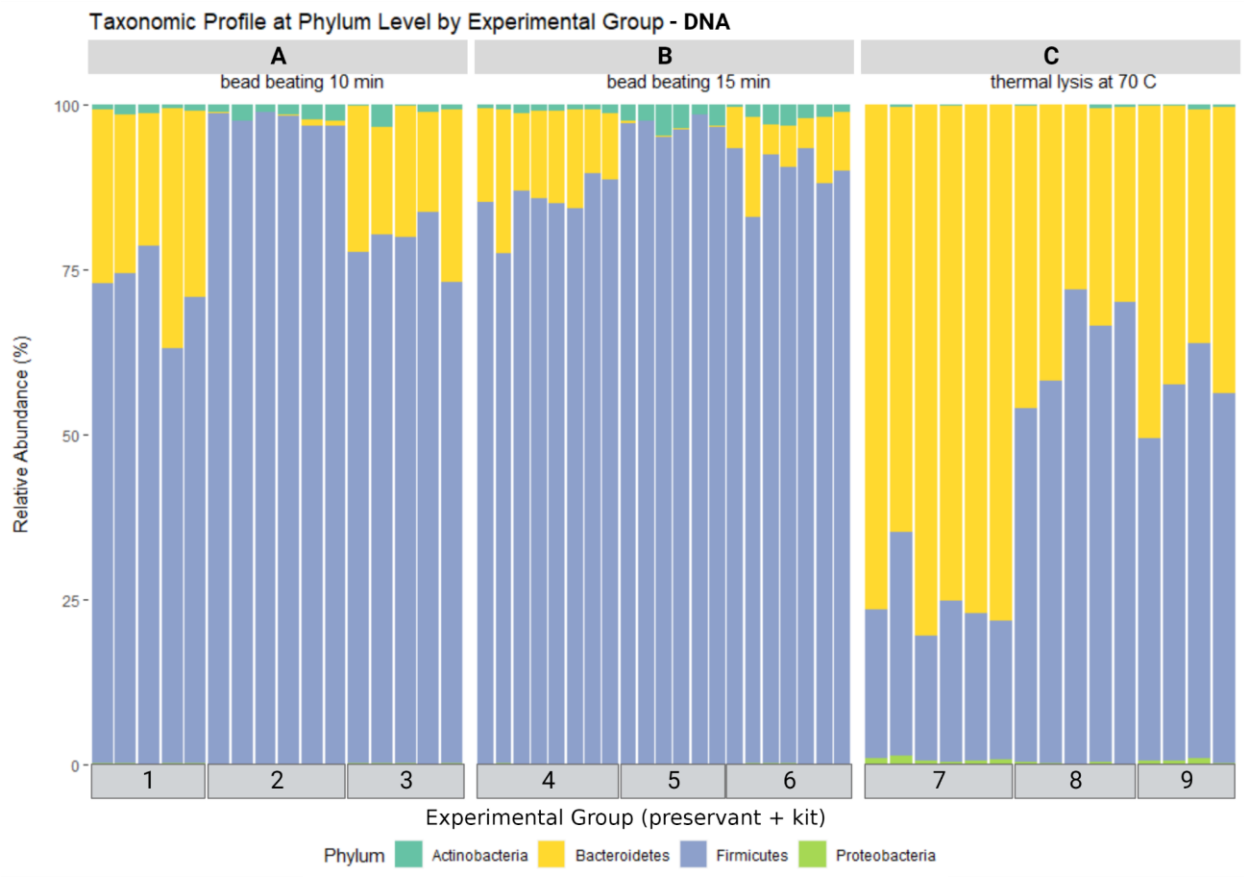


Figure 6. Phylum-level Taxonomic Profiles of Metagenomic Data from MetaPhlAn. Primary cell lysis approaches for each extraction kit (A-C) are annotated on the plot. Image modified with BioRender.

3.2.3 Firmicutes:Bacteroidetes Ratio is Significantly Impacted by Methodology

Phylum-level differences can be quantified by measuring the F/B ratio of each sample (Magne et al., 2020). For metagenomic data, Figure 7 illustrates how inflated the F/B ratio becomes when Bacteroidetes are detected in low abundance (Groups 2 and 5, mean = 6001 and 979). Based on the available literature (Magne et al., 2020), as well as our own comparison of public data from over 900 healthy human fecal metagenomes (Abu-Ali et al., 2018; Appendix 1, Figure S2), a reasonable expected range for F/B ratio in healthy adult North-American subjects is between 0.2–10. Although we cannot know the true F/B ratio of the underlying bacterial community in our stool sample, it is clear that Groups 2 and 5 are well outside of this reasonable range and are unlikely to be representative of the true community.

3.2.4 Alpha Diversity of Metagenomic Samples is Significantly Impacted by Method

Another standard metric in microbiome profiling is the alpha diversity of each sample. We found that the inflated F/B ratio previously seen in NP stool processed with bead-beating kits (Groups 2 and 5) corresponded to a significantly lower Shannon Index (mean = 2.14, Group 2; 2.29, Group 5) compared to NAS (mean = 2.78, Group 1; 2.74, Group 4) or RNAP (mean = 2.83, Group 3; 2.77, Group 6) stool processed with the same kit ($p < 0.01$, Figure 8). Additionally, a significantly reduced Shannon Index was observed in NAS stool processed with Kit C ($p < 0.01$, Group 7 vs Groups 1, 3, 4, 6, 8, and 9). This is likely due to the high detected abundance of Bacteroidetes in this group (Figure 7).

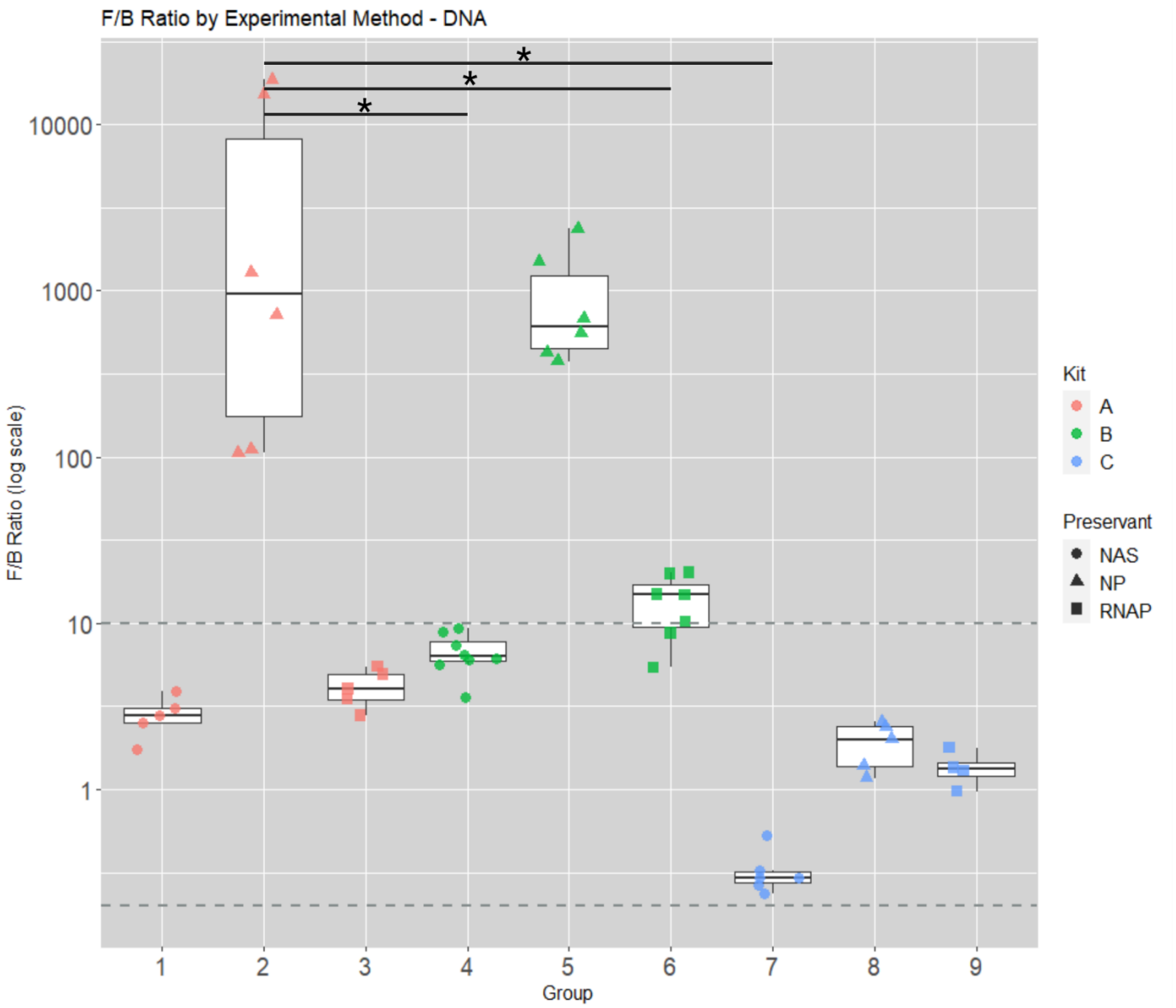


Figure 7. Firmicutes:Bacteroidetes Ratio from Metagenomic Data based on MetaPhlAn Taxonomic Profiles. Dashed lines indicate expected range (0.2–10). * $p < 0.05$ by pairwise T-tests with Bonferroni correction. Image modified with BioRender.

In order to investigate the factors contributing to this reduced Shannon Index for certain groups, we next investigated two components of alpha diversity: richness and evenness. Figure 9 illustrates that the underlying components of diversity were not uniformly impacted in Groups 2, 5, and 7, despite these groups having similar mean Shannon Indices (2.14, 2.29, and 2.19, respectively, Figure 8). NP stool processed with a bead-beating kit (Groups 2 and 5) shows reduced species richness (mean = 30.5, Group 2; 32.5, Group 5) and evenness (mean = 0.63, Group 2; 0.66, Group 5) when compared to NAS (mean richness = 43.5, mean evenness = 0.73) or RNAP (mean richness = 44.1, mean evenness = 0.74) stool extracted from the same kit. In contrast, NAS stool processed with the thermal lysis kit (Kit C, Group 7) has reduced evenness only (mean evenness = 0.61 vs. 0.70 and 0.73), where the richness is comparable to NP and RNAP stool from the same kit (mean richness = 37.5 vs. 38.4 and 46.8). It is not clear why Bacteroidetes is so grossly overrepresented in Group 7 alone. However, a possible explanation could be a vendor mismatch wherein incompatible reagents lead to incomplete lysis of Gram-positive Firmicutes. Due to their rigid cell wall structure, incomplete lysis of Firmicutes leading to a reduced relative abundance is a common phenomenon in microbiome research (Santiago et al., 2014; Lim et al., 2018).

A maximum richness of 51 species was observed for metagenomic data. Under the assumption of a low false discovery rate (species-level FDR < 0.05; Appendix 1, Figure S1), we can determine that NP stool processed with a bead-beating kit captures, on average, only ~60% of this maximum diversity (mean richness: 30.5, Group 2; 32.5, Group 5), even when the average read count is comparable or higher than preserved stool extracted with the same kit (Table 2).

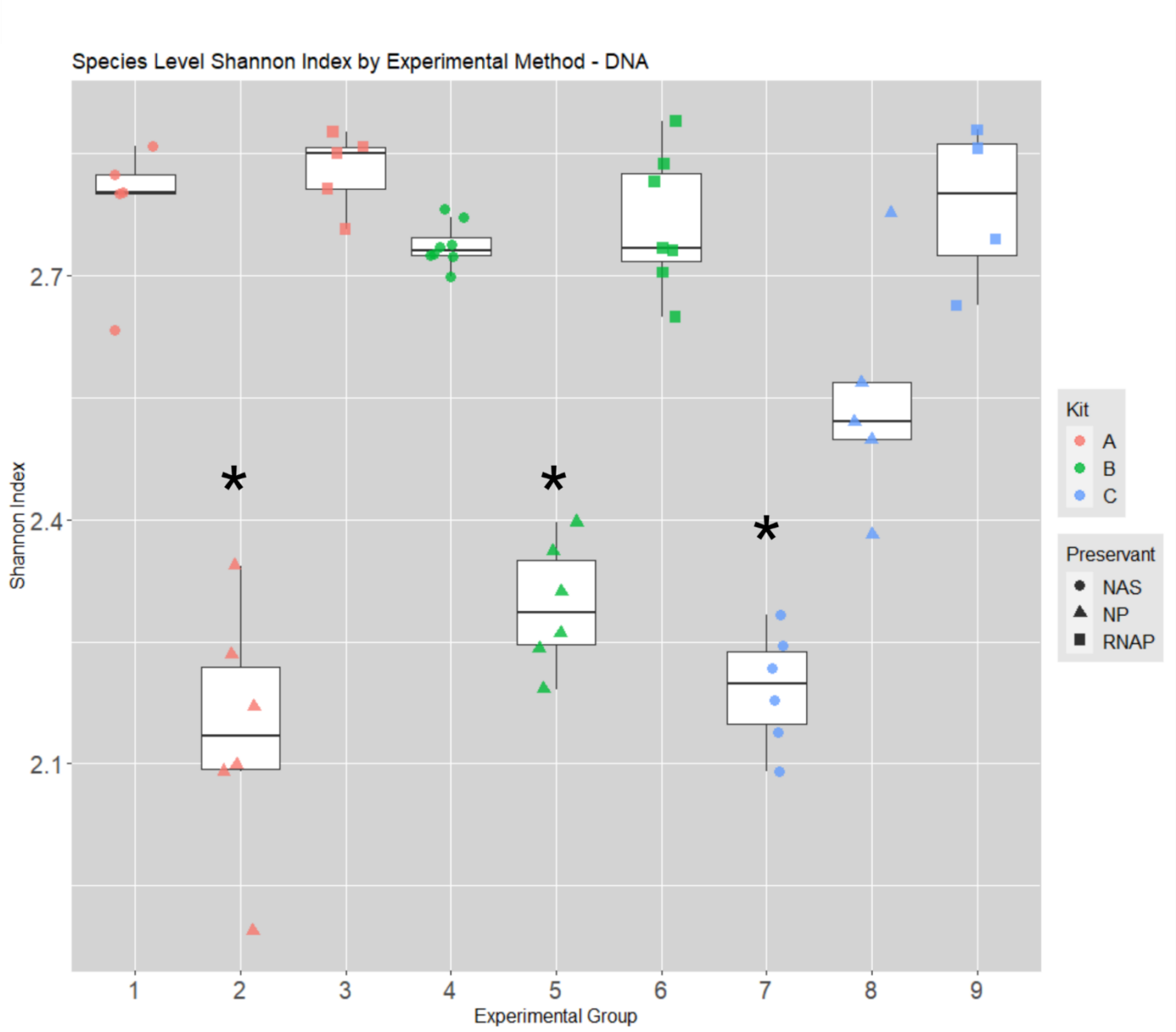


Figure 8. Species-level Shannon Index of Metagenomic Data based on MetaPhlAn Taxonomic Profiles. Overall alpha diversity, quantified by Shannon Index, is significantly reduced in experimental groups 2, 5, and 7. *$p < 0.01$ by pairwise two-tailed T-tests with Bonferroni correction. Image modified with BioRender.

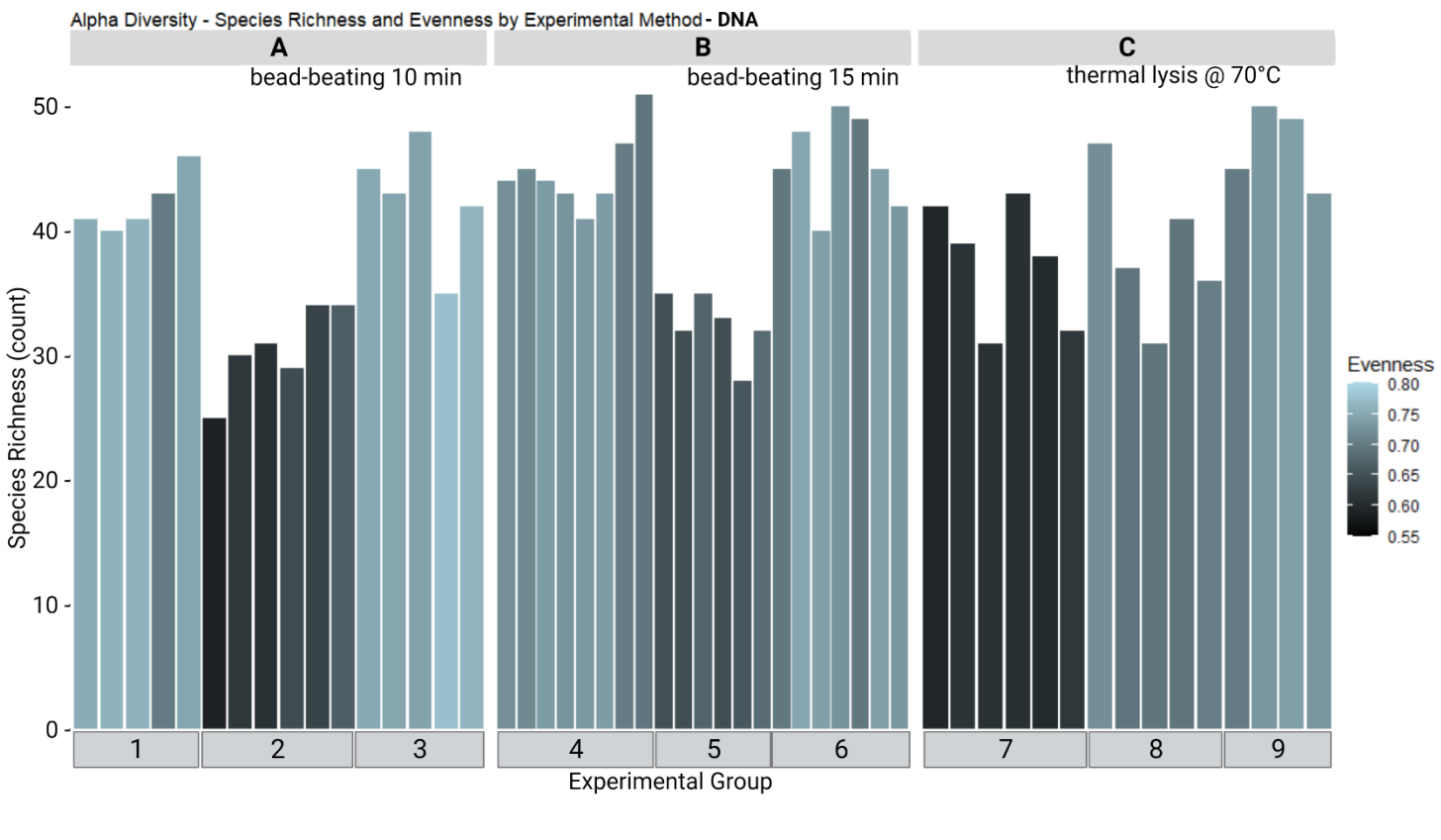


Figure 9. Alpha Diversity Components of Metagenomic Data based on MetaPhlAn Taxonomic Profiles. Microbiomes of unpreserved stool show reduced species richness and evenness when a bead-beating kit is used to extract DNA (Groups 2 and 5). Reduced evenness in Group 7 (preserved in DNA/RNA Shield) may indicate reagent incompatibility. Image modified with BioRender.

Table 2. Comparison of Average DNA concentration (ng/µl), read count following quality control, and species richness among Phase I Metagenomic Experimental Groups.

Metagenomics (Phase I)					
Experimental Group	**Preservant**	**Kit**	**[DNA] (ng/µl)**	**Read Count**	**Species Richness**
1	DNA/RNA Shield **(NAS)**	Quick-DNA Fecal/Soil Microbe **(A)**	**62.4**	**2,796,175**	**42.2**
2	None **(NP)**	Quick-DNA Fecal/Soil Microbe **(A)**	**54.3**	**2,513,694**	**30.5**
3	RNAprotect **(RNAP)**	Quick-DNA Fecal/Soil Microbe **(A)**	**38.3**	**2,678,613**	**42.6**
4	DNA/RNA Shield **(NAS)**	QIAamp PowerFecal Pro **(B)**	**4.7**	**3,471,792**	**44.8**
5	None **(NP)**	QIAamp PowerFecal Pro **(B)**	**142.4**	**3,404,240**	**32.5**
6	RNAprotect **(RNAP)**	QIAamp PowerFecal Pro **(B)**	**47.2**	**3,263,661**	**45.6**
7	DNA/RNA Shield **(NAS)**	QIAamp Fast DNA Stool Mini **(C)**	**9.1**	**2,961,825**	**37.5**
8	None **(NP)**	QIAamp Fast DNA Stool Mini **(C)**	**4.0**	**3,357,351**	**38.4**
9	RNAprotect **(RNAP)**	QIAamp Fast DNA Stool Mini **(C)**	**9.2**	**2,658,725**	**46.8**

3.2.5 Beta Diversity Reveals Differences due to Preservation and Lysis Approach

Next, we characterized microbiome profiles using beta diversity, also known as between-sample diversity. In order to satisfy compositional data analysis requirements (Gloor et al., 2017), Figure 10 was generated using CLR-transformed species counts to calculate Euclidean pairwise distances, which are projected onto a 2D space using principal components analysis (PCA). The first two principal components in this plot account for roughly 86% of the variation between all samples, indicating that the majority of variation in this dataset is attributable to the principal components plotted. An analysis of similarities (ANOSIM) of the taxonomic relative abundance profiles determined that the experimental groupings (1–9) are statistically distinct in terms of their taxonomic composition ($R = 0.5$, $p = 0.001$). In other words, the dissimilarity between groups is greater than the dissimilarity within groups.

Overall, the stool samples cluster on the PCA plot according to whether or not they were preserved (NAS or RNAP), and also based on the type of kit that was used to extract DNA (bead-beating; kits A and B, thermal lysis; Kit C, Figure 10). This is consistent with previous results; however, although the NP-thermal lysis samples (Group 8) had similar alpha diversity profiles to the preserved bead-beating samples (NAS or RNAP; Groups 1, 3, 4, and 6) in terms of species richness and evenness, their taxonomic profile is in fact unique. There is an exception of one NP sample (CQFMNP4) clustering closer to the preserved thermal-lysis group. Indicator species analysis revealed that this sample is more similar to RNAP samples processed with Kit C (Group 9) due to increased *Bacteroides spp.* and *Odoribacter splanchnicus*, and decreased *Roseburia intestinalis* and *Monoglobus pectinilyticus*. In addition, 5 bacterial species were strongly (Indicator Value Index > 0.75) associated with the NP-thermal lysis samples (Group 8, $p < 0.05$), 7 species were strongly associated with the preserved-thermal lysis samples (NAS or RNAP; Groups 7 and 9, $p < 0.05$), and 11 were strongly associated with the bead-beating clusters regardless of preservant (Groups 1–6, $p < 0.05$)(Appendix 1, Table S4). Therefore, most of the variation (85.7%) among taxonomic

profiles for all stool samples can be attributed to changes in the observed relative abundance of several taxa that are associated with both the preservation method used to store those samples, as well as the approach to cell lysis used in the respective extraction kits.

3.2.6 Gene Family Functional Profiles from DNA are Impacted by Preservant and Extraction Method

In order to evaluate changes in metagenomic functional potential (*ie.*, microbial gene detection), beta diversity analysis was conducted on gene family relative abundance tables generated by HUMAnN. A PCA plot of the results (Figure 11) shows that NP stool processed with a bead-beating kit (Groups 2 and 5) have distinct gene family profiles compared to all other experimental groupings, consistent with the taxonomic results. Additionally, NAS stool processed with Kit C (Group 7) produces a unique profile of gene families. Therefore, in addition to taxonomic profiles, DNA functional potential information, captured through gene family profiles, is also significantly (ANOSIM $R = 0.84$, $p = 0.001$) impacted by experimental methodology.

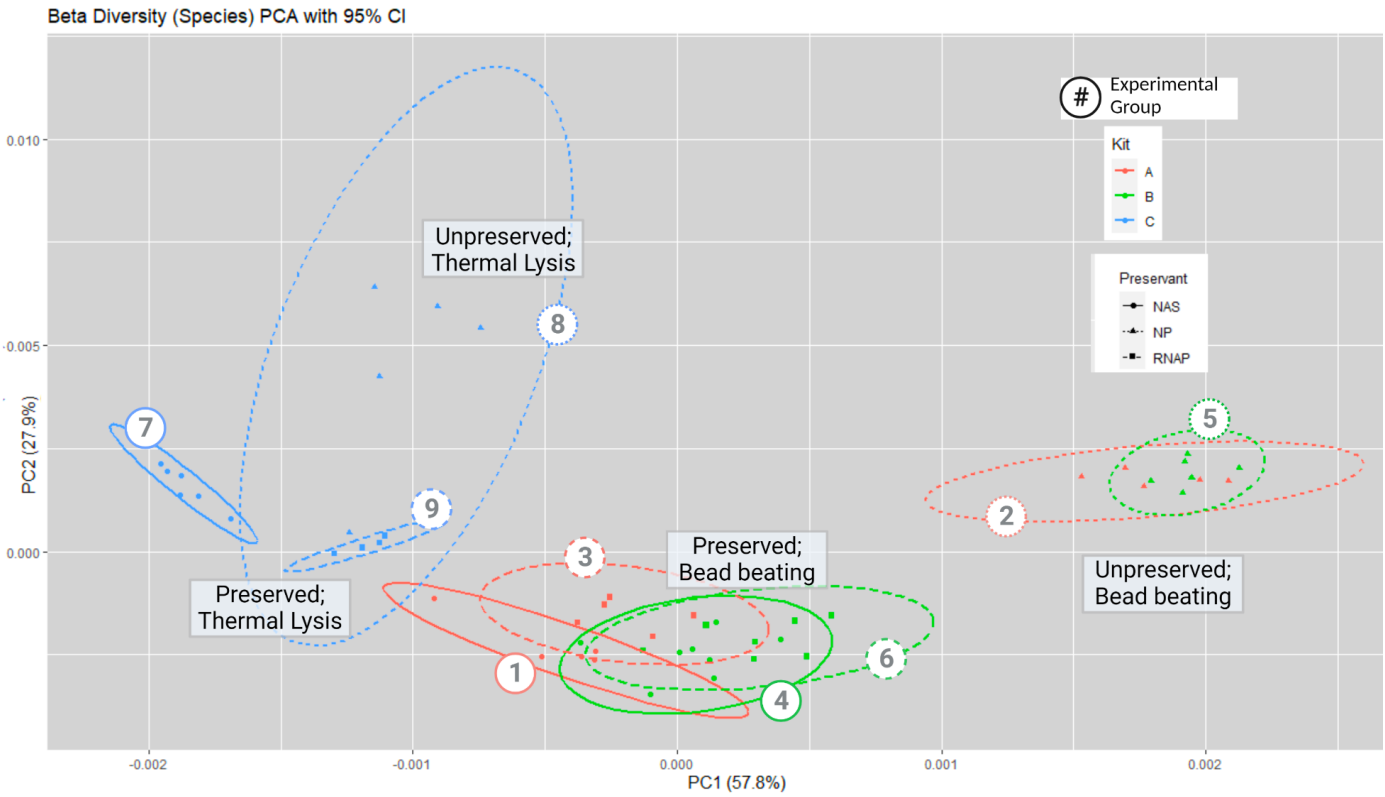


Figure 10. Phase I Species-Level Beta Diversity PCA Plot Principal component analysis (PCA) of CLR-transformed taxonomic stool microbiome profiles at the species level reveals that samples cluster according to i) whether or not samples are preserved and ii) the lysis approach of the extraction kit used. 95% confidence intervals are indicated by ellipses. Groups 1–9 are statistically distinct in terms of taxonomic composition (ANOSIM R = 0.5, p = 0.001). Image modified with BioRender.

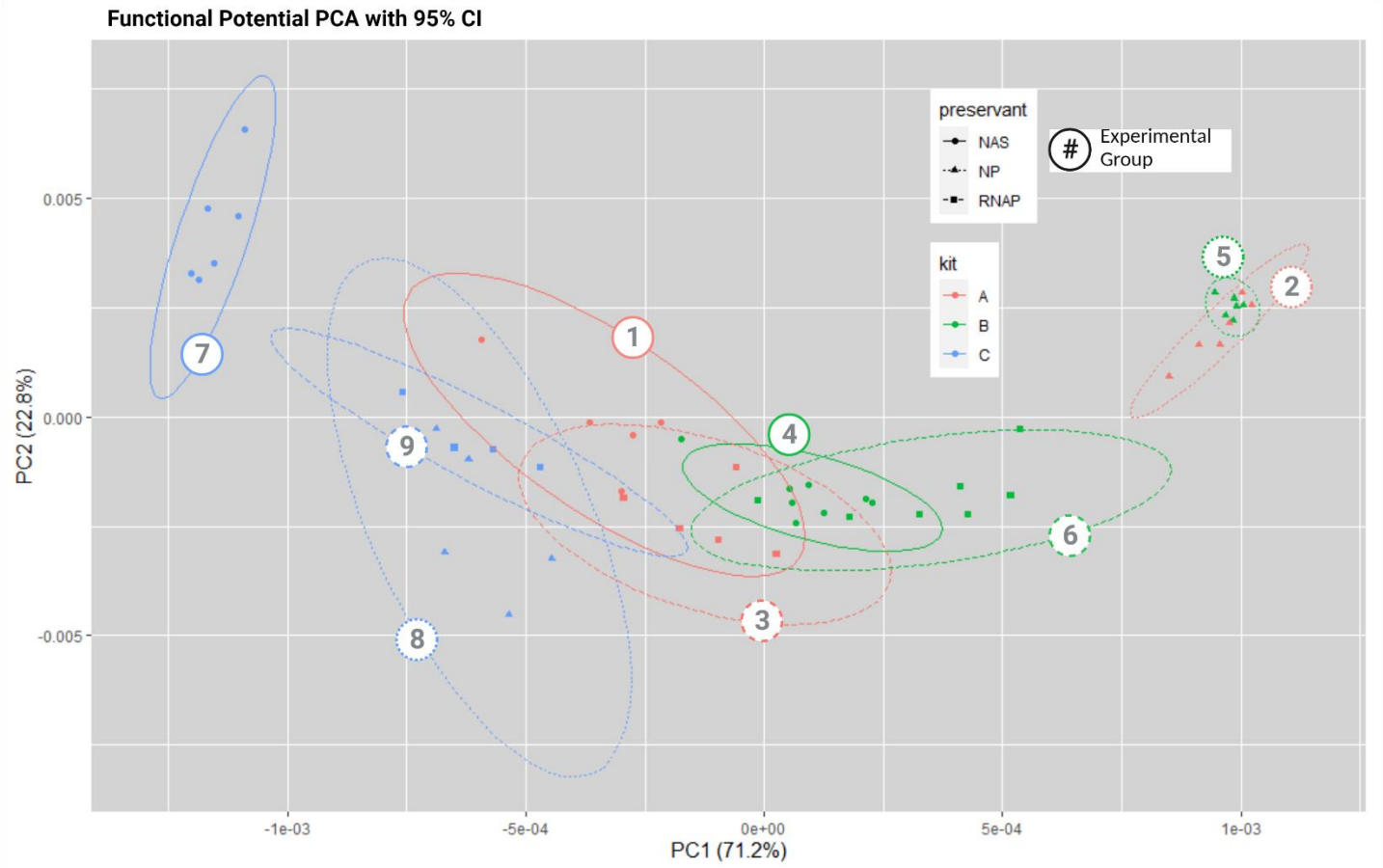


Figure 11. Phase I Functional Beta Diversity PCA Plot. Principal component analysis (PCA) of CLR-transformed functional stool microbiome gene family profiles reveals that Groups 1–9 are statistically dissimilar in terms of gene family composition (ANOSIM R = 0.84, p = 0.001). 95% confidence intervals are indicated by ellipses. Image modified with BioRender.

3.2.7 Pathway Abundances Differ by Method According to Taxonomic Differences.

In order to further investigate GI microbiome functional potential from metagenomic data, a multivariable association analysis was conducted (using MaAsLin) on the unstratified metabolic pathway relative abundance table generated by HUMAnN from metagenomic data. The top 50 features identified to be significantly associated ($p < 0.01$) with experimental grouping, named according to MetaCyc pathway definitions, are shown in Figure 12. Investigation of the features with significant associations to the experimental group reveals that the between-group differences are likely related to differential capture of particular bacterial phyla. For example, Figure 13a shows that pathway with the strongest association, the L-Histidine Degradation III pathway, was not detected in NP stool processed with bead-beating kits (Groups 2 and 5), which are the groupings with inflated F/B ratios and reduced alpha diversity due to low detection of Bacteroidetes (Sections 3.2.2–3.2.4 above). One bacterial species belonging to this phylum, *Bacteroides ovatus*, was identified as a contributor to this pathway (Figure 13a). Although *B. ovatus* was the only species identified contributing to this pathway, it is likely that the "unclassified" pathway contributions are from other species in the Bacteroidetes phylum based on their inability to be detected with experimental methods that are biased against this phylum. Thus, the observation of this metabolic pathway is highly associated with experimental grouping. Likewise, investigation of a metabolic pathway that is broadly conserved across bacterial phyla, Peptidoglycan Biosynthesis I (Figure 13b), shows that while this pathway is captured in all experimental groupings, diversity in terms of species-specific contributions is differentially captured; of 18 total contributing species identified, only 1-3 contributing species were identified in Groups 2, 7, 8, and 9. Therefore, by failing to capture a rich and balanced taxonomic community, metagenomic profiles generated from unpreserved stool, or stool processed with a thermal lysis kit, are additionally missing key components of community functional potential.

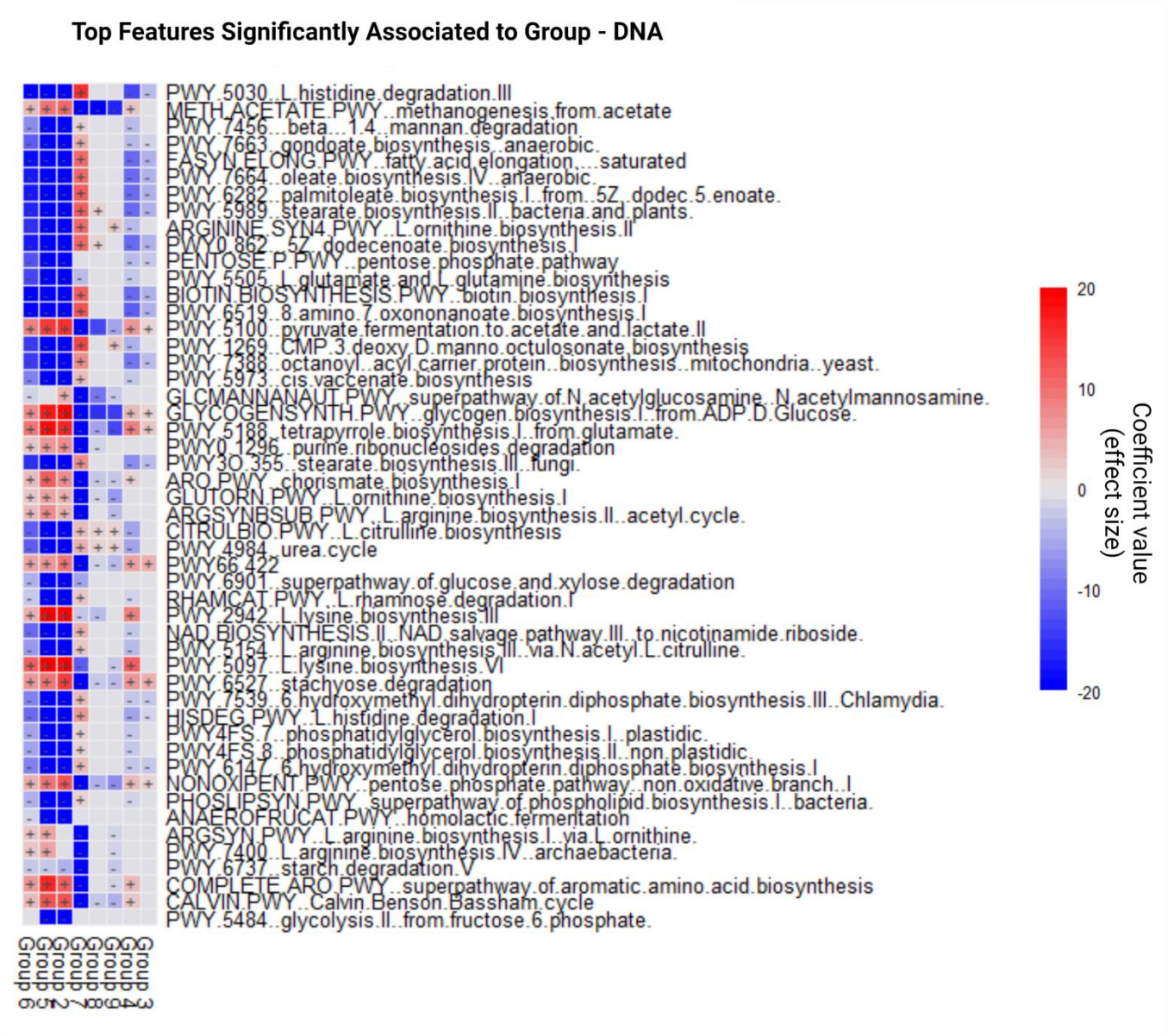


Figure 12. Top 50 Pathways Associated with Phase I Experimental Group identified by MaAsLin. Metabolic pathways, according to MetaCyc pathway definitions, with a significant association to the DNA experimental group, identified by linear modelling with MaAslin2. Group 1 is used as a comparator for pathway abundance. $p < 0.01$ for one or more experimental groupings per pathway.

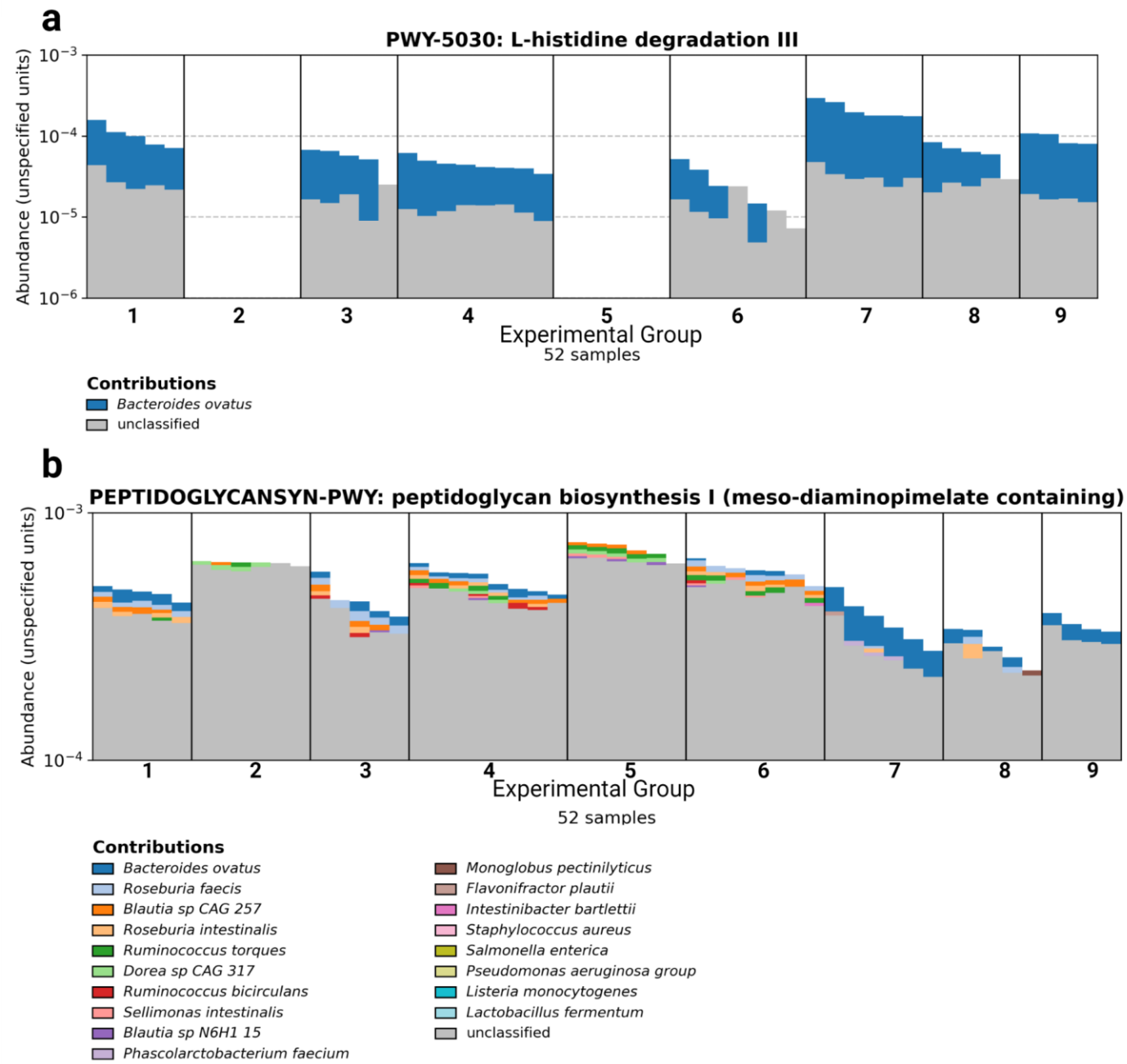


Figure 13. Observation of Functional Potential via HUMAnN Depends on Taxonomic Capture. Metabolic pathways with associations to experimental grouping correlate with differential capture of particular bacterial phyla. Samples from Groups 2 and 5 contain a low abundance of Bacteroidetes and are missing functional potential contributed by organisms in this phylum, such as the L-histidine degradation III pathway (a). Broadly conserved pathways, such as the Peptidoglycan biosynthesis I pathway (b), are captured in all experimental groupings. However, species-specific contributions are not uniformly represented. Image created with HUMAnN and modified with BioRender.

3.3 PHASE II Results - Metatranscriptomics

3.3.1 Taxonomic Profiles from RNA Cluster by Preservant and Extraction Kit

Metatranscriptomic data from six experimental groupings (Figure 3) was generated based on RNA extracted from stool aliquots originating from the same stool sample as the aliquots used to extract DNA in Phase I (Figure 2). In this phase of the study, rRNA reads were removed using KneadData software and the resulting transcriptional profiles were evaluated using the same community outcomes from Phase I. The microbiome profiles discussed Phase II reflect observed microbial gene expression under differing experimental conditions.

Taxonomic relative abundance profiles produced by MetaPhlAn were evaluated for differences at the phylum and species level. Clustering of samples according to average linkage (computed with Euclidean distances) of species-level taxonomic profiles is illustrated in Figure 14. The heatmap shows clustering of stool samples primarily based on the preservant condition (NAS, NP, or RNAP), except for one sample (ZRNAP1), which was the only replicate in its group (Group 3) with an assigned species profile and thus clustered separately. Within each preservant cluster there is clear separation between samples processed with Kit A and those processed with Kit B (Figure 14), except for a single sample (ZNP3), which clustered more closely with samples that are preserved similarly (NP), but processed with a different kit (Kit B). There were 5 bacterial species that were uniquely identified in this single metatranscriptomic sample: *Absiella dolichum, Eubacterium dolichum* CAG 375, *Anaerostipes hadrus, Ruminococcus lactaris,* and *Clostridium saccharolyticum*. Since these taxa were detected in metagenomic data from Phase I (Figure 5), this sample was included in subsequent analyses.

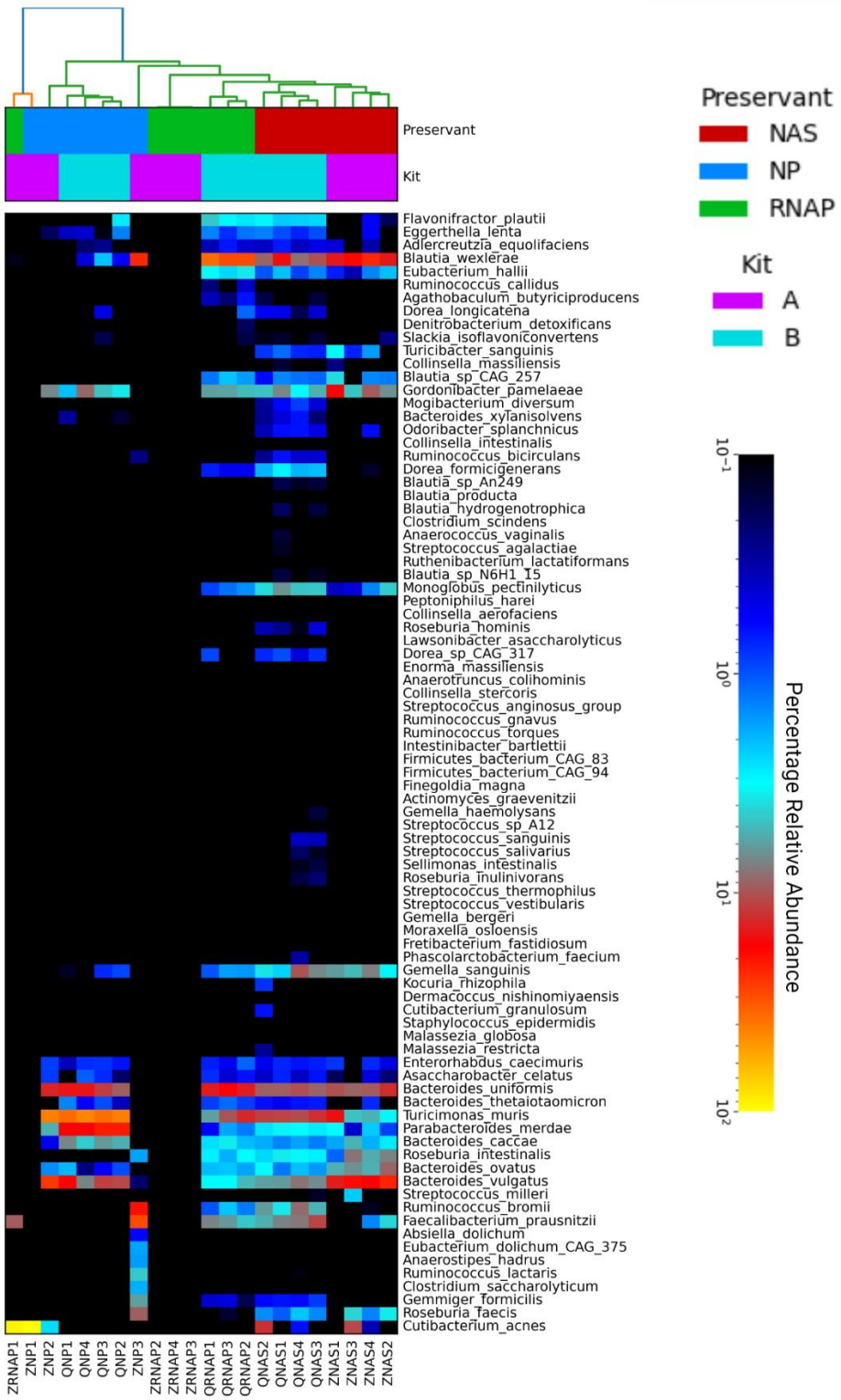


Figure 14. Species Abundance Heatmap of Metatranscriptomic Data based on MetaPhlAn Taxonomic Profiles. Sample names (bottom) and metadata (top) are arranged along the x-axis, which is clustered according to average linkage with Euclidean distances. Assigned species names are arranged along the y-axis. Heatmap colouring corresponds to relative abundance.

3.3.2 Phylum-level Profiles from RNA are Impacted by Experimental Approach and Sequencing Depth.

As a result of the low Gram-negative recovery presented in Phase I (section 3.2.2), the bead-beating approach was revised for RNA extraction kits (*e.g.*, 3 minutes of bead-beating vs. 15 minutes for Phase I or II Kit B, see Methods). This shorter bead-beating duration is likely responsible for an expansion of Gram-negative phyla at the expense of Gram-positive Firmicutes in NP stool (Group 5, Figure 15), contrary to what was observed in the metagenomic data. Therefore, it is important to acknowledge that taxonomic differences between metagenomic and metatranscriptomic data are not solely due to differences in functional potential vs. activity, but rather, are also impacted by alterations to methodology affecting taxonomic relative abundance. As such, we have refrained from making direct comparisons of functional potential to functional activity in this section (See Results 3.4).

Taxonomic profiles generated from metatranscriptomic data were inconsistent for stool samples in experimental Groups 2 and 3 (Figure 15): NP stool processed with Kit A (Group 2) had largely different phylum level profiles for each of the 3 replicates, whereas stool preserved in RNAP and processed with Kit A (Group 3) only produced a taxonomic profile for 1 out of 4 replicates. For this reason, Phase II experimental Group 3 was excluded from further analysis. These mixed taxonomic results are likely due to greatly reduced read counts for samples in these groupings compared to others (Table 3). Interestingly, stool samples in NAS that were processed with Kit A (Group 1) produced ~105 million reads on average (following QC), whereas stool preserved in NP or RNAP and processed with the same kit produced only a fraction of that amount of reads (Table 3). While this is not necessarily surprising for NP stool, the considerably reduced read count in RNAP stool is unusual. Since RNAP stool processed with Kit B (Group 6) did not result in a diminished read count, it may be that a vendor mismatch incompatibility is occurring between RNAP (Qiagen) and Kit A (Zymo Research).

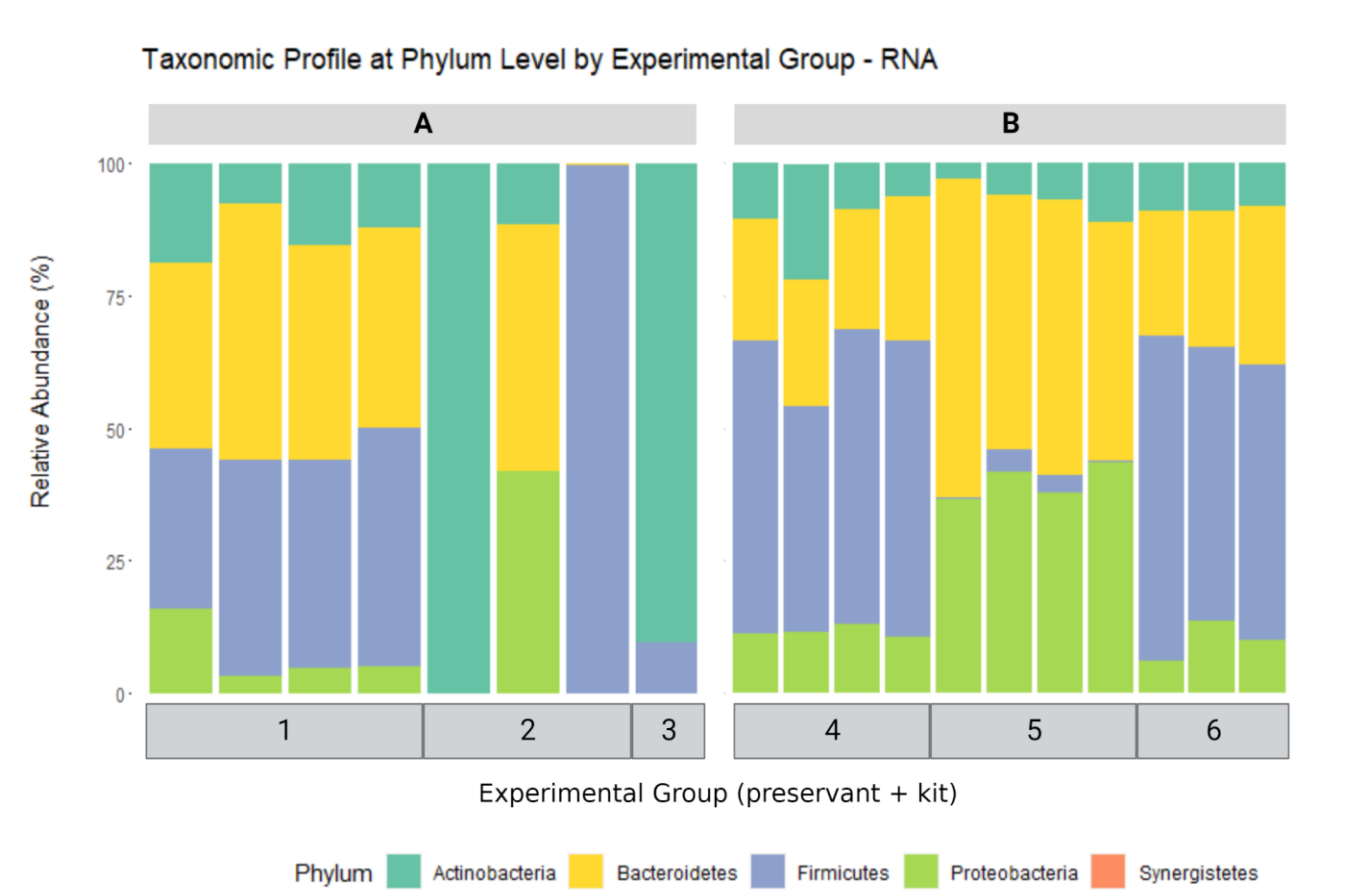


Figure 15. Phylum-level Taxonomic Profiles of Metatranscriptomic Data generated from MetaPhlAn. Stool samples are arranged along the x-axis according to RNA extraction approach (A, B) and nucleic acid preservant (Experimental Groups #1–6). Relative abundance of bacterial phyla is shown on the y-axis.

Table 3. Comparison of Average RNA concentration (ng/µl), read count following quality control, and species richness among Phase II Experimental Groups.

Metatranscriptomics (Phase II)					
Experimental Group	**Preservant**	**Kit**	**[RNA] (ng/µl)**	**Read Count**	**Species Richness**
1	DNA/RNA Shield **(NAS)**	Quick-RNA Fecal/Soil Microbe **(A)**	**325.5**	**104,998,665**	**21.0**
2	None **(NP)**	Quick-RNA Fecal/Soil Microbe **(A)**	**361.0**	**31,728,550**	**8.7**
3	RNAprotect **(RNAP)**	Quick-RNA Fecal/Soil Microbe **(A)**	**187.1**	**12,357,957**	**0.8**
4	DNA/RNA Shield **(NAS)**	RNeasy PowerMicrobiome **(B)**	**91.9**	**129,283,930**	**59.3**
5	None **(NP)**	RNeasy PowerMicrobiome **(B)**	**505.5**	**139,416,692**	**16.0**
6	RNAprotect **(RNAP)**	RNeasy PowerMicrobiome **(B)**	**890.0**	**137,704,436**	**27.7**

3.3.3 Firmicutes:Bacteroidetes Ratio is Impacted by Preservative Use

Regarding metatranscriptomic data, statistically significant differences in F/B ratio were not detected, likely due to the lower number of replicates in this experiment and the large spread of F/B ratio values seen in Group 2 (0.1–99.8). However, the reduced capture of Firmicutes (average relative abundance of Firmicutes is 2.06 for Group 5 vs. 52.54 and 55.09 for Groups 4 and 6, respectively) observed in NP stool processed with Kit B (Group 5, Figure 15) directly translates to a reduced F/B ratio compared to Groups 1, 4, and 6, for which the F/B ratios are all within the reasonable expected range discussed earlier (Figure 16).

3.3.4 Alpha Diversity is Significantly Altered by Metatranscriptomic Experimental Method.

Alpha diversity of the metatranscriptomic data was evaluated using the Shannon Index method according to the diversity function in the vegan v.2.5.7 R package. Significant differences in Shannon Index values were detected for several groupings: NP stool processed with Kit A (Group 2) had significantly lower alpha diversity compared to NAS stool processed with the same kit (mean Shannon Index: 1.16, Group 2; 2.42, Group 1) and also compared to preserved stool (NAS or RNAP) processed with Kit B (mean Shannon Index: 3.04, Group 4; 2.43, Group 6). Likewise, NP stool processed with Kit B (Group 5) had significantly lower alpha diversity than NAS stool processed with the same kit (mean Shannon Index: 1.76, Group 5; 3.04, Group 4)(Figure17). Therefore, observation of diverse species-specific gene expression is reduced in unpreserved stool samples, regardless of the extraction kit used.

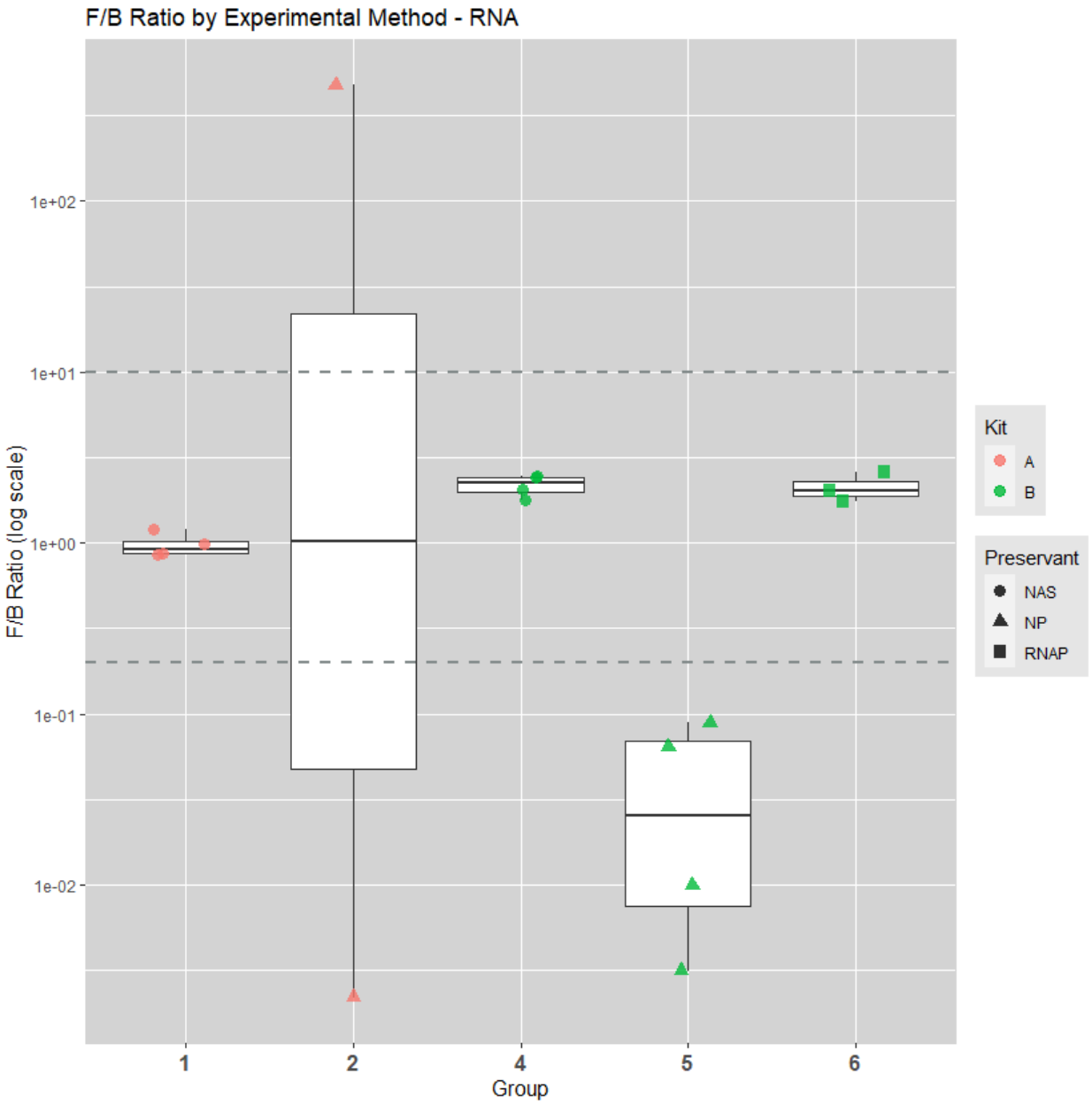


Figure 16. Firmicutes:Bacteroidetes Ratio from Metatranscriptomic Taxonomic Profiles generated with MetaPhlAn. Dashed lines indicate expected range (0.2–10). No statistically significant differences are observed between groupings.

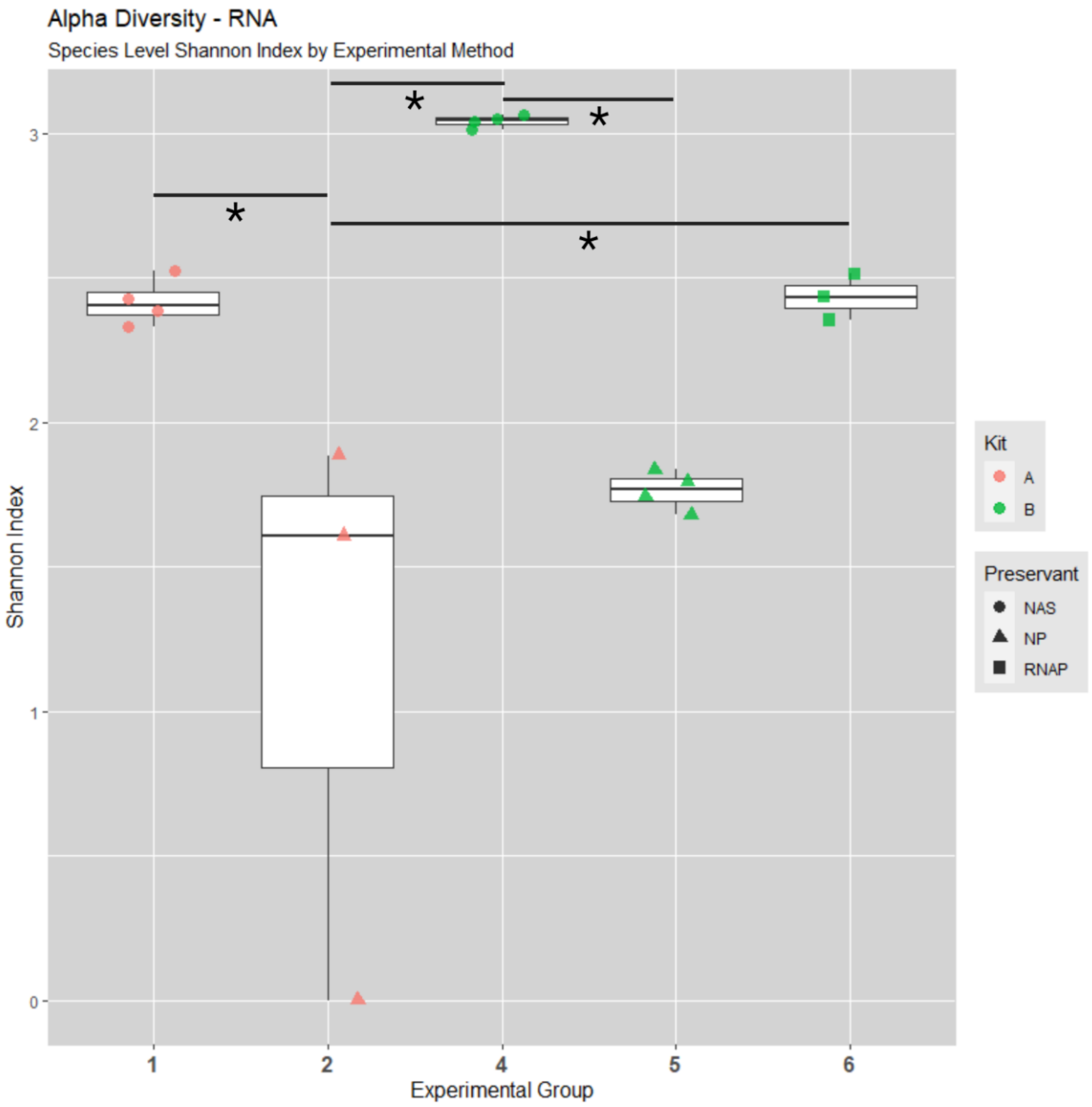


Figure 17. Species-level Shannon Index of Metatranscriptomic Taxonomic Profiles from MetaPhlAn. Overall alpha diversity, quantified by Shannon Index, is significantly different between experimental groups. *$p < 0.05$ by pairwise two-tailed T-tests with Bonferroni correction. Image modified with BioRender.

Investigation of alpha diversity components—species richness and evenness—revealed that NAS stool processed with Kit B (Group 4) was the only experimental method that captured comparable species richness to what was observed in the metagenomic data, despite having a similar quality-controlled read depth as Groups 5 and 6 (Figure 18, Table 3). In fact, the maximum species richness for samples in Group 4 exceeded what was captured from metagenomic data (71 vs. 51). All other experimental groupings captured only a fraction of this diversity, though patterns of reduced diversity in NP stool compared to preserved (NAS or RNAP) stool can still be observed, echoing what was seen in the metagenomic data. Interestingly, the experimental grouping of NAS (Zymo Research) stool processed with Kit B (Qiagen) represents a vendor mismatch in Group 4, though unlike previous results this mismatch does not appear to be detrimental to microbial capture (in fact, the opposite).

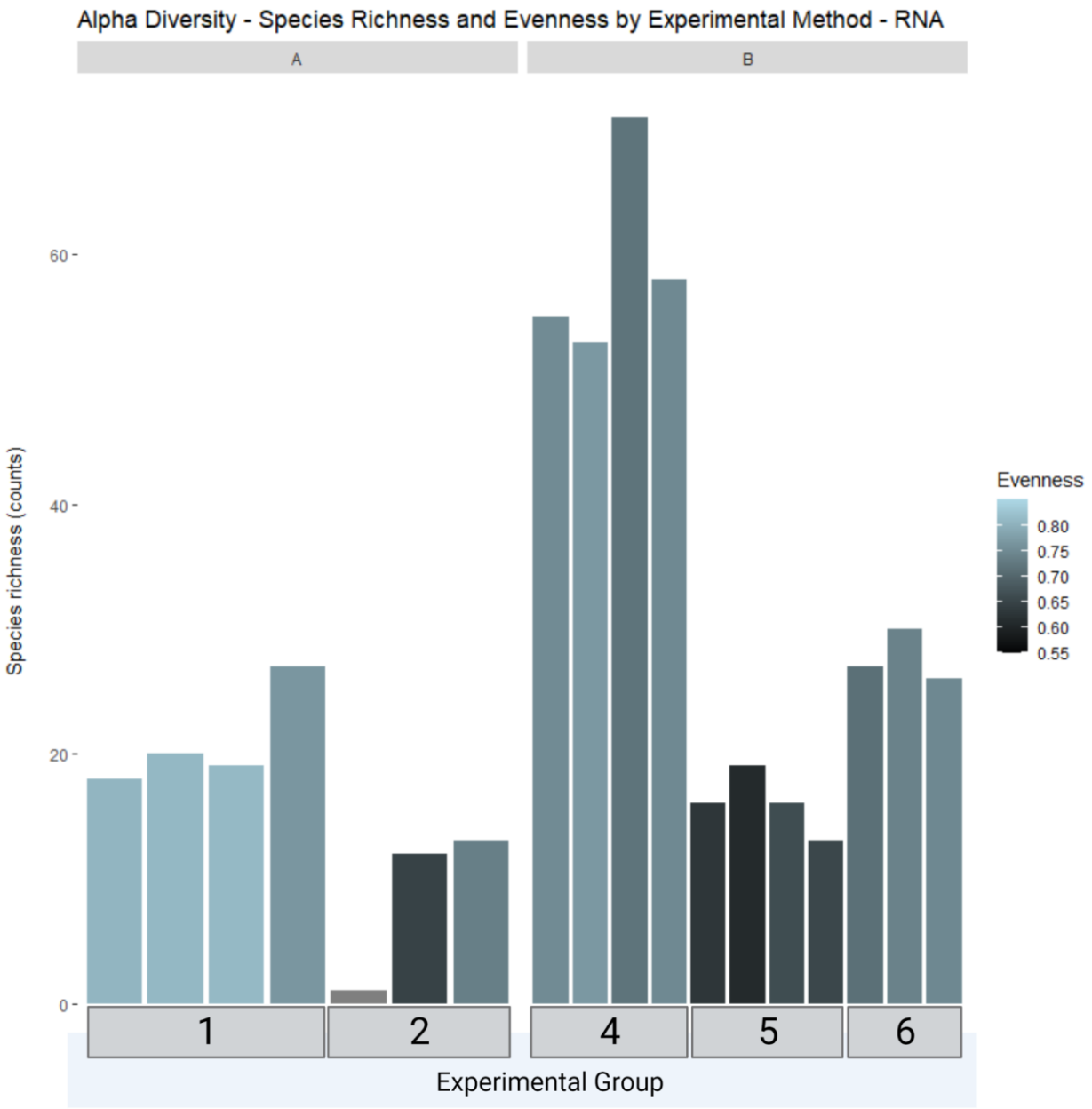


Figure 18. Alpha Diversity Components of Metatranscriptomic Taxonomic Profiles from MetaPhlAn. Microbiome profiles from unpreserved stool (Groups 2 and 5) show reduced alpha diversity in terms of species richness and evenness, compared to preserved stool. Despite vendor mismatch, Group 4 captures the highest amount of diversity. Image modified in BioRender.

3.3.5 Beta Diversity Illustrates Additional Species-level Distinctions Between Groupings.

Principal components analysis of CLR-transformed species counts from metatranscriptomic data reveals that ~63% of variation between samples is accounted for in the first two principal components (Figure 19), indicating that the majority of variation in this dataset is attributable to the principal components plotted. Although the separation between experimental groups is less clear than what was seen from metagenomic data, certain differences can be observed. For example, NP stool processed with Kit B (Group 5) is grouped distinctly from preserved stool (NAS or RNAP) processed with the same kit (Groups 4 and 6), similar to what was seen in metagenomic data. Despite overlapping confidence intervals, ANOSIM revealed that the within-group dissimilarity was lower than between-group dissimilarity for Groups 2, 4, and 6 ($R = 0.5$, $p < 0.01$). The latter two groups represent preserved stool (NAS or RNAP) processed with Kit B and are located proximally on the PCA plot. Unlike what was observed from metagenomic data, samples did not cluster according to whether or not a nucleic acid preservative was used; NAS stool samples processed with Kit A (Group 1) are located proximally to NP stool samples from either kit (Groups 2 and 5) along the first principal component. Therefore, despite similarities with preserved stool (Group 1 vs. Group 6) in terms of F/B ratio and Shannon Index, the species-level community composition in Group 1 samples more closely resemble that of unpreserved stool.

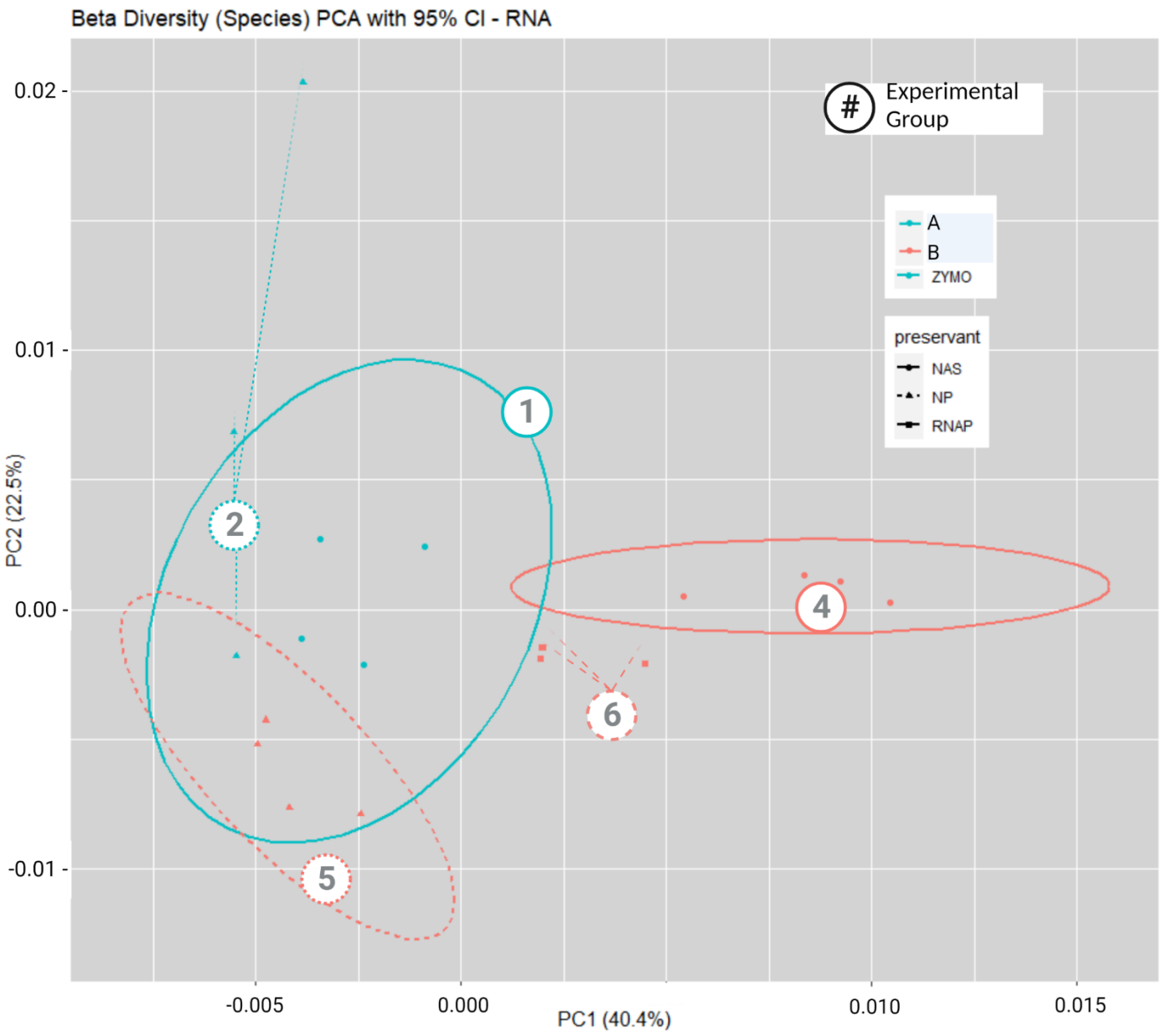


Figure 19. Phase II Species-level Beta Diversity PCA based on Taxonomic Profiles from MetaPhlAn. Groups 2 and 6 contained only 3 replicate samples and thus do not have 95% Confidence Interval (CI) ellipses drawn. NP stool processed with Kit B (Group 5) is grouped distinctly from preserved stool (NAS or RNAP) processed with the same kit (Groups 4 and 6). Image modified using BioRender.

3.3.6 Community Functional Activity is Differentially Captured by Experimental Groups

Multivariable association analysis with MaAsLin (Mallick et al., 2021) on metatranscriptomic pathway abundance data revealed many pathways that were significantly associated with experimental grouping, shown in Figure 20. Investigation of particular pathways illustrates that, unlike the metagenomic data, patterns of functional activity are not solely due to broad differences in capture of major GI phyla. For example, although experimental groups 1, 4, and 6 have similar phylum-level taxonomic distributions and calculated F/B ratios, several of the pathways with significant associations were differentially detected between these groups ($p < 0.01$, Figure 21). Overall, NAS stool processed with Kit B (Group 4) was the only experimental group to capture diverse species-specific contributions across multiple pathways (Figure 21a, b, c), a reflection of the high species richness detected in this grouping.

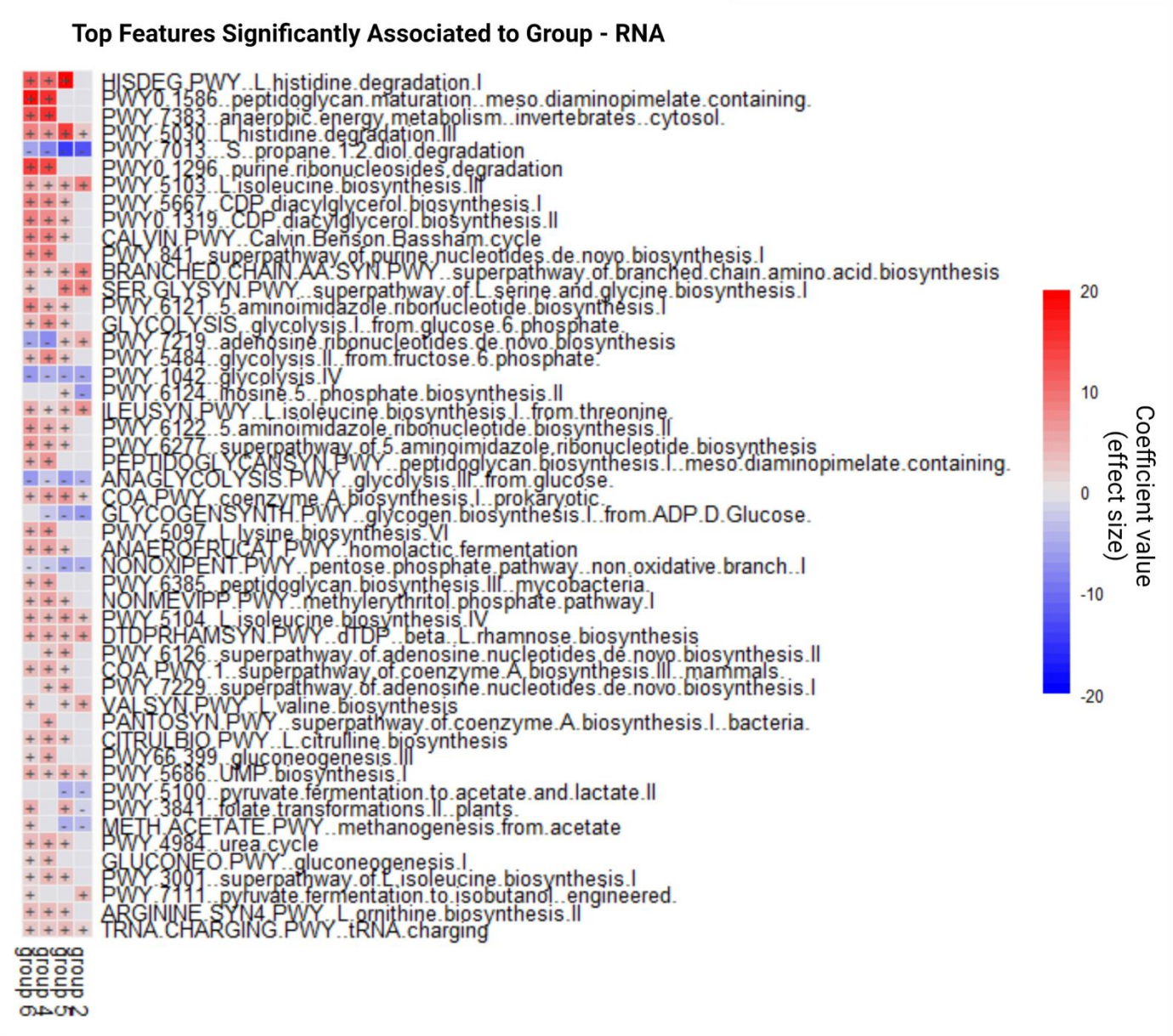


Figure 20. Top 50 Pathways Associated with Phase II Experimental Groups determined by MaAsLin. Metabolic pathways with significant associations to the RNA experimental groups, identified by linear modeling with MaAslin2. Group 1 is used as a comparator for pathway abundance. $p < 0.01$ for one or more experimental groupings per pathway.

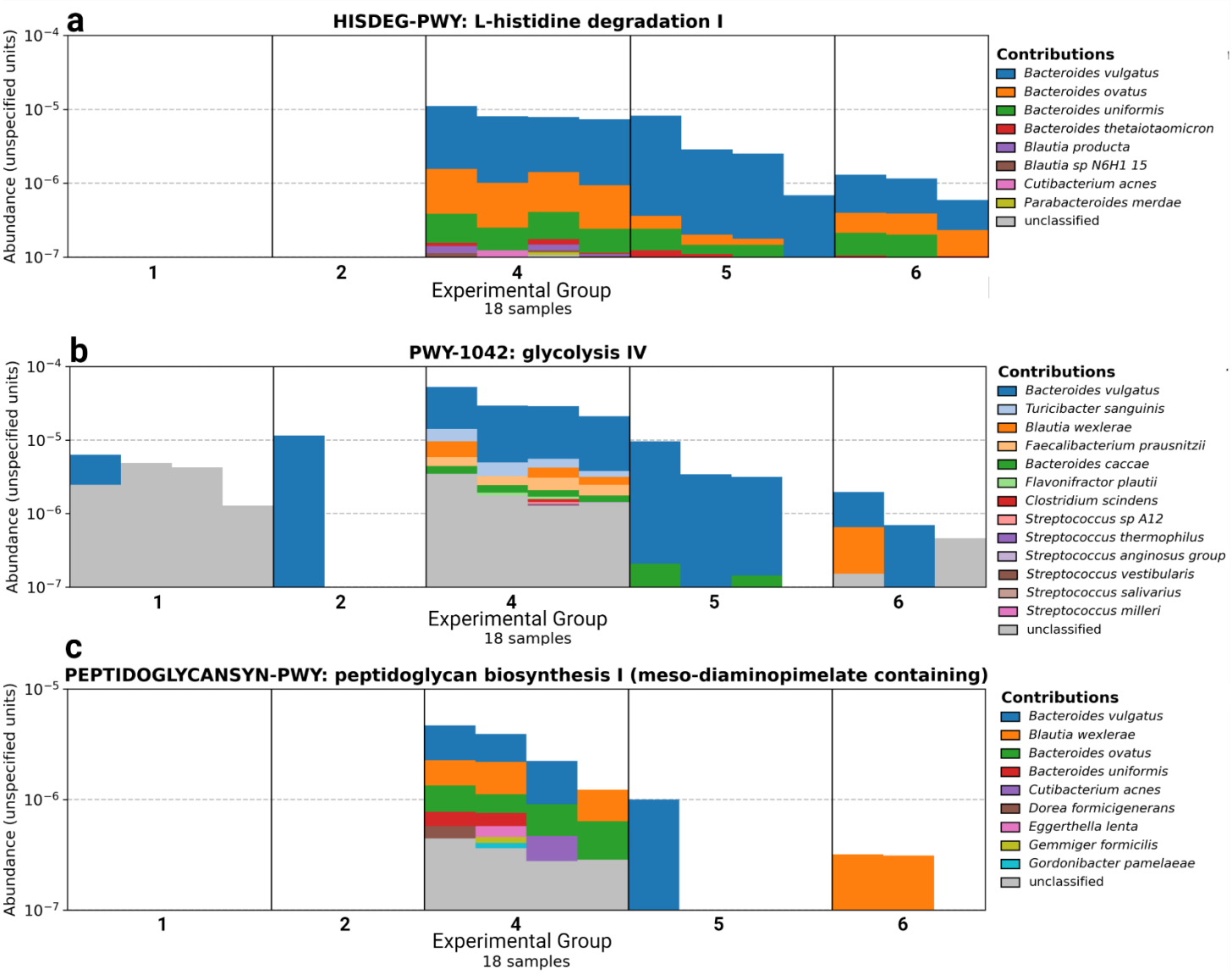


Figure 21. Observation of Functional Activity via HUMAnN Depends on Experimental Approach. The top pathway identified in multivariable association analysis, L-histidine degradation I (a), was captured only by experimental approaches using Kit B (Groups 4–6). Broadly conserved pathways such as glycolysis IV (b) and peptidoglycan biosynthesis (c) were detected in multiple groupings. However, no clear patterns according to experimental preservation or kit are observed. Increased capture of diverse species-specific contributions in Group 4 reflect increased species richness in this group.

3.4 Metatranscriptomics Captures a Subset of Metagenomic Functional Profiles

It would not be appropriate to directly compare functional pathway abundances from metagenomic and metatranscriptomic data in this study because the data were generated using different extraction and sequencing approaches (*i.e.,* bead-beating time and sequencing depth). However, based on the assumption of a low false discovery rate (FDR < 0.05), we can investigate features that are differentially present or absent from the metatranscriptomic data when compared to the metagenomic data.

Most features (gene families and pathways) detected from metatranscriptomic analysis were also detected in the metagenomic data (98.5% and 97.0%, respectively), but there were features detected in the latter that were not detected in the former. This observation supports the idea that mRNA profiling provides increased resolution into microbial community functional activity, as compared to metagenomics alone (Figure 22). A small proportion of features (< 1% of the total) were detected only in the metatranscriptomic data. This may be due to the greatly increased sequencing depth achieved in Phase II (~108 million reads) when compared to Phase I (~3.2 million reads). Sequencing depth is known to correlate with metagenomic feature richness (Gweon et al., 2019). Thus, Phase II benefited from an increased capacity for detecting low-abundance features. Investigation of the features detected only during Phase II reveals gene families and metabolic pathways that are expected from the GI microbiota based on available literature (De Vos et al., 2022). These included several bile acid reductases, choline trimethylamine-lyase, and functional pathways involved in glycolysis and fermentation. Several related or similar features (*e.g.,* enzymes involved in alternate pathways of choline catabolism) were detected in the overlapping features from Phase I and Phase II. This supports the idea that differential presence of these features in Phase II compared to Phase I is due to increased capacity for low-abundance feature detection, rather than false positive assignments or novel functional activity.

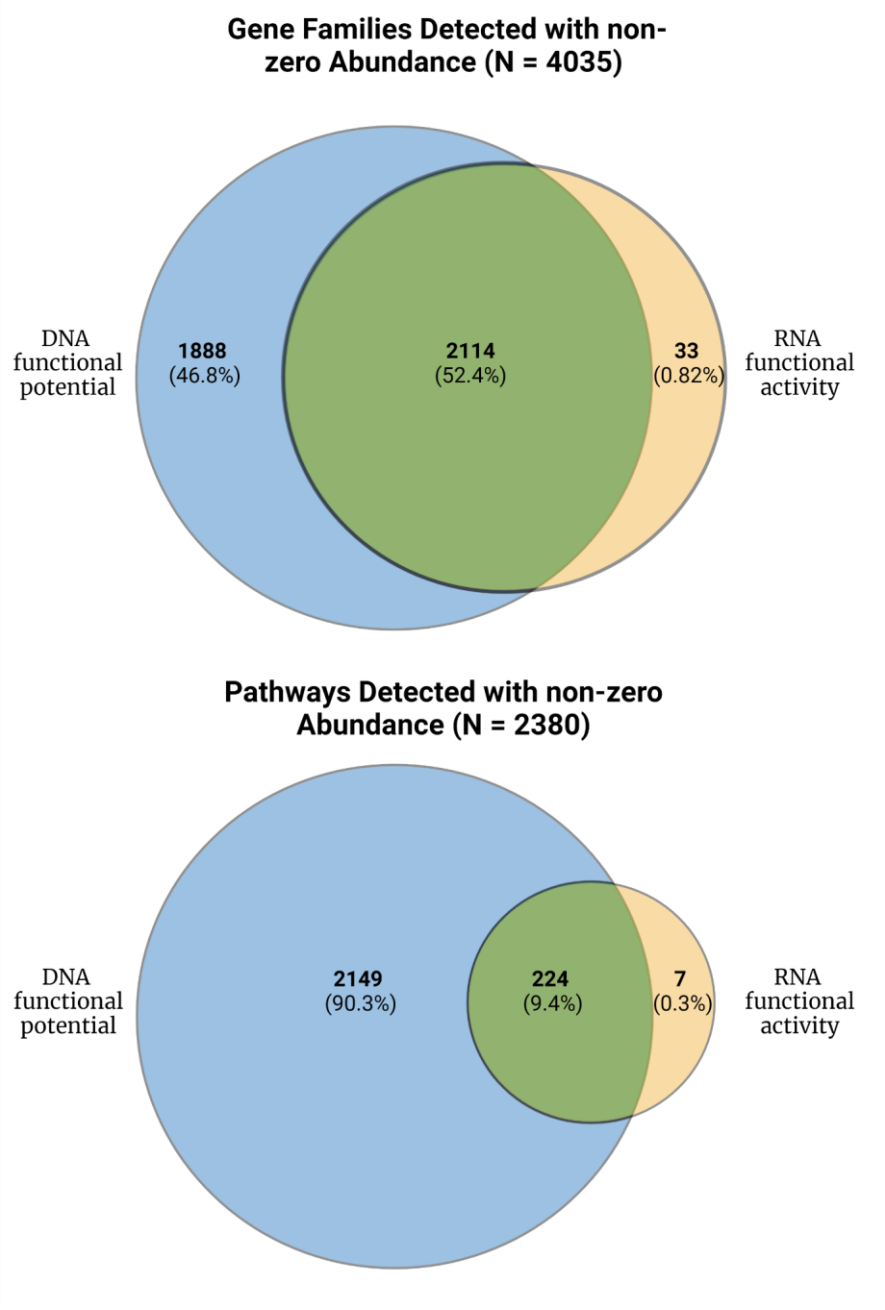


Figure 22. Metatranscriptomics Captures a Subset of Metagenomic Functional Potential. Metatranscriptomic data analysed with HUMAnN captures roughly half of the gene families detected in metagenomic data, and only about 10% of metabolic pathways. Image created in BioRender.

We additionally compared the evenness of microbial profiles across different data types using the Pielou metric, which has previously been used in meta-omics research to quantify variation among different types of microbiome data from different individuals (Franzosa et al., 2014). We observed a higher degree of variation (indicated by a lower evenness score) among taxonomic community profiles from DNA and among gene expression (functional activity) profiles from mRNA, as compared to DNA functional potential profiles. The latter produced consistently high evenness scores regardless of experimental group ($p < 0.01$, Figure 23). These results indicate that there is a high degree of redundancy within microbial community functional potential profiles, meaning that diverse gene families are captured evenly from metagenomic data even when species-level diversity is differentially captured by various experimental techniques. Lower evenness ($p < 0.01$) in functional activity profiles generated from metatranscriptomic data indicate that gene families are unevenly expressed in a majority of samples. Furthermore, the degree of variation, or unevenness, is differently captured by experimental approaches, resulting in a much wider range of evenness scores in the metatranscriptomic data (Figure 23).

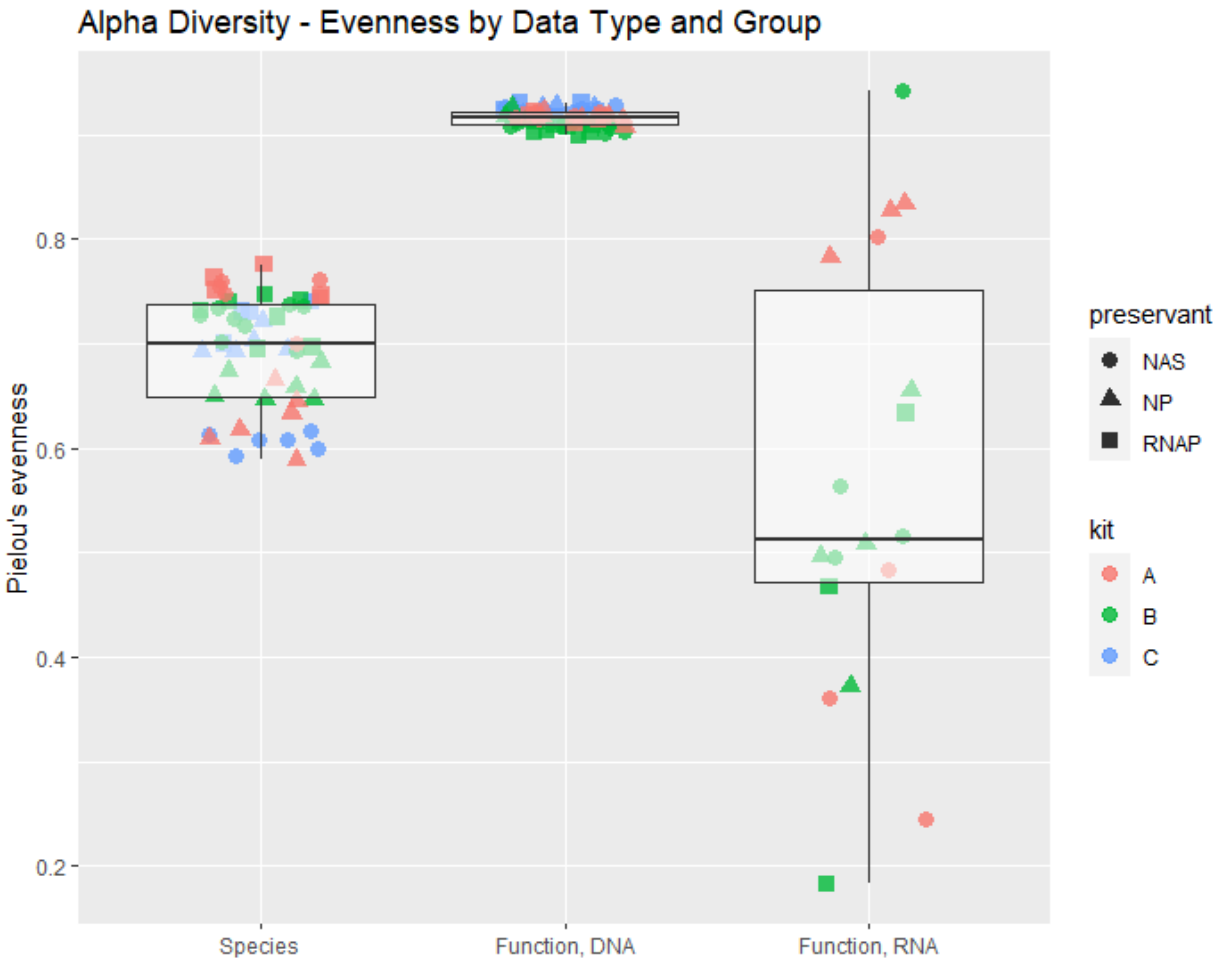


Figure 23. Within-sample Evenness of Taxonomic and Functional Profiles generated by MetaPhlAn and HUMAnN. All three data types have significantly different group means as determined by pairwise two-tailed T-tests with Bonferroni correction ($p < 0.01$).

4 DISCUSSION

The following sections will discuss key differences in taxonomic and functional microbiome profiles observed between the experimental approaches evaluated in Phases I and II. Significant differences in F/B ratio (sections 3.2.3 and 3.3.3), alpha diversity components (sections 3.2.4 and 3.3.4), beta diversity (sections 3.2.5, 3.2.6, and 3.3.5), and community functionality (sections 3.2.7 and 3.3.6) were observed between experimental approaches in both Phases, in support of the first research hypothesis. The importance of particular differentially observed features in the context of stool microbiome research is discussed in the following chapter. Our second research hypothesis is directly supported by the findings in section 3.4, indicating the usefulness of metatranscriptomics for high-resolution functional profiling of microbial communities.

4.1 Bacterial Lysis Approach is a Key Factor for Balanced Capture of GI Taxa

Regarding the importance of bead-beating for capturing Gram-positive microorganisms, our results are in partial agreement with those of Santiago et al. (2014) and Lim et al. (2018). However, there are some key differences: while we found that bead-beating approaches (Kits A and B) facilitated the capture of Gram-positive organisms (*i.e*, Firmicutes) from metagenomic data and increased alpha diversity in preserved stool (in NAS or RNAP), bead-beating approaches used with unpreserved stool (NP) were found to lower the recovery of Gram-negative organisms (Figure 6). This is possibly due to degradation of Gram-negative nucleic acids within the sample during intense mechanical lysis. This is reflected in the significantly altered F/B ratios for these experimental groups (Figure 7). Additionally, the length of bead-beating and the type of beads used may affect the relative proportion of Gram-negative and Gram-positive organisms. We found that a shorter bead-beating time (3 minutes) captured a relatively lower abundance of Gram-positive Firmicutes in NP stool, when compared to stool stored in NAS or RNAP, from

metatranscriptomic data (Figure 15). This is contrary to what was observed from the metagenomic data (Figure 6). Although bead-beating is useful for capturing DNA and RNA from tough-to-lyse Gram-positive organisms, we have found that balanced capture of diverse GI phyla from stool samples depends greatly on two factors: i) Whether or not a nucleic acid preservative is used during sample storage and extraction and ii) the length/intensity of mechanical disruption for cell lysis. In particular, unpreserved stool (NP) is sensitive to biased capture of Gram-positive or Gram-negative GI organisms based on the bacterial lysis approach (Groups 2 and 5, Figure 6; Group 5, Figure 15). Thus, stool samples collected for cohort studies or biobanks should be stabilized in a nucleic acid preservative where possible to facilitate meta-omic microbiome research. Alternatively, if stool samples have already been stored without a preservative, investigators should take care to evaluate optimal lysis approaches for balanced capture of GI taxa prior to engaging in comparative analyses.

4.2 A Highly Stringent Mapping Tool Facilitates Profile Comparisons between Experimental Groupings

As previously discussed in section 1.5.2, the choice of bioinformatic community profiling tool(s) is not arbitrary. The bioBakery analysis suite, developed by researchers for the HMP (Beghini et al., 2021), was chosen for the current study due to its capacity for highly stringent taxonomic and functional assignment (see Results section 3.2.1). The taxonomic profiling software included in the bioBakery analysis suite, MetaPhlAn, has been shown to produce highly precise taxonomic profiles from DNA and RNA sequence data with little-to-no false positive (FP) assignments across taxonomic ranks, compared to other modern taxonomic assignment tools, such as Kraken2, Kallisto, Clark-S, Bracken, and AGAMENON (Rajan et al., 2019; Beghini et al., 2021; Skoufos et al., 2022). Absent a ground truth community profile, the use of a highly stringent profiling tool such as MetaPhlAn allows for the assumption of a low FP rate, and under this assumption, we can increase our confidence in between-

sample taxonomic diversity comparisons. For example, differences in the relative abundance of particular GI taxa can be affected by changes in the abundance of all other taxa since the data are compositional. However, under the low FP assumption, we can determine that the samples with the highest absolute species richness (Figures 9 and 18) represent true taxonomic diversity within the stool source material.

In this study, stool samples stored in DNA/RNA Shield (Zymo Research) and processed with the RNeasy PowerMicrobiome Kit (Qiagen; Group 4) produced the richest taxonomic profiles from RNA (mean richness of 59.3 species) as compared to all other experimental groupings in Phase II (mean richness 8.7–27.7 species). Therefore, there is the potential to observe at least ~60 species using RNA extracted from the source material via the methodologies tested. Taking into account that all stool samples originated from the same source, the low taxonomic assignment FP rate, and consistent measures among technical replicates, we can attribute differential observation of microbiome species richness to underlying experimental biases. These experimental effects were less pronounced in taxonomic profiles from Phase I, where comparable species richness was observed in DNA data from several experimental groups (1, 3, 4, 6, 7, 8, and 9; mean richness 37.5–46.8 species). Despite this potential, the combination of unpreserved stool (NP) and the use of a bead-beating DNA extraction kit (Quick-DNA Fecal/Soil Microbe Kit, Group 2; Zymo Research, or QIAamp PowerFecal Pro Kit, Group 5; Qiagen) consistently produced taxonomic profiles from MetaPhlAn with comparatively lower species richness (mean richness 30.5, Group 2; 32.5, Group 5), despite similar sequence depths for samples in these groups (Table 2). Hence, the methodologies applied to stool samples in Groups 2 and 5 from Phase I hinder the observation of taxonomic diversity via MetaPhlAn from the stool source.

4.3 Vendor Mismatch May Be Detrimental to Taxonomic Characterization

Nucleic acid preservative reagents and extraction kits from two vendors—Qiagen and Zymo Research—were evaluated in this study. There are several reasons for selecting these vendors in particular:

Our laboratory routinely uses nucleic acid extraction kits from Qiagen or Zymo Research for genomic-based analyses, including microbiome investigations (Forbes et al., 2018), making them likely candidates for future in-house metagenomics and metatranscriptomics research. Furthermore, each of the DNA extraction kits tested in this study (or, a similar kit from the same vendor; see Table 1 and section 4.7) has been recently evaluated in the context of standardizing methodologies for microbial community measurements from stool (Tourlousse et al., 2021). Analogous RNA extraction kits were chosen from the same vendors where possible (see Methods 2.2 for detailed extraction approaches). As well, each of these vendors offers chaotropic reagents (see Appendix 4 for chemical specifications) for the stabilization of nucleic acids in biological samples as well as stool-specific microbiome extraction kits, allowing the current study to examine the effects of vendor matching/mismatching regarding these key components. The exact components and reaction chemistry are rarely disclosed for most commercial extraction kits, leading to uncertainty about whether the components of one product may impede another (though this is unlikely to be the case for products originating from the same vendor that are designed to be used in tandem). Therefore, we can define a "vendor mismatch" as occurring in any experimental approach that employs products from both vendors. We considered the effects of vendor mismatch important to evaluate because the storage conditions of stool samples used for microbiome research may be out of the investigator's control (*i.e.*, when using samples from a biobank). Thus, understanding the effects of vendor mismatch is useful for extraction kit selection in the context of sample storage conditions. For example, stool stored in DNA/RNA Shield (Zymo Research) and processed with QIAamp Fast DNA Stool Mini Kit (Qiagen) are considered candidates of a vendor mismatch. In this study, there were 3 such groupings in Phase I (Groups 3, 4, and 7) and 2 in Phase II (Groups 3 and 4). Interestingly, we observed that the effect of vendor mismatching was neither consistently detrimental nor advantageous across groupings; rather, our findings show examples of each case, depending on the specific combination of products.

The most detrimental vendor mismatch appears to have occurred in Phase II with stool stored in RNAprotect (Qiagen) and processed with the Quick-RNA Fecal/Soil Microbe Microprep Kit (Zymo Research). This experimental grouping (Group 3) produced the fewest high-quality sequence reads, disproportionate to RNA yield (Table 3). On average, the sequence count for Group 3 samples, following quality trimming and filtering, was only one-tenth (Group 3 mean = 12.4 million reads) of other groups we examined (Group 1 mean = 105 million reads, Group 4 mean = 129.3 million reads) and less than half of unpreserved stool samples processed with the same kit (Group 2 mean = 31.7 million reads; Table 3). Due to this greatly reduced read depth, three of the four technical replicates in this group were not assigned a taxonomic profile by MetaPhlAn. Hence, this experimental group was excluded from further data analysis. Of note, the analogous vendor mismatch in Phase I Group 3 affected neither the average read depth nor the downstream profiling, which was comparable to other groupings (Table 2, Figure 8). Therefore, if reagent incompatibility is responsible for these effects, it may be specific to components within the RNA isolation kit. To our knowledge, this detrimental vendor mismatch between RNAprotect (Qiagen) the Quick-RNA Fecal/Soil Microbe Microprep Kit (Zymo Research) has not been previously described for stool metatranscriptomics.

Another example of a detrimental vendor mismatch was observed from the reduced evenness in species-level taxonomic profiles of Phase I Group 7 samples, which represents stool stored in DNA/RNA Shield (Zymo Research) and processed with the QIAamp Fast DNA Stool Mini Kit (Qiagen). Six out of six total replicates in Group 7 produced evenness scores ≤ 0.62, lower than the evenness scores of any replicates from Groups 8 or 9 (minimum evenness = 0.69, N = 9), and an 18% reduction compared to the group with the highest mean evenness (Group 3 mean = 0.76). Compared to NP or RNAP stool samples processed with the same kit (Groups 8 and 9), Group 7 samples had comparable unique species counts (Figure 9), but the overrepresentation of Bacteroidetes in Group 7 results in such a large reduction in

overall evenness. What is striking about the two detrimental cases of vendor mismatch presented here is that in both cases, despite the use of a nucleic acid preservant, the samples in these groupings produced worse outcomes (regarding taxonomic richness or evenness, respectively; Table 3 and Figure 9) than unpreserved stool that was processed similarly. These results call into question the widespread utility of nucleic acid preservative agents and underscore the need for pilot studies in microbiome research in order to establish experimental methodologies that are optimized for capturing research outcomes.

However, in this study, we have also observed a case of vendor mismatch that appeared to be advantageous for capturing taxonomic diversity from stool samples. All replicates (4/4) of stool samples stored in DNA/RNA Shield (Zymo Research) and processed with the RNeasy PowerMicrobiome Kit (Qiagen; Group 4) captured the greatest amount of species diversity—in terms of richness—from the metatranscriptomic data (Figure 18). The analogous combination in Phase I Group 4 likewise produced diverse species-level taxonomic profiles (all 8 replicates from Phase I Group 4 produced species counts between 41-51) consistent with three other experimental groupings in terms of its beta diversity (Groups 1, 3, and 6; Figure 10). Together, the replicate samples from Groups 1, 3, 4, and 6 clustered proximally on the PCA plot and have overlapping 95% confidence intervals, indicating that the taxonomic profiles of these approaches (*i.e.*, using preserved stool processed with a bead-beating extraction kit) are highly similar. Thus, vendor mismatch is not universally detrimental, and reagent incompatibility may not be predictable, further emphasizing the need for pilot studies.

4.4 Experimental Bias Leads to Differential Observation of Meaningful Meta-Omic Features

Microbiome-mediated regulation of amino acid degradation, fermentation, and other metabolic processes in the GI tract has been identified as critical to host health, and disruptions to this regulatory symbiosis are considered to play a major role in metabolic disorders (de Vos et al., 2022). Our group has previously described common microbial-mediated metabolic dysregulation patterns, including amino acid

and fatty acid metabolism, between healthy and IBD or CRC cohorts (Pratt et al., 2021; Appendix 1, Table S3). All observed pathway abundances and their association to the tested experimental approaches were examined using multivariable association analysis (MaAsLin2; Mallick et al., 2021) in this study in order to evaluate whether the experimental approaches in Phases I and II are differentially suited to capture functional information from stool samples.

In the current study, we have identified several pathways involved in amino acid (Figures 12 and 20) and fatty acid metabolism (Figure 12), along with pathways involved in anaerobic respiration and metabolism, to be among the top associations identified by MaAsLin2 linear modelling with significant associations to the experimental approach used to store and process the stool sample ($p < 0.01$). The differential observation of pathways with associations to metabolic and GI diseases, namely amino acid and fatty acid metabolism, between experimental approaches was unexpected. Given that our data has been generated from a single stool sample from a healthy individual, this demonstrates how experimental bias and differences in nucleic acid integrity can lead to a differential observation of meaningful meta-omic features, possibly opening the door for spurious associations in disease research. Furthermore, our data highlights the challenges associated with comparison of meta-omic data from individual studies, and provides considerations for meta-analysis approaches that aim to reduce between-study noise and variation.

4.5 Metatranscriptomics Requires *a priori* Knowledge

Decision-making during nucleic acid library preparation requires knowledge of the integrity of the biological sample and any subsequently isolated nucleic acids, which will ultimately impact library diversity and sequence data quality (Shen, 2019). This information is used to inform library input amounts, the number of PCR cycles used, library pooling schemes, and sequencing controls such as % PhiX spike-in—a well-defined bacteriophage genome sequence with a diverse base composition that is used to

improve template registration when sequencing low-diversity libraries. For example, the current project involved several rounds of pilot RNA data generation, which were used to determine optimal laboratory procedures for metatranscriptomics. Without preliminary work and an understanding of the ways in which minor laboratory alterations can influence data capture and analysis, default methodologies may not be sufficient for the task at hand. Additionally, particular microbiome profile outcomes (*e.g.*, Shannon Index, F/B ratio, or differentially expressed genes) must be outlined prior to initiating lab work so that laboratory methods can be optimized to capture them. For instance, if one wishes to evaluate disease-associated changes in the alpha diversity of the GI microbiome, care must be taken to ensure that the lysis approach (bead size, instrument velocity, and time) is consistent across samples—and informed by knowledge of nucleic acid integrity—so as not to introduce a technical bias. In this study, we have observed that although several experimental approaches produced similar microbiome profiles from metagenomic data (Figures 10 and 11), the capture of functional activity from metatranscriptomic data was much more sensitive to the experimental approach used (Figure 21), highlighting the necessity of pilot studies for metatranscriptomic data capture. Such studies may wish to evaluate optimal cell lysis conditions (*e.g.*, by varying bead-beating time as in Tourlousse et al., 2021), methods of prokaryotic mRNA enrichment based on nucleic acid integrity (Petrova et al., 2017), RNA library input amounts (*e.g.*, although the Illumina protocol recommends 100 ng of input RNA, an input amount of 200 ng in the current study was informed by pilot work; see Methods 2.3), and also the number of libraries to be pooled and sequenced in a single run, based on instrument capacity as well as nucleic acid integrity.

4.6 Practical Considerations

Although a large area of human microbiome research is focused on characterizing disease-specific GI dysbioses (Lloyd-Price et al., 2019; Yachida et al., 2019), we have concentrated on the methodological evaluation of healthy stool in this study. Patients experiencing GI distress (*e.g.*, inflammation and/or

diarrhea) may produce stool that has much higher blood/water content (Santiago et al., 2014) or host DNA (Jiang et al., 2020), as compared to healthy stool. Therefore, the evaluation of disease-associated stool samples requires unique considerations regarding nucleic acid preservation, which are not explored here.

It is likely that the commercial extraction kits and stabilization reagents used in the current study may be altered or discontinued in the future. As such, we have taken care to describe each approach in as much detail as is available (see Methods 2.2) to facilitate reproducibility. We have demonstrated that despite limited knowledge of proprietary reagents, commercial offerings can be used to successfully capture diverse taxonomic and functional profiles from healthy stool (Figures 9, 13, 18, and 21). Based on the results of the current study, we re-emphasize the recommendations in Szamozi et al., (2020); namely, that publication of detailed sample processing methodology, alongside raw sequence data, is imperative for validation, interpretation, comparison, and reproducibility of microbiome research.

Other practical considerations for meta-omic microbiome research that are supported by the current project include the complexity of laboratory workflows for meta-omic research, as well as disjointed data generation and analysis. Regarding the former, it is important to acknowledge that the per-sample cost can rise exponentially each time a new step is added to laboratory workflows. Also, subsampling may compromise data capture with each additional step, and the added processing presents an opportunity for further nucleic acid degradation, particularly with RNA. We found this to be the case during pilot data generation, where superfluous sample processing in the form of an additional rRNA depletion step led to low-quality sequence reads rather than improving the sequence data. Finally, it is worth mentioning that meta-omic data generation and analyses are frequently handled by separate laboratories or teams, which presents opportunities for miscommunication and misunderstanding of how laboratory processing of biological samples can impact aspects of meta-omic data. In the current study, nucleic acid extractions, library preparations, and sequence data analysis were performed by the same

individual. Therefore, there is a particular need for accessible and multidisciplinary knowledge translation in the field to facilitate comprehensive microbiome research.

4.7 Overall Performance of Experimental Approaches

The Qiagen PowerX kits used in this study (QIAamp PowerFecal Pro/RNeasy PowerMicrobiome) are formerly known as MoBio PowerFecal DNA/RNA Isolation kits. These nucleic acid isolation kits have been validated in other meta-omic methodological studies and found to perform well with respect to nucleic acid yield and integrity (Reck et al., 2015; Koorakula et al., 2022). In the current study, we have observed that for metagenomics (Phase I) the QIAamp PowerFecal Pro Kit (Kit B) performed similarly to the Zymo Quick-DNA Fecal/Soil Microbe Kit (Kit A) in terms of alpha and beta diversity measures (Figures 8, 9, and 10). However, for metatranscriptomics (Phase II) the difference between extraction kits was much larger in terms of overall performance across several metrics (Results 3.3 - Phase II). Here, Kit B unquestionably outperformed Kit A in terms of capturing greater microbial taxonomic and functional diversity (Table 3, Figures 18 and 21).

Additionally, we evaluated the Qiagen QIAamp Fast DNA Stool Mini Kit (#51604, Kit C) in Phase I of this study. A predecessor of this kit, the Qiagen QIAamp DNA Stool Mini Kit (#51504), was found to be highly accurate for taxonomic profiling of human fecal samples (Costea et al., 2017; Tourlousse et al., 2021). However, unlike its predecessor, this kit primarily uses a thermal lysis approach. We found it to be the best approach for observing Gram-positive and Gram-negative bacterial phyla from unpreserved (NP) stool samples, specifically. However, it is important to note that the Kit C stool samples produced unique taxonomic profiles in terms of beta diversity analysis (Figure 10). It should be noted that the magnitude of this difference in the context of inter-individual microbiome variation was not explored here.

5 CONCLUSIONS

5.1 Overview of Main Findings

In this study, we have determined that the experimental methodology used to stabilize and isolate nucleic acids from human stool samples can significantly impact the ability to capture GI microbiome diversity in terms of phylum- and species-level taxonomic profiles and functional gene family or metabolic pathway profiles from metagenomic and metatranscriptomic data. In addition to providing support for current validated approaches, this research highlights the impact of combined approaches for sample storage and processing. Specifically, we observed that stool samples stored without a preservative reagent (NP) and subsequently processed using a mechanical lysis approach have a reduced capacity for balanced detection of major GI phyla, compared to stool stored in either DNA/RNA Shield (Zymo Research) or RNAprotect (Qiagen; Figures 6 and 15). This differential capture consequently affects the ability to identify downstream functional potential or functional activity from particular taxa (Figures 13 and 21) and as such, this combined approach should be avoided when unbiased community observation is desired.

Additionally, we observed that GI microbiome characteristics commonly used as health markers in disease research, including the Firmicutes:Bacteroidetes ratio (Figure 7) and alpha diversity (Figures 8, 9, 17, and 18), are differentially impacted by the specific combination of preservative reagent and nucleic acid extraction kit. Beta diversity analysis of between-sample taxonomic and functional profiles revealed significant differences between experimental approaches (Figures 10, 11, and 19). Regarding metatranscriptomic data, for example, stool stored in DNA/RNA Shield (Zymo Research) and processed with the RNeasy PowerMicrobiome kit (Qiagen; Group 4) was the only combination to produce rich and diverse species-level taxonomic and functional profiles on the level that was observed from the

metagenomic data (*i.e.*, representation from > 30 species; Figure 18). Notably, we observed significant differences (Figures 12 and 20, $p < 0.01$) between experimental approaches in terms of their ability to capture functional pathways involved in amino acid degradation and biosynthesis, cofactor synthesis, fatty acid metabolism, glycolysis, and fermentation—modules that have been identified as differentially expressed in disease cohorts compared to healthy controls (Pratt et al., 2021). This has implications for microbiome disease research since stool that is bloody or runny (*e.g.*, from an IBD patient) may have increased levels of nucleic acid degradation compared to healthy stool, and this degradation could subsequently hinder the observation of key microbial metabolic features and confound health/disease comparisons.

Thus, based on the gathered evidence herein, we can provide suggestions for future metagenomics and metatranscriptomics investigations when the aim is unbiased microbial community data capture from stool: For metagenomics, the use of a nucleic acid preservative reagent during sample storage is recommended, particularly when mechanical lysis is used during isolation of nucleic acids. In this study, DNA/RNA Shield (Zymo Research) and RNAprotect (Qiagen) performed similarly well for metagenomics. However, if stool has been stored without a stabilizing agent, we recommend that investigators consider less vigorous approaches to cell lysis, such as temperature-based. In the current study, unpreserved (NP) stool processed with the QIAamp Fast DNA Stool Mini Kit (Qiagen; Thermal lysis, Group 8) resulted in metagenomic MetaPhlAn taxonomic profiles with F/B ratios in the expected range for healthy individuals (Figure 7), as well as species richness and evenness measurements only slightly below those of preserved samples processed with a bead-beating kit (Groups 1, 3, 4, and 6; Table 2, Figure 9). Differences in group mean richness were not statistically significant between Group 8 and preserved, bead-beating groups (4–8 fewer species identified on average in Group 8; Table 2) and reduced

evenness in Group 8 was only statistically significant compared to Groups 1 and 3 (mean group evenness = 0.70 vs 0.74 and 0.76, $p < 0.05$).

Metatranscriptomic data is much more sensitive to combined experimental methodology; we observed that stool processed with the Quick-RNA Fecal/Soil Microbe kit (Zymo Research, Kit A) produced inconsistent data profiles when stored either in RNAprotect (Qiagen) or without a preservative (NP). Stool stored in DNA/RNA Shield (Zymo Research) appeared to perform best for metatranscriptomic data in this study, however, due to inconsistent performance between RNA extraction kits our recommendation is to always perform pilot studies for optimized metatranscriptomic data capture.

5.2 Limitations of This Study

This methodological study was designed with tightly controlled parameters in order to observe the isolated effects of specific laboratory methods on stool microbiome profiles. As such, there are several areas related to this study that remain to be explored. Firstly, the stool sample used in the current study was collected from a healthy individual. Stool that is bloody, for example, from IBD patients, may contain higher host DNA content (Jiang et al., 2020) or have increased degradation of nucleic acids compared to healthy stool. The stool used in this study, therefore, represents ideal study conditions for microbial nucleic acid extraction. Likewise, due to time constraints, only one stool sample was evaluated, so we cannot compare the variation introduced by the methodology described here, for example, to inter-individual microbiome variation or longitudinal intra-individual variation. Furthermore, in the interest of practicality and reproducibility, we have purposely chosen to evaluate commercially available products in this study. One limitation of this approach is that the chemical identities of particular reagents from commercial vendors often are not disclosed. Therefore, we cannot investigate the specific underlying components contributing to vendor mismatch to describe the mechanism of inhibition or enhancement. Additionally, because DNA and RNA were extracted and sequenced separately and with different methods (*i.e.*, bead

type and length of disruption, also sequencing platform and read depth) in Phases I and II, it would not be appropriate to directly compare relative abundances from the respective data, for example, in a quantitative comparison of DNA and RNA abundance for particular organisms, which has been done previously (Schirmer et al., 2018). We have shown that the extraction approach substantially impacts taxonomic and functional profiles, and thus, a direct quantitative comparison cannot be made.

Broadly, the current study is also subject to many of the same limitations as other GI microbiome investigations. Namely, the observation of a single kingdom (Bacteria) as a proxy for the entire GI microbiome will undoubtedly fail to capture the true complexity of microbial interactions in this microecosystem. Moreover, although taxonomic controls (Mock Community DNA Standards and Fecal Reference; Zymo Research) were used for monitoring bias in nucleic acid extraction and sequencing, and to ensure the taxonomic profiling tool was working as intended, the true community composition of the stool sample is unknown. Thus, our conclusions relate only to differences in observed, or measured microbiome profiles. As previously discussed, establishing a "ground truth" for stool samples, which inherently are heterogeneous, is difficult. Using alternative metrics, such as the proportion of sequence reads passing quality control, the observed Firmicutes:Bacteroidetes ratio, sample diversity (alpha/beta), as well as comparisons with published healthy human stool profiles (HMP 2012, Abu-Ali et al., 2018), we can robustly compare the various methods herein without establishing a "ground truth". We have also chosen microbiome profiling tools (bioBakery) with reported low FP rates (Rajan et al., 2019; Beghini et al., 2021; Skoufos et al., 2022) in order to increase confidence in our profile comparisons.

5.3 Future Directions

New or altered commercial approaches should always be validated in methodological studies such as this before they are applied to disease research in order to evaluate the capacity for capturing research outcomes and any incompatibilities with other reagents. Therefore, as vendors continue to develop new

products, the need to systematically evaluate them will continue as well. Other considerations include the underlying integrity and physical properties of stool samples used for research, which will affect optimal cell lysis approaches. It will therefore be valuable to recreate aspects of the current study using stool from a patient cohort such as IBD. There are likely going to be considerable differences compared to working with healthy stool, so the ability to capture particular taxonomic or functional outcomes from samples with higher blood or host DNA content needs to be evaluated. Likewise, separate protocols may need to be developed for data capture from healthy or diseased stool, for example, in order to answer the questions: is observed alpha diversity lower in IBD samples simply because the nucleic acids in the sample are more degraded? And, do nucleic acid stabilizing reagents preserve taxonomic diversity efficiently in stool that is bloody/runny? Further methodological investigations will help expand our understanding of how biological sample integrity and experimental bias affect observed meta-omic microbiome profiles. Informed experimental design will help to increase the quality and resolution of meta-omic data by decreasing inherent background noise due to experimental effects and subsequently assist in untangling complex disease etiologies via meta-omics.

REFERENCES

Abu-Ali, G. S., Mehta, R. S., Lloyd-Price, J., Mallick, H., Branck, T., Ivey, K. L., et al. (2018). Metatranscriptome of human faecal microbial communities in a cohort of adult men. *Nat. Microbiol.* 3, 356–366. doi: 10.1038/s41564-017-0084-4.

Alberti, A., Belser, C., Engelen, S., Bertrand, L., Orvain, C., Brinas, L., et al. (2014). Comparison of library preparation methods reveals their impact on interpretation of metatranscriptomic data. *BMC Genomics* 15, 912. doi: 10.1186/1471-2164-15-912.

Ananthakrishnan, A. N., Luo, C., Yajnik, V., Khalili, H., Garber, J. J., Stevens, B. W., et al. (2017). Gut microbiome function predicts response to anti-integrin biologic therapy in Inflammatory Bowel diseases. *Cell Host Microbe* 21, 603-610.e3. doi: 10.1016/j.chom.2017.04.010.

Atarashi, K., Tanoue, T., Shima, T., Imaoka, A., Kuwahara, T., Momose, Y., et al. (2011). Induction of Colonic Regulatory T Cells by Indigenous Clostridium Species. *Science* 331, 337–341. doi: 10.1126/science.1198469.

Bashiardes, S., Zilberman-Schapira, G., and Elinav, E. (2016). Use of Metatranscriptomics in Microbiome Research. *Bioinforma. Biol. Insights* 10, 19–25. doi: 10.4137/BBI.S34610.

Becattini, S., Sorbara, M. T., Kim, S. G., Littmann, E. L., Dong, Q., Walsh, G., et al. (2021). Rapid transcriptional and metabolic adaptation of intestinal microbes to host immune activation. *Cell Host Microbe* 29, 378-393.e5. doi: 10.1016/j.chom.2021.01.003.

Beghini, F., McIver, L. J., Blanco-Míguez, A., Dubois, L., Asnicar, F., Maharjan, S., et al. (2021). Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. *eLife* 10, e65088. doi: 10.7554/eLife.65088.

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120. doi: 10.1093/bioinformatics/btu170.

Boutin, R. C., Petersen, C., Woodward, S. E., Serapio-Palacios, A., Bozorgmehr, T., Loo, R., et al. (2021). Bacterial-fungal interactions in the neonatal gut influence asthma outcomes later in life. *eLife* 10. doi: 10.7554/eLife.67740.

Brady, A., and Salzberg, S. (2011). PhymmBL expanded: confidence scores, custom databases, parallelization and more. *Nat. Methods* 8, 367–367. doi: 10.1038/nmeth0511-367.

Cardona, S., Eck, A., Cassellas, M., Gallart, M., Alastrue, C., Dore, J., et al. (2012). Storage conditions of intestinal microbiota matter in metagenomic analysis. *BMC Microbiol.* 12, 158. doi: 10.1186/1471-2180-12-158.

Costea, P. I., Zeller, G., Sunagawa, S., Pelletier, E., Alberti, A., Levenez, F., et al. (2017). Towards standards for human fecal sample processing in metagenomic studies. *Nat. Biotechnol.* 35, 1069–1076. doi: 10.1038/nbt.3960.

Cullen, C. M., Aneja, K. K., Beyhan, S., Cho, C. E., Woloszynek, S., Convertino, M., et al. (2020).

Emerging Priorities for Microbiome Research. *Front. Microbiol.* 11. Available at: https://www.frontiersin.org/article/10.3389/fmicb.2020.00136 [Accessed May 6, 2022].

De Filippis, F., Paparo, L., Nocerino, R., Della Gatta, G., Carucci, L., Russo, R., et al. (2021). Specific gut microbiome signatures and the associated pro-inflamatory functions are linked to pediatric allergy and acquisition of immune tolerance. *Nat. Commun.* 12, 5958. doi: 10.1038/s41467-021-26266-z.

De Vos, W. M., Tilg, H., Hul, M. V., and Cani, P. D. (2022). Gut microbiome and health: mechanistic insights. *Gut* 71, 1020–1032. doi: 10.1136/gutjnl-2021-326789.

Deutscher, M. P. (2006). Degradation of RNA in bacteria: comparison of mRNA and stable RNA. *Nucleic Acids Res.* 34, 659–666. doi: 10.1093/nar/gkj472.

Dore, J., Ehrlich, S.D., Levenez, F., Pelletier, E., Alberti, A., Bertrand, L., Bork, P., Costea, P.I., Sunagawa, S., Guarner, F., Manichanh, C., Santiago, A., Zhao, L., Shen, J., Zhang, C., Versalovic, J., Luna, R.A., Petrosino, J., Yang, H., Li, S., Wang, J., Allen-Vercoe, E., Gloor, G., Singh, B. and IHMS Consortium (2015). IHMS_SOP 07 V1: Standard operating procedure for fecal samples DNA extraction, Protocol H. International Human Microbiome Standards. http://www.microbiome-standards.org

Eckburg, P. B., Bik, E. M., Bernstein, C. N., Purdom, E., Dethlefsen, L., Sargent, M., et al. (2005). Diversity of the Human Intestinal Microbial Flora. *Science* 308, 1635–1638. doi: 10.1126/science.1110591.

Ewels, P., Magnusson, M., Lundin, S., and Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* 32, 3047–3048. doi: 10.1093/bioinformatics/btw354.

Federici, S., Kredo-Russo, S., Valdés-Mas, R., Kviatcovsky, D., Weinstock, E., Matiuhin, Y., et al. (2022). Targeted suppression of human IBD-associated gut microbiota commensals by phage consortia for treatment of intestinal inflammation. *Cell* 185, 2879-2898.e24. doi: 10.1016/j.cell.2022.07.003.

Feng, Q., Liang, S., Jia, H., Stadlmayr, A., Tang, L., Lan, Z., et al. (2015). Gut microbiome development along the colorectal adenoma-carcinoma sequence. *Nat. Commun.* 6, 6528. doi: 10.1038/ncomms7528.

Forbes, J. D., Chen, C., Knox, N. C., Marrie, R.-A., El-Gabalawy, H., de Kievit, T., et al. (2018). A comparative study of the gut microbiota in immune-mediated inflammatory diseases—does a common dysbiosis exist? *Microbiome* 6, 221. doi: 10.1186/s40168-018-0603-4.

Forster, S. C., Kumar, N., Anonye, B. O., Almeida, A., Viciani, E., Stares, M. D., et al. (2019). A human gut bacterial genome and culture collection for improved metagenomic analyses. *Nat. Biotechnol.* 37, 186. doi: 10.1038/s41587-018-0009-7.

Franzosa, E. A., Morgan, X. C., Segata, N., Waldron, L., Reyes, J., Earl, A. M., et al. (2014). Relating the metatranscriptome and metagenome of the human gut. *Proc. Natl. Acad. Sci. U. S. A.* 111, E2329–E2338. doi: 10.1073/pnas.1319284111.

Franzosa, E. A., Sirota-Madi, A., Avila-Pacheco, J., Fornelos, N., Haiser, H. J., Reinker, S., et al. (2019). Gut microbiome structure and metabolic activity in inflammatory bowel disease. *Nat. Microbiol.* 4, 293. doi: 10.1038/s41564-018-0306-4.

Gevers, D., Kugathasan, S., Denson, L. A., Vázquez-Baeza, Y., Van Treuren, W., Ren, B., et al. (2014). The treatment-naïve microbiome in new-onset Crohn's disease. *Cell Host Microbe* 15, 382–392. doi: 10.1016/j.chom.2014.02.005.

Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V., and Egozcue, J. J. (2017). Microbiome Datasets Are Compositional: And This Is Not Optional. *Front. Microbiol.* 8, 2224. doi: 10.3389/fmicb.2017.02224.

Gloor, G. B., and Reid, G. (2016). Compositional analysis: a valid approach to analyze microbiome high-throughput sequencing data. *Can. J. Microbiol.* 62, 692–703. doi: 10.1139/cjm-2015-0821.

Goslee, S. C., and Urban, D. L. (2007). The ecodist Package for Dissimilarity-based Analysis of Ecological Data. *J. Stat. Softw.* 22, 1–19. doi: 10.18637/jss.v022.i07.

Guzzo, G. L., Mittinty, M. N., Llamas, B., Andrews, J. M., and Weyrich, L. S. (2022). Individuals with Inflammatory Bowel Disease Have an Altered Gut Microbiome Composition of Fungi and Protozoa. *Microorganisms* 10, 1910. doi: 10.3390/microorganisms10101910.

Gweon, H. S., Shaw, L. P., Swann, J., De Maio, N., AbuOun, M., Niehus, R., et al. (2019). The impact of sequencing depth on the inferred taxonomic composition and AMR gene content of metagenomic samples. Environmental Microbiome 14, 7. doi: 10.1186/s40793-019-0347-1.

Hall, A. B., Yassour, M., Sauk, J., Garner, A., Jiang, X., Arthur, T., et al. (2017). A novel Ruminococcus gnavus clade enriched in inflammatory bowel disease patients. *Genome Med.* 9, 103. doi: 10.1186/s13073-017-0490-5.

Hannigan, G. D., Duhaime, M. B., Ruffin, M. T., Koumpouras, C. C., and Schloss, P. D. (2018). Diagnostic Potential and Interactive Dynamics of the Colorectal Cancer Virome. *mBio* 9. doi: 10.1128/mBio.02248-18.

Hillman, E. T., Lu, H., Yao, T., and Nakatsu, C. H. (2017). Microbial Ecology along the Gastrointestinal Tract. *Microbes Environ.* 32, 300–313. doi: 10.1264/jsme2.ME17017.

Himmel, M. E., Hardenberg, G., Piccirillo, C. A., Steiner, T. S., and Levings, M. K. (2008). The role of T-regulatory cells and Toll-like receptors in the pathogenesis of human inflammatory bowel disease. *Immunology* 125, 145–153. doi: 10.1111/j.1365-2567.2008.02939.x.

Hoffmann, C., Dollive, S., Grunberg, S., Chen, J., Li, H., Wu, G. D., et al. (2013). Archaea and Fungi of the Human Gut Microbiome: Correlations with Diet and Bacterial Residents. *PLoS ONE* 8, e66019. doi: 10.1371/journal.pone.0066019.

Hold, G. L., Smith, M., Grange, C., Watt, E. R., El-Omar, E. M., and Mukhopadhya, I. (2014). Role of the gut microbiota in inflammatory bowel disease pathogenesis: what have we learnt in the past 10 years? *World J. Gastroenterol.* 20, 1192–1210. doi: 10.3748/wjg.v20.i5.1192.

Human Microbiome Project Consortium (2012a). A framework for human microbiome research. *Nature*

486, 215–221. doi: 10.1038/nature11209.

Human Microbiome Project Consortium (2012b). Structure, function and diversity of the healthy human microbiome. *Nature* 486, 207–214. doi: 10.1038/nature11234.

Huson, D. H., Auch, A. F., Qi, J., and Schuster, S. C. (2007). MEGAN analysis of metagenomic data. *Genome Res.* 17, 377–386. doi: 10.1101/gr.5969107.

Jiang, P., Lai, S., Wu, S., Zhao, X.-M., and Chen, W.-H. (2020). Host DNA contents in fecal metagenomics as a biomarker for intestinal diseases and effective treatment. *BMC Genomics* 21. doi: 10.1186/s12864-020-6749-z.

Kaplan, G. G., Bernstein, C. N., Coward, S., Bitton, A., Murthy, S. K., Nguyen, G. C., et al. (2019). The Impact of Inflammatory Bowel Disease in Canada 2018: Epidemiology. *J. Can. Assoc. Gastroenterol.* 2, S6–S16. doi: 10.1093/jcag/gwy054.

Kim, H., Sitarik, A. R., Woodcroft, K., Johnson, C. C., and Zoratti, E. (2019). Birth Mode, Breastfeeding, Pet Exposure, and Antibiotic Use: Associations With the Gut Microbiome and Sensitization in Children. *Curr. Allergy Asthma Rep.* 19, 22. doi: 10.1007/s11882-019-0851-9.

Knox, N. C., Forbes, J. D., Peterson, C.-L., Van Domselaar, G., and Bernstein, C. N. (2019a). The Gut Microbiome in Inflammatory Bowel Disease: Lessons Learned From Other Immune-Mediated Inflammatory Diseases. *Am. J. Gastroenterol.* 114, 1051. doi: 10.14309/ajg.0000000000000305.

Knox, N. C., Forbes, J. D., Van Domselaar, G., and Bernstein, C. N. (2019b). The Gut Microbiome as a Target for IBD Treatment: Are We There Yet? *Curr. Treat. Options Gastroenterol.* 17, 115–126. doi: 10.1007/s11938-019-00221-w.

Koorakula, R., Ghanbari, M., Schiavinato, M., Wegl, G., Dohm, J. C., and Domig, K. J. (2022). Storage media and RNA extraction approaches substantially influence the recovery and integrity of livestock fecal microbial RNA. *PeerJ* 10, e13547. doi: 10.7717/peerj.13547.

Kopylova, E., Noé, L., and Touzet, H. (2012). SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics* 28, 3211–3217. doi: 10.1093/bioinformatics/bts611.

Lagier, J.-C., Dubourg, G., Million, M., Cadoret, F., Bilen, M., Fenollar, F., et al. (2018). Culturing the human microbiota and culturomics. *Nat. Rev. Microbiol.* 16, 540–550. doi: 10.1038/s41579-018-0041-0.

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–359. doi: 10.1038/nmeth.1923.

Lee, J.-Y., Tsolis, R. M., and Bäumler, A. J. (2022). The microbiome and gut homeostasis. *Science* 377, eabp9960. doi: 10.1126/science.abp9960.

Li, D., Liu, C.-M., Luo, R., Sadakane, K., and Lam, T.-W. (2015). MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* 31, 1674–1676. doi: 10.1093/bioinformatics/btv033.

Lim, M. Y., Song, E.-J., Kim, S. H., Lee, J., and Nam, Y.-D. (2018). Comparison of DNA extraction methods for human gut microbial community profiling. *Syst. Appl. Microbiol.* 41, 151–157. doi: 10.1016/j.syapm.2017.11.008.

Lloyd-Price, J., Arze, C., Ananthakrishnan, A. N., Schirmer, M., Avila-Pacheco, J., Poon, T. W., et al. (2019). Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* 569, 655–662. doi: 10.1038/s41586-019-1237-9.

Macklaim, J. M., and Gloor, G. B. (2018). "From RNA-seq to Biological Inference: Using Compositional Data Analysis in Meta-Transcriptomics," in *Microbiome Analysis: Methods and Protocols* Methods in Molecular Biology., eds. R. G. Beiko, W. Hsiao, and J. Parkinson (New York, NY: Springer), 193–213. doi: 10.1007/978-1-4939-8728-3_13.

Magne, F., Gotteland, M., Gauthier, L., Zazueta, A., Pesoa, S., Navarrete, P., et al. (2020). The Firmicutes/Bacteroidetes Ratio: A Relevant Marker of Gut Dysbiosis in Obese Patients? *Nutrients* 12, 1474. doi: 10.3390/nu12051474.

Malele, I., Nyingilili, H., Lyaruu, E., Tauzin, M., Bernard Ollivier, B., Cayol, J.-L., et al. (2018). Bacterial diversity obtained by culturable approaches in the gut of Glossina pallidipes population from a non sleeping sickness focus in Tanzania: preliminary results. *BMC Microbiol.* 18, 164. doi: 10.1186/s12866-018-1288-3.

Mallick, H., Rahnavard, A., McIver, L. J., Ma, S., Zhang, Y., Nguyen, L. H., et al. (2021). Multivariable association discovery in population-scale meta-omics studies. *PLOS Comput. Biol.* 17, e1009442. doi: 10.1371/journal.pcbi.1009442.

Mandal, R., Cano, R., Davis, C. D., Hayashi, D., Jackson, S. A., Jones, C. M., et al. (2020). Workshop report: Toward the development of a human whole stool reference material for metabolomic and metagenomic gut microbiome measurements. *Metabolomics* 16, 119. doi: 10.1007/s11306-020-01744-5.

Mar Rodríguez, M., Pérez, D., Javier Chaves, F., Esteve, E., Marin-Garcia, P., Xifra, G., et al. (2015). Obesity changes the human gut mycobiome. *Sci. Rep.* 5, 14600. doi: 10.1038/srep14600.

Mariat, D., Firmesse, O., Levenez, F., Guimarăes, V., Sokol, H., Doré, J., et al. (2009). The Firmicutes/Bacteroidetes ratio of the human microbiota changes with age. *BMC Microbiol.* 9, 123. doi: 10.1186/1471-2180-9-123.

Martinez-Guryn, K., Leone, V., and Chang, E. B. (2019). Regional Diversity of the Gastrointestinal Microbiome. *Cell Host Microbe* 26, 314–324. doi: 10.1016/j.chom.2019.08.011.

Matsuoka, K., and Kanai, T. (2015). The gut microbiota and inflammatory bowel disease. *Semin. Immunopathol.* 37, 47–55. doi: 10.1007/s00281-014-0454-4.

McMurdie, P. J., and Holmes, S. (2014). Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible. *PLOS Comput. Biol.* 10, e1003531. doi: 10.1371/journal.pcbi.1003531.

Mehta, R. S., Abu-Ali, G. S., Drew, D. A., Lloyd-Price, J., Subramanian, A., Lochhead, P., et al. (2018). Stability of the human faecal microbiome in a cohort of adult men. *Nat. Microbiol.* 3, 347–355. doi: 10.1038/s41564-017-0096-0.

Méndez-García, C., Barbas, C., Ferrer, M., and Rojo, D. (2018). Complementary Methodologies To Investigate Human Gut Microbiota in Host Health, Working towards Integrative Systems Biology. *J. Bacteriol.* 200. doi: 10.1128/JB.00376-17.

Moons, K. G. M., Groot, J. A. H. de, Bouwmeester, W., Vergouwe, Y., Mallett, S., Altman, D. G., et al. (2014). Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies: The CHARMS Checklist. *PLOS Med.* 11, e1001744. doi: 10.1371/journal.pmed.1001744.

Nearing, J. T., Douglas, G. M., Hayes, M. G., MacDonald, J., Desai, D. K., Allward, N., et al. (2022). Microbiome differential abundance methods produce different results across 38 datasets. *Nat. Commun.* 13, 342. doi: 10.1038/s41467-022-28034-z.

Nielsen, C. C., Gascon, M., Osornio-Vargas, A. R., Shier, C., Guttman, D. S., Becker, A. B., et al. (2020). Natural environments in the urban context and gut microbiota in infants. *Environ. Int.* 142, 105881. doi: 10.1016/j.envint.2020.105881.

Ocvirk, S., Wilson, A. S., Posma, J. M., Li, J. V., Koller, K. R., Day, G. M., et al. (2020). A prospective cohort analysis of gut microbial co-metabolism in Alaska Native and rural African people at high and low risk of colorectal cancer. *Am. J. Clin. Nutr.* 111, 406–419. doi: 10.1093/ajcn/nqz301.

O'Hara, A. M., and Shanahan, F. (2006). The gut flora as a forgotten organ. *EMBO Rep.* 7, 688–693. doi: 10.1038/sj.embor.7400731.

Ohkusa, T., Okayasu, I., Ogihara, T., Morita, K., Ogawa, M., and Sato, N. (2003). Induction of experimental ulcerative colitis by Fusobacterium varium isolated from colonic mucosa of patients with ulcerative colitis. *Gut* 52, 79–83.

Oksanen, J., F. Guillaume Blanchet, Michael Friendly, Roeland Kindt, Pierre Legendre, Dan McGlinn, et al. (2020). vegan: Community Ecology Package. R package version 2.5-7. Available at: http://CRAN.R-project.org/package=vegan.

Ounit, R., Wanamaker, S., Close, T. J., and Lonardi, S. (2015). CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. *BMC Genomics* 16, 236. doi: 10.1186/s12864-015-1419-2.

Palarea-Albaladejo, J., and Martín-Fernández, J. A. (2015). zCompositions — R package for multivariate imputation of left-censored data under a compositional approach. *Chemom. Intell. Lab. Syst.* 143, 85–96. doi: 10.1016/j.chemolab.2015.02.019.

Patrick, D. M., Sbihi, H., Dai, D. L. Y., Mamun, A. A., Rasali, D., Rose, C., et al. (2020). Decreasing antibiotic use, the gut microbiota, and asthma incidence in children: evidence from population-based and prospective cohort studies. *Lancet Respir. Med.* 8, 1094–1105. doi: 10.1016/S2213-2600(20)30052-7.

Peloquin, J. M., and Nguyen, D. D. (2013). The Microbiota and Inflammatory Bowel Disease: Insights from Animal Models. *Anaerobe* 24. doi: 10.1016/j.anaerobe.2013.04.006.

Petrova, O. E., Garcia-Alcalde, F., Zampaloni, C., and Sauer, K. (2017). Comparative evaluation of rRNA depletion procedures for the improved analysis of bacterial biofilm and mixed pathogen

culture transcriptomes. *Sci. Rep.* 7, 41114. doi: 10.1038/srep41114.

Pielou, E. C. (1966). The measurement of diversity in different types of biological collections. *J. Theor. Biol.* 13, 131–144. doi: 10.1016/0022-5193(66)90013-0.

Pratt, M., Forbes, J. D., Knox, N. C., Bernstein, C. N., and Van Domselaar, G. (2021). Microbiome-Mediated Immune Signaling in Inflammatory Bowel Disease and Colorectal Cancer: Support From Meta-omics Data. *Front. Cell Dev. Biol.* 9, 3288. doi: 10.3389/fcell.2021.716604.

Pratt, M., Forbes, J. D., Knox, N. C., Van Domselaar, G., and Bernstein, C. N. (2022). Colorectal Cancer Screening in Inflammatory Bowel Diseases—Can Characterization of GI Microbiome Signatures Enhance Neoplasia Detection? *Gastroenterology* 162, 1409-1423.e1. doi: 10.1053/j.gastro.2021.12.287.

Proctor, L. M., Creasy, H. H., Fettweis, J. M., Lloyd-Price, J., Mahurkar, A., Zhou, W., et al. (2019). The Integrative Human Microbiome Project. *Nature* 569, 641–648. doi: 10.1038/s41586-019-1238-8.

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., et al. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 464, 59–65. doi: 10.1038/nature08821.

Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., et al. (2012). A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* 490, 55–60. doi: 10.1038/nature11450.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., et al. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* 41, D590–D596. doi: 10.1093/nar/gks1219.

R Core Team (2021). R: A Language and Environment for Statistical Computing. Available at: https://www.R-project.org/.

Rajan, S. K., Lindqvist, M., Brummer, R. J., Schoultz, I., and Repsilber, D. (2019). Phylogenetic microbiota profiling in fecal samples depends on combination of sequencing depth and choice of NGS analysis method. *PLoS ONE* 14, e0222171. doi: 10.1371/journal.pone.0222171.

Reck, M., Tomasch, J., Deng, Z., Jarek, M., Husemann, P., and Wagner-Döbler, I. (2015). Stool metatranscriptomics: A technical guideline for mRNA stabilisation and isolation. *BMC Genomics* 16. doi: 10.1186/s12864-015-1694-y.

Rehman, A., Lepage, P., Nolte, A., Hellmig, S., Schreiber, S., and Ott, S. J. (2010). Transcriptional activity of the dominant gut mucosal microbiota in chronic inflammatory bowel disease patients. *J. Med. Microbiol.* 59, 1114–1122. doi: 10.1099/jmm.0.021170-0.

Santiago, A., Panda, S., Mengels, G., Martinez, X., Azpiroz, F., Dore, J., et al. (2014). Processing faecal samples: a step forward for standards in microbial community analysis. *BMC Microbiol.* 14, 112. doi: 10.1186/1471-2180-14-112.

Schirmer, M., Franzosa, E. A., Lloyd-Price, J., McIver, L. J., Schwager, R., Poon, T. W., et al. (2018). Dynamics of metatranscription in the inflammatory bowel disease gut microbiome. *Nat.*

Microbiol. 3, 337–346. doi: 10.1038/s41564-017-0089-z.

Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O., and Huttenhower, C. (2012). Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat. Methods* 9, 811–814. doi: 10.1038/nmeth.2066.

Sellon, R. K., Tonkonogy, S., Schultz, M., Dieleman, L. A., Grenther, W., Balish, E., et al. (1998). Resident Enteric Bacteria Are Necessary for Development of Spontaneous Colitis and Immune System Activation in Interleukin-10-Deficient Mice. *Infect. Immun.* 66, 5224–5231. doi: 10.1128/IAI.66.11.5224-5231.1998.

Shanahan, F. (2002). The host–microbe interface within the gut. *Best Pract. Res. Clin. Gastroenterol.* 16, 915–931. doi: 10.1053/bega.2002.0342.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Syst. Tech. J.* 27, 379–423. doi: 10.1002/j.1538-7305.1948.tb01338.x.

Shapiro, H., Goldenberg, K., Ratiner, K., and Elinav, E. (2022). Smoking-induced microbial dysbiosis in health and disease. *Clin. Sci.* 136, 1371–1387. doi: 10.1042/CS20220175.

Shen, C.-H. (2019). "Chapter 6 - Extraction and Purification of Nucleic Acids and Proteins," in Diagnostic Molecular Biology, ed. C.-H. Shen (Academic Press), 143–166. doi: 10.1016/B978-0-12-802823-0.00006-7.

Skoufos, G., Almodaresi, F., Zakeri, M., Paulson, J. N., Patro, R., Hatzigeorgiou, A. G., et al. (2022). AGAMEMNON: an Accurate metaGenomics And MEtatranscriptoMics quaNtificatiON analysis suite. *Genome Biol.* 23, 39. doi: 10.1186/s13059-022-02610-4.

Stojanov, S., Berlec, A., and Štrukelj, B. (2020). The Influence of Probiotics on the Firmicutes/Bacteroidetes Ratio in the Treatment of Obesity and Inflammatory Bowel disease. *Microorganisms* 8, E1715. doi: 10.3390/microorganisms8111715.

Szamosi, J. C., Forbes, J. D., Copeland, J. K., Knox, N. C., Shekarriz, S., Rossi, L., et al. (2020). Assessment of Inter-Laboratory Variation in the Characterization and Analysis of the Mucosal Microbiota in Crohn's Disease and Ulcerative Colitis. *Front. Microbiol.* 11. doi: 10.3389/fmicb.2020.02028.

Tamames, J., Cobo-Simón, M., and Puente-Sánchez, F. (2019). Assessing the performance of different approaches for functional and taxonomic annotation of metagenomes. *BMC Genomics* 20, 960. doi: 10.1186/s12864-019-6289-6.

Ternes, D., Karta, J., Tsenkova, M., Wilmes, P., Haan, S., and Letellier, E. (2020). Microbiome in Colorectal Cancer: How to Get from Meta-omics to Mechanism? Trends in Microbiology 28, 401–423. doi: 10.1016/j.tim.2020.01.001.

Thaiss, C. A., Zmora, N., Levy, M., and Elinav, E. (2016). The microbiome and innate immunity. *Nature* 535, 65–74. doi: 10.1038/nature18847.

The Integrative HMP (iHMP) Research Network Consortium (2014). The Integrative Human Microbiome Project: Dynamic Analysis of Microbiome-Host Omics Profiles during Periods of

Human Health and Disease. *Cell Host Microbe* 16, 276–289. doi: 10.1016/j.chom.2014.08.014.

Tourlousse, D. M., Narita, K., Miura, T., Ohashi, A., Matsuda, M., Ohyama, Y., et al. (2022). Characterization and Demonstration of Mock Communities as Control Reagents for Accurate Human Microbiome Community Measurements. *Microbiol. Spectr.* 10, e01915-21. doi: 10.1128/spectrum.01915-21.

Tourlousse, D. M., Narita, K., Miura, T., Sakamoto, M., Ohashi, A., Shiina, K., et al. (2021). Validation and standardization of DNA extraction and library construction methods for metagenomics-based human fecal microbiome measurements. *Microbiome* 9. doi: 10.1186/s40168-021-01048-3.

Ungaro, R., Bernstein, C. N., Gearry, R., Hviid, A., Kolho, K.-L., Kronman, M. P., et al. (2014). Antibiotics associated with increased risk of new-onset Crohn's disease but not ulcerative colitis: a meta-analysis. *Am. J. Gastroenterol.* 109, 1728–1738. doi: 10.1038/ajg.2014.246.

Walker, W. A. (2017). "Chapter 25 - Dysbiosis," in *The Microbiota in Gastrointestinal Pathophysiology*, eds. M. H. Floch, Y. Ringel, and W. Allan Walker (Boston: Academic Press), 227–232. doi: 10.1016/B978-0-12-804024-9.00025-2.

Wesolowska-Andersen, A., Bahl, M. I., Carvalho, V., Kristiansen, K., Sicheritz-Pontén, T., Gupta, R., et al. (2014). Choice of bacterial DNA extraction method from fecal material influences community structure as evaluated by metagenomic analysis. *Microbiome* 2, 19. doi: 10.1186/2049-2618-2-19.

Whittaker, R. H. (1960). Vegetation of the Siskiyou Mountains, Oregon and California. Ecological Monographs 30, 407–407. doi: 10.2307/1948435.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York Available at: https://ggplot2.tidyverse.org/ [Accessed September 20, 2022].

Wilmes, P., Martin-Gallausiaux, C., Ostaszewski, M., Aho, V. T. E., Novikova, P. V., Laczny, C. C., et al. (2022). The gut microbiome molecular complex in human health and disease. *Cell Host Microbe* 30, 1201–1206. doi: 10.1016/j.chom.2022.08.016.

Wong, S. H., Zhao, L., Zhang, X., Nakatsu, G., Han, J., Xu, W., et al. (2017). Gavage of Fecal Samples From Patients With Colorectal Cancer Promotes Intestinal Carcinogenesis in Germ-Free and Conventional Mice. *Gastroenterology* 153, 1621-1633.e6. doi: 10.1053/j.gastro.2017.08.022.

Wood, D. E., and Salzberg, S. L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* 15, R46. doi: 10.1186/gb-2014-15-3-r46.

Wu, W.-K., Chen, C.-C., Panyod, S., Chen, R.-A., Wu, M.-S., Sheen, L.-Y., et al. (2019). Optimization of fecal sample processing for microbiome study — The journey from bathroom to bench. *J. Formos. Med. Assoc.* 118, 545–555. doi: 10.1016/j.jfma.2018.02.005.

Yachida, S., Mizutani, S., Shiroma, H., Shiba, S., Nakajima, T., Sakamoto, T., et al. (2019). Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer. *Nat. Med.* 25, 968–976. doi: 10.1038/s41591-019-0458-7.

Yang, H., Mirsepasi-Lauridsen, H. C., Struve, C., Allaire, J. M., Sivignon, A., Vogl, W., et al. (2020). Ulcerative Colitis-associated E. coli pathobionts potentiate colitis in susceptible hosts. *Gut Microbes* 12, 1847976. doi: 10.1080/19490976.2020.1847976.

Yang, Y., and Jobin, C. (2017). Novel insights into microbiome in colitis and colorectal cancer. *Curr. Opin. Gastroenterol.* 33, 422–427. doi: 10.1097/MOG.0000000000000399.

Yatsunenko, T., Rey, F. E., Manary, M. J., Trehan, I., Dominguez-Bello, M. G., Contreras, M., et al. (2012). Human gut microbiome viewed across age and geography. *Nature* 486, 222–227. doi: 10.1038/nature11053.

Zeller, G., Tap, J., Voigt, A. Y., Sunagawa, S., Kultima, J. R., Costea, P. I., et al. (2014). Potential of fecal microbiota for early-stage detection of colorectal cancer. *Mol. Syst. Biol.* 10, 766. doi: 10.15252/msb.20145645.

Zhang, X., Zhu, X., Cao, Y., Fang, J., Hong, J., and Chen, H. (2019). Fecal Fusobacterium nucleatum for the diagnosis of colorectal tumor: A systematic review and meta-analysis. *Cancer Med.* 8, 480–491. doi: 10.1002/cam4.1850.

Zhang, Y., Thompson, K. N., Branck, T., Yan Yan, Nguyen, L. H., Franzosa, E. A., et al. (2021a). Metatranscriptomics for the Human Microbiome and Microbial Community Functional Profiling. *Annu. Rev. Biomed. Data Sci.* doi: 10.1146/annurev-biodatasci-031121-103035.

Zhang, Y., Thompson, K. N., Huttenhower, C., and Franzosa, E. A. (2021b). Statistical approaches for differential expression analysis in metatranscriptomics. *Bioinformatics* 37, i34–i41. doi: 10.1093/bioinformatics/btab327.

Zoetendal, E. G., Booijink, C. C., Klaassens, E. S., Heilig, H. G., Kleerebezem, M., Smidt, H., et al. (2006). Isolation of RNA from bacterial samples of the human gastrointestinal tract. *Nat. Protoc.* 1, 954–959. doi: 10.1038/nprot.2006.143.

APPENDIX

a)

Balanced DNA Standard			Log Distribution DNA Standard		
Taxa	**Expected**	**Actual**	**Taxa**	**Expected**	**Actual**
Listeria monocytogenes	12%	9.71%	*Listeria monocytogenes*	89.1%	96.24%
Pseudomonas aeruginosa	12%	2.88%	*Pseudomonas aeruginosa*	8.9%	2.20%
Bacillus subtilis *	12%	15.20%	*Bacillus subtilis* *	0.89%	1.18%
Escherichia coli	12%	14.35%	*Saccharomyces cerevisiae*	0.89%	0.28%
Salmonella enterica	12%	12.63%	*Escherichia coli*	0.089%	0.02%
Lactobacillus fermentum	12%	21.05%	*Salmonella enterica*	0.089%	0.08%
Enterococcus faecalis	12%	14.72%	*Lactobacillus fermentum*	0.0089%	unidentified
Staphylococcus aureus	12%	8.52%	*Enterococcus faecalis*	0.00089%	unidentified
Saccharomyces cerevisiae	2%	0.66%	*Cryptococcus neoformans*	0.00089%	unidentified
Cryptococcus neoformans	2%	0.21%	*Staphylococcus aureus*	0.000089%	unidentified
Additional Taxa	0	1	Additional Taxa	0	0
Total	100%	99.93%	**Total**	99.97%	100%

b)

Taxa	**Expected**	**Detection at 0.1**	**0.05**	**0**
Listeria monocytogenes	89.1%	95.85%	94.22%	93.87%
Pseudomonas aeruginosa	8.9%	2.51%	2.70%	2.67%
Bacillus subtilis *	0.89%	1.18%	1.18%	1.15%
Saccharomyces cerevisiae	0.89%	0.31%	0.33%	0.33%
Escherichia coli	0.089%	0.05%	0.05%	0.06%
Salmonella enterica	0.089%	0.10%	0.12%	0.13%
Lactobacillus fermentum	0.0089%	unidentified	unidentified	0.01%
Enterococcus faecalis	0.00089%	unidentified	unidentified	unidentified
Cryptococcus neoformans	0.00089%	unidentified	unidentified	unidentified
Staphylococcus aureus	0.000089%	unidentified	unidentified	unidentified
Additional Taxa	0	0	2	20
Total	99.97%	99.99%	98.61%	98.22%

Figure S1. Comparison of ZymoBIOMICS Mock Community DNA Standards. Species-level taxonomic profiles of balanced and log-distributed mock community DNA standards were used to evaluate optimal MetaPhlAn stringency parameters via the stat_q value. The default stat_q value (= 0.2) detected only one false positive species in the balanced standard (a, left) and no false positives in the log distribution standard (a, right). The calculated FDR is 0.048. Different stat_q values were assessed in (b) to determine whether low-abundance taxa could be detected in the log distribution community. Lower values of stat_q were found to increase the rate of false positive assignments without significantly improving low-abundance detection. **Bacillus subtilis* was identified by MetaPhlan as either *B. subtilis* or *B. intestinalis*.

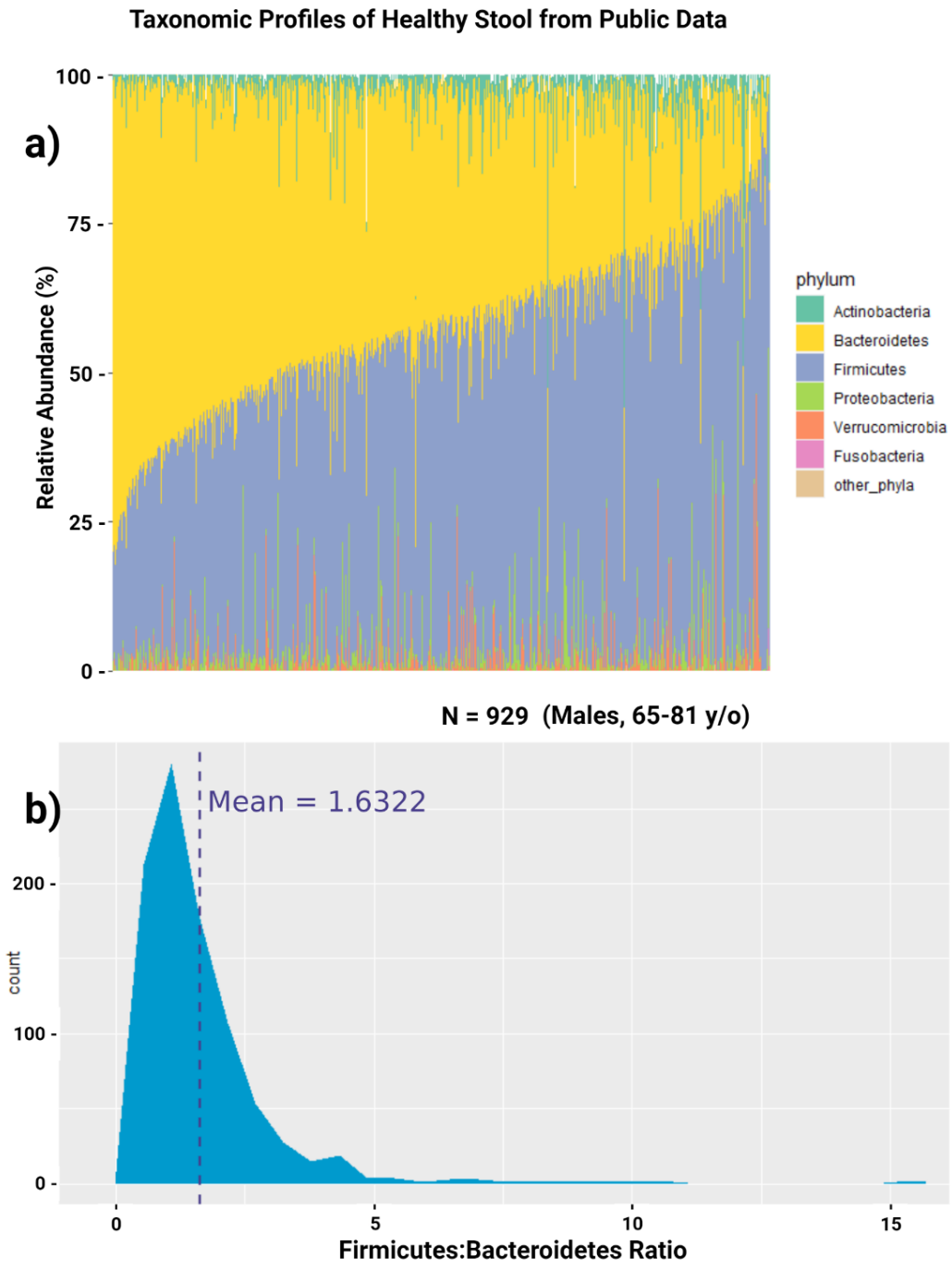
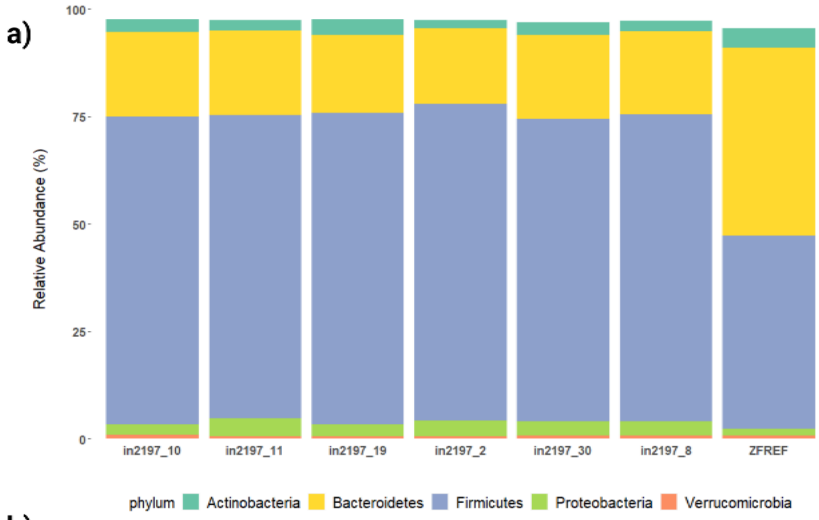


Figure S2. Metagenomic Comparison with Published Data. Taxonomic phylum-level profiles from healthy human stool, published in Abu-Ali et al., 2018 are shown in (a). The F/B ratio of all samples (b) was calculated and used for comparison with data from the current study; taking into consideration the age of participants as well as the large sample size, a broad range of 0-10 for healthy individuals was established for the current study, corresponding to Firmicutes abundance in the range of ~20-90% and Bacteroidetes abundance in the range of ~10-80%.



c)

Sample Name	Total Read Count	Species Richness
in2197_19	19.9 M	66
in2197_11	20.7 M	69
in2197_2	21.1 M	66
ZFREF	**21.1 M**	**80**
in2197_8	25.0 M	73
in2197_10	26.5 M	74
in2197_30	98.7 M	108

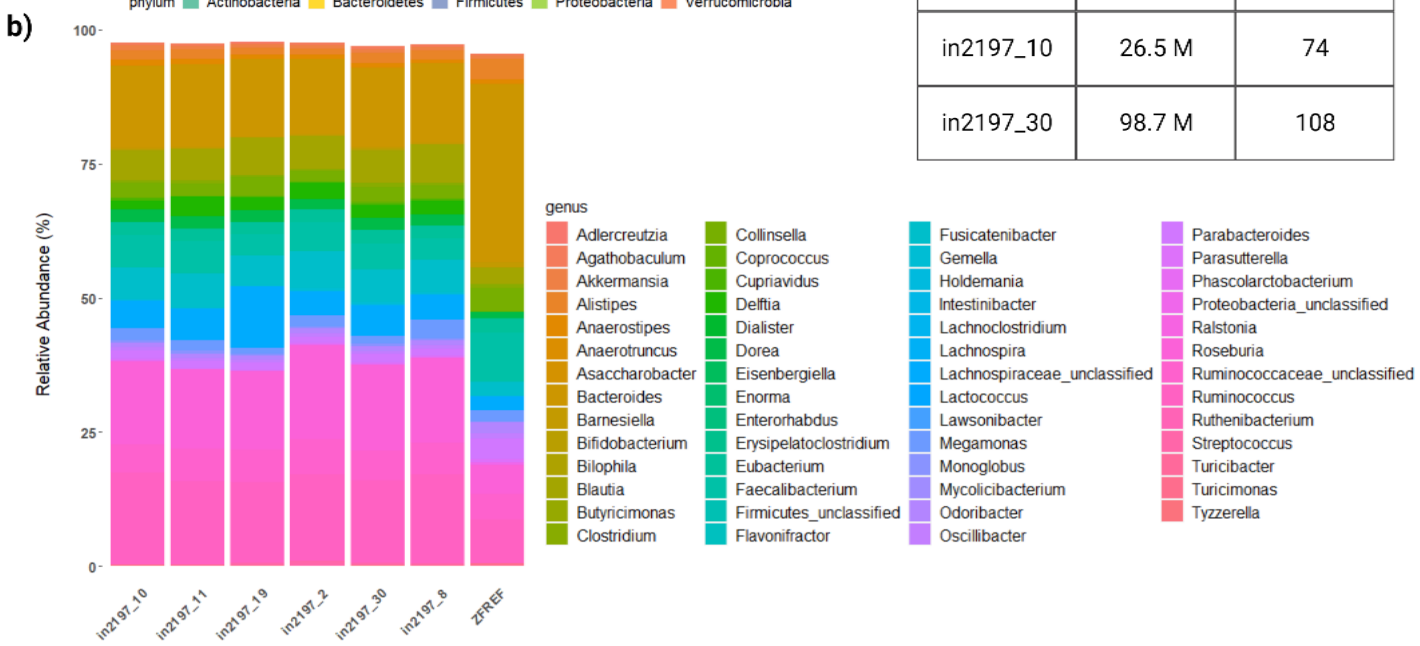


Figure S3. Fecal Reference Comparison. Paired-end metatranscriptomic sequence data from 6 samples (in2197_#) were downloaded from the ZymoBIOMICS™ Fecal Reference Database. Taxonomic profiles from MetaPhlAn are compared to the Fecal Reference sample prepared for this study (ZFREF) at the phylum (a) and genus (b) level. Separate extraction of the ZFREF sample highlights how differing approaches can impact relative abundance profiles. Species richness (c) of each sample corresponds with sample read depth; the ZFREF sample from this study captured more diversity than public data with a similar read depth.

Table S1. Stool microbiome signature models for classifying and detecting CRC and colorectal adenomas (CRA). (*continued on next 2 pages)*

	Microbiome Signature(s) used in Model			Model Performance				
Discovery cohort (n)	**Enriched in CRC**	**Enriched in Controls**	**Detection Method**	**Validation Cohort (n)**	**Sensitivity (TPR)**	**Specificity (1-FPR)**	**AUC**	**Reference**
American CRC (30), colorectal adenoma (CRA, 20), and healthy controls (HC, 30)	Enterobacteriaceae, *Fusobacterium*, Porphyromonadaceae, *Porphyromonas*	*Bacteroides*, Lachnospiraceae	OTUs assembled from 16S rRNA sequencing	*NA* (cross-classification only)	Microbiome, age, race, & BMI: **90%**	Microbiome, age, race, & BMI: **83.3%**	Microbiome only: **0.798** Microbiome, age, race, & BMI: **0.922**	Zackular et al., 2014
French CRC (53), CRA (42) and HC (61)	*Fusobacterium nucleatum (*subspp. *vincentii, animalis)*, *Peptostreptococcus stomatis*, *Porphyromonas asaccharolytica*, *Clostridium symbiosum*, *Clostridium hylemonae*, *Bacteroides fragilis*, *Lactobacillus salivarius*, *Fusobacterium gonidiaformans*	*Bacteroides caccae*, *Bifidobacterium angulatum*, *Butyrvibrio crossotus*, *Clostridium scindens*, *Dorea formicigenerans*, *Eubacterium eligens*, *Eubacterium ventriosum*, *Eubacterium rectale*, *Lactobacillus ruminis*, *Methanosphaera stadtmanae*, *Phascolarcobacterium succinatutens*, *Streptococcus salivarius*, Unclassified *Ruminococcus* sp.	Whole-genome shotgun sequencing	German CRC (38) and European HC (297); European IBD patients (25)	**58.5%** + FOBT: **72%** Validation: **52%**	**92%** + FOBT: **92%** Validation: **92.3%** IBD: **76%**	**0.84** + FOBT: **0.87** Validation: **0.85**	Zeller et al., 2014
Austrian CRC (46), advanced CRA (47), and HC (63) Training set: CRC (41) and HC (55)	Metagenomic linkage group (MLG) CRC markers: *Bacteroides massiliensis*, *Dialister invisus*, *Fusobacterium sp. oral taxon 370*, *Gemella morbillorum*, *Prevotella copri*, +10 taxonomically unclassified MLGs	NA	Whole-genome shotgun sequencing	Testing set: CRC (5), CRA (47), HC (8)	Not reported	Not reported	Validation: **0.96**	Feng et al., 2015

Table S1. *Continued from previous page…*

Chinese cohort, C1: CRC (74) and controls (54)	*Microbial gene biomarker models*: 4-marker model: 2 transposases (*P. anerobius*), m1704941 butyryl-CoA dehydrogenase (*F. nucleatum*), rpoB (*P. micra*) 2-marker model: m1704941 butyryl-CoA, dehydrogenase (*F. nucleatum*), rpoB (*P. micra*)	The authors additionally validated a 20-marker gene model, including 8 microbial genes enriched in CRC and 12 enriched in controls.	Targeted qPCR assays	Chinese cohort, C2: CRC (47) and controls (109); Danish cohort, D: CRC (16) and controls (24); French cohort, F: CRC (53) and controls (61); Austrian cohort, A: CRC (41) and controls (55)	Validation: (model, cohort) 2-marker model, C2: **72.3%**	Validation: (model, cohort) 2-marker model, C2: **92.7%**	Validation: (model, cohort) 2-marker model, C2: **0.84** 4-marker model, A: **0.77** D: **0.72** 20-marker model, D: **0.71**	Yu et al., 2015
North American CRC (120), CRA (198), and controls (172)	Microbiota model: 34 OTUs, including *F. nucleatum*, *Gemella spp.*, *P. asaccharolytica*, *P. stomatis*, *Parvimonas micra*, *Prevotella*	An additional model combines 23 OTU markers with FIT. The majority of OTUs included in this model were enriched in controls (e.g., Lachnospiraceae).	OTUs assembled from 16S rRNA sequencing	*NA* (cross-classification only)	Microbiota model: **51.7%** + FIT (MMT): **91.7%**	Microbiota model: **97.1%** + FIT (MMT): **90.1%**	Microbiota model: **0.847** + FIT (MMT): **0.952**	Baxter et al., 2016
Hong Kong* CRC (170) and HC (200) *A subset of the Hong Kong cohort was used for FIT testing (111 CRC, 119 HC)	*Clostridium hathewayi*, *F. nucleatum*, Undefined species (m7)	*Bacteroides clarus*	Targeted qPCR assays	Shanghai CRC (33) and HC (36)	**83.8%** + FIT: **92.8%** Validation: **84.9%**	**83.2%** + FIT: **81.5%** Validation: **61.1%**	**0.886** + FIT: not reported Validation: **0.756**	Liang et al., 2017
Hong Kong CRC (104), advanced CRA (103), and HC (102)	*F. nucleatum*	NA	Targeted qPCR assays	CRC (23), adenoma (62), HC (96)	**72.1%** + FIT: **92.3%** Validation: **91.3%** + FIT: **82.6%**	**91%** + FIT: **93%** Validation: **80.2%** + FIT: **94.8%**	**0.83** + FIT: **0.95** Validation: **0.89** + FIT: **0.96**	Wong et al., 2017

Table S1. *Continued from previous page…*

Japanese CRC stage 0 (73), stage I/II (111), stage III/IV (74), HC (251)	*Bilophila wadsworthia*, *Colinsella aerofaciens*, *D. longreachensis*, *Desulfovibrio vietnamensis*, *Dorea longicatena*, *F. nucleatum*, *G. morbillorum*, *Lactobacillus sanfranciscensis*, *Porphyromonas uenonis*, *P. stomatis*, *P. anaerobius*, *P. micra*, *Selenomonas sputigena*, *Streptococcus anginosus*, *S. moorei*	*Lachnospira multipara*, *Eubacterium eligens*	Whole-genome shotgun sequencing	*NA* (cross-classification only)	Not reported	Not reported	Early CRC detection: **0.73** Late-stage CRC detection: **0.83**	Yachida et al., 2019
Swedish FESCU cohort, colonoscopy patients: CRC (38), dysplasia (128), controls (63)	*clbA*+ bacteria, *P. micra*, *F. nucleatum*	*NA*	Targeted qPCR assays	Swedish U-CAN cohort: CRC patients (238) and controls (94)	**62.9%** + FIT: **73.9%** Validation: **56.8%**	**87.9%** + FIT: **85.7%** Validation: **87.9%**	Not reported	Löwenmark et al., 2020

Not all models from individual studies are included in this summary. AUC refers to the area under the receiver operating characteristics curve. FIT, fecal immunochemical testing. TPR, true positive rate. FPR, false positive rate. HC, healthy controls. OTU, operating taxonomic units. BMI, body mass index. qPCR, quantitative polymerase chain reaction. NA, not applicable. (Table from Pratt et al., 2022).

Table S2. CRC biomarkers identified from meta-analyses.

Parent studies	Commonly identified CRC markers (all studies)	Other CRC markers	Accuracy of CRC classification (AUC)	Ref.
Zeller et al., 2014, Feng et al., 2015, Yu et al., 2015, and Vogtmann et al., 2016	***F. nucleatum,*** ***P. micra,*** ***P. asaccharolytica***	*Alistipes finegoldii*, ***B. fragilis***, ***Prevotella intermedia***, *Thermanaerovibrio acidaminovorans*	**0.80**	Dai et al., 2018
Zeller et al., 2014, Feng et al., 2015, Yu et al., 2015, Vogtmann et al., 2016, Hannigan et al., 2018, and 2 new cohorts		*Anaerococcus vaginalis*, *A. obesiensis*, ***B. fragilis***, ***C. symbiosum***, *Granulicatella adiacens*, ***G. morbillorum***, ***Parvimonas* spp.**, *P. anaerobius*, ***P. intermedia***, *P. somerae*, ***P. stomatis***, ***P. uenonis***, *Streptococcus constellatus*, ***S. moorei***	**0.84**	Thomas et al., 2019
Zeller et al., 2014, Feng et al., 2015, Yu et al., 2015, Vogtmann et al., 2016, Thomas et al., 2019, and one new cohort		*A. obesiensis*, *A. vaginalis*, *C. bolteae*, ***C. symbiosum***, ***G. morbillorum***, *Hungatella hathewayi**, ***Parvimonas* spp.**, ***P. intermedia***, *P. nigrescens*, *P. somerae*, ***P. stomatis***, ***P. uenonis***, *Ruminococcus torques*, *Subdoligranulum* sp. ***S. moorei***, Unclassified *Dialister*, Unclassified *Porphyromonas*, Unclassified *Clostridiales*, Unclassified *Peptostreptococcaceae*, Unclassified Anaerotruncus,	**≥ 0.80**	Wirbel et al., 2019
Zeller et al., 2014, Feng et al., 2015, Yu et al., 2015, Vogtmann et al., 2016, Thomas et al., 2019, and Wirbel et al., 2019		***B. fragilis,*** *C. hathewayi*, ***C. symbiosum***, ***G. morbillorum***, *Lachnospiraceae bacterium 7 1 58FAA*, ***Parvimonas* spp.**, *Peptostreptococcus* spp. ***P. stomatis***, ***P. uenonis***, ***S. moorei***	**0.80**	Jiang et al., 2020

Bold text indicates markers that were identified in at least three of four meta-analyses. *Note: *H. hathewayi* is reported by some studies as *C. hathewayi*. (Table from Pratt et al., 2022).

Table S3. Meta-omic changes in the gut microbiome in CRC or IBD. *(Continued on next page)*

Cohort Type (N)	Sample type(s)	Method	Increased in disease	Decreased in disease	Reference
CRC					
CRC (31): colorectal tumour biopsy vs. normal tissue	Mucosal biopsy	**Metabolomics:** HR-MAS NMR and GC/MS	Choline-containing compounds, taurine, scyllo-inositol, lactate, phosphocholine, phosphate, L-glycine, 2-hydroxy-3-methyl valerate, L-proline, L-phenylalanine, palmitic acid, margaric acid, oleic acid, stearic acid, uridine, 11-eicosenoic acid, propyl octadecanoate, cholesterol	Lipids, polyethylene glycol, glucose, fumarate, malate, mannose, galactose, 1-hexadecanol, arachidonic acid	Chan *et al.*, 2009
CRC (11) vs. HC (10)	Stool, mucosal biopsy	**Metabolomics:** GC/ToFMS	Uracil, uridine, proline	Fructose, linoleic acid, nicotinic acid, glucose, galactose, 3-phosphoglycerate, citric acid, inosine, creatine	Phua *et al.*, 2014
Healthy Alaskan Natives (high-risk group for CRC) (32) vs. Healthy Rural Africans (low risk for CRC) (21)	Stool, urine	**Metataxonomics:** 16S rRNA sequencing **Metabolomics:** ^{1}H-NMR spectroscopy, GC, HPLC-MS	*Enriched in high-risk population:* Actinobacteria, Verrumicrobia, Lachnospiraceae, *Bifidobacterium spp.*, *Escherichia-Shigella spp.*, choline, formate, cholate, chenodeoxycholate, deoxycholate, conjugated bile acids, nicotinamide/niacin metabolites	*Enriched in low-risk population:* Ruminococcaceae, Prevotellaceae, *Prevotella 9*, Ruminococcaceae, *Succinivibrio*, *Eubacterium coprostanoligenes*, amino acids, purines[1], pyrimidines[1], butyrate, propionate, nicotinate (B3)	Ocvirk *et al.*, 2020
CRC (14) vs. HC (14)	Stool	**Metaproteomics:** LC-MS/MS	Desulfobacterales, *Methanobacteriaceae*, *Sporolactobacillaceae*, *Bacteroides fragilis*, *Peptostreptococcus anaerobius*, DNA replication, recombination, and repair proteins, iron intake and transport proteins, superoxide dismutases	*Sutterellaceae*, *Epulopiscium*, *Gordonibacter*, NADH:flavin oxidoreductases/NADH oxidases, energy production and conversion proteins, amino acid transport and metabolism, coenzyme transport and metabolism, lipid transport and metabolism, translation machinery, cell wall, membrane, and envelope biogenesis, cell motility, post-translational modification, protein turnover, and chaperones, inorganic ion transport and metabolism	Long *et al.*, 2020
IBD					
Identical twin pairs (N=17 pairs): discordant colonic CD*(4p); discordant ileal CD*(2p); concordant ICD*(2p); concordant CCD*(2p) vs. HC (7 pairs) *in remission	Stool	**Metabolomics:** ICR-FT/MS	Bile acid metabolism[2] (glycocholate, glycochenodeoxycholate taurocholate, Trihydroxy-6β-cholanate), amino acid metabolism[2], tyrosine[2], tryptophan(ICD only), phenylalanine (ICD only), fatty acid biosynthesis (ICD only; oleic acid, stearic acid, palmitic acid, linoleic acid, arachidonic acid), Urea cycle[1,2], vitamin B6 metabolism[2]	Arachidonic acid metabolism/prostaglandins[1,2] (PGs; PGF2a)	Jansson *et al.*, 2009
CD (83); UC (68); pouchitis (13) vs. HC (40)	Stool	**Metabolomics:** GC-MS	Styrene[3]	MCFAs, hexanoate[2], protein fermentation metabolites	De Preter *et al.*, 2015

Table S3. *Continued from previous page…*

Paediatric IBD in remission - CD (26); UC(10) vs. healthy 1st-degree relatives (54)	Stool	**Metataxonomics:** 16S rRNA sequencing **Metabolomics:** UPLC/ToFMS	Enterobacteriaceae, cholate[2], conjugated and sulphated bile acids[2], taurine[2], tryptophan[2], adrenate[2]	Stercobilin[2], acetyl-glutamic acid[2], boldione[2], estradiol[2], androstenedione[2], azelaic acid[2]	Jacobs *et al.*, 2016
Paediatric IBD (newly diagnosed, treatment naive)- CD (36); UC (20); IBD-U (13) vs. endoscopic non-IBD controls (29)	Stool, blood	**Metabolomics:** UPLC-MS/MS	Folate and pterine biosynthesis, purine metabolism[2], amino acid metabolism, nicotinate and nicotinamide metabolism[2], urea cycle, protein biosynthesis, bile acid biosynthesis[3], sphingolipid metabolism[3], ammonia recycling[3], taurine metabolism[3], oxidation of branched-chain FAs[3], phospholipid metabolism[3], glycerolipid metabolism[3]	L-tryptophan, kynurenic acid, aspartate, threonine, asparagine, cytosine, histidine[2], taurine[2]	Kolho *et al.*, 2017
Pediatric IBD (treatment naïve): CD (25); UC (22) vs. non-IBD (24)	Mucosal-luminal interface (MLI) biopsy	**Metaproteomics:** MS	DNA replication, recombination, and repair proteins, defence mechanism proteins (CRISPR/Cas), cell wall, membrane and envelope biogenesis proteins, amino acid transport and metabolism, mobilome, cysteine degradation, Proteobacteria, Verrucomicrobia, Ascomycota, Spirochetes, *Faecalibacterium prausnitzii* strain L2-6	Cysteine biosynthesis, *Bacteroides*	Zhang *et al.*, 2018
CD (68); UC (53) vs. non-IBD (34)	Stool	**Metagenomics:** WGS **Metabolomics:** LC-MS	Sphingolipids, carboximic acids[1], bile acids (cholate, chenodeoxycholate)[1], organonitrogen compounds, cholesteryl esters, phenylacetamides[2], phosphatidylcholines, α-amino acids, lactate	LCFAs, butyrate[4], propionate[4], secondary bile acids (lithocholate, deoxycholate) [4], flavonoids, indoles[1,2], cinnamic acids[1], triacylglycerols, tetrapyrroles[1,2], triterpenoids, alkyl-phenylketones, brassinolides[1,2], ergosterols[1], quinolines[1,2], vitamin D[1], stigmastanes[1], lactones, β-diketones[2], cholesterols[1], phenylbenzodioxanes, pantothenate (vitamin B5)	Franzosa *et al.*, 2019
CD (67); UC (38) vs. non-IBD (27)	Stool, colon biopsy, blood	**Metagenomics:** WGS **Metatranscriptomics** **Metabolomics:** LC-MS	Cholate[2], chenodeoxycholate[2], taurochenodeoxycholate[2], C8 carnitine[2], anti-ompc[2], calprotectin[3], adrenate, arachidonate, putrescine, taurine, *Escherichia coli*, *Klebsiella pneumoniae*, *Roseburia gnavus*[2]	Deoxycholate, lithocholate, propionate, C16:0 LPE, adipate, C20:4 carnitine, 3'-O-methyladenosine, suberate, nicotinate(B3), pantothenate(B5), *Faecalibacterium prausnitzii*, *Alistipes finegoldii*, *Alistipes shahii*, *Alistipes putredinis*, *Subdoligranulum unclassified*	Lloyd-Price *et al.*, 2019
Treatment-naive UC (18) vs. HC (14)	Mucosal biopsy	**Metabolomics:** GC-ToFMS, UPLC-MS	Lysophospholipids[3], acyl carnitines[3], arachidonate[3], asparagine[3], citrulline[3], dimethylarginine[3], glutamyl-L-amino acids[3], glutamate[3], kynurenine[3], L-valine[3], L-isoleucine[3], nicotinamide[3],	beta-alanine[3], creatine[3], eicosapentaenoate[3], fructose[3], glutaryl-carnitine[3], glycerol-3-phosphate[3], guanosine[3], leucylglycine[3], linoleate[3], L-glutamine[3], methylmalonyl carnitine[3]	Diab *et al.*, 2019

[1]Includes derivatives of the molecule class; [2] significantly different in CD only; [3]significantly different in UC only; [4]Difference not statistically significant in this cohort

Abbreviations: healthy controls (HC); high-resolution magic angle spinning nuclear magnetic resonance (HR-MAS NMR); gas chromatography (GC); ion cyclotron resonance Fourier transform mass spectrometry (ICR-FT/MS); time-of-flight mass spectrometry (ToFMS); liquid chromatography (LC); high performance LC (HPLC); ultra-performance LC (UPLC); medium-chain fatty acids (MCFA); long-chain fatty acids (LCFA). (Table from Pratt et al., 2021).

Table S4. Species Strongly Associated to Metagenomic Beta Diversity Clusters from the Current Study. Multilevel pattern analysis carried out using the indicspecies R package v.1.7.12.

Species	Indicator Value Index	p value
Preserved (NAS/RNAP) - Thermal Lysis		
Odoribacter splanchnicus	0.923	0.0001
Bacteroides thetaiotaomicron	0.887	0.0001
Bacteroides ovatus	0.874	0.0001
Turicimonas muris	0.827	0.0001
Bacteroides xylanisolvens	0.822	0.0001
Bacteroides uniformis	0.791	0.0003
Phascolarctobacterium faecium	0.789	0.0001
Unpreserved (NP) - Thermal Lysis		
Ruminococcus gnavus	0.928	0.0001
Streptococcus sp. A12	0.920	0.0001
Monoglobus pectinilyticus	0.915	0.0001
Roseburia intestinalis	0.874	0.0001
Ruthenibacterium lactatiformans	0.782	0.0001
All (NAS/NP/RNAP) - Mechanical Lysis		
Ruminococcus torques	0.912	0.0001
Agathobaculum butyriciproducens	0.910	0.0001
Dorea sp. CAG_317	0.895	0.0001
Blautia wexlerae	0.863	0.0001
Blautia sp. CAG_257	0.862	0.0001
Eubacterium hallii	0.858	0.0001
Sellimonas intestinalis	0.850	0.0001
Dorea formicigenerans	0.833	0.0001
Fusicatenibacter saccharivorans	0.808	0.0001
Turicibacter sanguinis	0.793	0.0001
Blautia sp. N6H1_15	0.775	0.0001

Appendix 2: Detailed SOPs for Laboratory Methods

Fecal Genomic DNA Extraction Protocol using Zymo Quick-DNA™ Fecal/Soil Microbe Miniprep Kit (D6010)

You will Need:

- Quick-DNA™ Fecal/Soil Microbe Miniprep Kit (50 Preps); Zymo Research, cat no. D6010
- Beta-mercaptoethanol
- Isopropanol

Before Starting:

1. For optimal performance, add beta-mercaptoethanol (user supplied) to the **Genomic Lysis Buffer** to a final dilution of 0.5%(v/v) i.e., 500 µl per 100 ml.
   Note: Fecal DNA Binding Buffer has been renamed to Genomic Lysis Buffer.

2. Heat appropriate amount (50 µl per sample processed) of **DNA Elution Buffer** to 60ºC.

Protocol:

1. Add 2 fecal punches to the **ZR BashingBead™ Lysis Tube**. Add 750 µl **Lysis Solution** to each. Cap tubes tightly to prevent leakage; without stripping.

2. Secure in a bead-beater fitted with a 2-ml tube holder assembly and process for 10 minutes at 1200 rpm.

3. Centrifuge the **ZR BashingBead™ Lysis Tube** in a microfuge at 10,000 x g for 1 minute.

4. Transfer 400 µl supernatant to a **Zymo-Spin™ IV Spin Filter (Orange Top)** in a Collection Tube and centrifuge at 8,000 x g for 1 minute.
   Note: Snap off the base of the Zymo-Spin™ IV Spin Filter (Orange Top) prior to use.

5. To the filtrate in the Collection Tube, add 467 µl of **Genomic Lysis Buffer** and 333.5 µl isopropanol. Pipette up and down to mix.
   Note: Fecal DNA Binding Buffer has been renamed to Genomic Lysis Buffer.
   Note: NML modification of the standard Zymo protocol.

6. Transfer 800 µl of the mixture from Step 5 to a **Zymo-Spin™ IIC Column** in a Collection Tube and centrifuge at 10,000 x *g* for 1 minute.
   Note: Zymo-Spin™ IIC Column has a maximum capacity of 800 µl.

7. Discard the flow through from the Collection Tube and repeat Step 6 until all the mixture from Step 5 has passed through a single column.

8. Add 200 µl **DNA Pre-Wash Buffer** to the **Zymo-Spin™ IIC Column** in a new Collection Tube and centrifuge at 10,000 x g for 1 minute.

9. Add 500 µl **g-DNA Wash Buffer** to the **Zymo-Spin™ IIC Column** and centrifuge at 10,000 x g for 1 minute.
   Note: Fecal DNA Wash Buffer has been renamed to g-DNA Wash Buffer.

10. Transfer the **Zymo-Spin™ IIC Column** to a clean 1.5 ml microcentrifuge tube and add 50 µl pre-heated **DNA Elution Buffer** directly to the column matrix. Wait 3 minutes before centrifuging the assembly at 10,000 x *g* for 30 seconds to elute the DNA.

11. Transfer the eluant back onto the same column and wait another 3 minutes before re-centrifuging the assembly at 10,000 x g for 30 seconds.
    Note: NML modification of the standard Zymo protocol.

12. Snap off the base of the **Zymo-Spin™ IV-HRC Spin Filter (Green Top)** and place into a clean Collection Tube. Centrifuge at 8,000 x *g* for 3 mins.
    Note: If the HRC matrix is dry, add 400-600 µl water prior to prepping the filter.

13. Transfer the eluted DNA (from Step 11), to the prepared **Zymo-Spin™ IV-HRC Spin Filter (Green Top)** in a clean 1.5 ml microcentrifuge tube and centrifuge at exactly 8,000 x g for 1 minute.

    The eluted, filtered DNA is now suitable for PCR and other downstream applications such as 16S V4 sequencing.

1/4th Volume NexteraXT Library Prep Protocol

Reference Materials

- Nextera-xt-library-prep-reference-guide
- Nextera-xt-troubleshooting-technical-note
- Index-hopping-white-paper
- Biomek i7 – Ampure protocol

IMPORTANT NOTES:

- This protocol is intended for use with Illumina's Nextera XT DNA Library Preparation Kit (FC-131-1096).
- Additional reagents required, not supplied within FC-131-1096, are KAPA HiFi HotStart Ready Mix (Roche KK2602), anhydrous ethanol and AMPure XP (or equivalent).
- There are a number of index combinations that can be purchased from Illumina or other suppliers that do not work with this 1/4th volume protocol; you must validate each index set before implementing the use of them in the lab with this protocol. We have had the most success with Nextera-compatible indexes ordered directly from IDT.
- For all AMPure XP (or equivalent) magnetic bead clean-ups, it is recommended that the 80% ethanol used for the washes is made at least monthly from anhydrous ethanol; DO NOT use 95% ethanol to make the 80% ethanol.

PREPARATION

1. Remove the following components from -20°C storage and thaw on ice:
    - ATM (Amplicon Tagment Mix)
    - TD (Tagment DNA buffer)
    - NPM (PCR master mix)
    - KAPA HiFi HotStart Ready Mix
    - Index plate(s)

NOTE: When selecting the indexes to use, you need to be mindful of which sequencing instrument you will be using and the indexing used during its last sequencing run. It is best practice to avoid using the same indexes on consecutive sequencing runs in order to eliminate possible carry-over contaminants demultiplexing with the data from the current run.

2. Remove NT from 4°C and bring to room temperature. Inspect the tube/plate to ensure that no precipitate has formed. If a precipitate has formed, vortex to resuspend.
3. Mix all reagents by inverting the tubes 3-5 times or by briefly vortexing. Spin briefly in microcentrifuge to collect contents.
4. Briefly vortex the index plate(s) or place on a plate shaker for 2 minutes at 1800 RPM and then centrifuge at 280 x g for 1 minute.
5. Label a 96-well plate with the appropriate NGS project number, submitting department, type of library, indexes to be used, and date.

TAGMENTATION OF INPUT DNA

NOTE: Tagmentation is time sensitive, it is important to carry out the subsequent steps rapidly.

1. Add **4.9 µl TD buffer** to each sample well.
2. Add **1.1 µl ATM buffer** to each sample well.

   Alternatively, a master mix of the above 2 reagents (TD buffer and ATM) can be made, mixed well, and 6 µl can be dispensed into the wells of the 96-well plate. Ex: for a 96-well plate of reactions, mix 490 µl TD buffer and 110 µl ATM.
3. Add **4 µl of 0.05 ng/µl template DNA** to each sample well (the template DNA was previously normalized and diluted down to 0.05 ng/µl during the Pre-Library Prep Protocol). Gently pipette up and down 5 times to mix and avoid introducing bubbles while pipetting. Alternatively, seal the plate and vortex or place on a plate shaker for 2 minutes at 1800 RPM.
4. Seal the plate with a Microseal B adhesive seal.
5. Centrifuge at 280 x g at room temperature for 1 minute.
6. Place the plate in a thermocycler and run the program "**5545**":

   Step 1: 55 ℃ for 5 minutes (ensure that lid is heated)
   Step 2: Hold at 10°C

NOTE: Once the samples reach 10℃, it is important to proceed immediately with the neutralization. The run time on the thermocycler is less than 7 minutes.

NEUTRALIZE TAGMENTATION

1. Remove the plate from the thermocycler and briefly centrifuge.
2. Carefully remove the Microseal B adhesive seal and add **1 µl NT** to each well of the plate containing sample.

NOTE: Pre-dispense NT into a 12-strip PCR tube or 96-well plate for multiple use and use a multichannel or liquidator pipette to dispense.

3. Gently pipette up and down 5 times to mix the sample, changing tips between samples.
4. Seal the plate tightly with a foil seal and place on plate shaker for 2 minutes at 1800 RPM.
5. Centrifuge at 280 x g for 1 minute.
6. Incubate the plate at room temperature for a total of 5 minutes. If the shaker was used for 2 minutes and centrifuge for 1 minute, you only need to incubate another 2 minutes.

PCR AMPLIFICATION OF TAGMENTED DNA

1. Add **3.75 µl KAPA HiFi HotStart Ready Mix** to each well of the plate containing samples.
2. Add **3.75 µl NPM** to each well of the plate containing samples.

3. Add **1.5 µl nuclease-free water** (NFW) to each well of the plate containing samples **OR 4.0 µl NFW** if using the 10 bp UDIs from IDT.

   Alternatively, a master mix of the above 3 reagents (KAPA, NPM and NFW) can be made, mixed well, and 9 µl be dispensed into the wells of the 96-well plate. Ex: for a 96-well plate of reactions, mix 375 µl KAPA mix, 375 µl NPM and 150 µl NFW.

   If using the 10 bp UDIs from IDT, the master mix for a 96-well plate of reactions will contain 375 µl KAPA mix, 375 µl NPM and 400 µl NFW. 11.5 µl of this master mix will be dispensed into each well.

4. Add **5 µl Index** (from the index plate with forward and reverse indexes already mixed) to each well containing sample. It is highly recommended Unique Dual Indexing is used for amplicon sequencing. If using the 10 bp UDIs from IDT, **2.5 µl Index** is used from the plate.
5. Gently pipette up and down 5 times to mix, while changing tips in between samples to avoid cross-contamination, then cover and seal the plate with a Microseal B adhesive seal. Alternative to tip mixing, seal plate tightly using a Microseal B and place on plate shaker for 2 minutes at 1800 RPM.
6. Centrifuge at 280 x g for 1 minute.
7. Place the plate in a thermocycler and run the program "**NEXTKAPA**".

   Step 1: 72°C for 3 minutes
   Step 2: 98°C for 45 seconds
   Step 3: 18 cycles of:
   98°C for 15 seconds
   55°C for 30 seconds
   72°C for 30 seconds
   Step 4: 72°C for 5 minutes
   Hold at 10°C
   25 µl reaction, with heated lid tracking 5°C higher than the block

NOTE: If the clean-up will be completed the same day as the PCR, remove AMPure XP (or equivalent) from 4°C storage and resuspension buffer (RSB) from -20°C storage and bring them to room temperature, this takes approximately 30 minutes.

NOTE: Upon completion of PCR, the plate can remain in the thermocycler overnight or be stored at 4°C for up to 3 days before proceeding to the PCR clean-up. If storing at 4°C, replace Microseal B with a foil seal before placing in the fridge.

PCR CLEAN–UP

NOTE: PCR clean-up may be performed with AMPureXP or an equivalent such as PCR CleanDX. From here on out, these will be referred to as magnetic beads.

NOTE: PCR clean-up elution can be performed with RSB or NFW, both are acceptable.

NOTE: PCR clean-up may be carried out on the Biomek i7 liquid handling robot (or equivalent) following the Biomek i7 – Ampure protocol.

1. Remove magnetic beads from 4°C storage and RSB from -20°C storage and bring to room temperature. This will take approximately 30 minutes.

2. Remove the plate from the thermocycler or 4°C storage and centrifuge at 280 x g for 1 min to collect contents.
3. Mix the magnetic beads well to ensure that beads are evenly dispersed.
4. Add **15 µl magnetic beads (for 0.6X clean-up)** to each well of the plate containing library.
5. Seal the plate tightly using foil, and place on plate shaker for 2 minutes at 1800 RPM to mix.
6. Incubate at room temperature for 5 minutes.
7. Place the plate onto magnetic plate stand (or magnetic tube holder) for 2 minutes or until the supernatant has cleared. If a plate shaker was used during Step 5, the plate must be briefly (30 seconds) centrifuged at 280 x g before placing on the magnet.
8. Keep the plate on the magnetic plate stand and carefully remove and discard supernatant using a pipette. Change tips between samples. If any beads are aspirated into the tips, dispense the beads back into the plate and let the plate rest on the magnetic plate stand until the supernatant has cleared and then repeat supernatant removal.
9. Keep the plate on the magnetic plate stand and add **200 µl 80% ethanol** to each well.
10. Incubate the plate on the magnetic plate stand for 30 seconds.
11. Remove and discard the supernatant using a pipette, ensuring tips are changed between samples.
12. Repeat steps 9 to 11 for a total of two ethanol washes.
13. Gently tap the plate on the bench and use a P20 pipette to remove any excess ethanol.
14. Keep the plate on the magnetic plate stand and allow the beads to air-dry at room temperature for 5-10 minutes in the fume hood (if no fume hood is available, you can let air dry on the bench).
15. Remove the plate from the magnetic plate stand and add **25-40 µl NFW or RSB** to each well of the plate containing beads, pipetting up and down at least 10 times to resuspend the bead pellet.

NOTE: If 25 µl is used for resuspension, you will have a higher library concentration than if 40 µl is used but more library (total nanograms) will be used during quantitation therefore leaving less nanograms available for pooling.

16. Incubate at room temperature for 2 minutes.
17. Place the plate on the magnetic plate stand for 2 minutes or until supernatant is cleared.
18. Transfer **supernatant** to a clean 96-well plate, ensuring that no beads are carried over. If some beads are carried over at this stage, perform subsequent library quantitation and pooling steps while on a magnetic plate stand.

Proceed to library quantitation and pooling.

This is a safe stopping point. If you are stopping, seal the plate very well with a foil seal and store at -25°C to -15°C for up to 7 days.

Appendix 3: Code Notebooks for RStudio and Jupyter

A.3.1 RStudio: Statistical Analyses

Detailed code is available at:

https://github.com/prattm1/MSC-Thesis/blob/main/Statistical-Analyses-RStudio

A.3.2 RStudio: Creating figures using ggplot2

Detailed code is available at:

https://github.com/prattm1/MSC-Thesis/blob/main/ggplot-Figures-RStudio

A.3.3 Jupyter Notebooks:

Detailed code for the following Jupyter notebooks:

- Meta-omics-QC-KneadData.ipynb
- Meta-omics-CommunityProfiling-Metaphlan.ipynb
- Meta-omics-FunctionalProfiling-Humann.ipynb
- Fecal-Ref-Comparison.ipynb

is available at:

https://github.com/prattm1/MSC-Thesis

Appendix 4: MSDS for Nucleic Acid Stabilizer Components

Below are excerpts from the Material Safety Data Sheets (MSDS) Composition/Information on Ingredients Section for the two nucleic acid stabilizing solutions used in the current study. Full MSDS can be accessed at the URLs provided.

1. **RNAProtect** (Qiagen)- available at https://www.qiagen.com/ca/knowledge-and-support/product-and-technical-support/quality-and-safety-data/sds-search/?l=CA&q=RNA%20Protect

SAFETY DATA SHEET

RNAprotect Tissue Reagent

Version 3.0 Revision Date: 10/15/2019 Date of last issue:08/11/2021 Date of first issue: 08/25/2017

SECTION 1. IDENTIFICATION

Product name : RNAprotect Tissue Reagent

Manufacturer or supplier's details

Company : QIAGEN GmbH
QIAGEN Str. 1

D-40724 Hilden

Telephone : +49-(0)2103-29-0

Responsible Department : QIAGEN Inc.
19300 Germantown Road

Germantown, MD 20874, USA Tel.: 800-426-8157
http://support.qiagen.com

E-mail : cpc@qiagen.com
addressResponsible/issuing person

Emergency telephone : CHEMTREC
USA & Canada 1-800-424-9300

Recommended use of the chemical and restrictions on use

Recommended use : Laboratory chemicals

SECTION 2. HAZARDS IDENTIFICATION

GHS classification in accordance with the OSHA Hazard Communication Standard (29 CFR 1910.1200)

Not a hazardous substance or mixture.

GHS label elements

Not a hazardous substance or mixture.

Other hazards None known.

SECTION 3. COMPOSITION/INFORMATION ON INGREDIENTS

Substance / Mixture : Mixture

Components

Chemical name	CAS-No.	Concentration (% w/w)
sulfuric acid	7664-93-9	>= 0.1 - < 1

Actual concentration is withheld as a trade secret

2. DNA/RNA Shield (Zymo Research) – available at https://files.zymoresearch.com/sds/dna-rna_shield_products_(us).pdf

Safety Data Sheet acc. to OSHA HCS

ZYMO RESEARCH
The Beauty of Science is to Make Things Simple

Printing date 01/16/2017 *Reviewed on 05/06/2016*

1 Identification

· ***Product identifier***

· ***Trade name: DNA/RNA Shield - 50 ml, 250 ml & sample; DNA/RNA Shield (2X Concentrate) - 25 ml & 125 ml; DNA/RNA Shield Blood Collection Tube, DNA/RNA Shield Fecal Collection Tube, DNA/RNA Shield Collection Tube, DNA/RNA Shield Lysis Tube (Microbe), DNA/RNA Shield Lysis Tube w/ Swab (Microbe), DNA/RNA Shield Lysis Tube (Tissue); DNA/RNA Shield Collection Tube w/ Swab***

· ***Article number:***
R1100-50, R1100-250, R1100-8-S, R1200-25, R1200-125, R1150, R1101, R1102, R1103, R1104, R1105, R1106, R1107, R1108, R1109

· ***Application of the substance / the mixture*** Laboratory Reagent

· ***Details of the supplier of the safety data sheet***

· ***Manufacturer/Supplier:***
Zymo Research Corp.
17062 Murphy Ave.
Irvine, CA 92614
USA
Phone: 1-949-679-1190 or 1-888-882-9682
sds@zymoresearch.com

· ***Information department:*** Product safety department

· ***Emergency telephone number:***
During normal business hours (8 am to 5 pm Pacific Standard Time): +1 (949) 679 1190

2 Hazard(s) identification

· ***Classification of the substance or mixture***


GHS07

Acute Tox. 4 H302 Harmful if swallowed.
Skin Irrit. 2 H315 Causes skin irritation.
Eye Irrit. 2B H320 Causes eye irritation.

· ***Label elements***
· ***GHS label elements*** The product is classified and labeled according to the Globally Harmonized System (GHS). ·
Hazard pictograms GHS07
· ***Signal word*** Warning
· ***Hazard statements***
Harmful if swallowed.
Causes skin and eye irritation.
· ***Precautionary statements*** Wear
protective gloves.
Wash thoroughly after handling.
Do not eat, drink or smoke when using this product.
If in eyes: Rinse cautiously with water for several minutes. Remove contact lenses, if present and easy to do.
Continue rinsing.

(Contd. on page 2)

***Safety Data Sheet** acc. to OSHA HCS*

Printing date 01/16/2017 *Reviewed on 05/06/2016*

Trade name: DNA/RNA Shield - 50 ml, 250 ml & sample; DNA/RNA Shield (2X Concentrate) - 25 ml & 125 ml; DNA/RNA Shield Blood Collection Tube, DNA/RNA Shield Fecal Collection Tube, DNA/RNA Shield Collection Tube, DNA/RNA Shield Lysis Tube (Microbe), DNA/RNA Shield Lysis Tube w/ Swab (Microbe), DNA/RNA Shield Lysis Tube (Tissue); DNA/RNA Shield Collection Tube w/ Swab

(Contd. of page 1)

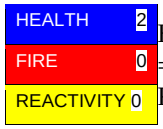
Specific treatment (see on this label).
IF SWALLOWED: Call a POISON CENTER/doctor if you feel unwell.
If skin irritation occurs: Get medical advice/attention.
If eye irritation persists: Get medical advice/attention.
Rinse mouth.
IF ON SKIN: Wash with plenty of water.
Take off contaminated clothing and wash it before reuse.
Dispose of contents/container in accordance with local/regional/national/international regulations. ·

Classification system:

· ***NFPA ratings (scale 0 - 4)***

Health = 2
Fire = 0
Reactivity = 0

· ***HMIS-ratings (scale 0 - 4)***



Health = 2
Fire = 0
Reactivity = 0

· ***Other hazards***

· ***Results of PBT and vPvB assessment*** ·

PBT: Not applicable.

· ***vPvB:*** Not applicable.

3 Composition/information on ingredients

· ***Chemical characterization: Mixtures***

· ***Description:*** This product is a proprietary solution.

· ***Dangerous components:*** Void

4 First-aid measures

· ***Description of first aid measures***

· ***After inhalation:*** Supply fresh air; consult a physician in case of complaints.

· ***After skin contact:*** Immediately wash with water and soap and rinse throughly.

· ***After eye contact:***

Rinsed opened eye for several minutes under running water. If symptons persist, consult a doctor. ·

After swallowing: Do not induce vomiting; immediately call for medical help.

· ***Information for doctor:***

· ***Most important symptoms and effects, both acute and delayed*** No further relevant information available.

(Contd. on page 3)