# METAGENOMIC ANALYSIS OF THE INFANT GUT MICROBIOME

MCMASTER UNIVERSITY

Department of Chemistry and Chemical Biology

# CREATING A METAGENOMIC DATA ANALYSIS PIPELINE USING SIMULATED INFANT GUT MICROBIOME DATA FOR GENOME-RESOLVED METAGENOMICS

By Bhavya Singh, H.B.S.c (McMaster University)

A Thesis Submitted to the School of Graduate Studies in Partial Fulfillment of the Requirements for the Degree of Master of Science

McMaster University **MASTER OF SCIENCE** (2021), Hamilton, Ontario (Chemical Biology)

**Title:** Creating a Metagenomic Data Analysis Pipeline Using Simulated Infant Gut Microbiome Data for Genome-Resolved Metagenomics

**Author**: Bhavya Singh

**Supervisor:** Dr. Jennifer C. Stearns

**Committee Members:** Dr. Andrew G. McArthur, Dr. Eileen K. Hutton, Dr. Katherine M. Morrison

**Number of Pages**: xxv, 191

## Lay Abstract

Solid food introduction to the infant diet brings new glycans to the gut environment, driving the selection of bacteria that are able to digest these compounds.  Studying the gut microbiome during this timepoint is essential to deciphering how and when beneficial bacteria colonize, how they evolve, and how the infant gut matures to an adult-like state. A widely used method to characterize microbial identity and metabolic function in the gut is metagenomic sequencing. However, dominant bacterial genera in the infant gut often have multiple closely related species and strains, making it difficult to decipher the essential metabolic differences between them. In this study, we simulated an infant gut metagenomic dataset to understand how the structure of the infant gut impacts commonly used metagenomic tools, and to quantify the quality of genomes and metabolic predictions at the end of common metagenomic analyses. We found that gut microbial community composition and metagenomic assembler choice both impact the quality of final genomes retrieved from the data, and the accuracy of metabolic gene predictions. Based on these results, we make several recommendations to use ensemble methods to improve metagenomic data analysis, and additionally propose a metagenomic pipeline to analyze infant gut data over the period of solid food introduction.

## Abstract

**Background:** Studying the infant gut microbiome during the period of solid food introduction may provide valuable insight into gut colonization, microbial evolution, and the ecological role of bacterial metabolic pathways in microbial succession. However, since infant gut microbial communities are made of bacterial genera with high relative abundance, within-genus and within-species diversity, the efficacy of current computational tools in elucidating strain-specific differences is not known.

**Methods:** 34 infant gut metagenomic samples were simulated with the CAMI-Simulator, using 16S rRNA gene profiles from subjects of the Baby & Mi study as a reference. Raw simulated reads were trimmed, assembled, and binned into metagenome-assembled genomes (MAGs) using mg_workflow, a Snakemake-based pipeline of current metagenomic analysis protocols. Results were compared to gold-standard references in order to benchmark the success of current computational methods in retrieving strain-level MAGs from the gut, and in predicting bacterial carbohydrate active enzymes. Real metagenomic samples from the Baby, Food & Mi cohort were processed through the bfm_mg_flow pipeline to study the taxonomic and metabolic changes in the infant gut microbiome during the solid food introduction period. Post-pipeline analyses were conducted in R.

**Results:** Misassemblies were significantly impacted by sample community composition, including Shannon diversity, number of strains in the sample, and relative abundance of the most dominant strain. MAG completeness, contamination, quality, and reference coverage were significantly impacted by choice of assembly software, and choice of single- or co-sample assembly. Different assemblies yielded different MAGs from the same samples. Reference coverage of MAGs recovered from co-assemblies were lower than for those from single assemblies and CAZyme predictions were more accurate from MetaSPAdes than from MEGAHIT assemblies at both the assembly-level and the MAG-level. Based on these results, we propose the MetAGenomic PIpelinE (MAGPIE), with recommendations for ensemble methods for assembly, binning, and gene predictions. Using these methods, we identified changes in microbial community composition before and after solid food introduction in real Baby & Mi infant gut samples. These changes included an increase in bacteria that can digest a wide variety of carbohydrates, such as *Bacteroides*, and a decrease in *Bifidobacterium*.

**Conclusions:** In this study, we characterized the current state of tools for genome-resolved metagenomics, and contributed a framework to tailor metagenomic data analysis for the unique composition of the infant gut microbiome. We further used this framework to study bacterial metabolism in the infant gut microbiome before and after the introduction of solid foods.

# Acknowledgments

First and foremost, I would like to thank my wonderful supervisor Dr. Jennifer Stearns, for her guidance, unwavering support, and invaluable insight. From the very first day, Jen encouraged me to take ownership of my project and be an independent thinker, while being the crucial guiding force for the conceptualization and execution of this thesis. She continues to be my inspiration as a scientist and mentor. I will always be grateful for the opportunity to do my Master of Science in the Stearns Lab. Thank you for helping me navigate the weeds, Jen!

I would like to thank Dr. Andrew McArthur of my supervisory committee for encouraging me to pursue my interest in bioinformatics, for always inspiring me think outside of the box, and for my being my mentor since my undergraduate years. I would also like to thank Dr. Katherine Morrison and Dr. Eileen Hutton of my supervisory committee for their guidance and direction, especially in the context of infant health and nutrition. They have always encouraged me to look at the bigger picture. I would like to especially thank my supervisory committee for their guidance throughout the pandemic.

I would like to thank Dr. Finlay Maguire, for helping me through my countless questions about simulations, bioinformatics, and Snakemake. I would also like to thank Dr. Paul McNicholas, for being my supervisor for the MacDATA Fellowship, and for generously sharing his computational resources. I would additionally like to thank the McArthur Lab, the members of the CSU Server Committee, including Amogelang Raphenya and Pardeep Singh, and members of the SHARCNET facilities, including Mark Hahn, for their generosity in sharing and providing support with computational resources. I would like to thank Salman Reza, Sara Dizzell, and Gayatri Nair for their time and efforts in processing infant samples for metagenomic sequencing.

Thank you to current and former Stearns Lab members, for their valuable friendship and knowledge: Colin Bruce, Chiara Homann, Sara Dizzell, Lucas Flett, Aman Patel, Sarisha Philip, and Salman Reza. I would also like to thank Kara Tsang, for being my friend and cheerleader.

Last, but certainly not the least, I would like to thank my family. Thank you, Mom and Dad, for everything.  And thank you, Papa, for cheering for me on from beyond.

# Table of Contents

# List of Figures

# List of Tables

## List of Abbreviations

| | |
|---|---|
| **AA** | Auxiliary Activities |
| **ANOVA** | Analysis of Variance |
| **ASV** | Amplicon Sequence Variant |
| **ATP** | Adenosine triphosphate |
| **Bifid-shunt** | Fructose-6-phosphate-shunt, a pathway unique to *Bifidobacterium* to metabolize glucose and fructose |
| **Bp** | Base pairs |
| **CAMI** | Critical Assessment of Metagenomic Interpretation |
| **CAMISIM** | CAMI simulator, a metagenomic data simulating tool |
| **CBM** | Carbohydrate Binding Module |
| **CE** | Carbohydrate Esterase |
| **CI** | Confidence interval |
| **COGs** | Cluster of orthologous genes |
| **DADA2** | Divisive Amplicon Denoising Algorithm 2 |
| **DNA** | Deoxyribonucleic acid |
| **EMP** | Embden-Meyerhof Parnas pathway |
| **FN** | False negative |
| **FP** | False positive |
| **GH** | Glycoside Hydrolase |
| **GPR** | G-protein-coupled receptor |
| **GSCA** | Gold standard co-assembly |
| **GSSA** | Gold standard single assembly |
| **GTs** | Glycol Transferase |
| **HDAC** | Histone Deacetylase |
| **HMO** | Human Milk Oligosaccharide |
| **HQ** | High quality |
| **LCA** | Lowest common ancestor |
| **MAG** | Metagenome assembled genome |
| **MHCA** | MEGAHIT co-assembly |
| **MHSA** | MEGAHIT single-sample assembly |
| **MSSA** | MetaSPAdes single-sample assembly |
| **NCBI** | National Center for Biotechnology Information |

| | |
|---|---|
| **PCR** | Polymerase chain reaction |
| **PE** | Paired-end |
| **PL** | Polysaccharide Lyase |
| **PUL** | Polysaccharide Utilization Loci |
| **rRNA** | Ribosomal ribonucleic acid |
| **SCFA** | Short chain fatty acid |
| **SE** | Single-end |
| **SNV** | Single-nucleotide variant |
| **TN** | True negative |
| **TP** | True positive |
| **V3** | 16S ribosomal gene variable region 3 |
| **V4** | 16S ribosomal gene variable region 4 |

# Declaration of Academic Achievement

I, Bhavya Singh, declare that the MSc thesis “Creating a Metagenomic Data Analysis Pipeline Using Simulated Infant Gut Microbiome Data for Genome-Resolved Metagenomics” is a product of my own work. This thesis has not been submitted to receive an academic degree at an any other institution. I am the independent author of this thesis and have written and performed the experiments for this degree without additional help, with the exception of named sources. My supervisor, Dr. Jennifer Stearns, and the members of my supervisory committee, Dr. Andrew McArthur, Dr. Katherine Morrison, and Dr. Eileen Hutton provided the support and guidance required for conceptualizing the research design of this thesis. Previous and current work essential to this publication was performed by Dr. Jennifer Stearns, Sara Dizzell, and Salman Reza.

# 1 Introduction

With the rise of metabolic and immune-related disorders, a growing body of research is pointing towards the role of the human microbiota in host health (Gilbert *et al*., 2018; Ogurtsova *et al*., 2017; Toniolo *et al*., 2019). Commensal gut microorganisms, which collectively harbor more genetic diversity than the human genome, have been found to contribute to host metabolism, immune development, and obesity (Duranti *et al*., 2017a; Turnbaugh *et al*., 2006; Turroni *et al*., 2014a). The human genome contains fewer than 20 unique enzymes for the breakdown and metabolism of complex carbohydrates, whereas gut microbes encode a variety of Carbohydrate Active Enzymes (CAZymes) to metabolize diverse complex carbohydrates and fiber from the host and the diet (Cantarel et al., 2012; Kaoutari et al., 2013). Established during infancy, bacterial communities in the gut continue to mature throughout the first three years of life, with the introduction to solid foods and new dietary carbohydrates being an important checkpoint in this process (Koenig *et al*., 2011). As such, studying the infant gut microbiome during the solid food introduction period may provide valuable insight into gut colonization, bacterial selection, and the ecological role of CAZymes and bacterial metabolic pathways in bacterial succession. With the continued rise of chronic metabolic disorders, understanding the contribution of specific microbial genes and pathways in human metabolism may serve as a crucial step towards the innovation of personalized treatment options.

Given the large amount of within-species variation in the infant gut microbiome, mapping the activity and evolution of gut bacteria over time is both a challenging and essential task (Van Rossum *et al*., 2020). The task is essential because different strains of the same species often display high phenotypic variance, and within-microbiome evolution and strain identity are known to contribute to host health (Van Rossum *et al*., 2020; Zhao *et al*., 2019). However, the challenge arises due to the high level of genomic sequence similarity between strains belonging to the same species, which cause significant errors in the bioinformatics analysis of these samples (Lugli *et al*., 2019a; Maguire *et al*., 2020). Unfortunately, it is not known how much strain diversity is retained or lost during standard metagenomic workflows for infant gut metagenomic data, or which types of community composition are likely to cause bioinformatics tools to underperform. Understanding and quantifying the limitations of current *in silico* workflows is an important step prior to using these tools for the infant gut microbiota, which has a unique microbial community composition.

## 1.1 The Infant Gut Microbiome

The gut microbiota encompasses microorganisms from any of the three domains of life—*Bacteria, Archaea,* and *Eukarya*—in addition to viruses, with complex ecological relationships that extend from competition to commensalism (Milani *et al*., 2017a). The majority of the gut microbiota is made up of *Bacteria* with estimates of 100 trillion microbial cells (Qin *et al*., 2010) in the gut. The coordinated activity and function of these microorganisms include the breakdown of incoming food substances, the modulation and development of the immune response, and protection from harmful pathogens (Duranti *et al*., 2017a; Laforest-Lapointe and Arrieta, 2017; Milani *et al*., 2017a; Turnbaugh *et al*., 2006; Turroni *et al*., 2014a).

The past decade has witnessed a remarkable increase in research involving gut microbial community dynamics and human health, attributed to the technological advancements in high-throughput sequencing, leading to the drastic decrease in sequencing costs (Gilbert *et al*., 2018). While the adult microbiota is typically dominated by bacteria of the phyla *Firmicutes* and *Bacteroidetes*, the infant gut is initially populated by the phyla *Proteobacteria* and *Actinobacteria* (Milani *et al*., 2017a). There is a high amount of variability in bacterial abundance across infants, with dominant bacterial groups in the gut including *Bifidobacterium*, *Veillonella, Streptococcus, Citrobacter, Escherichia, Bacteroides, and Clostridium* (Milani *et al*., 2017a). *Bifidobacterium* and *Lactobacillus* are commonly found in breastfed infants, while formula-fed infants have higher relative abundances of *Clostridium spp*., *Bacteroides spp*, and members of the *Enterobacteriaceae* family (Favier *et al*., 2002; Koropatkin *et al*., 2012). The introduction of solid food shifts bacterial abundances to an adult-like microbiota, dominated by *Bacteroides, Firmicutes*, and members of the *Actinobacteria* phylum other than *Bifidobacterium* (Milani *et al*., 2017a). However, the microbial profiles of gut bacteria vary depending on geographical location, diet, and environmental factors, making it difficult to make generalizations about bacterial abundances (David *et al*., 2014; Milani *et al*., 2017a; Sharon *et al*.; Yatsunenko *et al*., 2012).

### 1.1.1 Colonization

The first three years of life are characterized by rapid changes to microbial species and communities in the gut, until the microbiota reaches an adult-like state of stability and maturation. As per the defined patterns in community ecology and population genetics, colonization of the infant gut can be divided into four categories; dispersal of bacterial species into the environment, selection based on gene-level fitness traits, drift, and diversification of the individual species and the overall community. First, metabolic niches within a newborn infant are colonized by microbial species through the process of dispersal (Sprockett *et al*., 2018). Initial dispersal of these species into the gut is impacted by the maternal microbiota, mode of delivery, the infant's gestational age, and the physiological and genetic characteristics of the infant's gut (Laforest-Lapointe and Arrieta, 2017; La Rosa *et al*., 2014; Sprockett *et al*., 2018). For example, 50% of microbes found in the infant gut one day after birth are also found in the microbiota of various maternal body sites, such as the vagina, skin, and oral cavity (Ferretti *et al*., 2018). After this initial dispersal, selection drives the differences in bacterial species reproduction due to fitness and ecological niches, with major influences including the infant diet and immune system (Sprockett *et al*., 2018; Vellend, 2010). One example of selection is breast-feeding; breast milk contains several hundred complex human milk oligosaccharides (HMOs) that select for the growth of lactic acid bacteria, such as *Bifidobacterium*, *Lactobaccilus*, and *Bacteroides* (German *et al*., 2008; Martín *et al*., 2012). Breast milk drives bacterial dispersal by contributing to the initial colonization for microorganisms in the infant gut, while also promoting selection of species that are able to digest the human milk oligosaccharides that it is composed of (Martín *et al*., 2012; Sprockett *et al*., 2018). Dispersal is followed by drift, which is characterized by altered species abundances caused by random changes or events that are not impacted by the identity of the species (Sprockett *et al*., 2018). Lastly, diversification of the gut community is the rapid gene-level evolution and adaptation of microbial species to selective forces (Sprockett *et al*., 2018; Vellend, 2010). In this phase, the shifts in the species abundance—which are a result of selection based on gene-level fitness traits—can impact community diversity as a whole.

These ecological patterns collectively influence the colonization of the infant gut microbiome, with the host's external and internal environment often impacting multiple patterns at the same time. The introduction of solid food to the infant diet, which occurs between 4-6 months of age, is an example that leads to the rapid selection and diversification of the microbiota (Bäckhed *et al*., 2015; Sprockett *et al*., 2018). The solid food diet, which contains novel substrates for bacterial breakdown such as fiber, impacts selection of bacteria in the gut (Koropatkin *et al*., 2012).

### 1.1.2 Solid Food Introduction, Carbohydrate Active Enzymes, and Bacterial Metabolism

Glycans, or polysaccharides, are defined as monosaccharides linked by glycosidic bonds, and include the carbohydrate regions of glycoproteins, glycolipids, and other glycoconjugates (Dwek, 1996; Koropatkin et al., 2012; Varki, 2017). Dietary glycans are glycans derived from food (Koropatkin et al., 2012). The transition from breast milk to solid food is a major event in microbiome succession, causing a drastic increase in the diversity of new dietary glycans, including complex carbohydrates and fibers (Koropatkin *et al*., 2012). Prior to solid food introduction, available food glycans in the infant gut include commensal microorganisms, HMOs, and host mucus (Koropatkin *et al*., 2012). Solid food introduces additional glycans from plants and mammalian sources, including resistant starches, non-starch polysaccharides, unabsorbed sugars, sugar alcohols, oligosaccharides, and proteins (Ramakrishna, 2013). Some examples of resistant starches include starches encapsulated by indigestible plant matrices, while others include starches with high amylose content found in rice and maize. Furthermore, non-starch polysaccharides refer to compounds such as pectin, cellulose, and xylan, while unabsorbed sugars and sugar alcohols include maltitol, sorbitol, fructose, and more (Ramakrishna, 2013).

Carbohydrate-active enzymes (CAZymes) encoded by both humans and bacteria are able to break down the aforementioned compounds, enabling the transfer of ATP from the carbon-based food source to bacteria, as well as to the anaerobic cells in the gut (Kaoutari *et al*., 2013). CAZymes include glycoside hydrolases, polysaccharide lyases, glycosyltransferases, and carbohydrate esterases (Kaoutari *et al*., 2013).

### 1.1.3 Types of CAZymes

Carbohydrates harbor immense structural and functional variation, attributed to the sheer magnitude of theoretically possible structures that can be created by the rearrangement of glyosidic linkages by CAZymes (Bourne and Henrissat, 2001). Glycoside hydrolases (GHs), which represent enzymes that hydrolyze the O-glycosidic bond in the carbohydrates, are present in almost every living organism (Naumoff, 2011). The specificity and structure of bacterial GHs depends heavily on the substrates available in the environment, and their activity may often confer competitive advantage to the bacteria that have them (Naumoff, 2011). Even small changes in the structure of GHs can impact their substrate specificity (Naumoff, 2011). Currently, there are 167 GH families classified in the CAZy database, in addition to 18 clans of related families (Cantarel *et al*., 2009). Families within a clan tend to share an evolutionary origin, in addition to having significantly similar functional and structural characteristics, such as conserved tertiary structures and mechanism of enzymatic action (Naumoff, 2011). GHs within the same family typically have high sequence similarity and similar catalytic activity, but different substrate specificity (Van Den Broek and Voragen, 2008).

Glycosyltransferases (GTs) are responsible for forming glycoside bonds by transferring glycosyl groups from activated sugar phosphates to specific acceptors (Lairson *et al*., 2008). As such, GTs are essential for the biosynthesis of new carbohydrates. In contrast to the diversity of folds seen in glycoside hydrolases, the 110 classified GT families only have three potential folds; GT-A, GT-B, and GT-C (Gloster, 2014; Lairson *et al*., 2008; Schmid *et al*., 2016).

Due to the number of unique GHs and GTs, new families are characterized through bioinformatics methods, using sequence alignment and Hidden Markov Models. While this is an incredibly useful and powerful technique of finding new CAZymes, many gene classifications have not been biologically validated.

Polysaccharide Lyases (PLs) are also responsible for breaking down glyosidic bonds, and are evolutionarily related to GHs; however, unlike GHs, the catalytic mechanism involves an elimination as opposed to a hydrolytic bond cleavage(Lombard *et al*., 2010). PLs and GHs can often break down the same C-6 carboxylated polysaccharides to yield different products. While GHs break the glyosidic bonds through the addition of water, PLs instead utilize β-elimination (Lombard *et al*., 2010). Furthermore, both GHs and PLs can have additional modular structures, where the main catalytic region can be attached to supporting modules (Boraston *et al*., 2004; Lombard *et al*., 2010). This includes carbohydrate-binding modules (CBMs), which promote the initial recognition of the carbohydrate, along with its subsequent attachment to the enzyme’s active site (Boraston *et al*., 2004; Guillén *et al*., 2010). There are 40 PL families currently on the CAZy database, along with 86 CBM families(Cantarel *et al*., 2009).

Carbohydrate esterases (CEs) carry out the de-O or de-N-acylation of carbohydrates by removing the ester from the given polysaccharide (Cantarel *et al*., 2009; Nakamura *et al*., 2017). Since an ester is derived from an alcohol and an acid, CEs are divided into two classes, where the polysaccharide may play the role of the acid or the alcohol (Nakamura *et al*., 2017). Lastly, auxiliary activities (AA), or auxiliary redox enzymes, act in synchrony with other CAZymes to primarily degrade plant cell walls(Levasseur *et al*., 2013). This allows other CAZymes, such as GHs and CEs, to access the carbohydrates within the cell wall(Levasseur *et al*., 2013).

### 1.1.4 Relevance of CAZymes in the Gut Microbiota and Health

The human genome contains a limited number of CAZymes specific to the digestion of food glycans; out of the 97 glycoside hydrolases encoded by humans, fewer than 20 are responsible for the breakdown of complex carbohydrates (Cantarel *et al*., 2012; Koropatkin *et al*., 2012). Comparatively, a single gut bacterium like *Bacteroides thetaiotaomicron* contains 260 glycoside hydrolases (Cantarel *et al*., 2012). Even the carbohydrates in breast milk—which are not used by infants for nutrition—are utilized by lactic acid bacteria as energy sources. The carbohydrates in human milk include several hundred glycan structures and complex human milk oligosaccharides (HMOs), which are made of lactose, galactose, fucose, glucose, and N-acetylglucosamine (Gnoth *et al*., 2000; Koropatkin *et al*., 2012).

Complex carbohydrate digestion by CAZymes is a multi-step process involving trophic interactions between microbes, such as competition and cooperative sharing of resources (Milani *et al*., 2017a). For example, complex carbohydrates may be detected by bacteria known as primary degraders, which break down these glycans and provide oligosaccharides and monosaccharides to be broken down by other microbes in the gut (Milani *et al*., 2017a). Metabolites produced by primary degraders can then serve as substrates for microbial secondary degraders (Milani *et al*., 2017a).

Bacterial metabolism of carbohydrates leads to the production of molecules that have bioactive roles in the human body, including short-chain fatty acids (SCFAs) such as butyrate, acetate, and propionate (Bhattacharya *et al*., 2015). While butyrate serves as an energy source for gut epithelial cells and colonocytes, acetate and propionate are carried to the liver to serve as substrates for gluconeogenesis and lipogenesis (Tremaroli and Bäckhed, 2012). SCFAs play further roles in protein and cholesterol synthesis, in addition to impacting gene expression in the colon by regulating the enzyme histone deacetylase (HDAC) and certain G-protein-coupled receptors (GPRs) (Laforest-Lapointe and Arrieta, 2017; Tremaroli and Bäckhed, 2012). Consequently, the diversity and activity of bacterial CAZymes has important consequences for human nutrition, energy balance, and the regulation of adiposity (Bellahcene *et al*., 2013; Cantarel *et al*., 2012; Gnoth *et al*., 2000; Lin *et al*., 2012; Ramakrishna, 2013; Tremaroli and Bäckhed, 2012).

## 1.2 Bifidobacterium

*Bifidobacterium* are among the dominant members of the infant gut microbiota, and can make up to 90% of the relative abundance of infant stool samples. They are characterized as Gram-positive, anaerobic bacteria without motility or spore-forming activity (Duranti *et al*., 2020). Bifidobacteria were first isolated from infant feces in 1899 by H. Tissier, and were initially named *Bacillus bifidus* (Turroni *et al*., 2014b), Throughout the 20th century, bifidobacteria were classified under *Lactobacillus*, and were finally characterized as their own genus in 1974 (Buddingh, 1975). The genus consists of 80 taxa, which are comprised of 73 species and 7 subspecies, although some metagenomic studies have found over 89 novel bifidobacterial species in addition to the presently-described members (Duranti *et al*., 2017b, 2020; Laureys *et al*., 2016; Lugli *et al*., 2018; Michelini *et al*., 2016a, 2016b, 2018; Milani *et al*., 2017b; Modesto *et al*., 2018b, 2018a; Pechar *et al*., 2017). Based on sequence similarity, *Bifidobacterium* taxa typically cluster into ten distinct phylogenetic groups (Lugli *et al*., 2019b).

As gut commensals, bifidobacterial species carry Carbohydrate Active Enzymes (CAZymes) to metabolize incoming human milk oligosaccharides (HMOs) and other carbohydrates into human-digestible sugars. Although the relative abundance of *Bifidobacterium* species in infancy decreases as the microbiota transitions to an adult-like state, they persist in the gut throughout life (Laforest-Lapointe and Arrieta, 2017)·(Arboleya *et al*., 2016a; Odamaki *et al*., 2016). It is consistently reported in the literature that *B. longum* ssp. *infantis* is one of the key colonizers of the infant gut, while *B. adolescentis* is found primarily in adults (Avershina *et al*.; Underwood *et al*., 2015). In infancy, breastfeeding is strongly associated with *B. longum spp. infantis*, while formula-fed infants are often seen to have adult-like microbiotas with different bifidobacterial species, such as *B. adolescentis* and *B. longum spp. longum* (Davis *et al*., 2020). Different *B. longum* subspecies, *B. catenulatum*, *B. bifidum,* and *B. breve* are detected at all ages, with the exception of centenarians, who are often colonized by *B. dentium* (Kato *et al*., 2017).

The curious survival of this genus throughout life has led to the hypothesis that bifidobacteria may be keystone species in the gut, potentially conferring positive benefits to the gut microbiota in both infancy and adulthood (Garcia *et al*., 2019; Gotoh *et al*., 2019). Keystone groups in the microbiota exert major effects on the microbial community in a manner that is disproportionate to their abundance; the loss of these groups can be positive or negative, and can have cascading impacts, such as decreased diversity, extinction of microbial taxa, or disrupted ecosystem functions (Laforest-Lapointe and Arrieta, 2017). As keystone species, *Bifidobacterium* may be important for the survival of other species, particularly by producing SCFAs that are used by other microbes, or by acting against pathogenic species (Laforest-Lapointe and Arrieta, 2017). The majority of

*Bifidobacterium* species are isolated from the gastrointestinal tracts of humans, social insects, non-human mammals, and birds; surprisingly, all aforementioned habitats represent animals whose offspring receive greater parental care (Turroni *et al*., 2018a).

Bifidobacteria are detected in all mammalian species, with their overall average relative abundance in mammalian adults being roughly 3.5% (Milani *et al*., 2017b). This suggests that while bifidobacteria are not necessarily the dominant genera in adult mammals, they are widespread. The four most abundant and prevalent bifidobacterial species (*B. adolescentis, B. longum, B. pseudolongum,* and *B. bifidum*) have a wide range of adaptive capabilities, leading to their prevalence in the guts of 85% to 95% of mammalian species, at different abundances (Milani *et al*., 2017b). In contrast, there are other bifidobacterial species that are less abundant, but highly host-specific (Milani *et al*., 2017b). The prevalence of these species in gastrointestinal habitats suggests that their primary mode of ecological transmissions occurs from the maternal microbiota to the offspring (Turroni *et al*., 2018a). Importantly, studies suggest that *Bifidobacterium* species have coevolved to confer protective advantages to the infant microbiota, such as competitively excluding Gram negative enteropathogenic bacteria, in addition to inhibiting a variety of virulence factors (Delcaru *et al*., 2016; Vazquez-Gutierrez *et al*., 2016). Bifidobacteria produce metabolites like lactate and acetate that encourage the growth of other microorganisms through cross-feeding (Turroni *et al*., 2018a; Underwood *et al*., 2015).

### 1.2.1 Bifidobacterium metabolic activity in host health

*Bifidobacterium* species carry glycoside hydrolases to metabolize a large variety of carbohydrates, increasing the availability of nutrients to both the host and other gut microbes(Bottacini *et al*., 2014). In doing so, they also produce bioactive compounds—including SCFAs, vitamins, and fatty acids—that help lower intestinal pH, enable sodium and water absorption by the host, increase bioavailability of calcium and magnesium, and protect epithelial cells from pathogens(Bottacini *et al*., 2014; Scott *et al*., 2013).

During infancy, members of *B. longum spp. infantis* break down HMOs into acidic products, such as lactate and the SCFA acetate (Duar *et al*., 2020a). *B. longum spp. infantis* is the only known bacteria that has complete metabolic pathways to break down all HMO structures in human milk into acidic compounds (Duar *et al*., 2020a). Lactate and acetate go on to lower the pH of the gut, and increase colonization resistance to maintain a protective gut environment, which reduces the risk of enteric inflammation and autoimmune disorders (Duar *et al*., 2020a; Sela *et al*., 2008; Underwood *et al*., 2015). In the absence of *B. infantis*, HMOs are actually passed into stool, indicating that they are not used as a dietary resource. As a result, energy sources are not provided to the infant or to secondary degraders(Duar *et al*., 2020b). Acetate production by bifidobacteria is also important for anti-inflammatory activity through T-cell and cytokine regulation, in addition to improved protection of the gut mucosal epithelium (Fukuda *et al*., 2012;

Smith *et al*., 2013). On the other hand, lactate is transported by gut epithelial cells to the brain, where it is responsible for regulating neural activity (Duar *et al*., 2020b). By producing acetate and lactate, bifidobacteria engage in cross-feeding interactions that lead to the production of butyrate, which has a pivotal role as the preferred energy source of epithelial cells (Alessandri *et al*., 2019). Butyrate also encourages anti-inflammatory responses in the gut(Alessandri *et al*., 2019).

After the introduction of solid foods, the cross-feeding activities previously conducted by infant strains of bifidobacteria are subsequently handed off to adult strains (Roy *et al*., 2006; Turroni *et al*., 2018b). In addition to HMOs, bifidobacteria also degrade diet-derived sugars (glucans, fructans, xylans, resistant starches, pectins, etc), and mucins, which are the glycoprotein on the gut epithelial layer (Alessandri *et al*., 2019). As the gut microbiota matures, there is a transition towards adult metabolic pathways for SCFA production. For example, adult strains of *B. bifidum* begin breaking down mucins to produce acetate (Turroni *et al*., 2018b). *B. bifidum* strains such as PRL2010 are also known to be particularly altruistic by engaging in cross-feeding activities that encourage the growth of other *Bifidobacterium* species (Turroni *et al*., 2020). By breaking down mucin, bifidobacteria also encourage the secretion and recovery of more colonic mucin, which favorably thickens the gut epithelial barrier(Caballero-Franco *et al*., 2007).

### 1.2.2 The Bifidobacterium pan-genome and strain-level CAZyme activity

Selective pressure in the gut has led bifidobacteria to have a very large pan-genome of over 24,000 clusters of orthologous genes (COGs)(Milani *et al*., 2014, 2015; O'Callaghan and van Sinderen, 2016; Rodriguez and Martiny, 2020). In comparison, the strict bifidobacterial core genome that is found in 100% of species is comprised of just 438 genes, with 115 additional "soft core" genes that are found in 95% of species. Notably, while the *Bifidobacterium* core genome only contains the partial Embden-Meyerhof Parnas (EMP) pathway for glycolysis, it has all the genes necessary for the *Bifidobacterium*-specific "fructose-6-phosphate-shunt" (or "bifid-shunt") to break down glucose and fructose to lactic acid and acetate (Milani *et al*., 2014; O'Callaghan and van Sinderen, 2016). During the conversion of pyruvate to acetate, these pathways allow bifidobacteria to synthesize an additional ATP molecule per glucose, which means that the bifid-shunt produces more energy than the EMP pathway during glycolysis (Fushinobu, 2010; De Vuyst *et al*., 2014). In the gut microbiota, the bifidobacterial energetic yield is thus higher than that of lactic acid bacteria (Milani *et al*., 2014). Consequently, the bifid shunt may contribute to bifidobacterial persistence in the gut after infancy, and particularly after the introduction of solid foods.

Bifidobacterial species exhibit a large diversity of GHs for the breakdown of carbohydrates, with some GHs being a part of the core genome, and others being a part of the pan-genome (O'Callaghan and van Sinderen, 2016). For example, GH2 is a β-galactosidase that is predicted to be found in nearly all bifidobacterial genomes. For bifidobacterial species that reside in mammalian guts, GH13 is the most commonly-found glycoside hydrolase(O'Callaghan and van Sinderen, 2016). GH13 encodes α-1,4-glucosidases, amylopullanases, and α-amylase, and is responsible for breaking down a diverse range of carbohydrates, including plant-derived glycans such as starch, amylose, amylopectin, stachyose, raffinose, and melibiose(Katoh *et al*., 2020; Kumar, 2010; Turroni *et al*., 2019). The broad bifidobacterial glycobiome also consists of GH29 (fucosidases, exo-sialidase), GH95 (exo-sialidase), GH20 (hexosaminidase), GH112 (lacto-N-biosidases), GH38 (α-mannosidases), GH125 (mannosidases), and GH129 (α-N-acetylgalactosaminidases)(Katoh *et al*., 2020; Milani *et al*., 2015).

Bifidobacterial species also have their own core glycobiomes. For example, members of the *B. bifidum* species are known to have a large number of mucin-degrading GHs, due to their crucial roles in breaking down the gut epithelial layer(Turroni *et al*., 2018b). As such, some GHs responsible for the metabolism of host-derived glycans are always found within the *B. bifidium* species, including GH33 (exo-sialidases), GH34 (exo sialidases), GH29 (fucosidases, exo-sialidase), GH95 (fucosidases, exo-sialidases), and GH20 (hexosaminidase)(Turroni *et al*., 2018b).

While there have been many studies characterizing the GHs and other CAZymes in the *Bifidobacterium* pan-genome, all pan-genomic analyses indicate that there are a high number of COGs that are unknown(O'Callaghan and van Sinderen, 2016; Rodriguez and Martiny, 2020). Current knowledge points to the fact that the *Bifidobacterium* pan-genome has not yet been completely characterized, since rarefaction curves plotting the identification of new genes are yet to level off for the genus(Duranti *et al*., 2016; Milani *et al*., 2014).

### 1.2.3 Beyond *Bifidobacterium*: The importance of identifying strains in the gut microbiome

With the immense selective pressure in the gut microbiota, bacteria are constantly evolving to persist in the environment, especially during key timepoints such as solid food introduction. Since these changes are happening at a strain-level, many previously uncharacterized COGs and CAZymes are likely to be found in the pan-genome rather than the core genome. To fully understand the interactions and activity of bacteria in the infant gut—along with the impact of these bacterial interactions on host health—we must be able to identify known and unknown COGs in specific bacterial strains in a variety of genera. As we have outlined with *Bifidobacterium*, similar levels of diversity exist in all other major genera found in the infant gut microbiome, such as *Bacteroides* and *Lactobacillus* (Truong *et al*., 2017; Zhao *et al*., 2019). Given the close link between strain-level activity and host health, a species-level analysis is insufficient to fully understand how microbial communities are connected to disease states(Van Rossum *et al*., 2020).

## 1.3 Culture-independent methods of characterizing variation in the infant gut microbiome

The gut microbiota, which is represented by fecal stool samples in most studies, can be profiled using various approaches. While initial studies primarily involved culture-based techniques, it was formerly presumed that the majority of the microbiota is un-culturable, prompting an increase in culture-independent profiling techniques(Milani *et al*., 2017a; Walker *et al*., 2014). Although this theory has since been refuted, culture-independent techniques continue to provide an efficient, low cost mechanism to profile microbial communities (Browne *et al*., 2016; Lau *et al*., 2016; Walker *et al*., 2014). Given the sheer number of bacterial strains present in each sample, it is not possible to culture and sequence each isolate. While the challenges and limitations of purely *in silico* analyses have been extensively discussed in the gut microbiome research community, these protocols continue to be an invaluable tool for strain-level analysis(Maguire *et al*., 2020; Quince *et al*., 2017; Van Rossum *et al*., 2020). Two widespread methodologies to profile the microbiota include 16S ribosomal RNA gene sequencing and metagenomics.

### 1.3.1 16S ribosomal RNA gene Profiles

16S rRNA gene profiling involves high-throughput sequencing of the 16S rRNA gene, which serves as a conserved marker to profile microbial communities

(Milani *et al*., 2017a). PCR primers can bind to conserved DNA sequences bordering one of the nine hypervariable regions, V1 to V9, of the 16S rRNA gene; the amplification and analysis of these regions can be used to identify the microbial genomes they belong to (Chakravorty *et al*., 2007). While 16S rRNA gene profiling is a highly-characterized and tested approach, there are several caveats to relying solely on this methodology. Firstly, the V1 to V9 regions can vary based on their level of conservation across microbial taxa; as a result, it is essential to choose a PCR primer pair that can reliably differentiate between different taxa (Chakravorty *et al*., 2007). For example, while sequencing the V3 region is most suitable for profiling bacterial genera, V4 to V7 are less useful for genus or species-level differentiation (Chakravorty *et al*., 2007). Secondly, some bacteria can have multiple copies of the 16S rRNA gene, making the method less reliable for elucidating species abundances(Milani *et al*., 2017a). Lastly, this is a low-resolution method for characterizing within-species variation; while oligotyping, amplicon sequence variants (ASVs), and full gene single-nucleotide variants (SNVs) can sometimes capture within-species strains, the method is not optimized for high resolution profiling(Van Rossum *et al*., 2020). However, despite these caveats, 16S rRNA gene profiling remains to be one of the best methods for taxonomically profiling the gut microbiota(Milani *et al*., 2017a).

### 1.3.2 Shotgun Metagenomics

Shotgun metagenomic sequencing involves sequencing all DNA extracted from complex environmental samples, including microbes that have not been previously classified or cultured (Milani *et al*., 2017a). With the increased ease-of-access and lowering costs of sequencing technology, the past decade has made shotgun metagenomics an established protocol to study the gut microbiome. Unlike 16S rRNA gene profiling, metagenomic sequencing does not require primers or DNA amplification. Furthermore, metagenomic sequencing provides significantly more information about the microbiota, including functional information about metabolic processes and antibiotic resistance(Milani *et al*., 2017a). This data can also be assembled to extract full-length genome sequences of abundant bacteria, making it a highly valuable approach for microbial profiling(Milani *et al*., 2017a).

Unfortunately, while metagenomic sequencing is a powerful tool to profile the microbiome, analyzing metagenomic data is a highly variable and multi-step process, introducing potential for diversity in conclusions at every individual step(Quince *et al*., 2017). The nature and complexity the environmental sample, chosen sequencing platform, sequencing depth, and outputted data amount and length can all impact the choice of tools used to analyze the data (Quince *et al*., 2017). Choice of analysis software further impacts the quality of results(Quince *et al*., 2017). A summary of bioinformatics protocols for processing shotgun metagenomic sequencing data is outlined in **Figure 1**.

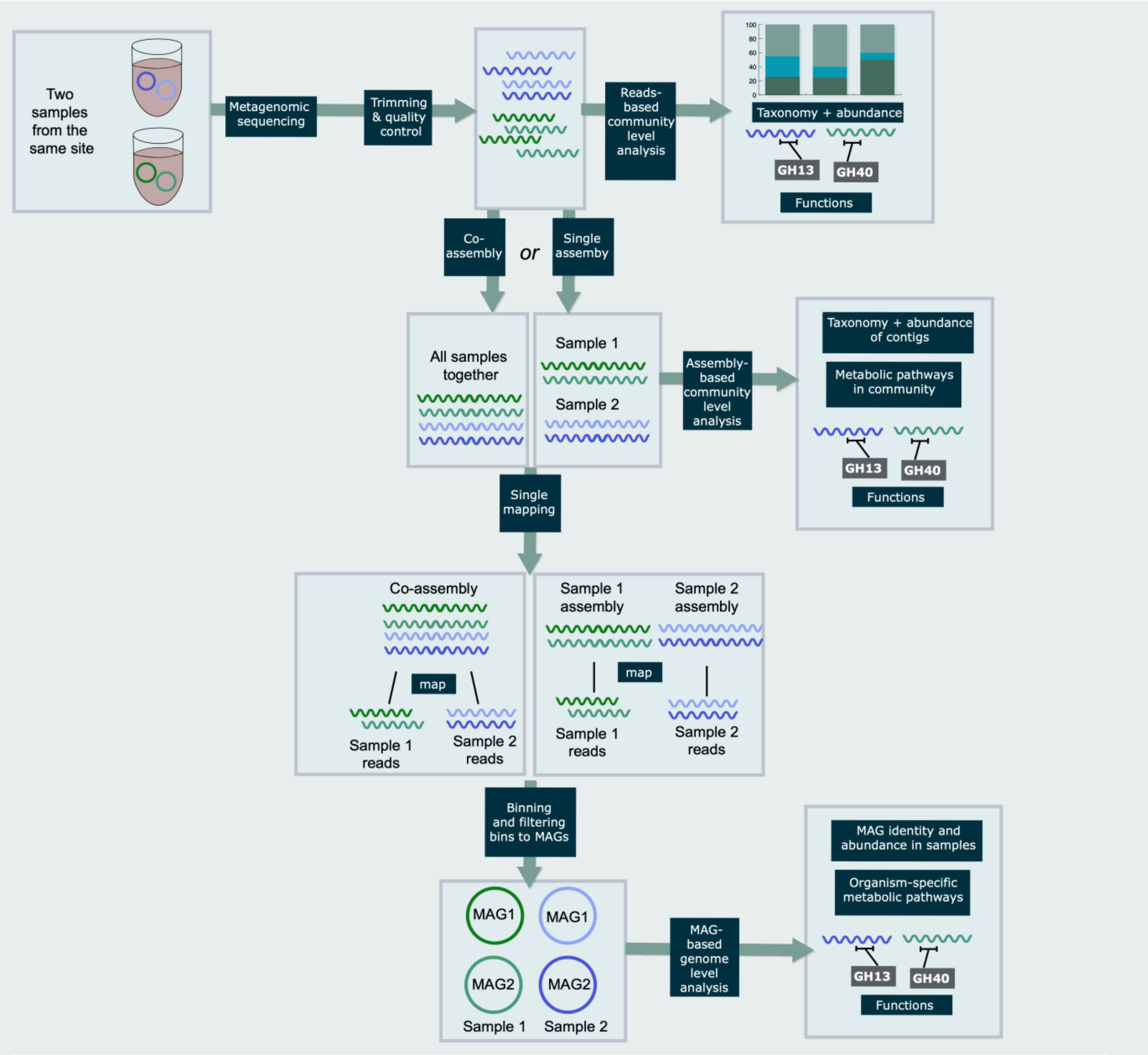

Two samples from the same site
Metagenomic sequencing
Trimming & quality control
Reads-based community level analysis
100
80
60
40
20
0
Taxonomy + abundance
GH13
GH40
Functions
Co-assembly
or
Single assemby
All samples together
Sample 1
Sample 2
Assembly-based community level analysis
Taxonomy + abundance of contigs
Metabolic pathways in community
GH13
GH40
Functions
Single mapping
Co-assembly
map
Sample 1 reads
Sample 2 reads
Sample 1 assembly
Sample 2 assembly
map
Sample 1 reads
Sample 2 reads
Binning and filtering bins to MAGs
MAG1
MAG1
MAG2
MAG2
Sample 1
Sample 2
MAG-based genome level analysis
MAG identity and abundance in samples
Organism-specific metabolic pathways
GH13
GH40
Functions

**Figure 1: Bioinformatics genome-resolved metagenomics workflow for shotgun metagenomics data**

### 1.3.3 Metagenomic sequencing

Sequencing platforms differ in the amount of data generated, and the maximum read lengths of the outputted DNA sequences (Quince *et al*., 2017). Currently, the most widely-used platforms are the Illumina sequencing instruments, including the Illumina MiSeq series and the Illumina NextSeq benchtop instrument, with previous studies also utilizing the now discontinued Illumina HiSeq series (Illumina, 2021; Quince *et al*., 2017). The Illumina MiSeq instrument is able to generate a maximum read length of 2 X 300 bp per read, with the other instruments being limited to 2 X 150 bp(Illumina, 2021; Quince *et al*., 2017). The NextSeq 1000 and 2000 can generate a maximum of 1.1 billion reads per run, while the others can generate up to 4 million, 25 million, or 400 million reads per run. Production-scale Illumina products, such as the NovaSeq 6000, can generate up to 20 billion reads per run, at a read length of 2 X 250 bp (Illumina, 2021). The maximum read length of DNA fragments greatly impacts the assembly process; for example, larger read lengths are easier to assemble than smaller read lengths(Quince *et al*., 2017). The aforementioned platforms can perform both single-end (SE) and paired-end (PE) sequencing;  PE-sequencing is more commonly used, since it allows researchers to elucidate the distance between DNA fragment ends (Almeida and De Martinis, 2019). The amount of data generated is impacted by sequencing depth, which refers to the number of times that a DNA fragment may be sequenced in a sample. Differences in coverage and sequencing depth determine the tools that can be used for data processing; for example, high coverage sequencing data can be both computationally intensive and time-consuming to analyze, which influences the type of metagenomic assemblers that would be most compatible for analysis (Quince *et al*., 2017).

### 1.3.4 Metagenomic quality control and assembly

DNA fragments generated with shotgun metagenomics are processed for quality control, by removing adapters and trimming low-quality reads(Ghurye *et al*., 2016). Data quality is assessed with tools such as FASTQC, and trimmed with tools such as Trimmomatic,, Cutadapt, and Trim Galore (Andrews, 2010; Bolger *et al*., 2014; Krueger, 2015; Martin, 2011). These tools use the PHRED score algorithm to assess quality, which determines the probability that a nucleotide has been inaccurately incorporated into sequencing reads (Almeida and De Martinis, 2019). Other parameters of quality control include GC-content, levels of sequence duplication, and sequence lengths (Andrews, 2010).

Trimmed data can be assembled into contiguous segments, or contigs, in a method known as assembly (Breitwieser *et al*., 2017). There are several advantages to using metagenomic assemblies as opposed to short-read analyses. Assemblies can be used to distinguish open reading frames, thus providing genomic context, and better predictions of phenotypes and metabolic activity. There are two types of metagenomic assembly methods, known as *de novo* or reference-based assembly. The former reconstructs the metagenome using only overlapping reads from raw sequencing data, and the latter uses previously-sequenced genomes as a reference for comparative assembly (Ghurye *et al*., 2016). *De novo* genomic assembly is a computational problem that cannot be solved efficiently, and thus requires heuristic methods (Ghurye *et al*., 2016). The three paradigms of *de novo* genomic assembly are Greedy, Overlap-Layout-Consensus, and de Bruijn graph. While Greedy algorithm assembly is not commonly used, Overlap-Layout-Consensus assembly is typically used for low coverage, longer reads with higher error rates, such as those produced by Pacific Biosciences or Oxford Nanopore sequencing instruments (Ghurye *et al*., 2016).

De Bruijn graph assemblers are currently the most commonly used, and are well-suited for the high coverage, short reads produced by Illumina sequencing instruments (Ghurye *et al*., 2016). This method is relatively more computationally efficient, which is an important factor when analyzing large amounts of sequencing data. The de Bruijn graph algorithm uses the overlaps and relationships between fixed-length substrings, or *k*-mers, found in the sequence reads (Ghurye *et al*., 2016). Shared *k*-mers are used to infer overlapping sequences and create a de Bruijn graph, which can then be resolved by finding a Eulerian path(Ghurye *et al*., 2016). Essentially, these overlapping reads are assembled into multiple branches of contiguous sequences, or contigs, which are then aligned into consensus sequences; gaps between different contigs are known as scaffolds, which can be used to order the contigs (Almeida and De Martinis, 2019). *K*-mer sizes impact the quality of the contigs and the finally assembly, with short *k*-mer sizes being ideal for extracting low-abundance genomes, and long *k*-mer sizes being ideal for higher quality contigs (Quince *et al*., 2017). Most metagenomic assemblers—such as IDBA-UD, MegaHit, and MetaSPAdes—use multiple *k*-mer sizes, thus circumventing the problem of having to choose the ideal *k*-mer size (Li *et al*., 2015; Nurk *et al*., 2017; Quince *et al*., 2017). Notably, the ability of assemblers to accurately create contigs is influenced by the relative abundances of microbial genera in these samples, in addition to the strain diversity of the species (Quince *et al*., 2017).

Reference-based assembly takes advantage of the fully-sequenced genomes available in genomic databases (Ghurye *et al*., 2016), used as reference sets, against which raw sequencing reads are aligned to create a consensus sequence. This kind of assembly is useful for datasets with low coverage or with many low-abundance organisms since these include many incomplete genomes that would be difficult to assemble (Ghurye *et al*., 2016). The quality of reference-based assembly relies on the quality of reference datasets, limiting its effectiveness for microbes where no reference genomes are available (Ghurye *et al*., 2016). Despite this, reference datasets can be a powerful tool to validate the quality of a *de novo* assembly, with the two approaches complementing each other when analyzing metagenomic data.

In addition to the different types of assembly algorithms and methods, there are also different strategies for employing these assembly methods. The two strategies for metagenomic assembly are single-sample assembly, where each sample is assembled individually, or co-assembly, where multiple samples are assembled together. While there is a lack of consensus on which of these two methods is beneficial, previous work has shown that co-assembling can be a useful technique when the experimental design includes multiple samples from the same site or organism, at either different time points or under different experimental conditions (Brown *et al*., 2013; Chen *et al*., 2020). In this case, time-series samples can provide additional data for bacterial strains unique to the community, potentially making it easier to resolve low abundance strains in the assembly.

While co-assembly is a useful strategy in some instances, a 2019 study reported that sample co-assembly can lead to consensus genomes, where strain-specific differences are often lost (Pasolli *et al*., 2019). For this reason, the authors reported that that co-assembly is less useful for cross-sectional data, and most useful for longitudinal data when there are more than five time-series samples from the same subject (Pasolli *et al*., 2019). Even in this case, if the goal of a study is to evaluate strain-specific evolution over time, co-assembly may not be the best strategy. However, co-assembly remains a widely used tool for metagenomic assembly, especially to recover genomic information from low abundance organisms in microbial communities.

### 1.3.5 Metagenomic assemblers and strain diversity

Metagenomic assembly is difficult due to a number of challenges. First, assembly algorithms cannot differentiate between genetic repeats from sequencing errors and repeats due to genuine biological differences (Ghurye *et al*., 2016). Secondly, less abundant organisms will have lower coverage in the sequencing data and may often be lost during the assembly process. One solution to this problem is to conduct depth normalization to correct for uneven coverage (Brown *et al*., 2012; Ghurye *et al*., 2016). Third, metagenomic assembly is computationally intensive, requiring a large amount of memory and generating large amounts of data that need to be stored (Ghurye *et al*., 2016). And lastly, the quality of assemblies is difficult to determine since it relies on assembly statistics, such as length of contigs, which do not necessarily equate with how well the assembly is representing the true genomic sequence. Conventionally, assembly quality has been determined using assembly statistics that rely on the length of assembled contigs. One such assembly statistic is the N50 value, or the median contig size, which represents the size of the largest contig in the genomic sample, such that all the contigs larger than the N50 are the sum of the genome size (Ghurye *et al*., 2016). However, assembly statistics that rely on contig size assume that longer contigs have better quality, which may not necessarily be true. Furthermore, metrics such as the N50 cannot be applied to metagenomic samples, since the size of the metagenome is not known(Ghurye *et al*., 2016).

### 1.3.6 Metagenomic assembly-based and reads-based analyses

The raw metagenomic sequences and assemblies can be used for a number of reads-based and assembly-based analyses. Taxonomic annotation is used to elucidate the relative abundances of microbial species in the microbiota sample (Prakash and Taylor, 2012). In the context of the infant gut microbiota, conducting these analyses at different time-points could shed light on the progression in microbial colonization throughout infancy. For example, this could include variances in the levels of keystone taxa, such as *Bifidobacterium* species. Taxonomic profiling is typically conducted on raw reads rather than the assembled metagenome (Quince *et al*., 2017). Lowest common ancestor (LCA) strategies are used to provide taxonomic classifications or functional annotations for metagenomic datasets with short reads, and include tools such as MEGAN (Huson *et al*., 2007). Reads are run through BLAST, and each identified gene is matched to a node on the NCBI taxonomy; reads are then assigned to the lowest common ancestor of the species known to have the identified gene. Marker gene-based approaches, such as mOTU and MetaPhlAn, enable phylogenetic and taxonomic analysis of a community by using conserved marker genes found in microorganisms(Sunagawa *et al*., 2013; Truong *et al*., 2015). Profilers such as Kaiju and Diamond instead use protein-coding sequences(Buchfink *et al*., 2015; Menzel *et al*., 2016).  On the other hand, tools such as Kraken and LMAT(Ames *et al*., 2013; Quince *et al*., 2017; Wood and Salzberg, 2014) use *k*-mer based identification in order to determine the taxonomy within a sample. Taxonomic classification prior to assembly can circumvent challenges such as the potential loss of low-abundance organisms caused by metagenomic assembly. However, since this approach relies entirely on reference databases, it can also make it difficult to profile microbes that have not been previously characterized(Quince *et al*., 2017).

Functional profiling is typically done after reads have been assembled into contiguous sequences. Functional annotation assigns known functions to genes in the metagenome and can be used to understand the activity of the microbiota(Prakash and Taylor, 2012). For example, the number and type of CAZymes in a metagenomic profile from an infant stool sample might differ between exclusive breastfeeding and consuming solid food. Most tools for functional annotation use homology-based approaches, while others use motif- or pattern-based approaches (Prakash and Taylor, 2012). When this data is combined with taxonomic information, genes can be linked to specific gut microbes (Prakash and Taylor, 2012). Predicted taxonomic, functional, and metabolic profiles can be combined with clinical data about each individual in order to understand the impact of different variables on the gut microbiome composition (Prakash and Taylor, 2012). Common post-assembly functional profiling tools include Prokka, Prodigal, and dbCAN, while reads-based profiling tools include HUMAnN3, MEGAN, and MG-Rast (Beghini *et al*., 2021; Beier *et al*., 2017; Hyatt *et al*., 2010; Meyer *et al*., 2019; Seemann, 2014; Yin *et al*., 2012).

### 1.3.7 Obtaining metagenome-assembled genomes from metagenomic assemblies

An assembled metagenome can be “binned” to computationally isolate full bacterial genomes from metagenomic data. Binning refers to the approach of assigning assembled contigs to draft genomes, or bins, based on identifying characteristics or signals within the genome (Chen *et al*., 2020; Roumpeka *et al*., 2017). Characteristics that are used for accurate binning include read coverage, sequence or tetranucleotide composition, or the taxonomic identity of single-copy bacterial marker genes within each scaffold (Chen *et al*., 2020). Binning highly abundant organisms in a sample is sometimes a relatively simple task, given the large number of fragments that share the same genomic characteristics, such as similar coverage within the sample, unique tetranucleotide frequencies, similar GC content, and consistent phylogenetic identities (Chen *et al*., 2020). However, binning becomes more complicated where there is increased within-species diversity, making it difficult to separate closely related genomes from each other. Often, if an experimental design includes multiple samples from the same site or organism at different timepoints, the time-series samples can provide a unique strategy for binning, where the shared patterns can be taken advantage of (Brown *et al*., 2013; Chen *et al*., 2020). Most binning tools use a combination of these aforementioned sample features; some examples of binning tools include CONCOCT, MetaBat, and MaxBin (Kang *et al*., 2015; Lu *et al*., 2017). With software pipelines such as Das Tool and MetaWrap, it is also possible to use multiple binning tools, evaluate the quality of bins from each tool, and aggregate the results into the best possible bins (Sieber *et al*., 2018; Uritskiy *et al*., 2018).

High-quality bins that may serve as draft genomes are known as metagenome-assembled genomes, or MAGs. The method of retrieving MAGs from metagenomic data is known as genome-resolved metagenomics. There are no strict rules in the field regarding when a “bin” becomes a “MAG”, or when a “MAG” becomes a “complete MAG” (CMAG).  Many studies report complete genomes if their MAGs contain all the essential marker genes expected to be found within the microbe that the MAG belongs to (Chen *et al*., 2020). However, using this metric to evaluate genome completeness disregards other important quality factors, such as fragmentation, gaps, or mis-assemblies. A recent 2020 paper attempted to formulate guidelines for MAG reporting (Chen *et al*., 2020). According to the authors, a MAG should be considered complete if it has (1) a single, circular chromosome, (2) “perfect” read coverage support, and (3) no gaps (Chen *et al*., 2020). If the reported MAG meets these criteria, it is a complete MAG, or a CMAG. It is very rare to obtain a CMAG from metagenomic data, especially when using short paired-end reads. As such, as of 2019, there were only 59 microbial CMAGs reported in publicly available datasets, with 36 of those belonging to microbes that have smaller than average genomes (Chen *et al*., 2020).

Genome-resolved metagenomics provides irreplaceable benefits in allowing for the study of genes and metabolic pathways in low-abundance organisms, and in previously uncultivated bacteria. This protocol has offered a solution to the bottleneck of culture-dependent approaches, where organisms had to be isolated and cultured prior to sequencing (Chen *et al*., 2020). MAGs can be a good alternative to the time-intensive task of sequencing bacterial isolates. Furthermore, good quality MAGs have led to various scientific breakthroughs, such as the discovery of "Candidate Phylum Rokubacteria", a previously unknown lineage on the tree of life (Becraft *et al*., 2017; Chen *et al*., 2020). Another example is the characterization of the complete nitrification pathway of the genus *Nitrospira*, made possible by the identification of a single gene from a MAG (Chen *et al*., 2020; Daims *et al*., 2015). Furthermore, binning to retrieve MAGs is an important part of metagenomic data analysis, since using assembly-level data as a proxy for microbial communities can lead to incorrect interpretations (Chen *et al*., 2020). For example, a 2017 study that conducted a contig-level analysis erroneously claimed to have identified "hundreds" of novel microbes in the human blood metagenome, using single-copy genes (Chen *et al*., 2020; Kowarsky *et al*., 2017). Analysis of the same data after binning revealed that the single-copy genes clustered around a small number of contigs, belonging to the superphylum Parcubacteria (Chen *et al*., 2020). Rather than hundreds of novel microbes, the dataset contained just one.

However, to obtain biologically relevant information, a genome must be as complete as possible. Due to gaps in metagenomic assemblies, errors or misassemblies, chimeras, contamination, and strain diversity, achieving a fully complete MAG from metagenomic data is not only challenging, but also rare (Chen *et al*., 2020). Some limitations of using incomplete, draft MAGs include missing information about genes, loss of gene order, and difficulty in differentiating between chromosomes and plasmids (Chen *et al*., 2020). Furthermore, MAGs may be “composite” or “consensus” genomes, without the representative within-species variation that exists in the complex community that they were reconstructed from (Chen *et al*., 2020; Pasolli *et al*., 2019). Low quality, fragmented, or composite genomes are more common when the samples are complex and heterogenous, and less common when the within-sample diversity is low (Chen *et al*., 2020). There is also the danger of contigs being mis-binned, or genomic sequences being inserted into a bin belonging to the incorrect organism (Chen *et al*., 2020). The standard way to determine a bin’s quality is with the CheckM software package, which uses the occurrence of genus-specific bacterial single-copy genes to determine the % completeness of a bin, along with the % contamination within the bin (Parks *et al*., 2015). Unfortunately, this is not a robust method to determine MAG quality, and relying solely on single copy genes can lead to erroneous conclusions. For example, if the MAG contains mis-binned regions that do not contain any single-copy genes, this contamination would never be detected (Chen *et al*., 2020).

### 1.3.8 Challenges of studying the infant gut microbiome

Compared to most environmental or adult gut microbiome samples, the infant gut microbiome is relatively simple, with a few abundant microbial species in each sample (Milani *et al*., 2017a). While this may initially appear as an advantage, it amplifies one of the previously mentioned challenges of metagenomic assembly and binning. The abundant microbial species in the infant gut often have many species variants, predictably causing confusion between sequence repeats and creating multiple assembly graphs that must be reconciled (Breitwieser *et al*., 2017). These challenges are caused by the abundance and diversity of genera such as *Bifidobacterium* (Milani *et al*., 2017a) and *Bacteroides* in the gut.

Specifically, *Bifidobacterium* species have a large pan-genome. As a result, in addition to the conserved regions of high similarity in each *Bifidobacterium* strain, they also have pan-genomic regions with many unique genes. This can potentially cause errors in metagenomic assembly, where unique regions may be un-assembled or misassembled. Lastly, highly abundant taxa in the infant gut microbiota can often mask low-abundance species.

### 1.3.9 Previous work in assessing metagenomic analysis quality with simulated data

Given the possibility of erroneous and incomplete results at the major stages of taxonomic prediction, assembly, binning, and gene annotation, several studies have attempted to validate the accuracy of metagenomic tools. The Critical Assessment of Metagenome Interpretation (CAMI) challenge benchmarked multiple assembly and binning tools with simulated metagenomic data(Sczyrba *et al*., 2017). Unsurprisingly, the study found that the overall quality of final bins varied largely based on community complexity; average genome completeness could be as low as 38% (Sczyrba *et al*., 2017). A recent paper also evaluated the performance of different methods specifically on data assembled with MetaSPAdes, finding that most binners were not well-optimized for similar strains (Yue *et al*., 2020). There have also been attempts to quantify the performance of MAG binning tools for plasmids and genomic islands, with results showing that less than 30% of plasmids and 45% of genomic islands could be identified with good coverage in the genome they originally belonged to (Maguire *et al*., 2020). Furthermore, when it came to predicting antimicrobial resistance (AMR) genes in MAGs, only 53% of chromosomal, 16% of plasmid, and 45% of genomic island AMR genes could be identified (Maguire *et al*., 2020).

It is important to note that a major caveat in many benchmarking studies is the lack of consideration towards the difference in how values are reported by different tools (Sun *et al*., 2021). One example of this is bacterial abundance values, which are not reported the same way across different taxonomic profilers. *K*-mer-based profilers report sequence abundance, while marker gene-based profilers report taxonomic abundance (Sun *et al*., 2021). A recent study used 144 simulated metagenomic samples to test the performance of taxonomic profiling pipelines, finding that *k*-mer based profilers, such as Centrifuge and Kraken, outperformed marker gene-based profilers, such as MetaPhlAn2 (Miossec *et al*., 2020). However, given the fact that Centrifuge and Kraken report sequence abundance, while MetaPhlAn2 reports taxonomic abundance, these tools cannot be directly compared. Depending on which abundance value is being evaluated, different profilers will outperform each other. These considerations are incredibly important when validating protocols, and several tutorials have been published to establish best practices for benchmarking bioinformatics tools (Bokulich *et al*., 2020; Meyer *et al*., 2021).

## 1.4 Research Paradigm

### 1.4.1 Purpose

The purpose of this thesis is to understand the effectiveness of current metagenomic protocols on infant gut samples, and to provide a framework and bioinformatics pipeline for future analyses of infant gut metagenomic data. Particularly, due to the persistence of *Bifidobacterium* in the gut both before and after solid food introduction, we want to understand the effectiveness of these pipelines in samples dominated by *Bifidobacterium*. By providing a framework for analyzing these samples, we aim to provide a consistent protocol for mapping functional and metabolic activity for genera with high within-species diversity, in addition to low abundance organisms that might be masked by overrepresented genera.

### 1.4.2 Research Question

Based on the gaps identified in the metagenomic analysis of infant gut microbiome data, we had the following research questions:

**Research question I:** How effective is a standard metagenomic workflow in reconstructing high quality metagenome-assembled genomes from infant gut metagenomic samples? How do microbial community composition and metagenomic assembly quality impact the accuracy of strain-level CAZyme predictions in MAGs?

I. **Research question II:** Can we build a metagenomic analysis pipeline geared towards attaining high quality metagenome-assembled genomes and CAZyme predictions from infant gut microbiome data?

II. **Research question III:** How does the number and type of predicted metabolic genes and pathways of abundant and rare bacteria in the infant gut microbiome change over the period of solid food introduction?

### 1.4.3 Aims

I. **Aim I:** Use a simulated metagenomic dataset of infant gut microbiome data to understand and quantify the role of community composition, the presence of closely related strains, and assembly quality on 1) retrieving high quality MAGs from metagenomic data, and 2) accurately predicting CAZymes from assembly-level and MAG-level data.

II. **Aim II:** Based on the results of Objective I, build a reproducible and computationally feasible standard metagenomic pipeline to analyze infant gut microbiome data, and

III. **Aim III:** Use real metagenomic data to understand how the metabolic activity of bacterial species change over the solid food introduction period.

### 1.4.4 Hypotheses

I. **Hypothesis I:** Higher community complexity, as characterized by sample Shannon Diversity Index, observed species count, and high inter-genus diversity, will lead to lower quality MAGs and less accurate CAZyme predictions for abundant genera.

II. **Hypothesis II:** Samples obtained after solid food introduction will display an increased abundance of bacterial strains with the ability to metabolize non-HMO food glycans, and a higher number of carbohydrate degrading genes and pathways.

# 2 Methodology

## 2.1 Data Usage

The data used for the completion of this thesis includes 34 metagenomic samples that were simulated using 16S rRNA gene profiles from the Baby & Mi cohort, and 30 metagenomic samples from 15 infants belonging to the Baby & Mi cohort.

### 2.1.1 Baby, Food & Mi Study

The majority of the data analysis was conducted on samples from the Baby, Food & Mi sub-study, which falls under the Baby & Mi cohort (Dizzell *et al*., 2021). The cohort included 15 infants, who were intensively sampled for approximately two weeks following the introduction of solid foods and during weaning from breastmilk (**Figure 2**). Sample collection began 3-4 days before the introduction of solid foods, and continued daily for 17 days. To be included in the Baby & Mi study, the infants had to be full-term, vaginally born, and breastfed. Exclusion criteria included caesarian section at birth, high-risk pregnancies, and exposure to antibiotics within 28 days of starting the study (Dizzell *et al*., 2021).

### 2.1.2 Shotgun metagenomic sequencing

Paired-end shotgun metagenomic sequencing was conducted on 2 stool samples from each of the 15 infants; the first sample was immediately before the introduction of solid food, while the second sample was towards the end of the collection period **(Figure 2**). 100mg of solid stool was used during the DNA extraction; when applicable, stool was collected from the diaper liner instead (Dizzell *et al*., 2021; Homann *et al*., 2021; Stearns *et al*., 2017). The DNA extraction was involved mechanical lysis with ceramic beads that were 2.8 mm and glass

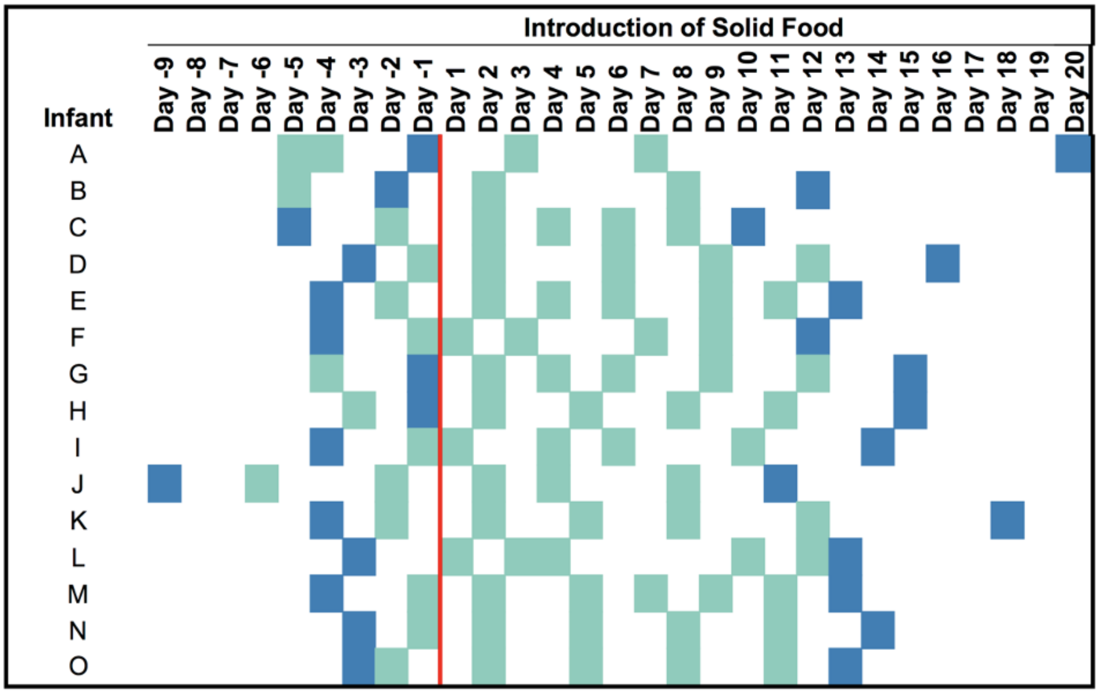


**Figure 2: All Baby, Food & Mi samples collected during the introduction to solid foods. Green squares represent days on which stool samples were collected. Blue squares represent samples that were sent for metagenomic sequencing. The red line represents the day that solid foods were introduced to the infant diet.**

beads that were 0.1 mm, for a total of 3 minutes at 3000 rmp, in sodium phosphate monobasic and guanidinium thiocyanate EDTA N-lauroylsarcosine buffer (Homann *et al*., 2021; Stearns *et al*., 2015, 2017). The MagMAX-96 DNA Multi-Sample Kit was used to purify the DNA extracts, using the MagMAX Express-96 Deep Well Magnetic Particle Processor (Homann *et al*., 2021).

2 samples belonging to the same infant, JCSA4 and JCSA5, were sequenced at the Genomics Core Facility at Dalhousie University, with the Illumina NextSeq 500 instrument. Paired-end reads were 150 bp in length, and the two samples had a read count of 97,762,039 bp for JCSA4 and 21,377,794 bp for JCSA5. The remaining 28 samples were sequenced at the McMaster Genome Facility, with the Illumina Hiseq1500 instrument. Paired-end reads were 250 bp, and each forward and reverse file had a mean read count of 9,809,250 bp. JCSA4 and JCSA5 were sub-sampled to 17 million reads to accommodate the differences in read count and length.

### 2.1.3 16S rRNA gene Profiles

34 infant samples from Baby, Food & Mi study were chosen at random for the metagenomic simulation, using 16S microbial profiles from infant stool, which were based on the v3 region of the 16S rRNA genes as part of the Baby, Food & Mi study (Dizzell *et al*., 2021). Originally, libraries were sequenced with 2x250 bp reads on the Illumina MiSeq instrument. Adapter, primer, and barcode sequences were trimmed from sequencing reads with cutadapt and ASVs (amplicon sequence

variants) were inferred with the Divisive Amplicon Denoising Algorithm 2 (DADA2) package (Callahan *et al*., 2016; Homann *et al*., 2021; Martin, 2011) Taxonomy was assigned to ASV sequences using the Divisive Amplicon Denoising Algorithm 2 (DADA2) package, which used the RDP Bayesian classification method against the Silva 2013 full length 16S rRNA gene tree (Callahan *et al*., 2016; Homann *et al*., 2021; McDonald *et al*., 2012; Wang *et al*., 2007). Abundance values and taxonomic metadata were exported into BIOM format using QIIME2. Samples were rarefied to achieve even sample depth, and relative abundance plots were made with the phyloseq package (v1.32.0) in R (McMurdie and Holmes, 2013).

### 2.1.4 CAMISIM metagenomic simulations from infant gut 16S rRNA gene profiles

The Critical Assessment of Metagenome Annotation (CAMI-) Simulator (SIM) was used to simulate the 34 metagenomes, using taxonomic profiles derived from the 16S rRNA gene profiles (Fritz *et al*., 2019; Sczyrba *et al*., 2017). CAMISIM inputs included the summary .biom file, along with a configuration file with paths to all samples, relevant databases, and additional specifications. Since 16S profiles were used for the simulation, CAMISIM was run for both community design and read simulation. The ART simulator was chosen within CAMISIM to generate Illumina HiSeq 150bp reads (Huang *et al*., 2012). The mean fragment size for the paired-end reads was 270, with a standard deviation of 27, while the genome masking 'N' cutoff frequency was one in 150. As a part of the simulation, sequencing errors that are common to the Illumina platform were also introduced to the reads. CAMISIM was run without anonymization, with a random seed. A maximum of three strains were set to be simulated per ASV. In addition to generating raw metagenomic reads, the simulator was also used to generate individual gold-standard assemblies and a pooled gold-standard co-assembly Once the raw forward and reverse reads of sample-specific genomes had been generated, the resulting reads were concatenated together and shuffled for each sample using BBMap (Bushnell, 2014). CAMISIM taxonomy, abundance, and community data for all samples were formatted with labdsv package in R. Relative abundance was calculated using the CAMISIM-generated absolute abundance and taxonomy files for each sample. Absolute counts were transformed as a proportion of the total counts. To create relative abundance stacked bar charts, the data was filtered to the Order level.

## 2.2 Software availability

All bioinformatics analyses and pipelines used and created for this thesis are outlined in **Table 1**.

**Table 1: Bioinformatics analyses and GitLab repositories for all analyses conducted in this thesis**

| METHOD SECTION | RESULTS SECTION | ANALYSIS | GITLAB REPOSITORY |
|---|---|---|---|
| 2.3 | Chapter 1 | mg_workflow, a pipeline to analyze | **mg_workflow**, accessible at: |

| | | | |
|---|---|---|---|
| | | simulated metagenomic samples | https://gitlab.com/bhavyasingh/mg_workflow |
| 2.4 | Chapter 1 | Documentation for post-pipeline simulated data analysis | **metagenomics_simulation_project**, accessible at: https://gitlab.com/bhavyasingh/metagenomics_simulation_project |
| 2.5 | Chapter 2-3 | MAGPIE, a minimal and customizable metagenomic pipeline for infant gut samples | **MAGPIE** (MetAGenomics PipelInE), accessible at: https://gitlab.com/bhavyasingh/magpie |
| 2.6 | Chapter 2-3 | bfm_mg_flow, An instance of the MAGPIE pipeline, modified for Baby, Food & Mi metagenomic samples. | **bfm_mg_flow**, accessible at: https://gitlab.com/bhavyasingh/bfm_mg_flow |

## 2.3 mg_workflow: Snakemake pipeline to derive metagenome-assembled genomes from simulated data

The simulated samples were analyzed with a custom pipeline, mg_workflow, made with Snakemake (v5.23.0) (Koster and Rahmann, 2012). The Snakefile used for the metagenomic data analysis is accessible on GitLab (**Table 1**). In addition to software-specific outputs, documentation includes software benchmarks, log files, and version numbers for each run of the pipeline. Any instances where post-workflow analysis had to be conducted are documented in the post-processing GitLab repository, metagenomics_simulation_project (see section 2.4).

### 2.3.1 Snakemake pipeline environments

Prior to generating the pipeline, Miniconda3 environments were created for the various software used in the analysis. Environments were exported as .yaml files, which were then used by Snakemake to carry out the individual rules. A total of 9 environments were created, for pre-assembly quality control, metagenomic assembly with MEGAHIT and MetaSPAdes, post-assembly quality control, mapping, binning, post-binning quality control, BLAST alignments, and CAZyme predictions. All .yaml files for the mg_workflow pipeline are also accessible on GitLab.

### 2.3.2 Metagenomic assembly

Forward and reverse reads were first checked for quality with FastQC (v0.11.9). Reads were trimmed using Trimmomatic (v0.39) with default parameters and all files were again checked for quality with FastQC (Andrews, 2010; Bolger *et al*., 2014). The samples were individually assembled with MEGAHIT and MetaSPAdes (Li *et al*., 2015; Nurk *et al*., 2017). All samples were then concatenated for co-assembly with MEGAHIT and MetaSPAdes. Contigs from the MEGAHIT, MetaSPAdes, and gold-standard individual assemblies were simplified using Anvi’o (v6.2.0) (Eren *et al*., 2015). Assembly quality was evaluated using MetaQUAST, with a custom reference database consisting of all genomes used to generate the simulated reads (Mikheenko *et al*., 2016).

### 2.3.3 Contig binning and identification

All assemblies, including the gold-standard assemblies, were indexed using bowtie2-build (v2.4.1), with bowtie2 also being used to map raw reads back to the assemblies (Langmead and Salzberg, 2012). The resulting .sam files were converted to .bam format, sorted, and indexed using samtools (v1.11) (Li *et al*., 2009). Assembly coverage information and paired contigs were calculated and assemblies then binned with MetaBat2 (v2.2.15) (Kang *et al*., 2015). Finally, the checkm lineage_wf (v1.0.16) pipeline was used to identify bin identity, and to assess bin completeness, contamination, and strain heterogeneity (Parks *et al*., 2015). CheckM results were exported for further analysis in R.

### 2.3.4 BLAST alignment of bins to references

A BLAST database (v2.10.1+) was made using all CAMISIM reference genomes, along with a genome mapping file to link contigs to their respective accession numbers and taxonomic IDs(Madden, 2013; Ye *et al*., 2006). Bins were queried against the database with BLASTN, with default parameters. The results were outputted with output format 6, and included the following output metrics: query accession and version, subject accession and version, query length, subject length, percent identity, length, mismatches, gaps, length, mismatches, number of gap openings, query start, query end, subject start, subject end, e-values, bit scores, taxonomic IDs, scientific names, query coverage, query coverage per subject, and query coverage per HSP.

### 2.3.5 CAZyme predictions

CAZymes were predicted using dbCAN (v 2.0.11), with default settings (Yin *et al*., 2012), using the 2020 update of the dbCAN databases. Since dbCAN uses DIAMOND, Hotpep, and Hmmer for predicting CAZymes in the query sequences, the outputs include a concatenated file of predictions from all three methods. The confidence of a CAZyme prediction is based on the number of tools that identified the gene in the queried contig. For the assembly-level analysis, dbCAN was run on all MEGAHIT, MetaSPAdes, and gold-standard single-assemblies. For the MAG-level analysis, dbCAN was run on all predicted bins, in addition to all the reference genomes used for the simulation. The gold-standard assemblies and reference genomes served as positive controls for the CAZyme analysis.

## 2.4 Post-pipeline data analysis of actual and gold-standard results for simulated data

Once the bins had been retrieved from the simulated data, all post-pipeline analyses were carried out in R. This included comparison of assembling, binning, and CAZyme prediction performance across samples with different community compositions, and across different assembly and binning methods (**Figure 3)**.

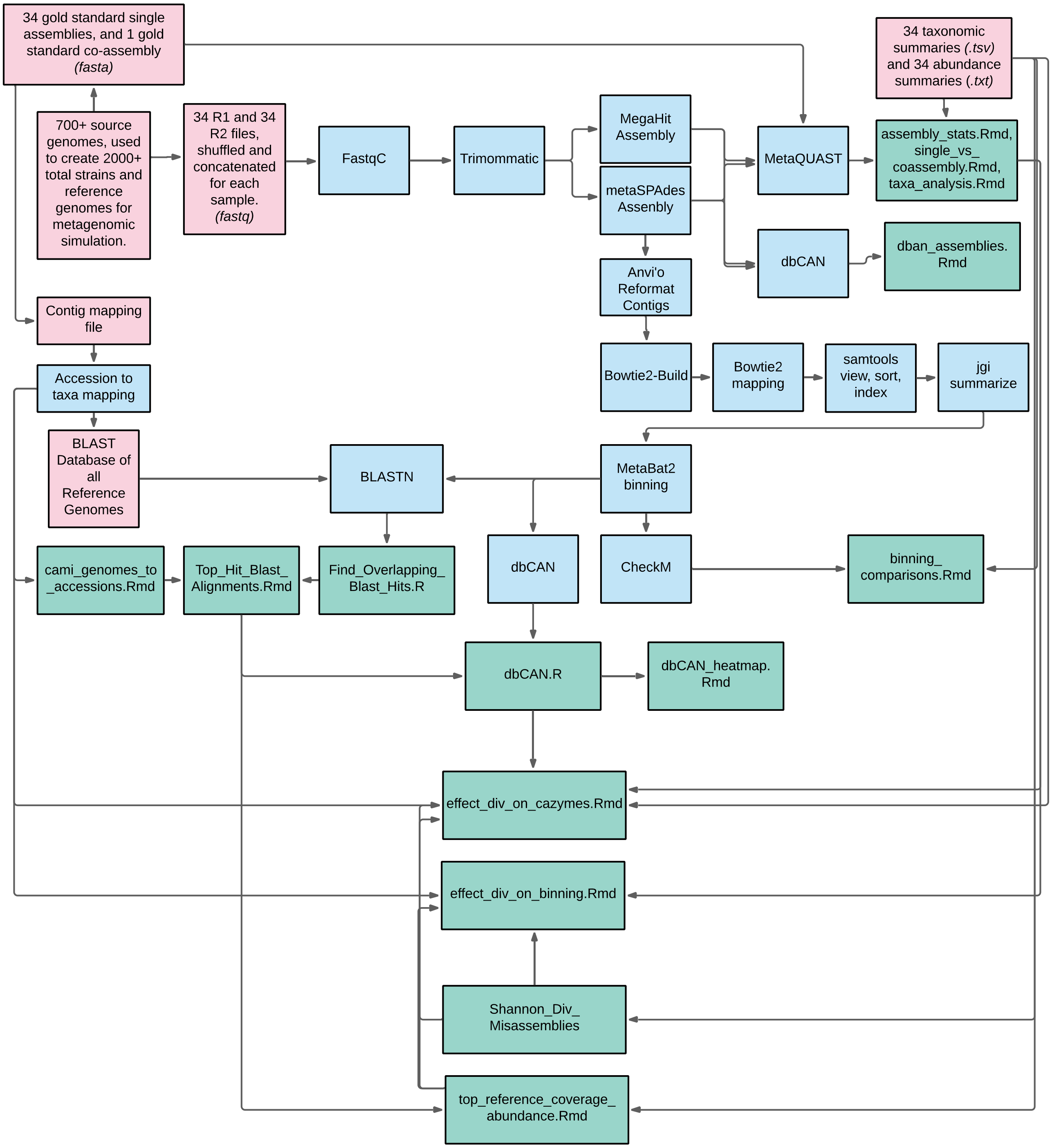

34 gold standard single assemblies, and 1 gold standard co-assembly (fasta)
700+ source genomes, used to create 2000+ total strains and reference genomes for metagenomic simulation.
34 R1 and 34 R2 files, shuffled and concatenated for each sample. (fastq)
FastqC
Trimommatic
MegaHit Assembly
metaSPAdes Assembly
MetaQUAST
34 taxonomic summaries (.tsv) and 34 abundance summaries (.txt)
assembly_stats.Rmd, single_vs_coassembly.Rmd, taxa_analysis.Rmd
dbCAN
dban_assemblies.Rmd
Anvi'o Reformat Contigs
Contig mapping file
Accession to taxa mapping
Bowtie2-Build
Bowtie2 mapping
samtools view, sort, index
jgi summarize
BLAST Database of all Reference Genomes
BLASTN
MetaBat2 binning
cami_genomes_to_accessions.Rmd
Top_Hit_Blast_Alignments.Rmd
Find_Overlapping_Blast_Hits.R
dbCAN
CheckM
binning_comparisons.Rmd
dbCAN.R
dbCAN_heatmap.Rmd
effect_div_on_cazymes.Rmd
effect_div_on_binning.Rmd
Shannon_Div_Misassemblies
top_reference_coverage_abundance.Rmd

**Figure 3: Summary of post-pipeline data processing after retrieving main outputs from mg_workflow.** Pink boxes refer to raw data outputted from CAMISIM. Blue boxes refer to metagenomic tools used for data processing within a metagenomic pipeline. Green boxes refer to R scripts used for the analysis.

### 2.4.1 Generating sample and species mapping files

The unique identifiers for simulated strains included the following: 1) accessions of reference genomes used for the simulation, and 2) unique CAMI identifiers for each strain created for the reference genomes. The accessions were used to download NCBI Taxa IDs and full scientific names for each strain, using the Entrez Direct command line tools (Kans, 2010) The CAMISIM output generated sample mapping files, which outlined the reference genomes used to create each CAMI strain. Abundance files were mapped to taxonomic profiles, Taxa IDs, and strain identifiers to create a master sample identification file.

### 2.4.2 Alpha Diversity and Beta Diversity

Shannon Diversity Index, Simpson Diversity Index, and Observed Species for 1) the whole dataset and 2) the *Bifidobacterium* genus were calculated with the vegan and phyloseq packages (Dixon, 2003; McMurdie and Holmes, 2013). Prior to calculating sample diversity, all organisms that were not present in any samples were removed from the analysis. Beta diversity was calculated with the phyloseq package, using the ordinate function, with the Bray-Curtis Dissimilarity method.

### 2.4.3 Assembly and binning quality metrics

Assembly quality metrics (total contig numbers, the size of the largest contigs, misassembled regions, N50, NA50, etc) were derived from the MetaQUAST tabular output, and imported into R. Similarly, binning quality metrics (completeness, contamination, strain heterogeneity, number of marker genes identified), were derived from the CheckM tabular (.tsv) outputs. Bin quality was calculated as the % completeness - % contamination of each bin. All raw tabular output files used for the post-pipeline analyses are also accessible in the metagenomics_simulation_project GitLab repository.

### 2.4.4 Calculating bin reference coverage

The bin reference coverage is defined as the amount that a bin covers the reference genome of its identified organism. BLASTN alignment results were used to calculate the coverage of the original reference genomes in the bins, as per previous methods (Maguire *et al*., 2020). The script Find_Overlapping_Blast_Hits.R was used to establish the identity of contigs in a bin. Regions that aligned to the same taxa were merged, and taxa that represented the majority of the query were assigned as that contig's identity. Overall, if the majority of the contigs in a bin belonged to the same taxa, it was classified as such. To calculate how much of a reference genome was covered by a bin, the sum of the total length of alignments to the reference genome was divided by the length of the reference genome. Since alignments can often be misleading due to the number of closely related strains in the database, the highest aligned reference genome was only considered to be the identity of the bin if it was directly used to simulate a strain in the sample that the bin was constructed from.

### 2.4.5 Classifying binned and unbinned organisms

Once bin identity had been established, the bins that had been reconstructed for each sample were compared to the abundance and taxonomy of all the organisms in the sample, in order to determine which original strains in each sample were binned or un-binned.

### 2.4.6 Filtering MAGs and assigning MAG identities

Using the CheckM outputs, bins were considered MAGs if: (1) Their overall quality (CheckM % completeness - % contamination) was over 50%, and (2) If a BLASTN alignment of the bin against our local database of reference genomes used in the simulation returned a hit for an organism that was in the same sample that the bin came from.

### 2.4.7 Assembly-level and MAG-level comparison of CAZyme prediction accuracy

The script dbCAN.r was used to concatenate all individual dbCAN results, and to create a combined output to consolidate the DIAMOND, HMMER, and HotPep predictions. Using CAZyme predictions for gold-standard assemblies as the reference, the Caret package in R was used to create confusion matrices comparing gene predictions between the assemblers. Mean true positives, true negatives, false positives, and false negatives for MEGAHIT and MetaSPAdes single assemblies were determined using caret's binary comparisons between predicted and reference values. If a particular CAZyme was predicted in both the reference (gold standard) and prediction (MEGAHIT or MetaSPAdes), then both would have a value of "1", classifying a true positive prediction. Similarly, if the reference had a "0", while the prediction had a "1", then the prediction was a false positive. Similarly, to compare *Bifidobacterium* CAZyme predictions, dbCAN results for all *Bifidobacterium* MAGs and references were filtered out from the combined output. When confusion matrices were made for each *Bifidobacterium* MAG, the original genome was used as the positive control and reference for the matrix. Thus, the CAZyme predictions for each MAG were compared to correct CAZyme predictions from the reference genome that the MAG represented. The predictions for each MAG were formatted as a tabular output, which included the following quality metrics: Prediction Accuracy, Kappa, AccuracyPValue, McnemarPValue, Sensitivity, Specificity, Pos.Pred.Value, Neg.Pred.Value, Precision, Recall, F1, Prevalence. The Pos.Pred.Value refers to the percent of positive predictions that were true positives, while the Neg.Pred.Value refers to the percent of negative predictions that were true negatives.

## 2.5 MAGPIE: A Snakemake MetAGenomic PIpelinE for infant gut microbiome metagenomic metagenomic samples

MAGPIE, or the Infant Gut MetAGenomic PIpelinE, was modeled after the mg_workflow pipeline, which had been created to analyze simulated metagenomic data. The trimming, quality check, assembly, binning, and binning quality steps are common between MAGPIE and mg_workflow. MAGPIE was based on results from **Chapter 1: Evaluating current metagenomic protocols on simulated infant gut samples,** and is outlined in detail in **Chapter 2: Development of a Metagenomic Pipeline**.

### 2.5.1 Binning and bin aggregation

Two different binning tools were used in MAGPIE. In addition to MetaBat2, contigs are binned with MaxBin2 as well, with DAS Tool being used to integrate the results from both binners (Kang *et al*., 2015; Sieber *et al*., 2018; Wu *et al*., 2016). Prior to binning, contig depth, coverage and pairs are calculated for MetaBat2 with the script jgi_summarize_bam_contig_depths. For MaxBin2, coverage is calculated with the script pileup.sh from BBMap. To create the final bins with DAS Tool, the MatBat2 and MaxBin2 bins from each sample are converted to scaffolds using the Fasta_to_Scaffolds2Bin.sh script.

## 2.6 Analysis of Real Metagenomic Data

### 2.6.1 bfm_mg_flow: A cloned instance of MAGPIE for the Baby, Food & Mi Study

An instance of MAGPIE was cloned and customized for infant gut metagenomic samples from the Baby, Food & Mi study, which were longitudinal. The following additional analyses were added to the pipeline:

#### 2.6.1.1 Read-level analyses: trimming, taxonomy, and functional predictions

The first analysis to be added was for read-level trimming, taxonomy, and functional predictions, using the bioBakery3 workflows. Raw FASTQ paired-end reads were processed through the bioBakery3 Whole Metagenome Shotgun (wmgx) pipeline (Beghini *et al*., 2021). The environment and rules used to run the pipeline is available in the bfm_mg_flow GitLab repository. Within the pipeline, reads were trimmed using KneadData, which uses Trimmomatic (v0.36), Bowtie2, TRF, and FastQC. Taxonomic profiles for the trimmed reads were determined using MetaPhlAn3, and functional profiles were determined using HUMAnN3. The final outputs included trimmed and original read counts for all samples, merged taxonomic profiles, merged pathway abundance profiles, and HUMAnN3 alignment and feature counts.

#### 2.6.1.2 Post-processing of read-level analysis

The outputs from the previous analysis were used in the bioBakery3 Whole Metagenome Shotgun Visualization workflow (wmgx_vis) (Beghini *et al*., 2021). The wmgx_vis workflow was run twice, once to compare differences between the Baby & Mi infants by participant ID, and once to compare all infants before and after the introduction of solid foods.

#### 2.6.1.3 Post-pipeline processing of HUMAnN3 metabolic predictions before and after solid food introduction

BioCyc IDs were downloaded for all bacterial carbohydrate degradation genes (Caspi *et al*., 2020). Merged HUMAnN3 outputs were filtered based on metabolic pathways defined by the BioCyc IDs, and visualized before and after solid food introduction using the humann_barplot program, with participant ID and sample timepoint as categorical variables. All outputs were scaled using the pseudolog argument. Merged HUMAnN3 outputs were then filtered based on organisms of interest. Heatmaps of MetaCyc pathways before and after solid food introduction were created with pheatmap.

## 2.7 Statistics

### 2.7.1 Comparison of means

The Shapiro test was used to determine whether data were normally distributed. To compare means, we used Analysis of Variance (ANOVA) for parametric data, and Kruskal-Wallis for non-parametric data. After each ANOVA, if the residuals were not normally distributed, we did a Kruskal-Wallis for confirmation of significance.

### 2.7.2 Linear regressions and comparison of correlations

Multiple linear regression analyses were carried out using the lme4 package (Bates *et al*., 2014). The $R^2$ values for multiple regressions were determined using the MuMIn package. To compare correlations between groups in linear models, we did a Fisher's Z transformation of Spearman correlations for non-parametric data, and Pearson correlations for parametric data (Berry and Mielke, 2000; Bishara and Hittner, 2017).

# 3 Chapter 1: Evaluating current metagenomic protocols on simulated infant gut samples

Given the infant gut community structure of samples dominant in abundant genera with high inter-genus diversity, our first aim was to quantify the effectiveness of standard metagenomic workflows in infant gut metagenomic samples with high within-species diversity of abundant genera. We aimed to answer our first research question: (I) How effective is a standard metagenomic workflow in reconstructing high quality metagenome-assembled genomes belonging to (i) abundant organisms, and (ii) rare organisms, and how do microbial community composition and metagenomic assembly quality impact the accuracy of strain-level CAZyme predictions in MAGs? The research design for this aim was to process simulated reads with a general metagenomic workflow, and to compare

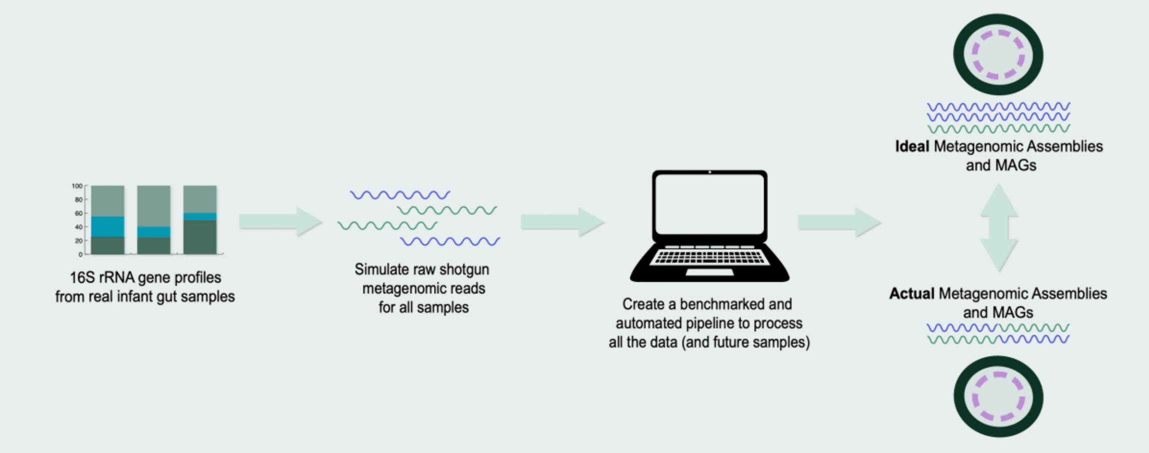

100
80
60
40
20
0
16S rRNA gene profiles from real infant gut samples
Simulate raw shotgun metagenomic reads for all samples
Create a benchmarked and automated pipeline to process all the data (and future samples)
Ideal Metagenomic Assemblies and MAGs
Actual Metagenomic Assemblies and MAGs

**Figure 4: Research design to quantify the performance of standard metagenomic analysis workflows for the infant gut microbiome**

the actual results of the pipeline to the “gold-standard” reference results (**Figure 4**). We simulated infant gut metagenomic samples using 16S rRNA gene profiles, and created a general metagenomic data analysis pipeline, mg_workflow. The simulation included raw metagenomic reads, taxonomic and abundance profiles for each sample, gold-standard single sample assemblies and co-assemblies, and gold-standard MAGs, represented by the reference genomes used to simulate the raw reads. We hypothesized that higher community complexity, observed species count, and high inter-genus diversity would lead to lower quality MAGs and less accurate CAZyme predictions for abundant genera.

## 3.1 34 infant gut metagenomic samples were simulated from 16S rRNA gene profiles

34 infant gut metagenomic samples were simulated with CAMISIM, using 16S rRNA gene profiles from real infant samples (**Figure 5**). 255 reference genomes were used to simulate 405 strains in the simulation, with a maximum of three strains being simulated per species. Within the simulated metagenomic samples, 11 samples were dominant in *Bifidobacterium angulatum*, nine were dominant in *Bifidobacterium longum subsp. infantis ATCC 15697 = JCM 1222 = DSM 20088*, one dominant in *Bifidobacterium longum subsp. infantis 157F*, five dominant in *Escherichia coli O111:H- str. 11128*, four dominant in *Eubacterium eligens ATCC 27750*, and four samples dominant in *Akkermansia muciniphila*, *Bacteroides fragilis*, *Clostridium aceticum*, and *Parabacteroides distasonis ATCC*

*8503* (**Table 2**). 21 samples were dominant in *Bifidobacterium* (**Supplementary Table 1**). The relative abundance of the most dominant strain was over 60% in nine samples, between 40% and 60% in 13 samples, and under 40% in 12 samples (**Table 2**). The observed species count for the simulated metagenomes ranged

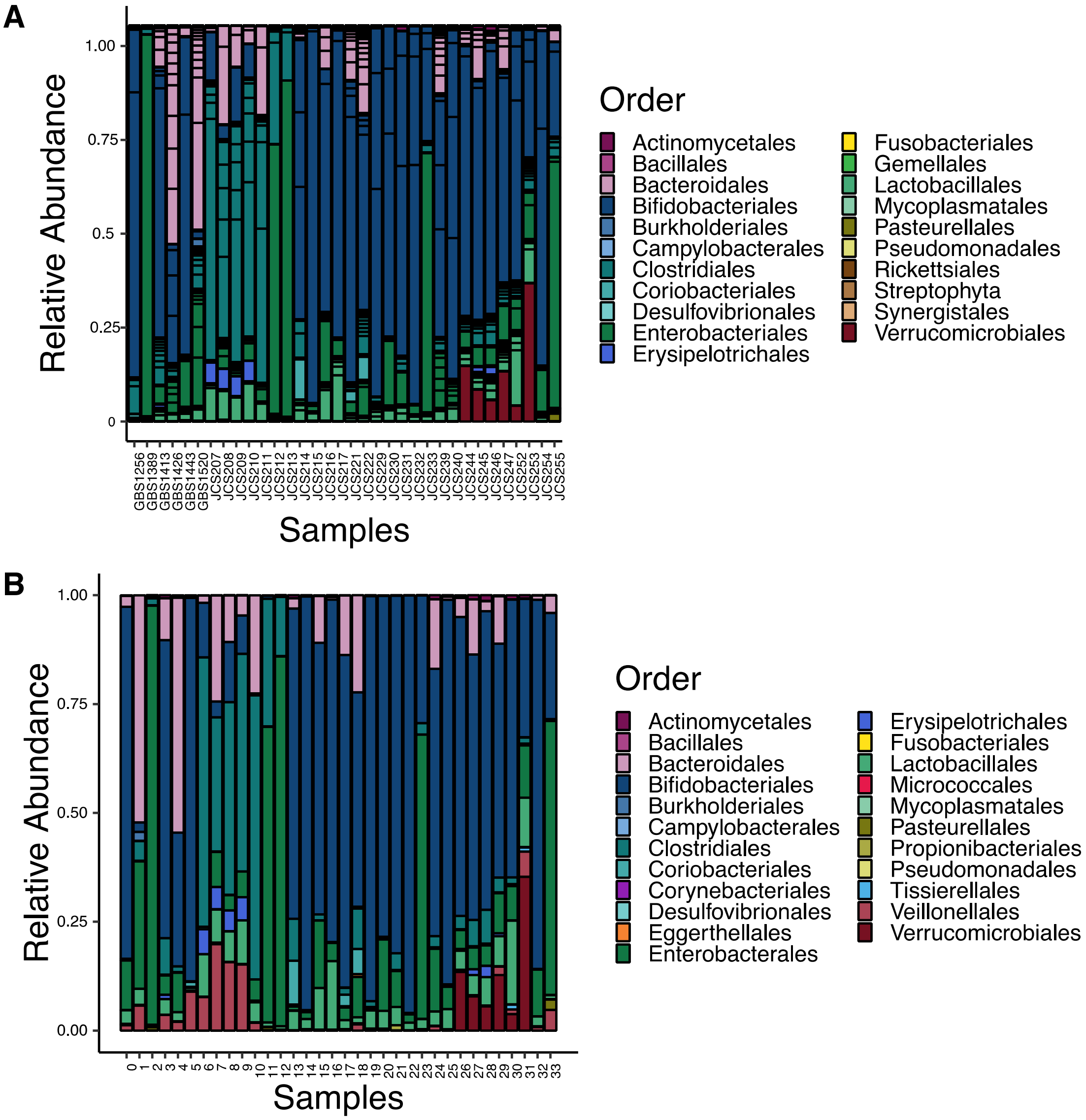


**Figure 5: Relative abundances of original 16S rRNA profiles and the simulated metagenome. A.** Order-level relative abundance of ASVs from original 16S rRNA profiles derived from infant gut stool samples, and **B.** Order-level relative abundance of organism in metagenomes simulated using the CAMI-Simulator.

from 21 to 82, with a mean observed species count of 50. The Shannon Diversity Index for the simulated metagenomes ranged from 0.24 to 2.88, with a mean of 1.74. Lastly, the Simpson Diversity Index for the metagenomes ranged from 0.07 to 0.92, with a mean of 0.68 (**Table 2**). Along with raw metagenomic reads for the 34 samples, the simulation output also included gold-standard single- and co-assemblies per sample.

**Table 2: Microbiome composition for each simulated metagenomic sample**

| SAMPLE | SHANNON DIVERSITY | SIMPSON DIVERSITY | NUMBER OF OBSERVED STRAINS | MOST ABUNDANT STRAIN IN SAMPLE | RELATIVE ABUNDANCE OF MOST ABUNDANT STRAIN |
|---|---|---|---|---|---|
| **sample_0** | 1.27 | 0.58 | 56 | *Bifidobacterium longum subsp. longum BBMN68* | 60.98 |
| **sample_1** | 2.79 | 0.88 | 82 | *Bacteroides fragilis* | 27.59 |
| **sample_10** | 1.83 | 0.76 | 50 | *Clostridium aceticum* | 38.93 |
| **sample_11** | 0.91 | 0.47 | 28 | *Escherichia coli O111:H- str. 11128* | 67.91 |
| **sample_12** | 0.55 | 0.26 | 31 | *Escherichia coli O111:H- str. 11128* | 84.94 |
| **sample_13** | 1.99 | 0.80 | 42 | *Bifidobacterium longum subsp. infantis ATCC 15697 = JCM 1222 = DSM 20088* | 33.21 |
| **sample_14** | 0.93 | 0.49 | 30 | *Bifidobacterium angulatum* | 65.87 |
| **sample_15** | 1.89 | 0.77 | 50 | *Bifidobacterium angulatum* | 40.94 |
| **sample_16** | 1.46 | 0.65 | 41 | *Bifidobacterium angulatum* | 53.38 |
| **sample_17** | 1.99 | 0.73 | 49 | *Bifidobacterium angulatum* | 47.12 |
| **sample_18** | 2.76 | 0.87 | 62 | *Bifidobacterium angulatum* | 30.85 |
| **sample_19** | 1.30 | 0.62 | 21 | *Bifidobacterium longum subsp. infantis ATCC 15697 = JCM 1222 = DSM 20088* | 52.76 |

| | | | | | |
|---|---|---|---|---|---|
| **sample_2** | 0.24 | 0.07 | 25 | *Escherichia coli O111:H- str. 11128* | 96.17 |
| **sample_20** | 1.42 | 0.67 | 30 | *Bifidobacterium longum subsp. infantis ATCC 15697 = JCM 1222 = DSM 20088* | 51.27 |
| **sample_21** | 1.52 | 0.68 | 23 | *Bifidobacterium longum subsp. infantis ATCC 15697 = JCM 1222 = DSM 20088* | 48.48 |
| **sample_22** | 1.10 | 0.56 | 22 | *Bifidobacterium longum subsp. infantis ATCC 15697 = JCM 1222 = DSM 20088* | 59.82 |
| **sample_23** | 1.09 | 0.52 | 28 | *Escherichia coli O111:H- str. 11128* | 65.25 |
| **sample_24** | 2.60 | 0.86 | 67 | *Bifidobacterium longum subsp. infantis 157F* | 26.96 |
| **sample_25** | 1.75 | 0.75 | 59 | *Bifidobacterium longum subsp. infantis ATCC 15697 = JCM 1222 = DSM 20088* | 34.74 |
| **sample_26** | 1.87 | 0.73 | 64 | *Bifidobacterium angulatum* | 45.99 |
| **sample_27** | 2.15 | 0.78 | 65 | *Bifidobacterium angulatum* | 41.34 |
| **sample_28** | 2.06 | 0.74 | 63 | *Bifidobacterium angulatum* | 46.23 |
| **sample_29** | 2.33 | 0.81 | 72 | *Bifidobacterium angulatum* | 36.34 |
| **sample_3** | 2.14 | 0.76 | 58 | *Bifidobacterium angulatum* | 44.29 |
| **sample_30** | 1.94 | 0.74 | 72 | *Bifidobacterium longum subsp. infantis ATCC 15697 = JCM 1222 = DSM 20088* | 46.47 |
| **sample_31** | 2.13 | 0.80 | 71 | *Akkermansia muciniphila* | 35.28 |
| **sample_32** | 1.15 | 0.57 | 45 | *Bifidobacterium longum subsp. longum BBMN68* | 60.07 |
| **sample_33** | 1.35 | 0.57 | 59 | *Escherichia coli O111:H- str. 11128* | 61.82 |

| | | | | | |
|---|---|---|---|---|---|
| **sample_4** | 2.89 | 0.92 | 58 | *Parabacteroides distasonis ATCC 8503* | 13.42 |
| **sample_5** | 1.43 | 0.67 | 69 | *Bifidobacterium angulatum* | 50.52 |
| **sample_6** | 1.50 | 0.60 | 58 | *[Eubacterium] eligens ATCC 27750* | 60.58 |
| **sample_7** | 2.47 | 0.88 | 61 | *[Eubacterium] eligens ATCC 27750* | 21.04 |
| **sample_8** | 2.22 | 0.82 | 55 | *[Eubacterium] eligens ATCC 27750* | 36.07 |
| **sample_9** | 2.16 | 0.81 | 44 | *[Eubacterium] eligens ATCC 27750* | 38.40 |

## 3.2 The effect of assembly tools and microbial community diversity on metagenomic assembly

In order to determine the impact of microbial diversity and closely related species on both single- and co-assembly quality, we first assembled simulated metagenomic data with MEGAHIT and MetaSPAdes, and subsequently compared the quality of these assemblies based on sample community composition. MetaQUAST results for single-sample assemblies by MEGAHIT (MHSA) and MetaSPAdes (MSSA) were compared against CAMISIM gold-standard single assemblies (GSSA), which served as a positive control to determine the impact of assembler on overall assembly quality (**Table 3**) Similarly, the MEGAHIT co-assembly (MGCA) was compared to the CAMISIM gold-standard co-assembly (GSCA). Co-assembling with MetaSPAdes failed due to memory requirements (>1TB RAM).

The mean total single-assembly length and largest contig size for both MHSA and MSSA were significantly lower than the GSSAs ($p < 0.05$) (**Table 3**). Misassembled regions were identified by aligning the assemblies to the original reference genomes with MetaQUAST. Overall, the misassemblies made up a mean of 4.8% (+/- 0.7) and 0.7% (+/- 0.2) of the total MHSA and MSSA assemblies respectively. As expected, zero misassemblies were reported for all GSSAs, which served as a positive control when comparing assembly quality (**Table 3**). From both the mean percent of misassemblies and the number of misassemblies per simulated sample, it is clear that MHSA assemblies had more misassemblies than MSSA (**Figure 6A**). For the co-assemblies, the length of the MHCA was roughly half that of the GSCA, and with misassembled regions making up 6.5% of the total MHCA. (**Figure 6B**).

In terms of the role of community composition on misassemblies, assemblers performed best with low-Shannon Diversity samples. Higher number of strains per sample was positively associated with more misassemblies for both MHSA ($\rho = 0.48$, $p < 0.001$) and MSSA ($\rho = 0.64$, $p < 0.001$) (**Figure 6C-D**). Furthermore, a higher relative abundance of the most dominant strain in each sample was correlated with fewer number of misassemblies for MEGAHIT ($\rho = -0.83$ $p < 0.001$) and MetaSPades ($\rho = -0.59$, $p < 0.001$), with the association being significantly stronger for MEGAHIT than for MetaSPAdes ($z=3.19$) (**Figure 6E-F**). The sample Shannon Diversity Index was also correlated with higher number of misassemblies per sample for both MEGAHIT ($\rho = 0.86$, $p < 0.001$) and MetaSPAdes ($\rho = 0.71$, $p < 0.001$), with the association being significantly stronger for MEGAHIT than for MetaSPAdes ($z=3.8$) (**Figure 6G-H**). This demonstrates that samples with lower community diversity with one clearly dominant strain—as opposed to a dominant genus with multiple strains—had fewer misassemblies, with this relationship being significantly impacted by the metagenomic assembler of choice.

**Table 3: Raw assembly quality metrics from MetaQUAST to compare MEGAHIT, MetaSPAdes, and gold-standard single- and co-assemblies**

| | CO-ASSEMBLY | | SINGLE ASSEMBLY | | |
|---|---|---|---|---|---|
| **Metric** | **MHCA** | **GSCA** | **MHSA** | **MSSA** | **GSSA** |
| **Total length of assembly bp (95% CI)** | 291,698,138 | 624,020,247 | 28,962,308 (5,396,646) | 27,542,948 (5,013,836) | 45,38,0295 (9,504,826) |
| **Total length of assembly as a proportion of expected (95% CI)** | 0.50 | 1 | 0.67 (0.03) | 0.64 (0.03) | 1 |
| **Longest contig bp (95% CI)** | 766,344 | 6,475,289 | 607,155.65 (87883.19) | 712,688.59 (107053.30) | 4,721,608.77 (396648.61) |
| **N50 bp (95% CI)** | 8,140 | 69,764 | 26,430.71 (10031.07) | 34,612.27 (15164.79) | 691,983.61 (478,068.79) |
| **NA50 bp (95% CI)** | 7,873 | 69,764 | 25,389.91 (9,683.46) | 34,590.24 (15,169.72) | 691,983.62 (478,068.79) |
| **% Missassemblies (95% CI)** | 6.50 | 0 | 4.82 (0.69) | 0.71 (0.23) | 0 |

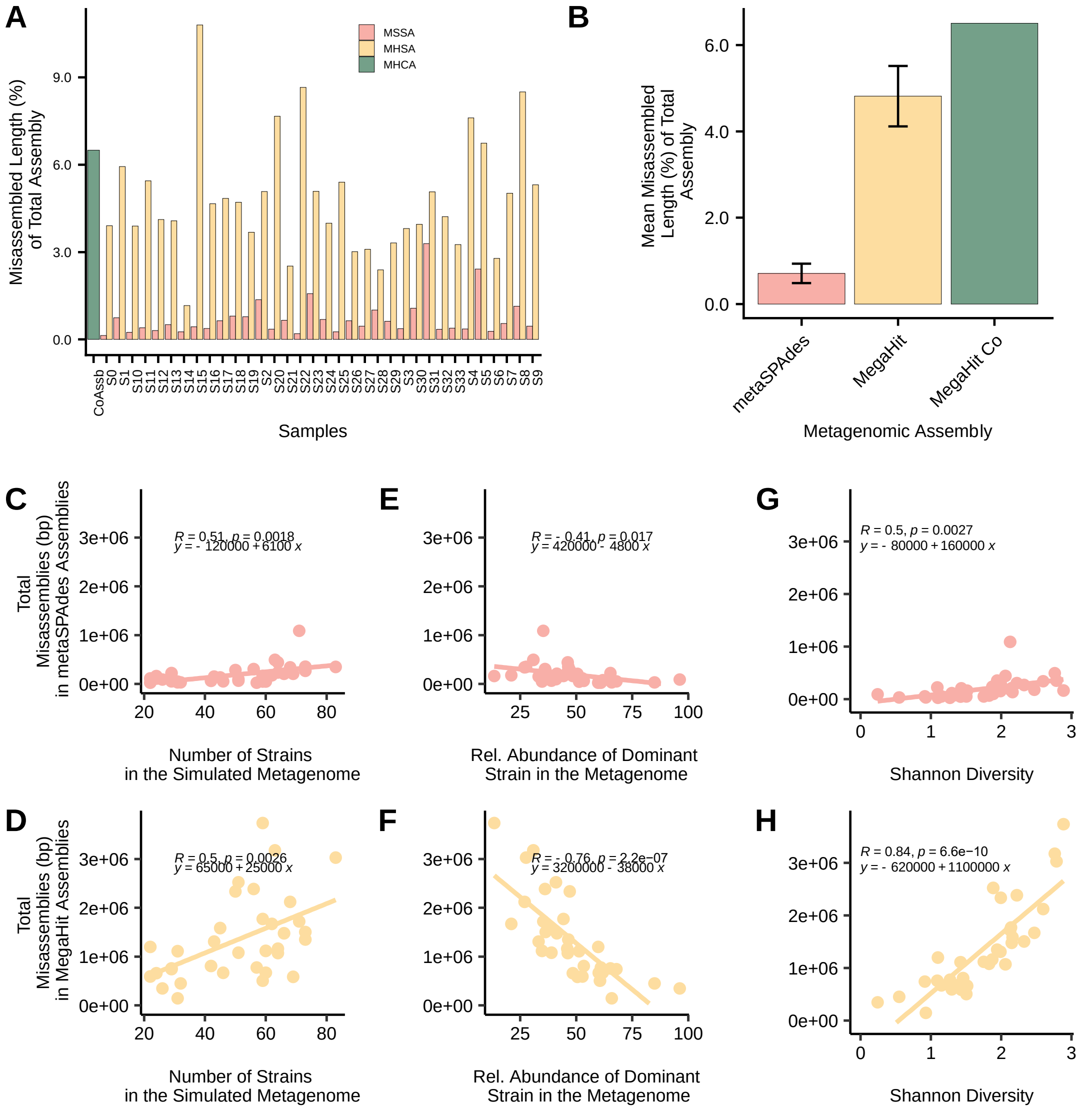


**Figure 6: Metagenomic assembly quality metrics for 104 metagenomic assemblies from 34 CAMISIM-simulated metagenomic samples. A.** Misassembled lengths of total MSSA and MHSA assemblies**. B.** Mean misassembled assembly length as a percent of the total MSSA

## 3.3 Accuracy of assembly-level CAZyme predictions compared to the gold standard

dbcAN was used to predict CAZymes for MEGAHIT, MetaSPAdes, and GSSA single-sample assemblies, showing that predictions were more reliable for MSSAs than for MHSAs. The GSSA predictions were used as a reference to determine the accuracy of MEGAHIT and MetaSPAdes predictions. MHSAs had a significantly higher false positive rate at 18%, compared to 12.4% in MSSAs assemblies (p < 0.05) (**Figure 7A**). Overall, CAZyme predictions in MHSAs were

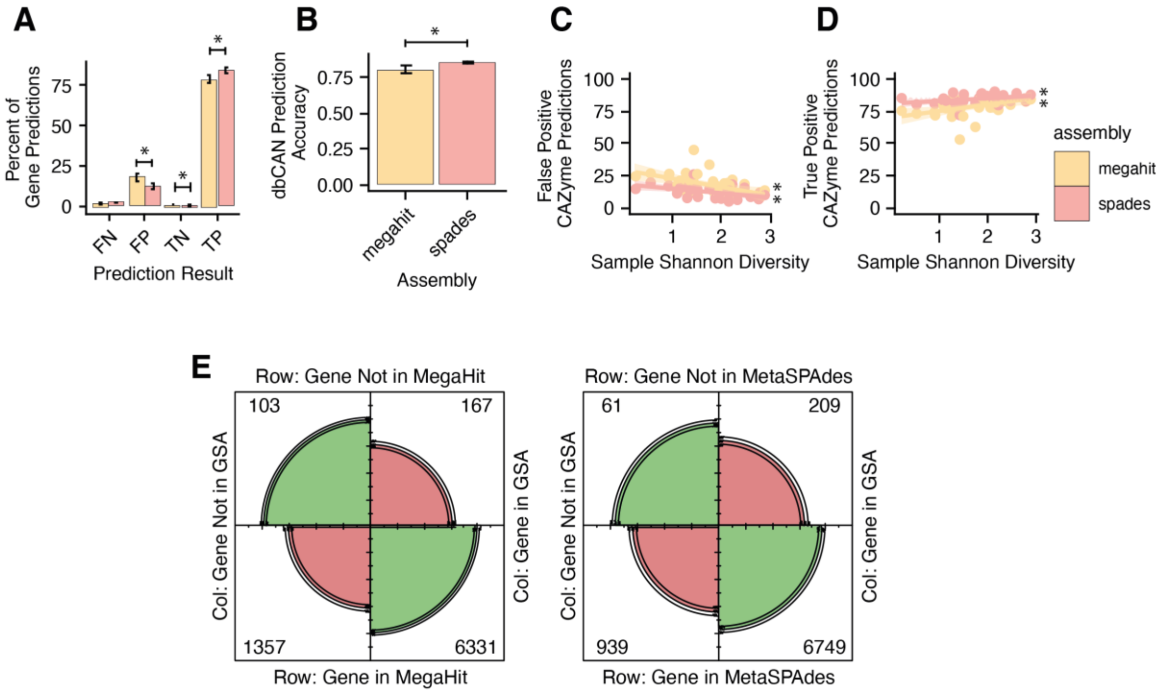


**Figure 7: MEGAHIT and MetaSPAdes single assembly-level CAZyme predictions compared to the gold-standard control. A.** Distribution of true positive (TP), false negative (FN), false positive (FP), and true negative (TN) CAZyme predictions for MHSA and MSSA. **B.** Bar plot of mean prediction accuracy for CAZyme predictions in MHSA and MSSA across samples. **C-D.** Linear regression model of sample Shannon Diversity and CAZyme prediction accuracy for MHSA and MSSA. **E.** Number of true positive, false negative, false positive, and true negative CAZyme predictions in a confusion matrix.

80.1% (+/- 0.03) accurate compared to the gold-standard, while MSSAs were 85.2% accurate (+/- 0.01), with a significant difference between the two ($p < 0.05$) (Fig 4B). For both MEGAHIT and MetaSPAdes, increased Shannon Diversity was positively associated with the percentage of true positives ($p < 0.05$), and negatively associated with the percentage of false positives ($p < 0.05$) (**Figure 7 C-D**). A distribution of the total number of FP, FN, TP, and TN are outlined in **Figure 7E**.

### 3.4 Differences in bin and MAG quality by assembly method

Our next research question was to determine the impact of the assemblers, assembly quality, and single and co-assemblies on the resulting MAGs. All single- and co-assemblies were binned with MetaBat2, including the GSSAs and GCSAs, which served as positive controls to determine the impact of the binning software alone, without the added effect of the assemblers. The identity of each bin was determined by aligning bins to a database of the reference genomes used to simulate the metagenomic samples. Bins were considered MAGs if: (1) Their overall quality (CheckM % completeness - % contamination) was over 50%, and (2) If a BLASTN alignment of the bin against our local database of reference genomes used in the simulation returned a hit for an organism that was in the same sample that the bin came from. The reference coverage of each MAG, or the percent alignment of the MAG to its reference, was also used as a quality metric.

A total of 2,049 bins were filtered down to 1,026 MAGs. The total number of bins obtained for all samples was the highest for MHCA at 429, followed by 378 for GSCA, 319 for GSSA, 242 for MHSA, and 232 for MSSA (**Table 4**). Overall, GSSAs yielded an average of 6.74 MAGs per sample (+/- 0.55) and 210 in total, and GSCAs yielded 6.17 MAGs (+/- 0.42) per sample and 229 in total. MEGAHIT co- and single-sample assemblies yielded 7.8 (+/- 0.7) and 5 (+/- 0.42) per sample, and 265 and 170 in total (**Table 4, Figure 8**).  Lastly, MetaSPades single-sample assemblies yielded 4.47 (+/- 0.35) MAGs per sample, and 152 in total. The number of bins yielded by the MEGAHIT coassembly and the gold-standard coassembly were both significantly higher than the number of bins yielded by the MEGAHIT and MetaSPAdes single-sample assemblies ($p < 0.05$). There was no significant difference in the number of organisms that were not binned by any of the assembly methods. Interestingly, the MEGAHIT co-assembly also yielded more bins than the gold-standard (**Table 4, Figure 8**).

**Table 4: Bin and MAG quality metrics from CheckM, alignment to reference genomes, and abundance of organisms reconstructed as MAGs**

| | CO-ASSEMBLY | | SINGLE ASSEMBLY | | |
|---|---|---|---|---|---|
| **METRIC** | **MHCA** | **GSCA** | **MHSA** | **MSSA** | **GSSA** |
| **Total Bins** | 429 | 378 | 242 | 232 | 319 |
| **Total MAGs** | 265 | 229 | 170 | 152 | 210 |
| **Mean number of MAGs per sample (95% CI)** | 7.8 (0.7) | 6.74 (0.55) | 5 (0.42) | 4.47 (0.35) | 6.17 (0.42) |
| **Mean CheckM quality of MAGs % (95% CI)** | 86.5 (1.7) | 95.1 (1.3) | 90.5 (1.9) | 92.8 (1.7) | 92.1 (1.7) |
| **Mean CheckM contamination of MAGs % (95% CI)** | 3.3 (0.5) | 1.8 (0.7) | 1.4 (0.4) | 1.3 (0.6) | 2.1 (0.8) |
| **Average % CheckM strain heterogeneity of all MAGs (95% CI)** | 30.37 (3.78) | 11.00 (3.81) | 26.15 (4.93) | 16.36 (4.61) | 8.08 (3.06) |
| **Mean reference coverage of all bins % (95% CI)** | 30.9 (1.6) | 76.5 (3.6) | 38.1 (2.2) | 44.8 (3.3) | 68.5 (3.9) |
| **Mean reference coverage of final MAGs % (95% CI)** | 38.9 (1.4) | 93.4 (1.8) | 44.6 (2.2) | 57.8 (3.3) | 88 (2.1) |
| **Sensitivity: Average lowest, average highest relative abundance of organisms reconstructed in MAGs (95% CI)** | 0.21 (0.1), 19.89 (5.22) | 0.69 (0.3), 42.9 (7.52) | 4.00 (5.75), 23.24 (6.01) | 7.62 (6.1), 40.21 (7.44) | 4.6 (5.69)-44.41 (6.77) |
| **Avg relative abundance of organisms reconstructed in MAGs (95% CI)** | 5.35 (1.1) | 10.66 (2.2) | 5.35 (1.11) | 14.8 (2.87) | 12.32 (2.36) |

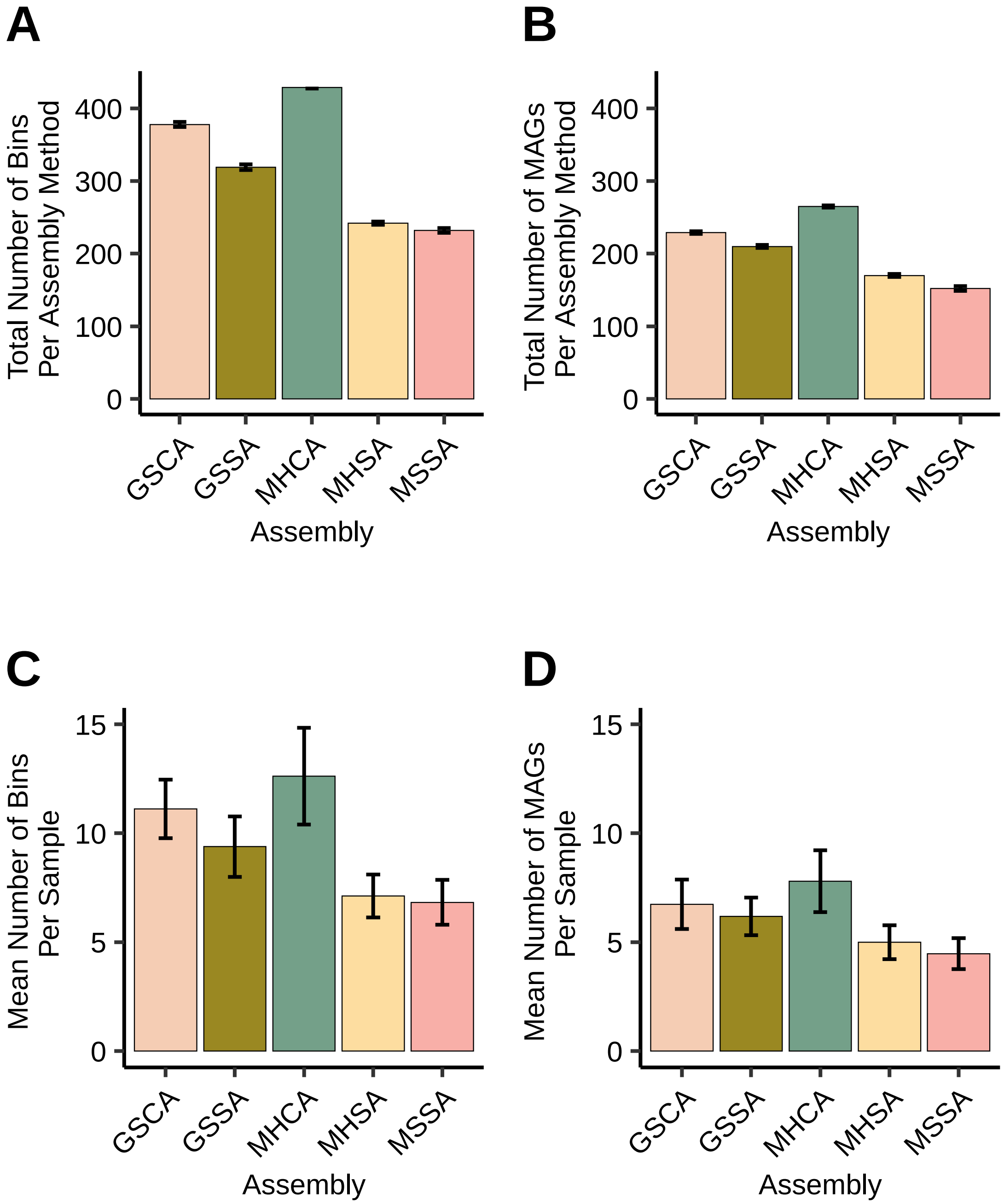


**Figure 8: The number of bins and MAGs obtained in total and per sample from each assembly. A.** Bar plot of the total number of bins obtained per assembly, **B.** Bar plot of the total number of MAGs retrieved per assembly, **C.** Bar plot of the mean number of bins retrieved per sample, with 95 % CI, and **D.** Bar plot of the number of MAGs retrieved per sample, with 95% CI.

### 3.4.1 MAG reference coverage and CheckM quality metrics

For each reference genome, there could be multiple sub-strains simulated by CAMISIM, at different relative abundances. Since each sample had multiple closely related species and strains, there could be more than one organism in each MAG. Given the importance of obtaining high quality, strain-level MAGs to accurately map the metabolic differences between organisms, we chose to evaluate the following quality metrics: quality (CheckM % completeness - % contamination), CheckM % contamination by itself, Checkm % strain heterogeneity, and mean coverage of the reference genome in each MAG. As expected, MAGs reconstructed from the GSCA and GSSAs had the highest quality MAGs, at 95.1% and 92.1%. This was followed by the MSSAs, MHSAs, and MHCAs, at 92.8%, 90.5%, and 86.5%. There was a significant difference between the quality of MAGs from the gold standard assemblies and all MEGAHIT assemblies. However, there was no significant difference between the quality of the gold standard single assembly and the MetaSPAdes single assembly, and between the MetaSPAdes and MEGÅHIT single assemblies (**Figure 9**).

MAGs reconstructed from the MHCA had the highest mean contamination at 3.3%, while MAGs constructed from the MSSAs had the lowest mean contamination at 1.4%. The GSSA, GSCA, and MHSA had mean contamination values of 2.1%, 1.8%, and 1.4% respectively. The MHCA and MHSA had the highest strain heterogeneity, at 30.37% and 26.15%, compared to 11.00% for GSCA, 8.08% for GSSA, and 16.36% for MSSA (**Table 4**).

It is worth noting that while MAGs from all assemblies had quality scores above 85%, these CheckM-based quality metrics were not consistent with MAG reference coverage, or the amount that each MAG covered its reference genome. MAGs reconstructed from the GSCA and GSSAs had the highest mean reference coverage, at 93.4% and 88%. In comparison, the mean reference coverage values for MAGs from MSSAs, MHSAs, and the MHCA were 57.8%, 44.6%, and 38.9%. Thus, while a MAG may be considered complete by CheckM, it may not be fully complete when considering its alignment rate to the original reference genome it represents.

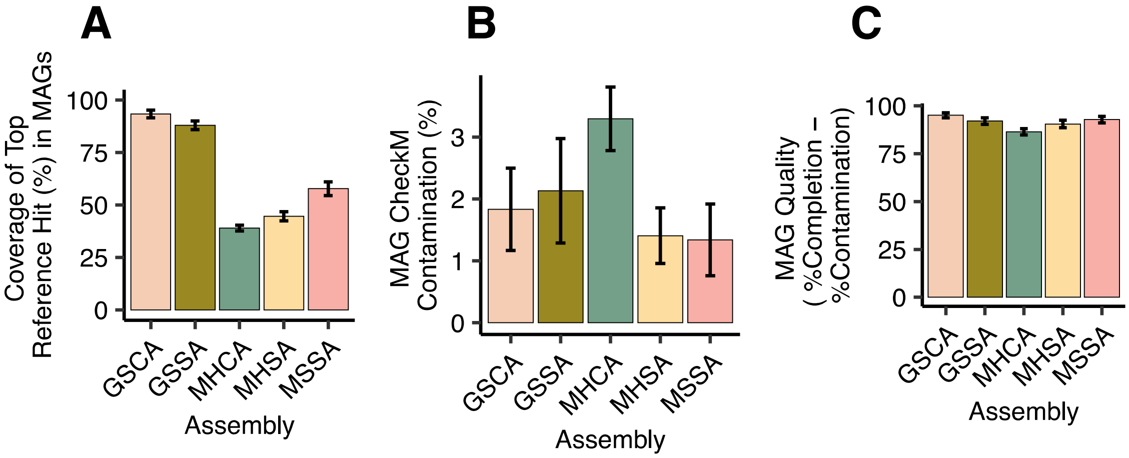


**Figure 9: Mean reference coverage, CheckM contamination, and quality for MAGs reconstructed from GSCA, GSSAs, MHCA, MHSAs, and MSSAs. A.** Mean % of reference genome covered by the MAG, **B.** Mean % CheckM contamination in MAG, and **C.** Mean MAG quality (% completeness - % contamination), for MAGs retrieved from the GSCA, GSSA, MHCA, MHSA, and MSSA.

### 3.4.2 The impact of relative abundance on MAG completeness and reference coverage

To understand how the relative abundance of a strain may play a role in MAG quality, we plotted the distribution of MAG completeness and reference coverage versus the relative abundance of the original strain that the MAG belonged to. For the MSSAs, MHSAs, and MHCA, most strains with a relative abundance more than 25% had MAGs with consistently high levels of completeness, ranging from 90% to 100%, with little difference in MAG completeness distribution between the three assemblies (**Table 4**). However, for the same strains, higher relative abundances did not guarantee high reference coverage for the MHSA and MSSA assemblies (**Figure 10A-D**), and lower reference coverage values (less than or equal to 50% reference coverage) could be seen even for strains with relative abundances over 25%. From these results, it is once again evident that while CheckM completeness and contamination metrics are a very useful tool to gauge MAG completeness, a MAG that may seem “complete” for having all the core marker genes may still have gaps and missing genes.

For MAGs that have high CheckM completeness but low reference coverage, it could be possible that while the core genome is assembled and binned, the species- and strain-specific pan-genome is not assembled and binned. If we compare the distribution of MAG completeness and reference coverage for the GSSAs (**Supplementary Figures 2A-C**, **3A-C**), they look very similar. Most MAGs that are 100% complete also have 100% reference coverage. Furthermore, the density graphs of the proportion of MAGs as per their % completeness and reference coverage are also similar (**Supplementary Figures 2D, 3F**). Since achieving this overlap in completeness and reference coverage is possible given perfect single- and assemblies, it is likely that gaps in the MAGs are a product of the assembly step rather than the binning step.

It is also evident that each assembler was binning organisms of different relative abundances, as demonstrated by **Figure 11E**, showcasing that the MHCA had the highest proportion of MAGs from lower-abundance strains, followed by GSSAs and MSSAs. Furthermore, **Figure 11F** also demonstrates that the MHSA yielded more MAGs with lower reference coverage, with most MHSA MAGs falling below 50% coverage of their reference genomes. In comparison, the majority of MAGs reconstructed from the GSSAs had reference coverage higher than 80%.

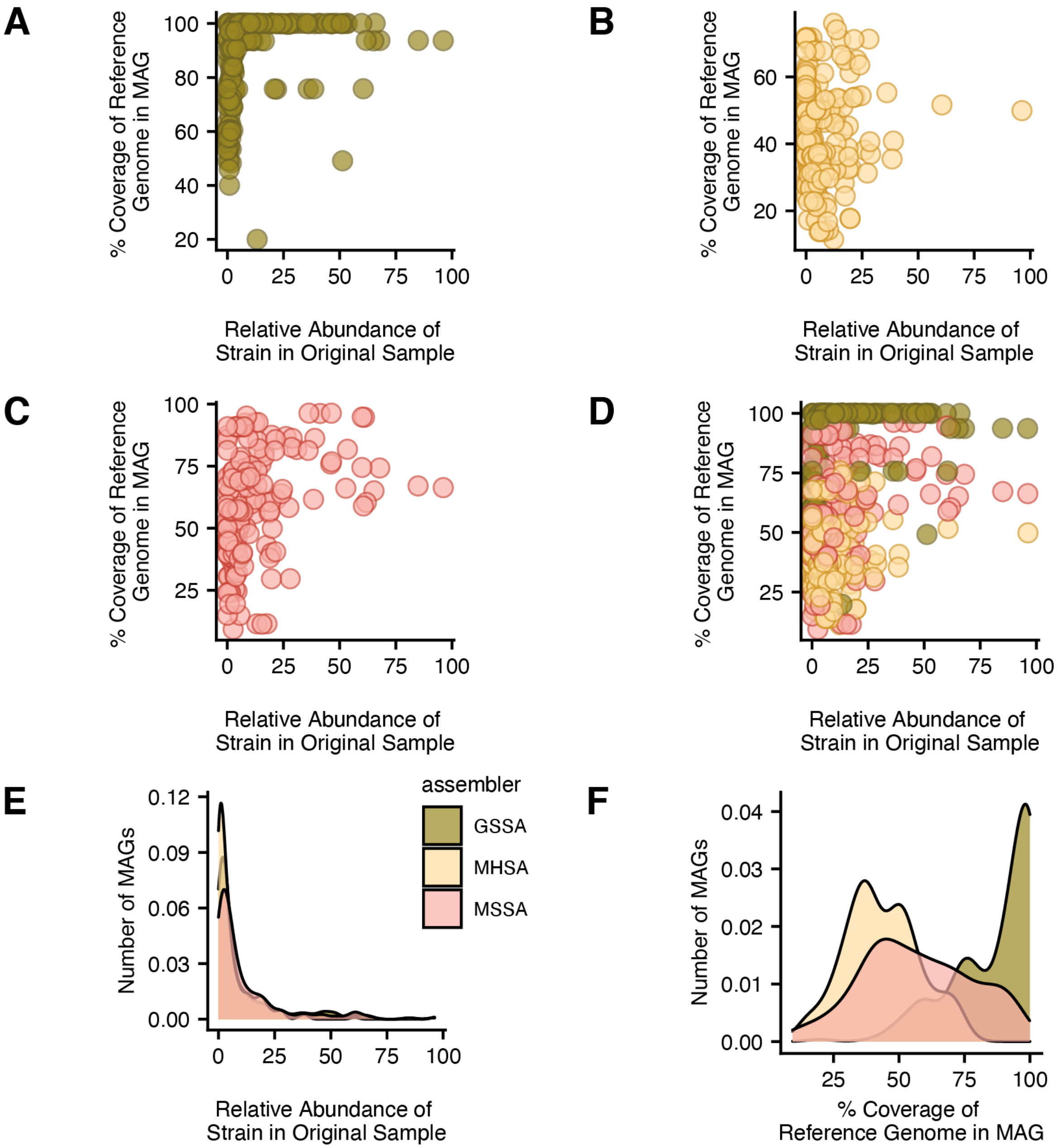


**Figure 10: Distribution of MAG reference coverage as a function of the relative abundance of original strains.** Distribution of the reference coverages of MAGs from **A.** just GSSA**, B.** MHSA**, C.** MSSA**, or D.** all three assemblers plotted against the relative abundance of the original strains reconstructed within the MAGs. **E.** Density plot of the number of MAGs yielded by GSSA, MHSA, and MSSA by relative abundance. **F.** Density plot of the number of MAGs yielded by the three assemblies by their reference coverage.

### Assembly strategy impacts binning of low- and high-abundance organisms into MAGs

MAGs constructed from the MEGAHIT assemblies, MHCA and MHSA, captured more low-abundance species (**Table 4, Figure 11**). As expected, the GSSAs and GSCAs captured the largest range of strains of different relative abundances; the average relative abundance of strains reconstructed as a MAG was between 0.69% and 42.9% for the GSCA, and between 4.6% and 44.41% for the GSSA. The MSSA was less sensitive, capturing organisms with an average relative abundance of between 7.62% and 40.21%. The average relative abundance of organisms captured by MHCA was between 0.21% and 19.89%, and between 4% and 23.34% for the MHSA (**Figure 11A, 11C**). MEGAHIT co-assembly was thus able to capture very low abundance organisms but not high abundance organisms, whereas MetaSpades single assembly was best for high abundance organisms but missed many low abundance strains (**Figure 11B**). MEGAHIT assemblies yielded the highest number of bins and MAGs, however, bins from the MEGAHIT assemblies also had the lowest reference coverage, highest contamination, and highest strain heterogeneity. This contamination was also reflected in the false positive CAZyme predictions (section **3.3**).

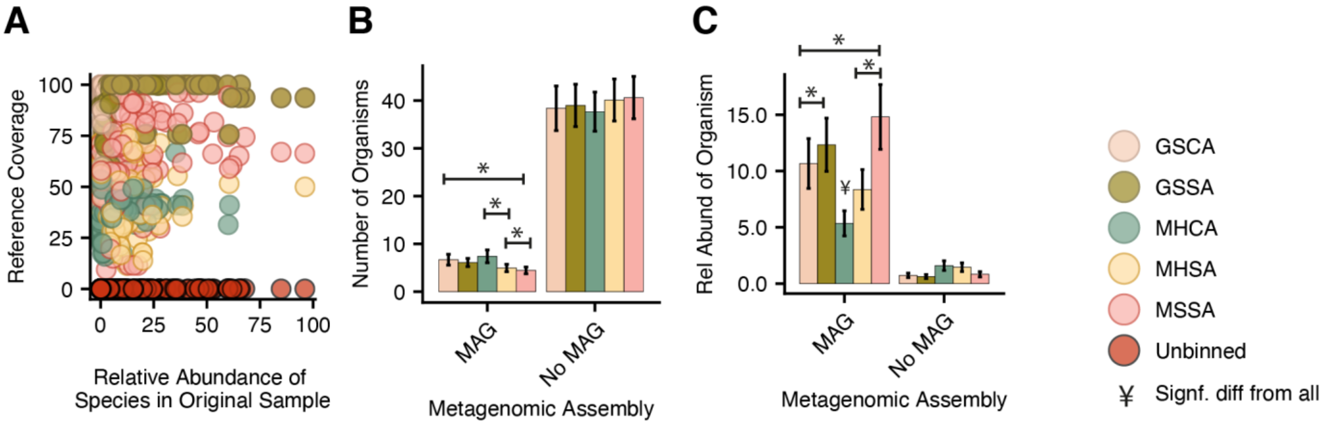


**Figure 11: The mean reference coverage, number and relative abundance of organisms reconstructed or not reconstructed as MAGs for each assembly type. A.** Scatter plot highlighting the distribution of MAGs by the relative abundanc of their represented strain versus the % reference coverage. **B.** Bar plot comparing the number of organisms recovered or not recovered as MAGs by the different assembly types. **C.** The mean relative abundance of organisms recovered or not recovered as MAGs by the different assembly methods.

### 3.4.3 Difference in MAGs retrieved from single-sample assemblies and coassembly

To strictly determine the impact of single- and co-assemblies on reconstructing high-quality bins, we can compare the performance of the GSCA and GSSAs, without looking at any of the assemblers. Given a perfect assembly, both the co-assembly and single-sample assemblies yield qualitatively similar MAGs, with the GSCA conferring a mild advantage (**Table 4**). Overall, the GSCA yielded more bins and MAGs (378 and 229) compared to the GSSAs (319 and 210), with higher reference coverage (93.4% compared to 88%), higher quality (95.1% compared to 92.1%), lower contamination (1.8% compared to 2.1%), and with more sensitivity to low abundance strains (average lowest abundance MAG retrieved was 0.69% abundance compared to 4.6%). The only metric where the GSSAs outperformed the coassembly was strain heterogeneity, where MAGs from the single-sample assemblies had 8.08% strain heterogeneity, compared to 11% for the coassembly. In theory, the coassembly was superior in MAG reconstruction. However, in practice, when comparing the real MEGAHIT coassembly with MEGAHIT single assemblies, the MHSA outperformed the MHCA.

## 3.5 The effect of microbial community diversity on retrieving strain-level MAGs from abundant genera

Our next research question was to determine whether samples containing abundant genera with high strain diversity would affect the number and quality of MAGs for that genus, and whether this effect would be different based on assemblers and assembly method. For this, we chose to look at *Bifidobacterium*, since 21 out of 34 simulated metagenomes were dominated by a high relative abundance of bifidobacterial species, with most samples having multiple species, and often multiple strains per species. To answer this question, we calculated overall abundance of the *Bifidobacterium* genus per sample, the number of *Bifidobacterium* strains per sample, and bifidobacterial inter-genus Shannon Diversity per sample (**Supplementary Table 1).**

Higher diversity within the *Bifidobacterium* genus was associated with lower MAG reference coverage, demonstrating that MAGs belonging to diverse genera were of lower quality, covering a lesser percent of their reference genome (p < 0.05) (**Figure 12**). On the other hand, higher relative abundance of the *Bifidobacterium* genus was associated with higher reference coverage (p < 0.05).

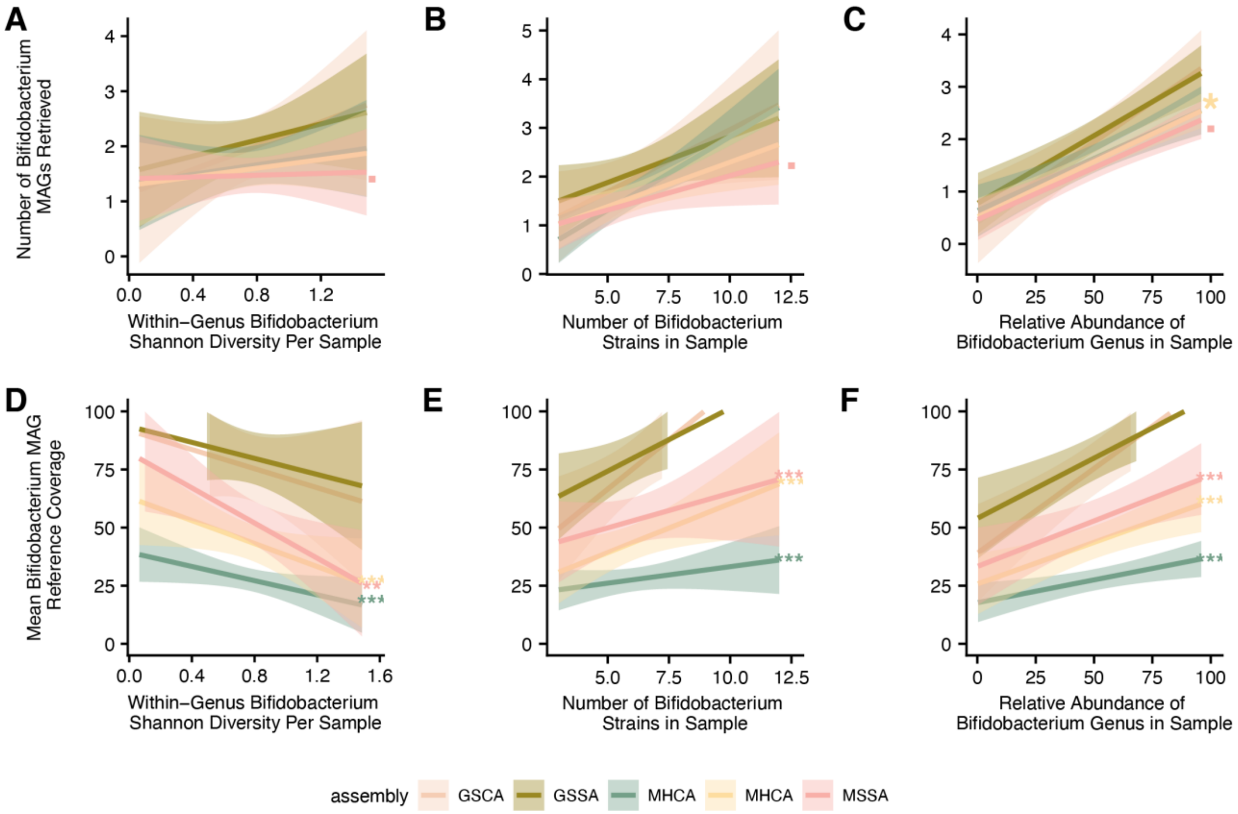


**Figure 12: The effect of within-genus Bifidobacterium Shannon Diversity, number of strains, and relative abundance on the number and reference coverage of MAGs retrieved. A**. The correlation between higher within-genus Shannon Diversity and the number of MAGs was trending for MHSA. There were no significant correlations. **B.** The correlation between higher *Bifidobacterium* strains and the number of MAGs was trending for MHSA **C.** Higher relative abundance of *Bifidobacterium* was positively correlated with the number of MAGs obtained from MSSA (p < 0.05), and trending for MHSA. **D.** Higher within-genus Shannon Diversity was negatively correlated with the reference coverage of MAGs obtained from MHSA, MHCA, and MSSA (p < 0.05). **E.** Higher *Bifidobacterium* observed species was positively correlated with the reference coverage of MAGs obtained from MHSA, MHCA, and MSSA (p < 0.05**). F.** Higher relative abundance of the *Bifidobacterium* genus was positively correlated with the reference coverage of MAGs obtained from MHSA, MHCA, and MSSA. (p < 0.05).

For MHCA and MHSA, the effect of relative abundance on reference coverage was significantly different compared to the gold-standard assemblies ($p < 0.05$). For MHSA, higher relative abundance of *Bifidobacterium* within the sample was also correlated with the number of *Bifidobacterium* MAGs retrieved.

Lastly, higher number of *Bifidobacterium* strains in the sample was associated with higher reference coverage. For MHCA, this association was significantly different compared to the gold-standard ($p < 0.05$).

We thus found that the number and reference coverage of Bifidobacterium MAGs were significantly affected by Bifidobacterium relative abundance within the sample, along with within-genus Shannon Diversity and number of observed bifidobacterial strains and species.

The same analysis was carried out for exclusively high-quality (HQ) MAGs that had a quality score of more than 90%, yielding similar results (**Supplementary Figure 4**). It is worth noting that out of all assemblies, the MHCA yielded the least number of HQ *Bifidobacterium* MAGs, with only 5 HQ MAGs from 34 samples.

## 3.6 The effect of assembler choice on CAZyme prediction quality of strain-level MAGs from abundant genera

In order to determine the impact assembly quality on MAG-level gene predictions, we chose to compare the accuracy of CAZyme predictions across all *Bifidobacterium* strains in all 34 samples. CAZyme prediction profiles for each of the 298 Bifidobacterium MAGs from the five assemblies were compared against

CAZyme profiles of their respective reference genomes, belonging to one of 33 *Bifidobacterium* strains. Out of the 33 *Bifidobacterium* strains in the original sample, we were able to obtain MAGs for 19.

The mean CAZyme prediction accuracy was 0.98 for Bifidobacterium MAGs reconstructed from the GSCA, 0.95 for MAGs from GSSA, 0.93 for MAGs from MSSA, 0.87 for MHSA, and 0.76 for MHCA (**Table 5**, **Figure 13**). The prediction accuracy for MAGs from MHCA was significantly lower than all other assemblers ($p < 0.05$). The mean sensitivity, which refers to the rate of true positives captured by each method, was 0.98 for GSCA, 0.87 for GSCA, 0.86 for MSSA, 0.72 for MHSA, and 0.62 for MHCA. The performance of all assemblers was consistent across different CAZyme prediction quality metrics, including positive predictive value, negative predictive value, precision, specificity, and F1 (**Table 5**, **Figure 13**). The MHCA consistently had the worst performance compared to the gold-standard, while the MSSA was closest to the gold-standard.

**Table 5: Bifidobacterium MAG-level CAZyme prediction quality metrics compared to CAZyme predictions of gold-standard reference genomes**

| ***BIFIDOBACTERIUM* MAGs FROM:** | **CO-ASSEMBLY** | | **SINGLE ASSEMBLY** | | |
|---|---|---|---|---|---|
| **METRIC (95% CI)** | **MHCA** | **GSCA** | **MHSA** | **MSSA** | **GSSA** |
| **% True positive** | 45.8 (5) | 74.7 (4.1) | 65.4 (5.6) | 69 (5.6) | 72.7 (3.95) |
| **% True negative** | 30.8 (3.6) | 24.4 (3.9) | 21.7 (4.9) | 24.2 (4.8) | 22.7 (3.95) |
| **% False positive** | 16.9 (2.3) | 0.5 (0.5) | 8.1 (2.7) | 4.4 (2.3) | 2.8 (1.4) |
| **% False negative** | 6.5 (1.2) | 0.4 (0.4) | 4.8 (1.6) | 2.4 (1.3) | 1.8 (1.2) |
| **Accuracy (0-1)** | 0.76 (0.03) | 0.98 (0.01) | 0.87 (0.03) | 0.93 (0.03) | 0.95 (0.02) |
| **Pos. Predictive Value (0-1)** | 0.84 (0.04) | 0.98 (0.01) | 0.80 (0.08) | 0.92 (0.05) | 0.92 (0.05) |
| **Neg. Predictive Value (0-1)** | 0.71 (0.04) | 0.99 (0.01) | 0.89 (0.03) | 0.94 (0.03) | 0.96 (0.02) |
| **Sensitivity (0-1)** | 0.62 (0.05) | 0.99 (0.01) | 0.72 (0.08) | 0.86 (0.06) | 0.87 (0.06) |
| **Specificity (0-1)** | 0.86 (0.02) | 0.99 (0.01) | 0.92 (0.02) | 0.96 (0.02) | 0.98 (0.02) |
| **Precision (0-1)** | 0.84 (0.04) | 0.98 (0.01) | 0.80 (0.08) | 0.92 (0.05) | 0.92 (0.05) |
| **F1 (0-1)** | 0.70 (0.04) | 0.98 (0.01) | 0.74 (0.08) | 0.88 (0.06) | 0.88 (0.06) |

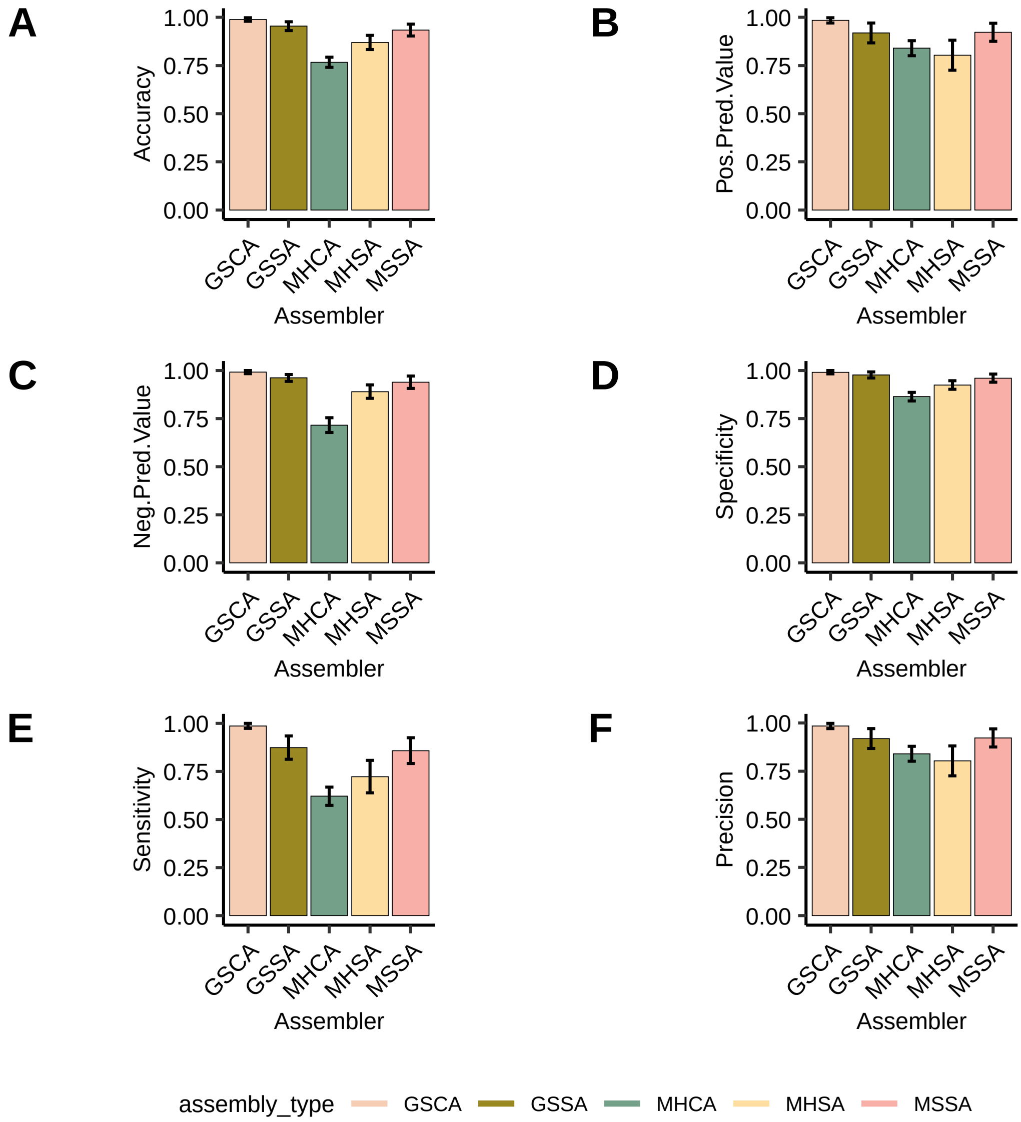


**Figure 13: Distribution of CAZyme prediction quality metrics across *Bifidobacterium* MAGs from GSCA, GSSA, MHCA, MHSA, and MSSA. A.** CAZyme prediction accuracy, **B.** Positive Predictive Value, **C.** Negative Predictive value, **D.** Specificity, **E.** Sensitivity, and **F.** Precision. Values are reported as means, with 95% confidence intervals.

Furthermore, conducting CAZyme predictions at the MAG-level proved to have higher accuracy and lower FP predictions compared to conducting CAZyme predictions at the assembly-level (**Figure 14**). At the MAG-level, the percent of predictions from MEGAHIT and MetaSPAdes assemblies that were FP ranged from 4.4% to 16.9%, while at the assembly-level, the FP ranged from 12.4% to 18%. In both instances, the MEGAHIT assemblies and MAGs had higher FP values.

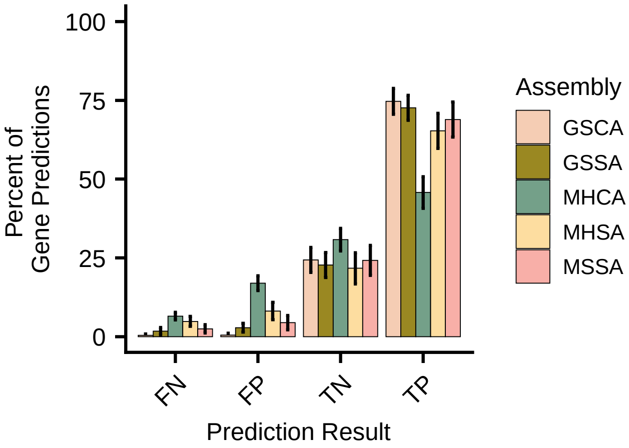


**Figure 14: Distribution of true positive, true negative, false positive, and false negative CAZyme predictions across *Bifidobacterium* MAGs from GSCA, GSSA, MHCA, MHSA, and MSSA.** The mean % (95% CI) of FP, FN, TP, TN in MAGs are plotted for each assembly type.

## 3.7 The effect of microbial community diversity on CAZyme prediction quality of strain-level MAGs from abundant genera

Given the differences between the GSAs, MEGAHIT, and MetaSPAdes in terms of CAZyme prediction quality, our next research question was to determine whether the quality of predictions was also impacted by microbiome community

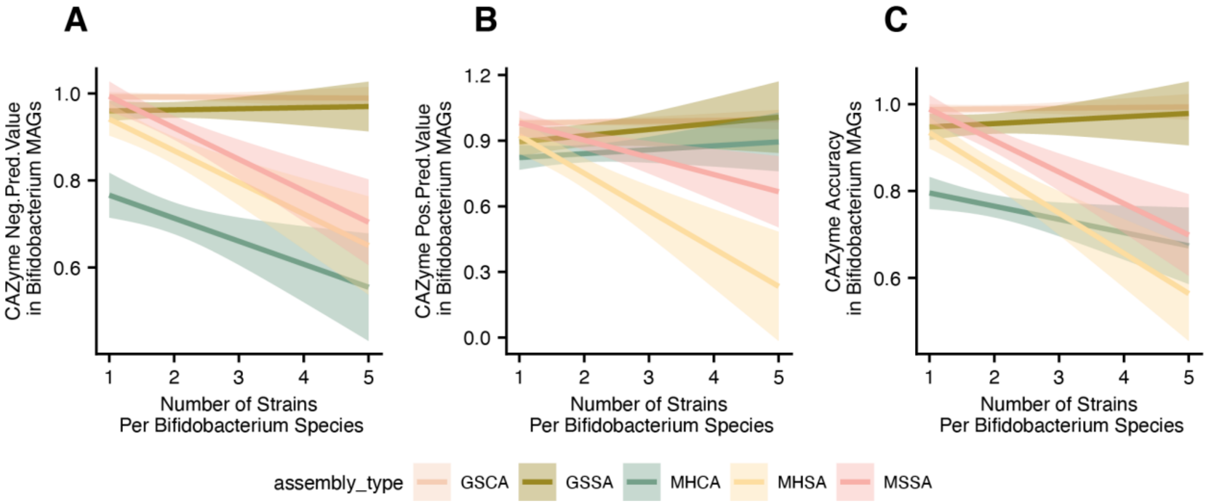


**Figure 15: The impact of *Bifidobacterium* within-species diversity on Bifidobacterium MAG-level CAZyme prediction quality metrics. A. Mean** CAZyme negative predictive value**, B.** mean CAZyme positive predictive value**,** and **C.** mean CAZyme prediction accuracy are plotted against the number of strains per *Bifidobacterium* species in each sample.

composition. For all samples, higher number of *Bifidobacterium* strains within each species was correlated with decreased CAZyme prediction accuracy (**Figure 15**). The relative abundance of the *Bifidobacterium* strains represented by the MAGs was also positively associated with CAZyme prediction accuracy for their respective MAGs ($p < 0.05$). Within-genus *Bifidobacterium* Shannon Diversity was negatively associated with CAZyme prediction ($p < 0.05$). Interestingly, while *Bifidobacterium* within-genus Shannon Diversity was negatively associated with CAZyme prediction quality, sample Shannon Diversity was generally positively correlated with the same quality metrics.

Decreased CAZyme prediction accuracy in *Bifidobacterium* MAGs with increased sample complexity is also evident in **Figure 16**, particularly for MAGs from the MEGAHIT coassembly. After ordering samples based on bifidobacterial abundance, it was demonstrated that samples with higher abundance of *Bifidobacterium* had a higher number of MAGs with better CAZyme prediction accuracy. However, prediction accuracy was also observed to be impacted by 1) the assembler, 2) the number of bifidobacterial species and strains in the sample, and 3) phylogenetic distance.

The number and type of *Bifidobacterium* strains obtained by MSSA, MHSA, and MHCA were unique (**Figure 16**). Furthermore, despite the presence of multiple strains per species, one *Bifidobacterium* sp. usually only yielded one MAG, as opposed to a MAG for each strain. For example, if a sample had three strains of *B. longum*, only one *B. longum* MAG was retrieved (**Figure 16**). Given the fact that MAGs from the MSSA, MHSA, and MHCA had 16.36%, 26.15%, and 30.37% strain heterogeneity respectively, it is possible that these MAGs had multiple strains within them for each bifidobacterial species.

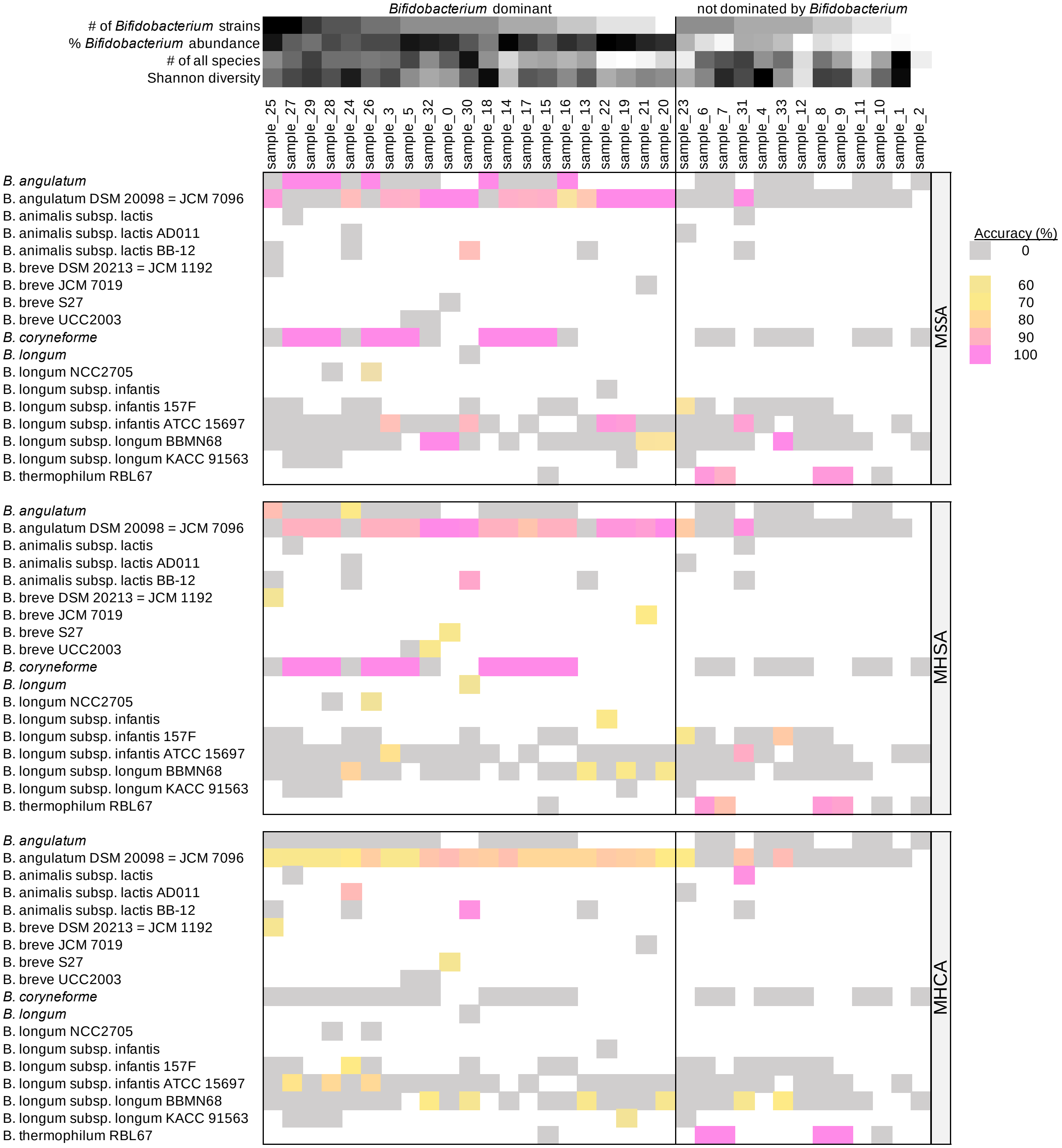


**Figure 16: Heatmap of CAZyme prediction accuracy within Bifidobacterium MAGs reconstructed from MSSA, MHSA, and MHCA for each sample.** Each cell represents the CAZyme prediction accuracy of a MAG retrieved from MSSA, MHSA, or MHCA for one of 34 samples. CAZyme prediction accuracy was calculated against the reference genome for the strain that each MAG was identified to be derived from. Sample community composition metrics are reported above the heatmap, including the number of *Bifidobacterium* strains in the sample, the relative abundance of the *Bifidobacterium* genus in the sample, number of total species in the sample, and sample Shannon Diversity.

Lastly, bifidobacterial strains that were phylogenetically distinct consistently had high CAZyme prediction accuracy, regardless of relative abundance. One example is *B. thermophilium RBL67*, which was present as a singular strain in five non-*Bifidobacterium* dominant samples, and was successfully binned in four samples, in all three assembly types. *B. coryneforme* was another lone strain that was successfully binned by MHSA and MSSA but not binned at all in MHCA. Both lone strains had high CAZyme prediction accuracy (**Figure 16)**, with their CAZyme profiles strongly matching with that of their reference genome. This can be seen in **Figure 17** and **Figure 18**, which represents CAZyme predictions for all *B. thermophilium* and *B. coryneforme* MAGs and reference genomes. Curiously, *B. coryneforme* was binned by all assembly types except MHCA.

In comparison, with some samples having as many as four strains of *B. longum*, high CAZyme prediction accuracy for MAGs belonging to these strains was inconsistent, especially for MAGs reconstructed from MEGAHIT assemblies. In some occasions, samples had multiple strains of subspecies such as *B. longum subsp. infantis*. Not only did these MAGs have low CAZyme prediction accuracy, but CAZyme profiles had missing genes (false negatives), or additional genes that did not belong to a particular strain (false positives). Such “hybrid” MAGs can be seen in **Figure 19**, which represents CAZyme predictions for all *B. longum subsp. infantis* MAGs and reference genomes across different samples and assembler types.

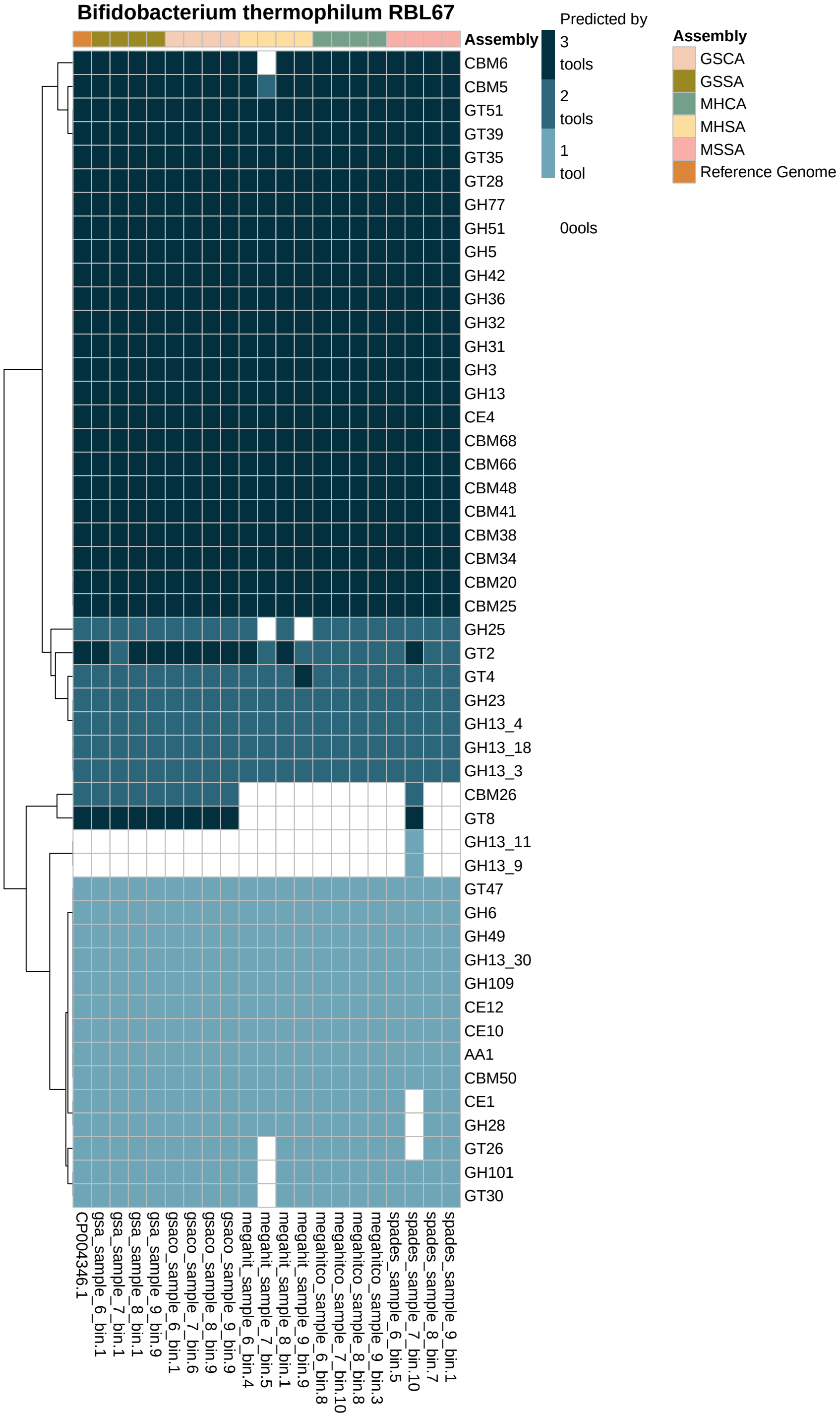


**Figure 17: CAZyme predictions for all *B. thermophilium* MAGs and reference genomes across different samples and assembler types.** The bottom axis represents MAGs, while the right axis represents predicted CAZymes. Each MAG is color-coded with the assembly-type it was derived from.

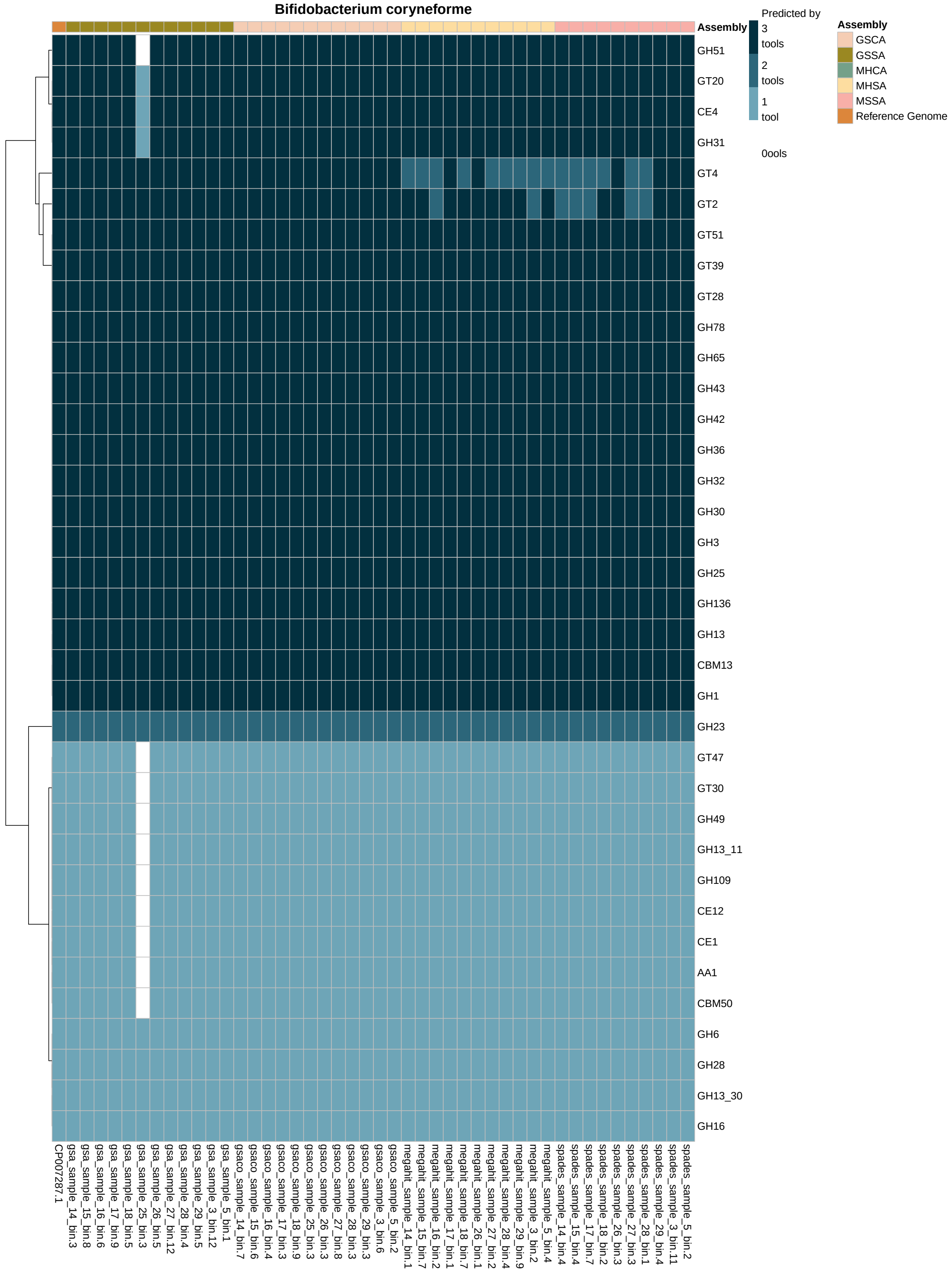


**Figure 18: CAZyme predictions for all *B. coryneforme* MAGs and reference genomes across different samples and assembler types.** The bottom axis represents MAGs, while the right axis represents predicted CAZymes. Each MAG is color-coded with the assembly-type it was derived from.

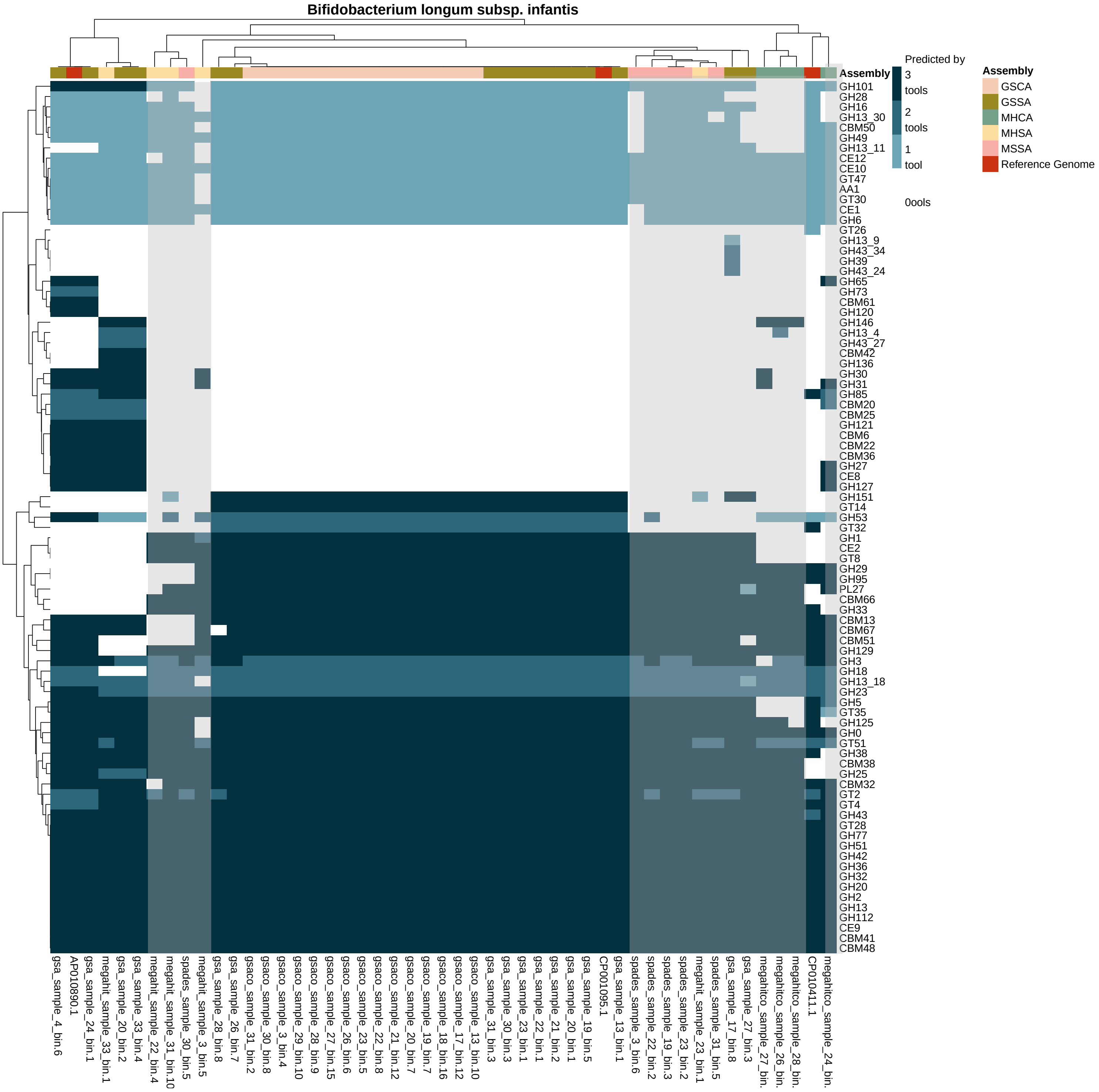


**Figure 19: CAZyme predictions for all *B. longum subsp. infantis* MAGs and reference genomes across different samples and assembler types.** This heatmap includes the visualization of three strains of *B. longum subsp. infantis*, demonstrating the the prevelance of “hybrid” MAGs. MAGs with potentially more than 1 strain are highlighted in grey.

## 3.8 Overall comparison of MAGs and MAG-level CAZyme predictions and gold-standard reference genome CAZyme predictions

Given the assembler-specific differences in *Bifidobacterium* MAG recovery and CAZyme prediction accuracy, we decided to broaden this analysis to include MAG recovery profiles from each sample, in addition to CAZyme prediction profiles for each reference genome and its respective MAGs. Consistent with our previous findings from section **3.7**, the five assembly methods recovered different MAGs within each sample. While this was observed in all samples, two specific examples are Sample 19 and Sample 21 (**Figure 20A-B**). MAGs from MHCA and MHCA had lower reference coverage than GSCA, GSSA, and MSSA. The MSSA recovered the fewest MAGs, but with the highest reference coverage. Overall, since each assembly method yielded different MAGs, their combined outputs recovered a larger share of strains from the metagenome than any single method.

A unique example of MAG recovery in samples with high strain diversity is Sample 25, where two strains of *B. longum subsp. infantis* were the first- and second-most abundant. Curiously, neither of the two abundant strains were binned, while two less abundant but phylogenetically distinct strains of *Bifidobacterium*—*B. angulatum* and *B. breve*—were recovered as MAGs instead (**Supplementary Figure 5)**. More examples from samples with different sample complexities can be seen in **Supplementary Figures 6-7**.

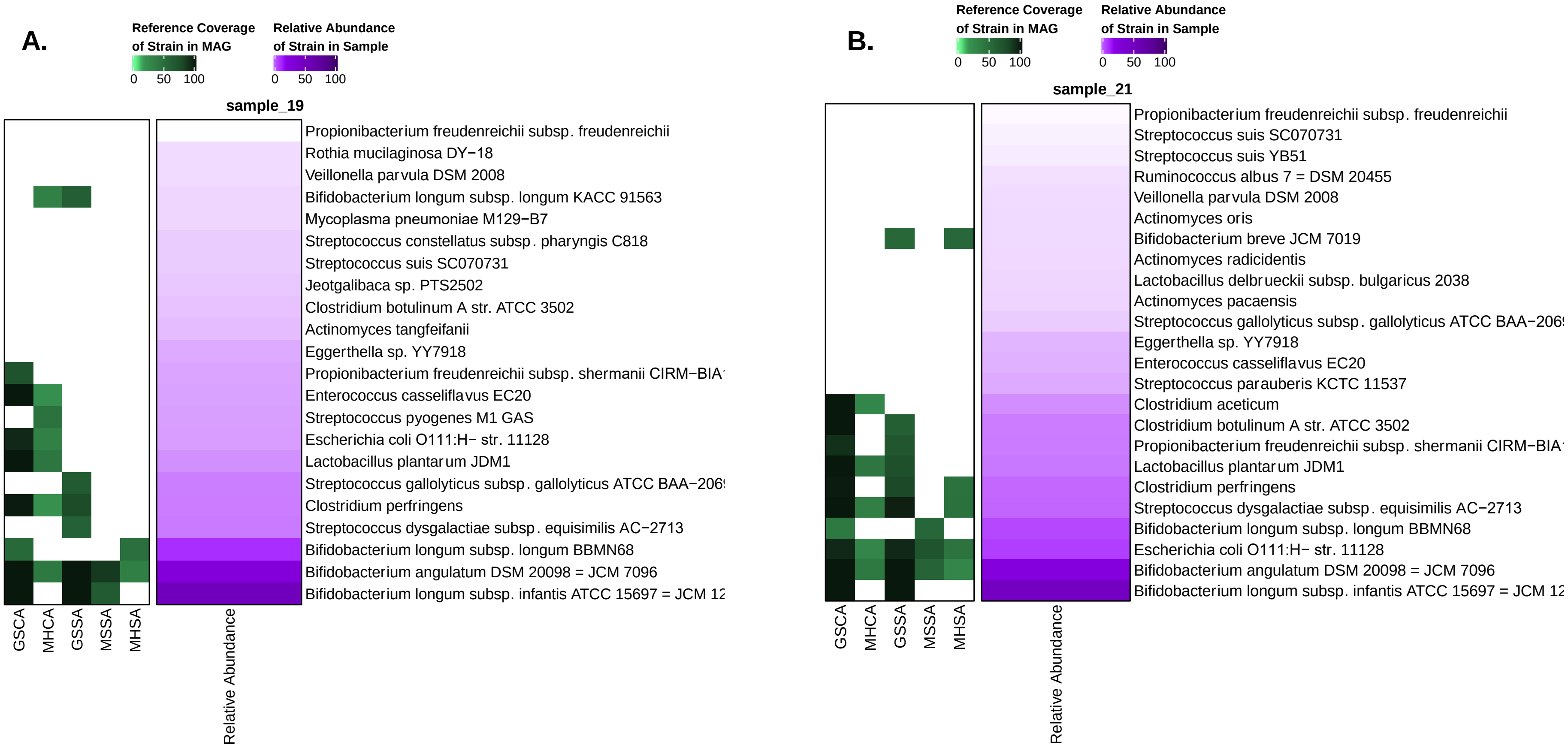


**Figure 20: Sample-specific MAG recovery profiles, with relative abundances of all strains in each sample and reference coverages of all retrieved MAGs. A.** Sample 19. **B.** Sample 21. For each heatmap, the right-most panel represents a list of all strains in the sample. The middle heatmap panel represents the relative abundance of samples. Each cell in the left-most panel represents whether a MAG was recovered for a particular strain by a particular assembly. Darker shades of purple and green represent relative abundance and reference coverage respectively, between 0% and 100%.

CAZyme profiles from non-*Bifidobacterium* MAGs and strains were also consistent with our previous results. MAG CAZyme predictions were most complete for species with singular strains that were phylogenetically distinct, such as *Akkermansia muciniphila* **(Supplementary Figure 8)**. Abundant strains of dominant species, such as the strain *E. coli O111:H− str. 11128,* also had complete CAZyme profiles. (**Supplementary Figure 9**). In contrast, rare strains of dominant species with high strain diversity had more gaps in their MAGs. Two examples of such strains are *E. coli BW2952* and E*. coli O104:H4 str. 2011C−3493* (**Supplementary Figure 10**). Overall, while MAGs from rare strains had the advantage of less contamination and fewer false positives, these MAGs were less complete, with more false negative CAZyme predictions.

# 4 Chapter 2: Development of a Metagenomic Pipeline

Bioinformatics analysis of metagenomic data is a complex process, and protocols for such data analysis can rarely be generalized (Chen *et al*., 2020). Even when studying a relatively simple community such as the infant gut, there are different challenges, such as the high within-species and within-strain diversity of abundant genera (Van Rossum *et al*., 2020). Most often, data analysis pipelines must be tailored to the research questions, organisms of interest, community composition, sample size, and study design.

After evaluating and quantifying the effects of microbial community composition and assembly quality on MAG recovery and CAZyme prediction, our objective was to build a reproducible and computationally feasible standard metagenomic pipeline to analyze infant gut metagenomic data. The MetAGenomic Analysis PIpelinE, or MAGPIE, is a minimal pipeline that can be customized for samples of different community composition. An instance of MAGPIE was cloned at customized for samples from the Baby & Mi study, which includes samples from the Baby, Food & Mi and Baby & Pre-Mi cohorts.

## 4.1 Rationale for MAGPIE and bfm_mg_flow

Mg_workflow was designed to evaluate the success of metagenomic tools in recovering strain-level MAGs for all organisms. While MAGPIE was modeled after mg_workflow, essential changes were made to tailor this pipeline for infant gut microbiome data, based on the results from **Chapter 1**. Similarly, bfm_mg_flow was further adapted for Baby, Food & Mi samples. For the evaluation of Baby, Food & Mi metagenomic data, the primary bacterial genus of interest was *Bifidobacterium*, followed by other carbohydrate-degrading bacteria, such as *Bacteroides*. The *Bifidobacterium* genus was chosen due to the well-established role of bifidobacterial species in infant gut microbial succession, and their presence in the gut throughout life (Arboleya *et al*., 2016b; Milani *et al*., 2017a). Consequently, bfm_mg_flow was tailored to recover high quality MAGs and CAZymes from abundant carbohydrate-degrading bacterial genera with high levels of within-genus diversity, based on the results from **Chapter 1**.

## 4.2 MAGPIE: MetAGenomic Analysis PIpelinE

MAGPIE is an automated metagenomic pipeline, designed to run on high-performance computing (HPC) environments, with and without a workflow management system. The default workflow manager for the pipeline is Slurm, although this can be modified in the configuration. MAGPIE is not designed for any specific datasets, and is intended to be adapted as per the research questions of the user. As input, the pipeline requires raw forward and reverse FASTQ files, and a sample file with at least one column containing sample names. The rule graph for the pipeline can be seen in **Figure 21**. The following adaptations were made to MAGPIE after mg_workflow:

### 4.2.1 Separation of MetaSPAdes read error correction and assembly

Prior to running, MAGPIE was tested on a small subset of samples. Unlike mg_workflow, since MAGPIE was tested—and intended to be used on—real metagenomic samples, additional errors were noted. Metagenomic samples with smaller insert sizes were often not assembled, or responsible for causing delays in MetaSPAdes's error correction and k-mer counting steps. After trial and error, separating the MetaSPAdes assembly into separate rules of error correction and assembly resolved the problem.

### 4.2.2 Addition of MaxBin2 and Das Tool

The second major change was the inclusion of three binning tools instead of one. This change was made based on the results from **Chapter 1** regarding MAG completeness and reference coverage metrics for MHCA, MHSA, and MSSA (**Table 4**), in addition to MAG contamination and false positives in CAZyme prediction accuracy. While MAGs from the MSSA had the best mean (95% CI) reference coverage out of the three at 57.8% (3.3), a higher reference coverage could potentially be retrieved with better binning. A second binning tool, MaxBin2, was thus added to the workflow. DAS Tool was added to create combined bins based on the results from MetaBat2 and MaxBin2.

## 4.3 bfm_mg_flow

Derived from MAGPIE, bfm_mg_flow is an automated metagenomic pipeline, designed to run in HPC environments, with and without a workflow management system. Unlike MAGPIE, bfm_mg_flow has been adapted to run specifically on the Baby, Food & Mi samples, and has been customized for our specific research questions. Bfm_mg_flow has been successfully used on all Baby, Food & Mi samples. The rule graph for the pipeline can be seen in **Figure 22**. As input, the pipeline requires raw forward and reverse FASTQ files, and a sample metadata file with sample groupings, such as participant ID, and date of sample collection. The following additions and changes were made to bfm_mg_flow after being cloned from MAGPIE:

### 4.3.1 Addition of reads-based analyses

The first major change to bfm_mg_flow was the addition of reads-based taxonomic and metabolic predictions with the Whole Metagenome Shotgun (wmgx) pipeline from the bioBakery3 workflows. Given the high number of false positive CAZyme predictions at the assembly- and MAG-level analyses noted in **Chapter 1**, a reads-based marker gene approach was included, in the case that metagenomic assemblies and bins would be of lower quality than expected. MetaPhlAn3 was added to the pipeline for taxonomic profiling, and HUMAnN3 to profile the identity and abundance of microbial metabolic pathways. The inclusion of the latter was also in line with the research goal of identifying bacterial carbohydrate degrading genes during the introduction of solid foods to the infant diet.

### 4.3.2 Addition of sample metadata-based statistics

The second change to the bfm_mg_flow pipeline was the addition of the bioBakery3 wmgx_vis workflow, to agglomerate the results from wmgx based on Baby & Mi sample meta-data. The two factors included in the analysis were participant IDs, to evaluate infant-specific changes, and the time point of sample collection, to compare results before and after the introduction of solid food. The reason for this change was to answer our original research questions regarding the impact of solid food introduction on the infant gut microbiota.

### 4.3.3 Omission of MEGAHIT single- and co-assemblies

Lastly, all MEGAHIT assemblies were omitted from the workflow due to the high levels of contamination and low reference coverage in MAGs from MHCA and MHSA, as noted in **Chapter 1**. MEGAHIT assemblies had been less successful at obtaining rarer strains of our target *Bifidobacterium* species and had lower overall CAZyme prediction accuracy. MHCA and MHSA also had the highest levels of false positive gene predictions, which would be a significant disadvantage for prediction bacterial metabolism with carbohydrate-active enzymes.

While a MetaSPAdes co-assembly could have been included, it was omitted due to computational memory requirements, and due to previous results from **Chapter 1** reporting higher levels of strain heterogeneity in MAGs from co-assemblies. As such, the assembly strategy used for bfm_mg_flow was MSSA. Since the research goal of this project was to study strain-level metabolic and CAZyme activity, co-assembling samples by participant ID may hinder the ability to recover strain-level MAGs

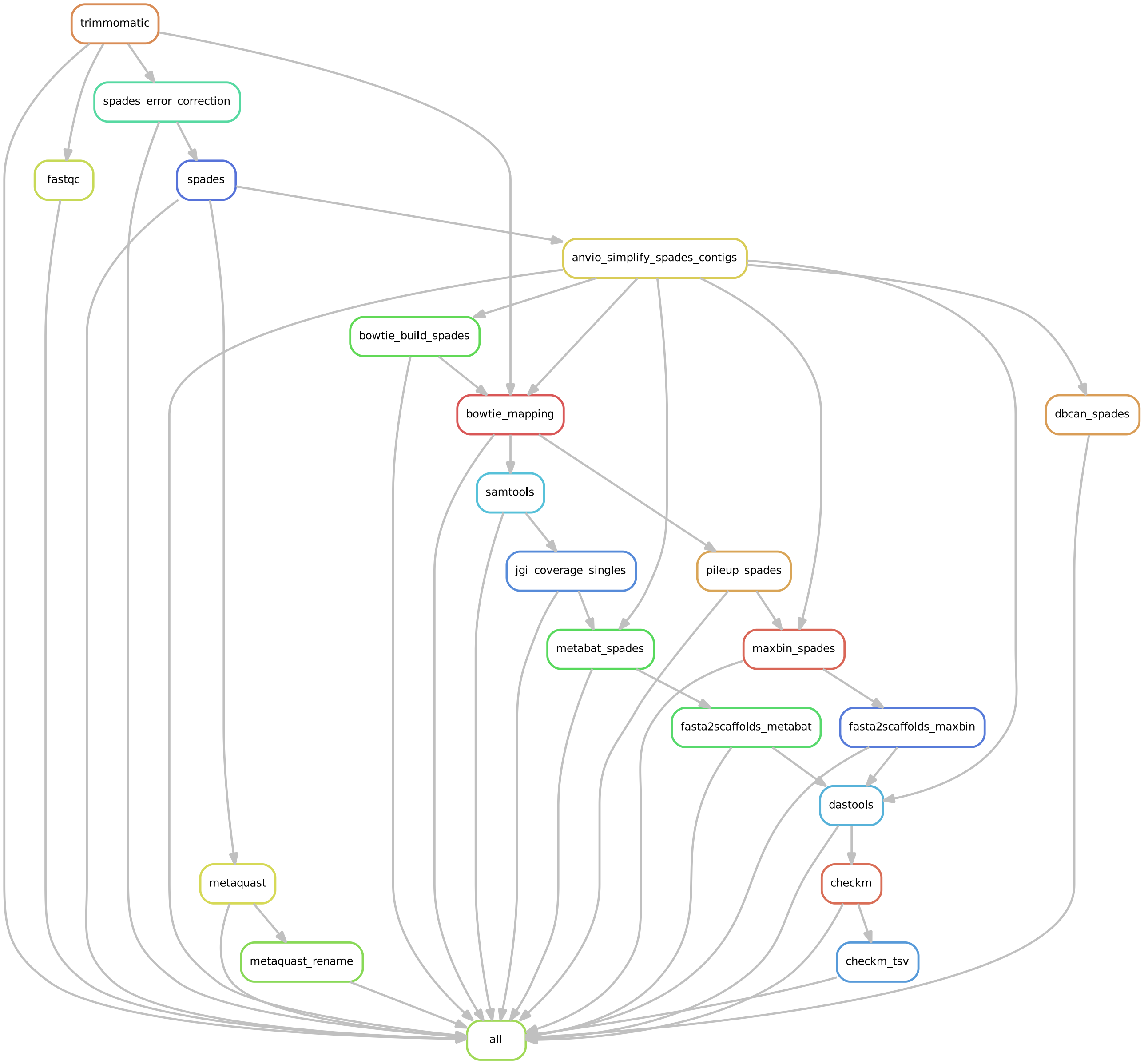


**Figure 21: Snakemake rule-graph for the MAGPIE metagenomic analysis pipeline**

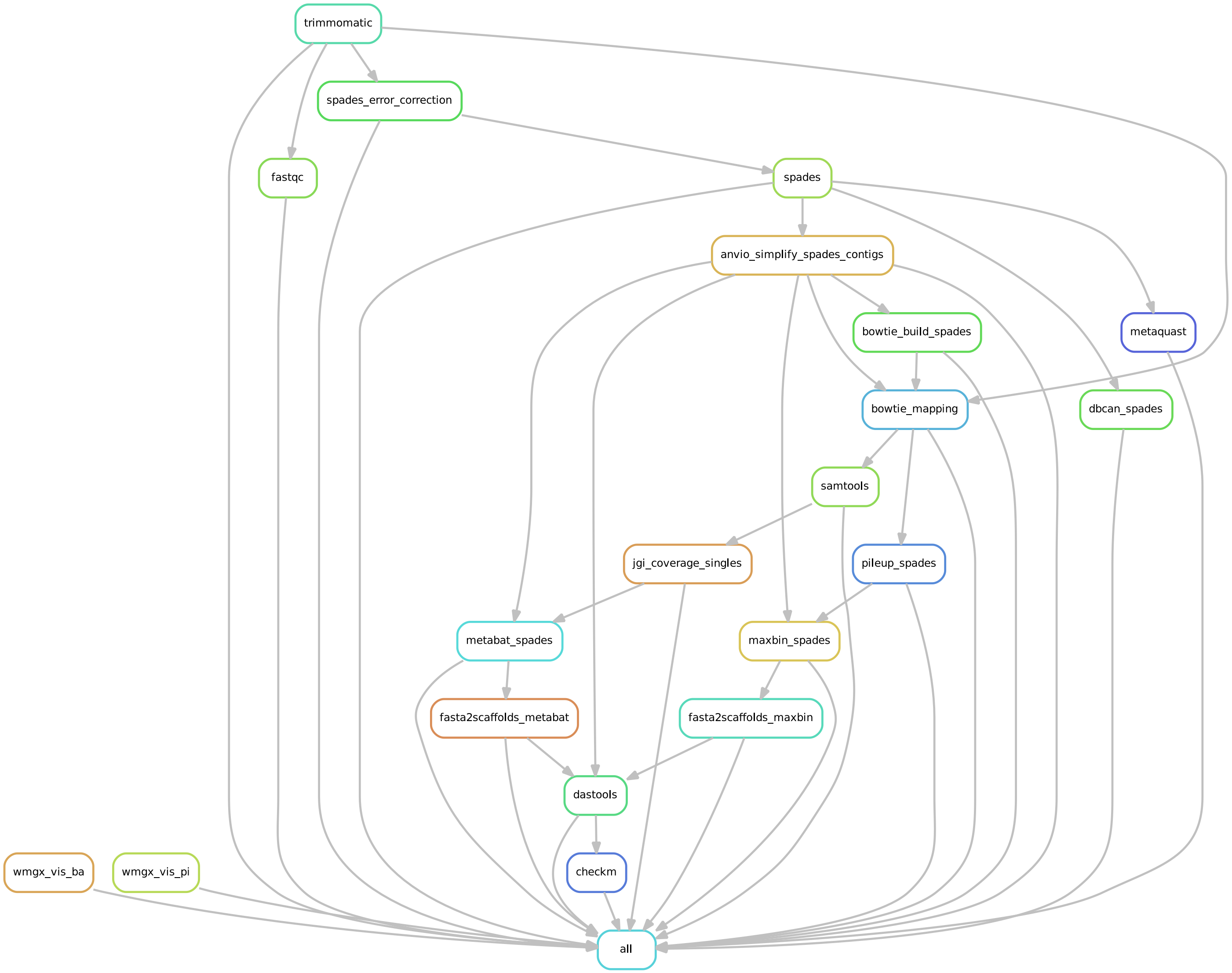


**Figure 22: Snakemake rule-graph for the bfm_mg_flow metagenomic analysis pipeline**

# 5 Chapter 3: The infant gut microbiome during solid food introduction

Prior to the introduction of solid foods, the gut microbiota of breastfed infants is composed primarily of bacteria that are able to metabolize and derive energy from HMOs. Solid food introduction brings new food glycans to the gut environment, including dietary fibers and starches, and drives the selection of bacteria that have the CAZymes to degrade these glycans (Katoh *et al*., 2020; Koropatkin *et al*., 2012). Bacteria that are able to degrade the dietary fibers from solid food survive this transitionary period, while others are outcompeted (Ioannou *et al*., 2021).

With the development of a metagenomic pipeline, our next objective was to identify the changes in the infant gut microbiome over the introduction of solids, and to specifically understand how the metabolic activity of genera such as *Bifidobacterium* change over the solid food introduction period, allowing them to persist from infancy to adulthood. We hypothesized that samples obtained after solid food introduction will display an increased abundance of bacterial strains with the ability to metabolize non-HMO food glycans, and a higher number of carbohydrate-degrading genes.

## 5.1 Baby, Food & Mi: The infant gut microbiome during the introduction to solid foods

30 samples belonging to 15 infants from the Baby, Food & Mi cohort of the Baby & Mi study were processed with the bfm_mg_flow pipeline (**Supplementary Figure 1**, **Supplementary Table 2**). Two samples were collected for each infant, before and after the introduction of solid foods. The samples were analyzed on two different levels; firstly, samples were grouped based on participant ID to compare the microbiome between infants, and secondly, samples were grouped based on their timepoint of collection, to analyze broader changes based on solid food introduction.

### 5.1.1 Read-level taxonomy and abundance over solid food introduction

At the Order level, samples were primarily dominated by *Bifidobacteriales* and *Bacteroidales,* in addition to *Enterobacterales*, *Ruminococcus*, *Clostridia*, *Vellionellales*, with some samples having *Erysipelotrichales* (**Figure 23)**. Before the introduction of solid foods, 11 samples were dominant in *Bifidobacteriales*, two samples were dominant in *Bacteroidales*, one sample was dominant in *Enterobacterales*, and one dominant in *Erysipelotrichales*. After the introduction of solids, eight samples were dominant in *Bifidobacteriales*, six were dominant in *Bacteroidales*, and one dominant in *Clostridiales* (**Figure 24A)**.

Before solid food introduction, samples had a high relative abundance of *Bifidobacterium longum*, *Bifidobacterium bifidum*, *Bifidobacterium breve*, *Bacteroides fragilis*, *Bacteroides vulgatus*, *Erysipelatoclostridium ramosum*, and *Escherichia coli* **(Figure 24B-C).** After the introduction of solids, five samples were dominant in *B. bifidum*, two in *B. breve*, one in *B. longum*, two in *B. fragilis*, two in *Bacteroides uniformis*, one in *Bacteroides faecis*, one in *Bacteroides dorei,* and one in *Ruminococcus gnavus* (**Supplementary Table 3, Figure 24B-C).** Species-level changes before and after the introduction of solid foods can be seen in **Figure 24B**.

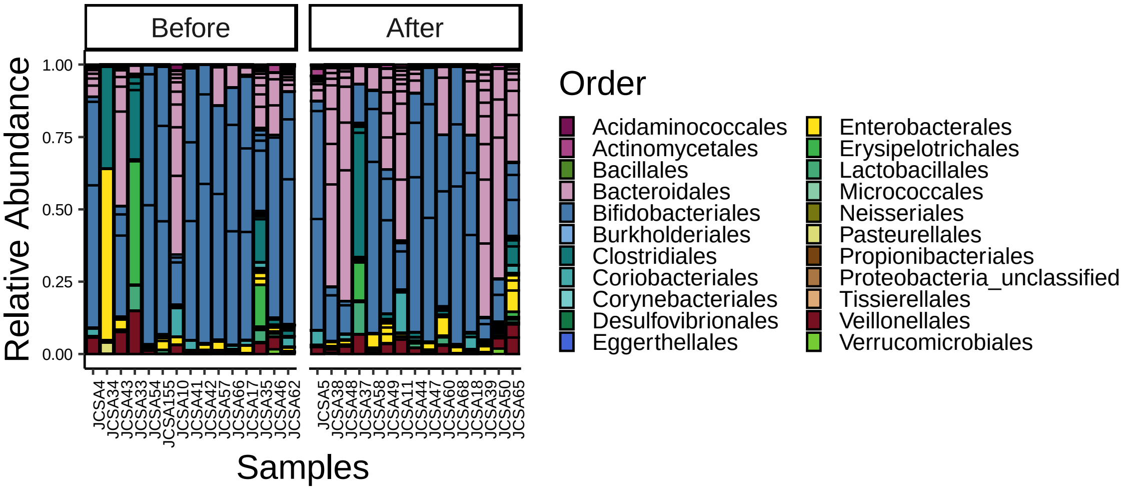


**Figure 23: Order-level taxonomic profiles of all Baby, Food & Mi metagenomic samples, before and after the introduction of solid foods.**

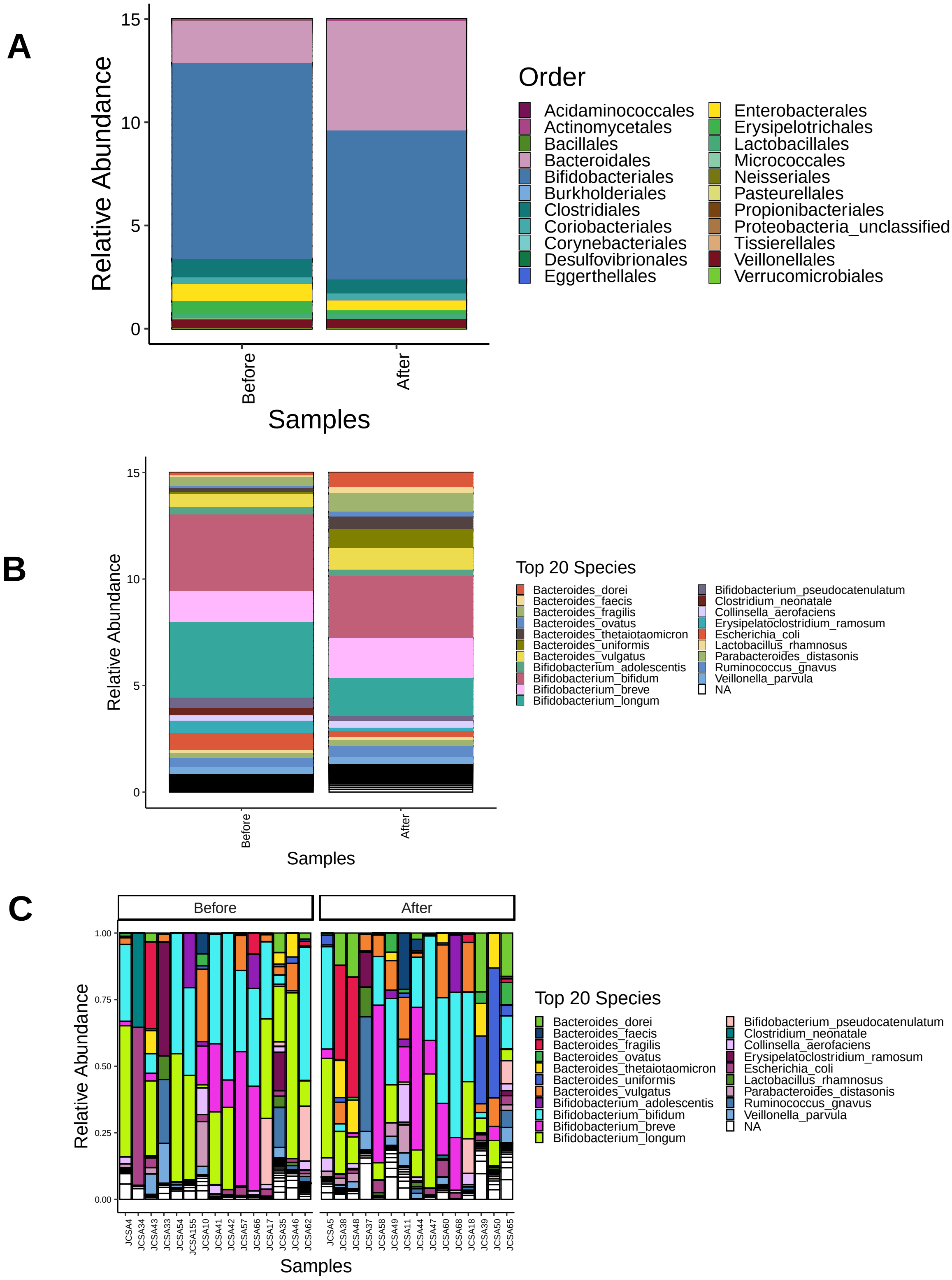


**Figure 24: Taxonomy changes in microbiome composition before and after the introduction of solid foods. A.** Order-level, **B.** species level, and **C.** species level per sample.

Despite *Bifidobacterium* being dominant in 11 samples prior to solid food introduction, they were only dominant in 8 after solid food introduction (**Supplementary Table 4, Figure 24C)**. Within the genus, the number of samples dominant in *B. longum* dropped from six to one, the dominance of *B. bifidum* increased from three to five, while the number of samples dominant in *B. breve* remained the same. The relative abundance of most *Bifidobacterium* species decreased after the introduction of solids. The diversity of *Bifidobacterium* remained the same.

While the abundance of *Bifidobacterium* decreased, the abundance of *Bacteroides* increased after the introduction of solids. Prior to the introduction of solid food, two samples were dominant in in *B. fragilis* and *B. vulgatus* respectively. After solid food introduction, six samples were dominant in *Bacteroides*. This change in distribution is also evident in **Figure 24**. The mean abundance of *B. fragilis*, *B. uniformis*, *B. fragilis*, *B. faecis*, *B. dorei* all increased after the introduction of solid foods.

There was also an overall decrease of abundance in *Erysipelatoclostridium* and *Escherichia*. The samples that had the highest abundance of by *E. ramosum* and *E. coli* showed increased abundance of *Bacteroides* and *Bifidobacterium* after the introduction of solids. The presence of *Veillonella* stayed consistent at both timepoints. In addition, there were few *Lactobacillus* species identified in this study.

Principal Coordinates Analysis (PCoA) plots were calculated for the taxonomic profiles of all samples, using Bray Curtis Dissimilarity. In **Figure 25A**, samples are labeled based on whether stool collection occurred before or after solid food introduction, while in **Figure 25B,** samples are labeled based on participant ID. The PCoA demonstrated that there was no significant separation of samples based on their time point, although the taxonomic profiles of individual participants were significantly different from each other (p < 0.001, PERMANOVA).

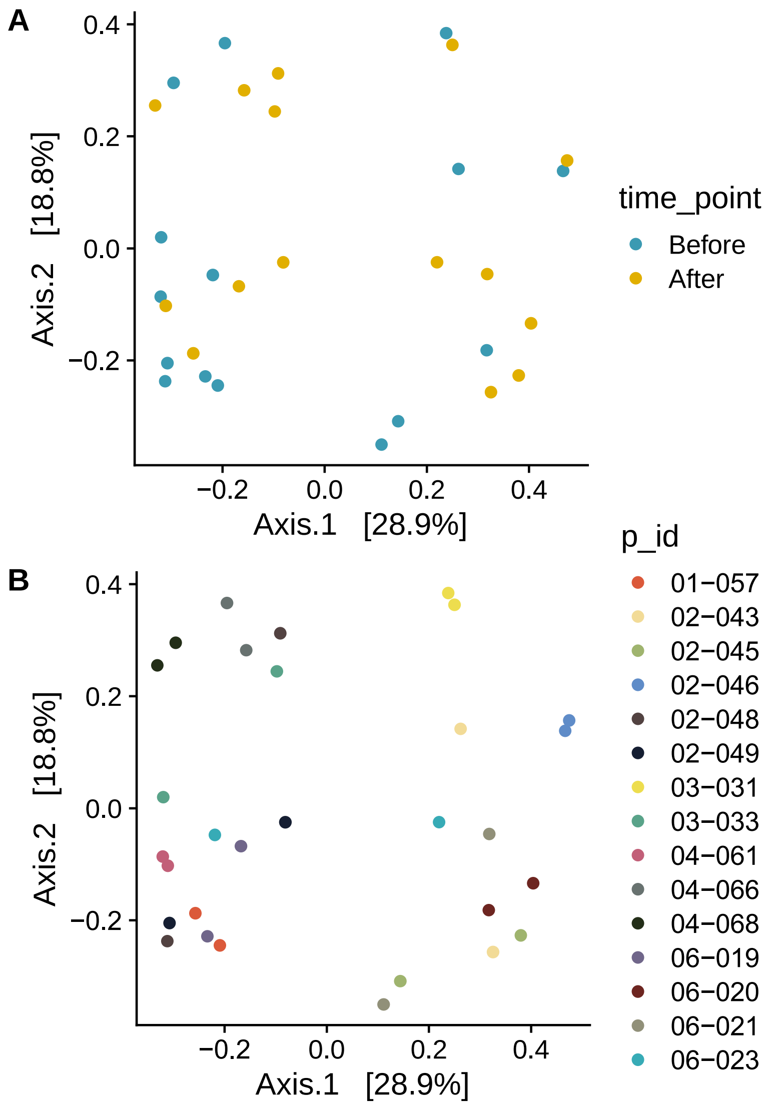


**Figure 25: PCoA plots of all samples in the Baby, Food & Mi cohort, labeled by participant IDs and stool collection time points (before and after the introduction of solid foods)**

### 5.1.2 Read-level metabolic predictions over solid food introduction

Humann3 was used to calculate the abundance of bacterial metabolic pathways in all samples. As a starting point, the top bacterial metabolic pathways across all samples were calculated and compared before and after solid food introduction (**Supplementary Table 5; Supplementary Figure 11**). While differences based on sample timepoint were not immediately clear, the predictions were consistent with our expectations of characterized bacterial carbohydrate degradation pathways. Among the top abundant pathways were the various steps of the superpathway of branched chain amino acid biosynthesis, and pathways for Uridine Monophosphate biosynthesis, starch degradation, glycolysis III (from glucose), glycolysis IV (plant cytosol), glycogen degradation, and sucrose degradation IV (sucrose phosphorylase).

Since the mean abundant pathways primarily included core microbial metabolic genes, we subsequently extracted the top pathways for our genera of interest. Based on these results, metabolic predictions of interest were also filtered to specifically characterize bacterial carbohydrate degrading genes and metabolic pathways of interest. All MetaCyc carbohydrate degrading pathways identified in the samples can be seen in **Appendix D**.

Our first genus of interest was *Bifidobacterium*, due to the high abundance and dominance of bifidobacterial species in our samples (**Figure 26**). Predictably, the most abundant *Bifidobacterium* pathway across all samples was the bifid-shunt, or the “fructose-6-phosphate-shunt”, to break down glucose and fructose to lactic acid and acetate (Milani *et al*., 2014; O'Callaghan and van Sinderen, 2016) (**Figure 27**). The bifid-shunt was detected in *B. longum*, *B. dentium*, and *B. pseudocatenulatum,* and partially in *B. breve* (**Figures 27-28, Supplementary Figures 12-13**). Consistent with the overall reduction in abundance of bifidobacterial species after solid food introduction, the bifid-shunt pathway also decreased in abundance. The starch degradation VI pathway was seen in *B. longum*, *B. bifidum*, and *B. breve*. There was a decrease in abundance of *B. longum* species carrying this pathway after solid food introduction, with a corresponding increase of abundance of the pathway in an unclassified species of bacteria.

*B. breve, B. longum*, *B. bifidum*, and *B. adolescentis* had the pathway for pyruvate fermentation to the SCFAs acetate and lactate. In addition, *B. breve* was the only *Bifidobacterium* to have the glucose and glucose-1-phosphate degradation pathway (**Figures 28-32**). More pathways of interest included the presence of the sucrose degradation IV pathway, which is a part of the bifid-shunt pathway, in *B. bifidum*, *B. dentium*, and *B. breve*. The glycogen degradation I pathway was also one of the most abundant in *B. breve* and *B. bifidum*.

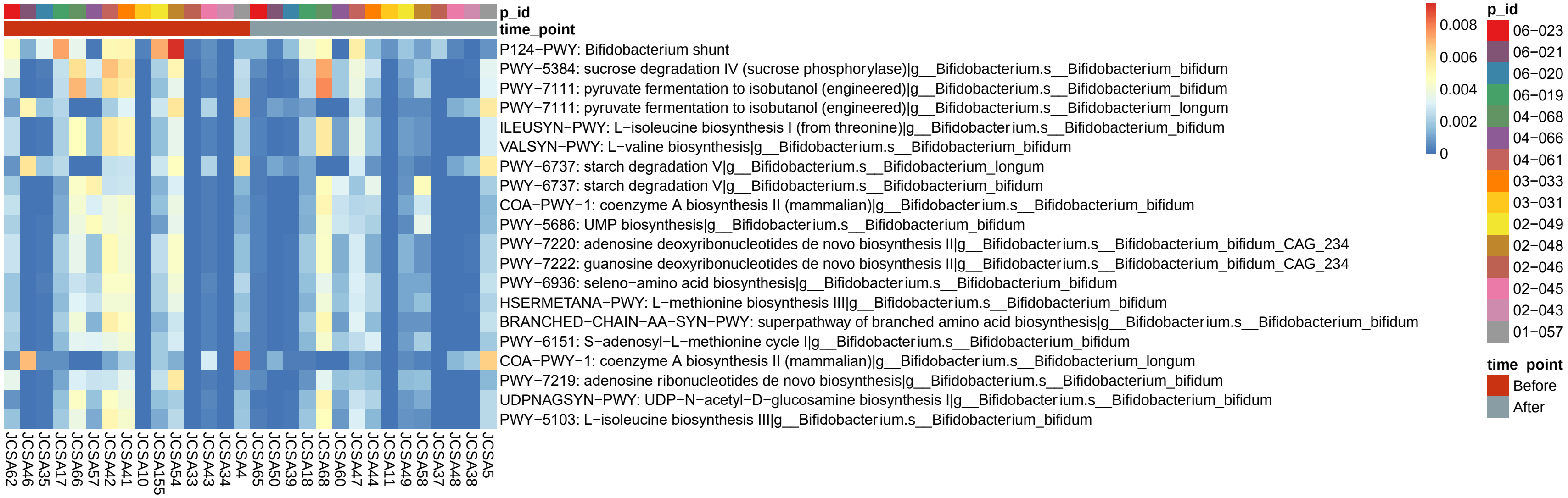


**Figure 27: 20 most abundant microbial metabolic pathways predicted for *Bifidoabacterium*, before and after the introduction of solid foods. The p_id refers to each individual study participant. Blue represents zero pathway abundance, in a gradient to higher pathway abundance in red.**

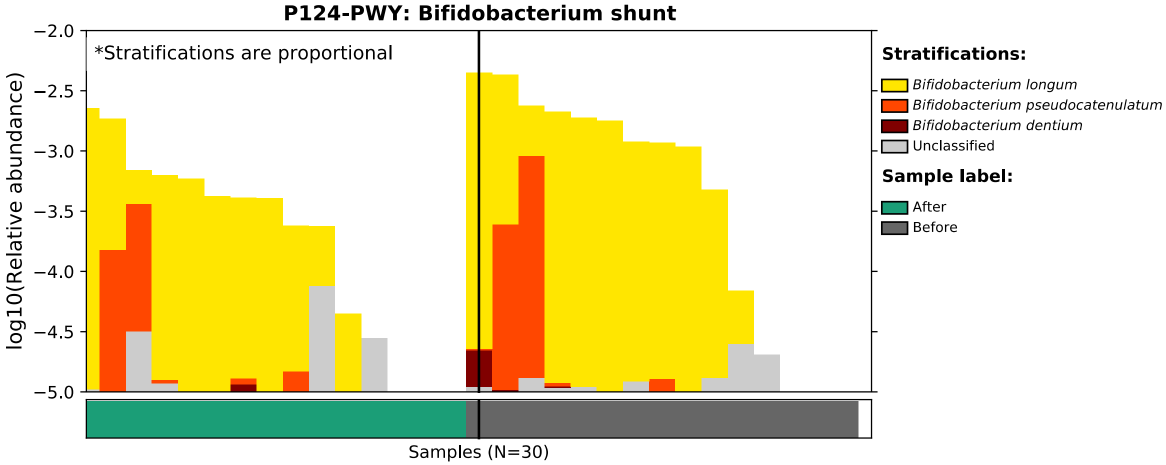


**Figure 26: Abundance of the bifid-shunt pathway before and after solid food introduction.**

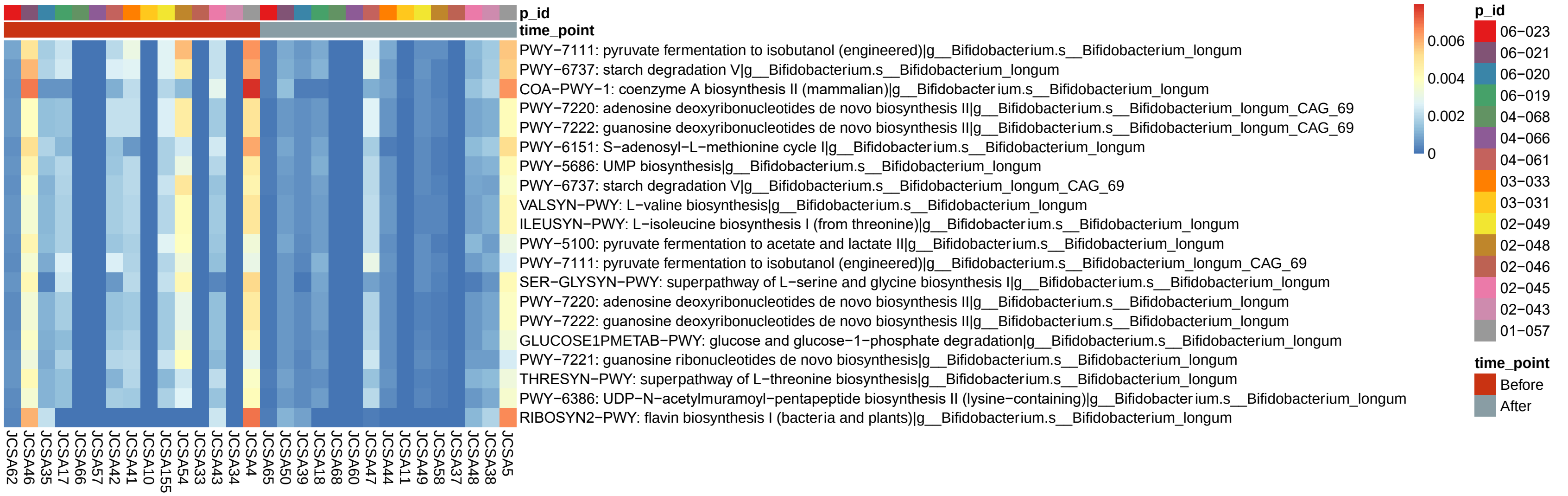


**Figure 28: 20 most abundant microbial metabolic pathways predicted for *B. longum*, before and after the introduction of solid foods. The p_id refers to each individual study participant. Blue represents zero pathway abundance, in a gradient to higher pathway abundance in red.**

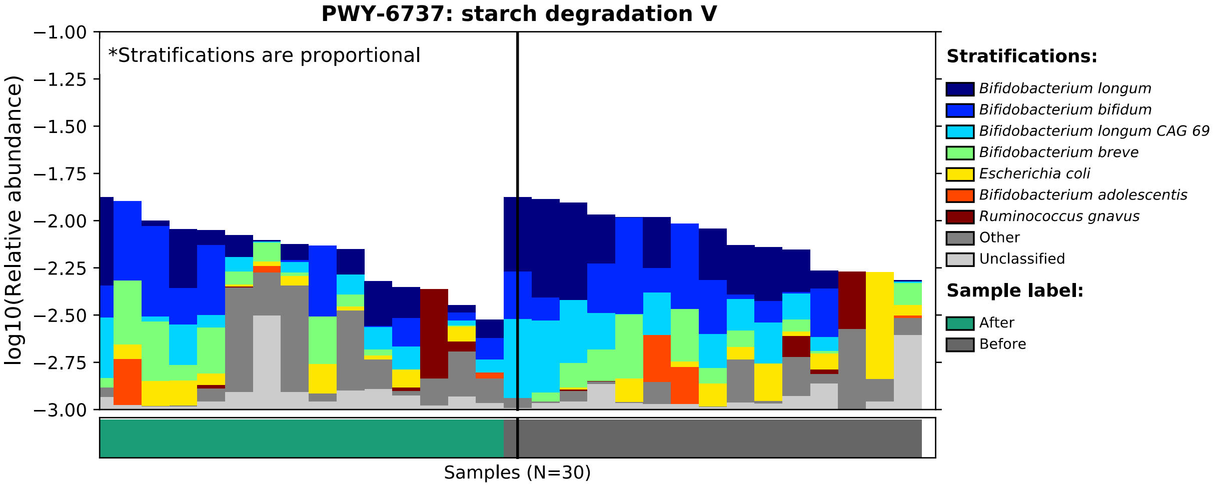


**Figure 29: Abundance of the starch degradation V/VI pathway, before and after solid food introduction.**

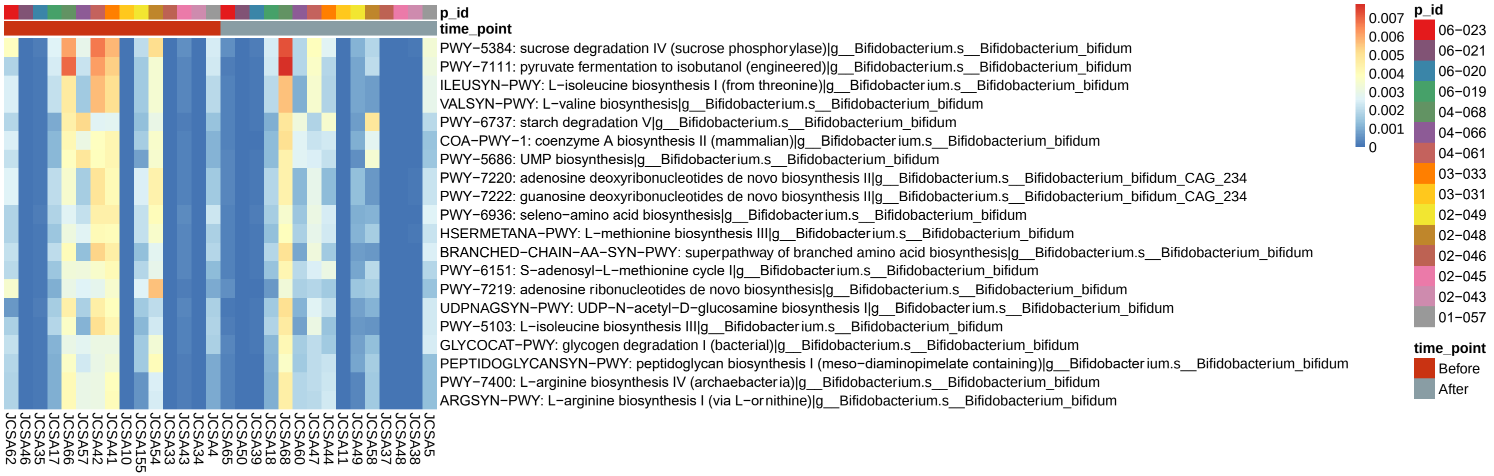


**Figure 30: 20 most abundant microbial metabolic pathways predicted for *B. bifidum*, before and after the introduction of solid foods. The p_id refers to each individual study participant. Blue represents zero pathway abundance, in a gradient to higher pathway abundance in red.**

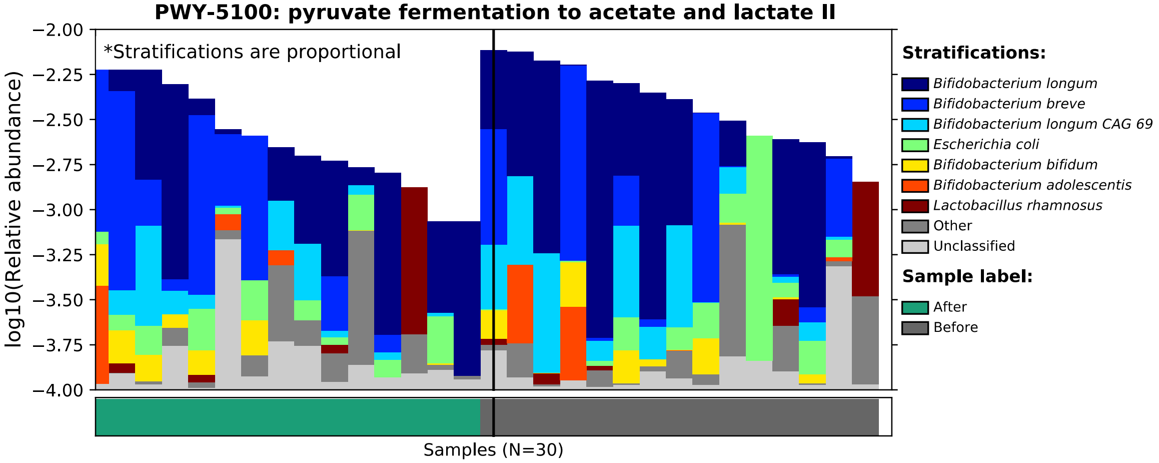


**Figure 31: Abundance of the pyruvate fermentation to acetate and lactate II pathway, before and after solid food introduction.**

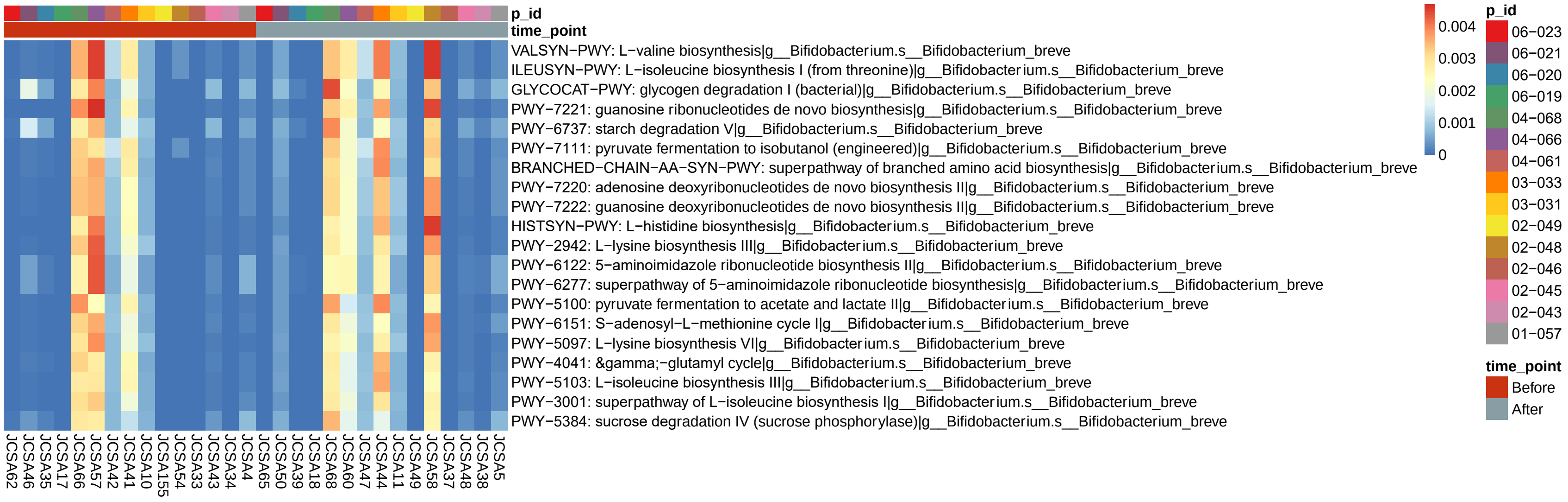


**Figure 32: 20 most abundant microbial metabolic pathways predicted for *B. breve*, before and after the introduction of solid food. The p_id refers to each individual study participant. Blue represents zero pathway abundance, in a gradient to higher pathway abundance in red.**

Our next genus of interest was *Bacteroides*, due to their increase in abundance, dominance, and diversity after the introduction of solid foods. Consistent with the taxonomic profiles, there was an increase in abundance of microbial metabolic predictions after solid food introduction in *B. dorei, B. fragilis, B. uniformis,* and *B. vulgatus* (**Supplementary Figures 14-16**). Unlike *Bifidobacterium*, two separate species of *Bacteroides—B. dorei and B. vulgatus–*–had the pathway for glycolysis IV. *B. uniformis* additionally had a pathway for L-histidine degradation.

Interestingly, multiple species of *Bacteroides*, *E. coli*, and *Ruminoccocus gnavus* all had an L−rhamnose degradation pathway, which is a deoxy-hexose sugar that is commonly found in pectins and hemicelluloses (Rodionova *et al*., 2013). Consistent with the increased dominance of Bacteroides after solid food introduction, the pathway also increased in abundance in *B. uniformis*, *B. thetaiotaomicron, B. ovatus*, and *B. faecis* (**Figure 33).** *B. fragilis* was also involved

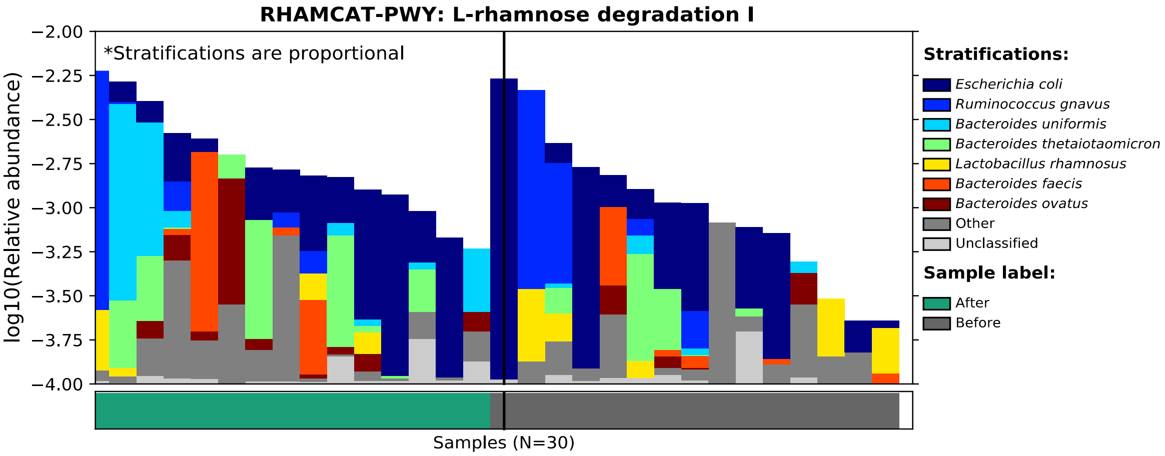


**Figure 33: Abundance of L-rhamnose degradation I pathway, before and after solid food introduction.**

in dTDP-β-L-rhamnose biosynthesis and dTDP-N-acetylviosamine biosynthesis, which are important for the formation of bacterial lipopolysaccharides (Coyne *et al*., 2000).

*E. coli* contained pathways for D−galactarate and D−galactarate degradation into pyruvate. In addition, both *E. coli* and *Lactobacillus rhamnosus* contained pathways for hexitol fermentation to lactate, formate, ethanol and acetate (**Figure 34**). Hexitols include sugar alcohols such as D-mannitol, D-sorbitol, and galactitol (Lengeler, 1975). The abundance of the hexitol fermentation pathway was higher prior to the introduction of solid foods. Both organisms also included a pathway for the homolactic fermentation of sugars into lactate.

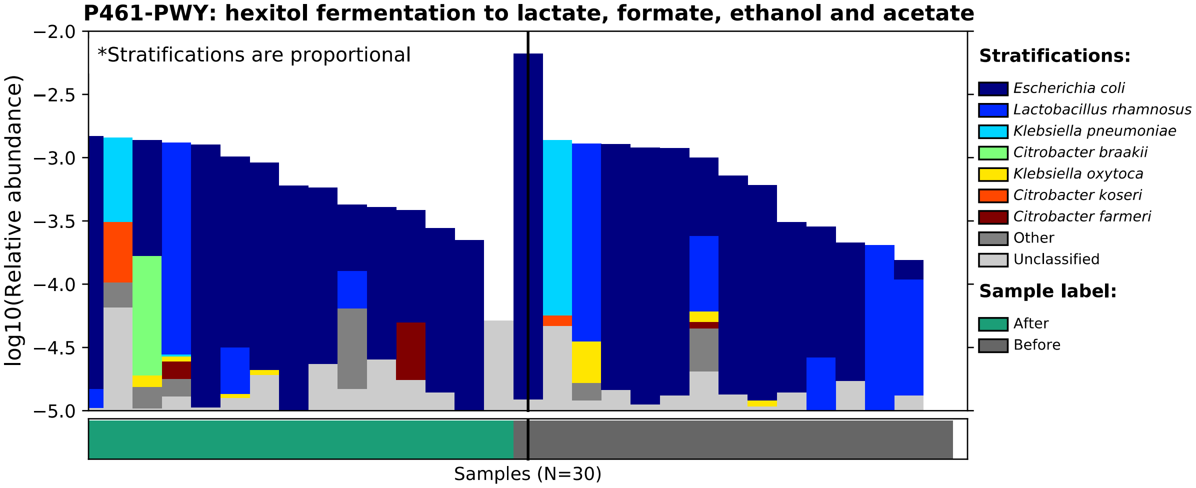


**Figure 34: Abundance of the hexitol fermentation to lactate, formate, ethanol, and acetate pathway, before and after solid food introduction**

# 6 Discussion

When studying complex microbial communities such as the gut microbiota, having access to bacterial genomic information is crucial to understanding the overall community function and activity in the context of their environment (Chen *et al*., 2020). With the technological advances in sequencing, the limiting step of culturing and sequencing individual isolates from environmental and metagenomic samples has largely been eliminated (Van Rossum *et al*., 2020; Tyson *et al*., 2004) However, the various shortcomings of metagenomic analysis—such as errors in assembly due to closely-related strains, and loss of genetic information due to un-assembled regions—are a major caveat when using MAGs to derive biologically-relevant conclusions (Chen *et al*., 2020; Marbouty and Koszul, 2015). While various efforts have been made to quantify the extent that assembly, binning, and gene prediction tools are impacted by sample complexity, and how much they individually impact the accuracy of the final predictions, previous benchmarking efforts have not typically focused on the infant gut microbiome.

The objective of this thesis was to quantify the performance of standard bioinformatics tools on infant gut metagenomic data, build a robust pipeline to obtain high quality metagenome-resolved genomes and bacterial metabolic pathways, and to use the pipeline to evaluate the infant gut microbiome and carbohydrate-active enzymes over the period of solid food introduction.

## 6.1 Assembly and binning performance in simulated infant gut metagenomic samples

Often dominated by very few highly abundant genera, with high levels of inter-genus species and strain diversity, the infant gut poses a unique challenge to metagenomic data analysis (Milani *et al*., 2017a). While computational tools for metagenomic datasets have previously been benchmarked, their performance has not previously been evaluated specifically for infant gut microbiome data. Furthermore, the accuracy of strain-level CAZyme predictions in genome-resolved metagenomics has not previously been quantified, especially in the context of sample diversity.

Our research demonstrates that there is a relationship between infant gut microbial community composition and the performance of the computational tools that are used to study those very communities. Assembly tools are impacted by sample Shannon Diversity, the number of strains in the sample, and the relative abundance of the most dominant strain. In turn, assemblies impact the quality of recovered MAGs and strain-level CAZyme prediction accuracy, proving our initial hypothesis that higher community complexity leads to lower quality MAGs and less accurate CAZyme predictions for abundant genera.

We found a significant difference in assembler performance between MetaSPAdes and MEGAHIT, with assemblies and MAGs from the latter having a significantly higher number of misassemblies and false positive gene predictions respectively. However, the MHSA assembly was also larger than the MSSA, likely recovering a higher proportion of the original metagenome. This was consistent with previous findings from the Critical Assessment of Metagenomic Challenge, where MEGAHIT had the highest number of misassemblies, but also the largest assembly length, ultimately recovering the larges total number of MAGs per assembly (Sczyrba *et al*., 2017). While we did not experiment with assembly parameters, the authors of the CAMI Challenge found that changing MEGAHIT parameters did not significantly impact assembly length or genome fraction (Sczyrba *et al*., 2017).

Consistent with previous work, strain diversity posed a challenge for assemblers (Sczyrba *et al*., 2017; Yue *et al*., 2020). Higher sample Shannon Diversity was correlated with higher number of misassemblies, while higher within-genus *Bifidobacterium* Shannon Diversity had a significant negative impact on the number and quality of final MAGs for MEGAHIT and MetaSPAdes assemblies, but not the gold-standard assemblies. In our results, MEGAHIT assemblies were also disproportionally impacted by sample strain diversity, showcased by the inferior performance of the MHCA and MHSA in recovering *Bifidobacterium* MAGs of higher CAZyme accuracy and higher reference coverage, especially in samples with abundant and diverse bifidobacterial strains. One potential reason for this could be additional read-correction step employed by the MetaSPAdes assembler (Nurk *et al*., 2017). However, it is important to note that other studies have demonstrated results where MEGAHIT outperformed MetaSPAdes in recovering MAGs with higher reference coverage (Maguire *et al*., 2020), and it may thus be possible that our results are specific only to the infant gut microbiome, due to its unique community composition.

Given a realistic assembly of simulated samples with MEGAHIT or MetaSPAdes, MAG-level CAZyme prediction quality decreased with every additional strain belonging to the same genus. In comparison, given a perfect assembly, MAG CAZyme prediction quality metrics stayed consistently high even with added inter-genus strain diversity. This implied that the assembly step, rather than the binning step, has a larger role in the correct prediction of metabolic genes. However, while we did not specifically compare the performance of binning tools, previous research has found significant differences between binning tools in generating MAGs with higher completeness and lower contamination (Yue *et al*., 2020). Yue *et al* additional found that binning performance was better for unique strains compared to closely related strains. The CAMI project similarly found that binning performance varied based on high- versus low-complexity datasets (Sczyrba *et al*., 2017). This is consistent with our comparison of *Bifidobacterium* MAGs across samples, where MAGs from unique bifidobacterial species such as *B. coryneforme, B. thermophilium*, and *B. angulatum* had higher CAZyme prediction accuracy compared to MAGs from similar strains of *B. longum*, such as different strains of *B. longum subsp. infatis*. Similarly, organisms with no or low within-species strain diversity, such as *A. municiphilia*, were observed to have more complete CAZyme predictions compared to organisms with high within-species diversity.

When comparing gold-standard, “ideal” co-assemblies to single-sample assemblies in overall assembly and MAG quality metrics, we found that the GSCAs modestly outperformed the GSSAs in every metric, except for the strain heterogeneity in recovered MAGs. But in practice, when comparing the ”real” MEGAHIT co-assembly to single-sample assembles, the MHSA and MSSA dramatically outperformed the MHCA, recovering MAGs of higher reference coverage and lower strain heterogeneity. While there have not been many comparisons of the merits of co-assembly versus single-sample assembly, previous work has reported that improvements from the co-assembling are moderate, and are most useful when co-assembling a larger number of samples (>5) from the same site, or alternately, a larger number of longitudinal samples from the same subject (Pasolli *et al*., 2019). Pasolli *et al*. have additionally reported that co-assembling does potentially recover more low-abundance strains, those MAGs tend to be consensus genomes, lacking strain-specific differences. They concluded that co-assembly may be most advantageous when the goal is to extract a larger amount of information from metagenomic samples, rather than when studying strain-specific differences (Pasolli *et al*., 2019).

Importantly, for all metagenomic samples, we found that the five different types of assemblies recovered a different subset of strains as MAGs. In general, MEGAHT assemblies recovered strains of lower relative abundance, while MetaSPAdes assemblies were biased towards strains of higher relative abundance.  For example, despite being present in 22 samples, *B. coryneforme* was never reconstructed as a MAG from the MHCA assembly. In contrast, *B. coryneforme* MAGs were retrieved from the MHSA and MSSA assemblies in ten separate samples each. Similarly, the *Bifidobacterium* MAGs retrieved for most samples differed based on the assembler. One example is sample 25, which contained eight *Bifidobacterium* strains, including *B. angulatum* and *B. angulatum DSM 20098 = JCM* 7096. Binning the MSSA assembly for sample 25 yielded only *B. angulatum DSM 20098 = JCM 7096*, while both MHCA and MSSA yielded only *B. angulatum*. In addition, MHCA and MSSA yielded a *B. breve DSM 20213 = JCM 1192* strain that was missed by MSSA. In both of the above cases, certain strains were only reconstructed into MAGs by certain assemblers.

## 6.2 Improvements and recommendations for infant gut metagenomic data analysis

Based on our results in the context of current metagenomic research in the literature, our first proposed recommendation for analyzing gut metagenomic data is to use multiple assembly and binning methods, contingent on the availability of computational resources. Since different assemblies yielded different MAGs from the same samples, the advantage of using multiple assemblers is that combined collection of MAGs would recover a larger share of strains from the original metagenomes than any singular output. Furthermore, while we did not use multiple binning tools, other studies have showcased improvements in binning when ensemble approaches are used and agglomerated (Maguire *et al*., 2020; Yue *et al*., 2020).

If computational resources are a limiting factor and only a few approaches can be used, our second recommendation is to tailor assemblers and assembly-methods based on organisms of interest in a metagenomic sample. Our results showed that assembly methods were impacted differently by sample community composition, including Shannon diversity, number of strains in the sample, and relative abundance of the most dominant strain. Furthermore, MAG completeness, contamination, quality, and reference coverage were significantly impacted by choice of assembly software. Consequently, if the goal is to study strain-specific differences and evolution in longitudinal samples, co-assembling will not be useful due to the reconstruction of composite genomes across samples (Pasolli *et al*., 2019). If the goal is to amass a strain-level catalog of microbial genomes from different sites and individuals, single-assembly would be the preferred approach (Xie *et al*., 2021). On the other hand, if the data belongs to different individuals or sites, but with high levels of similarity between sites, then co-assembly may be beneficial if the goal of the study is to amass a large number of MAGs specific to sites, but not if the goal is to study individual-specific differences (Hofmeyr *et al*., 2020). Based on our own results, if the strains of interest are of lower abundance, MEGAHIT assemblies would be more useful than MetaSPAdes, although the resulting MAGs would have lower reference coverage.

Our third proposed recommendation is to conduct multiple levels of analysis—read-based, assembly-based, or MAG-based—when appropriate for the research question. Based on our results, when making broad comparisons between sites or individuals, we found that assembly-level predictions can often have higher false positive gene predictions than MAG-level predictions. As such, when applying assembly-level predictions, it may be appropriate to take measures to protect against predictions errors due to misassemblies. This may include stringently removing low-quality or small contigs, or it may include a second read-level analysis for comparison to the assembly-level predictions. Similarly, when conducting a MAG-level analysis, it is useful to have access to isolate genomes from the same site, or general reference genomes for comparison.

Our last proposed recommendation is to conduct manual curation of MAGs when possible. While genome-resolved metagenomics has greatly reduced the bottleneck of culture-dependent sequencing, it comes with the disadvantages of incomplete, or incorrectly assembled MAGs. CMAGs, or complete MAGs are incredibly rare. A 2020 paper reported that as of September 10, 2019, there were only 59 CMAGs on GenBank (Chen *et al*., 2020). In comparison, it is not uncommon to see studies reporting thousands of MAGs at time, which may be of high quality, but are not necessarily complete (Parks *et al*., 2017).

Furthermore, quality for most MAGs is calculated with the CheckM bin quality software, which our results have shown to be useful, but not entirely accurate for regions of the MAGs that do not contain core marker genes (Parks *et* *al*., 2015). CheckM statistics are determined based on the number and identity of multi-copy marker genes detected in a MAG (Parks *et al*., 2015). Thus, a *Bifidobacteriaceae* MAG that is 100% complete would imply that it has 100% of the marker genes that one would expect to find in a genome belonging to the family *Bifidobacteriaceae*. If the same MAG is reported to have 0% contamination, that implies that no marker genes outside of the expeced set were found in the MAG. However, the genome could still be contaminated with sequences that do no belong to the set of marker genes from the CheckM database. Alternately, if the MAG has 30% contamination, this would mean that 30% of the single-copy marker genes are present more than once in the MAG, or that 30% of the MAG has marker genes belonging to a different bacterial family (Parks *et al*., 2015). Lastly, if the MAG had 50% strain heterogeneity, that would be imply that half of the contamination can be explained by the presence of closely-related strains in the MAG. Strain heterogeneity is separated from contamination if the detected "contaminant" marker gene has more than 90% amino acid similarity to the expected marker gene (Parks *et al*., 2015). Thus, when evaluating MAGs with CheckM metrics, it is important to note that quality estimates are not based on whole-genome comparisons, but marker gene presence, counts, and amino acid similarity. This may unfortunately lead to erroneous conclusions regarding MAG completeness and contamination. As a solution to this problem, the manual curation of genomes with comparisons to reference genomes may be necessary.

For predictions of gene function, culture-based approaches on strains of interest may also be necessary.

On the MAG-level, manual curation may include discarding misassembled regions and contaminations, filling gaps with un-assembled reads, and matching paired reads to newly assembled regions (Nadalin *et al*., 2012). If gaps are not filled, then the sample may have to be re-assembled with different parameters to generate a new scaffold (Chen *et al*., 2020). While manual curation is an intensive task, it may be necessary for the evaluation of strain-level MAGs when cultured isolates are not available.

## 6.3 The gut microbiome and bacterial metabolism over the period of solid food introduction

In characterizing the gut microbiome of 15 vaginally born, breastfed infants from the Baby, Food & Mi cohort, we found changes in the taxonomic and metabolic profiles of infants before and after the introduction of solid foods. We had initially hypothesized that samples obtained after solid food introduction would display an increased abundance of bacterial strains with the ability to metabolize non-HMO food glycans, and a higher number of carbohydrate degrading genes. We had further expected to see a continued persistence of *Bifidobacterium* after the introduction of solids, with the aim to potentially characterize metabolic pathways that contribute to their persistence.

In analyzing the taxonomy of samples before and after solid food introduction, it was observed that all *Bifidobacterium* decreased in abundance after solid food introduction, except *B. breve*, which increased, and all *Bacteroides* increased in abundance. Prior to solid food introduction, 11 samples had been dominant in *Bifidobacterium*, two dominant in *Bacteroides*, one in *E. coli*, and one in *E. ramosum*. After solid food introduction, 6 samples were dominant in *Bifidobacterium*, 8 in *Bacteroides,* and 1 in *Ruminococcus*.

Both the increase of *Bacteroides* and decrease of *Bifidobacterium* are consistent with the literature. *Bacteroides* belong to the phylum Bacteroidetes, which are known for using thousands of enzyme combinations for glycan degradation (Lapébie *et al*., 2019). Bacteroidetes have clusters of CAZymes called polysaccharide utilization loci, or PULs, with each PUL having the functionality to degrade a specific glycan. PULs will generally have a few GHs, PLs, and *susC/D* gene pair (Lapébie *et al*., 2019). Typically, CAZymes are secreted by Bacteroidetes for glycan binding and partial glycan breakdown (by amylases such as SusG, or GH13), transported to the bacterial periplasm, and subsequently broken down in the periplasm by other enzymes (such as SusB, or GH97) (Koropatkin *et al*., 2012). In comparison, *Bifidobacterium* initially colonize the infant gut due to their HMO-degrading abilities, and are known to persist due to their non-HMO glycan degrading abilities, although the exact reason for their persistence is not known. Certain infant bifidobacterial of *B. longum* and *B. bifidum* that are able to directly degrade HMOs (Koropatkin *et al*., 2012; Turroni *et al*., 2018b) are generally replaced by adult bifidobacterial strains after solid food introduction. For example, *B. bifidum* strains in adults are known for their mucin-degrading abilities (Roy *et al*., 2006; Turroni *et al*., 2018b). The short timeframe in which the taxonomy of the infant gut microbiome changed before and after solid food introduction is because non-HMO degrading bacteria already exist in the gut during exclusive breastfeeding. Certain bacteria like *B. thetaiotaomicron*—which increased after solid food introduction in our samples—are known as “glycan generalists” and harbor the ability to degrade a variety of glycans, allowing them to adapt and survive. *B. thetaiotaomicron* alone contain 260 glycoside hydrolases, contributing to their persistence (Cantarel *et al*., 2012). This suggests that samples after solid food introduction may be characterized by an increased abundance of bacterial strains with the ability to metabolize non-HMO food glycans.

In evaluating read-level metabolic predictions, the top predicted pathway across all samples was pyruvate fermentation to isobutanol (BioCyc ID: PWY-7111), which is an engineered pathway that is not supposed to naturally occur in bacteria, since isobutanol is a fuel source. However, this same pathway has been previously found in other gut microbiome studies, including in the human gut microbiome, mouse maternal microbiome, pig microbiome, and the human salivary microbiome (Deng *et al*., 2021; Haque *et al*., 2021; Huang *et al*., 2021; Yan *et al*., 2021). Species-specific versions of this pathway were also frequently detected, including for *B. dorei, B. fragilis, B. vulgatus, B. bifidum*, and more. Given that fermentation of pyruvate is a common reaction in the infant gut, in addition to the degradation of pyruvate to butyrate (also known as butanoate), it is possible that a different pathway that uses similar genes is being predicted as PWY7111 (Silva *et al*., 2020). If this is the case, it is not yet possible to know whether bacteria predicted to have this pathway actually do, or if uncharacterized bacteria are being erroneously being assigned to taxonomic ranks that they do not belong to.

A major pathway of interest for this study was the bifid-shunt, or the fructose-6-phosphate-shunt pathway, which breaks down glucose and fructose to acetate and lactate (Devika and Raman, 2019). In the read-level analysis of our samples, the bifid-shunt was the most abundant pathway in the *Bifidobacterium* genus, followed by sucrose degradation IV, which is also a part of the bifid-shunt pathway. According to previous research, the presence of HMOs in the environment causes an upregulation of enzymatic activity related to the bifid shunt pathway (Walsh *et al*., 2020). This was consistent in our own data, where the majority of *Bifidobacterium* in the breastfed infants harbored the bifid shunt pathway as their most abundant pathway. *B. breve*, which were the only bifidobacteria without a detectable bifid-shunt pathway, instead had the highest abundance of the glycogen degradation I pathway out of all bifidobacteria. It was also the only bifidobacteria to have the glucose and glucose-1-phosphate degradation pathway. However, given that the bifid-shunt pathway is a core carbohydrate degradation mechanism of *Bifidobacterium*, its absence from the *B. breve* genome is unlikely. One alternative explanation for these results may be that *B. breve* pathways are being mis-characterized as unclassified bacteria.

*Bifidobacterium* also had pathways for L−isoleucine biosynthesis I and L−valine biosynthesis, which has also been previously seen in the literature (Liu *et al*., 2021; Senizza *et al*., 2020; Zhang *et al*., 2020). The continued, but decreased, presence of bifidobacterial species in the infant gut after solid food introduction, implies that bifidobacteria are able to adapt to the changing gut environment.

*Bacteroides*, which increased in abundance and dominance after the introduction of solid foods, had metabolic pathways not seen in *Bifidobacterium*. Such pathways included glycolysis IV and the L−rhamnose degradation pathway, the latter of which is used in the degradation of pectins and hemicelluloses (Rodionova *et al*., 2013). *R. gnavus*, which became the dominant species in one sample after solid food introduction, had degradation pathways for L- rhamnose as well, in addition to the starch degradation V pathway.

Overall, the metabolic profiles of bacteria that increased *(Bacteroides)* or decreased *(Bifidobacterium)* were in line with our initial hypothesis that solid food introduction would bring an increased abundance of bacterial strains with the ability to metabolize non-HMO food glycans.

## 6.4 Limitations and Strengths

One of the greatest strengths of our simulated infant gut metagenomic dataset is that it was simulated using real 16S rRNA gene data from infants, and thus it closely resembled the complexity of real metagenomic samples. Access to the original reference genomes and gold-standard assembles allowed us to use proper positive controls to compare against the results of computational tools. Furthermore, a strength of the analysis is the use of predicted CAZymes as practical and relevant gene predictions. We also constructed a pipeline that will be publicly available to the research community. A strength of the Baby, Food & Mi study is the longitudinal nature of the project, potentially allowing us to sequence and compare long-term changes in the taxonomic and functional profiles after solid food introduction. Access to sequenced cultured isolates from each infant sample broaden the scope of sequencing sample-specific strains of interest in the future. Availability of multiple data types—including 16S rRNA gene profiles, genomes from cultured isolates, and metagenomic samples—allow for further benchmarking and validation of our methods. The limitations of this study include the use of Shannon Diversity to determine whether sample diversity was impacting the accuracy of final assemblies, MAGs, and CAZyme predictions. This diversity metric does not take into account the phylogenetic distance between species and strains, which is an important point to consider when evaluating samples with different levels of strain diversity. Secondly, the samples simulated for the evaluation of computational tools were 3 million reads each. This is relatively low compared to our own metagenomic samples, which are usually upwards of 10 million reads each. Evaluating larger datasets would provide a more accurate substitution for real gut metagenomic samples. Thirdly, only *Bifidobacterium* MAGs and genomes were used for strain-level CAZyme and MAG quality analyses in the simulated dataset. As such, there are limits to how much we can generalize these strain-level results to other genera. For example, some *Bacteroides* species often have over 200 CAZymes in their genomes, while *Bifidobacterium* are relatively less complex in their CAZyme profiles (Cantarel *et al*., 2012).

## 6.5 Future Directions

All Baby, Food & Mi samples have been processed through the bfm_mg_flow pipeline, and the first step after this thesis is to conduct assembly-level and MAG-level analyses, including the manual curation of MAGs. This may allow for CAZyme-level comparison of individual strains over the introduction of solid food. Once strains of interest have been identified and analyzed computationally, these strains can also be isolated from infant samples to study their metabolic activities in culture.

Given the increase in abundance of *Bacteroides*, a potential avenue for further research is to predict PULs in *Bacteroides* MAGs, using the dbCAN-PUL database (Ausland *et al*., 2021). For *Bifidobacterium*, MAGs can be used to analyze species-level differences in the bifido-shunt pathway, in addition to unique pathways in *B. breve* that might have contributed to its increased abundance after solid food introduction.

Finally, a major future step would be to compare sample-specific MAGs from this study to fiber-utilization profiles of sample-specific strains in culture. Furthermore, dietary data from infants can be compared to the subsequent changes in CAZyme predictions in the gut microbiome.

## 6.6 Significance

In recent years, many studies have used metagenome-assembled genomes to study bacterial metabolism in complex communities with high strain diversity, such as the infant gut microbiome. However, the caveat of *de novo* analysis of highly variable microbial communities without access to isolate or reference genomes is that we are often unaware of the extent to which our final MAGs and gene predictions are accurate (Sczyrba *et al*., 2017). Here, we have reported the scenarios in which one can expect a generic metagenomic workflow to fail, along with scenarios where it is less likely to fail. We have provided a realistic outlook of the quality of final MAGs and strain-level CAZyme predictions one can expect to receive when using a general metagenomic analysis workflow for infant gut data. We have also provided recommendations for creating an appropriate metagenomic analysis workflow based on the community composition of samples, based on the research questions surrounding the samples.

# 7 Conclusions

Dominated by abundant genera with high levels of strain diversity, the infant gut poses a unique challenge to metagenomic data analysis, especially during the highly dynamic timepoint of solid food introduction. To effectively characterize the changes in gut bacterial carbohydrate-active enzymes over time, it is essential to have access to high quality metagenome-assembled genomes. In this study, we used a simulated infant gut metagenomic dataset to report the shortcomings of current metagenomic computational tools. Our findings showed that both microbial diversity and choice of software impact the number and quality of predicted MAGs and genes. Based on these results, a robust bioinformatics pipeline was tailored to the specific gut microbial community composition of the breastfed, vaginally born infants from the Baby, Food & Mi cohort. Using this pipeline on infant gut samples before and after solid food introduction, it was reported that the infant gut microbiome after solid food introduction showcased an increased abundance of bacterial strains with the ability to metabolize non-human milk oligosaccharide dietary carbohydrates. These results show that solid food introduction changes the taxonomy and abundance of bacteria in the gut, ultimately selecting for bacteria that are able to digest new dietary food glycans. However, further studies are required to characterize the specific changes that occur in bacterial genomes as a result of the selective pressure induced by solid food introduction. CAZyme annotation of complete MAGs from infant samples throughout exclusive breastfeeding, introduction of solid foods, weaning, and exclusive solid food consumption would highlight the characteristics of bacteria that are dominant in the gut at each stage. This would contribute to our understanding of bacterial colonization and succession, and ultimately how the composition of the infant gut impacts health in later life.

This study provided a framework for studying longitudinal infant gut metagenomic samples, in addition to providing new knowledge regarding the characteristics of gut bacterial composition and metabolism over the period of solid food introduction.

# Works Cited

Alessandri, G., Ossiprandi, M.C., MacSharry, J., van Sinderen, D., and Ventura, M. (2019). Bifidobacterial Dialogue With Its Human Host and Consequent Modulation of the Immune System. Front. Immunol. *10*, 2348.

Almeida, O.G.G., and De Martinis, E.C.P. (2019). Bioinformatics tools to assess metagenomic data for applied microbiology. Appl. Microbiol. Biotechnol. *103*, 69–82.

Ames, S.K., Hysom, D.A., Gardner, S.N., Lloyd, G.S., Gokhale, M.B., and Allen, J.E. (2013). Scalable metagenomic taxonomy classification using a reference genome database. Bioinformatics *29*, 2253–2260.

Andrews, S. (2010). FastQC - A quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Arboleya, S., Watkins, C., Stanton, C., and Ross, R.P. (2016a). Gut Bifidobacteria Populations in Human Health and Aging. Front. Microbiol. *7*, 1204.

Arboleya, S., Watkins, C., Stanton, C., and Ross, R.P. (2016b). Gut Bifidobacteria Populations in Human Health and Aging. Front. Microbiol. *7*, 1204.

Ausland, C., Zheng, J., Yi, H., Yang, B., Li, T., Feng, X., Zheng, B., and Yin, Y. (2021). dbCAN-PUL: a database of experimentally characterized CAZyme gene clusters and their substrates. Nucleic Acids Res. *49*, D523–D528.

Avershina, E., Lundgå Rd, K., Sekelja, M., Dotterud, C., Storrø, O., Øien, T., Johnsen, R., and Rudi, K. Transition from infant-to adult-like gut microbiota.

Bäckhed, F., Roswall, J., Peng, Y., Feng, Q., Jia, H., Kovatcheva-Datchary, P., Li, Y., Xia, Y., Xie, H., Zhong, H., et al. (2015). Dynamics and Stabilization of the Human Gut Microbiome during the First Year of Life. Cell Host Microbe *17*, 690–703.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2014). Fitting Linear Mixed-Effects Models using lme4.

Becraft, E.D., Woyke, T., Jarett, J., Ivanova, N., Godoy-Vitorino, F., Poulton, N., Brown, J.M., Brown, J., Lau, M.C.Y., Onstott, T., et al. (2017). Rokubacteria: Genomic Giants among the Uncultured Bacterial Phyla. Front. Microbiol. *8*, 2264.

Beghini, F., McIver, L.J., Blanco-Míguez, A., Dubois, L., Asnicar, F., Maharjan, S., Mailyan, A., Manghi, P., Scholz, M., Thomas, A.M., et al. (2021). Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. Elife *10*.

Beier, S., Tappu, R., and Huson, D.H. (2017). Functional Analysis in Metagenomics Using MEGAN 6. In Functional Metagenomics: Tools and Applications, (Cham: Springer International Publishing), pp. 65–74.

Bellahcene, M., O'Dowd, J.F., Wargent, E.T., Zaibi, M.S., Hislop, D.C., Ngala, R.A., Smith, D.M., Cawthorne, M.A., Stocker, C.J., and Arch, J.R.S. (2013).

Male mice that lack the G-protein-coupled receptor GPR41 have low energy expenditure and increased body fat content. Br. J. Nutr. *109*, 1755–1764.

Berry, K.J., and Mielke, P.W. (2000). A Monte Carlo investigation of the Fisher Z transformation for normal and nonnormal distributions. Psychol. Rep. *87*, 1101–1114.

Bhattacharya, T., Ghosh, T.S., and Mande, S.S. (2015). Global Profiling of Carbohydrate Active Enzymes in Human Gut Microbiome. PLoS One *10*.

Bishara, A.J., and Hittner, J.B. (2017). Confidence intervals for correlations when data are not normal. Behav. Res. Methods *49*, 294–309.

Bokulich, N.A., Ziemski, M., Robeson, M.S., and Kaehler, B.D. (2020). Measuring the microbiome: Best practices for developing and benchmarking microbiomics methods. Comput. Struct. Biotechnol. J. *18*, 4048–4062.

Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics.

Boraston, A.B., Bolam, D.N., Gilbert, H.J., and Davies, G.J. (2004). Carbohydrate-binding modules: fine-tuning polysaccharide recognition. Biochem. J. *382*, 769–781.

Bottacini, F., Ventura, M., Sinderen, D., and Motherway, M. (2014). Diversity, ecology and intestinal function of bifidobacteria. Microb. Cell Fact. *13*, S4.

Bourne, Y., and Henrissat, B. (2001). Glycoside hydrolases and glycosyltransferases: families and functional modules. Curr. Opin. Struct. Biol. *11*, 593–600.

Breitwieser, F.P., Lu, J., and Salzberg, S.L. (2017). A review of methods and databases for metagenomic classification and assembly. Brief. Bioinform.

Van Den Broek, L.A.M., and Voragen, A.G.J. (2008). Bifidobacterium glycoside hydrolases and (potential) prebiotics. Innov. Food Sci. Emerg. Technol. *9*, 401–407.

Brown, C.T., Howe, A., Zhang, Q., Pyrkosz, A.B., and Brom, T.H. (2012). A Reference-Free Algorithm for Computational Normalization of Shotgun Sequencing Data.

Brown, C.T., Sharon, I., Thomas, B.C., Castelle, C.J., Morowitz, M.J., and Banfield, J.F. (2013). Genome resolved analysis of a premature infant gut microbial community reveals a Varibaculum cambriense genome and a shift towards fermentation-based metabolism during the third week of life. Microbiome *1*, 30.

Browne, H.P., Forster, S.C., Anonye, B.O., Kumar, N., Neville, B.A., Stares, M.D., Goulding, D., and Lawley, T.D. (2016). Culturing of ‘unculturable’ human microbiota reveals novel taxa and extensive sporulation. Nature *533*, 543–546.

Buchfink, B., Xie, C., and Huson, D.H. (2015). Fast and sensitive protein alignment using DIAMOND. Nat. Methods *12*, 59–60.

Buddingh, G.J. (1975). Bergey’s Manual of Determinative Bacteriology. Am. J. Trop. Med. Hyg.

Bushnell, B. (2014). BBMap: a fast, accurate, splice-aware aligner.

Caballero-Franco, C., Keller, K., De Simone, C., and Chadee, K. (2007). The VSL#3 probiotic formula induces mucin gene expression and secretion in colonic epithelial cells. Am. J. Physiol. Liver Physiol. *292*, G315–G322.

Callahan, B.J., McMurdie, P.J., Rosen, M.J., Han, A.W., Johnson, A.J.A., and Holmes, S.P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. Nat. Methods *13*, 581–583.

Cantarel, B.L., Coutinho, P.M., Rancurel, C., Bernard, T., Lombard, V., and Henrissat, B. (2009). The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. Nucleic Acids Res. *37*, D233–D238.

Cantarel, B.L., Lombard, V., and Henrissat, B. (2012). Complex Carbohydrate Utilization by the Healthy Human Microbiome. PLoS One *7*, e28742.

Caspi, R., Billington, R., Keseler, I.M., Kothari, A., Krummenacker, M., Midford, P.E., Ong, W.K., Paley, S., Subhraveti, P., and Karp, P.D. (2020). The MetaCyc database of metabolic pathways and enzymes - a 2019 update. Nucleic Acids Res. *48*, D445–D453.

Chakravorty, S., Helb, D., Burday, M., Connell, N., and Alland, D. (2007). A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. J. Microbiol. Methods *69*, 330–339.

Chen, L.-X., Anantharaman, K., Shaiber, A., Eren, A.M., and Banfield, J.F. (2020). Accurate and complete genomes from metagenomes. Genome Res. *30*, 315–333.

Coyne, M.J., Kalka-Moll, W., Tzianabos, A.O., Kasper, D.L., and Comstock, L.E. (2000). Bacteroides fragilis NCTC9343 produces at least three distinct capsular polysaccharides: cloning, characterization, and reassignment of polysaccharide B and C biosynthesis loci. Infect. Immun. *68*, 6176–6181.

Daims, H., Lebedeva, E. V., Pjevac, P., Han, P., Herbold, C., Albertsen, M., Jehmlich, N., Palatinszky, M., Vierheilig, J., Bulaev, A., et al. (2015). Complete nitrification by Nitrospira bacteria. Nature *528*, 504–509.

David, L.A., Maurice, C.F., Carmody, R.N., Gootenberg, D.B., Button, J.E., Wolfe, B.E., Ling, A. V., Devlin, A.S., Varma, Y., Fischbach, M.A., et al. (2014). Diet rapidly and reproducibly alters the human gut microbiome. Nature *505*, 559–563.

Davis, E.C., Dinsmoor, A.M., Wang, M., and Donovan, S.M. (2020). Microbiome Composition in Pediatric Populations from Birth to Adolescence: Impact of Diet and Prebiotic and Probiotic Interventions. Dig. Dis. Sci. *65*, 706–722.

Delcaru, C., Alexandru, I., Podgoreanu, P., Cristea, V.C., Bleotu, C., Chifiriuc, M.C., Bezirtzoglou, E., and Lazar, V. (2016). Antagonistic activities of some Bifidobacterium sp. strains isolated from resident infant gastrointestinal microbiota on Gram-negative enteric pathogens. Anaerobe *39*, 39–44.

Deng, F., Li, Y., Peng, Y., Wei, X., Wang, X., Howe, S., Yang, H., Xiao, Y., Li, H., Zhao, J., et al. (2021). The Diversity, Composition, and Metabolic Pathways of Archaea in Pigs. Animals *11*, 2139.

Devika, N.T., and Raman, K. (2019). Deciphering the metabolic capabilities of Bifidobacteria using genome-scale metabolic models. Sci. Rep. *9*, 18222.

Dixon, P. (2003). VEGAN, a package of R functions for community ecology. J. Veg. Sci. *14*, 927–930.

Dizzell, S., Stearns, J.C., Li, J., van Best, N., Bervoets, L., Mommers, M., Penders, J., Morrison, K.M., Hutton, E.K., and Partners, on behalf of the G.-M.C. (2021). Investigating colonization patterns of the infant gut microbiome during the introduction of solid food and weaning from breastmilk: A cohort study protocol. PLoS One *16*, e0248924.

Duar, R.M., Kyle, D., and Casaburi, G. (2020a). Colonization Resistance in the Infant Gut: The Role of B. infantis in Reducing pH and Preventing Pathogen Growth. High-Throughput *9*, 7.

Duar, R.M., Henrick, B.M., Casaburi, G., and Frese, S.A. (2020b). Integrating the Ecosystem Services Framework to Define Dysbiosis of the Breastfed Infant Gut: The Role of B. infantis and Human Milk Oligosaccharides. Front. Nutr. *7*, 33.

Duranti, S., Milani, C., Lugli, G.A., Mancabelli, L., Turroni, F., Ferrario, C., Mangifesta, M., Viappiani, A., Sánchez, B., Margolles, A., et al. (2016). Evaluation of genetic diversity among strains of the human gut commensal Bifidobacterium adolescentis. Sci. Rep. *6*, 23971.

Duranti, S., Ferrario, C., van Sinderen, D., Ventura, M., and Turroni, F. (2017a). Obesity and microbiota: an example of an intricate relationship. Genes Nutr. *12*, 18.

Duranti, S., Mangifesta, M., Lugli, G.A., Turroni, F., Anzalone, R., Milani, C., Mancabelli, L., Ossiprandi, M.C., and Ventura, M. (2017b). Bifidobacterium vansinderenii sp. nov., isolated from faeces of emperor tamarin (Saguinus imperator). Int. J. Syst. Evol. Microbiol. *67*, 3987–3995.

Duranti, S., Lugli, G.A., Viappiani, A., Mancabelli, L., Alessandri, G., Anzalone, R., Longhi, G., Milani, C., Ossiprandi, M.C., Turroni, F., et al. (2020). Characterization of the phylogenetic diversity of two novel species belonging to the genus Bifidobacterium: Bifidobacterium cebidarum sp. nov. and Bifidobacterium leontopitheci sp. nov. Int. J. Syst. Evol. Microbiol. *70*, 2288–2297.

Dwek, R.A. (1996). Glycobiology: Toward understanding the function of sugars. Chem. Rev. *96*.

Eren, A.M., Esen, Ö.C., Quince, C., Vineis, J.H., Morrison, H.G., Sogin, M.L., and Delmont, T.O. (2015). Anvi’o: an advanced analysis and visualization platform for ‘omics data. PeerJ *3*, e1319.

Favier, C.F., Vaughan, E.E., De Vos, W.M., and Akkermans, A.D.L. (2002). Molecular monitoring of succession of bacterial communities in human neonates. Appl. Environ. Microbiol. *68*, 219–226.

Ferretti, P., Pasolli, E., Tett, A., Asnicar, F., Gorfer, V., Fedi, S., Armanini, F., Truong, D.T., Manara, S., Zolfo, M., et al. (2018). Mother-to-Infant Microbial Transmission from Different Body Sites Shapes the Developing Infant Gut Microbiome. Cell Host Microbe *24*, 133-145.e5.

Fritz, A., Hofmann, P., Majda, S., Dahms, E., Dröge, J., Fiedler, J., Lesker, T.R., Belmann, P., DeMaere, M.Z., Darling, A.E., et al. (2019). CAMISIM: simulating metagenomes and microbial communities. Microbiome *7*, 17.

Fukuda, S., Toh, H., Taylor, T.D., Ohno, H., and Hattori, M. (2012). Acetate-producing bifidobacteria protect the host from enteropathogenic infection via carbohydrate transporters. Gut Microbes *3*, 449–454.

Fushinobu, S. (2010). Unique Sugar Metabolic Pathways of Bifidobacteria. Biosci. Biotechnol. Biochem. *74*, 2374–2384.

Garcia, S.N., Osburn, B.I., and Cullor, J.S. (2019). A one health perspective on dairy production and dairy food safety. One Heal. *7*, 100086.

German, J.B., Freeman, S.L., Lebrilla, C.B., and Mills, D.A. (2008). Human Milk Oligosaccharides: Evolution, Structures and Bioselectivity as Substrates for Intestinal Bacteria. In Personalized Nutrition for the Diverse Needs of Infants and Children, (Basel: KARGER), pp. 205–222.

Ghurye, J.S., Cepeda-Espinoza, V., and Pop, M. (2016). Metagenomic Assembly: Overview, Challenges and Applications. Yale J. Biol. Med. *89*, 353–362.

Gilbert, J.A., Blaser, M.J., Caporaso, J.G., Jansson, J.K., Lynch, S. V, and Knight, R. (2018). Current understanding of the human microbiome. Nat. Med. *24*, 392–400.

Gloster, T.M. (2014). Advances in understanding glycosyltransferases from a structural perspective. Curr. Opin. Struct. Biol. *28*, 131–141.

Gnoth, M.J., Kunz, C., Kinne-Saffran, E., and Rudloff, S. (2000). Human Milk Oligosaccharides Are Minimally Digested In Vitro. J. Nutr. *130*, 3014–3020.

Gotoh, A., Ojima, M.N., and Katayama, T. (2019). Minority species influences microbiota formation: the role of Bifidobacterium with extracellular glycosidases in bifidus flora formation in breastfed infant guts. Microb. Biotechnol. *12*, 259.

Guillén, D., Sánchez, S., and Rodríguez-Sanoja, R. (2010). Carbohydrate-binding domains: multiplicity of biological roles. Appl. Microbiol. Biotechnol. *85*, 1241–1249.

Haque, M., Koski, K.G., and Scott, M.E. (2021). A gastrointestinal nematode in pregnant and lactating mice alters maternal and neonatal microbiomes. Int. J. Parasitol.

Hofmeyr, S., Egan, R., Georganas, E., Copeland, A.C., Riley, R., Clum, A., Eloe-Fadrosh, E., Roux, S., Goltsman, E., Buluç, A., et al. (2020). Terabase-scale metagenome coassembly with MetaHipMer. Sci. Rep. *10*, 10689.

Homann, C.-M., Rossel, C.A.J., Dizzell, S., Bervoets, L., Simioni, J., Li, J., Gunn, E., Surette, M.G., de Souza, R.J., Mommers, M., et al. (2021). Infants' First Solid Foods: Impact on Gut Microbiota Development in Two Intercontinental Cohorts. Nutrients *13*, 2639.

Huang, K., Gao, X., Wu, L., Yan, B., Wang, Z., Zhang, X., Peng, L., Yu, J., Sun, G., and Yang, Y. (2021). Salivary Microbiota for Gastric Cancer Prediction: An Exploratory Study. Front. Cell. Infect. Microbiol. *11*, 640309.

Huang, W., Li, L., Myers, J.R., and Marth, G.T. (2012). ART: a next-generation sequencing read simulator. Bioinformatics *28*, 593–594.

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

Hyatt, D., Chen, G.-L., LoCascio, P.F., Land, M.L., Larimer, F.W., and Hauser, L.J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics *11*, 119.

Illumina (2021). Illumina sequencing platforms.

Ioannou, A., Knol, J., and Belzer, C. (2021). Microbial Glycoside Hydrolases in the First Year of Life: An Analysis Review on Their Presence and Importance in Infant Gut. Front. Microbiol. *12*.

Kang, D.D., Froula, J., Egan, R., and Wang, Z. (2015). MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. PeerJ *3*, e1165.

Kans, J. (2010). Entrez Direct: E-utilities on the UNIX Command Line.

Kaoutari, A. El, Armougom, F., Gordon, J.I., Raoult, D., and Henrissat, B. (2013). The abundance and variety of carbohydrate-active enzymes in the human gut microbiota. Nat. Rev. Microbiol. *11*, 497–504.

Kato, K., Odamaki, T., Mitsuyama, E., Sugahara, H., Xiao, J., and Osawa, R. (2017). Age-Related Changes in the Composition of Gut Bifidobacterium Species. Curr. Microbiol. *74*, 987.

Katoh, T., Ojima, M.N., Sakanaka, M., Ashida, H., Gotoh, A., and Katayama, T. (2020). Enzymatic Adaptation of Bifidobacterium bifidum to Host Glycans, Viewed from Glycoside Hydrolyases and Carbohydrate-Binding Modules. Microorganisms *8*, 481.

Koenig, J.E., Spor, A., Scalfone, N., Fricker, A.D., Stombaugh, J., Knight, R., Angenent, L.T., and Ley, R.E. (2011). Succession of microbial consortia in the developing infant gut microbiome. Proc. Natl. Acad. Sci.

Koropatkin, N.M., Cameron, E.A., and Martens, E.C. (2012). How glycan metabolism shapes the human gut microbiota. Nat. Rev. Microbiol. *10*, 323–335.

Koster, J., and Rahmann, S. (2012). Snakemake--a scalable bioinformatics workflow engine. Bioinformatics *28*, 2520–2522.

Kowarsky, M., Camunas-Soler, J., Kertesz, M., De Vlaminck, I., Koh, W., Pan, W., Martin, L., Neff, N.F., Okamoto, J., Wong, R.J., et al. (2017). Numerous uncharacterized and highly divergent microbes which colonize humans are revealed by circulating cell-free DNA. Proc. Natl. Acad. Sci. U. S. A. *114*, 9623–9628.

Krueger, F. (2015). Trim Galore!: A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files.

Kumar, V. (2010). Analysis of the key active subsites of glycoside hydrolase 13 family members. Carbohydr. Res. *345*, 893–898.

Laforest-Lapointe, I., and Arrieta, M.-C. (2017). Patterns of Early-Life Gut Microbial Colonization during Human Immune Development: An Ecological Perspective. Front. Immunol.

Lairson, L.L., Henrissat, B., Davies, G.J., and Withers, S.G. (2008).

Glycosyltransferases: Structures, Functions, and Mechanisms. Annu. Rev. Biochem. *77*, 521–555.

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

Lapébie, P., Lombard, V., Drula, E., Terrapon, N., and Henrissat, B. (2019). Bacteroidetes use thousands of enzyme combinations to break down glycans. Nat. Commun. *10*, 2043.

Lau, J.T., Whelan, F.J., Herath, I., Lee, C.H., Collins, S.M., Bercik, P., and Surette, M.G. (2016). Capturing the diversity of the human gut microbiota through culture-enriched molecular profiling. Genome Med. *8*, 72.

Laureys, D., Cnockaert, M., De Vuyst, L., and Vandamme, P. (2016). Bifidobacterium aquikefiri sp. nov., isolated from water kefir. Int. J. Syst. Evol. Microbiol. *66*, 1281–1286.

Lengeler, J. (1975). Mutations affecting transport of the hexitols D-mannitol, D-glucitol, and galactitol in Escherichia coli K-12: isolation and mapping. J. Bacteriol. *124*, 26–38.

Levasseur, A., Drula, E., Lombard, V., Coutinho, P.M., and Henrissat, B. (2013). Expansion of the enzymatic repertoire of the CAZy database to integrate auxiliary redox enzymes. Biotechnol. Biofuels *6*, 41.

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.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Subgroup, 1000 Genome Project Data Processing (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

Lin, H. V., Frassetto, A., Kowalik Jr, E.J., Nawrocki, A.R., Lu, M.M., Kosinski, J.R., Hubert, J.A., Szeto, D., Yao, X., Forrest, G., et al. (2012). Butyrate and Propionate Protect against Diet-Induced Obesity and Regulate Gut Hormones via Free Fatty Acid Receptor 3-Independent Mechanisms. PLoS One *7*, e35240.

Liu, Z., Li, L., Fang, Z., Lee, Y., Zhao, J., Zhang, H., Chen, W., Li, H., and Lu, W. (2021). Integration of Transcriptome and Metabolome Reveals the Genes and Metabolites Involved in Bifidobacterium bifidum Biofilm Formation. Int. J. Mol. Sci. *22*, 7596.

Lombard, V., Bernard, T., Rancurel, C., Brumer, H., Coutinho, P.M., and Henrissat, B. (2010). A hierarchical classification of polysaccharide lyases for glycogenomics. Biochem. J. *432*, 437–444.

Lu, Y.Y., Chen, T., Fuhrman, J.A., Sun, F., and Sahinalp, C. (2017). COCACOLA: Binning metagenomic contigs using sequence COmposition, read CoverAge, CO-alignment and paired-end read LinkAge. Bioinformatics *33*.

Lugli, G.A., Mangifesta, M., Duranti, S., Anzalone, R., Milani, C., Mancabelli, L., Alessandri, G., Turroni, F., Ossiprandi, M.C., van Sinderen, D., et al. (2018). Phylogenetic classification of six novel species belonging to the genus

Bifidobacterium comprising Bifidobacterium anseris sp. nov., Bifidobacterium criceti sp. nov., Bifidobacterium imperatoris sp. nov., Bifidobacterium italicum sp. nov., Bifidobacterium margollesii sp. nov. and Bifidobacterium parmae sp. nov. Syst. Appl. Microbiol. *41*, 173–183.

Lugli, G.A., Milani, C., Mancabelli, L., Turroni, F., Sinderen, D., and Ventura, M. (2019a). A microbiome reality check: limitations of *in silico* -based metagenomic approaches to study complex bacterial communities. Environ. Microbiol. Rep. *11*, 1758-2229.12805.

Lugli, G.A., Milani, C., Duranti, S., Alessandri, G., Turroni, F., Mancabelli, L., Tatoni, D., Ossiprandi, M.C., van Sinderen, D., and Ventura, M. (2019b). Isolation of novel gut bifidobacteria using a combination of metagenomic and cultivation approaches. Genome Biol. *20*, 96.

Madden, T. (2013). The BLAST Sequence Analysis Tool.

Maguire, F., Jia, B., Gray, K.L., Lau, W.Y.V., Beiko, R.G., and Brinkman, F.S.L. (2020). Metagenome-assembled genome binning methods with short reads disproportionately fail for plasmids and genomic Islands. Microb. Genomics *6*, e000436.

Marbouty, M., and Koszul, R. (2015). Metagenome Analysis Exploiting High-Throughput Chromosome Conformation Capture (3C) Data. Trends Genet. *31*, 673–682.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.Journal *17*, 10–12.

Martín, V., Maldonado-Barragán, A., Moles, L., Rodriguez-Baños, M., Campo, R. del, Fernández, L., Rodríguez, J.M., and Jiménez, E. (2012). Sharing of Bacterial Strains Between Breast Milk and Infant Feces. J. Hum. Lact. *28*, 36–44.

McDonald, D., Price, M.N., Goodrich, J., Nawrocki, E.P., DeSantis, T.Z., Probst, A., Andersen, G.L., Knight, R., and Hugenholtz, P. (2012). An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. ISME J. *6*, 610–618.

McMurdie, P.J., and Holmes, S. (2013). phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. PLoS One *8*, e61217.

Menzel, P., Ng, K.L., and Krogh, A. (2016). Fast and sensitive taxonomic classification for metagenomics with Kaiju. Nat. Commun. *7*, 11257.

Meyer, F., Bagchi, S., Chaterji, S., Gerlach, W., Grama, A., Harrison, T., Paczian, T., Trimble, W.L., and Wilke, A. (2019). MG-RAST version 4—lessons learned from a decade of low-budget ultra-high-throughput metagenome analysis. Brief. Bioinform. *20*, 1151–1159.

Meyer, F., Lesker, T.-R., Koslicki, D., Fritz, A., Gurevich, A., Darling, A.E., Sczyrba, A., Bremges, A., and McHardy, A.C. (2021). Tutorial: assessing metagenomics software with the CAMI benchmarking toolkit. Nat. Protoc. *16*, 1785–1801.

Michelini, S., Modesto, M., Pisi, A.M., Filippini, G., Sandri, C., Spiezio, C., Biavati, B., Sgorbati, B., and Mattarelli, P. (2016a). Bifidobacterium eulemuris sp. nov., isolated from faeces of black lemurs (Eulemur macaco). Int. J. Syst. Evol. Microbiol. *66*, 1567–1576.

Michelini, S., Oki, K., Yanokura, E., Shimakawa, Y., Modesto, M., Mattarelli, P., Biavati, B., and Watanabe, K. (2016b). Bifidobacterium myosotis sp. nov., Bifidobacterium tissieri sp. nov. and Bifidobacterium hapali sp. nov., isolated from faeces of baby common marmosets (Callithrix jacchus L.). Int. J. Syst. Evol. Microbiol. *66*, 255–265.

Michelini, S., Modesto, M., Filippini, G., Spiezio, C., Sandri, C., Biavati, B., Pisi, A., and Mattarelli, P. (2018). Corrigendum to "Bifidobacterium aerophilum sp. nov., Bifidobacterium avesanii sp. nov. and Bifidobacterium ramosum sp. nov.: Three novel taxa from the faeces of cotton-top tamarin (Saguinus oedipus L.)" [Syst. Appl. Microbiol. 39 (2016) 229–236]. Syst. Appl. Microbiol. *41*, 528.

Mikheenko, A., Saveliev, V., and Gurevich, A. (2016). MetaQUAST: evaluation of metagenome assemblies. Bioinformatics *32*, 1088–1090.

Milani, C., Lugli, G.A., Duranti, S., Turroni, F., Bottacini, F., Mangifesta, M., Sanchez, B., Viappiani, A., Mancabelli, L., Taminiau, B., et al. (2014). Genomic encyclopedia of type strains of the genus Bifidobacterium. Appl. Environ. Microbiol. *80*, 6290–6302.

Milani, C., Lugli, G.A., Duranti, S., Turroni, F., Mancabelli, L., Ferrario, C., Mangifesta, M., Hevia, A., Viappiani, A., Scholz, M., et al. (2015). Bifidobacteria exhibit social behavior through carbohydrate resource sharing in the gut. Sci. Rep. *5*, 15782.

Milani, C., Duranti, S., Bottacini, F., Casey, E., Turroni, F., Mahony, J., Belzer, C., Delgado Palacio, S., Arboleya Montes, S., Mancabelli, L., et al. (2017a). The First Microbial Colonizers of the Human Gut: Composition, Activities, and Health Implications of the Infant Gut Microbiota. Microbiol. Mol. Biol. Rev. *81*, e00036-17.

Milani, C., Mangifesta, M., Mancabelli, L., Lugli, G.A., James, K., Duranti, S., Turroni, F., Ferrario, C., Ossiprandi, M.C., van Sinderen, D., et al. (2017b). Unveiling bifidobacterial biogeography across the mammalian branch of the tree of life. ISME J. *11*, 2834–2847.

Miossec, M.J., Valenzuela, S.L., Pérez-Losada, M., Johnson, W.E., Crandall, K.A., and Castro-Nallar, E. (2020). Evaluation of computational methods for human microbiome analysis using simulated data. PeerJ *8*, e9688.

Modesto, M., Michelini, S., Oki, K., Biavati, B., Watanabe, K., and Mattarelli, P. (2018a). Bifidobacterium catulorum sp. nov., a novel taxon from the faeces of the baby common marmoset (Callithrix jacchus). Int. J. Syst. Evol. Microbiol. *68*, 575–581.

Modesto, M., Michelini, S., Sansosti, M.C., De Filippo, C., Cavalieri, D., Qvirist, L., Andlid, T., Spiezio, C., Sandri, C., Pascarelli, S., et al. (2018b). Bifidobacterium callitrichidarum sp. nov. from the faeces of the emperor tamarin (Saguinus imperator). Int. J. Syst. Evol. Microbiol. *68*, 141–148.

Nadalin, F., Vezzi, F., and Policriti, A. (2012). GapFiller: a de novo assembly approach to fill the gap within paired reads. BMC Bioinformatics *13 Suppl 14*, S8.

Nakamura, A.M., Nascimento, A.S., and Polikarpov, I. (2017). Structural diversity of carbohydrate esterases. Biotechnol. Res. Innov. *1*, 35–51.

Naumoff, D.G. (2011). Hierarchical classification of glycoside hydrolases. Biochem. *76*, 622–635.

Nurk, S., Meleshko, D., Korobeynikov, A., and Pevzner, P.A. (2017). metaSPAdes: a new versatile metagenomic assembler. Genome Res. *27*, 824–834.

O'Callaghan, A., and van Sinderen, D. (2016). Bifidobacteria and Their Role as Members of the Human Gut Microbiota. Front. Microbiol. *7*, 925.

Odamaki, T., Kato, K., Sugahara, H., Hashikura, N., Takahashi, S., Xiao, J., Abe, F., and Osawa, R. (2016). Age-related changes in gut microbiota composition from newborn to centenarian: a cross-sectional study. BMC Microbiol. *16*, 90.

Ogurtsova, K., da Rocha Fernandes, J.D., Huang, Y., Linnenkamp, U., Guariguata, L., Cho, N.H., Cavan, D., Shaw, J.E., and Makaroff, L.E. (2017). IDF Diabetes Atlas: Global estimates for the prevalence of diabetes for 2015 and 2040. Diabetes Res. Clin. Pract.

Parks, D.H., Imelfort, M., Skennerton, C.T., Hugenholtz, P., and Tyson, G.W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. *25*, 1043–1055.

Parks, D.H., Rinke, C., Chuvochina, M., Chaumeil, P.-A., Woodcroft, B.J., Evans, P.N., Hugenholtz, P., and Tyson, G.W. (2017). Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nat. Microbiol. *2*, 1533–1542.

Pasolli, E., Asnicar, F., Manara, S., Zolfo, M., Karcher, N., Armanini, F., Beghini, F., Manghi, P., Tett, A., Ghensi, P., et al. (2019). Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. Cell *176*, 649-662.e20.

Pechar, R., Killer, J., Salmonová, H., Geigerová, M., Švejstil, R., Švec, P., Sedláček, I., Rada, V., and Benada, O. (2017). Bifidobacterium apri sp. nov., a thermophilic actinobacterium isolated from the digestive tract of wild pigs (Sus scrofa). Int. J. Syst. Evol. Microbiol. *67*, 2349–2356.

Prakash, T., and Taylor, T.D. (2012). Functional assignment of metagenomic data: challenges and applications. Brief. Bioinform. *13*, 711–727.

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K.S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., et al. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. Nature *464*, 59–65.

Quince, C., Walker, A.W., Simpson, J.T., Loman, N.J., and Segata, N. (2017). Shotgun metagenomics, from sampling to analysis. Nat. Biotechnol. *35*, 833–844.

Ramakrishna, B.S. (2013). Role of the gut microbiota in human nutrition and metabolism. J. Gastroenterol. Hepatol. *28*, 9–17.

Rodionova, I.A., Li, X., Thiel, V., Stolyar, S., Stanton, K., Fredrickson, J.K., Bryant, D.A., Osterman, A.L., Best, A.A., and Rodionov, D.A. (2013). Comparative genomics and functional analysis of rhamnose catabolic pathways and regulons in bacteria. Front. Microbiol. *4*, 407.

Rodriguez, C.I., and Martiny, J.B.H. (2020). Evolutionary relationships among bifidobacteria and their hosts and environments. BMC Genomics *21*, 26.

La Rosa, P.S., Warner, B.B., Zhou, Y., Weinstock, G.M., Sodergren, E., Hall-Moore, C.M., Stevens, H.J., Bennett, W.E., Shaikh, N., Linneman, L.A., et al. (2014). Patterned progression of bacterial populations in the premature infant gut. Proc. Natl. Acad. Sci. U. S. A. *111*, 12522–12527.

Van Rossum, T., Ferretti, P., Maistrenko, O.M., and Bork, P. (2020). Diversity within species: interpreting strains in microbiomes. Nat. Rev. Microbiol. *18*, 491–506.

Roumpeka, D.D., Wallace, R.J., Escalettes, F., Fotheringham, I., and Watson, M. (2017). A Review of Bioinformatics Tools for Bio-Prospecting from Metagenomic Sequence Data. Front. Genet. *8*, 23.

Roy, C.C., Kien, C.L., Bouthillier, L., and Levy, E. (2006). Short-Chain Fatty Acids: Ready for Prime Time? Nutr. Clin. Pract. *21*, 351–366.

Schmid, J., Heider, D., Wendel, N.J., Sperl, N., and Sieber, V. (2016). Bacterial Glycosyltransferases: Challenges and Opportunities of a Highly Diverse Enzyme Class Toward Tailoring Natural Products. Front. Microbiol. *7*, 182.

Scott, K.P., Gratz, S.W., Sheridan, P.O., Flint, H.J., and Duncan, S.H. (2013). The influence of diet on the gut microbiota. Pharmacol. Res. *69*, 52–60.

Sczyrba, A., Hofmann, P., Belmann, P., Koslicki, D., Janssen, S., Dröge, J., Gregor, I., Majda, S., Fiedler, J., Dahms, E., et al. (2017). Critical Assessment of Metagenome Interpretation—a benchmark of metagenomics software. Nat. Methods *14*, 1063–1071.

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. Bioinformatics *30*, 2068–2069.

Sela, D.A., Chapman, J., Adeuya, A., Kim, J.H., Chen, F., Whitehead, T.R., Lapidus, A., Rokhsar, D.S., Lebrilla, C.B., German, J.B., et al. (2008). The genome sequence of Bifidobacterium longum subsp. infantis reveals adaptations for milk utilization within the infant microbiome. Proc. Natl. Acad. Sci. U. S. A.

Senizza, A., Rocchetti, G., Callegari, M.L., Lucini, L., and Morelli, L. (2020). Linoleic acid induces metabolic stress in the intestinal microorganism Bifidobacterium breve DSM 20213. Sci. Rep. *10*, 5997.

Sharon, I., Morowitz, M.J., Thomas, B.C., Costello, E.K., Relman, D.A., and Banfield, J.F. Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization.

Sieber, C.M.K., Probst, A.J., Sharrar, A., Thomas, B.C., Hess, M., Tringe, S.G., and Banfield, J.F. (2018). Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. Nat. Microbiol. *3*, 836–843.

Silva, Y.P., Bernardi, A., and Frozza, R.L. (2020). The Role of Short-Chain Fatty

Acids From Gut Microbiota in Gut-Brain Communication. Front. Endocrinol. (Lausanne). *11*, 25.

Smith, P.M., Howitt, M.R., Panikov, N., Michaud, M., Gallini, C.A., Bohlooly-Y, M., Glickman, J.N., and Garrett, W.S. (2013). The Microbial Metabolites, Short-Chain Fatty Acids, Regulate Colonic Treg Cell Homeostasis. Science (80-. ). *341*, 569–573.

Sprockett, D., Fukami, T., and Relman, D.A. (2018). Role of priority effects in the early-life assembly of the gut microbiota. Nat. Rev. Gastroenterol. Hepatol. *15*, 197–205.

Stearns, J.C., Davidson, C.J., McKeon, S., Whelan, F.J., Fontes, M.E., Schryvers, A.B., Bowdish, D.M.E., Kellner, J.D., and Surette, M.G. (2015). Culture and molecular-based profiles show shifts in bacterial communities of the upper respiratory tract that occur with age. ISME J. *9*, 1246–1259.

Stearns, J.C., Simioni, J., Gunn, E., McDonald, H., Holloway, A.C., Thabane, L., Mousseau, A., Schertzer, J.D., Ratcliffe, E.M., Rossi, L., et al. (2017). Intrapartum antibiotics for GBS prophylaxis alter colonization patterns in the early infant gut microbiome of low risk infants. Sci. Rep. *7*, 16527.

Sun, Z., Huang, S., Zhang, M., Zhu, Q., Haiminen, N., Carrieri, A.P., Vázquez-Baeza, Y., Parida, L., Kim, H.-C., Knight, R., et al. (2021). Challenges in benchmarking metagenomic profilers. Nat. Methods *18*, 618–626.

Sunagawa, S., Mende, D.R., Zeller, G., Izquierdo-Carrasco, F., Berger, S.A., Kultima, J.R., Coelho, L.P., Arumugam, M., Tap, J., Nielsen, H.B., et al. (2013). Metagenomic species profiling using universal phylogenetic marker genes. Nat. Methods *10*, 1196–1199.

Toniolo, A., Cassani, G., Puggioni, A., Rossi, A., Colombo, A., Onodera, T., and Ferrannini, E. (2019). The diabetes pandemic and associated infections: suggestions for clinical microbiology. Rev. Med. Microbiol. *30*, 1–17.

Tremaroli, V., and Bäckhed, F. (2012). Functional interactions between the gut microbiota and host metabolism. Nature *489*, 242–249.

Truong, D.T., Franzosa, E.A., Tickle, T.L., Scholz, M., Weingart, G., Pasolli, E., Tett, A., Huttenhower, C., and Segata, N. (2015). MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nat. Methods *12*, 902–903.

Truong, D.T., Tett, A., Pasolli, E., Huttenhower, C., and Segata, N. (2017). Microbial strain-level population structure and genetic diversity from metagenomes. Genome Res. *27*, 626–638.

Turnbaugh, P.J., Ley, R.E., Mahowald, M.A., Magrini, V., Mardis, E.R., and Gordon, J.I. (2006). An obesity-associated gut microbiome with increased capacity for energy harvest. Nature *444*, 1027–1031.

Turroni, F., Taverniti, V., Ruas-Madiedo, P., Duranti, S., Guglielmetti, S., Lugli, G.A., Gioiosa, L., Palanza, P., Margolles, A., van Sinderen, D., et al. (2014a). Bifidobacterium bifidum PRL2010 modulates the host innate immune response. Appl. Environ. Microbiol. *80*, 730–740.

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

Turroni, F., Milani, C., Duranti, S., Ferrario, C., Lugli, G.A., Mancabelli, L., van Sinderen, D., and Ventura, M. (2018a). Bifidobacteria and the infant gut: an example of co-evolution and natural selection. Cell. Mol. Life Sci. *75*, 103–118.

Turroni, F., Milani, C., Duranti, S., Mahony, J., van Sinderen, D., and Ventura, M. (2018b). Glycan Utilization and Cross-Feeding Activities by Bifidobacteria. Trends Microbiol. *26*, 339–350.

Turroni, F., Duranti, S., Milani, C., Lugli, G.A., van Sinderen, D., and Ventura, M. (2019). Bifidobacterium bifidum: A Key Member of the Early Human Gut Microbiota. Microorganisms *7*, 544.

Turroni, F., Milani, C., Duranti, S., Lugli, G.A., Bernasconi, S., Margolles, A., Di Pierro, F., van Sinderen, D., and Ventura, M. (2020). The infant gut microbiome as a microbial organ influencing host well-being. Ital. J. Pediatr. *46*, 16.

Tyson, G.W., Chapman, J., Hugenholtz, P., Allen, E.E., Ram, R.J., Richardson, P.M., Solovyev, V. V., Rubin, E.M., Rokhsar, D.S., and Banfield, J.F. (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature *428*, 37–43.

Underwood, M.A., German, J.B., Lebrilla, C.B., and Mills, D.A. (2015). Bifidobacterium longum subspecies infantis: champion colonizer of the infant gut. Pediatr. Res. *77*, 229–235.

Uritskiy, G. V., DiRuggiero, J., and Taylor, J. (2018). MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis. Microbiome *6*, 158.

Varki, A. (2017). Biological roles of glycans. Glycobiology *27*.

Vazquez-Gutierrez, P., de Wouters, T., Werder, J., Chassard, C., and Lacroix, C. (2016). High Iron-Sequestrating Bifidobacteria Inhibit Enteropathogen Growth and Adhesion to Intestinal Epithelial Cells In vitro. Front. Microbiol. *7*, 1480.

Vellend, M. (2010). Conceptual Synthesis in Community Ecology. Q. Rev. Biol. *85*, 183–206.

De Vuyst, L., Moens, F., Selak, M., Rivière, A., and Leroy, F. (2014). Summer Meeting 2013: growth and physiology of bifidobacteria. J. Appl. Microbiol. *116*, 477–491.

Walker, A.W., Duncan, S.H., Louis, P., and Flint, H.J. (2014). Phylogeny, culturing, and metagenomics of the human gut microbiota. Trends Microbiol. *22*, 267–274.

Walsh, C., Lane, J.A., van Sinderen, D., and Hickey, R.M. (2020). Human milk oligosaccharides: Shaping the infant gut microbiota and supporting health. J. Funct. Foods *72*, 104074.

Wang, Q., Garrity, G.M., Tiedje, J.M., and Cole, J.R. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl. Environ. Microbiol. *73*, 5261–5267.

Wood, D.E., and Salzberg, S.L. (2014). Kraken: Ultrafast metagenomic sequence classification using exact alignments. Genome Biol.

Wu, Y.-W., Simmons, B.A., and Singer, S.W. (2016). MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. Bioinformatics *32*, 605–607.

Xie, F., Jin, W., Si, H., Yuan, Y., Tao, Y., Liu, J., Wang, X., Yang, C., Li, Q., Yan, X., et al. (2021). An integrated gene catalog and over 10,000 metagenome-assembled genomes from the gastrointestinal microbiome of ruminants. Microbiome *9*, 137.

Yan, S., Ma, Z., Jiao, M., Wang, Y., Li, A., and Ding, S. (2021). Effects of Smoking on Inflammatory Markers in a Healthy Population as Analyzed via the Gut Microbiota. Front. Cell. Infect. Microbiol. *11*, 646.

Yatsunenko, T., Rey, F.E., Manary, M.J., Trehan, I., Dominguez-Bello, M.G., Contreras, M., Magris, M., Hidalgo, G., Baldassano, R.N., Anokhin, A.P., et al. (2012). Human gut microbiome viewed across age and geography. Nature *486*, 222–227.

Ye, J., McGinnis, S., and Madden, T.L. (2006). BLAST: improvements for better sequence analysis. Nucleic Acids Res. *34*, W6–W9.

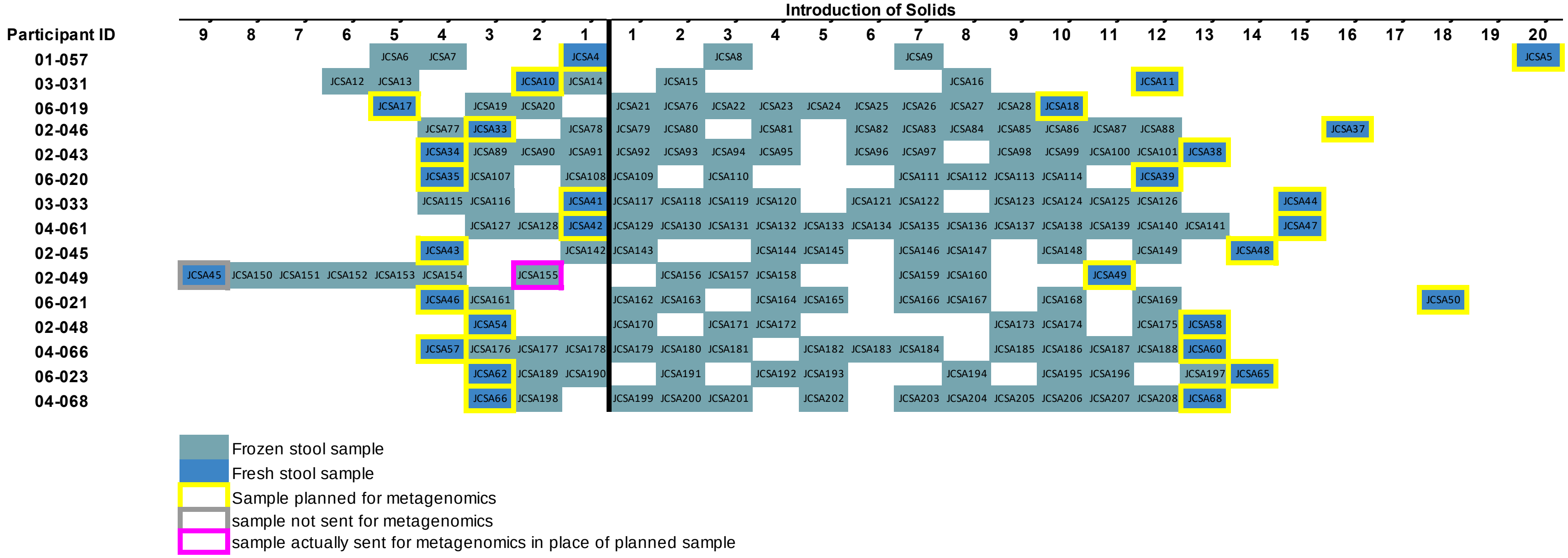
Yin, Y., Mao, X., Yang, J., Chen, X., Mao, F., and Xu, Y. (2012). dbCAN: a web resource for automated carbohydrate-active enzyme annotation. Nucleic Acids Res. *40*, W445–W451.

Yue, Y., Huang, H., Qi, Z., Dou, H.-M., Liu, X.-Y., Han, T.-F., Chen, Y., Song, X.-J., Zhang, Y.-H., and Tu, J. (2020). Evaluating metagenomics tools for genome binning with real metagenomic datasets and CAMI datasets. BMC Bioinformatics *21*, 334.

Zhang, R., Zhang, J., Dang, W., Irwin, D.M., Wang, Z., and Zhang, S. (2020). Unveiling the Biogeography and Potential Functions of the Intestinal Digesta- and Mucosa-Associated Microbiome of Donkeys. Front. Microbiol. *11*, 3051.

Zhao, S., Lieberman, T.D., Poyet, M., Kauffman, K.M., Gibbons, S.M., Groussin, M., Xavier, R.J., and Alm, E.J. (2019). Adaptive Evolution within Gut Microbiomes of Healthy People. Cell Host Microbe *25*, 656-667.e8.

# Appendix to Methodology



Introduction of Solids

| Participant ID | 9 | 8 | 7 | 6 | 5 | 4 | 3 | 2 | 1 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 01-057 | | | | | JCSA6 | JCSA7 | | | JCSA4 | | | JCSA8 | | | | JCSA9 | | | | | | | | | | | | | JCSA5 |
| 03-031 | | | | JCSA12 | JCSA13 | | | JCSA10 | JCSA14 | | JCSA15 | | | | | | JCSA16 | | | | JCSA11 | | | | | | | | |
| 06-019 | | | | | JCSA17 | | JCSA19 | JCSA20 | | JCSA21 | JCSA76 | JCSA22 | JCSA23 | JCSA24 | JCSA25 | JCSA26 | JCSA27 | JCSA28 | JCSA18 | | | | | | | | | | |
| 02-046 | | | | | | JCSA77 | JCSA33 | | JCSA78 | JCSA79 | JCSA80 | | JCSA81 | | JCSA82 | JCSA83 | JCSA84 | JCSA85 | JCSA86 | JCSA87 | JCSA88 | | | | JCSA37 | | | | |
| 02-043 | | | | | | JCSA34 | JCSA89 | JCSA90 | JCSA91 | JCSA92 | JCSA93 | JCSA94 | JCSA95 | | JCSA96 | JCSA97 | | JCSA98 | JCSA99 | JCSA100 | JCSA101 | JCSA38 | | | | | | | |
| 06-020 | | | | | | JCSA35 | JCSA107 | | JCSA108 | JCSA109 | | JCSA110 | | | | JCSA111 | JCSA112 | JCSA113 | JCSA114 | | JCSA39 | | | | | | | | |
| 03-033 | | | | | | JCSA115 | JCSA116 | | JCSA41 | JCSA117 | JCSA118 | JCSA119 | JCSA120 | | JCSA121 | JCSA122 | | JCSA123 | JCSA124 | JCSA125 | JCSA126 | | | JCSA44 | | | | | |
| 04-061 | | | | | | | JCSA127 | JCSA128 | JCSA42 | JCSA129 | JCSA130 | JCSA131 | JCSA132 | JCSA133 | JCSA134 | JCSA135 | JCSA136 | JCSA137 | JCSA138 | JCSA139 | JCSA140 | JCSA141 | | JCSA47 | | | | | |
| 02-045 | | | | | | JCSA43 | | | JCSA142 | JCSA143 | | | JCSA144 | JCSA145 | | JCSA146 | JCSA147 | | JCSA148 | | JCSA149 | | JCSA48 | | | | | | |
| 02-049 | JCSA45 | JCSA150 | JCSA151 | JCSA152 | JCSA153 | JCSA154 | | JCSA155 | | | JCSA156 | JCSA157 | JCSA158 | | | JCSA159 | JCSA160 | | | JCSA49 | | | | | | | | | |
| 06-021 | | | | | | JCSA46 | JCSA161 | | | JCSA162 | JCSA163 | | JCSA164 | JCSA165 | | JCSA166 | JCSA167 | | JCSA168 | | JCSA169 | | | | | | JCSA50 | | |
| 02-048 | | | | | | | JCSA54 | | | JCSA170 | | JCSA171 | JCSA172 | | | | | JCSA173 | JCSA174 | | JCSA175 | JCSA58 | | | | | | | |
| 04-066 | | | | | | JCSA57 | JCSA176 | JCSA177 | JCSA178 | JCSA179 | JCSA180 | JCSA181 | | JCSA182 | JCSA183 | JCSA184 | | JCSA185 | JCSA186 | JCSA187 | JCSA188 | JCSA60 | | | | | | | |
| 06-023 | | | | | | | JCSA62 | JCSA189 | JCSA190 | | JCSA191 | | JCSA192 | JCSA193 | | | JCSA194 | | JCSA195 | JCSA196 | | JCSA197 | JCSA65 | | | | | | |
| 04-068 | | | | | | | JCSA66 | JCSA198 | | JCSA199 | JCSA200 | JCSA201 | | JCSA202 | | JCSA203 | JCSA204 | JCSA205 | JCSA206 | JCSA207 | JCSA208 | JCSA68 | | | | | | | |

▨ Frozen stool sample
■ Fresh stool sample
□ Sample planned for metagenomics
□ sample not sent for metagenomics
□ sample actually sent for metagenomics in place of planned sample

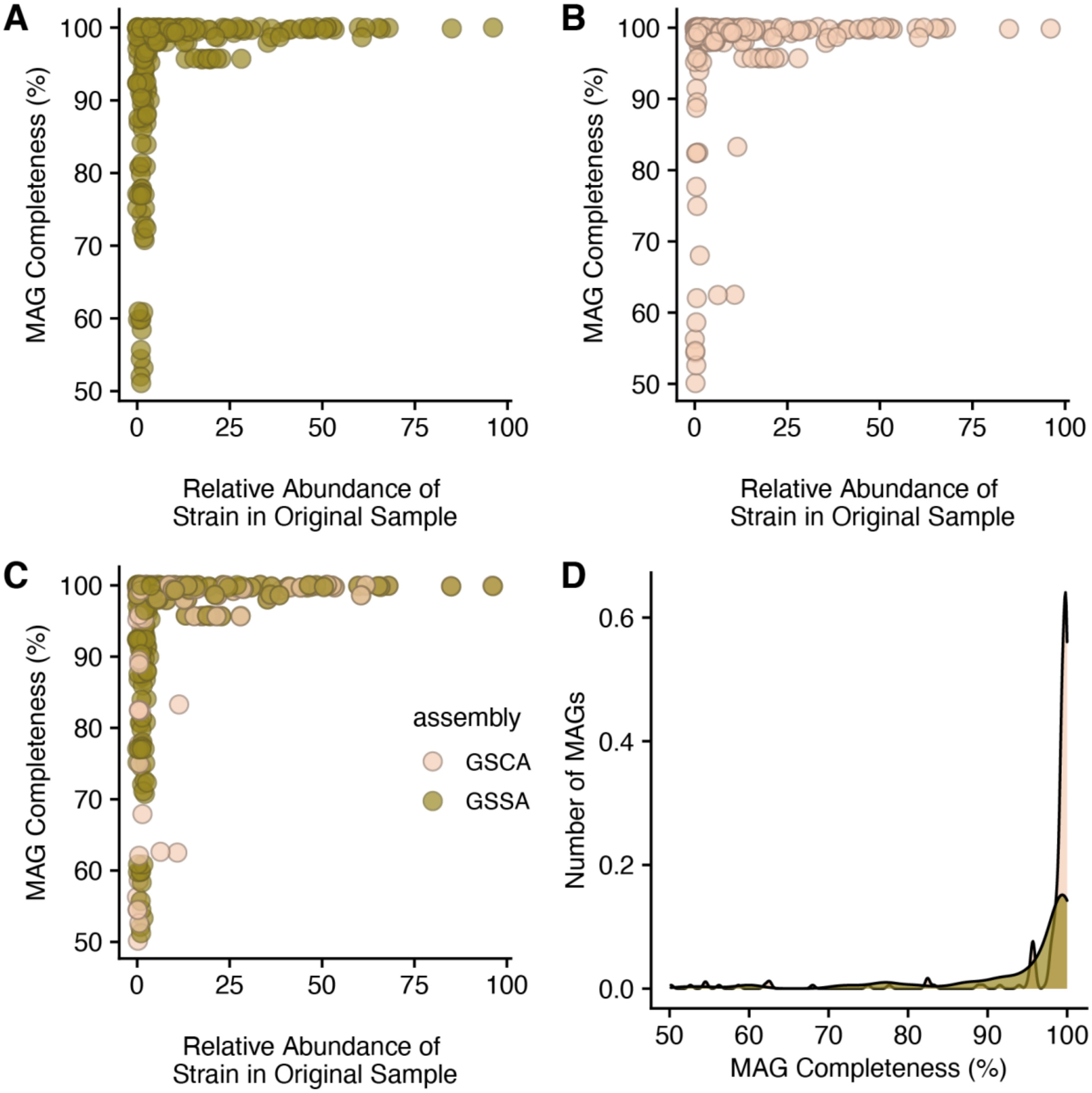
**Supplementary Figure 1: Baby, Food & Mi log of sample availability and documentation of samples sent for shotgun metagenomic sequencing**

# Appendix to Chapter 1

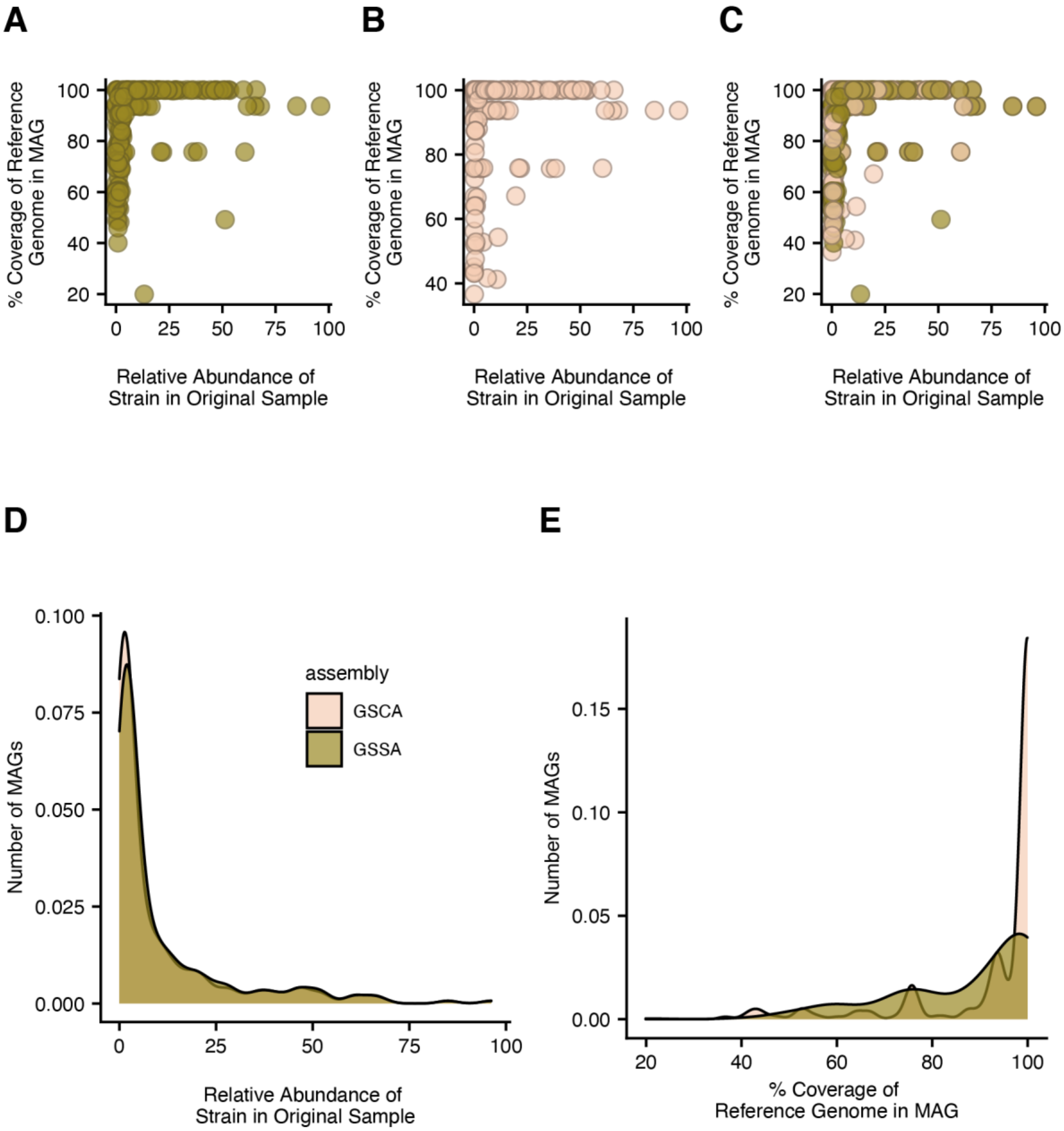
**Supplementary Table 1: The number of total species, total species, and Shannon Diversity of each simulated sample, compared to the number of *Bifidobacerium* strains, species, and bifidobacterial within-genus diversiy in each sample**

| | | **Total** | | | ***Bifidobacterium*** | | |
|---|---|---|---|---|---|---|---|
| | | **Number of Species** | **Number of Strains** | **Shannon Diversity** | **Relative Abundance** | **Number of Species** | **Number of Strains** |
| *Bifidobacterium* dominant | sample_22 | 20 | 23 | 1.100 | 0.959 | 2 | 4 |
| | sample_14 | 26 | 31 | 0.929 | 0.950 | 5 | 6 |
| | sample_19 | 19 | 22 | 1.298 | 0.931 | 2 | 4 |
| | sample_25 | 39 | 60 | 1.753 | 0.884 | 6 | 12 |
| | sample_5 | 51 | 69 | 1.433 | 0.881 | 5 | 7 |
| | sample_32 | 39 | 46 | 1.151 | 0.848 | 5 | 7 |
| | sample_21 | 21 | 24 | 1.515 | 0.821 | 3 | 4 |
| | sample_0 | 47 | 57 | 1.271 | 0.809 | 5 | 7 |
| | sample_16 | 32 | 42 | 1.458 | 0.787 | 3 | 5 |
| | sample_20 | 26 | 31 | 1.423 | 0.784 | 2 | 3 |
| | sample_17 | 34 | 50 | 1.994 | 0.764 | 4 | 6 |
| | sample_13 | 35 | 43 | 1.991 | 0.713 | 4 | 5 |
| | sample_26 | 42 | 65 | 1.874 | 0.687 | 3 | 8 |
| | sample_28 | 44 | 64 | 2.060 | 0.686 | 4 | 9 |
| | sample_3 | 43 | 59 | 2.141 | 0.685 | 4 | 8 |

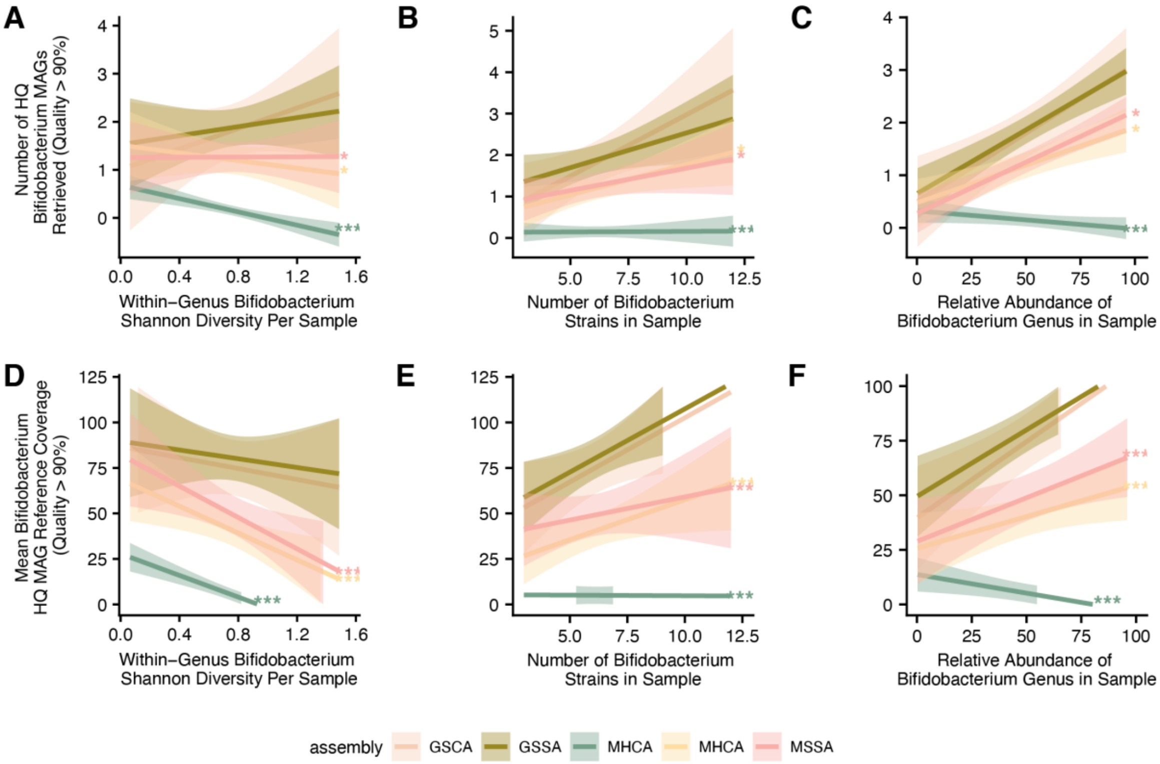
| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | sample_30 | 60 | 73 | 1.942 | 0.638 | 4 | 7 |
| | sample_15 | 35 | 51 | 1.887 | 0.624 | 4 | 6 |
| | sample_24 | 44 | 68 | 2.596 | 0.614 | 4 | 9 |
| | sample_27 | 45 | 66 | 2.150 | 0.610 | 5 | 12 |
| | sample_29 | 52 | 73 | 2.326 | 0.538 | 6 | 10 |
| | sample_18 | 38 | 63 | 2.762 | 0.493 | 5 | 7 |
| Not dominated by *Bifidobacterium* | sample_31 | 52 | 72 | 2.128 | 0.318 | 3 | 6 |
| | sample_4 | 38 | 59 | 2.886 | 0.307 | 3 | 6 |
| | sample_23 | 22 | 29 | 1.091 | 0.294 | 3 | 7 |
| | sample_33 | 44 | 60 | 1.353 | 0.244 | 3 | 6 |
| | sample_8 | 43 | 56 | 2.223 | 0.138 | 3 | 5 |
| | sample_6 | 45 | 59 | 1.505 | 0.125 | 4 | 7 |
| | sample_9 | 35 | 45 | 2.162 | 0.088 | 4 | 5 |
| | sample_7 | 49 | 62 | 2.471 | 0.036 | 5 | 7 |
| | sample_1 | 63 | 83 | 2.786 | 0.022 | 3 | 3 |
| | sample_11 | 22 | 29 | 0.914 | 0.007 | 3 | 4 |
| | sample_2 | 22 | 26 | 0.242 | 0.005 | 3 | 3 |
| | sample_10 | 35 | 51 | 1.828 | 0.004 | 3 | 4 |
| | sample_12 | 25 | 32 | 0.550 | 0.003 | 3 | 6 |



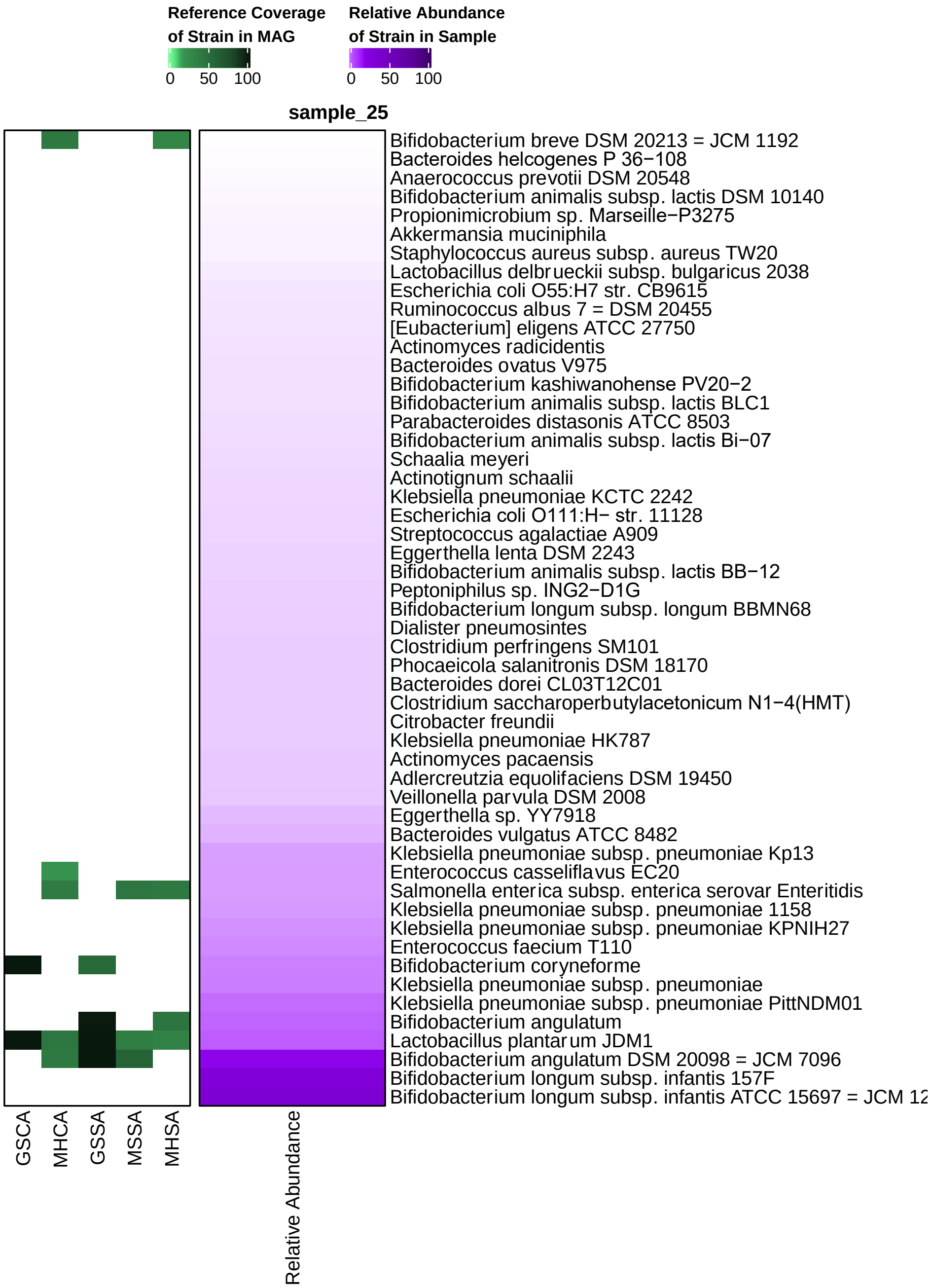
**Supplementary Figure 2: Distribution of MAG completeness as a function of the relative abundance of original strains. A.** MAGs from GSSA**, B.** MAGs from GSCA**, C.** MAGs from GSSA and GSCA, plotted against % completeness for each MAG. **E.** Density plot of the number of MAGs yielded by each assembly by MAG completeness.



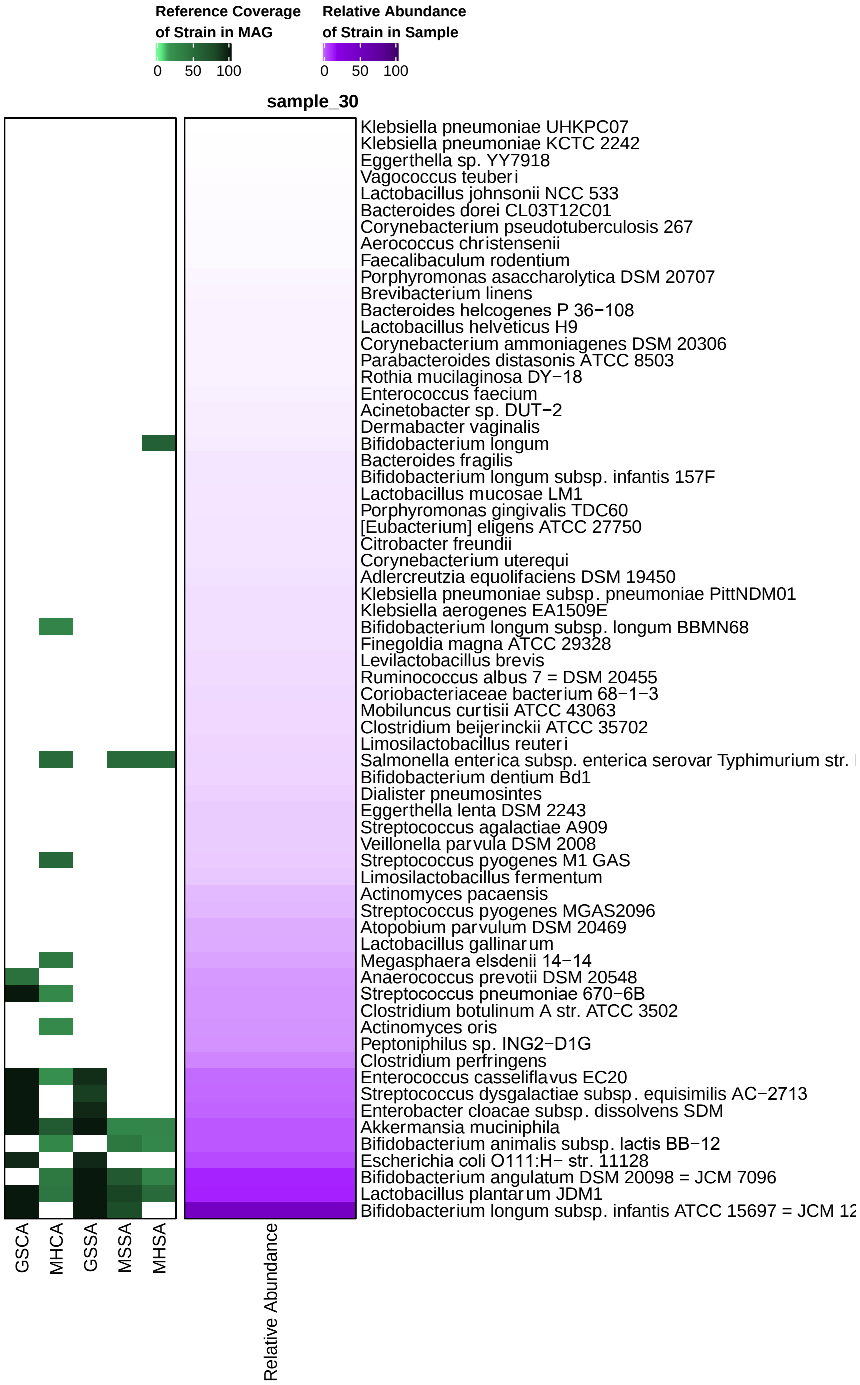
**Supplementary Figure 3: Distribution of MAG reference coverage as a function of the relative abundance of original strains.** Distribution of the reference coverages of MAGs from **A.** just GSSA**, B.** GSCA**, C.** both GSSA and GSCA plotted against the relative abundance of the original strains reconstructed within the MAGs. **E.** Density plot of the number of MAGs yielded by GSSA and GSCA by relative abundance. **F.** Density plot of the number of MAGs yielded by the two assemblies by their reference coverage



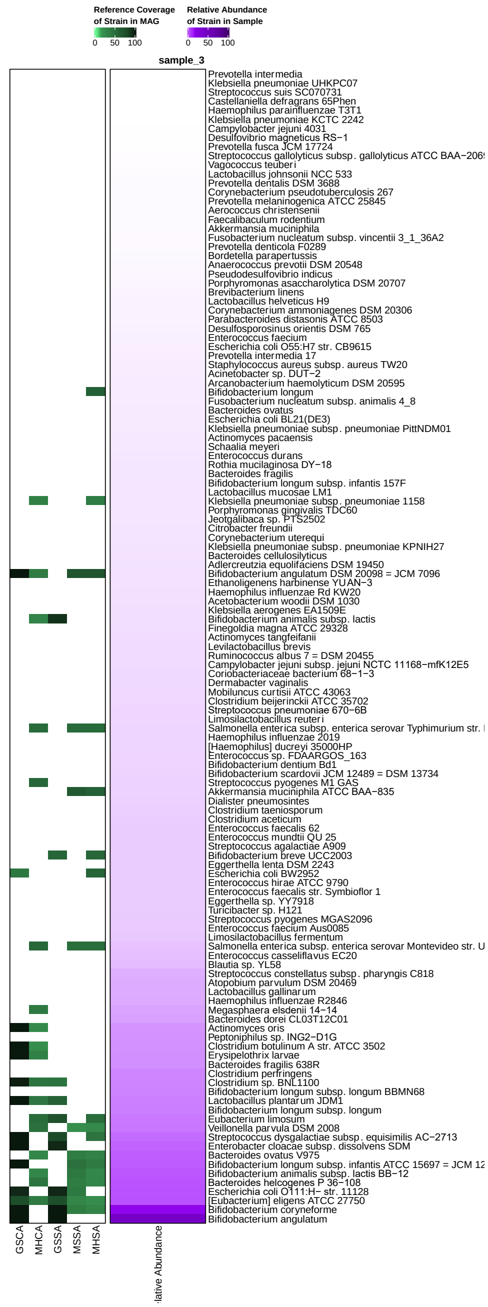
**Supplementary Figure 4: The effect of within-genus *Bifidobacterium* Shannon Diversity, observed species count, and relative abundance on the number and reference coverage of high quality (> 90% quality) *Bifidobacterium* MAGs**. A**.** Higher within-genus Shannon Diversity was negatively correlated with the number of high-quality MAGs obtained from MHSA, MHCA, and MSSA ($p < 0.05$). **B.** Higher *Bifidobacterium* observed species count was positively correlated with the number of high-quality MAGs obtained from MHSA, MHCA, and MSSA ($p < 0.05$). **C.** Higher relative abundance of the *Bifidobacterium* genus was positively correlated with the number of high-quality MAGs obtained from MHSA, MHCA, and MSSA ($p < 0.05$). **D.** Higher within-genus Shannon Diversity was negatively correlated with the reference coverage of MAGs obtained from MHSA, MHCA, and MSSA ($p < 0.05$)**. E.** Higher *Bifidobacterium* observed species was positively correlated with the reference coverage of MAGs obtained from MHSA, MHCA, and MSSA ($p < 0.05$). **F.** Higher relative abundance of the *Bifidobacterium* genus was positively correlated with the reference coverage of MAGs obtained from MHSA, MHCA, and MSSA ($p < 0.05$).



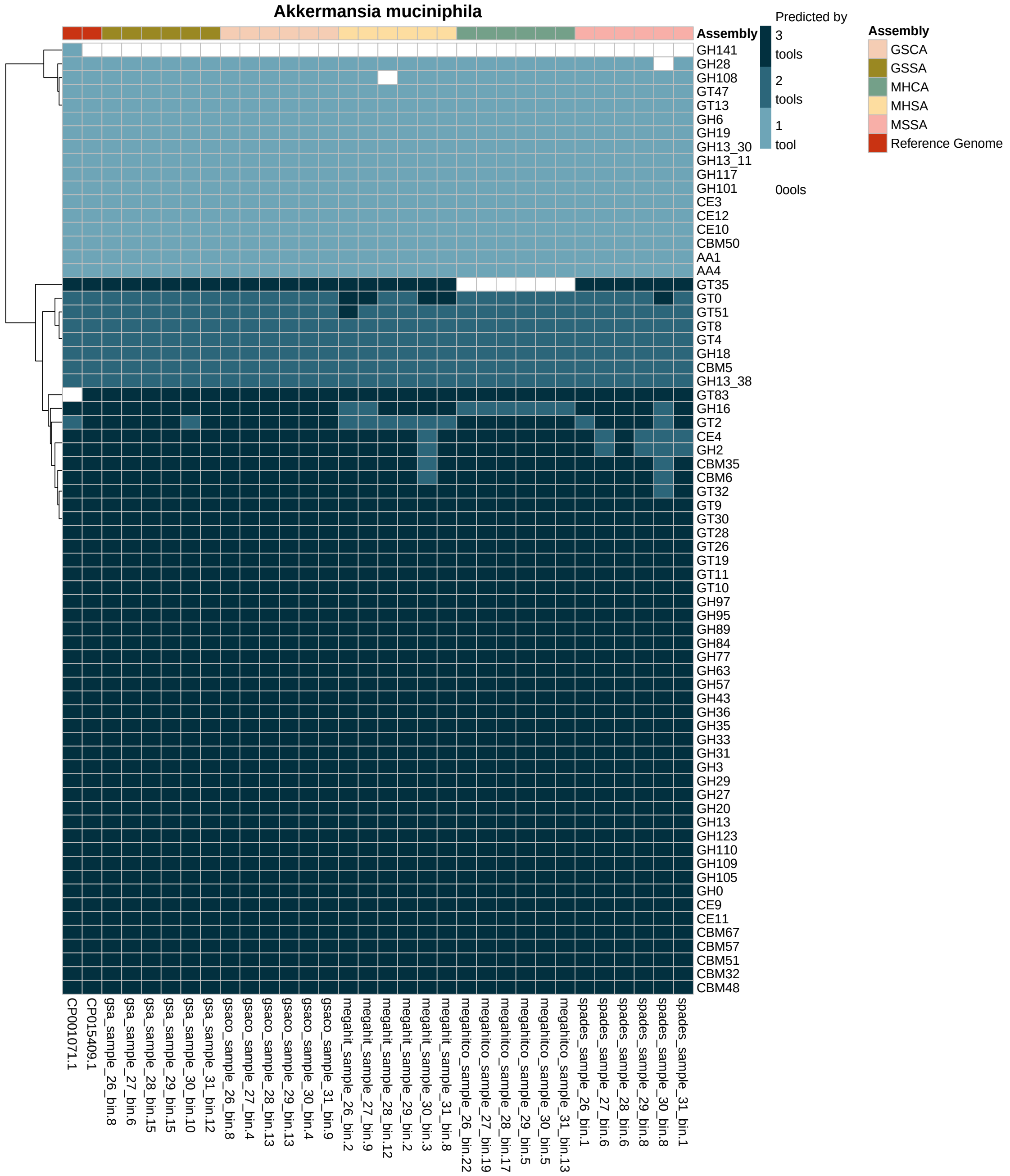
**Supplementary Figure 5: Sample 25 MAG recovery profile, with relative abundances of all strains in the sample and reference coverages of all retrieved MAGs.** The right-most panel represents a list of all strains in the sample. The middle heatmap panel represents the relative abundance of samples. Each cell in the left-most panel represents whether a MAG was recovered for a particular strain by a particular assembly. Darker shades of purple and green represent relative abundance and reference coverage respectively, between 0% and 100%.



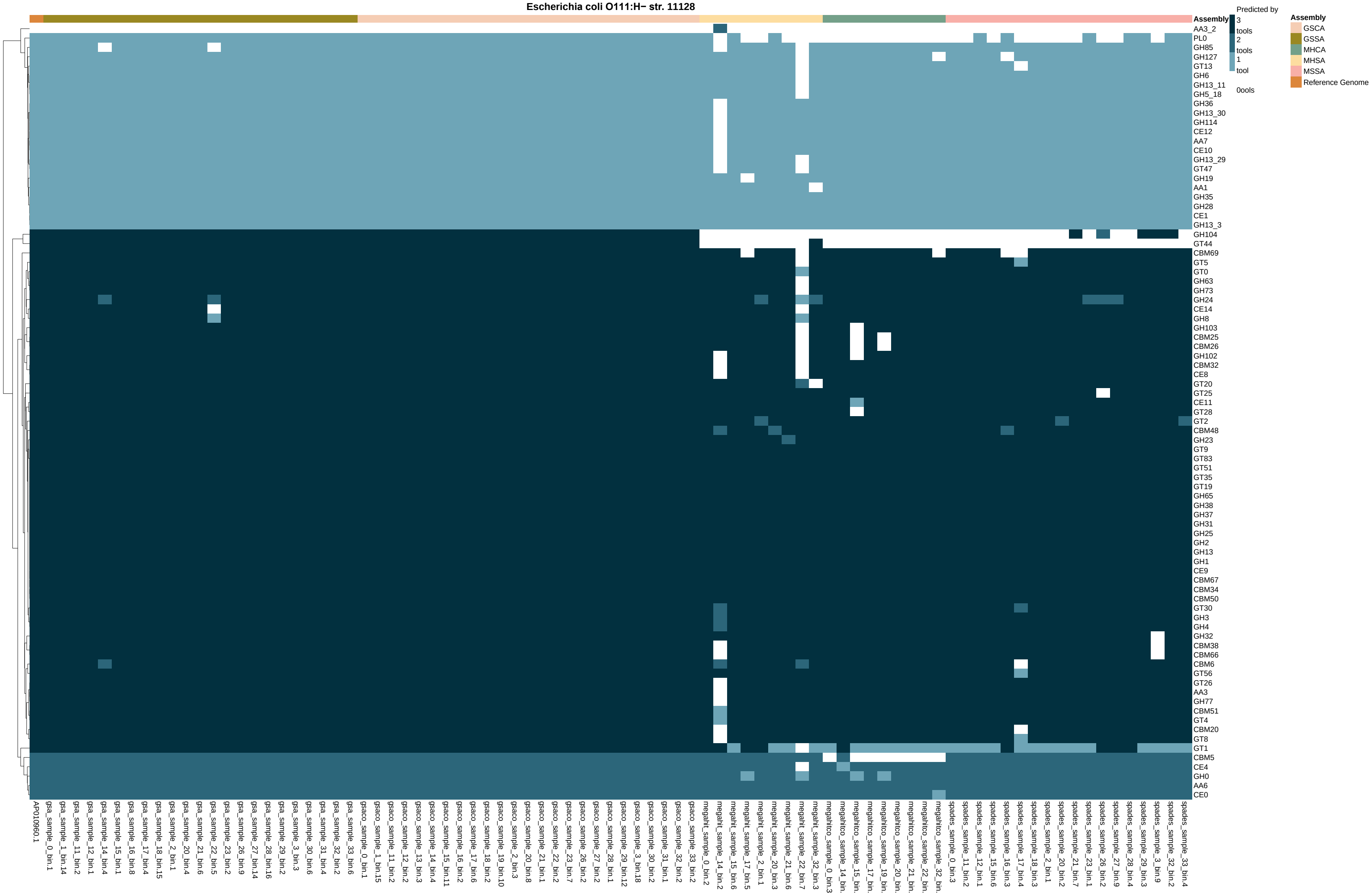
**Supplementary Figure 6: Sample 30 MAG recovery profile.** The right-most panel represents a list of all strains in the sample. The middle heatmap panel represents the relative abundance of samples. Each cell in the left-most panel represents whether a MAG was recovered for a particular strain by a particular assembly. Darker shades of purple and green represent relative abundance and reference coverage respectively, between 0% and 100%.



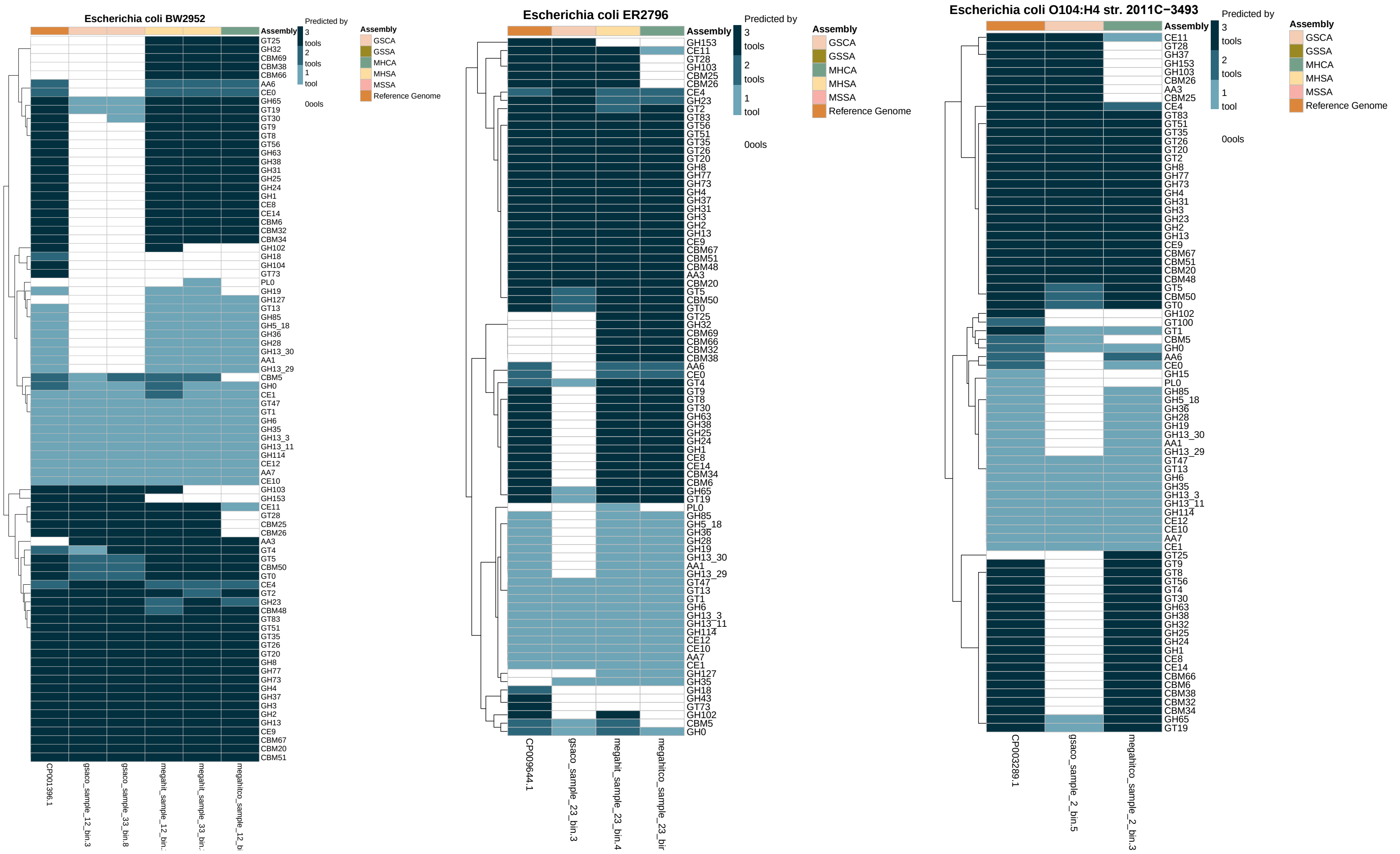
**Supplementary Figure 7: Sample 3 MAG recovery profile, with relative abundances of all strains in the sample and reference coverages of all retrieved MAGs.** The right-most panel represents a list of all strains in the sample. The middle heatmap panel represents the relative abundance of samples. Each cell in the left-most panel represents whether a MAG was recovered for a particular strain by a particular assembly. Darker shades of purple and green represent relative abundance and reference coverage respectively, between 0% and 100%.



**Supplementary Figure 8: CAZyme predictions for all *Akkermansia muciniphilia* MAGs and reference genomes across different samples and assembler types.** The bottom axis represents MAGs, while the right axis represents predicted CAZymes. Each MAG is color-coded with the assembly-type it was derived from.



**Supplementary Figure 9: CAZyme predictions for all *Escherichia coli O111:H- str. 11128* MAGs and reference genomes across different samples and assembler types.** The bottom axis represents MAGs, while the right axis represents predicted CAZymes. Each MAG is color-coded with the assembly-type it was derived from.



**Supplementary Figure 10: CAZyme predictions for all *E. coli BW2952,* E. *coli O104:H4 str. 2011C−3493*, and *E. coli ER2796* MAGs and reference genomes across different samples and assembler types.** The bottom axis represents MAGs, while the right axis represents predicted CAZymes. Each MAG is color-coded with the assembly-type it was derived from.

# Appendix to Chapter 3

**Supplementary Table 2: Thirty Baby, Food & Mi samples processed with the bfm_mg_flow pipeline. Two samples were included for each infant, collected before and after the introduction of solid food.**

| SAMPLE ID | SURETTE SAMPLE ID | PARTICIPANT ID | TIME POINT | AGE IN DAYS |
|---|---|---|---|---|
| **JCSA4** | JCS35 | 01-057 | Before | 182 |
| **JCSA5** | JCS36 | 01-057 | After | 202 |
| **JCSA62** | JCS202 | 06-023 | Before | 135 |
| **JCSA65** | JCS203 | 06-023 | After | 151 |
| **JCSA46** | JCS193 | 06-021 | Before | 174 |
| **JCSA50** | JCS197 | 06-021 | After | 195 |
| **JCSA39** | JCS187 | 06-020 | After | 162 |
| **JCSA35** | JCS184 | 06-020 | Before | 147 |
| **JCSA18** | JCS177 | 06-019 | After | 201 |
| **JCSA17** | JCS176 | 06-019 | Before | 187 |
| **JCSA66** | JCS204 | 04-068 | Before | 121 |
| **JCSA68** | JCS205 | 04-068 | After | 136 |
| **JCSA57** | JCS199 | 04-066 | Before | 178 |
| **JCSA60** | JCS201 | 04-066 | After | 194 |
| **JCSA42** | JCS189 | 04-061 | Before | 169 |
| **JCSA47** | JCS194 | 04-061 | After | 183 |
| **JCSA41** | JCS188 | 03-033 | Before | 186 |
| **JCSA44** | JCS191 | 03-033 | After | 201 |
| **JCSA10** | JCS171 | 03-031 | Before | 241 |
| **JCSA11** | JCS172 | 03-031 | After | 254 |
| **JCSA49** | JCS196 | 02-049 | After | 194 |
| **JCSA155** | JCS240 | 02-049 | Before | 182 |
| **JCSA54** | JCS198 | 02-048 | Before | 144 |
| **JCSA58** | JCS200 | 02-048 | After | 159 |
| **JCSA33** | JCS182 | 02-046 | Before | 165 |
| **JCSA37** | JCS185 | 02-046 | After | 183 |
| **JCSA48** | JCS195 | 02-045 | After | 201 |

| | | | | |
|---|---|---|---|---|
| **JCSA43** | JCS190 | 02-045 | Before | 184 |
| **JCSA34** | JCS183 | 02-043 | Before | 162 |
| **JCSA38** | JCS186 | 02-043 | After | 178 |

**Supplementary Table 3: Dominant species in Baby, Food & Mi metagenomic samples before and after the introduction of solid foods.**

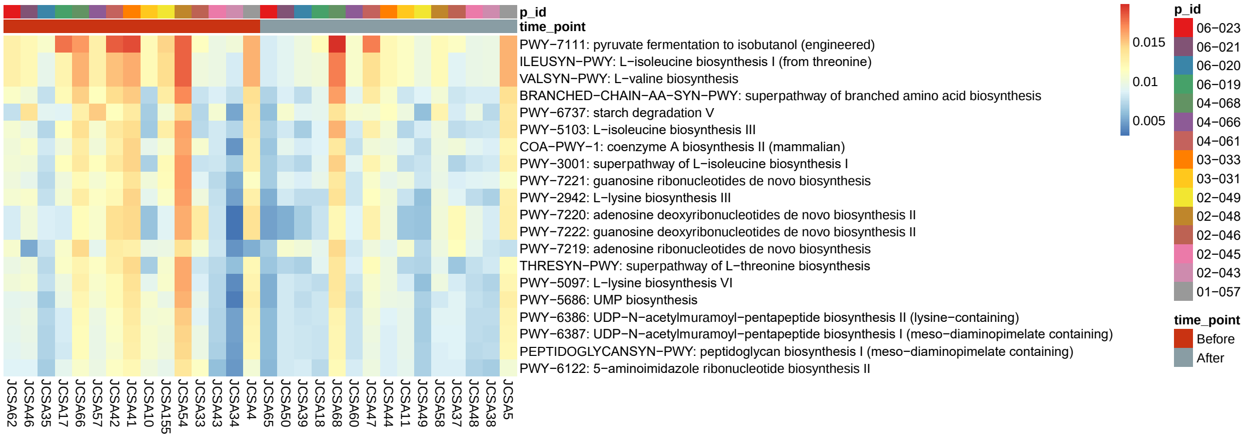
| TIME POINT (BEFORE OR AFTER THE INTRODUCTION OF SOLID FOODS) | SPECIES | NUMBER OF SAMPLES WHERE SPECIES IS MOST ABUNDANT |
|---|---|---|
| **Before** | *Bifidobacterium longum* | 6 |
| **Before** | *Bifidobacterium bifidum* | 3 |
| **Before** | *Bifidobacterium breve* | 2 |
| **Before** | *Bacteroides fragilis* | 1 |
| **Before** | *Bacteroides vulgatus* | 1 |
| **Before** | *Erysipelatoclostridium ramosum* | 1 |
| **Before** | *Escherichia coli* | 1 |
| **After** | *Bifidobacterium bifidum* | 5 |
| **After** | *Bacteroides fragilis* | 2 |
| **After** | *Bacteroides uniformis* | 2 |
| **After** | *Bifidobacterium breve* | 2 |
| **After** | *Bacteroides dorei* | 1 |
| **After** | *Bacteroides faecis* | 1 |
| **After** | *Bifidobacterium longum* | 1 |
| **After** | *Ruminococcus gnavus* | 1 |

**Supplementary Table 4: Most abundant microbial metabolic pathways, predicted with HUMAnN3**

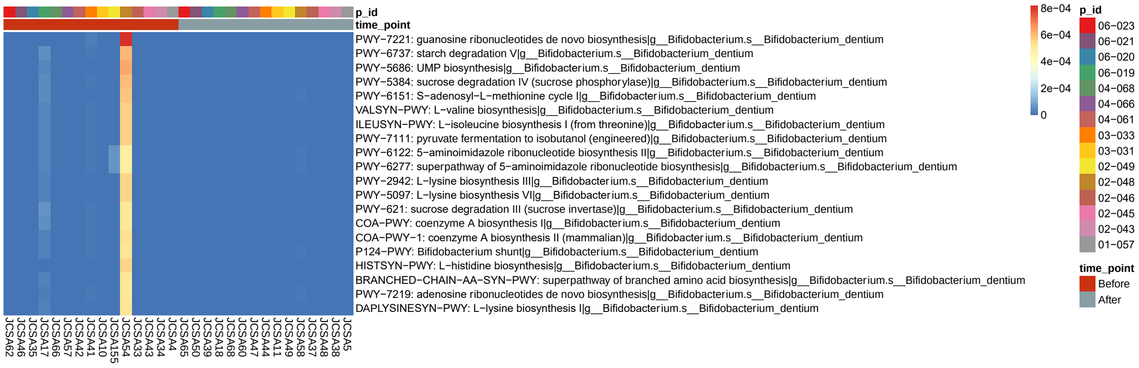
| PATHWAY | | AVERAGE ABUNDANCE |
|---|---|---|
| PWY-7111 | Pyruvate fermentation to isobutanol (engineered) | 0.013 |

| | | |
|---|---|---|
| ILEUSYN-PWY | L-isoleucine biosynthesis I (from threonine) | 0.0121 |
| VALSYN-PWY: | L-valine biosynthesis | 0.0121 |
| BRANCHED-CHAIN-AA-SYN-PWY | Superpathway of branched amino acid biosynthesis | 0.0106 |
| PWY-6737 | Starch degradation V | 0.0103 |
| PWY-5103 | L-isoleucine biosynthesis III | 0.0101 |
| COA-PWY-1 | Coenzyme A biosynthesis II (mammalian) | 0.0101 |
| PWY-3001 | Superpathway of L-isoleucine biosynthesis I | 0.01 |
| PWY-7221 | Guanosine ibonucleotides de novo biosynthesis | 0.00987 |
| PWY-2942 | L-lysine biosynthesis III | 0.00981 |
| PWY-7220 | Adenosine deoxyribonucleotides de novo biosynthesis II | 0.00961 |
| PWY-7222 | Guanosine deoxyribonucleotides de novo biosynthesis II | 0.00961 |
| PWY-7219 | Adenosine ribonucleotides de novo biosynthesis | 0.00959 |
| THRESYN-PWY | Superpathway of L-threonine biosynthesis | 0.00955 |
| PWY-5097 | L-lysine biosynthesis VI | 0.00955 |
| PWY-5686 | UMP biosynthesis | 0.00946 |
| PWY-6386 | UDP-N-acetylmuramoyl-pentapeptide biosynthesis II (lysine-containing) | 0.00933 |
| PWY-6387 | UDP-N-acetylmuramoyl-pentapeptide biosynthesis I (meso-diaminopimelate containing) | 0.00925 |
| PEPTIDOGLYCANSYN-PWY | Peptidoglycan biosynthesis I (meso-diaminopimelate containing) | 0.00919 |
| PWY-6122 | 5-aminoimidazole ribonucleotide biosynthesis II | 0.00892 |
| PWY-6277 | Superpathway of 5-aminoimidazole ribonucleotide biosynthesis | 0.00892 |
| PWY-724 | Superpathway of L-lysine, L-threonine and L-methionine biosynthesis II | 0.0089 |
| DTDPRHAMSYN-PWY | dTDP-L-rhamnose biosynthesis I | 0.00878 |
| ANAGLYCOLYSIS-PWY | Glycolysis III (from glucose) | 0.00873 |

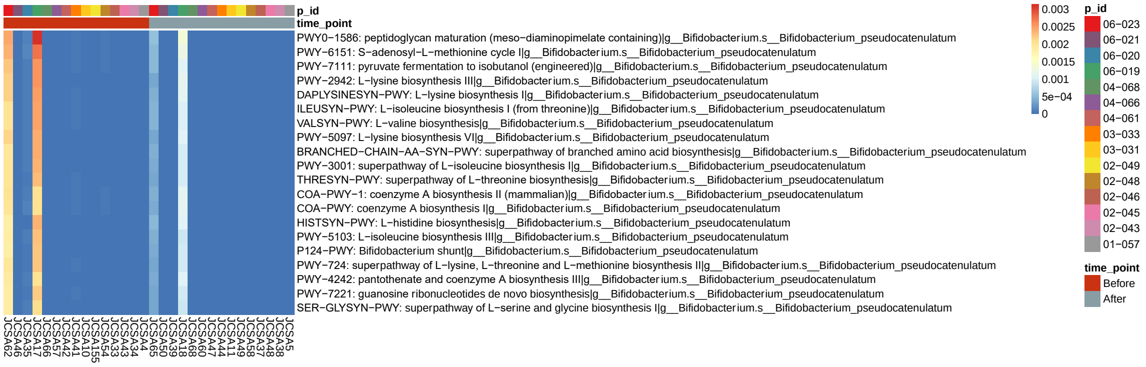
| | | |
|---|---|---|
| HISTSYN-PWY | L-histidine biosynthesis | 0.00866 |
| PWY-6151 | S-adenosyl-L-methionine cycle I | 0.00866 |
| PWY-6385 | Peptidoglycan biosynthesis III (mycobacteria) | 0.00844 |
| PWY-6121 | 5-aminoimidazole ribonucleotide biosynthesis I | 0.00833 |
| ARGSYNBSUB-PWY | L-arginine biosynthesis II (acetyl cycle) | 0.00801 |
| PWY-5913 | TCA cycle VI (obligate autotrophs) | 0.00785 |
| PWY-6936 | seleno-amino acid biosynthesis | 0.0078 |
| PWY-3841 | folate transformations II | 0.00779 |
| NONMEVIPP-PWY | methylerythritol phosphate pathway I | 0.00774 |
| GLYCOCAT-PWY | glycogen degradation I (bacterial) | 0.00771 |
| ARGSYN-PWY | L-arginine biosynthesis I (via L-ornithine) | 0.00768 |
| PWY-7400: | L-arginine biosynthesis IV (archaebacteria) | 0.00766 |
| PYRIDNUCSYN-PWY | NAD biosynthesis I (from aspartate) | 0.00761 |
| COMPLETE-ARO-PWY | superpathway of aromatic amino acid biosynthesis | 0.00742 |
| ARO-PWY | chorismate biosynthesis I | 0.00731 |
| PWY-7229 | superpathway of adenosine nucleotides de novo biosynthesis I | 0.00729 |
| HSERMETANA-PWY | L-methionine biosynthesis III | 0.00716 |
| 1CMET2-PWY | N10-formyl-tetrahydrofolate biosynthesis | 0.00714 |
| PWY-6163 | chorismate biosynthesis from 3-dehydroquinate | 0.00712 |
| PWY-1042 | glycolysis IV (plant cytosol) | 0.0071 |
| GLUTORN-PWY | L-ornithine biosynthesis | 0.00709 |
| UDPNAGSYN-PWY | UDP-N-acetyl-D-glucosamine biosynthesis I | 0.0069 |
| PWY0-1586 | peptidoglycan maturation (meso-diaminopimelate containing) | 0.00662 |
| GLYCOGENSYNTH-PWY | glycogen biosynthesis I (from ADP-D-Glucose) | 0.00656 |
| PWY-6123 | inosine-5'-phosphate biosynthesis I | 0.00636 |
| PWY-5384 | sucrose degradation IV (sucrose phosphorylase) | 0.00635 |



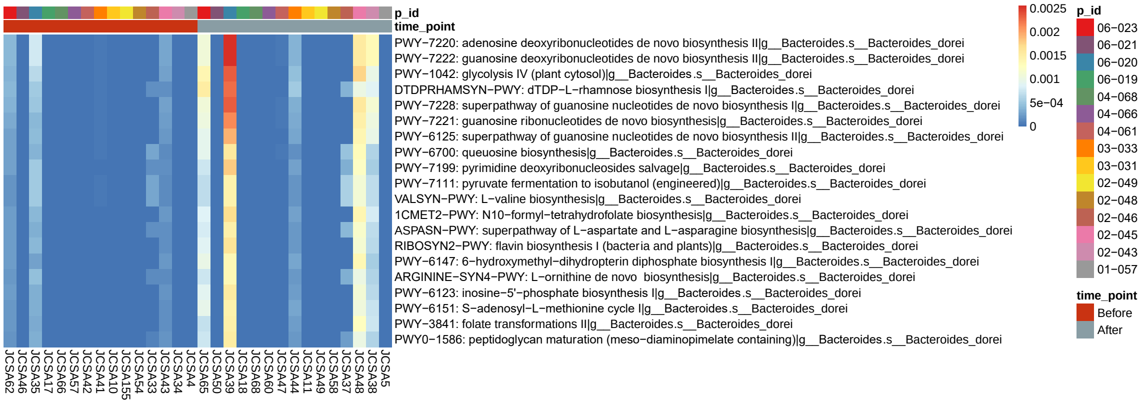
**Supplementary Figure 11: Most abundant bacterial metabolic pathways in Baby, Food & Mi samples before and after the introduction of solid foods. Samples are separated based on sample collection time point and participant ID.**



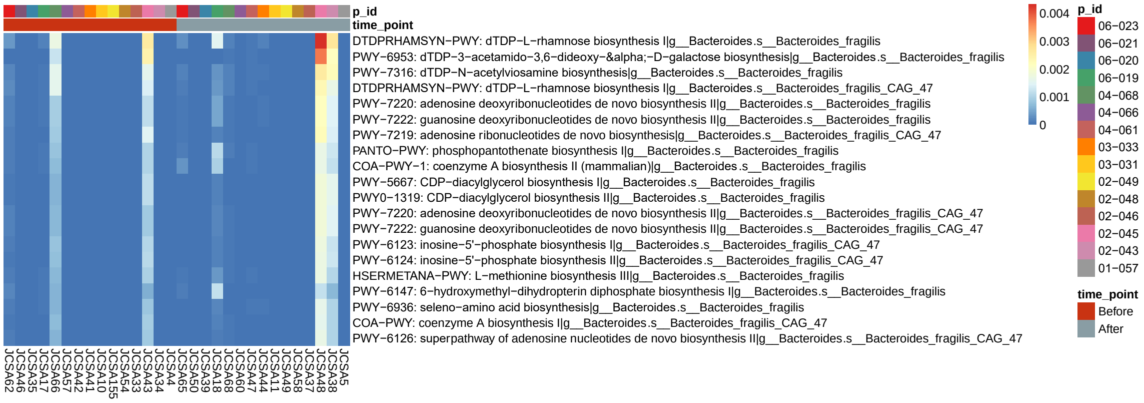
**Supplementary Figure 12: 20 most abundant microbial metabolic pathways predicted for *B. dentium*, before (left) and after (right) the introduction of solid foods**



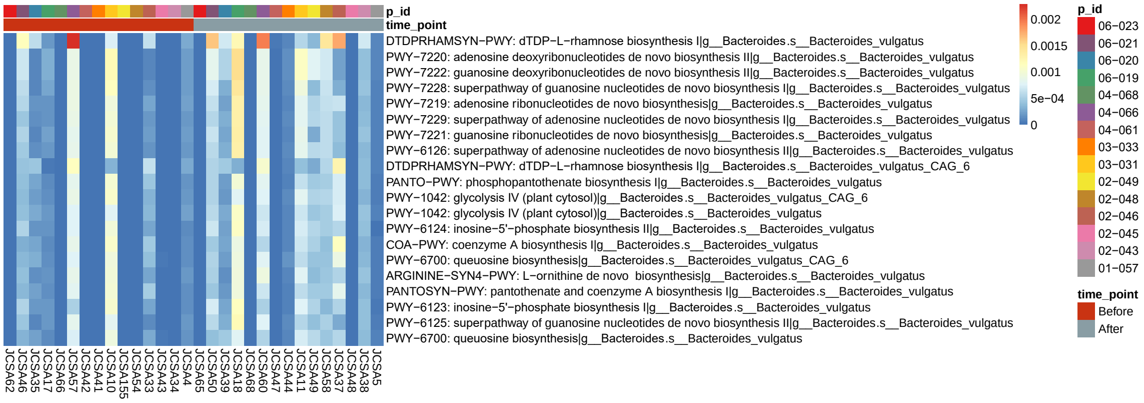
**Supplementary Figure 13: 20 most abundant microbial metabolic pathways predicted for *B. pseudocatenulatum*, before (left) and after (right) the introduction of solid foods**



**Supplementary Figure 14: 20 most abundant microbial metabolic pathways predicted for *B. dorei*, before (left) and after (right) the introduction of solid foods**

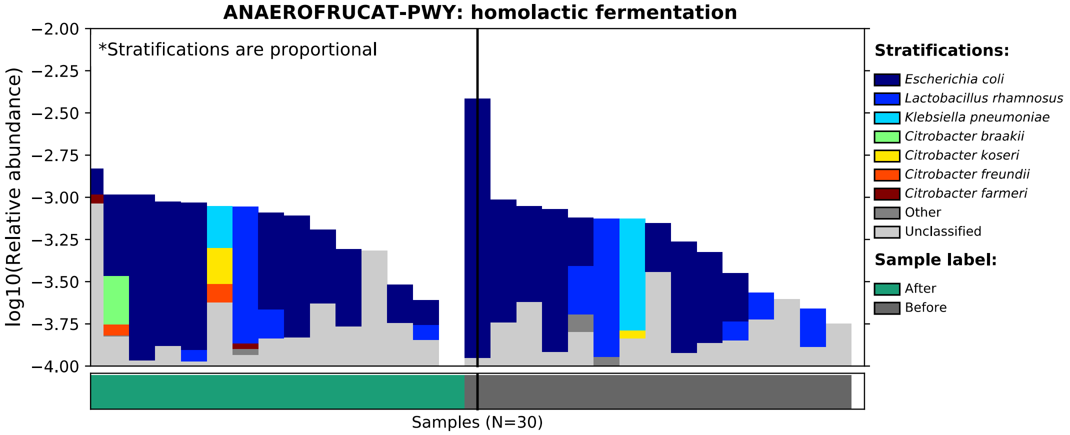


**Supplementary Figure 15: 20 most abundant microbial metabolic pathways predicted for *B. fragilis*, before (left) and after (right) the introduction of solid foods**

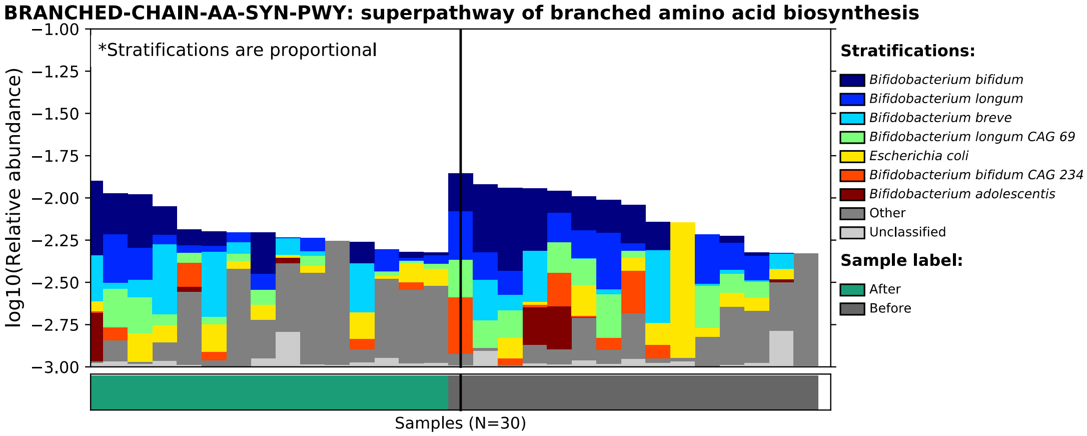


**Supplementary Figure 16: 20 most abundant microbial metabolic pathways predicted for *B. vulgatus*, before (left) and after (right) the introduction of solid foods**

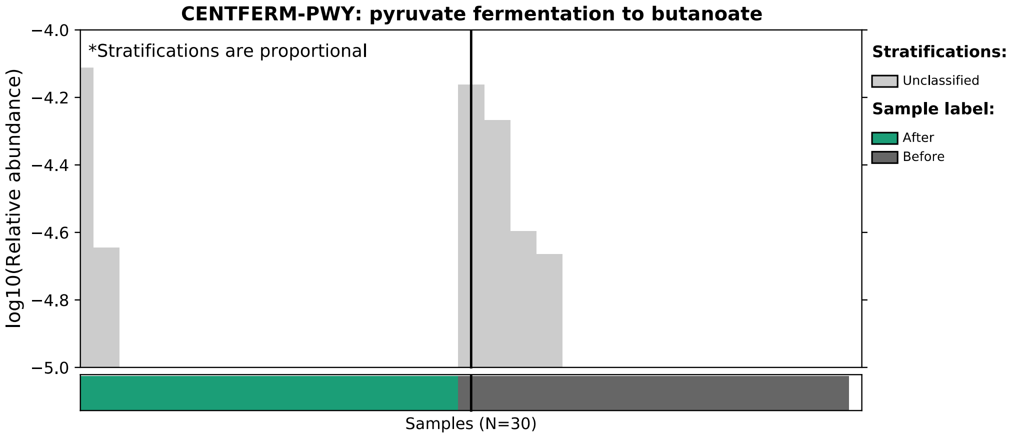
# Appendix to Chapter 3B: Metabolic Pathway Abundance



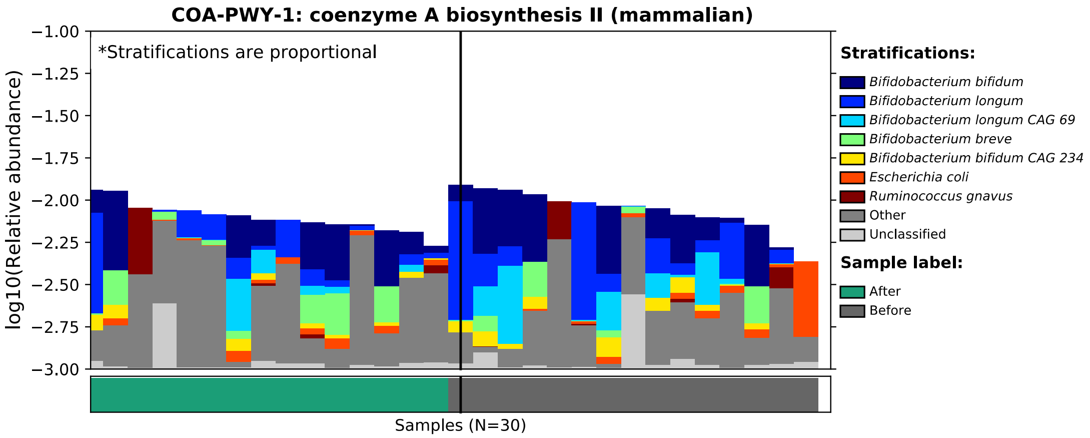
**Supplementary Figure 17: Abundance of bacteria containing the homeolactic fermentation pathway, before (right) and after (left) solid food introduction**



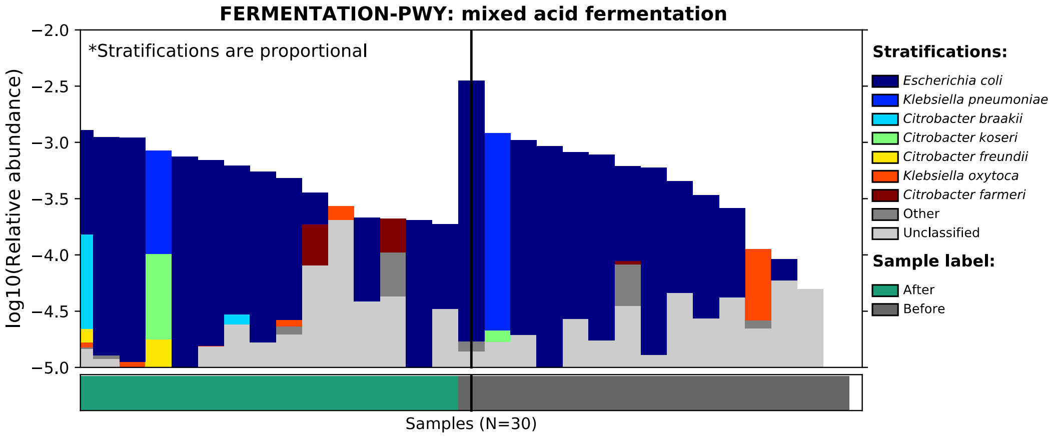
**Supplementary Figure 18: Abundance of bacteria containing the superpathway of branched amino acid biosynthesis pathway, before (right) and after (left) solid food introduction**



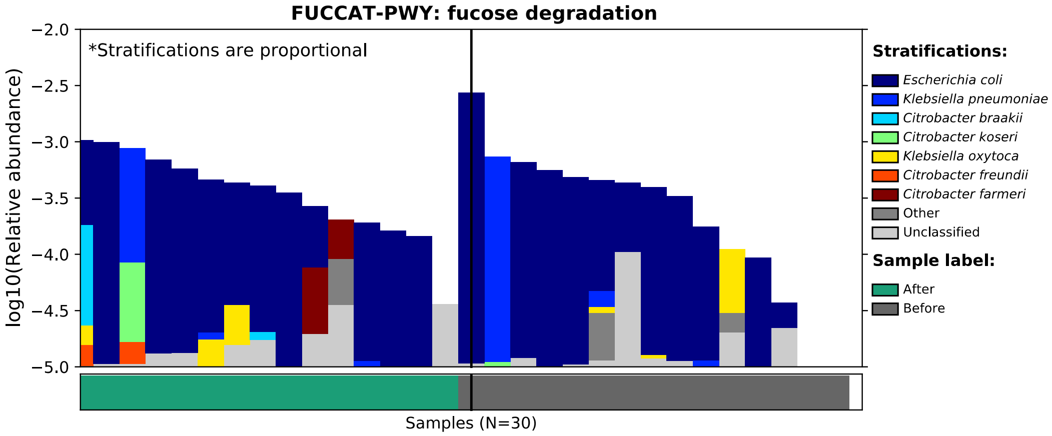
**Supplementary Figure 19: Abundance of bacteria containing the pyruvate fermentation to butanoate, before (right) and after (left) solid food introduction**



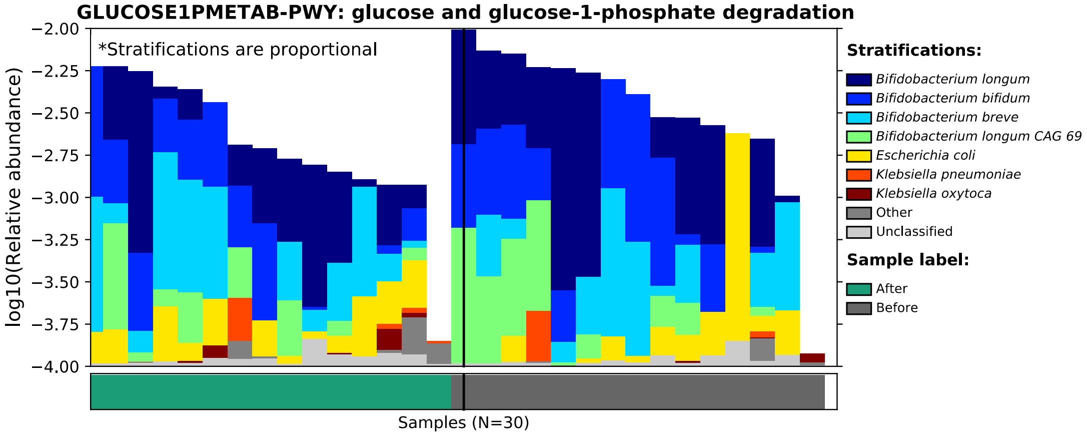
**Supplementary Figure 20: Abundance of bacteria containing the coenzyme A biosynthesis II pathway, before (right) and after (left) solid food introduction**



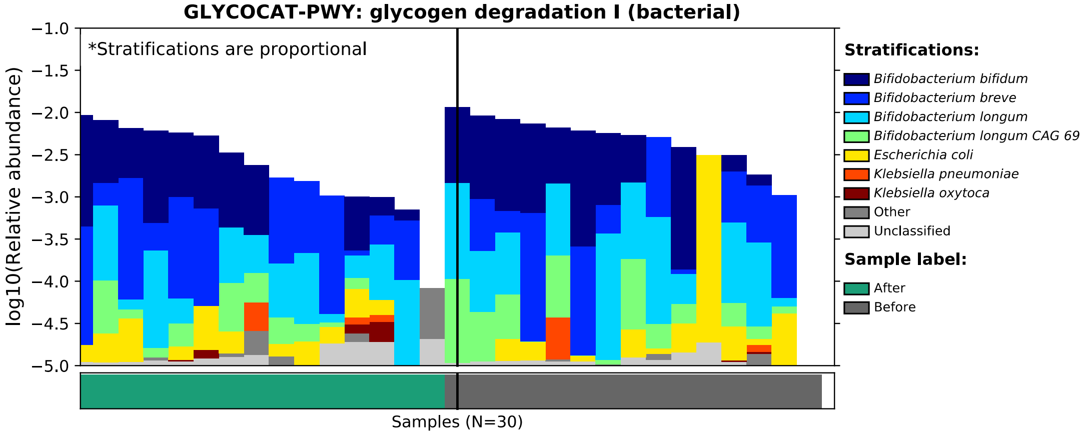
**Supplementary Figure 21: Abundance of bacteria containing the mixed acid fermentation pathway, before (right) and after (left) solid food introduction**



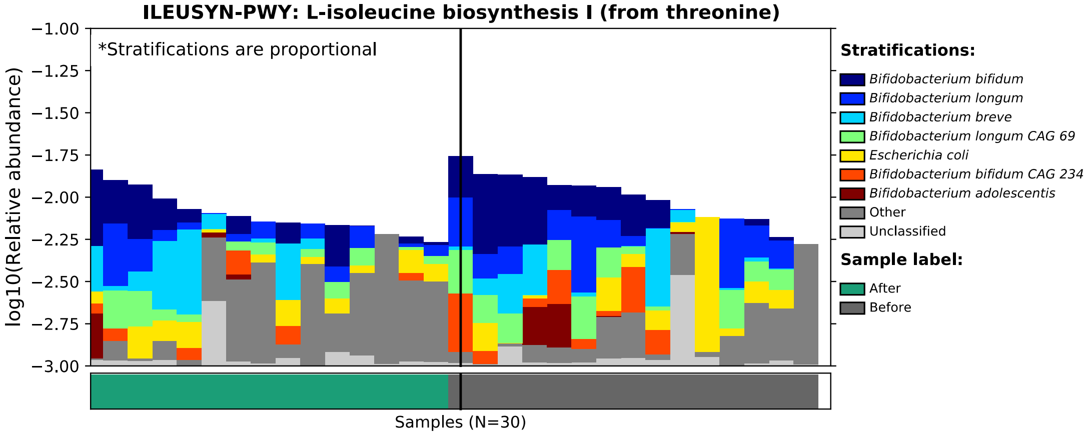
**Supplementary Figure 22: Abundance of bacteria containing the fucose degradation pathway, before (right) and after (left) solid food introduction**



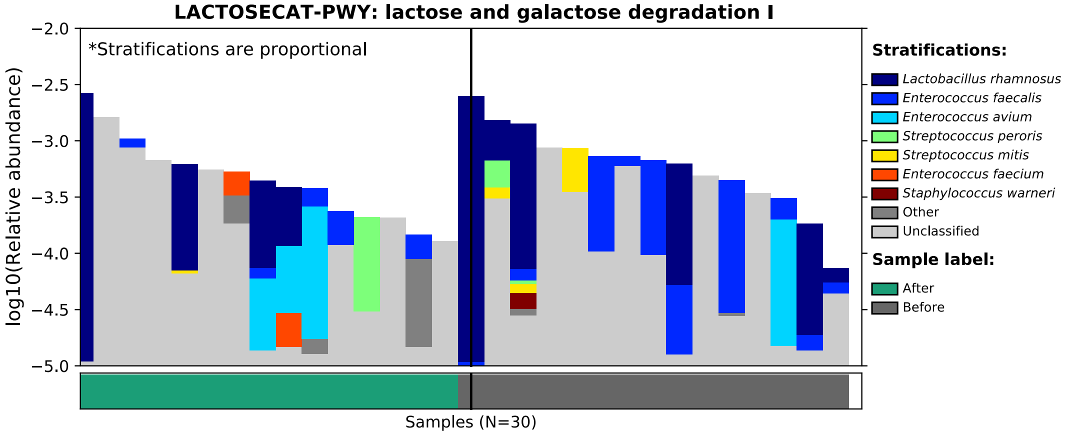
**Supplementary Figure 23: Abundance of bacteria containing the glucose and glucose-1-phosphate degradation pathway, before (right) and after (left) solid food introduction**



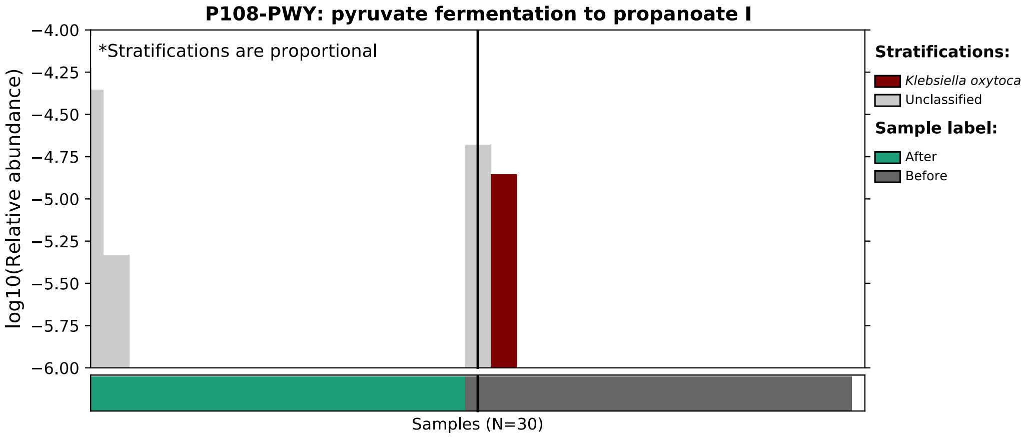
**Supplementary Figure 24: Abundance of bacteria containing the glycogen degradation I pathway, before (right) and after (left) solid food introduction**



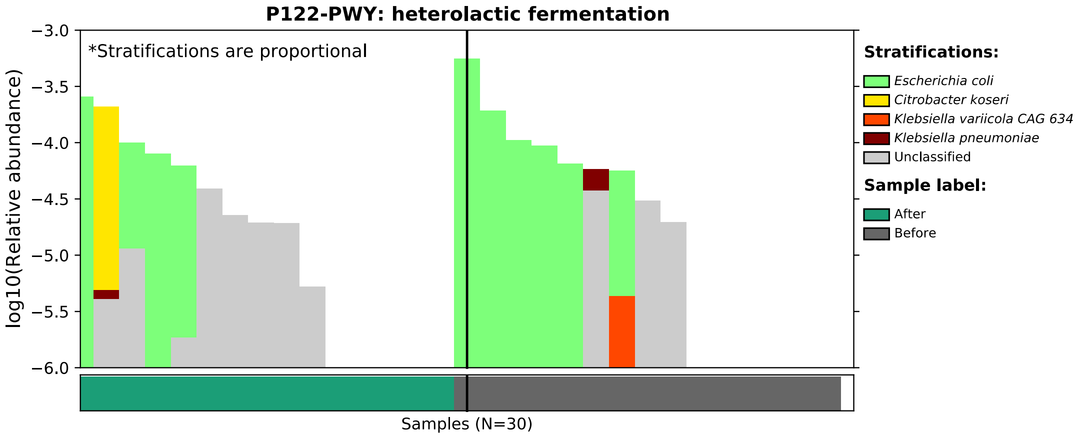
**Supplementary Figure 25: Abundance of bacteria containing L-isoleucine biosynthesis pathway, before (right) and after (left) solid food introduction**



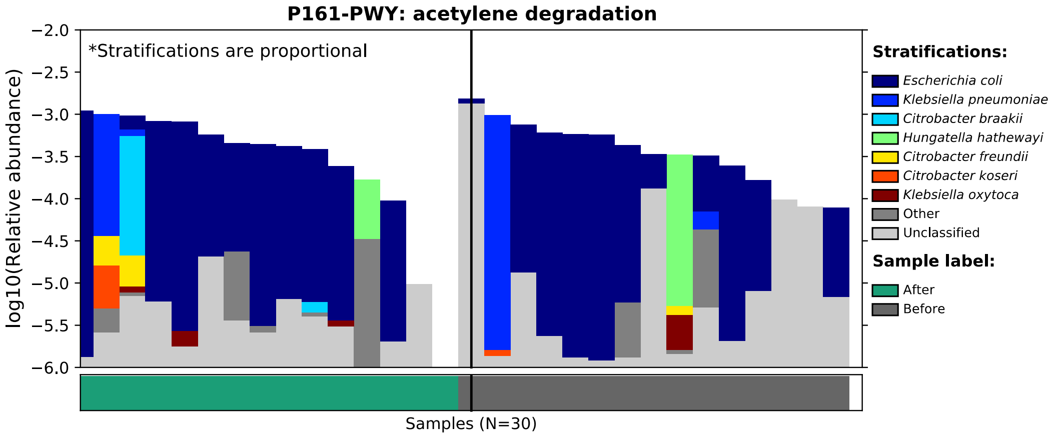
**Supplementary Figure 26: Abundance of bacteria containing the lactose and galactose degradation I pathway, before (right) and after (left) solid food introduction**



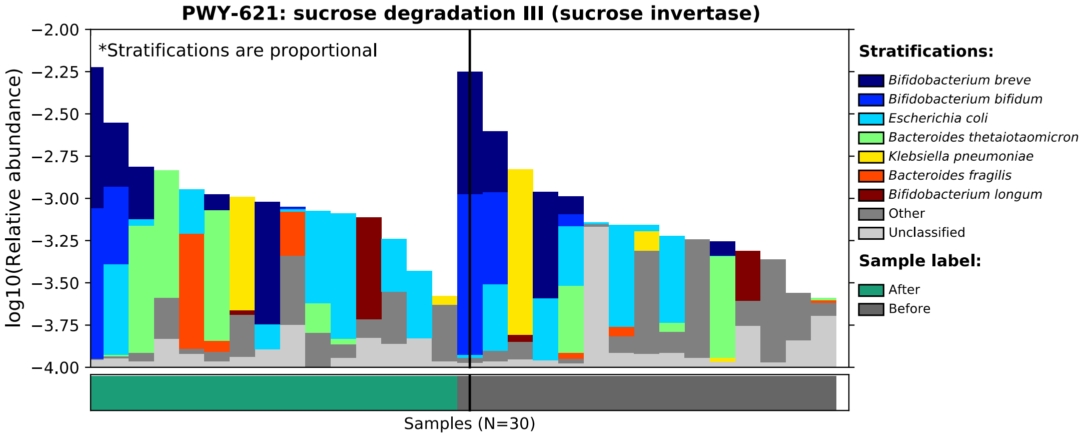
**Supplementary Figure 27: Abundance of bacteria containing the pyruvate fermentation to propanoate pathway, before (right) and after (left) solid food introduction**



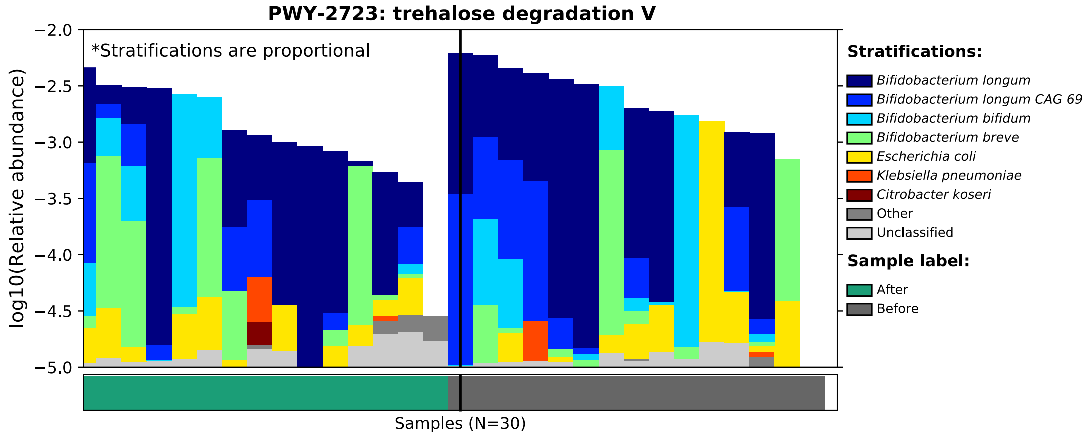
**Supplementary Figure 28: Abundance of bacteria containing the heterolactic fermentation pathway, before (right) and after (left) solid food introduction**



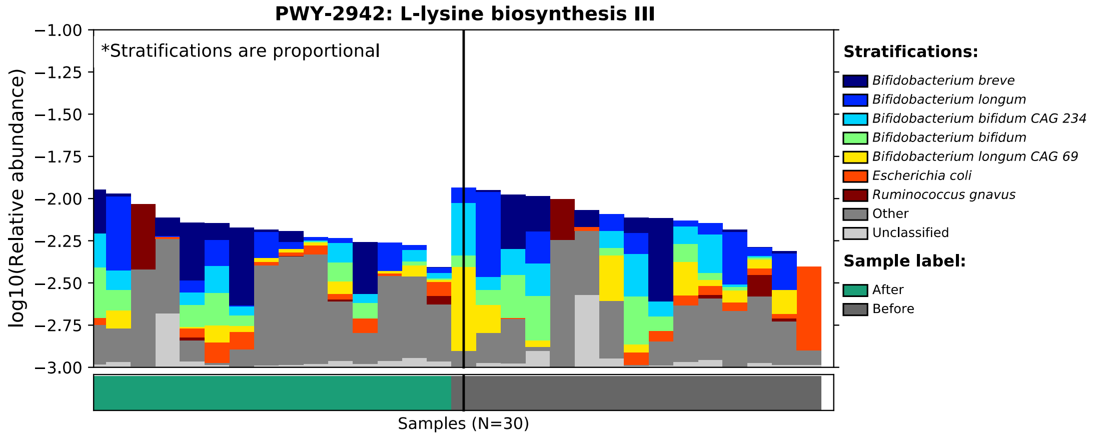
**Supplementary Figure 29: Abundance of bacteria containing the acetylene degradation pathway, before (right) and after (left) solid food introduction**



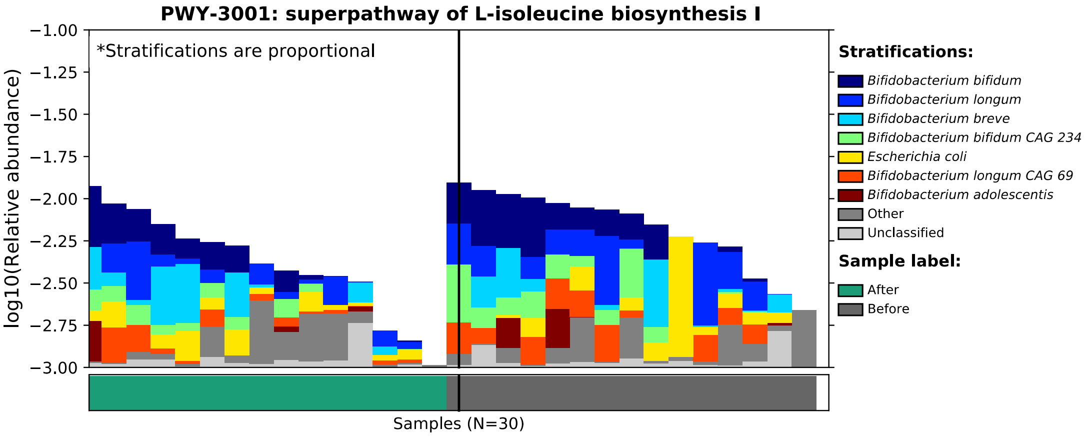
**Supplementary Figure 30: Abundance of bacteria containing the sucrose degradation III pathway, before (right) and after (left) solid food introduction**



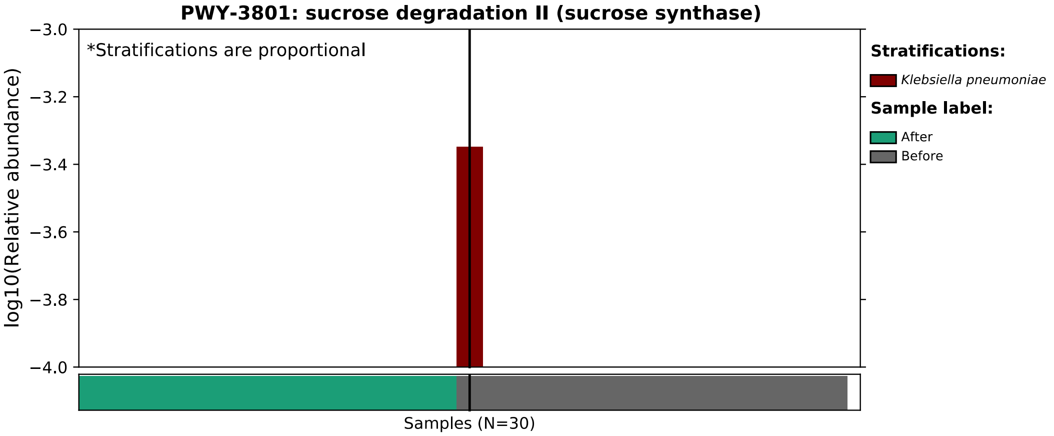
**Supplementary Figure 31: Abundance of bacteria containing the trehalose degradation V pathway, before (right) and after (left) solid food introduction**



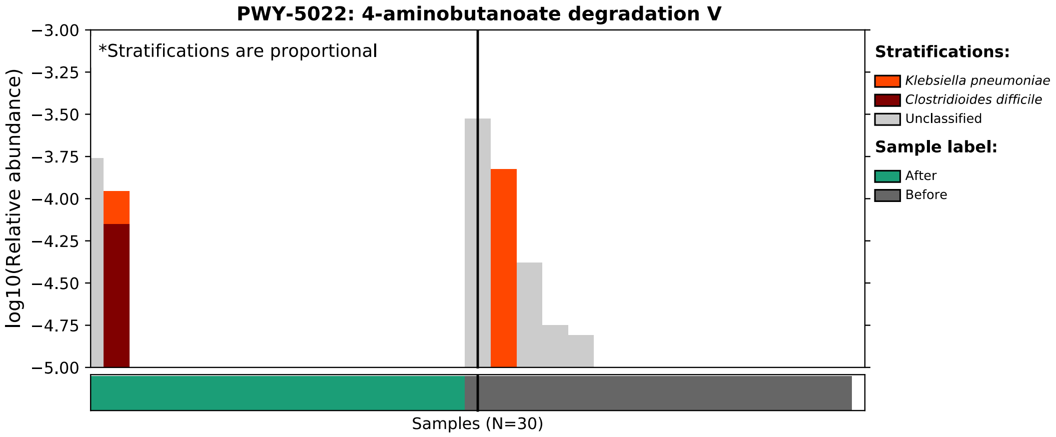
**Supplementary Figure 32: Abundance of bacteria containing the L-lysine biosynthesis III pathway, before (right) and after (left) solid food introduction**



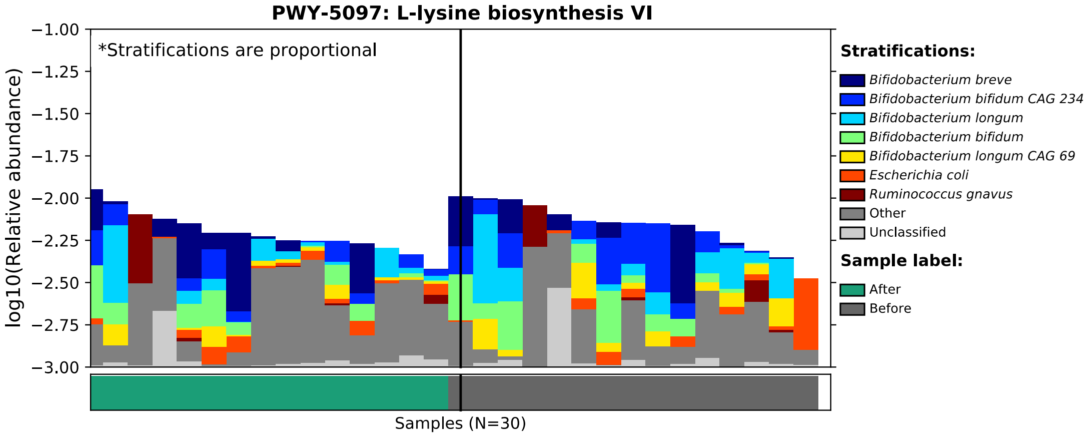
**Supplementary Figure 33: Abundance of bacteria containing the superpathway of L-isoleucine biosynthesis I pathway, before (right) and after (left) solid food introduction**



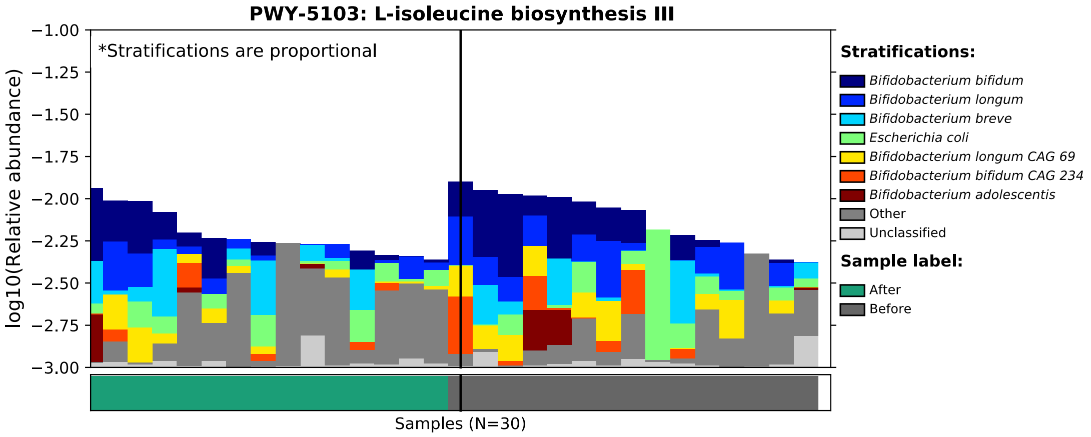
**Supplementary Figure 34: Abundance of bacteria containing the sucrose degradation II pathway, before (right) and after (left) solid food introduction**



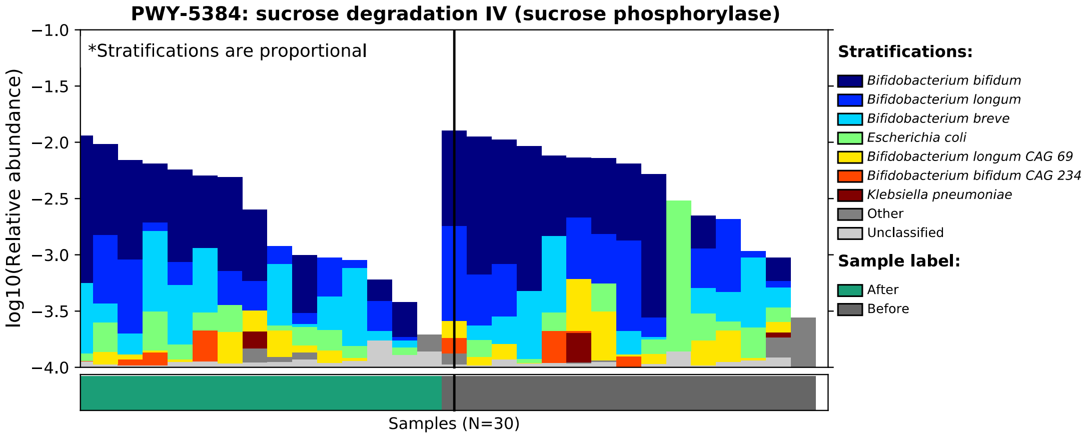
**Supplementary Figure 35: Abundance of bacteria containing the 4-aminobutanoate degradation V pathway, before (right) and after (left) solid food introduction**



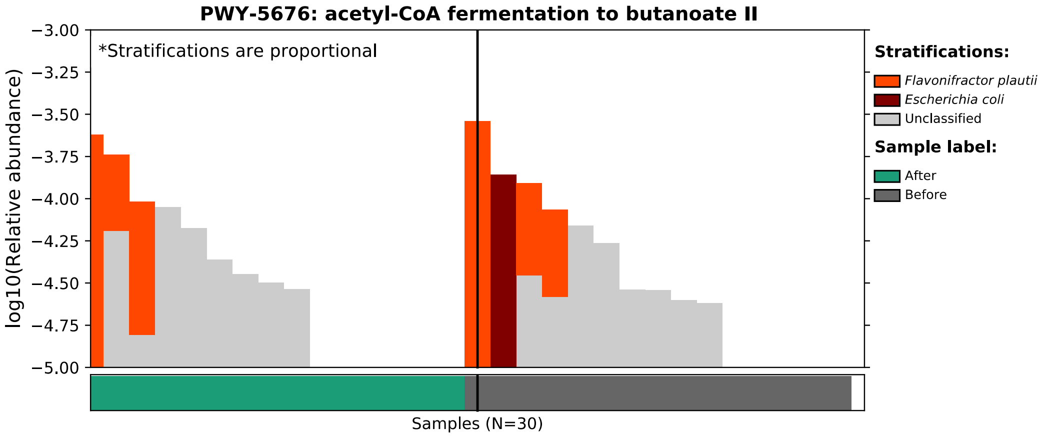
**Supplementary Figure 36: Abundance of bacteria containing the L-lysine biosynthesis VI pathway, before (right) and after (left) solid food introduction**



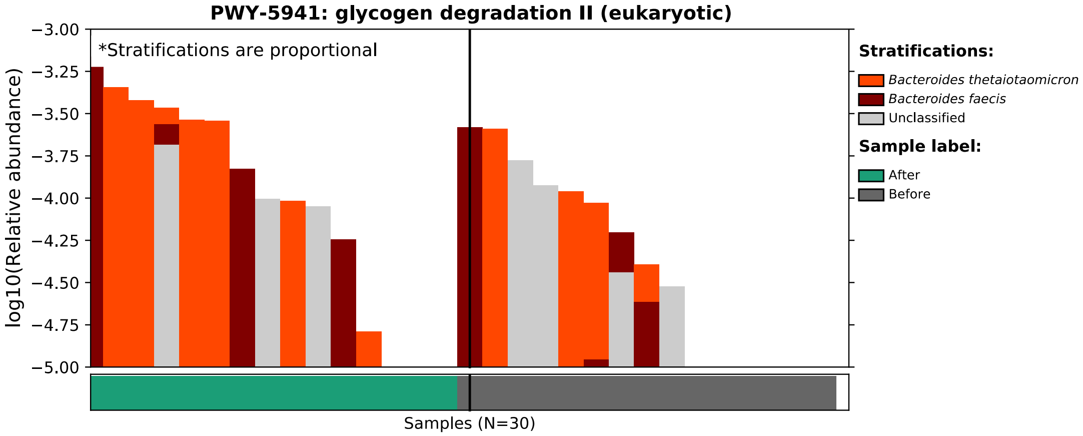
**Supplementary Figure 37: Abundance of bacteria containing the L-isoleucine biosynthesis III pathway, before (right) and after (left) solid food introduction**



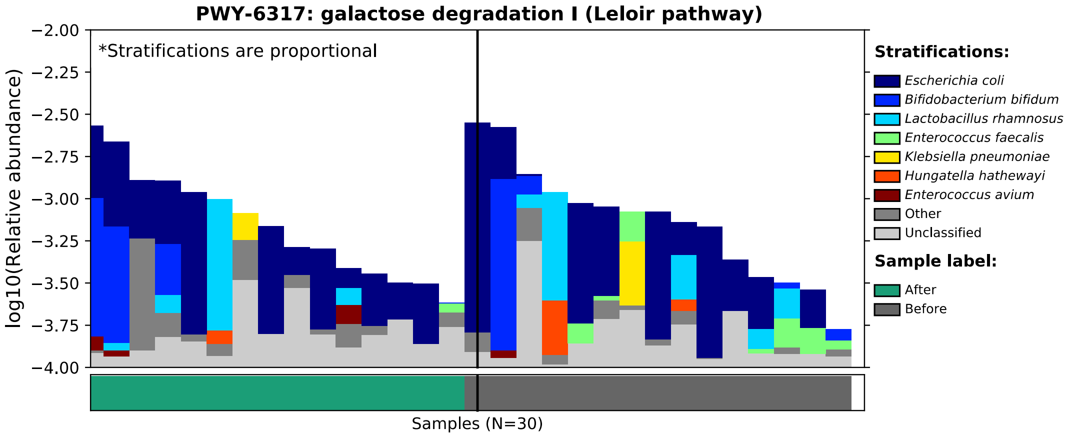
**Supplementary Figure 38: Abundance of bacteria containing the sucrose degradation IV pathway, before (right) and after (left) solid food introduction**



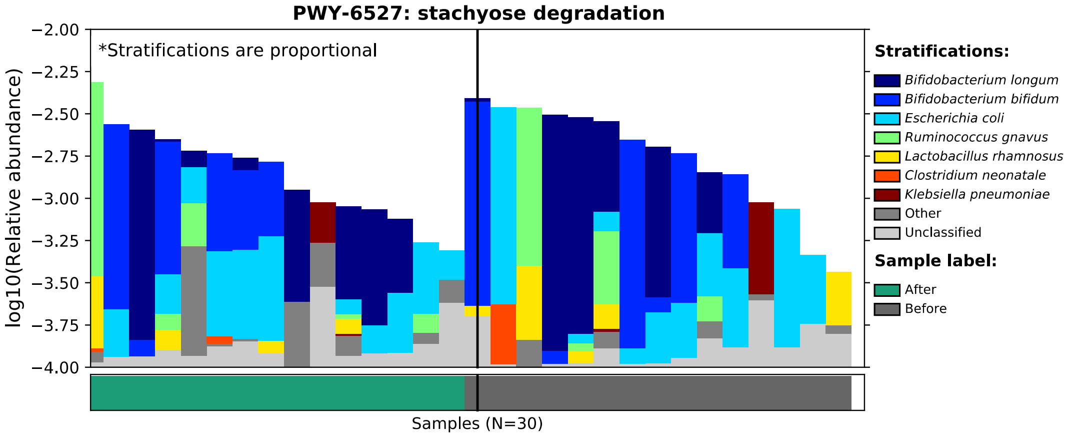
**Supplementary Figure 39: Abundance of bacteria containing the acetyl-CoA fermentation to butanoate II pathway, before (right) and after (left) solid food introduction**



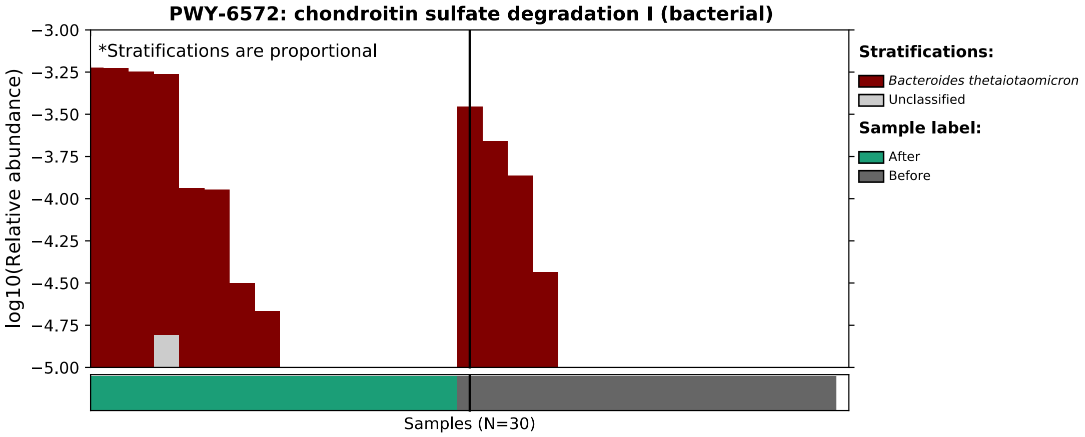
**Supplementary Figure 40: Abundance of bacteria containing the glycogen degradation II pathway, before (right) and after (left) solid food introduction**



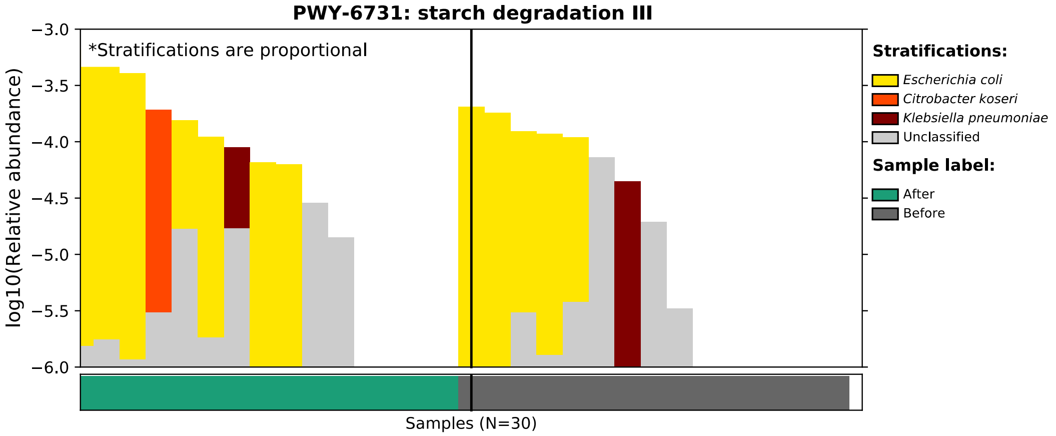
**Supplementary Figure 41: Abundance of bacteria containing the galactose degradation I pathway, before (right) and after (left) solid food introduction**



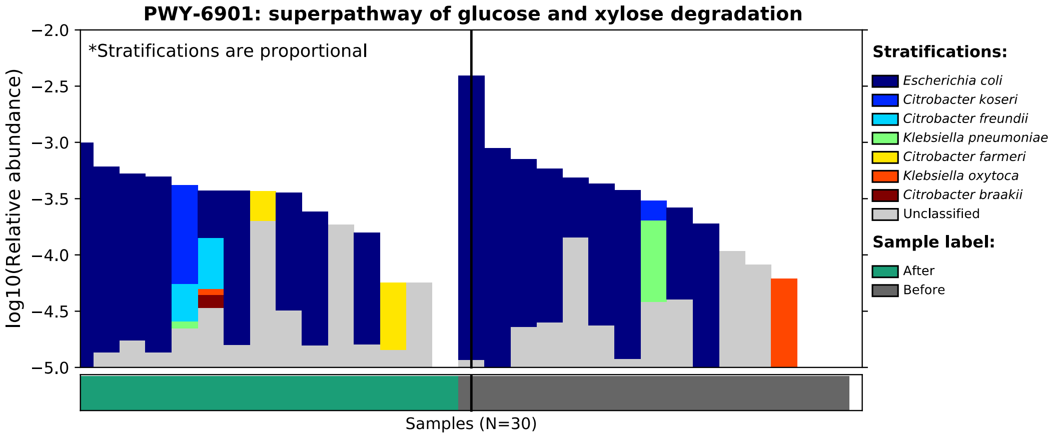
**Supplementary Figure 42: Abundance of bacteria containing the stachyose degradation pathway, before (right) and after (left) solid food introduction**



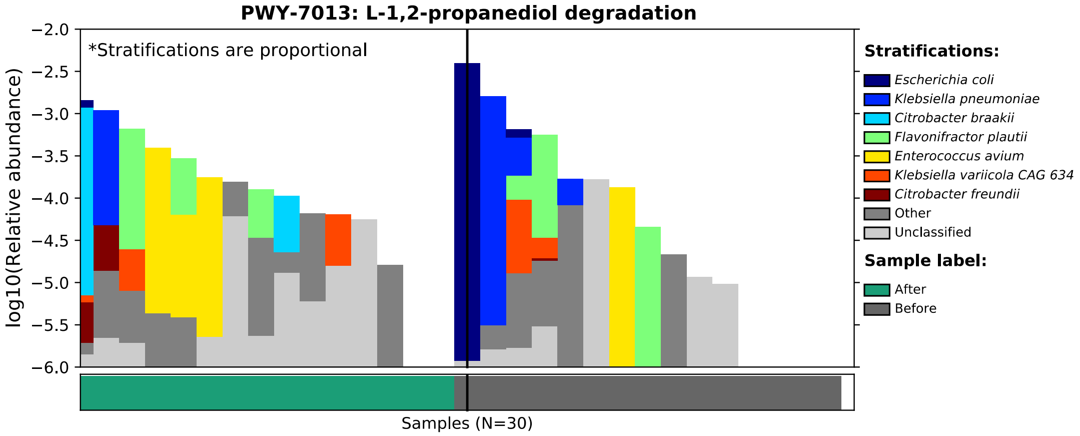
**Supplementary Figure 43: Abundance of bacteria containing the chondrltin sulfate degradation pathway, before (right) and after (left) solid food introduction**



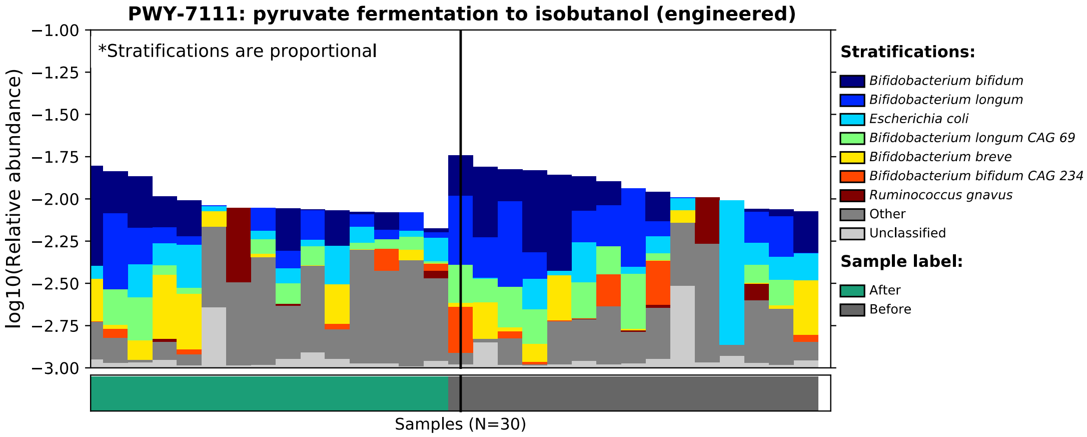
**Supplementary Figure 44: Abundance of bacteria containing the starch degradation III pathway, before (right) and after (left) solid food introduction**



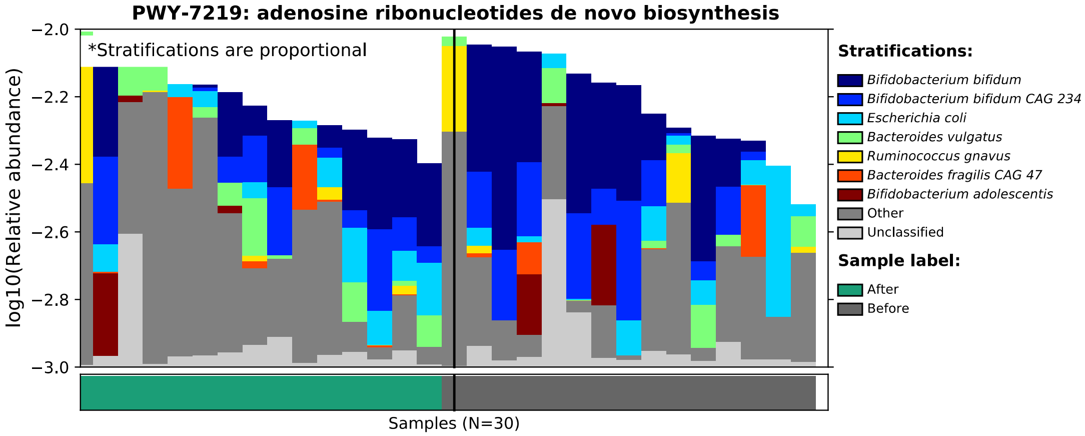
**Supplementary Figure 45: Abundance of bacteria containing the superpathway of glucose and xylose degradation, before (right) and after (left) solid food introduction**



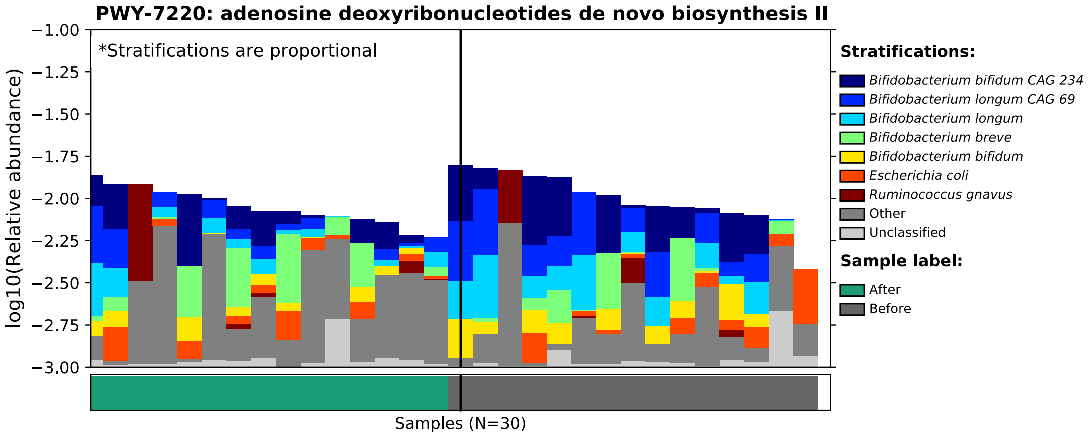
**Supplementary Figure 46: Abundance of bacteria containing the L-1,2-propanediol degradation pathway, before (right) and after (left) solid food introduction**



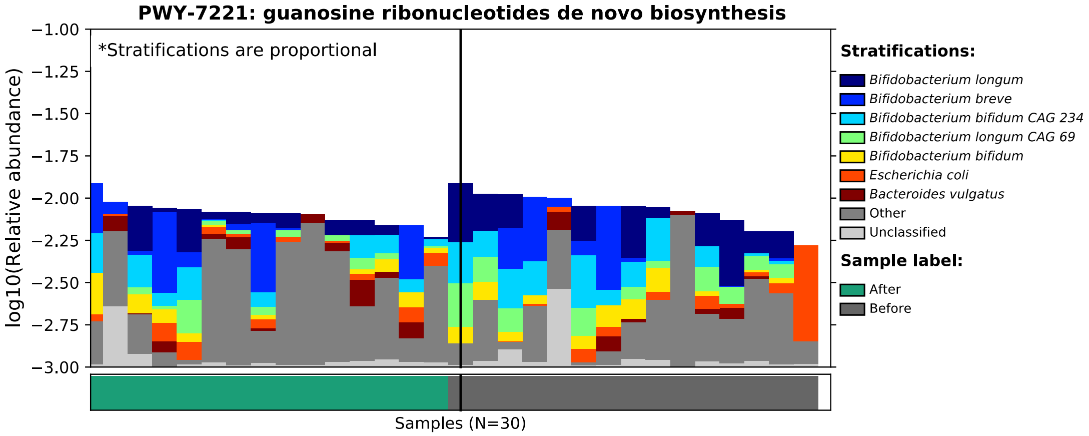
**Supplementary Figure 47: Abundance of bacteria containing the pyruvate fermentation to isobutanol (engineered) pathway, before (right) and after (left) solid food introduction**



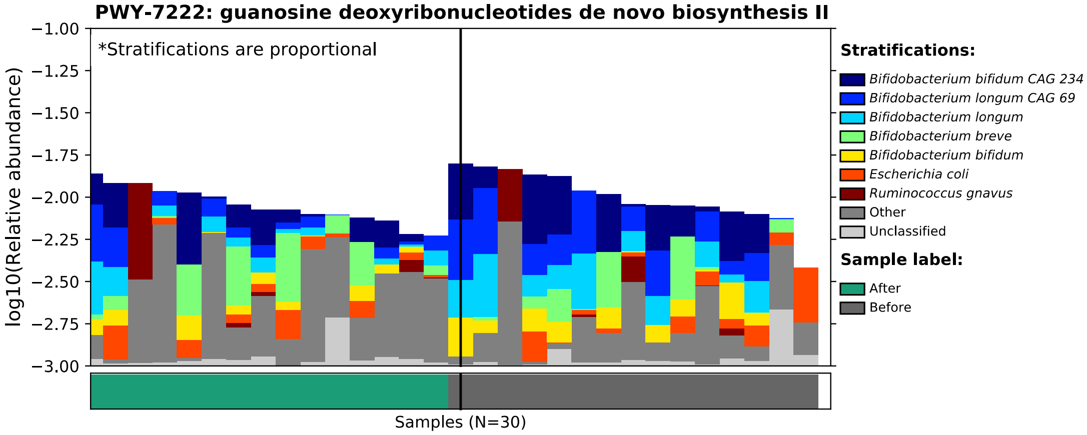
**Supplementary Figure 48: Abundance of bacteria containing the adenosine ribonucleotides de novo biosynthesis pathway, before (right) and after (left) solid food introduction**



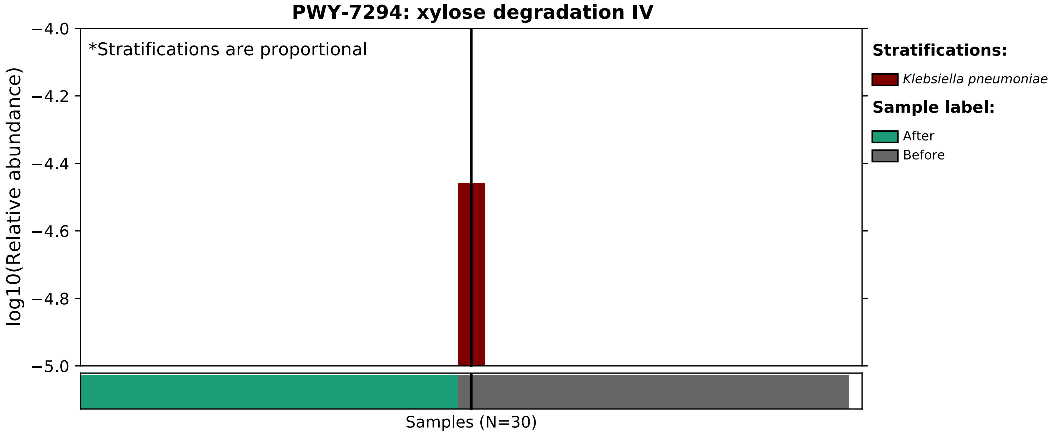
**Supplementary Figure 49: Abundance of bacteria containing the adenosine deoxyribonucleotides de novo biosynthesis II pathway, before (right) and after (left) solid food introduction**



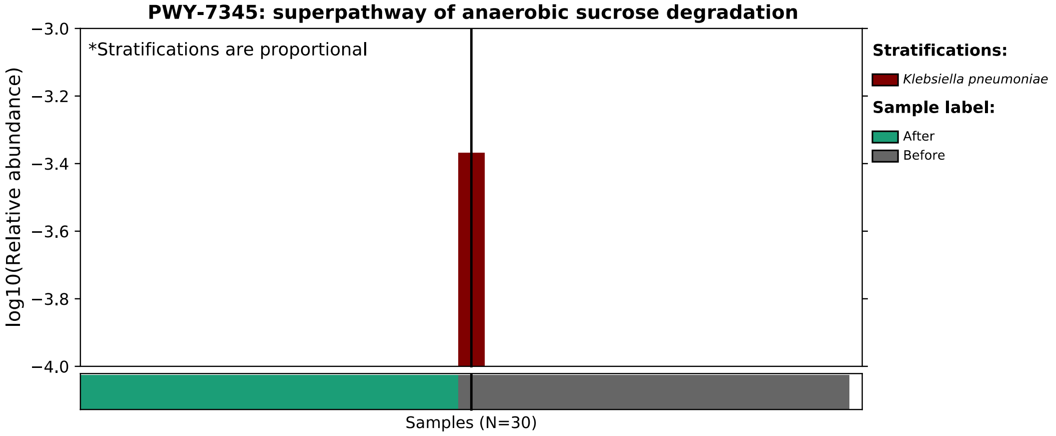
**Supplementary Figure 50: Abundance of bacteria containing the guanosine ribonucleotides de novo biosynthesis pathway, before (right) and after (left) solid food introduction**



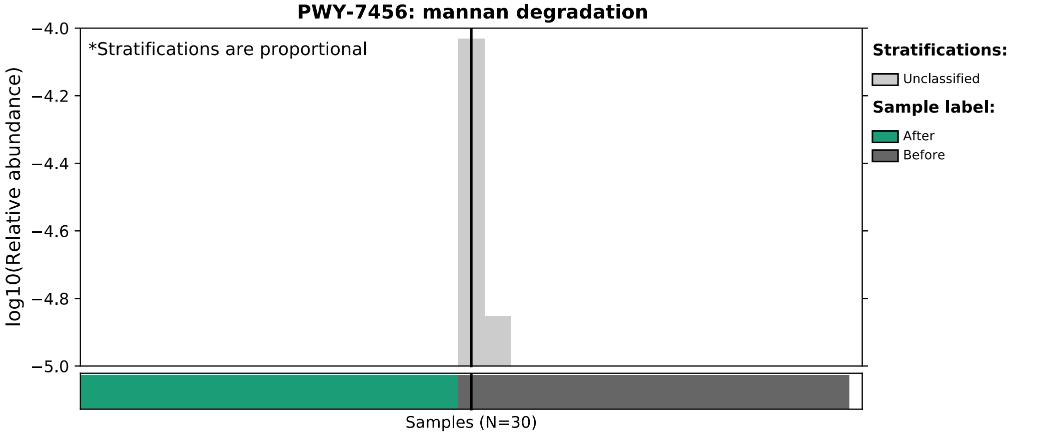
**Supplementary Figure 51: Abundance of bacteria containing the guanosine deoxyribonucleotides de novo biosynthesis II pathway, before (right) and after (left) solid food introduction**



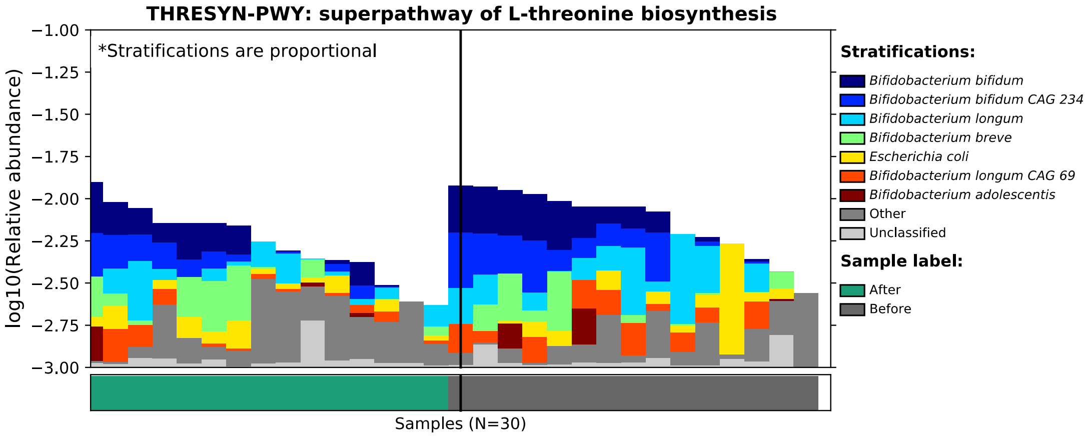
**Supplementary Figure 52: Abundance of bacteria containing the xylose degradation IV pathway, before (right) and after (left) solid food introduction**



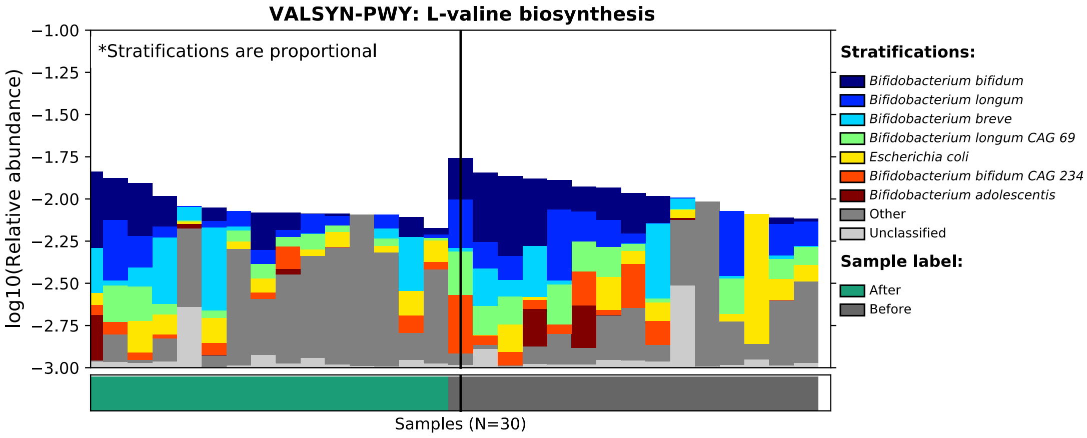
**Supplementary Figure 53: Abundance of bacteria containing the superpathway of anaerobic sucrose degradation, before (right) and after (left) solid food introduction**



**Supplementary Figure 54: Abundance of bacteria containing the mannan degradation pathway, before (right) and after (left) solid food introduction**



**Supplementary Figure 55: Abundance of bacteria containing the superpathway of L-threonine biosynthesis, before (right) and after (left) solid food introduction**



**Supplementary Figure 56: Abundance of bacteria containing the L-valine biosynthesis pathway, before (right) and after (left) solid food introduction**