Ollscoil na hÉireann, Corcaigh

National University of Ireland, Cork

Mother and Infant Microbiomes During the Perinatal Period: Compositional Analyses, Identification of Strain Transfer Events and Implications for Pregnancy Outcomes.

Thesis presented by

Naga Sai Ravi Chandra Nori, B. Tech

for the degree of

Doctor of Philosophy

UCC student number: 121100236

University College Cork

School of Microbiology, University College Cork, Cork, Ireland

Teagasc Food Research Centre, Moorepark, Fermoy, Co. Cork, Ireland

Supervisor(s): Prof. Paul Cotter, Prof. Douwe Van Sinderen, Dr Conor Feehily and Dr Calum Walsh

Head of School/Department: Dr David Clarke

February 2025

Table of contents

Declaration

This is to certify that the work I am submitting is my own and has not been submitted for another degree, either at University College Cork or elsewhere. All external references and sources are clearly acknowledged and identified within the contents. I have read and understood the regulations of University College Cork concerning plagiarism and intellectual property.

Signature: 

Date: 05-02-2025

Student number: 121100236

Acknowledgements

I would first like to express my deepest gratitude to my supervisor, Prof. Paul Cotter, for graciously accepting me as his student and for offering constant support throughout my PhD. His guidance and encouragement have been invaluable, and I am truly grateful for his belief in me and my work. I also want to extend a heartfelt thank you to Dr Conor Feehily, whose supervision, patience, and unwavering push kept me moving forward, and to Dr Calum Walsh for providing insightful feedback even at late hours, despite the challenges posed by being in a different time zone in Australia. I am equally indebted to Prof. Douwe van Sinderen, whose encyclopaedic knowledge and willingness to clarify my questions in great detail were instrumental in expanding my understanding of the field. My supervisors made an extraordinary team, guiding me through the many ups and downs of this four-year journey and enabling me to reach its conclusion.

I would also like to thank Prof. Remco Kort for welcoming me as an exchange student in his laboratory in Amsterdam. This experience provided me with exposure to different projects and allowed me to connect with researchers from diverse backgrounds. I am grateful to my funding bodies, Science Foundation Ireland – Genomics Centre for Research Training and Marie Skłodowska-Curie Actions, for their financial support throughout my doctoral studies.

A special thank you goes out to all the Vision 1 colleagues for their assistance and the many conversations and walks that enriched my life. My friends have also been a constant source of strength, making the past four years smoother and more enjoyable through their companionship and support.

I am profoundly thankful to God for being my guiding light during this challenging period. Without Krishna, I would not have found the resilience to continue when times were tough. Lastly, I would like to extend my deepest thanks to my mother for her unwavering support, for laying the strong foundation that brought me to this point, and for always being there when I needed guidance and reassurance. Each of these individuals has played a vital role in my doctoral journey, and I owe my success to their generous help and encouragement.

Abstract

During the perinatal period, the transfer of microbial communities from mother to infant is critical in shaping the infant's early-life microbiome and broader health outcomes. While it is well established that maternal sources such as the gut, breastmilk, and vaginal tract each contribute to the neonate's microbiota, fewer studies have examined the fine-scale strain-level differences that can govern colonisation success or influence pregnancy outcomes. This thesis addresses that gap by adopting metagenomic approaches to identify and compare specific bacterial strains across mother-infant pairs, offering novel perspectives on how functional traits may underpin microbe-host interaction and vertical transmission. Chapter 2 highlights how computational choices can markedly affect the reconstruction of metagenome-assembled genomes (MAGs) and, in turn, our interpretations of microbial sharing events. Comparing different binning tools demonstrates that method selection shapes both the number and quality of MAGs recovered, including the detection of critical genes like 16S rRNA and antimicrobial resistance determinants. These findings underscore the necessity of rigorous bioinformatic pipelines for accurately tracing maternal strains into infants. Chapter 3 highlights how functional traits, particularly genes involved in human milk oligosaccharide (HMO) metabolism may confer a competitive advantage in the HMO-rich infant gut. Such insights illuminate why some maternal strains flourish in early infancy even when they are not dominant in the maternal microbiome. Chapters 4 and 5 focus on the maternal vaginal microbiome, a niche strongly linked to pregnancy outcomes but relatively understudied at the microbial strain level. Investigating *Lactobacillus* strains and their genetic variation reveals that even minor within-species distinctions correlate with gestational duration, offering novel insights on how vaginal microbiome composition may influence pregnancy risks. Taken together, these results underscore the critical importance of examining microbial communities at the strain level to gain a more refined view of mother-infant transmission. Ultimately, these findings open avenues for targeted investigation to understand the mechanisms governing maternal/infant microbial interaction and to tailor interventions for improved health outcomes during the perinatal period.

Publications

1. Nori, S. R. C. *et al.* Profiling of vaginal *Lactobacillus jensenii* isolated from preterm and full-term pregnancies reveals strain-specific factors relating to host interaction. *Microbial Genomics* 9, 001137 (2023).
2. Nori, S. R. C. *et al.* Strain-level variation among vaginal *Lactobacillus crispatus* and *Lactobacillus iners* as identified by comparative metagenomics. *npj Biofilms Microbiomes* 11, 1–11 (2025).

Abbreviations

1M	1 month
1W	1 week
3M	3 months
ABC	ATP-binding cassette
ABX	Antibiotics
ackA	Acetate kinase
AMR	Antimicrobial resistance
ANI	Average nucleotide identity
BLAST	Basic local alignment search tool
BV	Bacterial vaginosis
CRT	Centre for Research Training
CST	Community state type
CW_MucBP	Cell wall anchor motif
DEL	Delivery
DO	Delivery outcome
E	Early
ENA	European Nucleotide Archive
ET	Ethiopian
FTB	Full-term birth
GH	Glycosyl hydrolase
Gram_pos_anchor	Gram positive cell wall anchor
GTs	Glycosyltransferases
HMO	Human milk oligosaccharide
HQ	High-quality
iTOL	Interactive Tree Of Life
L	Late
ldhD	D-lactate dehydrogenase
LLETZ	Large loop excision of the transformation zone
MAGs	Metagenome-assembled genomes
Mbp	Megabase pairs
MFS	Mahor facilitator superfamily
mgCST	Metagenomic community state types
mgSs	Metagenomic subspecies
MISAG	Minimum information about a single amplified genome
MM	MicrobeMom
MQ	Medium quality
MRS	DeMan Rogosa Sharpe
MucBP	Mucin binding domain
PCoA	Principal coordinate analysis
PTB	Preterm birth

RF	Robinson-Foulds
rRNA	Ribosomal RNA
SFI	Science Foundation Ireland
SLAP	S-layer associated protein
SROM	Spontaneous rupture of membranes
SWE	Swedish
VOGs	Vaginal orthologous groups

1 Chapter 1

Introduction

1.1 Introduction

The human microbiome comprises a vast collection of microorganisms including bacteria, viruses and fungi that reside in and on the human body[1–3]. These complex and dynamic microbial communities play essential roles in human physiology, extending far beyond digestion to immune system regulation, metabolism, and neurodevelopment[4,5]. Importantly, environmental factors such as diet, antibiotic use, stress, and lifestyle choices can influence microbial composition, underscoring the delicate balance and adaptability of host–microbe interactions[6–9].

The microbiome exhibits remarkable site-specific diversity across the human body, each with its own characteristic microbial consortia[1,10]. For instance, the skin microbiome is typically dominated by genera such as *Staphylococcus, Corynebacterium,* and *Cutibacterium*, which have adapted to the skin's acidic pH and relatively low moisture levels[11,12] and the oral cavity is highly colonised by genera including *Streptococcus*, *Veillonella*, and *Prevotella*[13–15]. In the vaginal microbiome, *Lactobacillus* species dominate the environment producing lactic acid that maintains a low pH and protects against pathogenic invasion[16,17]. The gut microbiome is by far the most extensive, containing trillions of microbes from diverse phyla such as *Bacillota*, *Bacteroidota*, *Actinomycetota*, and *Pseudomonadota* [1,18,19]. Disruptions in gut microbial communities have been linked to wide-ranging conditions, including inflammatory bowel disease, obesity, diabetes, and various neurodevelopmental disorders[20–24].

Historically, microbial research relied heavily on culture-based techniques that focused on isolating and identifying individual bacterial species. This limited our understanding of microbial communities that were difficult to cultivate under laboratory conditions. With the emergence of high-throughput sequencing, particularly 16S rRNA gene amplicon-based sequencing and shotgun metagenomics, entire microbial ecosystems can be studied uncovering taxa that were previously missed by cultivation[25,26]. These molecular approaches allow for comprehensive characterisation of microbial diversity, community structure, and functional potential, illuminating how microbiomes develop, maintain health, and contribute to disease processes across the human lifespan.

The perinatal period, encompassing pregnancy, delivery, and the early postnatal phase is pivotal for shaping the infant microbiome[27–29]. During pregnancy, the maternal microbiome undergoes notable shifts, such as increases in *Bifidobacterium* and *Akkermansia* in the gut driven by hormonal changes and reduced microbial diversity influenced by factors including diet, antibiotic use, and other lifestyle factors[30–34]. Emerging evidence indicates that the maternal gut microbiome is a major contributor to the infant's initial microbial repertoire[35,36]. Recent high-resolution strain-level analyses suggest that infants acquire *Bifidobacterium* and *Bacteroides* species strains from maternal sources[37–39]. An improved understanding of this transmission process could aid in designing intervention strategies such as targeted probiotic supplementation or microbiome restoration protocols that are particularly valuable for high-risk groups, such as preterm or cesarean-delivered infants, who may not experience the typical mode of maternal microbial transfer[40–42]. Disruptions to microbial seeding have been associated with an increased risk of childhood conditions such as asthma, obesity, and metabolic disorders[43–45].

Under typical conditions, the infant gut microbiome undergoes a characteristic developmental trajectory. Over the first few weeks, the infant gut microbiome shifts from a predominance of facultative anaerobes (e.g., *Enterobacteriaceae*) to strict anaerobes, with *Bifidobacterium* often becoming dominant and *Bacteroides* remaining in lower abundance[46,47]. A major transition occurs as a result of weaning and the concomitant introduction of solid foods, when the relative abundance of *Bacteroides* and other adult-like taxa (e.g., *Ruminococcus*) increases, while *Bifidobacterium* levels may decline[48,49]. By approximately one year of age, the infant microbiome more closely resembles that of an adult in terms of phylogenetic composition and functional potential[50,51].

1.2 Maternal Gut Microbiome: A Major Reservoir for Infant Colonization

During pregnancy, the maternal gut microbiome undergoes considerable shifts driven by hormonal fluctuations, dietary patterns, and metabolic adaptations[52,53]. These changes may have a lasting influence on infant health, given that many of the microorganisms initially acquired by newborns originate from the maternal gut.

Elevated levels of hormones such as estrogen and progesterone during pregnancy have been associated with alterations in maternal gut microbial composition and diversity[53,54]. For instance, increased progesterone levels have been shown to stimulate growth of *Bifidobacterium*, a genus crucial for maintaining gut homeostasis and supporting early immune development[55]. Notably, *Bifidobacterium* species are among the most frequently shared microbes in mother–infant pairs and play a key role in immune maturation and early-life immunological imprinting[56].

Maternal diet is another critical determinant of gut microbiome composition, with high-fat or high-sugar diets commonly linked to reduced microbial diversity and a higher abundance of pro-inflammatory taxa[57,58]. Such obesogenic patterns can pose risks for both maternal and offspring metabolic health. In contrast, diets rich in fruits, vegetables, and other fiber sources are correlated with more favorable maternal gut microbiota profiles which can in turn be reflected in the infant's gut microbiome as early as two months postpartum[59].

Among the taxa that are most commonly found in both mothers and infants are *Bacteroides* spp., *Phocaeicola* spp., and *Bifidobacterium* spp[35,38,60]. These microbes are pivotal for the infant's gut development and immune system maturation[61–63]. Several studies have demonstrated that maternal strains of these bacteria may persist in the infant gut for months to years, with vertical transmission efficiency varying among species[35,36,38,60,64]. Indeed, it has been estimated that a substantial proportion (40-70%) of the infant's gut microbiome can be traced back to maternal sources, highlighting the maternal gut as a primary reservoir for microbial colonization[46,65]. However, it is important to note that these estimates are largely

based on sequence-based strain-tracking approaches, which are prone to false positives due to limited resolution in distinguishing closely related strains.

Although species from these genera are frequently transferred, strain-level differences can significantly affect which variants successfully establish in the infant gut. For example, in the case of *Bacteroides uniformis,* a strain that is not the most dominant representative of the species in the mother may become dominant in the infant due to possessing advantageous genetic features, such as a starch utilization cluster[38]. Furthermore, differences in *Bifidobacterium longum* between adult (*B. longum* subsp*. longum*) and infant (*Bifidobacterium longum* subsp*. infantis*) strains were reported, with those isolated from infants being capable of utilising human milk oligosaccharides whereas adult-derived strains were better adapted to fermenting plant oligosaccharides[19]. Moreover, it has been highlighted that maternally derived strains generally persist longer in infants compared to those acquired from other sources, underscoring the importance of maternal-infant microbial exchange[35].

High-resolution strain-tracking techniques have revealed complex patterns of microbial persistence and competition. Lou et al. demonstrated that bacterial strains exhibiting high persistence in infant guts, defined as those maintained across longitudinal samples (n = 59), were characterised by the enrichment of genes associated with carbohydrate metabolism, adhesion mechanisms, and iron uptake systems[66]. Certain maternal strains become established in the infant gut, while others are gradually outcompeted by strains from the environment or other external sources[37,67,68]. This selective transfer and persistence of specific strains may reflect an evolutionary adaptation that favors functionally beneficial microbes, especially those adept at metabolising breast milk-derived compounds or modulating infant immune responses in early life[56,67]. Together, these findings highlight the maternal gut microbiome as a central reservoir for the infant's initial microbial seeding.

1.3 Role of breastmilk in infant microbiome development

Breastmilk serves as a critical source of bioactive compounds that can shape the early infant gut ecosystem. Its microbial composition is modulated by a range of maternal factors, including diet, health status, and mode of delivery[69,70]. Human milk oligosaccharides (HMOs), in particular, act as selective prebiotics by promoting the growth of beneficial bacteria such as *Bifidobacterium* while inhibiting pathogenic species[71–73].

Recent investigations have revealed diverse bacterial communities in breastmilk, often dominated by *Staphylococcus*, *Streptococcus*, and *Corynebacterium*, with lower levels of *Bifidobacterium* compared to other taxa[74,75]. Although the composition of breastmilk can shift during the transition from colostrum to mature milk[76,77], Mastromatina *et al*. reported no significant differences in the counts of *Lactobacillus* or *Bifidobacterium* between these two stages[78]. Other studies likewise suggest that the breastmilk microbiome remains relatively stable over time[79–81]. Notably, one study reported that up to 27% of the infant gut microbiota during the first month of life can be traced back to breastmilk, though it should be noted that this study was conducted using 16S rRNA amplicon sequencing which limits the taxonomic resolution to species level[81]. Specific strain-level transmissions have also been documented. For instance, Benito *et al*. reported that *Staphylococcus aureus* strains were transferred from mothers to infants in 4 out of 21 cases[82]. Furthermore, Makino et al. used multi-locus sequence typing to investigate transmission of *Bifidobacterium* species and identified strain-level matches in 16 of 18 mother-infant pairs[83]. Specifically, they observed identical strains in 3 of 18 pairs for *B. bifidum*, 11 of 18 for *B. breve*, and 4 of 18 for *B. longum subsp. longum*. Although similar *Lactobacillus* phylotypes have been detected in both breastmilk and infant feces, evidence for *Bifidobacterium* phylotype matching appears less consistent[84,85]. Most comparative analyses of breastmilk and infant gut microbiota focus on overall community profiles rather than fine-scale strain resolution[74,86,87].

Breastfeeding itself substantially influences the composition and diversity of the infant gut microbiome. Ma *et al*. found that exclusively breast-fed infants had lower alpha diversity at 40 days of age than formula-fed counterparts but significantly

higher abundances of *Bifidobacterium* and *Bacteroides*, alongside reduced *Streptococcus* and *Enterococcus*[88]. In cases where cesarean-delivered infants exhibit *Bifidobacterium*-depleted profiles, breastfeeding can partially restore these communities[89]. This restorative effect is likely mediated through the provision of HMOs, which selectively enrich beneficial bacterial taxa. Despite recent progress, further research employing high-resolution strain-level tracking is needed to illuminate the specific mechanisms and evolutionary pressures that govern microbial transfer via breastmilk, as well as the long-term health implications for the child.

1.4 Maternal Vaginal Microbiome and Its Contribution to Infant Gut Colonisation

The maternal vaginal microbiome is characterised by its predominance of *Lactobacillus* species - most commonly *Lactobacillus crispatus, Lactobacillus iners, Lactobacillus gasseri*, and *Lactobacillus jensenii* - which help maintain a low pH and protect against pathogenic invasion[16,17,90]. Although this community has been hypothesised to contribute to infant gut colonisation in vaginal deliveries, evidence so far suggests that the extent of vaginal-to-gut microbial transmission is relatively modest and often transient[35,36,91,92]. For example, Ferretti *et al*. reported that 16.3% of the infant gut microbiome at delivery could be traced to maternal vaginal sources, with *Lactobacillus* species dominating these transferred taxa; however, these organisms were largely lost by day four of life[35]. Mitchell *et al*. found no significant evidence of transmission from the maternal vaginal microbiota to the infant gut, with only 1 out of 75 mother-infant pairs sharing a *Lactobacillus crispatus* strain at the 1-week timepoint[92]. Similarly, our own findings indicate that out of 135 strain-level transmissions, only 8 strains were transmitted from the maternal vaginal tract to the infant gut, comprising 6 strains of *Bifidobacterium breve*, 1 strain of *Lactobacillus gasseri*, and 1 strain of *Lactobacillus paragasseri*[36]. More recently, Santos *et al*. showed that the composition of the maternal vaginal microbiome at delivery does not notably affect the infant stool microbiome, nor does it significantly alter the trajectory of microbial development in the early postnatal period[91].

While some *Lactobacillus* and *Bifidobacterium* strains that transfer from the vaginal environment can persist, or at least be transiently present in the infant gut, these species generally do not establish stable, long-term communities compared to gut-derived maternal strains. Despite this, delivery mode can exert a marked influence on the number and type of maternally transferred strains, particularly in the early days of life[38,92]. Cesarean (C-) section, for instance, reduces direct exposure to vaginal microbes and may thus alter initial colonisation patterns. Further research is needed to elucidate the specific selective pressures ranging from nutrient availability in the infant gut to local immune responses that determine whether and how vaginal microbes modulate the neonatal microbiome.

1.5 Maternal Oral and Skin Microbiomes: Additional Routes of Microbial Transmission

Although the maternal gut is often viewed as the primary reservoir for early infant microbial seeding, maternal oral and skin microbiomes also contribute to shaping the neonatal microbiota. In a study of 25 mother-infant pairs, Ferretti *et al*. reported that, on average, 11.2% and 5.8% of the infant gut microbiota at one and three days postpartum comprised species also found in the maternal oral and skin microbiomes, respectively, compared to 22.1% from the maternal gut[35]. However, these figures represent the percentage of shared species between mother and infant, not strain-level transmissions, and are limited to very early postnatal timepoints. Kageyama *et al*., using 16S rRNA gene sequencing, reported similar proportions of oral species sharing (approximately 9.7%).[93] Oral colonisation in infants begins at birth, coinciding with the rupture of the amniotic membrane, and is typically dominated by *Streptococcus* and *Staphylococcus*[94,95].

Recent findings suggest that the similarity between maternal and infant oral microbiomes can increase over the first six months, potentially influenced by dietary changes and the eruption of primary teeth[96]. Feeding practices also seem to shape oral microbial similarity; Kageyama *et al*. reported less resemblance between breastfed infants and their mothers compared to formula-fed

infants[93]. *Streptococcus*, *Staphylococcus*, *Gemella*, *Rothia*, and *Veillonella* are among the most commonly shared genera in breastmilk and maternal-child oral microbiomes[97]. However, in our previous study, we detected only minimal (1 out of 135) strain-level transmission (*Veillonella parvula*) between mother and infant at 1-month, indicating that instances of oral microbiome transfer are relatively rare[36].

Maternal skin also represents a potential source of infant microbial colonisation. At birth, the neonate's skin is seeded primarily by strains belonging to *Staphylococcus* and *Lactobacillus*[98–100], which are gradually replaced by *Acinetobacter*, *Streptococcus*, *Veillonella*, *Bifidobacterium*, and *Enhydrobacter* as the infant ages[101]. Even in later childhood, skin microbial communities in specific sites remain more similar to those of the mother than to unrelated individuals[102]. In addition, the areolar skin microbiome influences the composition of human milk, and an estimated 10.4% of breastmilk bacterial communities may derive from the areolar skin during the first month of lactation[81].

Various factors including antibiotic exposure, delivery mode, and breastfeeding practices can modulate this early skin colonisation process and have long-term health consequences[101,103]. Some studies indicate that prenatal and postnatal probiotic supplementation can reduce the risk of pediatric atopic dermatitis[104], although such supplementation does not necessarily benefit established atopic dermatitis cases[105]. These findings highlight the significance of the initial seeding window and its potential influence on the gut–skin axis. Taken together, there is evidence of microbial overlap between maternal oral and skin niches and the infant microbiome, although much of the existing data is based on 16S rRNA gene amplicon sequencing, which limits taxonomic resolution. Therefore, strain-level metagenomic analyses are needed to establish definitive links.

1.6 Factors Influencing Maternal-Infant Strain Transmission

The initial seeding of maternal microbes in infants is influenced by various factors, including delivery mode, antibiotic exposure, feeding practices, maternal diet, lifestyle, and environmental conditions. Studies have repeatedly shown that delivery mode and antibiotic use are among the most significant determinants of mother-infant strain transmission[35,36,39,47,106]. Disruption of this early colonisation window can have lasting implications for infant immune maturation and metabolic programming[107,108].

1.6.1 Delivery Mode

Vaginally delivered infants are exposed to the maternal birth canal and perianal regions, leading to higher initial microbial diversity compared to cesarean delivered infants[109]. In these vaginal deliveries, *Bacteroides* spp., *Phocaeicola* spp., and *Bifidobacterium* spp. are frequently transferred from mother to infant, whereas C-section delivered infants commonly acquire *Staphylococcus epidermidis*, *Klebsiella pneumoniae*, *Bacteroides fragilis*, and *Escherichia coli*. Shao *et al*. found that C-section delivered infants inherit approximately 12.56% (vs 74.39% in vaginally delivered infants) of their microbiota from maternal strains, though the transmission of key taxa in the phylum Bacteroidetes is substantially impaired in these neonates[64].

Multiple studies confirm that vaginally delivered neonates initially harbor more *Lactobacillus* and *Bacteroides*, whereas such taxa are underrepresented in C-section infants [110,111]. Although significant microbial differences diminish by one month postpartum in some cohorts observed persistent distinctions at one year of age in the cumulative distribution of *Lactobacillus*, *Prevotella*, and *Sneathia* between infants born vaginally and those delivered by C-section[69,109]. Introduction of solid foods eventually reduces these disparities. Notably, “vaginal seeding” interventions where C-section infants are swabbed with vaginal fluids can partially normalize the microbiome trajectories in these infants[112].

This mode of delivery-associated variations extend beyond the gut. Vaginally delivered infants often develop skin microbiota that more closely resembles their mother’s vaginal community, whereas infants born by C-section display skin bacterial communities similar to maternal skin[100]. Nonetheless, these delivery-mode

signatures tend to wane by three months of age[113]. Given the immunomodulatory roles of *Bacteroides* and *Bifidobacterium*[56,114], examining how perturbations in early life alter long-term health outcomes remains a critical area for future research.

1.6.2 Antibiotic Exposure

Maternal antibiotic use during pregnancy and the perinatal period can alter the maternal microbiome in ways that significantly affect infant colonization[115,116]. In particular, prenatal or intrapartum antibiotics can disrupt the vaginal microbiota, thereby reducing the transfer of beneficial taxa to the neonate[117,118]. Specific antibiotic regimens may exert differential effects; for instance, certain types of antibiotics have been linked to a reduction in *Bifidobacterium* and *Lactobacillus*[119,120].

Beyond immediate compositional shifts, these alterations can have lasting consequences. Patangia *et al*. reported that infants can acquire maternal and environmental antibiotic resistance genes (ARGs) even before they are personally exposed to antibiotics[9]. Strain transmission is often reduced when mothers receive antibiotics during delivery, with lower transfers of *Bacteroides* and *Bifidobacterium*[36]. Li *et al*. found that 72% of the infant gut microbiome could be traced back to maternal sources in control groups, compared to only 25% in infants whose mothers received antibiotics[106]. Furthermore, early exposure to antibiotics can have long-lasting effects, which are significant even at 12 months after birth with more species persisting in the control group (21 species) compared to the antibiotic-treated group (2 species). Zou *et al*. reported decreased levels of *Bacteroides* and *Bifidobacterium* in the antibiotic-treated group and this altered microbiota resembled the resistant bacteria found in the neonatal intensive care unit[121]. These findings underscore the substantial influence of delivery mode and antibiotic exposure on the vertical transfer of strains from mother to infant.

1.7 Computational Approaches to Investigate Strain Sharing in Mother–Infant Microbiomes

High-throughput sequencing has revolutionised our ability to study microbial communities at fine-scale resolution enabling to track specific strains across a wide range of settings from food microbiomes to pathogen detection in hospitals[26,37,122,123]. Traditionally, microbiologists have relied on isolated bacterial genomes obtained via cultivation or single colony sequencing to identify and characterise microbial species and strains. However, such approaches may miss a substantial fraction of the community, especially those microbes that are difficult to culture. By contrast, metagenomic sequencing (shotgun sequencing of all microbial DNA in a sample) captures a broader range of organisms, including those not culturable with standard techniques. This approach is especially relevant for studying maternal–infant microbiomes, which typically exhibit a high diversity of taxa that may vary considerably among individuals and across body sites[1].

Shotgun metagenomic sequencing generates short DNA reads typically ranging from 150-300 base pairs in length representing the genome content of the whole community. These reads are further analysed using various downstream methods. With read-based approaches, we can directly analyse the DNA sequences to identify strain-level variations without the need for genome assembly. These methods typically involve mapping reads to reference genomes or species-specific marker genes to compare single nucleotide polymorphism (SNP) patterns. Tools like StrainPhlAn and PanPhlAn have emerged as popular choices for strain-level analysis using reads to track specific bacterial strains across multiple samples and determine their phylogenetic relationships within mother-infant pairs[124–126]. Such tools are less computationally intensive than assembly-based approaches and suited for widely studied microbiomes with reliable reference genomes and marker gene sets.

In assembly-based pipelines, sequencing reads are assembled into contiguous sequences (contigs) and then are binned and annotated using tools to identify genomic and functional differences at strain-level resolution. Common assemblers include MEGAHIT, SPAdes, and IDBA-UD that are developed to handle complex metagenomic datasets[127–129]. However, strains present at low-abundance or with

uneven coverage can be difficult to reconstruct in datasets from high-diversity samples such as gut microbiome. Furthermore, these contigs are grouped together based on the sequence composition and coverage patterns using binning tools like MetaBAT, MaxBin, or CONCOCT[130–132]. These bins, now called metagenome assembled genomes (MAGs), represent individual genomes or strains and can then be assessed for quality and completeness using tools like CheckM[133].

After reconstructing MAGs from metagenomic samples, these can be compared at strain-level resolution using genomic similarity metrics (average nucleotide identity - ANI) to compare the entire genomes in paired samples. Additional functional annotation of these genomes using tools like PROKKA or eggNOG-mapper provides insights into the metabolic capabilities of the strains. This comparative genomic analysis can reveal strain-specific features, including unique genes, metabolic pathways, and virulence factors. For example, certain HMO utilisation genes may be identified that could influence host-microbe interactions in the context of driving mother-infant strain sharing[134,135].

Strain diversity within a particular species can be comprehensively studied using pan genome approaches that analyse the core genome (genes present in all strains) and accessory genome (genes present in some strains) which can reveal functional adaptations and evolutionary relationships between strains. Tools like Roary and Panaroo can efficiently construct pan-genomes from multiple strains, identifying core and accessory genes that may contribute to colonisation success[136,137]. This analysis is particularly valuable for understanding how different strains of the same species adapt to various host environments and how these adaptations might influence vertical transmission patterns between mothers and infants.

Furthermore, tools like InStrain have enabled us to profile and identify strain-level genetic variations across metagenomic samples with high precision and also for samples without reference genomes using MAGs[37]. This tool integrates population genetics metrics with genome-wide coverage analysis to detect subtle genomic differences between closely related strains. When applied to mother-infant pairs, InStrain can reveal fine-scale evolutionary dynamics and transmission patterns that

can reveal how bacterial populations adapt and diversify during vertical transfer from mother to child.

1.8 Conclusion

The mother-infant microbiome interface has emerged as a pivotal area of research in early-life health and development. Research so far have established that the maternal gut microbiome is the major reservoir for initial microbial seeding; however, other maternal niches including the vaginal, breastmilk, oral, and skin microbiomes contribute to the infant's microbiome development[35,36,38,93]. These diverse and sometimes transient microbial inputs highlight the complexity of vertical transmission and underscore the need to examine each route of transfer. It is well established that maternal and environmental factors such as delivery mode, antibiotic use, diet, and lifestyle can all modulate strain sharing, with potentially lifelong ramifications for infant health[9,59,111,119].

Several factors modulate the efficiency and persistence of vertical transmission. Depending on the delivery mode, whether vaginal or caesarean section, the infant's initial microbial exposure changes leading to distinct microbial profiles that may last for weeks or months[64,111]. Antibiotics use significantly diminish key taxa such as *Bifidobacterium* and *Bacteroides*, and foster expansion of resistant or opportunistic species[36,120,121]. Although many initial imbalances seem to diminish with age, certain variations in microbial composition and function may persist, potentially impacting long-term immune and metabolic health outcomes.

Technological advancements have substantially enhanced the resolution at which we can investigate these complex interactions. High-throughput sequencing, particularly shotgun metagenomics, allows examination of microbial DNA from clinical or environmental samples without the need for culturing[26]. Computational approaches now permit identification of strain-level variation to identify specific single nucleotide polymorphisms (SNPs) and track individual strains[37,126]. Furthermore, pangenomic methods offer deeper insights into core and accessory genes,

highlighting which genetic features might give certain strains a competitive edge in the infant gut[136,137].

Emerging evidence indicates that microbial functionality, rather than mere abundance, may shape vertical transmission patterns[38]. Strains equipped with advantageous traits such as the ability to utilise human milk oligosaccharides or adhere effectively to the gut epithelium may have a better chance of persisting in the infant's gut[66]. Understanding these functional traits is critical to developing targeted interventions for restoring optimal microbial profiles in vulnerable groups, including preterm or caesarean-delivered infants who may not experience normal microbial seeding.

Ultimately, the complex interplay between mother and infants, and external factors underscores the need for longitudinal investigations that combine high-resolution sequencing with well-characterised clinical data. This thesis aims to advance our understanding of these maternal–infant microbiome dynamics by (i) employing cutting-edge bioinformatic strategies to pinpoint strain-level transmission events, (ii) investigating functional genes that underpin the establishment and persistence of key microbial species, and (iii) examining how these factors correlate with pregnancy outcomes. In the ensuing chapters, we compare different bioinformatic approaches for metagenomic assembly and binning, investigate into carbohydrate utilisation and other functional determinants driving early colonisation in the infant gut, and examine whether specific vaginal microbial strain profiles are linked to gestational outcomes. Such research holds promise for informing strategies aimed at optimising microbial colonisation, whether through tailored probiotic supplementation, vaginal seeding, or other microbiome restoration efforts. By elucidating the drivers of successful maternal-infant transmission at the strain level, this work may pave the way for interventions that help protect against adverse health outcomes and enhance overall well-being in early life.

1.9 References

1. Huttenhower, C. *et al.* Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–214 (2012).

2. Mousa, W. K., Chehadeh, F. & Husband, S. Recent Advances in Understanding the Structure and Function of the Human Microbiome. *Front. Microbiol.* **13**, (2022).

3. Gilbert, J. *et al.* Current understanding of the human microbiome. *Nat Med* **24**, 392–400 (2018).

4. Zheng, D., Liwinski, T. & Elinav, E. Interaction between microbiota and immunity in health and disease. *Cell Res* **30**, 492–506 (2020).

5. Dash, S., Syed, Y. A. & Khan, M. R. Understanding the Role of the Gut Microbiome in Brain Development and Its Association With Neurodevelopmental Psychiatric Disorders. *Front Cell Dev Biol* **10**, 880544 (2022).

6. Madison, A. & Kiecolt-Glaser, J. K. Stress, depression, diet, and the gut microbiota: human–bacteria interactions at the core of psychoneuroimmunology and nutrition. *Curr Opin Behav Sci* **28**, 105–110 (2019).

7. Clarke, S. F. *et al.* Exercise and associated dietary extremes impact on gut microbial diversity. *Gut* **63**, 1913–1920 (2014).

8. Singh, R. K. *et al.* Influence of diet on the gut microbiome and implications for human health. *J Transl Med* **15**, 73 (2017).

9. Patangia, D. V., Anthony Ryan, C., Dempsey, E., Paul Ross, R. & Stanton, C. Impact of antibiotics on the human microbiome and consequences for host health. *Microbiologyopen* **11**, e1260 (2022).

10. Zhou, X. *et al.* Longitudinal profiling of the microbiome at four body sites reveals core stability and individualized dynamics during health and disease. *Cell Host & Microbe* **32**, 506-526.e9 (2024).

11. Byrd, A. L., Belkaid, Y. & Segre, J. A. The human skin microbiome. *Nat Rev Microbiol* **16**, 143–155 (2018).

12. Yang, Y., Qu, L., Mijakovic, I. & Wei, Y. Advances in the human skin microbiota and its roles in cutaneous diseases. *Microbial Cell Factories* **21**, 176 (2022).

13. Baker, J. L., Mark Welch, J. L., Kauffman, K. M., McLean, J. S. & He, X. The oral microbiome: diversity, biogeography and human health. *Nat Rev Microbiol* **22**, 89–104 (2024).

14. Deo, P. N. & Deshmukh, R. Oral microbiome: Unveiling the fundamentals. *J Oral Maxillofac Pathol* **23**, 122–128 (2019).

15. Willis, J. R. & Gabaldón, T. The Human Oral Microbiome in Health and Disease: From Sequences to Ecosystems. *Microorganisms* **8**, 308 (2020).

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

17. Amabebe, E. & Anumba, D. O. C. The vaginal microenvironment: The physiologic role of Lactobacilli. *Frontiers in Medicine* **5**, 181 (2018).

18. Hou, K. *et al.* Microbiota in health and diseases. *Sig Transduct Target Ther* **7**, 1–28 (2022).

19. Lozupone, C. A., Stombaugh, J. I., Gordon, J. I., Jansson, J. K. & Knight, R. Diversity, stability and resilience of the human gut microbiota. *Nature* **489**, 220–230 (2012).

20. Halfvarson, J. *et al.* Dynamics of the human gut microbiome in inflammatory bowel disease. *Nat Microbiol* **2**, 1–7 (2017).

21. Shan, Y., Lee, M. & Chang, E. B. The Gut Microbiome and Inflammatory Bowel Diseases. *Annual review of medicine* **73**, 455 (2021).

22. Davis, C. D. The Gut Microbiome and Its Role in Obesity. *Nutr Today* **51**, 167–174 (2016).

23. Geng, J., Ni, Q., Sun, W., Li, L. & Feng, X. The links between gut microbiota and obesity and obesity related diseases. *Biomedicine & Pharmacotherapy* **147**, 112678 (2022).

24. Sadagopan, A. *et al.* Understanding the Role of the Gut Microbiome in Diabetes and Therapeutics Targeting Leaky Gut: A Systematic Review. *Cureus* **15**, e41559.

25. Wensel, C. R., Pluznick, J. L., Salzberg, S. L. & Sears, C. L. Next-generation sequencing: insights to advance clinical investigations of the microbiome. *J Clin Invest* **132**, e154944.

26. Kim, N. *et al.* Genome-resolved metagenomics: a game changer for microbiome medicine. *Exp Mol Med* **56**, 1501–1512 (2024).

27. Catassi, G. *et al.* The importance of gut microbiome in the perinatal period. *Eur J Pediatr* **183**, 5085–5101 (2024).

28. Sinha, T., Brushett, S., Prins, J. & Zhernakova, A. The maternal gut microbiome during pregnancy and its role in maternal and infant health. *Current Opinion in Microbiology* **74**, 102309 (2023).

29. Dunlop, A. L. *et al.* The Maternal Microbiome and Pregnancy Outcomes that Impact Infant Health: A Review. *Adv Neonatal Care* **15**, 377–385 (2015).

30. Demmelmair, H., Jiménez, E., Collado, M. C., Salminen, S. & McGuire, M. K. Maternal and Perinatal Factors Associated with the Human Milk Microbiome. *Current Developments in Nutrition* **4**, nzaa027 (2020).

31. Wang, S. *et al.* A good start in life is important—perinatal factors dictate early microbiota development and longer term maturation. *FEMS Microbiology Reviews* **44**, 763–781 (2020).

32. Ryan, N., Leahy-Warren, P., Mulcahy, H., O'Mahony, S. & Philpott, L. The impact of perinatal maternal stress on the maternal and infant gut and human milk microbiomes: A scoping review protocol. *PLoS One* **19**, e0304787 (2024).

33. Gorczyca, K., Obuchowska, A., Kimber-Trojnar, Ż., Wierzchowska-Opoka, M. & Leszczyńska-Gorzelak, B. Changes in the Gut Microbiome and Pathologies in Pregnancy. *International Journal of Environmental Research and Public Health* **19**, 9961 (2022).

34. SMID, M. C. *et al.* Maternal gut microbiome biodiversity in pregnancy. *Am J Perinatol* **35**, 24–30 (2018).

35. Ferretti, P. *et al.* Mother-to-Infant Microbial Transmission from Different Body Sites Shapes the Developing Infant Gut Microbiome. *Cell Host and Microbe* **24**, 133-145.e5 (2018).

36. Feehily, C. *et al.* Detailed mapping of Bifidobacterium strain transmission from mother to infant via a dual culture-based and metagenomic approach. *Nat Commun* **14**, 3015 (2023).

37. Olm, M. R. *et al.* inStrain profiles population microdiversity from metagenomic data and sensitively detects shared microbial strains. *Nat Biotechnol* **39**, 727–736 (2021).

38. Yassour, M. *et al.* Strain-Level Analysis of Mother-to-Child Bacterial Transmission during the First Few Months of Life. *Cell Host and Microbe* **24**, 146-154.e4 (2018).

39. Yang, R. *et al.* Inspecting mother-to-infant microbiota transmission: disturbance of strain inheritance by cesarian section. *Front. Microbiol.* **15**, (2024).

40. Ríos-Covian, D., Langella, P. & Martín, R. From Short- to Long-Term Effects of C-Section Delivery on Microbiome Establishment and Host Health. *Microorganisms* **9**, 2122 (2021).

41. Kim, G. *et al.* Delayed Establishment of Gut Microbiota in Infants Delivered by Cesarean Section. *Front. Microbiol.* **11**, (2020).

42. Cuna, A., Morowitz, M. J., Ahmed, I., Umar, S. & Sampath, V. Dynamics of the preterm gut microbiome in health and disease. *Am J Physiol Gastrointest Liver Physiol* **320**, G411–G419 (2021).

43. Alsharairi, N. A. The Infant Gut Microbiota and Risk of Asthma: The Effect of Maternal Nutrition during Pregnancy and Lactation. *Microorganisms* **8**, 1119 (2020).

44. Tun, H. M. *et al.* Roles of Birth Mode and Infant Gut Microbiota in Intergenerational Transmission of Overweight and Obesity From Mother to Offspring. *JAMA Pediatrics* **172**, 368–377 (2018).

45. Fujimura, K. E. *et al.* Neonatal gut microbiota associates with childhood multisensitized atopy and T cell differentiation. *Nat Med* **22**, 1187–1191 (2016).

46. Enav, H., Bäckhed, F. & Ley, R. E. The developing infant gut microbiome: A strain-level view. *Cell Host and Microbe* **30**, 627–638 (2022).

47. Browne, H. P., Shao, Y. & Lawley, T. D. Mother–infant transmission of human microbiota. *Current Opinion in Microbiology* **69**, 102173 (2022).

48. Laursen, M. F., Bahl, M. I., Michaelsen, K. F. & Licht, T. R. First Foods and Gut Microbes. *Front. Microbiol.* **8**, (2017).

49. Zafar, H. & Saier, M. H. Gut Bacteroides species in health and disease. *Gut Microbes* **13**, 1848158.

50. Yao, Y. *et al.* The Role of Microbiota in Infant Health: From Early Life to Adulthood. *Front Immunol* **12**, 708472 (2021).

51. Avershina, E. *et al.* Transition from infant- to adult-like gut microbiota. *Environ Microbiol* **18**, 2226–2236 (2016).

52. Koren, O., Konnikova, L., Brodin, P., Mysorekar, I. U. & Collado, M. C. The maternal gut microbiome in pregnancy: implications for the developing immune system. *Nat Rev Gastroenterol Hepatol* **21**, 35–45 (2024).

53. Edwards, S. M., Cunningham, S. A., Dunlop, A. L. & Corwin, E. J. The Maternal Gut Microbiome during Pregnancy. *MCN Am J Matern Child Nurs* **42**, 310–317 (2017).

54. Jiang, G., Zhou, Z., Li, X., Qian, Y. & Wang, K. The Gut Microbiome During Pregnancy. *Maternal-Fetal Medicine* **5**, 36 (2023).

55. Nuriel-Ohayon, M. *et al.* Progesterone Increases *Bifidobacterium* Relative Abundance during Late Pregnancy. *Cell Reports* **27**, 730-736.e3 (2019).

56. Henrick, B. M. *et al.* Bifidobacteria-mediated immune system imprinting early in life. *Cell* **184**, 3884-3898.e11 (2021).

57. Ratsika, A. *et al.* Maternal high-fat diet-induced microbiota changes are associated with alterations in embryonic brain metabolites and adolescent behaviour. *Brain, Behavior, and Immunity* **121**, 317–330 (2024).

58. Di Gesù, C. M. *et al.* Maternal gut microbiota mediate intergenerational effects of high-fat diet on descendant social behavior. *Cell Rep* **41**, 111461 (2022).

59. Fan, H.-Y. *et al.* Maternal Vegetable and Fruit Consumption during Pregnancy and Its Effects on Infant Gut Microbiome. *Nutrients* **13**, 1559 (2021).

60. Li, M. *et al.* Relationship between infant gastrointestinal microorganisms and maternal microbiome within 6 months of delivery. *Microbiology Spectrum* **12**, e03608-23 (2024).

61. Lin, C. *et al.* Intestinal 'Infant-Type' Bifidobacteria Mediate Immune System Development in the First 1000 Days of Life. *Nutrients* **14**, 1498 (2022).

62. Gavzy, S. J. *et al.* Bifidobacterium mechanisms of immune modulation and tolerance. *Gut Microbes* **15**, 2291164.

63. Tanaka, M. & Nakayama, J. Development of the gut microbiota in infancy and its impact on health in later life. *Allergology International* **66**, 515–522 (2017).

64. Shao, Y. *et al.* Stunted microbiota and opportunistic pathogen colonization in caesarean-section birth. *Nature* **574**, 117–121 (2019).

65. Korpela, K. & de Vos, W. M. Early life colonization of the human gut: microbes matter everywhere. *Current Opinion in Microbiology* **44**, 70–78 (2018).

66. Lou, Y. C. *et al.* Infant gut strain persistence is associated with maternal origin, phylogeny, and traits including surface adhesion and iron acquisition. *Cell Reports Medicine* **2**, 100393 (2021).

67. Vatanen, T. *et al.* The human gut microbiome in early-onset type 1 diabetes from the TEDDY study. *Nature* **562**, 589–594 (2018).

68. Wang, S. *et al.* Metagenomic analysis of mother-infant gut microbiome reveals global distinct and shared microbial signatures. *Gut Microbes* **13**, 1911571 (2021).

69. Bhattacharyya, C. *et al.* Influence of Maternal Breast Milk and Vaginal Microbiome on Neonatal Gut Microbiome: a Longitudinal Study during the First Year. *Microbiology Spectrum* **11**, e04967-22 (2023).

70. Kumar, H. *et al.* Distinct Patterns in Human Milk Microbiota and Fatty Acid Profiles Across Specific Geographic Locations. *Front. Microbiol.* **7**, (2016).

71. Salli, K. *et al.* Selective Utilization of the Human Milk Oligosaccharides 2'-Fucosyllactose, 3-Fucosyllactose, and Difucosyllactose by Various Probiotic and Pathogenic Bacteria. *J. Agric. Food Chem.* **69**, 170–182 (2021).

72. Marcobal, A. *et al.* Bacteroides in the Infant Gut Consume Milk Oligosaccharides via Mucus-Utilization Pathways. *Cell Host & Microbe* **10**, 507–514 (2011).

73. Kiely, L. J., Busca, K., Lane, J. A., van Sinderen, D. & Hickey, R. M. Molecular strategies for the utilisation of human milk oligosaccharides by infant gut-associated bacteria. *FEMS Microbiology Reviews* **47**, fuad056 (2023).

74. Kordy, K. *et al.* Contributions to human breast milk microbiome and enteromammary transfer of Bifidobacterium breve. *PLOS ONE* **15**, e0219633 (2020).

75. Edwards, C. A. *et al.* A systematic review of breast milk microbiota composition and the evidence for transfer to and colonisation of the infant gut. *Beneficial Microbes* **13**, 365–382 (2022).

76. Wan, Y. *et al.* Human milk microbiota development during lactation and its relation to maternal geographic location and gestational hypertensive status. *Gut Microbes* **11**, 1438–1449.

77. Xie, W., Zhang, H., Ni, Y. & Peng, Y. Contrasting Diversity and Composition of Human Colostrum Microbiota in a Maternal Cohort With Different Ethnic Origins but Shared Physical Geography (Island Scale). *Front. Microbiol.* **13**, (2022).

78. Mastromarino, P. *et al.* Correlation between lactoferrin and beneficial microbiota in breast milk and infant's feces. *Biometals* **27**, 1077–1086 (2014).

79. Hunt, K. M. *et al.* Characterization of the Diversity and Temporal Stability of Bacterial Communities in Human Milk. *PLOS ONE* **6**, e21313 (2011).

80. Simpson, M. R. *et al.* Breastfeeding-associated microbiota in human milk following supplementation with Lactobacillus rhamnosus GG, Lactobacillus

acidophilus La-5, and Bifidobacterium animalis ssp. lactis Bb-12. *Journal of Dairy Science* **101**, 889–899 (2018).

81. Pannaraj, P. S. *et al.* Association Between Breast Milk Bacterial Communities and Establishment and Development of the Infant Gut Microbiome. *JAMA Pediatrics* **171**, 647–654 (2017).

82. Benito, D. *et al.* Characterization of Staphylococcus aureus strains isolated from faeces of healthy neonates and potential mother-to-infant microbial transmission through breastfeeding. *FEMS Microbiology Ecology* **91**, fiv007 (2015).

83. Makino, H. *et al.* Multilocus sequence typing of bifidobacterial strains from infant's faeces and human milk: are bifidobacteria being sustainably shared during breastfeeding? *Benef Microbes* **6**, 563–572 (2015).

84. Zhang, X. *et al.* The Composition and Concordance of Lactobacillus Populations of Infant Gut and the Corresponding Breast-Milk and Maternal Gut. *Front. Microbiol.* **11**, (2020).

85. Yan, W., Luo, B., Zhang, X., Ni, Y. & Tian, F. Association and Occurrence of Bifidobacterial Phylotypes Between Breast Milk and Fecal Microbiomes in Mother–Infant Dyads During the First 2 Years of Life. *Front. Microbiol.* **12**, (2021).

86. Jost, T., Lacroix, C., Braegger, C. P., Rochat, F. & Chassard, C. Vertical mother-neonate transfer of maternal gut bacteria via breastfeeding. *Environ Microbiol* **16**, 2891–2904 (2014).

87. Tuzun, F., Kumral, A., Duman, N. & Ozkan, H. Breast Milk Jaundice. *Journal of Pediatric Gastroenterology and Nutrition* **56**, 328–332 (2013).

88. Ma, J. *et al.* Comparison of gut microbiota in exclusively breast-fed and formula-fed babies: a study of 91 term infants. *Sci Rep* **10**, 15792 (2020).

89. Guo, C. *et al.* Breastfeeding restored the gut microbiota in caesarean section infants and lowered the infection risk in early life. *BMC Pediatrics* **20**, 532 (2020).

90. Lebeer, S. *et al.* A citizen-science-enabled catalogue of the vaginal microbiome and associated factors. *Nat Microbiol* **8**, 2183–2195 (2023).

91. Dos Santos, S. J. *et al.* Maternal vaginal microbiome composition does not affect development of the infant gut microbiome in early life. *Front. Cell. Infect. Microbiol.* **13**, (2023).

92. Mitchell, C. M. *et al.* Delivery Mode Affects Stability of Early Infant Gut Microbiota. *Cell Reports Medicine* **1**, 100156 (2020).

93. Kageyama, S. *et al.* High-Level Acquisition of Maternal Oral Bacteria in Formula-Fed Infant Oral Microbiota. *mBio* **13**, e03452-21 (2022).

94. Wang, S. *et al.* Maternal Vertical Transmission Affecting Early-life Microbiota Development. *Trends in Microbiology* **28**, 28–45 (2020).

95. Nelson-Filho, P. *et al.* Dynamics of Microbial Colonization of the Oral Cavity in Newborns. *Braz. Dent. J.* **24**, 415–419 (2013).

96. Azevedo, M. J. *et al.* The contribution of maternal factors to the oral microbiota of the child: Influence from early life and clinical relevance. *Japanese Dental Science Review* **59**, 191–202 (2023).

97. Williams, J. E. *et al.* Strong Multivariate Relations Exist Among Milk, Oral, and Fecal Microbiomes in Mother-Infant Dyads During the First Six Months Postpartum. *The Journal of Nutrition* **149**, 902–914 (2019).

98. Schoch, J. J. *et al.* Skin microbiome sampling in the preterm neonate. *Pediatric Dermatology* **40**, 129–131 (2023).

99. Hellmann, K. T., Tuura, C. E., Fish, J., Patel, J. M. & Robinson, D. A. Viability-Resolved Metagenomics Reveals Antagonistic Colonization Dynamics of Staphylococcus epidermidis Strains on Preterm Infant Skin. *mSphere* **6**, 10.1128/msphere.00538-21 (2021).

100. Dominguez-Bello, M. G. *et al.* Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. *Proceedings of the National Academy of Sciences* **107**, 11971–11975 (2010).

101. Rapin, A. *et al.* The skin microbiome in the first year of life and its association with atopic dermatitis. *Allergy* **78**, 1949–1963 (2023).

102. Zhu, T. *et al.* Age and Mothers: Potent Influences of Children's Skin Microbiota. *Journal of Investigative Dermatology* **139**, 2497-2505.e6 (2019).

103. Dwyer, L. R. & Scharschmidt, T. C. Early life host-microbe interactions in skin. *Cell Host & Microbe* **30**, 684–695 (2022).

104. Li, L. *et al.* Probiotic Supplementation for Prevention of Atopic Dermatitis in Infants and Children: A Systematic Review and Meta-analysis. *Am J Clin Dermatol* **20**, 367–377 (2019).

105. Kong, H. H. Skin microbiome: genomics-based insights into the diversity and role of skin microbes. *Trends in Molecular Medicine* **17**, 320–328 (2011).

106. Li, W. *et al.* Vertical Transmission of Gut Microbiome and Antimicrobial Resistance Genes in Infants Exposed to Antibiotics at Birth. *J Infect Dis* **224**, 1236–1246 (2020).

107. Stokholm, J. *et al.* Delivery mode and gut microbial changes correlate with an increased risk of childhood asthma. *Science Translational Medicine* **12**, eaax9929 (2020).

108. Patrick, D. M. *et al.* Decreasing antibiotic use, the gut microbiota, and asthma incidence in children: evidence from population-based and prospective cohort studies. *The Lancet Respiratory Medicine* **8**, 1094–1105 (2020).

109. Akagawa, S. *et al.* Effect of Delivery Mode and Nutrition on Gut Microbiota in Neonates. *Ann Nutr Metab* **74**, 132–139 (2019).

110. Dominguez-Bello, M. G. *et al.* Partial restoration of the microbiota of cesarean-born infants via vaginal microbial transfer. *Nat Med* **22**, 250–253 (2016).

111. Zhang, C. *et al.* The Effects of Delivery Mode on the Gut Microbiota and Health: State of Art. *Front Microbiol* **12**, 724449 (2021).

112. Song, S. J. *et al.* Naturalization of the microbiota developmental trajectory of Cesarean-born neonates after vaginal seeding. *Med* **2**, 951-964.e5 (2021).

113. Capone, K. A., Dowd, S. E., Stamatas, G. N. & Nikolovski, J. Diversity of the Human Skin Microbiome Early in Life. *J Invest Dermatol* **131**, 2026–2032 (2011).

114. Mazmanian, S. K., Liu, C. H., Tzianabos, A. O. & Kasper, D. L. An Immunomodulatory Molecule of Symbiotic Bacteria Directs Maturation of the Host Immune System. *Cell* **122**, 107–118 (2005).

115. Zhou, P. *et al.* Perinatal Antibiotic Exposure Affects the Transmission between Maternal and Neonatal Microbiota and Is Associated with Early-Onset Sepsis. *mSphere* **5**, 10.1128/msphere.00984-19 (2020).

116. Wong, W. S. W. *et al.* Prenatal and Peripartum Exposure to Antibiotics and Cesarean Section Delivery Are Associated with Differences in Diversity and Composition of the Infant Meconium Microbiome. *Microorganisms* **8**, 179 (2020).

117. Dobbler, P. *et al.* The vaginal microbial communities of healthy expectant Brazilian mothers and its correlation with the newborn's gut colonization. *World J Microbiol Biotechnol* **35**, 159 (2019).

118. Stokholm, J., Sevelsted, A., Bønnelykke, K. & Bisgaard, H. Maternal propensity for infections and risk of childhood asthma: a registry-based cohort study. *The Lancet Respiratory Medicine* **2**, 631–637 (2014).

119. Azad, M. *et al.* Impact of maternal intrapartum antibiotics, method of birth and breastfeeding on gut microbiota during the first year of life: a prospective cohort study. *BJOG: An International Journal of Obstetrics & Gynaecology* **123**, 983–993 (2016).

120. Coker, M. *et al.* Specific class of intrapartum antibiotics relates to maturation of the infant gut microbiota: a prospective cohort study. *BJOG: An International Journal of Obstetrics & Gynaecology* **127**, 217–227 (2020).

121. Zou, Z.-H. *et al.* Prenatal and postnatal antibiotic exposure influences the gut microbiota of preterm infants in neonatal intensive care units. *Annals of Clinical Microbiology and Antimicrobials* **17**, 9 (2018).

122. Pasolli, E. *et al.* Large-scale genome-wide analysis links lactic acid bacteria from food with the gut microbiome. *Nat Commun* **11**, 2610 (2020).

123. Pham, V. D., Simpson, D. J. & Gänzle, M. G. Strain-level identification of bacteria persisting in food and in food processing facilities: when do two isolates

represent the same strain and which tools identify a strain? *Current Opinion in Food Science* **61**, 101245 (2025).

124. Blanco-Míguez, A. *et al.* Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. *Nat Biotechnol* **41**, 1633–1644 (2023).

125. Truong, D. T., Tett, A., Pasolli, E., Huttenhower, C. & Segata, N. Microbial strain-level population structure and genetic diversity from metagenomes. *Genome Res.* gr.216242.116 (2017) doi:10.1101/gr.216242.116.

126. Beghini, F. *et al.* Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. *eLife* **10**, e65088 (2021).

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

128. Prjibelski, A., Antipov, D., Meleshko, D., Lapidus, A. & Korobeynikov, A. Using SPAdes De Novo Assembler. *Current Protocols in Bioinformatics* **70**, e102 (2020).

129. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth | Bioinformatics | Oxford Academic. https://academic.oup.com/bioinformatics/article/28/11/1420/266973.

130. Kang, D. D. *et al.* MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* **7**, e7359 (2019).

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

132. Alneberg, J. *et al.* Binning metagenomic contigs by coverage and composition. *Nat Methods* **11**, 1144–1146 (2014).

133. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* (2015) doi:10.1101/gr.186072.114.

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

135. Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *bioRxiv* 2021.06.03.446934 (2021) doi:10.1101/2021.06.03.446934.

136. Page, A. J. *et al.* Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31**, 3691–3693 (2015).

137. Tonkin-Hill, G. *et al.* Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biology* **21**, 180 (2020).

2 Chapter 2

Metagenomic Binning Tool Selection Substantially Impacts Microbiome Strain-Sharing Inferences in Mother-Infant Pairs

This chapter has been submitted to Gut microbes journal and is currently under review.

2.1 Abstract

Advances in sequencing, and shotgun metagenomic sequencing in particular, have improved our understanding of the human microbiome, especially in the case of the mother-infant axis. An important aspect of this research involves the generation of high-quality metagenome-assembled genomes (MAGs), which are valuable for detailed analysis but can be challenging to recover. This study focused on comparing the effectiveness of two key binning tools, MetaBAT2 and MetaWRAP. We assessed their ability to recover MAGs from three distinct paired mother/infant metagenomic datasets and evaluated the impact on downstream analyses. We found that MetaWRAP generated more than double the number of high-quality MAGs compared to MetaBAT2, particularly for *Bacteroides* and *Phocaeicola* genera. A notable difference was found in the recovery of 16S rRNA genes, with only 6% of MetaBAT2 MAGs harbouring the gene versus 16.5% of MetaWRAP MAGs. Additionally, MAGs from MetaWRAP uncovered antimicrobial resistance (AMR) genes in species undetected by MetaBAT2, enabling more sensitive detection of AMRs pertaining to clinically significant antibiotic classes such as cephalosporins and tetracyclines. Importantly, using a reference genome-guided approach for profiling strain-sharing in mother-infant pairs revealed a 25% increase in detection of sharing events using MetaWRAP MAGs versus MetaBAT2. The results of this investigation underscore the significant impact of binning tool selection in metagenomic studies of the human microbiome. The observed superior performance of MetaWRAP in generating high-quality MAGs notably enhances our understanding of the intricacies of microbiome structure. Importantly, these findings highlight the potential for analytical biases due to tool selection and the need for standardized, robust methodologies in metagenomic research.

2.2 Introduction

Advances in sequencing technologies have led to a large expansion in human microbiome data. This, coupled with modern analysis tools, has profoundly enhanced our ability to explore microbiomes of relevance to both health and disease [1,2]. Shotgun metagenomics stands out as the preferred method for detailed taxonomic and functional analysis of microbial communities [3–5]. Especially when used to generate metagenome-assembled genomes (MAGs), shotgun metagenomics allows for a more nuanced comparison of the functional potential of both cultivated and uncultivated microbes, offering insights that are both rapid and increasingly cost-effective. This technique becomes particularly useful in deciphering the microbial dynamics of the mother-infant microbiome axis. Recent reviews have underscored the pressing need to understand antimicrobial resistance in infants and the mechanisms of vertical microbial inheritance, both of which are important for assessing the early development of the infant's microbiome and its long-term health implications [6,7].

The task of reconstructing complete or near-complete MAGs is challenging and influenced by factors including sequencing depth, sample diversity, and the methodologies of assembly and binning. The development of various tools and pipelines has vastly improved genome reconstruction from metagenomes [11–14], with the further advantage that strain-tracking tools have in turn benefited by virtue of no longer relying solely on marker-based databases [15]. A key development in this domain is the InStrain tool, which can use high-quality MAGs as a dataset-specific reference database for strain profiling in metagenomes [15]. This has the potential to not only enhance the accuracy of strain tracking, but also detect a broader range of microbial strain transfers. These advancements are vital for precise microbial analysis, especially in clinical settings where identifying antimicrobial resistance-associated determinants is critical [9,16], and where the quality and diversity of MAGs are of utmost importance.

Recent research has noted that some gene families are often absent from MAGs, most notably 16S rRNA and ribosomal protein genes [17–19]. The former were present in only 7% of 270,000 high-quality MAGs [17], while the latter were missing from more than 20-40% of near complete MAGs analysed [19]. The difficulty in reconstruction of 16S rRNA genes most likely arises due to the presence of multiple gene copies within a single genome, their high sequence similarity, and ubiquitous presence across bacteria [18]. The issue of absent ribosomal protein genes is particularly pronounced in fast-growing genera such as *Bacteroides* and has been attributed to atypical tetranucleotide frequencies surrounding ribosomal proteins due to their higher rate of translation [19]. In line with this, Nelson *et al.* noted that genes with atypical tetranucleotide frequencies are often missed in MAGs [20]. Moreover, Checkm, a tool widely used to assess genome quality, uses lineage specific marker genes that may label genomes as high-quality despite the absence of 16S rRNA and ribosomal protein genes [19,21].

Benchmarking studies have shown that MetaWRAP outperforms MetaBAT2 in recovering high-quality MAGs from mock and *in silico* communities, showing a 23% increase in MAG reconstruction, mainly due to the additional refinement steps developed in MetaWRAP. [14,22]. While these simulated datasets validate the reliability of such tools, they do not fully replicate the complexities inherent in real datasets, such as population diversity and intraspecies heterogeneity, which significantly influence the quality of MAG reconstruction [23]. Additionally, the impact of MAG-recovery tool selection on downstream study outcomes remains unexamined, particularly the quality of MAGs, the recovery rates of ribosomal marker genes, and further detection of antimicrobial resistance and strain transmission events. As the field progresses towards high-resolution, strain-resolved metagenomics, it is vital to understand the potential biases in microbiome data analysis methods and their implications for the field.

To address these questions, we generated 20,834 MAGs from three existing mother-infant gut metagenomic datasets using two tools to explore the effect of binning tool selection on MAG reconstruction and to examine its downstream impacts.

2.3 Results

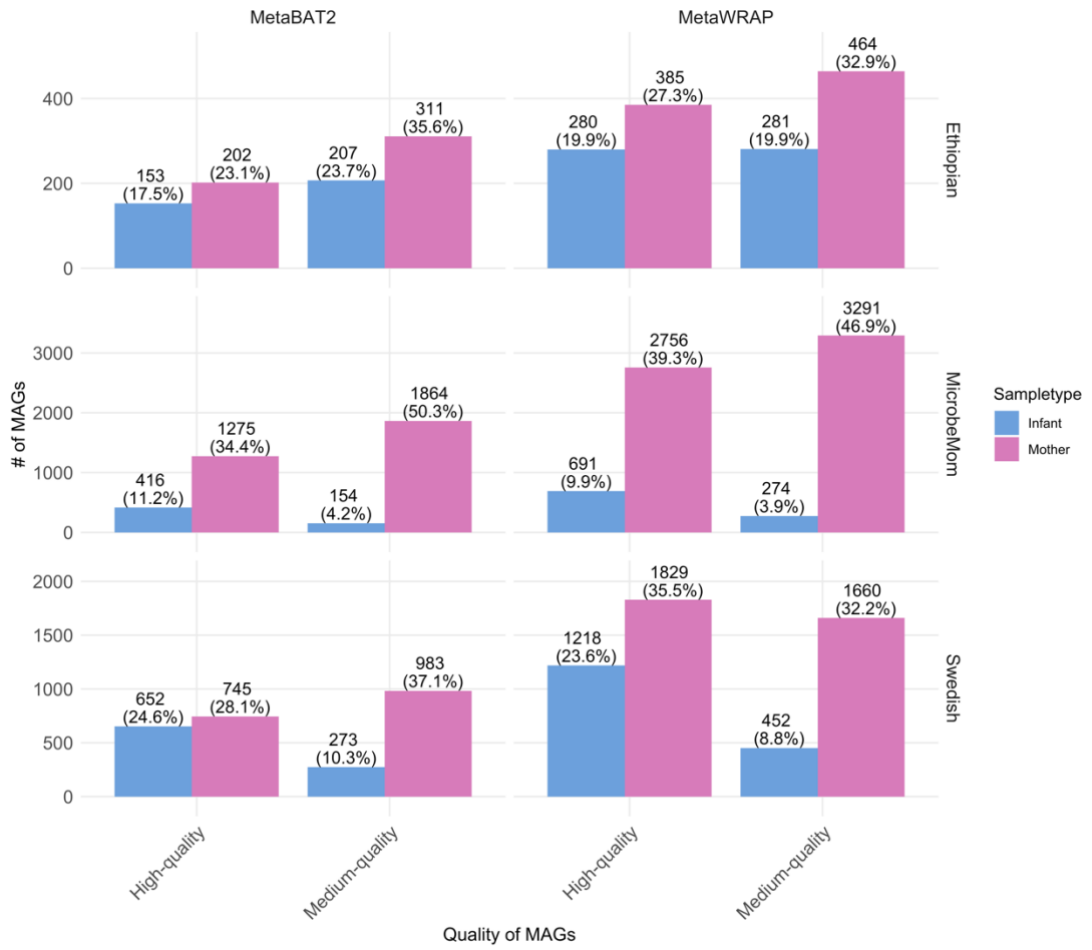
2.3.1 Binning Tool Selection Impacts the Number, Diversity, and Quality of MAGs Reconstructed from Gut Metagenomes

We utilised three shotgun metagenomic datasets from studies involving mother-infant pairs to compare the performance of two distinct binning tools: MetaBAT2, a standalone tool, and MetaWRAP, which combines three binning tools. These datasets were from the MicrobeMom study, focusing on an Irish cohort (n=643) [24], one Swedish cohort (n=300) and one Ethiopian cohort (n=50) [25,26]. Using MetaBAT2 and MetaWRAP, metagenome-assembled genomes (MAGs) were reconstructed from each dataset. MAGs with greater than 90% completeness and less than 5% contamination were classified as high-quality (HQ), while those with greater than 50% completeness and less than 10% contamination were designated as medium quality (MQ) [27].

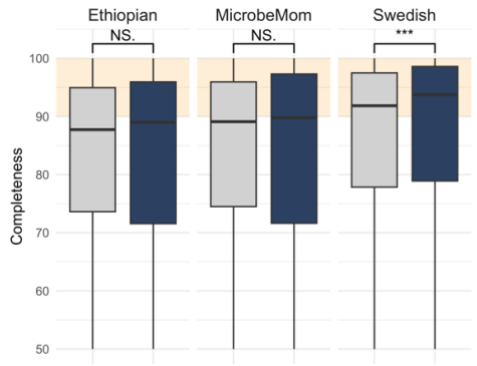
MetaWRAP significantly outperformed MetaBAT2 in the reconstruction of MAGs across all datasets used in the study. In the Ethiopian (ET), MicrobeMom (MM), and Swedish (SWE) datasets, MetaWRAP recovered 61% (1410 vs. 873), 89% (7012 vs. 3709), and 94% (5159 vs. 2653) more combined HQ- and MQ-MAGs than MetaBAT2 (Fig. 2.1a). An average of 62% more MQ-MAGs were reconstructed by MetaWRAP compared to MetaBAT2 across the three datasets. MetaWRAP recovered 117% and 78% more HQ-MAGs cumulatively across all datasets from mother and infant samples, respectively, with a mean completeness of 96% (Fig. 2.1b, Supplementary Table 2.1a). We also found that MetaWRAP reconstructed a greater diversity of MAGs, with a cumulative total of 233 high-quality, more diverse species MAGs across all datasets compared to MetaBAT2. In all three datasets, contamination rates were low (mean <1%) for both sets of HQ MAGs; however, contamination was lower when MetaWRAP was used (mean of 0.66% compared to 0.86%) (Fig. 2.1c, Supplementary Table 2.1b). Distinct differences between the tools were also evident in the N50 and number of contigs, with MetaWRAP performing significantly better with respect to these parameters across all datasets (Supplementary Figure 2.1a & 2.1b). For all downstream analyses, only high-quality MAGs were used.

We also investigated whether either tool reconstructed MAGs belonging to certain bacterial taxa more effectively. MAGs from the *Bacteroides* and *Phocaeicola* genera, were recovered at an approximately 7- and 8-fold higher frequency with MetaWRAP compared to MetaBAT2 across all datasets (Figure 2.1d). Additionally, the *Bifidobacterium* and *Agathobacter* genera also showed notable differences, with MetaWRAP recovering roughly twice as many MAGs as MetaBAT2. Further comparison of the HQ-MAGs reconstructed by MetaWRAP to both the MQ- and HQ-MAGs from MetaBAT2 revealed that there were MAGs reconstructed by MetaWRAP which had no equivalent MetaBAT2 MAG (MetaWRAP-exclusive MAGs). MAGs were considered MetaWRAP-exclusive if there was no MAG of the same species reconstructed from the same sample by MetaBAT2. Within the SWE dataset 80.3% (n = 167) of *Bacteroides* and 86.1% (n = 87) of *Phocaeicola* were MetaWRAP-exclusive, whilst in the MM dataset this MetaWRAP-exclusivity was 77.7% (n = 139) and 83.8% (n = 161) for *Bacteroides* and *Phocaeicola*, respectively. (Supplementary Figure 2.1c). Additionally, on average, 43% of high-quality *Bifidobacterium* MAGs derived from MetaWRAP in the SWE and MM cohorts lacked any counterpart in MetaBAT2, whether medium- or high-quality. We hypothesised that this discrepancy in recovery was linked to sequencing depth variations for these genera, but no significant differences were observed (Supplementary Figure 2.1d). Taken together, MetaWRAP was able to reconstruct approximately twice the number of high-quality MAGs when compared to MetaBAT2, while also improving on genome completeness and reducing contamination.

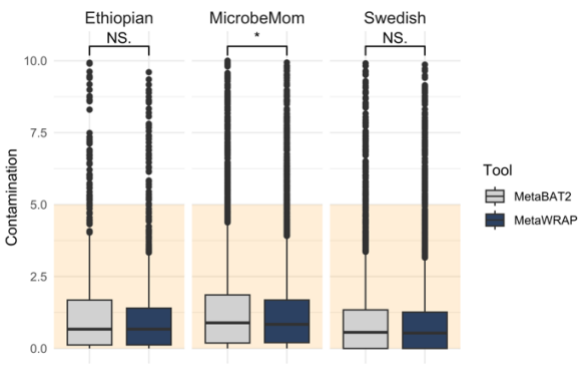
a)


MetaBAT2
MetaWRAP
Ethiopian
MicrobeMom
Swedish
# of MAGs
Quality of MAGs
High-quality
Medium-quality
Sampletype
Infant
Mother
153 (17.5%)
202 (23.1%)
207 (23.7%)
311 (35.6%)
280 (19.9%)
385 (27.3%)
281 (19.9%)
464 (32.9%)
416 (11.2%)
1275 (34.4%)
154 (4.2%)
1864 (50.3%)
691 (9.9%)
2756 (39.3%)
274 (3.9%)
3291 (46.9%)
652 (24.6%)
745 (28.1%)
273 (10.3%)
983 (37.1%)
1218 (23.6%)
1829 (35.5%)
452 (8.8%)
1660 (32.2%)

b)


Ethiopian
MicrobeMom
Swedish
NS.
NS.
***
Completeness

c)


Ethiopian
MicrobeMom
Swedish
NS.
*
NS.
Contamination
Tool
MetaBAT2
MetaWRAP

d)

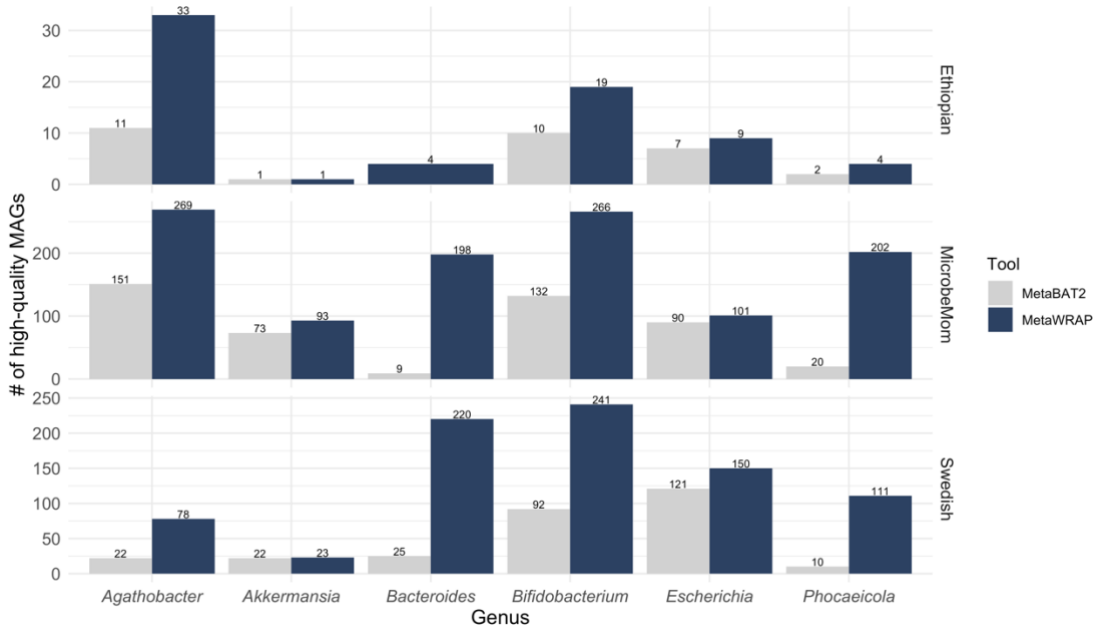
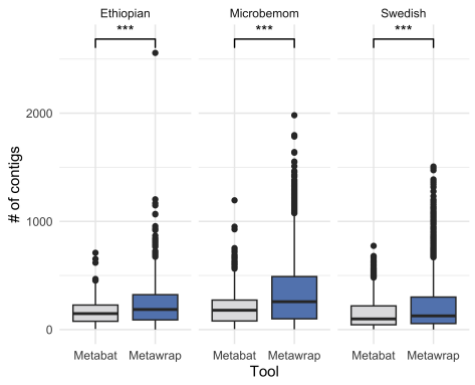
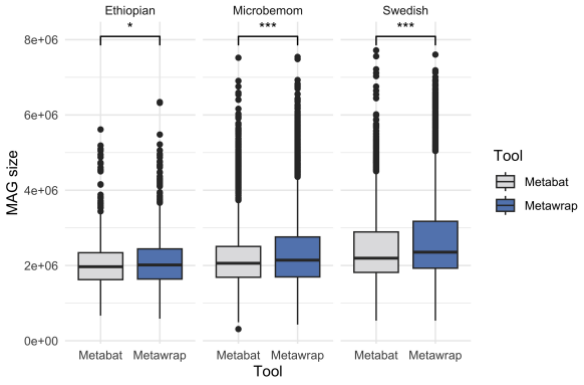


Figure 2.1. Comparison of MAGs recovered by MetaBAT2 and MetaWRAP. **a.** Barplots showing the number of MAGs recovered from each sample type and dataset, also shown are the percentage of MAGs from each dataset-binner combination stratified by quality and sample source. **b, c.** The completeness and contamination values of MAGs across three datasets represented as boxplots. Orange shaded regions representing the completeness and contamination values for high-quality MAGs. The significant differences between the tools, as determined by Wilcoxon test, are indicated by asterisks. **d.** Barplots representing the number of high-quality MAGs recovered per genus by each tool. The values on each bar represents the MAGs recovered in each category.

a)


Ethiopian
Microbemom
Swedish
***
# of contigs
2000
1000
0
Metabat
Metawrap
Tool

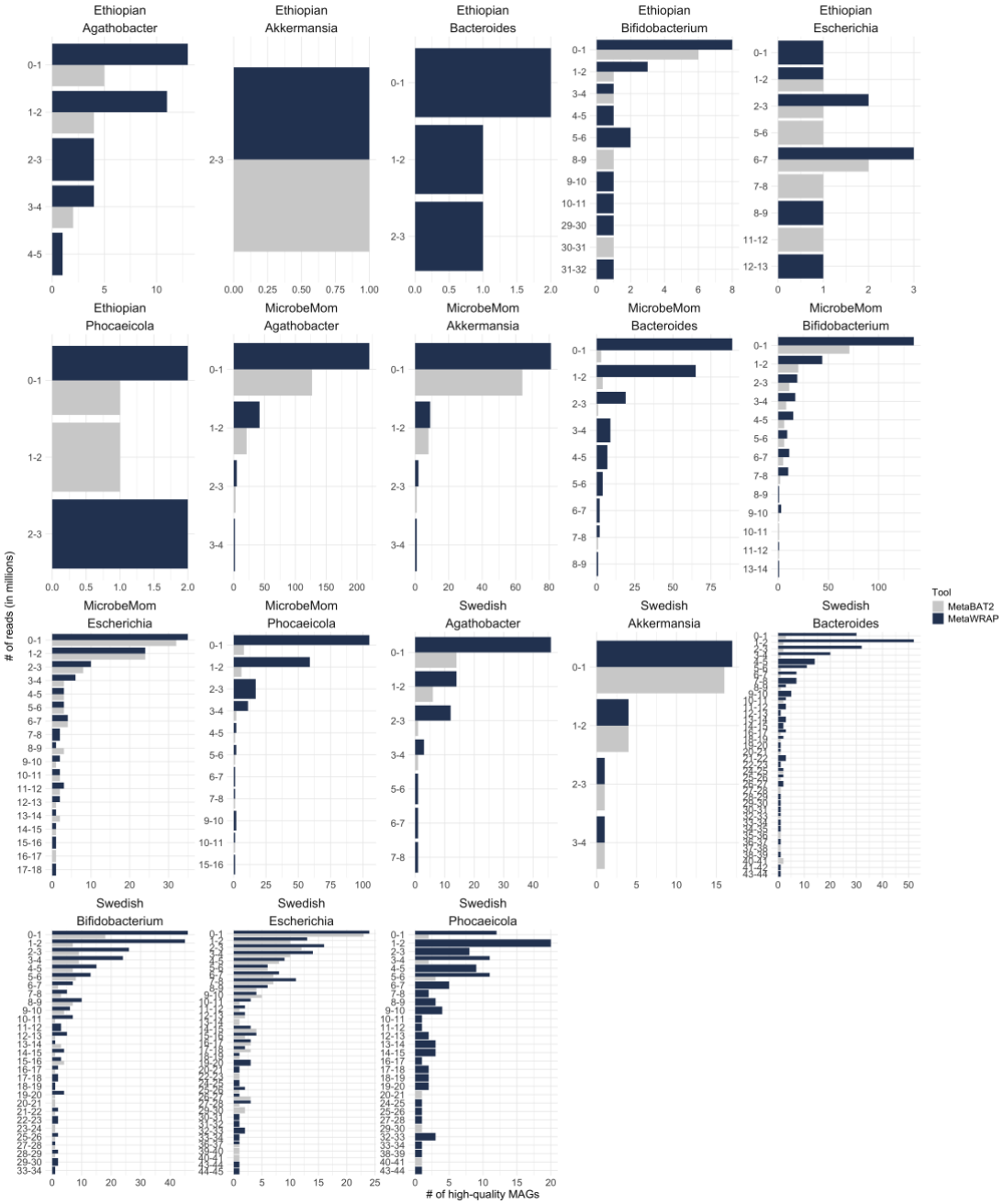
b)


Ethiopian
Microbemom
Swedish
*
***
MAG size
8e+06
6e+06
4e+06
2e+06
0e+00
Metabat
Metawrap
Tool
Tool
Metabat
Metawrap

c)


# of MAGs
Ethiopian
Microbemom
Swedish
Distribution of HQ MetaWRAP MAGs within MetaBAT2 MAGs
Not reconstructed by MetaBAT2
Reconstructed as HQ MetaBAT2 MAGs
Reconstructed as MQ MetaBAT2 MAGs
Agathobacter
Akkermansia
Bacteroides
Bifidobacterium
Escherichia
Phocaeicola
Genus

d)



Supplementary Figure 2.1. a, b. Boxplots showing the number of contigs and N50 values of MAGs across three datasets. The significant differences between the tools were indicated by asterisks. **c.** Bar plots represents the high-quality MAGs recovered by MetaWRAP and how they are classified by MetaBAT2 **d.** Barplots showing the number of high-quality MAGs recovered at various sequencing depth across different genera.

Supplementary Table 2.1a | Mean completeness values of MAGs

Tool	Dataset	Quality	mean
Metabat	Ethiopian	High-quality	95.4303944
Metabat	Ethiopian	Medium-quality	75.0967375
Metabat	Microbemom	High-quality	95.8840036
Metabat	Microbemom	Medium-quality	74.3655847
Metabat	Swedish	High-quality	96.5307516
Metabat	Swedish	Medium-quality	74.9334395
Metawrap	Ethiopian	High-quality	95.8202105
Metawrap	Ethiopian	Medium-quality	72.3849262
Metawrap	Microbemom	High-quality	96.5305367
Metawrap	Microbemom	Medium-quality	71.0823815
Metawrap	Swedish	High-quality	97.0846439
Metawrap	Swedish	Medium-quality	73.1666099

Supplementary Table 2.1b | Mean contamination values of MAGs

Binner	Dataset	Quality	mean
Metabat	Ethiopian	High-quality	0.88769014
Metabat	Ethiopian	Medium-quality	1.67839768
Metabat	Microbemom	High-quality	0.99969249
Metabat	Microbemom	Medium-quality	1.84387017
Metabat	Swedish	High-quality	0.70185397
Metabat	Swedish	Medium-quality	1.50848726
Metawrap	Ethiopian	High-quality	0.67732481
Metawrap	Ethiopian	Medium-quality	1.44739329
Metawrap	Microbemom	High-quality	0.73051262
Metawrap	Microbemom	Medium-quality	1.80129004
Metawrap	Swedish	High-quality	0.59200689
Metawrap	Swedish	Medium-quality	1.50274148

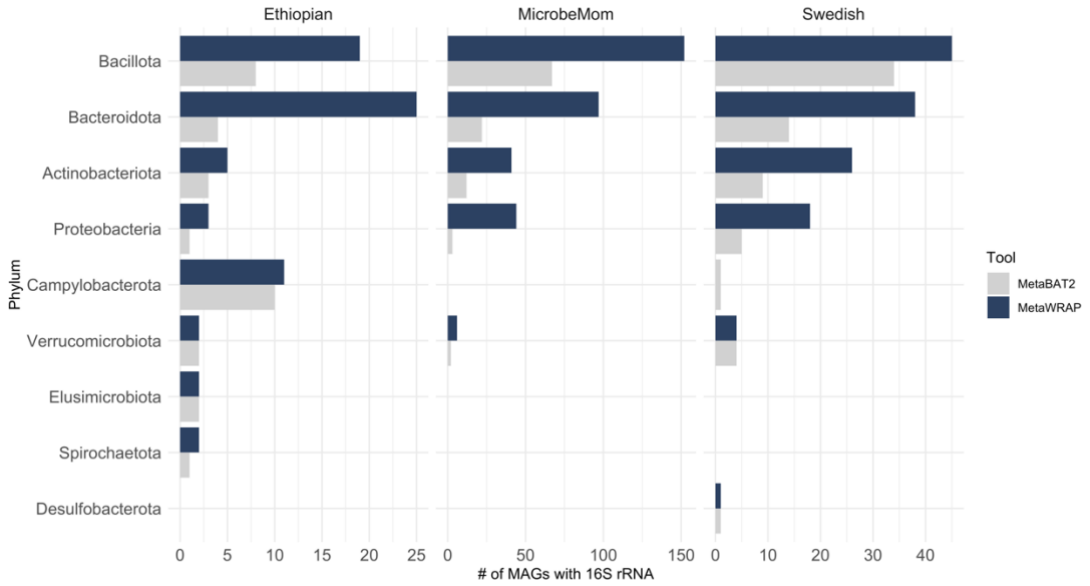
2.3.2 Binning Tool Selection Impacts Ribosomal Gene Reconstruction

To assess the recovery rates of 16S rRNA genes within MAGs, we investigated the presence of at least one 16S rRNA gene in common HQ-MAGs, i.e., MAGs from the same species and sample that were reconstructed by both MetaWRAP and MetaBAT2. For the ET, MM, and SWE datasets, the recovery rates of 16S rRNA genes by MetaBAT2 were relatively low: 8.7% (31/356), 6.4% (106/1656), and 5.03% (68/1351), respectively (Figure 2.2a). On the other hand, MetaWRAP showed a notable improvement in recovery rates: 19.3% (69/356), 20.5% (340/1656), and 9.7% (132/1351), respectively.

The difference between the two tools was especially noticeable for MAGs from the Bacteroidota and Bacillota phyla, with MetaWRAP recovering approximately four- and two-fold, respectively, as many MAGs containing at least one 16S rRNA gene relative to MetaBAT2 and comparable numbers were also observed in MetaWRAP-exclusive MAGs. We found the 16S rRNA gene in 14.1% (44/310), 17.4% (307/1756), and 10.3% (170/1650) of the MAGs in the ET, MM, and SWE datasets, respectively (Supplementary Figure 2.2a).

To assess the distribution of genes encoding other core ribosomal proteins across the three datasets, we again used the common HQ-MAGs and found similar results to when 16S rRNA gene profiles were assessed (Fig. 2.2b). More specifically, MetaWRAP was particularly effective, recovering around 10% more MAGs with at least 90% of core ribosomal proteins (38/42) on average across the datasets. Additionally, about 80% of the MetaWRAP-specific MAGs contained >90% of these core proteins (Supplementary Figure 2.2b), highlighting MetaWRAP's ability to recover high-quality MAGs with ribosomal marker genes more effectively than MetaBAT2.

a)



b)

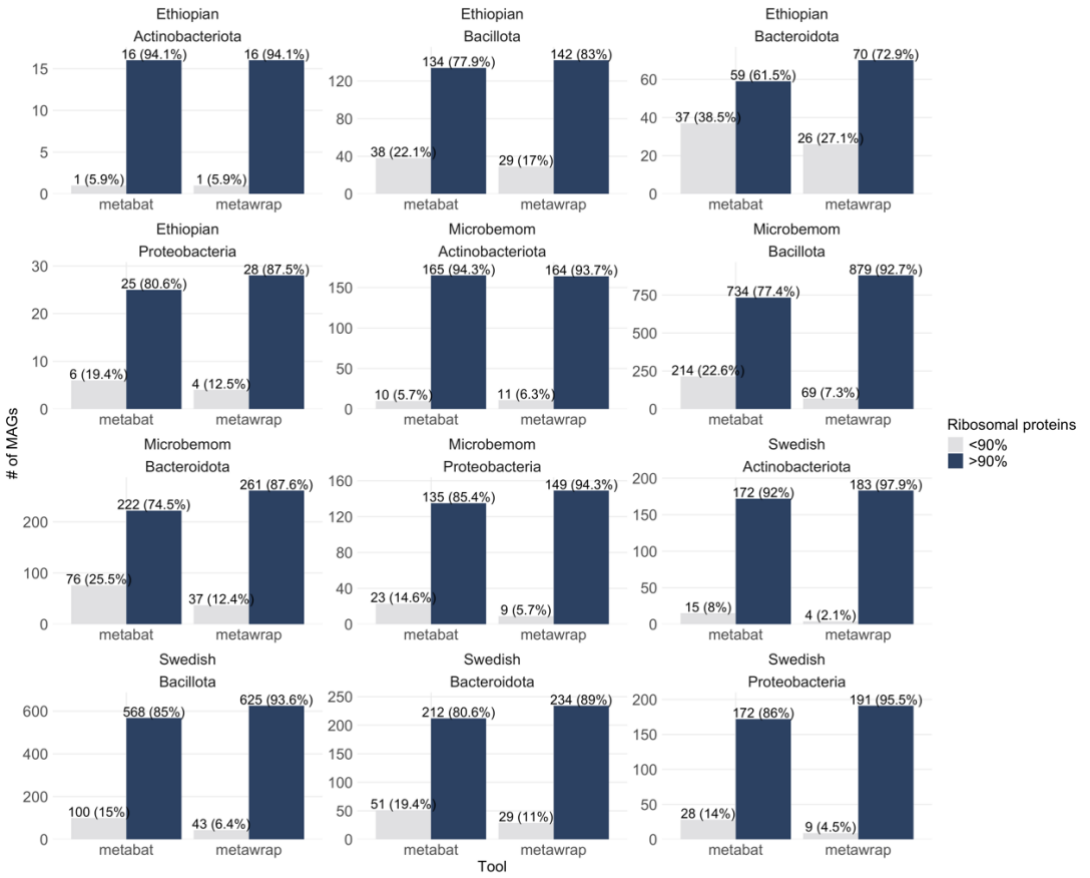
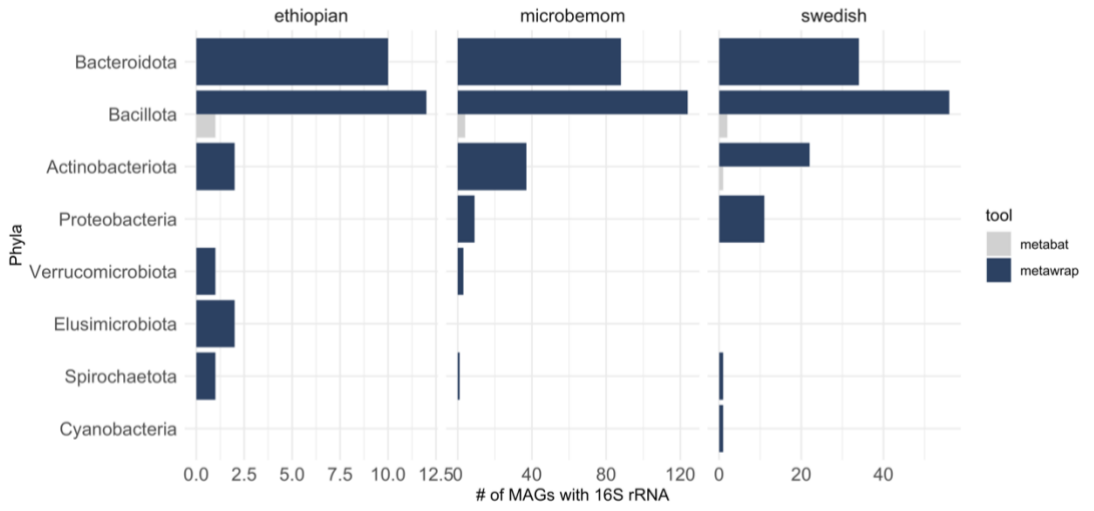


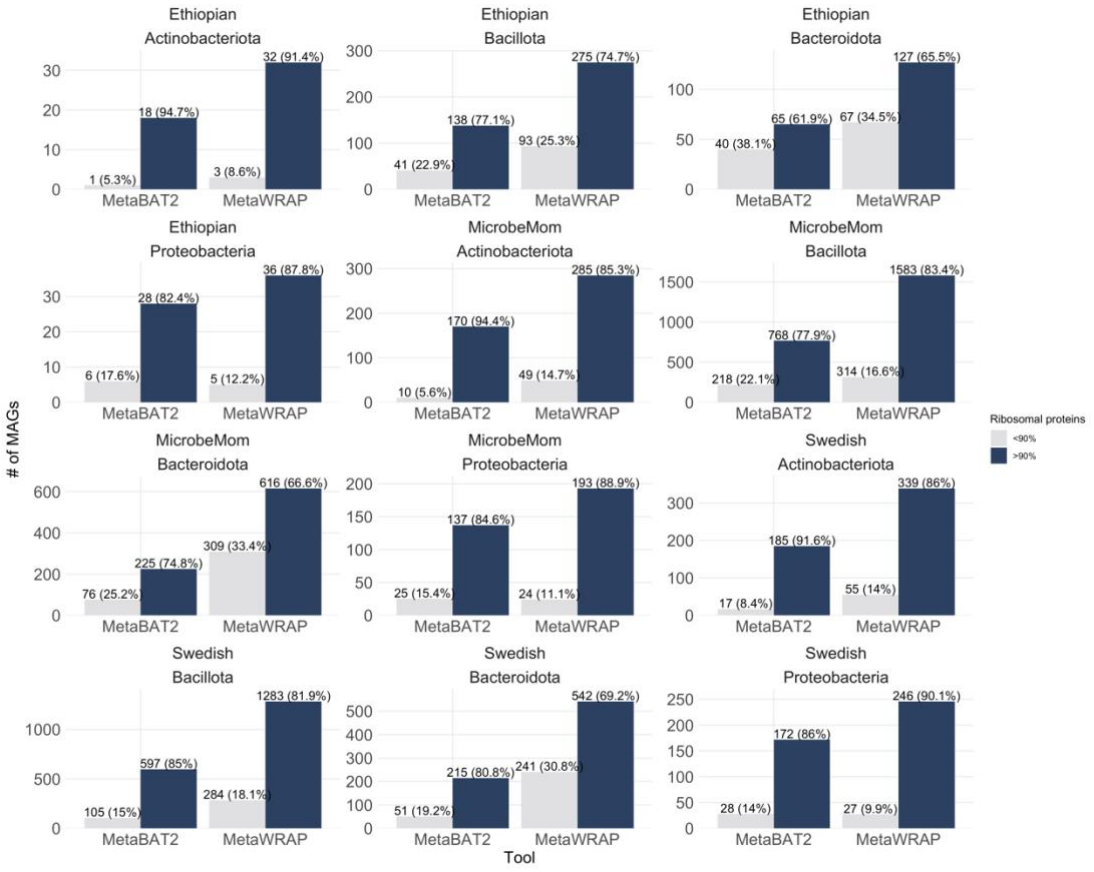
Figure 2.2. Comparison of ribosomal gene recovery in MAGs reconstructed by both MetaWRAP and MetaBAT2. **a.** Barplots comparing the number of MAGs containing at least one 16S rRNA gene across different phyla. **b.** Grouped barplots showing the

presence of core ribosomal protein genes across three datasets in the most abundant phyla. The values on each bar represent the number of MAGs and its proportion in each category.

a)



b)



Supplementary Figure 2.2. a. Barplots showing the presence of 16S rRNA gene in high-quality MAGs exclusively recovered by MetaWRAP and MetaBAT2 across three datasets. **b.** Distribution of core ribosomal proteins in high-quality MAGs recovered by MetaWRAP and MetaBAT2 across multiple phyla represented as barplots.

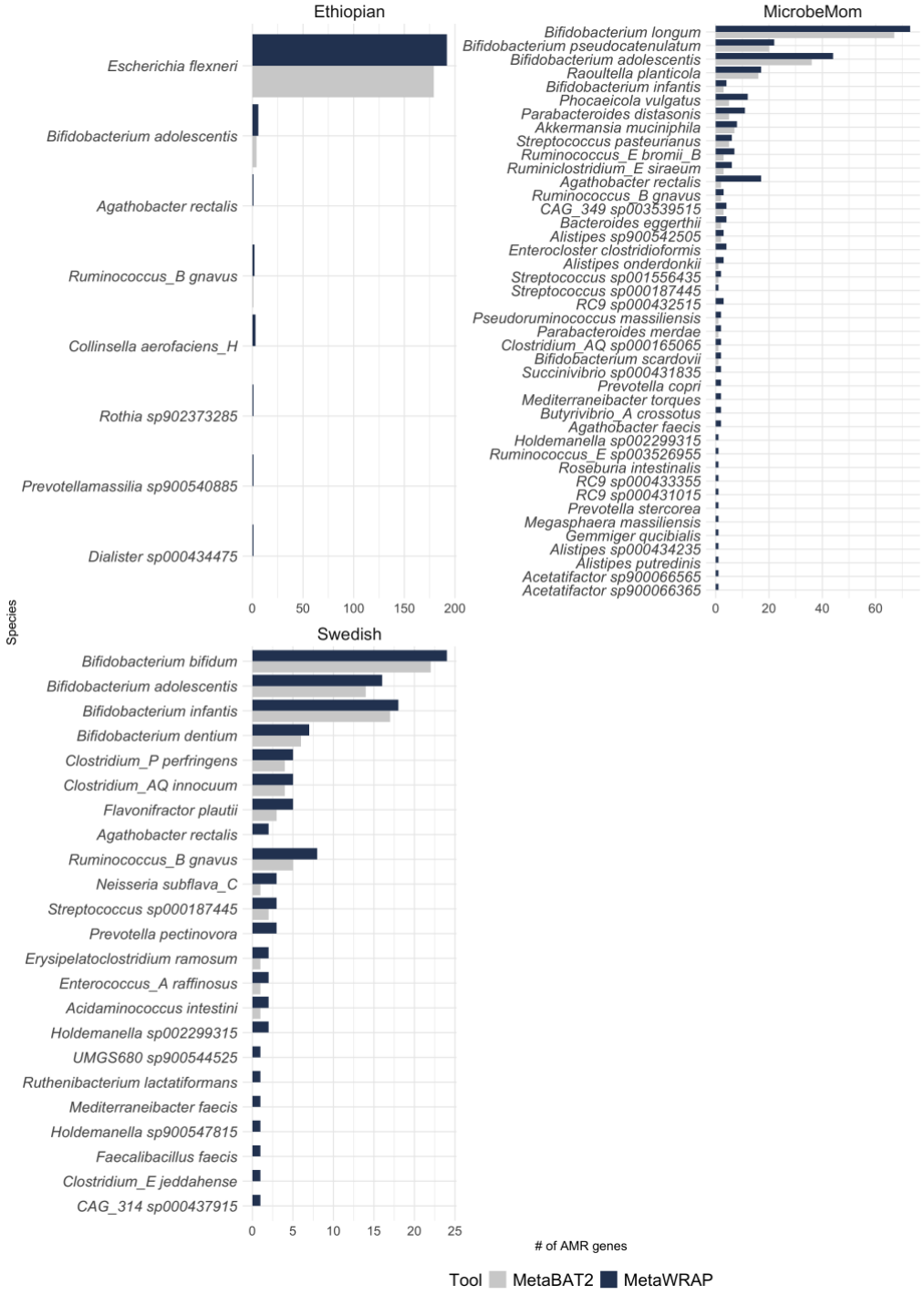
2.3.3 Binning Tool Selection Impacts the Efficiency of Antibiotic Resistance Gene Detection

Next, we examined the antimicrobial resistance (AMR) gene profiles within common MAGs across the three datasets. We found 55 more AMR genes from 17 species in the MM dataset, and 23 additional AMR genes from 8 species in the SWE dataset, that were exclusively detected in MetaWRAP MAGs and not in those generated using MetaBAT2 (Figure 2.3a). In particular, MetaWRAP MAGs from the species *Agathobacter rectalis* and *Bifidobacterium adolescentis* generated from the MM dataset revealed an additional 15 and 10 AMR genes, respectively, compared to MetaBAT2. While, 13 more AMR genes were identified in *Escherichia flexneri* from the Ethiopian MetaWRAP MAGs.

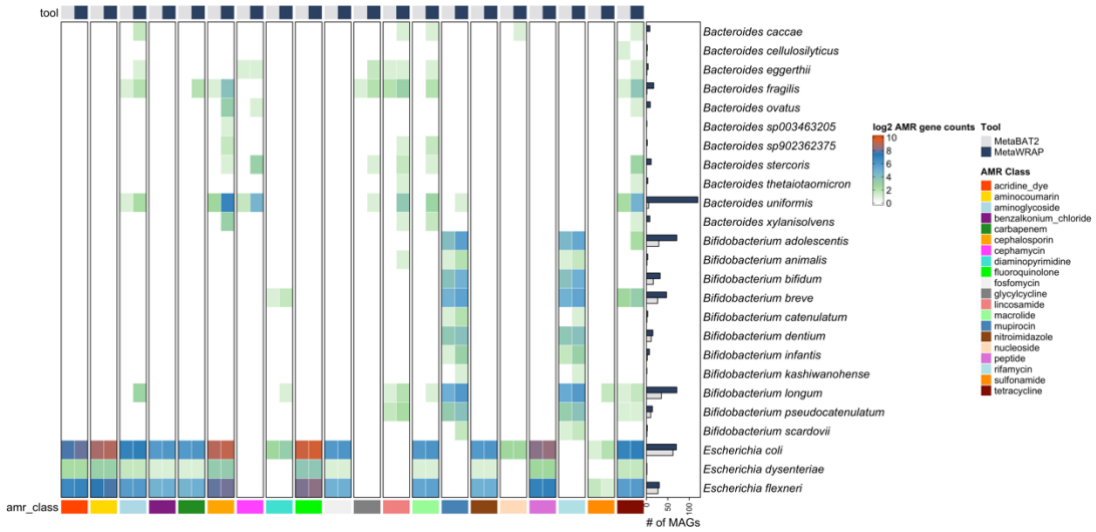
Assessing the broader impact of the MetaWRAP-specific MAGs on AMR gene detection at the genome level, we reconstructed the resistome profiles for gut bacterial genera with at least 100 MAGs across all three datasets. MetaWRAP-specific MAGs from the *Bacteroides* genus contributed to a more comprehensive detection of AMR genes, with significant increases observed in the number of cephalosporin- (24-fold increase from 6 to 151 genes, and 7.5-fold increase from 18 to 147 genes) and tetracycline- (7-fold increase from 6 to 47 genes, and 1.6-fold increase from 13 to 34 genes) resistance genes in the MM and SWE datasets, respectively compared to MetaBAT2 MAGs (Figures 2.3b & 2.3c). Additionally, MetaWRAP-derived *Bifidobacterium* MAGs contained twice as many genes related to mupirocin and rifamycin resistance across all three datasets compared to their MetaBAT2-derived counterparts (Figures 2.3b. & 2.3c, Supplementary Figure 2.3). These differences were also observed in the genus *Escherichia*, particularly with respect to the fluoroquinolone and cephalosporin classes. Within the MM dataset, approximately 100 additional AMR genes were identified in each class, whereas the SWE dataset

revealed about 300 more genes in each class (Supplementary Tables 2.2a & 2.2b). These findings reveal MetaWRAP's improved efficiency in recovering MAGs with AMR genes relative to MetaBAT2, thereby providing a better understanding of the gut resistome.

a)



b)



c)

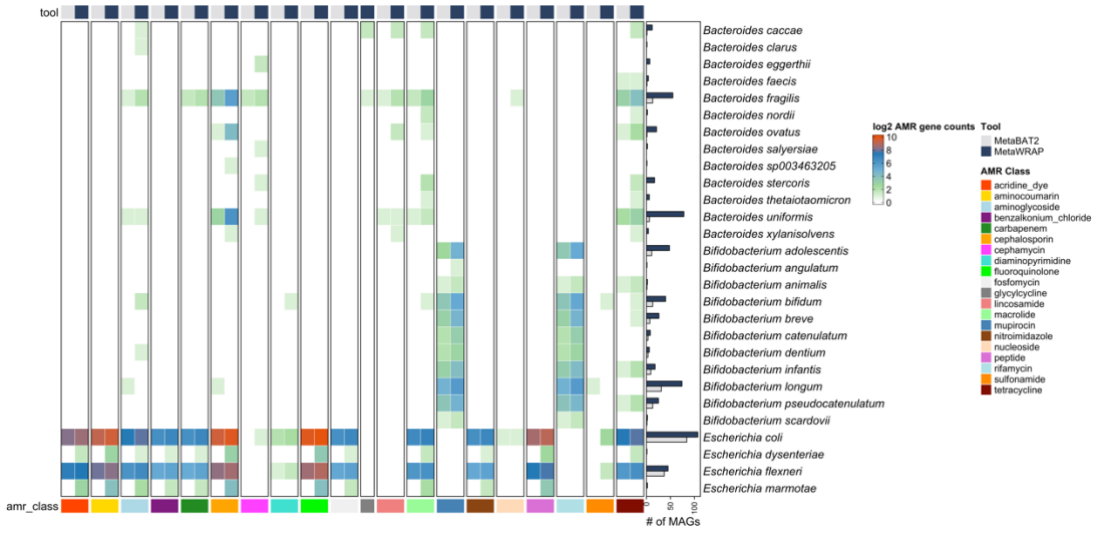
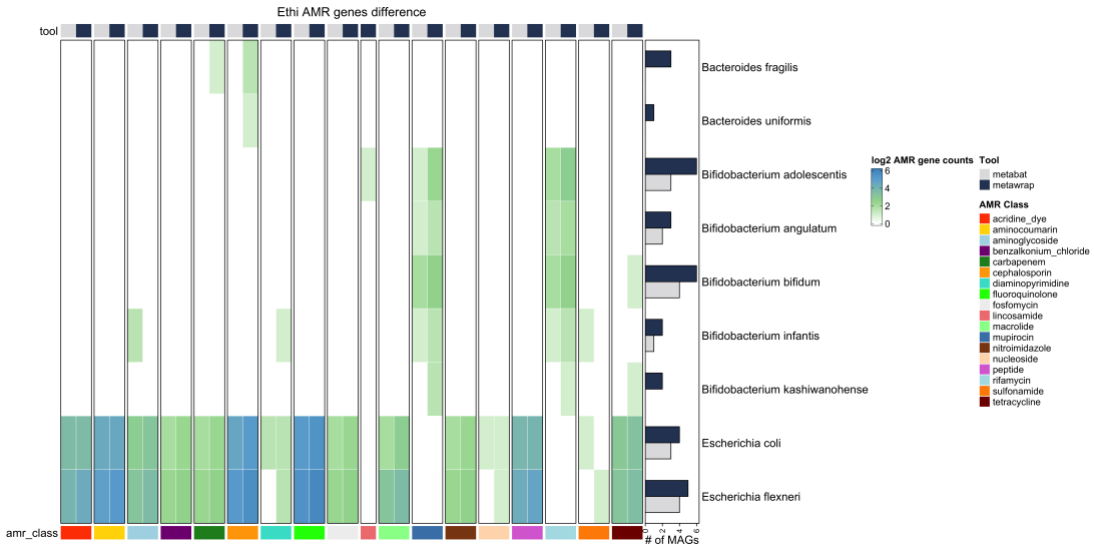


Figure 2.3. Comparison of antimicrobial resistance determinant detection in MAGs recovered by MetaWRAP and MetaBAT2. **a.** Barplots faceted by dataset showing the AMR genes detected in common MAGs per species – only genes with **>**5% difference are plotted. **b, c.** Heatmaps showing the AMR genes detected in bacterial genera with at least 100 MAGs in the (**b**) MicrobeMom and (**c**) Swedish datasets. The barplot on the right indicates the number of MAGs recovered in the corresponding species by each tool. The colour strip on the top represents the binning tool, and the bottom strip indicates the antimicrobial resistance classes.



Supplementary Figure 2.3. Comparison of antimicrobial resistance detections in MAGs recovered by MetaWRAP and MetaBAT2. Heatmap showing the AMR genes detected in common gut bacterial genera from Ethiopian cohort. The barplot on the right indicates the number of MAGs recovered in the corresponding species by each tool. The colour strip on the top represents the binning tool and the bottom strip indicates the antimicrobial resistance classes.

Supplementary Table 2.2a | Antibiotic resistance gene counts per genera in Microbemom dataset

Genera	Binner	AMR class	n
Bacteroides	MetaBAT2	aminoglycoside	2
Bacteroides	MetaBAT2	cephalosporin	6
Bacteroides	MetaBAT2	cephamycin	3
Bacteroides	MetaBAT2	glycylcycline	1
Bacteroides	MetaBAT2	lincosamide	4
Bacteroides	MetaBAT2	tetracycline	6
Bacteroides	MetaWRAP	aminoglycoside	10
Bacteroides	MetaWRAP	carbapenem	3
Bacteroides	MetaWRAP	cephalosporin	151
Bacteroides	MetaWRAP	cephamycin	32
Bacteroides	MetaWRAP	glycylcycline	7
Bacteroides	MetaWRAP	lincosamide	24
Bacteroides	MetaWRAP	macrolide	16
Bacteroides	MetaWRAP	mupirocin	1
Bacteroides	MetaWRAP	nucleoside	1
Bacteroides	MetaWRAP	tetracycline	47

Bifidobacterium	MetaBAT2	diaminopyrimidine	1
Bifidobacterium	MetaBAT2	lincosamide	3
Bifidobacterium	MetaBAT2	mupirocin	108
Bifidobacterium	MetaBAT2	rifamycin	119
Bifidobacterium	MetaBAT2	tetracycline	7
Bifidobacterium	MetaWRAP	aminoglycoside	6
Bifidobacterium	MetaWRAP	diaminopyrimidine	3
Bifidobacterium	MetaWRAP	lincosamide	8
Bifidobacterium	MetaWRAP	macrolide	1
Bifidobacterium	MetaWRAP	mupirocin	227
Bifidobacterium	MetaWRAP	rifamycin	211
Bifidobacterium	MetaWRAP	sulfonamide	2
Bifidobacterium	MetaWRAP	tetracycline	15
Escherichia	MetaBAT2	acridine_dye	308
Escherichia	MetaBAT2	aminocoumarin	615
Escherichia	MetaBAT2	aminoglycoside	185
Escherichia	MetaBAT2	benzalkonium_chloride	85
Escherichia	MetaBAT2	carbapenem	86
Escherichia	MetaBAT2	cephalosporin	750
Escherichia	MetaBAT2	diaminopyrimidine	5
Escherichia	MetaBAT2	fluoroquinolone	958
Escherichia	MetaBAT2	fosfomycin	90
Escherichia	MetaBAT2	macrolide	104
Escherichia	MetaBAT2	nitroimidazole	88
Escherichia	MetaBAT2	nucleoside	5
Escherichia	MetaBAT2	peptide	436
Escherichia	MetaBAT2	sulfonamide	3
Escherichia	MetaBAT2	tetracycline	175
Escherichia	MetaWRAP	acridine_dye	360
Escherichia	MetaWRAP	aminocoumarin	692
Escherichia	MetaWRAP	aminoglycoside	207
Escherichia	MetaWRAP	benzalkonium_chloride	98
Escherichia	MetaWRAP	carbapenem	100
Escherichia	MetaWRAP	cephalosporin	842
Escherichia	MetaWRAP	diaminopyrimidine	8
Escherichia	MetaWRAP	fluoroquinolone	1089
Escherichia	MetaWRAP	fosfomycin	100
Escherichia	MetaWRAP	macrolide	120
Escherichia	MetaWRAP	nitroimidazole	100
Escherichia	MetaWRAP	nucleoside	5
Escherichia	MetaWRAP	peptide	483
Escherichia	MetaWRAP	sulfonamide	4

Escherichia	MetaWRAP	tetracycline	201

Supplementary Table 2.2b | Antibiotic resistance gene counts per genera in Swedish dataset

Genera	Binner	AMR class	n
Escherichia	MetaWRAP	fluoroquinolone	1656
Escherichia	MetaWRAP	cephalosporin	1313
Escherichia	MetaBAT2	fluoroquinolone	1288
Escherichia	MetaWRAP	aminocoumarin	1046
Escherichia	MetaBAT2	cephalosporin	1030
Escherichia	MetaBAT2	aminocoumarin	819
Escherichia	MetaWRAP	peptide	738
Escherichia	MetaBAT2	peptide	577
Escherichia	MetaWRAP	acridine_dye	553
Escherichia	MetaBAT2	acridine_dye	436
Escherichia	MetaWRAP	aminoglycoside	327
Escherichia	MetaWRAP	tetracycline	300
Escherichia	MetaBAT2	aminoglycoside	238
Escherichia	MetaBAT2	tetracycline	229
Bifidobacterium	MetaWRAP	rifamycin	212
Escherichia	MetaWRAP	macrolide	210
Bifidobacterium	MetaWRAP	mupirocin	203
Escherichia	MetaBAT2	macrolide	165
Escherichia	MetaWRAP	nitroimidazole	155
Escherichia	MetaWRAP	benzalkonium_chloride	153
Escherichia	MetaWRAP	carbapenem	153
Escherichia	MetaWRAP	fosfomycin	153
Bacteroides	MetaWRAP	cephalosporin	147
Escherichia	MetaBAT2	fosfomycin	121
Escherichia	MetaBAT2	nitroimidazole	121
Escherichia	MetaBAT2	benzalkonium_chloride	120
Escherichia	MetaBAT2	carbapenem	116
Bifidobacterium	MetaBAT2	rifamycin	86
Bifidobacterium	MetaBAT2	mupirocin	76
Bacteroides	MetaWRAP	tetracycline	34
Bacteroides	MetaBAT2	cephalosporin	18
Bacteroides	MetaWRAP	macrolide	17
Bacteroides	MetaBAT2	tetracycline	13
Bifidobacterium	MetaWRAP	tetracycline	10
Bacteroides	MetaWRAP	lincosamide	9

Bacteroides	MetaWRAP	cephamycin	8
Escherichia	MetaWRAP	sulfonamide	7
Bacteroides	MetaWRAP	aminoglycoside	6
Escherichia	MetaWRAP	diaminopyrimidine	6
Escherichia	MetaBAT2	diaminopyrimidine	4
Bacteroides	MetaBAT2	macrolide	3
Bacteroides	MetaWRAP	carbapenem	3
Bacteroides	MetaWRAP	glycylcycline	3
Bifidobacterium	MetaBAT2	tetracycline	3
Bifidobacterium	MetaWRAP	aminoglycoside	3
Bacteroides	MetaBAT2	aminoglycoside	2
Bacteroides	MetaBAT2	carbapenem	2
Bacteroides	MetaBAT2	cephamycin	2
Bacteroides	MetaBAT2	lincosamide	2
Bacteroides	MetaWRAP	nucleoside	1
Bifidobacterium	MetaBAT2	aminoglycoside	1
Bifidobacterium	MetaBAT2	cephalosporin	1
Bifidobacterium	MetaBAT2	sulfonamide	1
Bifidobacterium	MetaWRAP	diaminopyrimidine	1
Bifidobacterium	MetaWRAP	macrolide	1
Bifidobacterium	MetaWRAP	sulfonamide	1
Escherichia	MetaBAT2	nucleoside	1
Escherichia	MetaWRAP	cephamycin	1
Escherichia	MetaWRAP	nucleoside	1

2.3.4 Binning Tool Selection Impacts the Detection of Mother-Infant Strain Sharing Events

Given the marked differences in MAG recovery and AMR profiling between MetaWRAP and MetaBAT2, we extended our investigation to the phenomenon of strain sharing events between paired mother-infant metagenomes. Utilizing the InStrain tool, we profiled strain sharing in mother-infant pairs, comparing the outcomes between databases constructed from representative MAGs generated by each binning tool for the three datasets. MetaWRAP-derived MAGs produced a more diverse InStrain database, yielding on average approximately 1.8 times more representative MAGs across all datasets than the MetaBAT2-derived database. Notably, there was an approximate 25% increase in detected strain-sharing events in MM (37 vs 47) and SWE (76 vs 95) mother-infant samples using the database constructed from MetaWRAP MAGs compared to MetaBAT2, while in the Ethiopian dataset, only 2 additional events (25 vs 27) were detected (Figure 2.4).

Additionally, when using MetaWRAP-derived MAGs as a reference database for strain sharing, we found strain sharing events in species that were not identified using MetaBAT2 derived database. Specifically, in the Ethiopian, MicrobeMom, and Swedish datasets, we identified 4, 8, and 10 unique shared species, respectively. The majority of these additional strain sharing events in the MM (8/10) and SWE (17/19) datasets were from the *Bacteroides* and *Phocaeicola* genera. Of the 138 sharing events detected across all datasets using MetaBAT2 MAGs for reference database construction, 10 were missed by MetaWRAP. Notably, species such as *Bifidobacterium pseudocatenulatum*, *Bacteroides cellulosilyticus* and *Phocaecola plebeius* were exclusively detected when using the MetaBAT2 MAGs database in MM dataset. Furthermore, there was a completely distinct strain sharing profile in the Ethiopian cohort when compared to the Swedish and MicrobeMom cohorts with more uncharacterized species being shared between mother-infant pairs in ET (20/25) compared to SWE (5/34) and MM (0/24) samples in line with the previous study [25].

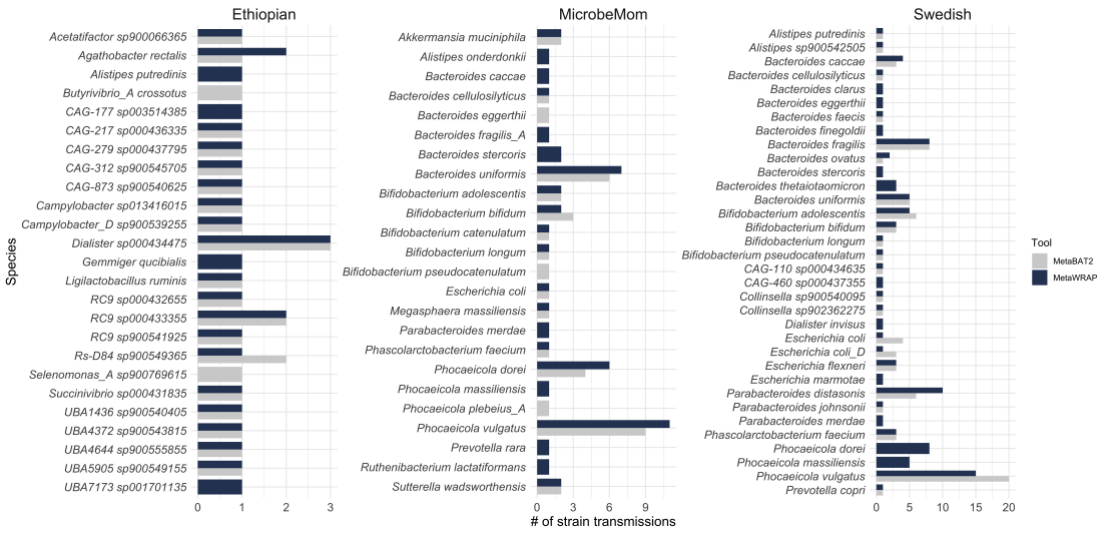


Figure 2.4. Comparison of strain sharing identified by InStrain using databases constructed from MAGs generated by MetaWRAP and MetaBAT2. Barplots showing the number of strain sharing events detected per species in each dataset by InStrain using MAGs from both tools.

2.4 Discussion

This work emphasises the fundamental impact of binning tool selection on research outcomes, as shown through the analysis of three mother-infant metagenomic datasets. MetaWRAP, a wrapper tool combining multiple binning algorithms with quality control and refinement modules, convincingly outperformed the standalone tool MetaBAT2 in terms of the number, quality, and diversity of reconstructed MAGs. It also outperformed MetaBAT2 in recovering genes traditionally underrepresented in MAG catalogues, namely those encoding the 16S rRNA and ribosomal proteins. While these are interesting findings themselves, the most important observation was the substantial impact MAG reconstruction had on downstream analyses, such as the detection of resistance determinants to clinically important antibiotics, and the detection of strain sharing events between mother and infant.

Shotgun metagenomic sequencing is a robust method for exploring the human microbiome's taxonomic and functional potential, including carbohydrate utilization, antibiotic resistance, and strain-level transmission dynamics [10,24,28]. However, the

recovery of high-quality metagenome-assembled genomes is a challenging undertaking. Various tools and pipelines have been developed to produce high-quality MAGs, but recent reports show that only about 7% of these contain the 16S rRNA gene [12,14,17]. The assembly of these genes presents a significant challenge due to their redundant nature and the prevalence of multiple copies within a single genome, frequently resulting in incorrect assemblies [18]. Additionally, high-quality bacterial MAGs often lack core ribosomal protein genes, especially in the Gammaproteobacteria class and Bacteroidota phyla, due to atypical tetranucleotide frequencies that are hypothesised to have arisen from the evolutionary pressures of fast-growing bacteria [19]. These findings raise concerns about the reliability and consistency of high-quality MAGs for recovering ribosomal proteins and other crucial genes like AMR, impacting subsequent analyses.

We noted that the additional MAGs recovered by MetaWRAP were predominantly from the Bacteroidota phylum, and the recovery rates of ribosomal genes followed a similar pattern. This could be attributed to MetaWRAP's bin refinement and reassembly steps producing better assemblies, thus reducing contamination and enhancing completion [14]. Our investigation into the sequencing depth of *Bacteroides* and *Phocaeicola* genera revealed no notable differences compared to other genera. The Bacteroidota phylum contains many fast-growing species, for example *Bacteroides* spp., suggesting that their corresponding genome sequences possess atypical tetranucleotide frequencies, complicating MAG recovery [19,29]. MetaBAT2 tends to split genomes into multiple bins to minimize contamination, resulting in lower quality MAGs, a limitation addressed by MetaWRAP's additional refinement steps [12,14,22]. Our findings, supported by a study from Mise *et al*., suggest that evolutionary pressures in fast-growing bacteria, which have been reported to cause atypical tetranucleotide frequencies, make MAG recovery challenging for MetaBAT2 [19]. Ultimately, our results highlight a hidden bias in microbiome data analysis, as the choice of binning tools can significantly affect the recovery of specific taxa MAGs, leading to data underrepresentation and impacting further analysis.

An example of the importance of MAGs quality is in the detection of antimicrobial resistance genes. Antimicrobial resistance poses a growing threat, complicating

treatment as bacteria develop resistance through resistance mutations and horizontal gene transfer [30,31] and shotgun metagenomic sequencing is commonly used to study antimicrobial resistance in the human microbiome [32,33]. Recent reviews have emphasized the importance of studying AMR in the human microbiome, particularly in the mother-infant axis, as early-life colonization by antibiotic-resistant strains can make infections harder to treat [6,34]. Our previous work has indicated that intrapartum antibiotic use negatively impacts strain transfer between mothers and infants [24]. Highly accurate and complete genomes are crucial for discerning subtle differences at the genome/strain level in AMR studies, and here we identified 78 AMR genes in 25 species using MetaWRAP, which were not detected in the corresponding MetaBAT2 MAG. These differences in AMR detection can be attributed to MetaWRAP's ability to recover MAGs with higher completeness, thus facilitating the identification of additional antibiotic resistance genes.

When reconstructing AMR profiles using MetaWRAP-derived MAGs, we observed significant differences in AMR detection, identifying more strains in the gut microbiome that harbour AMR determinants. It is particularly concerning to note the increased AMR genes in *Bacteroides* spp., as *Bacteroides fragilis* is a common gut commensal but also among the most pathogenic taxa in this group, with increasing resistance to cephalosporins, commonly used during pregnancy [35,36]. Conversely, the *Bifidobacterium* genus, known for its beneficial species, showed more AMR genes related to the mupirocin and rifamycin antibiotic classes. Previous studies have reported that species in this genus exhibit widespread resistance to mupirocin and this antibiotic is widely used as a selective agent when isolating *Bifidobacterium* from faecal samples [37]. Given that *Bifidobacterium* is generally considered safe [38] and naturally resistant to certain antibiotics like rifamycin, strains of bifidobacteria, and especially those with probiotic properties, have been suggested for use alongside antibiotic treatments to mitigate side effects [39]. Our findings underscore an improved detection of antibiotic resistance genes in MetaWRAP MAGs in mother-infant samples, providing valuable insights into the structure of the human gut resistome.

Strain tracking, a key aspect of microbiome studies, is crucial for understanding transmission dynamics in various settings, such as monitoring the spread of virulent

strains in hospitals, or vertical transfer events in mother-infant pairs [40–42]. The quality of MAG databases for reference genome-based strain tracking is critical for accurately capturing a complete set of sharing events. Using the improved set of MAGs reconstructed by MetaWRAP resulted in a 25% increase in sharing events detected in the MicrobeMom and Swedish datasets. This improvement can be attributed to (i) the recovery of high-quality MAGs of specific species exclusively by MetaWRAP, leading to the identification of uniquely shared species, and (ii) a higher diversity of MAGs from the same species recovered by MetaWRAP, resulting in more species subgroups (groups of genomes with > 98% ANI) in the reference database, thereby improving real alignments and SNV profiling. For example, in the Swedish dataset, MetaWRAP recovered seven *Bacteroides ovatus* representative genomes compared to only one by MetaBAT2, enabling the capture of a wider range of strains from the same species. However, ten strain sharing events were identified using the MetaBAT2 database which are not detected with MetaWRAP database. Further investigation showed that the increased genome size of the more complete MetaWRAP MAGs led to more bases being compared between samples than using the MetaBAT2 MAGs, causing borderline sharing events to fall below the cutoffs. For instance, *Bifidobacterium pseudocatenulatum* sharing events were detected in the MicrobeMom dataset using MetaBAT2 MAGs as a reference, but not when using the MetaWRAP MAGs. The *B. pseudocatenulatum* MetaWRAP MAG used as the reference was 2.13 megabase pairs (Mbp) in length whereas the MetaBAT2 MAG size was 2.04 Mbp. This 90 kilobase pairs difference reduced the percentage of genome being compared from 53% to < 50% which falls below the recommended threshold (50%). It is evident that a higher number of high-quality MAGs facilitates a greater strain sharing detection rate in mother-infant pairs while at the same time reducing identification of false-positive sharing events. This is particularly important when tracking strains in clinical samples where high accuracy is needed [43]. Overall, this finding is highly interesting, further underscoring the importance of tool selection in accurately tracing microbial transmissions.

Taken together, this work not only highlights the importance of tool selection in metagenomic data analysis but also emphasizes the need to compare and

standardize a wide range of tools for future studies, incorporating isolate genomes and long-read datasets as sequencing technologies continue to advance rapidly. While these MAGs are high-quality draft genomes, they may not have incorporated particular genes that are present in these microorganisms, yet still offer valuable strain-level insights at metagenomic scale in a relatively rapid and efficient manner. This study underlines the importance of tool selection in obtaining accurate and comprehensive insights into microbial structure.

2.5 Methods

2.5.1 Reconstruction of metagenome assembled genomes

Raw metagenomic sequencing reads of MicrobeMom (PRJEB48251), Swedish (PRJNA504891) and Ethiopian (PRJEB6456) samples were downloaded from European Nucleotide Archive. Quality filtering and host contamination removal was performed using read_qc module of MetaWRAP (v 1.2.1) [14,44–46] with default options. The contigs were assembled using the MetaWRAP assembly module with metaspades option [47,48]. The MetaWRAP MAGs were reconstructed using the binning module with three binning algorithms - MetaBAT2, MaxBin2 and Concoct, and the resulting bins with >50% completeness and <10% contamination were refined and reassembled using MetaWRAP's bin_refinement and reassemble_bins modules with default options [11–13]. The MetaBAT2 MAGs were reconstructed using quality filtered reads with MetaBAT2 (v 2.12.1).

2.5.2 Taxonomy and MAGs quality metrics

MAG quality was assessed with CheckM (v 1.0.18) [21] Based on the standards recommended by Bowers *et al.* MAGs with >90% completeness & <5% contamination were considered as high-quality, and those with >50% completeness & <10% contamination were considered medium quality [27]. All other MAGs not meeting these criteria were discarded. Statistical tests were performed in R (v 4.3.1) [49]. Wilcoxon test from ggsignif package (v 0.6.4) [50] was used to determine the significant differences between tools with a p-value threshold of 0.05. Assembly quality metrics were determined using Seqkit (v 1.4) [51], and GTDB-TK (v 1.5.0) with reference

database (v r202) was used to assign taxonomy to each MAG [52,53]. The sequencing depth for each MAG was determined by mapping metagenomic reads of each sample using BBMap (v 38.22)[54].

2.5.3 Identification of ribosomal genes, AMR and strain sharing events

Prodigal was used to predict the coding sequences from each MAG [55]. The 16S rRNA genes were identified from high quality MAGs using Barrnap (v 0.9) [56]. The 42 core ribosomal protein genes described in Mise *et al*. was used as a database for KofamScan (v 1.3.0) to identify these genes in MAGs [19,57]. Antimicrobial resistance genes were identified using Abricate (v 1.0.1) with the CARD database [58,59]. Strain sharing events were identified using InStrain (v 1.8.0) with quality filtered metagenomic reads as input [15]. High quality MAGs were used to construct the InStrain database after dereplication of MAGs at 98% identity using dRep (v 3.2.0) [60]. Bowtie2 (v 2.4.4) was used to map the metagenomic reads to the MAG database [61]. The thresholds of 99.999% popANI and 50% reference genome coverage recommended by InStrain authors were used to identify strain sharing events.

2.6 Data availability

No new data were generated as part of this study, and previously published publicly available data were used for the analysis. The analysis scripts used in this study are available at https://github.com/Metagymomics/MetaBAT2vsMetaWRAP.

2.7 References

1. Consortium, H. M. P. Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–14 (2012).

2. Valles-Colomer, M. *et al.* The person-to-person transmission landscape of the gut and oral microbiomes. *Nature* **614**, 125–135 (2023).

3. Wallen, Z. D. *et al.* Metagenomics of Parkinson's disease implicates the gut microbiome in multiple disease mechanisms. *Nat Commun* **13**, 6958 (2022).

4. Walsh, L. H. *et al.* Comparison of the relative impacts of acute consumption of an inulin-enriched diet, milk kefir or a commercial probiotic product on the human gut microbiome and metabolome. *npj Sci Food* **7**, 41 (2023).

5. Franzosa, E. A. *et al.* Gut microbiome structure and metabolic activity in inflammatory bowel disease. *Nat Microbiol* **4**, 293–305 (2019).

6. Patangia, D. V., Ryan, C. A., Dempsey, E., Stanton, C. & Ross, R. P. Vertical transfer of antibiotics and antibiotic resistant strains across the mother/baby axis. *Trends Microbiol* **30**, 47–56 (2022).

7. Samarra, A. *et al.* Maternal-infant antibiotic resistance genes transference: what do we know? *Gut Microbes* **15**, 2194797 (2023).

8. Sabater, C., Calvete-Torre, I., Ruiz, L. & Margolles, A. Arabinoxylan and Pectin Metabolism in Crohn's Disease Microbiota: An In Silico Study. *Int J Mol Sci* **23**, 7093 (2022).

9. Lee, K. *et al.* Population-level impacts of antibiotic usage on the human gut microbiome. *Nat Commun* **14**, 1191 (2023).

10. Brooks, B. *et al.* Strain-resolved analysis of hospital rooms and infants reveals overlap between the human and room microbiome. *Nat Commun* **8**, 1814 (2017).

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

12. Kang, D. D. *et al.* MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* **7**, e7359 (2019).

13. Alneberg, J. *et al.* Binning metagenomic contigs by coverage and composition. *Nat Methods* **11**, 1144–1146 (2014).

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

15. Olm, M. R. *et al.* inStrain profiles population microdiversity from metagenomic data and sensitively detects shared microbial strains. *Nat Biotechnol* **39**, 727–736 (2021).

16. Diebold, P. J. *et al.* Clinically relevant antibiotic resistance genes are linked to a limited set of taxa within gut microbiome worldwide. *Nat Commun* **14**, 7366 (2023).

17. Hiseni, P., Snipen, L., Wilson, R. C., Furu, K. & Rudi, K. Questioning the Quality of 16S rRNA Gene Sequences Derived From Human Gut Metagenome-Assembled Genomes. *Frontiers in Microbiology* **12**, (2022).

18. Zhang, Y., Ji, P., Wang, J. & Zhao, F. RiboFR-Seq: a novel approach to linking 16S rRNA amplicon profiles to metagenomes. *Nucleic Acids Research* **44**, e99 (2016).

19. Mise, K. & Iwasaki, W. Unexpected absence of ribosomal protein genes from metagenome-assembled genomes. *ISME Communications* **2**, 118 (2022).

20. Nelson, W. C., Tully, B. J. & Mobberley, J. M. Biases in genome reconstruction from metagenomic data. *PeerJ* **8**, e10119 (2020).

21. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* (2015) doi:10.1101/gr.186072.114.

22. Mori, H. *et al.* Assessment of metagenomic workflows using a newly constructed human gut microbiome mock community. *DNA Research* **30**, dsad010 (2023).

23. Meziti, A. *et al.* The Reliability of Metagenome-Assembled Genomes (MAGs) in Representing Natural Populations: Insights from Comparing MAGs against Isolate

Genomes Derived from the Same Fecal Sample. *Applied and Environmental Microbiology* **87**, (2021).

24. Feehily, C. *et al.* Detailed mapping of Bifidobacterium strain transmission from mother to infant via a dual culture-based and metagenomic approach. *Nat Commun* **14**, 3015 (2023).

25. Manara, S. *et al.* Maternal and food microbial sources shape the infant microbiome of a rural Ethiopian population. *Curr Biol* **33**, 1939-1950.e4 (2023).

26. Bäckhed, F. *et al.* Dynamics and Stabilization of the Human Gut Microbiome during the First Year of Life. *Cell Host & Microbe* **17**, 690–703 (2015).

27. Bowers, R. M. *et al.* Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat Biotechnol* **35**, 725–731 (2017).

28. Zeng, S. *et al.* A compendium of 32,277 metagenome-assembled genomes and over 80 million genes from the early-life human gut microbiome. *Nat Commun* **13**, 5139 (2022).

29. Adamberg, K. & Adamberg, S. Selection of fast and slow growing bacteria from fecal microbiota using continuous culture with changing dilution rate. *Microb Ecol Health Dis* **29**, 1549922 (2018).

30. Anthony, W. E., Burnham, C.-A. D., Dantas, G. & Kwon, J. H. The Gut Microbiome as a Reservoir for Antimicrobial Resistance. *The Journal of Infectious Diseases* **223**, S209–S213 (2021).

31. John, J., George, S., Nori, S. R. C. & Nelson-Sathi, S. Phylogenomic Analysis Reveals the Evolutionary Route of Resistant Genes in Staphylococcus aureus. *Genome Biology and Evolution* **11**, 2917–2926 (2019).

32. Hu, Y. *et al.* Metagenome-wide analysis of antibiotic resistance genes in a large cohort of human gut microbiota. *Nat Commun* **4**, 2151 (2013).

33. Bengtsson-Palme, J. *et al.* The Human Gut Microbiome as a Transporter of Antibiotic Resistance Genes between Continents. *Antimicrobial Agents and Chemotherapy* **59**, 6551–6560 (2015).

34. Patangia, D. V., Anthony Ryan, C., Dempsey, E., Paul Ross, R. & Stanton, C. Impact of antibiotics on the human microbiome and consequences for host health. *Microbiologyopen* **11**, e1260 (2022).

35. Rong, S. M. M., Rodloff, A. C. & Stingu, C.-S. Diversity of antimicrobial resistance genes in Bacteroides and Parabacteroides strains isolated in Germany. *Journal of Global Antimicrobial Resistance* **24**, 328–334 (2021).

36. Bookstaver, P. B. *et al.* A Review of Antibiotic Use in Pregnancy. *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy* **35**, 1052–1062 (2015).

37. Serafini, F. *et al.* Insights into Physiological and Genetic Mupirocin Susceptibility in Bifidobacteria. *Applied and Environmental Microbiology* **77**, 3141–3146 (2011).

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

39. Graziano, T. *et al.* The Possible Innovative Use of Bifidobacterium longum W11 in Association With Rifaximin: A New Horizon for Combined Approach? *J Clin Gastroenterol* **50 Suppl 2, Proceedings from the 8th Probiotics, Prebiotics&New Foods for Microbiota and Human Health meeting held in Rome, Italy on September 13-15, 2015**, S153–S156 (2016).

40. Siranosian, B. A. *et al.* Rare transmission of commensal and pathogenic bacteria in the gut microbiome of hospitalized adults. *Nat Commun* **13**, 586 (2022).

41. Li, W. *et al.* Vertical Transmission of Gut Microbiome and Antimicrobial Resistance Genes in Infants Exposed to Antibiotics at Birth. *J Infect Dis* **224**, 1236–1246 (2021).

42. Ferretti, P. *et al.* Mother-to-Infant Microbial Transmission from Different Body Sites Shapes the Developing Infant Gut Microbiome. *Cell Host and Microbe* **24**, 133-145.e5 (2018).

43. Gorrie, C. L. *et al.* Key parameters for genomics-based real-time detection and tracking of multidrug-resistant bacteria: a systematic analysis. *The Lancet Microbe* **2**, e575–e583 (2021).

44. Agarwala, R. BMTagger: Best Match Tagger for removing human reads from metagenomics datasets. https://ftp.ncbi.nlm.nih.gov/pub/agarwala/bmtagger/ (2011).

45. Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

46. Krueger, F. *et al.* FelixKrueger/TrimGalore: v0.6.10 - add default decompression path. (2023) doi:10.5281/ZENODO.7598955.

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

48. Prjibelski, A., Antipov, D., Meleshko, D., Lapidus, A. & Korobeynikov, A. Using SPAdes De Novo Assembler. *Current Protocols in Bioinformatics* **70**, e102 (2020).

49. R: The R Project for Statistical Computing. https://www.r-project.org/.

50. Ahlmann-Eltze, C. & Patil, I. ggsignif: R Package for Displaying Significance Brackets for 'ggplot2'. (2024) doi:10.31234/osf.io/7awm6.

51. Shen, W., Le, S., Li, Y. & Hu, F. SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. *PLoS ONE* **11**, e0163962 (2016).

52. Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* **36**, 1925–1927 (2019).

53. Matsen, F. A., Kodner, R. B. & Armbrust, E. V. pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics* **11**, 538 (2010).

54. BBMap. *SourceForge* https://sourceforge.net/projects/bbmap/ (2023).

55. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).

56. Seemann, T. tseemann/barrnap. (2024).

57. Aramaki, T. *et al.* KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics* **36**, 2251–2252 (2020).

58. Seemann, T. tseemann/abricate. (2024).

59. Jia, B. *et al.* CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic Acids Res* **45**, D566–D573 (2017).

60. Olm, M. R., Brown, C. T., Brooks, B. & Banfield, J. F. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J* **11**, 2864–2868 (2017).

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

3 Chapter 3

Enrichment of Human Milk Oligosaccharide Utilisation Genes in Bacteria Shared Between Mothers and Infants

3.1 Abstract

The establishment of the infant gut microbiome is influenced by several factors including maternal microbial transfer and infant diet. Studies have shown that human milk oligosaccharides (HMOs) in breast milk can enrich bacteria such as *Bifidobacterium* and *Bacteroides* in the neonatal gut. The genetic makeup of bacterial strains, especially when their genome contains genes involved in HMO metabolism, may play a crucial role in not only determining which strains are successfully transmitted from mother to infant, but also their ability to persist within the infant gut. We explored this phenomenon through a metagenomics-based analysis of samples collected 3-months postpartum from a mother-infant dyad cohort. Using metagenome assembled genomes and microdiversity-aware strain-level profiling, we found that *Phocaeicola vulgatus, Phocaeicola dorei, Bacteroides uniformis* and *Bifidobacterium bifidum* were frequently shared between mother-infant dyads despite not being among the dominant species in samples collected from mothers. These shared species were also predicted to possess a higher number of HMO-related GHs with signal peptides when compared to other maternal microbes. Our analysis indicates that bacterial strains encoding secreted HMO utilisation enzymes are more likely to be transferred from mother to infant. Overall, our study provides insights into the genetic factors that play a determining role in bacterial strain transmission among mother-infant pairs.

3.2 Introduction

The seeding and establishment of the infant gut microbiome has the potential to influence long-term health outcomes[1–4]. This microbial ecosystem is believed to play pivotal roles in immune system maturation, pathogen protection and nutrient utilisation[5–7], and perturbations in the early life gut microbiome have been associated with various adverse health outcomes including asthma, metabolic disorders and allergies[8–12]. The composition of the gut microbiome, particularly the dominance of *Bifidobacterium* and *Bacteroides* species, is well documented and these microorganisms are generally believed to be primarily acquired through vertical transmission from mother to infant[13–18]. However, the factors that influence which bacterial strains successfully transfer and colonise the infant gut require further elucidation.

Diet is one of the major factors impacting gut-associated microbial communities, with the typical infant diet predominantly consisting of human milk, which contains complex human milk oligosaccharides (HMOs) that are known to influence gut microbiome composition by selectively feeding certain bacterial species[19–21]. In some instances, the ability to directly or indirectly utilise HMOs varies between strains of the same species, indicating that strain-level differences in their genetic makeup influence their ability to establish and persist in the infant gut[22,23].

Bifidobacteria are the most widely studied of the microorganisms that dominate the infant gut microbiota and it is well established that certain species can utilise a wide variety of HMOs[24–26]. Bifidobacterial species such as *Bifidobacterium breve* and *Bifidobacterium longum* typically transport oligosaccharides into the cell and metabolise HMOs using intracellular glycosyl hydrolases (GHs)[27,28]. Recent reports suggest that certain species of the closely-related genera *Bacteroides* and *Phocaeicola* also possess the ability to utilise these complex sugars from human milk[29,30]. Such species appear to degrade complex sugars extracellularly by cell envelope-associated GHs, which allows them to metabolise a wide array of glycans[30,31]. This capability may have a selective advantage in long-term persistence even after the introduction of solid foods[16,32,33].

In our previous study, we showed that species from the genera *Bacteroides* and *Bifidobacterium* were the most frequently transferred from mother to infant up to 1-month postpartum[14]. To understand the dynamics of strain sharing and the bacterial genetic factors associated with transfer, we extended our previous dataset through shotgun metagenomics-based analysis of infant and mother stool samples collected from the same cohort 3 months postpartum.

3.3 Results

3.3.1 Taxonomic landscape of the early life microbiome

Shotgun sequencing-based analysis of additional faecal samples from the 3-month post-partum timepoint of the MicrobeMom cohort (ISRCTN53023014) expanded the original dataset of 1012 samples to 1217 through the addition of 109 maternal and 96 infant stool samples, comprising 87 mother-infant pairs, 9 unmatched infant samples, and 22 unmatched maternal samples. All unmatched samples had a matched sample at an earlier timepoint, allowing us to infer strain sharing information for that mother-infant pair. A median of 9.3 and 8.6 million quality filtered microbial reads were obtained through sequencing of the DNA isolated from the infant and mother faecal samples, respectively, at the 3-month (post-partum) timepoint. As expected, the Shannon diversity of the infant microbiome samples was significantly lower than maternal samples at both 1-month and 3-month post-partum timepoints. This significant difference was observed in both unpaired analyses (Wilcoxon rank-sum test, 1-month = $p < 2.2x10^{-16}$ and 3-months = $p < 2.2x10^{-16}$) and paired analyses (Wilcoxon signed-rank paired test, $p < 2.2 \times 10^{-16}$ at 1 month and $p = 2.37 \times 10^{-14}$ at 3 months). Additionally, there was a significant difference in the beta diversity of the infant microbiome samples compared to maternal samples (PERMANOVA, $p = 0.001$) (Fig 3.1a and Fig 3.1b).

As previously reported, taxonomic profiling of infant faecal sample-derived metagenomes identified *Escherichia coli* and *Enterococcus faecalis* as the major early residents of the infant stool with mean relative abundances of 18.5% and 5.7%, respectively, at delivery and 11.9% and 6.9%, respectively, at 1-week (Fig 3.1c & Supplementary Figure 3.1). At the 1-month and 3-month post-partum timepoints, certain *Bifidobacterium* species, in particular *Bifidobacterium longum*,

Bifidobacterium breve and *Bifidobacterium bifidum*, were highly prevalent and abundant with a mean relative abundance of 17.1%-26.2%, 17%-14% and 8.8%-9%, respectively. *Phocaeicola vulgatus* was the fifth most prevalent and abundant species on average across all timepoints, with a mean relative abundance of 3%. We further investigated the impact of factors such as breast feeding at hospital discharge, maternal exposure to antibiotics in labour and delivery mode on infant microbiota composition at species level. This analysis indicated significant reductions in Shannon diversity in instances of maternal antibiotic exposure during labour at the 1-month post-partum timepoint ($p = 0.044$) and delivery mode at the 1-week timepoint ($p = 0.014$) (Supplementary Figure 3.2), although these differences were not apparent at the 3-month timepoint. No significant compositional differences were found among breastfeeding groups at any timepoint. The impact of these factors is only evident at early timepoints, indicating that the infant microbiome has recovered by the 3-month post-partum mark.

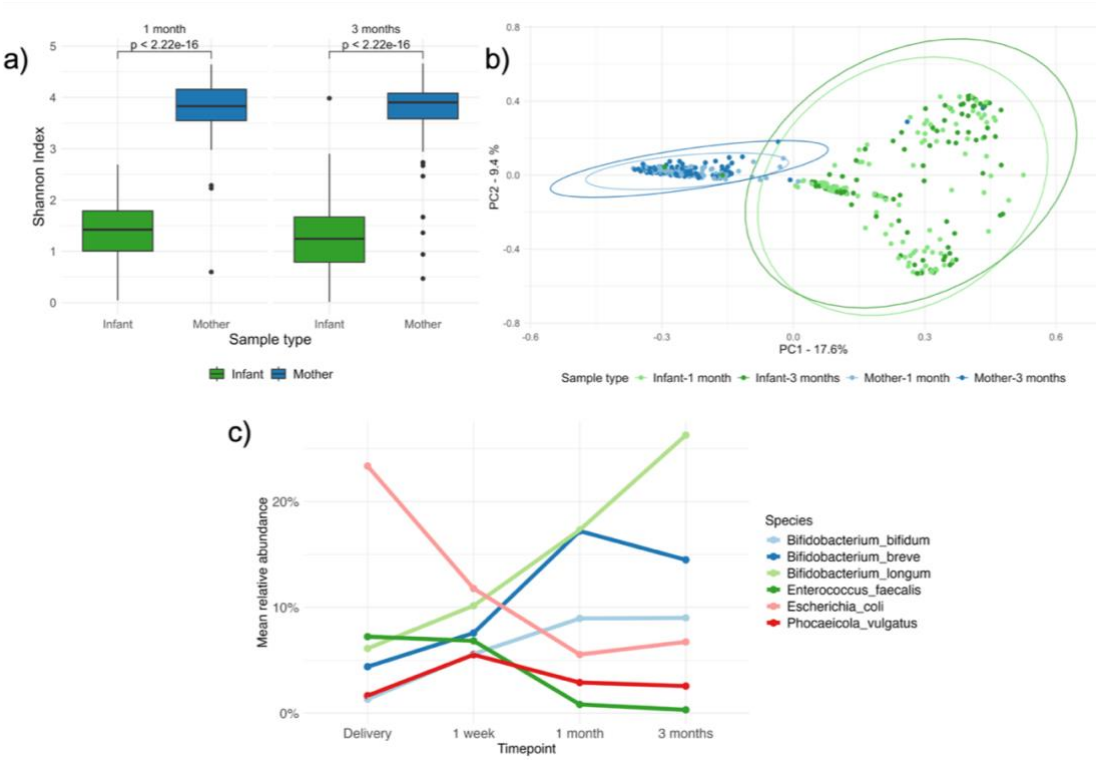
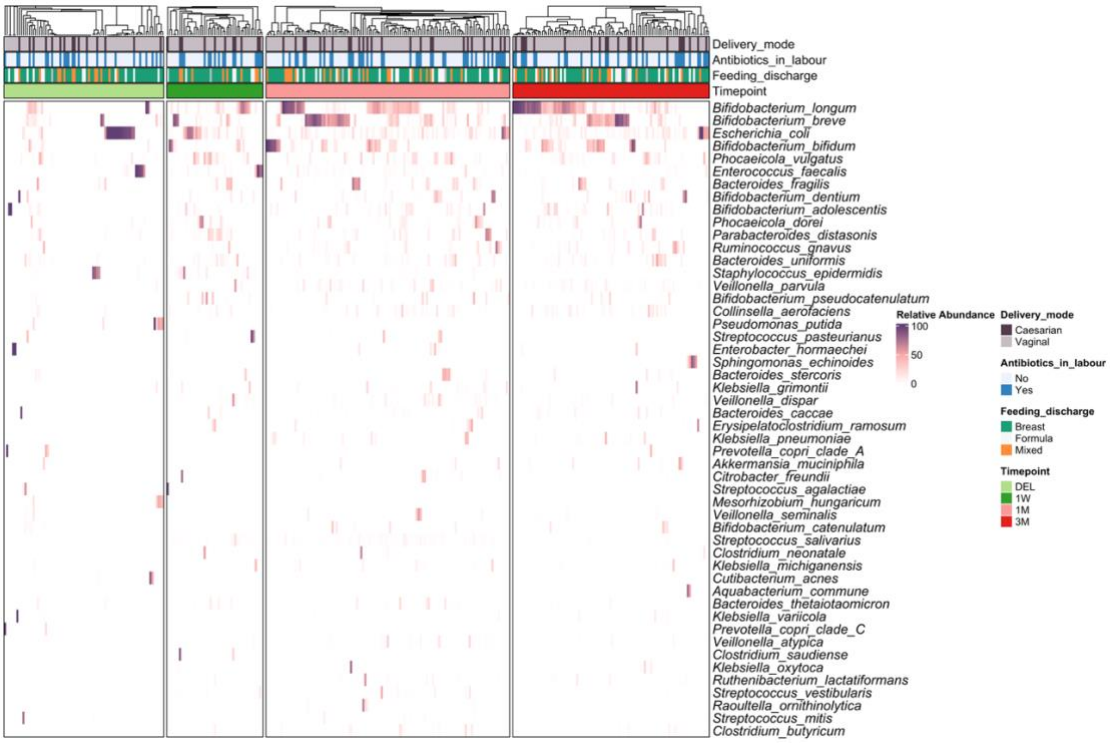


Figure 3.1 | Taxonomic composition of early life infant microbiome

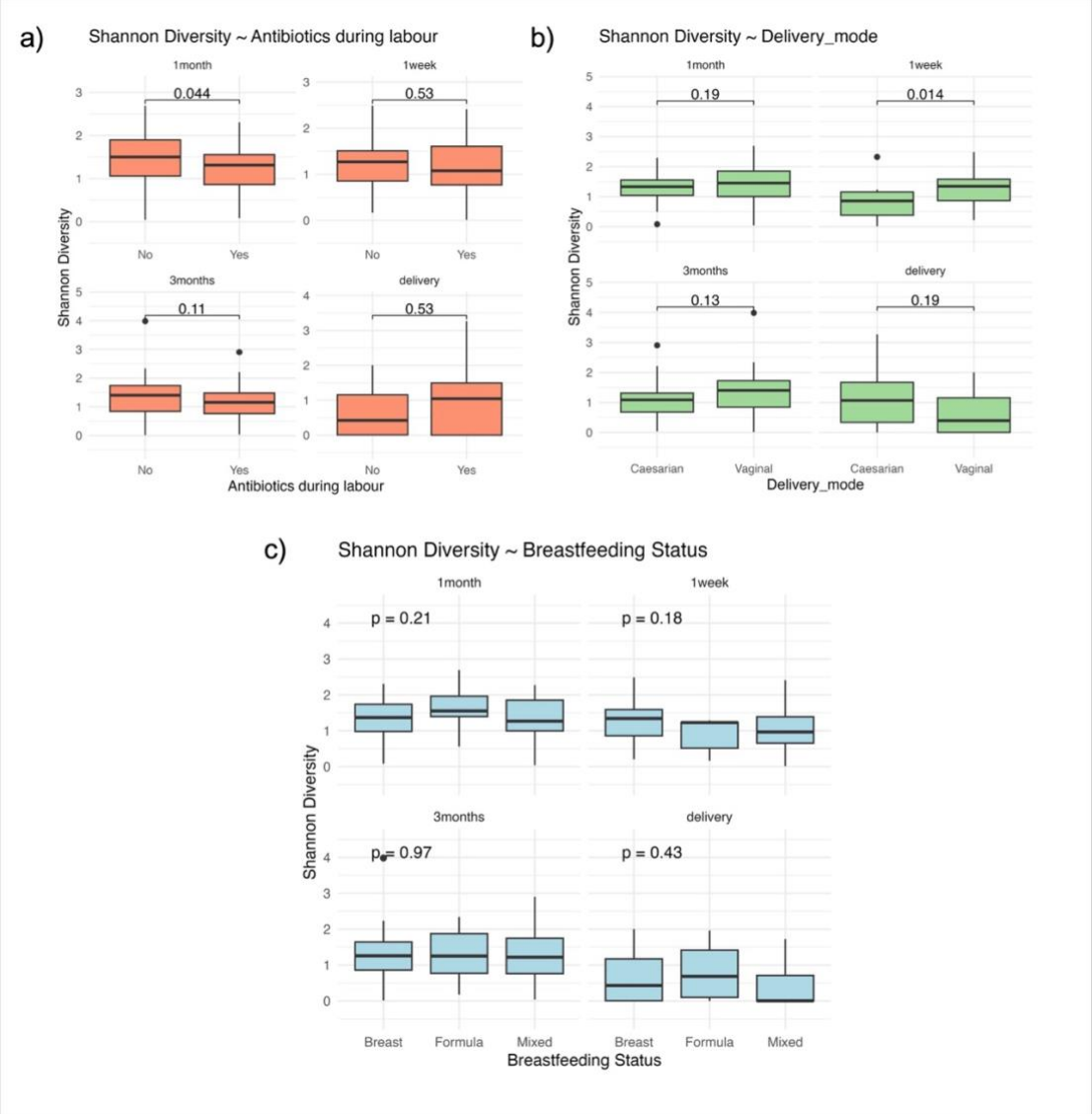
a). Boxplots showing the Shannon diversity index in mother and infant samples at 1-month and 3-months timepoints. The significance was calculated using Wilcoxon test

in R with **** representing a p-value of < 0.0001. b). PCoA ordination of beta diversity (Bray-Curtis dissimilarity) at species level compositional profiles of mother and infant samples. c). Line plot showing the mean relative abundance of the most abundant species in the infant gut microbiome at each timepoint.



Supplementary Figure 3.1 | Taxonomic composition of infant gut microbiome

Heatmap representing taxonomic composition of top 50 species of infant stool samples at various timepoints (E – Early (16 weeks), L- late (34 weeks), 1M- 1 month and 3M – 3 months). Taxonomic profiling was performed using MetaPhlAn based on clade-specific marker genes. *Bifidobacterium longum* subsp. *longum* and *B. longum* subsp. *infantis* could not be distinguished due to the resolution limits of the method.



Supplementary Figure 3.2 | Impact of covariates on infant microbiome composition

Boxplots representing association between infant shannon diversity index and with factors such as a) Antibiotics during labour, b) Delivery mode and c) type of feeding at discharge at various timepoints. The p-values were shown on top of each plot.

3.3.2 Species common to mother and infant gut microbiomes

We compared the taxonomic profiles of all maternal gut metagenomes with those of all infant gut metagenomes to identify common species. Focussing on the 1-month and 3-month post-partum timepoints, we detected 39 species in common between all mother and infant stool metagenomes at >0.1% relative abundance (Fig 3.2a and Supplementary Table 3.1). At the 1-month timepoint, we identified 25 species that are commonly present between the observed mother and infant taxonomic profiles, while 33 species were commonly present at the 3-month timepoint, 19 of which overlapped with 1-month. Species from the genera *Bacteroides* (20%, 8/39), *Bifidobacterium* (13%, 5/39) and *Phocaeicola* (8%, 3/39) were most frequently found in both mother and infant gut metagenomes. When comparing the relative abundances of these genera between infant and mother samples, we observed an increase in the abundance of *Bifidobacterium* from 1-month to 3-months post-partum (Infant – 49% to 57% and Mother – 3% to 6%) (Fig 3.2b). In contrast, *Phocaeicola* and *Bacteroides* abundance decreased from 1-month to 3-months post-partum in both mother and infants.

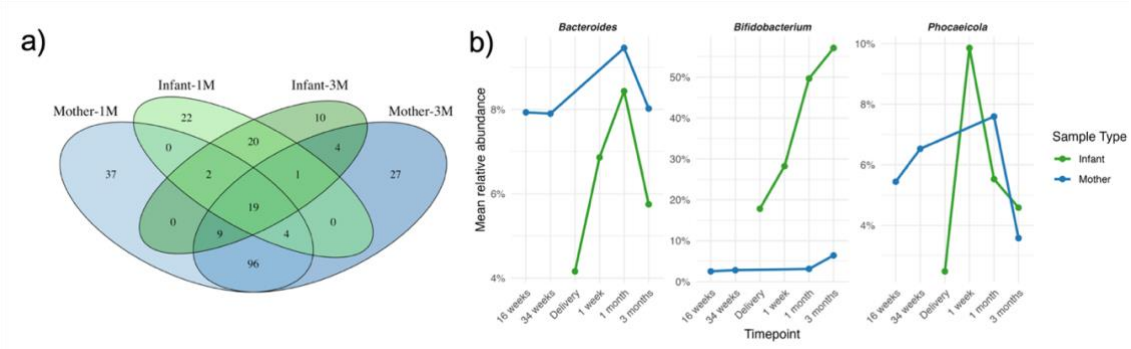


Figure 3.2 | Common species in mother and infant gut microbiomes

a) Venn diagram showing the number of common species between mother and infant samples at the 1-month and 3-month post-partum timepoints. b) Line plots showing the mean relative abundance of the most abundant genera in the infant gut microbiome in comparison with maternal data.

Supplementary table 3.1 | Common species between mother and infant stool metagenomes

species	timepoint	mother_mean_relative_abundance	infant_mean_relative_abundance
Akkermansia_muciniphila	3M	2.16306762	1.19606635
Alistipes_onderdonkii	3M	2.51910624	0.14299
Alistipes_putredinis	3M	3.97702037	0.15176563
Aquabacterium_commune	3M	0.14231743	1.37958927
Bacteroides_caccae	3M	0.32039404	0.50734396
Bacteroides_fragilis	3M	0.22349844	2.33774802
Bacteroides_ovatus	3M	0.3753144	0.1138001
Bacteroides_stercoris	3M	1.36619725	0.35540958
Bacteroides_uniformis	3M	3.95934835	2.09202479
Bifidobacterium_adolescentis	3M	1.66014706	2.45986729
Bifidobacterium_bifidum	3M	0.50982092	9.0045074
Bifidobacterium_catenulatum	3M	0.10482413	0.74016
Bifidobacterium_longum	3M	3.4858678	26.2744644
Bifidobacterium_pseudocatenulatum	3M	0.46140872	0.6240749
Bilophila_wadsworthia	3M	0.26210028	0.10756958
Blautia_wexlerae	3M	0.63087422	0.15427063
Collinsella_aerofaciens	3M	0.88449725	1.47102125
Eggerthella_lenta	3M	0.13724339	0.20785792
Escherichia_coli	3M	0.25112495	6.73769896
Faecalibacterium_prausnitzii	3M	7.42313495	0.13028896
Flavonifractor_plautii	3M	0.1975311	0.25881521
Hungatella_hathewayi	3M	0.27164083	0.58056229
Intestinimonas_massiliensis	3M	0.32943064	0.3231175
Parabacteroides_distasonis	3M	0.45359945	0.97663969
Parabacteroides_merdae	3M	0.45336092	0.19327094
Phocaeicola_dorei	3M	0.62117505	1.78993729
Phocaeicola_vulgatus	3M	2.32306257	2.56291042

Prevotella_copri_clade_A	3M	2.58654982	0.41059479
Prevotella_marseillensis	3M	0.3927478	0.18591333
Roseburia_faecis	3M	0.54102385	0.12217385
Ruminococcus_torques	3M	0.28158294	0.12816292
Ruthenibacterium_lactatiformans	3M	0.25869147	0.24782958
Sutterella_wadsworthensis	3M	0.65034862	0.35107792

3.3.3 Shared strains in mother-infant pairs

Next, we characterised strain sharing events from mother to infant using InStrain with high-quality metagenome-assembled genomes as a reference database with a strain identity threshold of 99.999% population-average nucleotide identity (popANI). A sharing event was considered when the same strain was identified in an infant sample and a paired maternal sample at an earlier timepoint. In total, we identified 65 strain sharing events across 14 different genera at delivery, and 1-week, 1-month and 3-month post-partum timepoints (Figure 3.3a). At the delivery and 1-week timepoints combined, we identified 8 sharing events, with 5 of these strains persisting in infants at later timepoints – 2 strains persisted to the 1-month timepoint, and 3 strains persisted to the 3-month timepoint (Figure 3.3b). At the 1-month timepoint, we identified 27 new sharing events that had not been observed at the earlier time point, of which 6 (6/27) strains persisted to the 3-month timepoint. At the 3-month timepoint, we identified 30 new sharing events that had not been identified in previous timepoints.

In total, 55% of the identified sharing events correspond to strains from the genera *Bacteroides*, *Phocaeicola* or *Bifidobacterium*. Among members of the *Phocaeicola* genus, we found 11 sharing events for *P. vulgatus* and 7 for *P. dorei*. For the *Bacteroides* genus, there was evidence of 9 sharing occurrences for *B. uniformis* and 2 sharing events for *Bacteroides stercoris*. In the *Bifidobacterium* genus, 6 strain sharing occurrences were observed for *B. bifidum*, 3 for *B. adolescentis* and 2 for *B. longum* subsp. *longum.* We also compared the relative abundance of frequently shared species and observed that *Bifidobacterium* species (*B. bifidum*, *B. adolescentis*

and *B. longum*) abundance increased from 1-month to 3-month timepoint in both mother and infant samples (Fig 3.3c), while *Phocaeicola* species (*P. dorei and P. vulgatus*) relative abundance decreased. We observed an increase in the relative abundance of *B. uniformis* in both mother and infant faecal samples at 1-month and 3-month timepoints. When investigating the impact of covariates on strain sharing in mother-infant pairs, our analysis identified 4 factors that were significantly associated. Delivery mode showed the most significant association (Chi-square test, $\chi^2 = 13.2,\ p = 0.00028$) (Supplementary Figure 3.4). Antibiotics during labour ($\chi^2 = 10.61, p = 0.0011$) was the next most significantly associated factor impacting strain sharing along with labour onset ($\chi^2 = 10.58, p = 0.014$) and membrane rupture ($\chi^2 = 11.47, p = 0.022$).

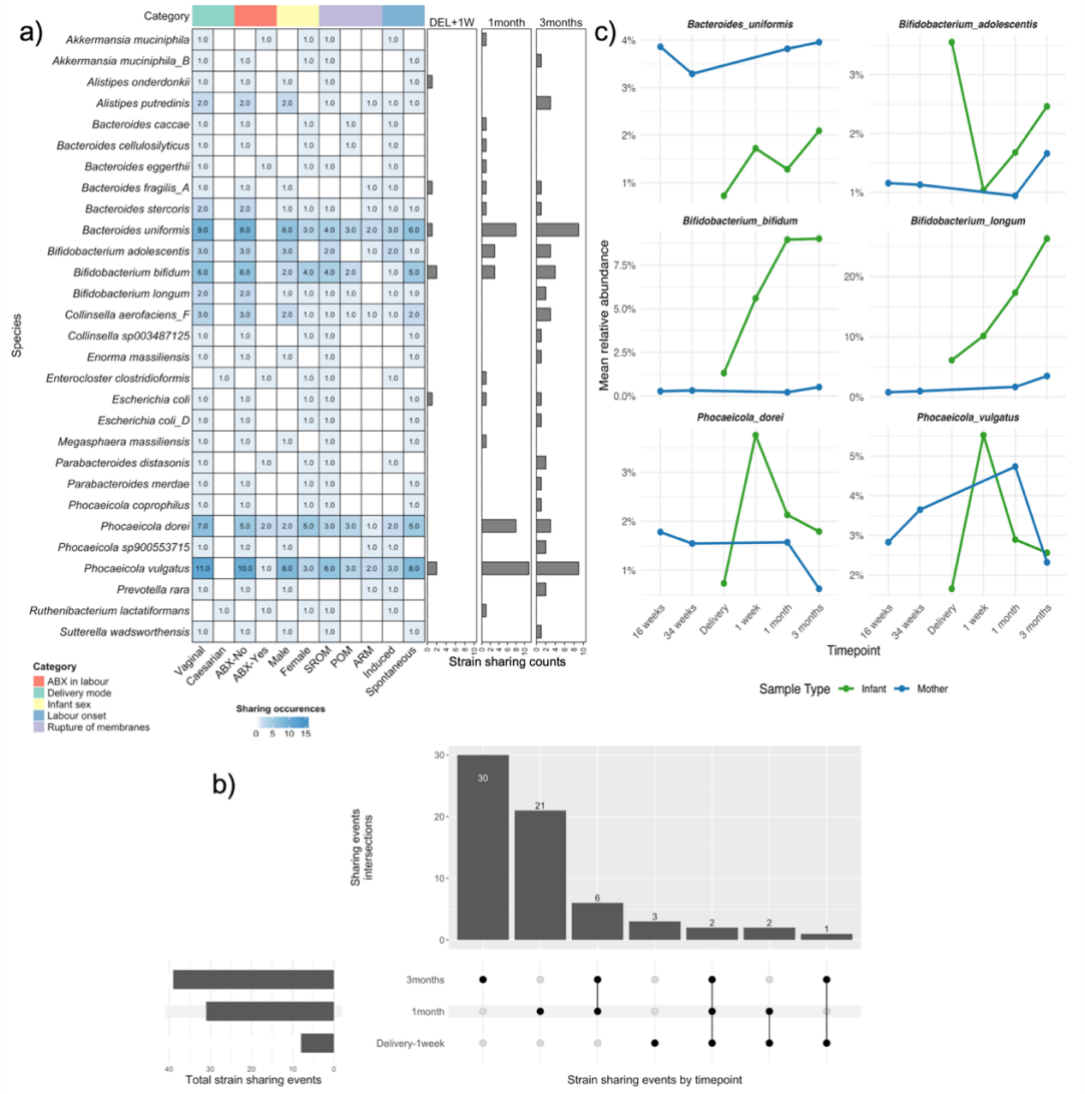
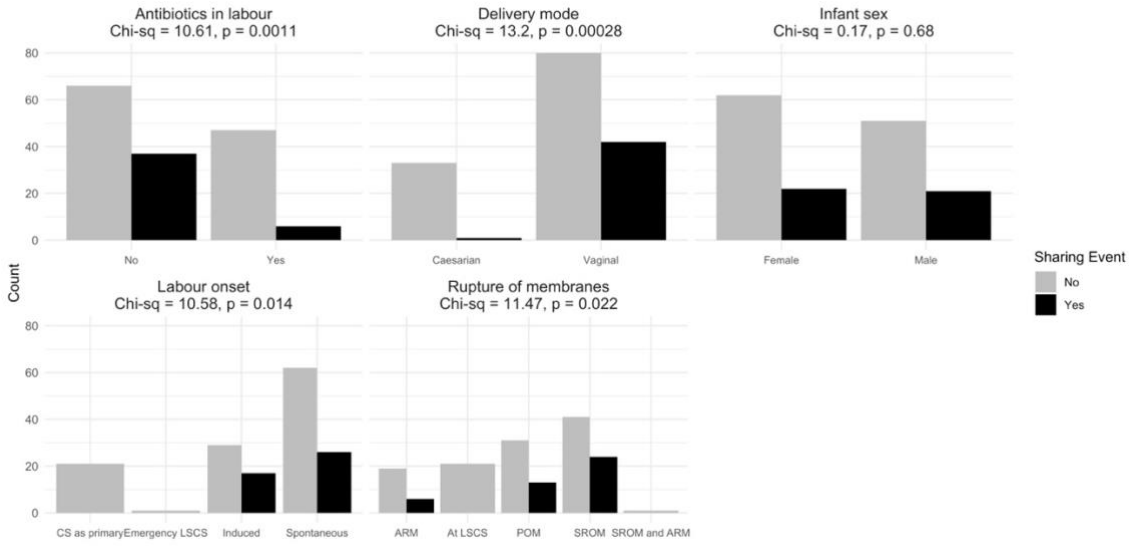


Figure 3.3 | Shared species in mother-infant pairs

a) Heatmap showing identified strain transfer events from mother to infant with available metadata (ABX – Antibiotics, SROM - Spontaneous rupture of membranes, POM - Puncture of membranes and ARM - Artificial rupture of membranes). The bars on the right represent the total number of sharing events for a given species at Delivery-1week (DEL+1W), 1month and 3-month timepoints. b) Upset plot showing strain sharing events identified at each timepoint and their overlap. Horizontal bars show the total number of sharing events detected at each timepoint. Vertical bars show the total number of sharing events detected at each timepoint and their overlap. c) Relative abundance of frequently shared species in infant and maternal faecal samples from 16 weeks pregnancy to 3-months post-partum.



Supplementary Figure 3.4 | Association of covariates with sharing events

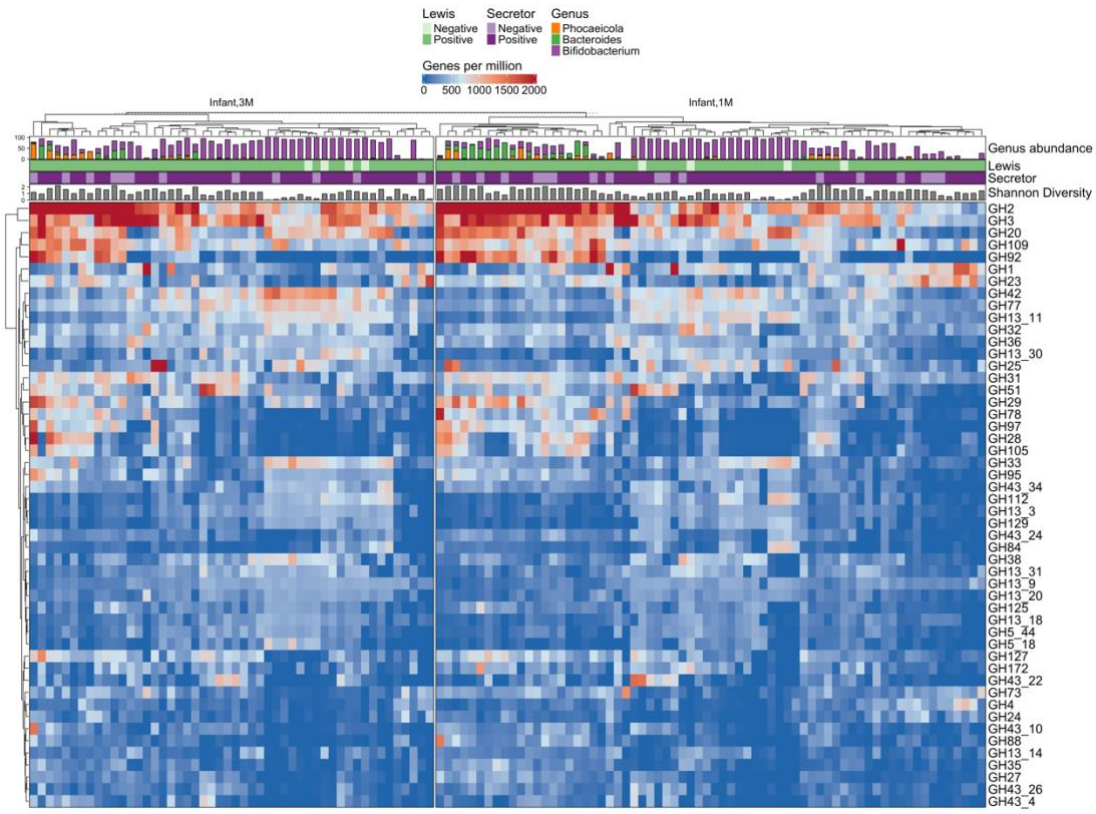
Each barplot facet represents an association with covariates and strain sharing events between mothers and infants pairs. The significant differences were determined using two-sided, Chi-square test. Strain-sharing events were first identified using strain-resolved metagenomic analysis (inStrain) based on the MAGs from maternal and infant stool samples. Each mother-infant pair was then categorised into binary outcomes, shared (1-Yes) or not shared (0-No), based on the presence or absence of at least one shared strain for each covariate category. Counts of these binary

outcomes across covariate groups were subsequently analysed using two-sided Chi-square tests to assess statistical significance.

3.3.4 Characterisation of glycosyl hydrolases in early life infant gut microbiome

We determined the glycosyl hydrolase (GH) gene profile from the metagenomes of the infant stool samples at 1-month and 3-month post-partum to understand the carbohydrate utilisation potential of the early life microbiome. This analysis revealed 207 and 195 GH-encoding genes corresponding to infant stool samples at the 1-month and 3-month timepoints, respectively (Fig 3.4a). A total of 52 GH genes were significantly (Supplementary Table 3.2) enriched in infant samples with a higher relative abundance (>30%) of *Bacteroides* and *Phocaeicola* species compared to *Bifidobacterium*-dominant samples. Of these 52 GH genes, 7 were predicted to be HMO utilisation genes based on the 11 known HMO metabolism-associated GHs[19,20,29] (Supplementary table 3.3). The HMO utilisation gene families most positively associated with *Bacteroides*/*Phocaeicola* rich samples were GH2 and GH92 (*p = 2.08e-15* and *p = 3.72e-6*). In contrast, GH1, GH25, GH77, GH42, GH36 and GH13 family genes were more abundant in *Bifidobacterium* dominant stool samples.

a)



b)

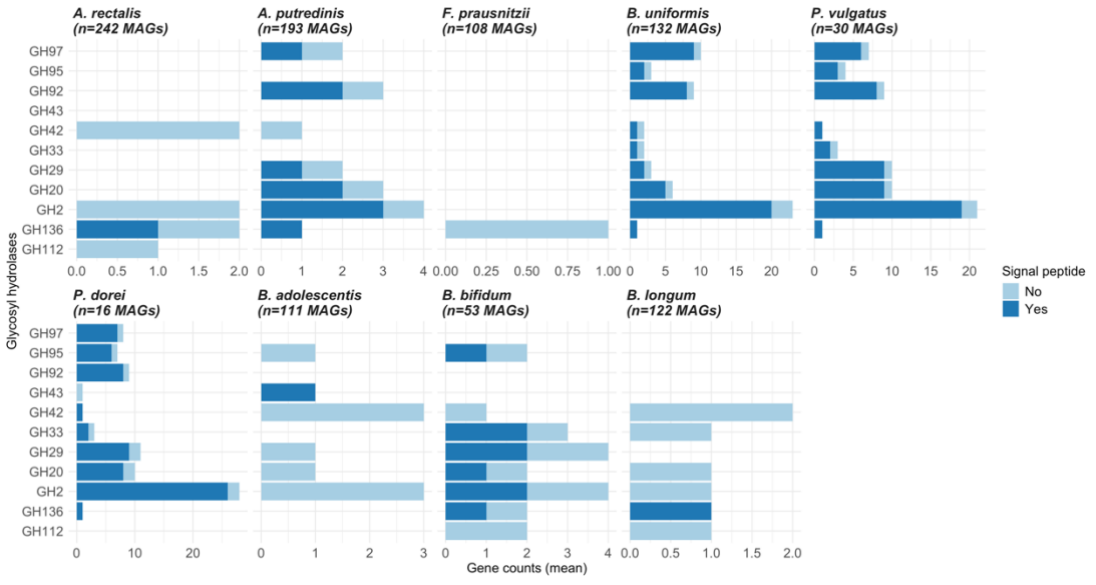


Figure 3.4 | Glycosyl hydrolase profiles in early life infant microbiome

a) Heatmap showing the glycosyl hydrolase abundance in infant metagenomes at 1-month and 3-month timepoints. The bars on the top represent the Shannon diversity index, secretor status, Lewis status and within-sample abundance of the three most

shared genera across the entire dataset. **b)** Distribution of HMO-related glycosyl hydrolases with signal peptides found in metagenome-assembled genomes shown as stacked barplots.

Supplementary table 3.2 | Glycosyl hydrolases that are enriched in *Bacteroides* and *Phoaceaicola* abundant samples. Columns indicate enzyme family (CAZyme), sum of squares (sumsq), mean sum of squares (meansq), test statistic (statistic), p-value (p.value), and regression coefficient indicating direction and magnitude of enrichment (coef).

Cazyme	sumsq	meansq	statistic	p.value	coef
GH2	40216740.4	40216740.4	84.1065344	2.07712343568306e-15	1588.36402
GH92	8951406.29	8951406.29	25.2367945	3.72471604197102e-06	800.013191
GH29	7312901.11	7312901.11	79.2116901	1.78498725636768e-14	684.098444
GH20	6817639.77	6817639.77	40.5849573	4.30066864841641e-09	656.191874
GH3	5374865.04	5374865.04	13.287786	0.00040119	580.671335
GH97	2678957.15	2678957.15	37.343272	7.2222338557561e-08	447.805902
GH28	1874310.39	1874310.39	9.35049535	0.00301002	364.042576
GH95	1123685.23	1123685.23	28.3596528	5.68955927879403e-07	267.893644
GH127	1019829.09	1019829.09	18.5340061	3.84235170647364e-05	256.267838
GH16_3	833116.08	833116.08	52.1299905	6.10652470633026e-10	245.988895
GH109	902904.291	902904.291	6.54820066	0.01178601	237.994798
GH35	797470.079	797470.079	56.9299245	2.45190451575812e-11	227.904553
GH105	730335.064	730335.064	8.21318608	0.005177	220.098336
GH31	665673.634	665673.634	8.70793588	0.00385283	204.864723
GH76	484987.837	484987.837	19.8662136	4.33798033584587e-05	203.203184
GH23	632292.972	632292.972	5.69636052	0.01873924	200.566328

GH130	487004.712	487004.712	26.7558869	2.09945681884655e-06	186.602774
GH78	466998.686	466998.686	4.16141632	0.04410182	174.614587
GH144	325489.957	325489.957	20.2758369	4.15276067496645e-05	169.468882
GH171	361844.258	361844.258	105.300472	3.77616789935952e-15	163.533632
GH88	384500.247	384500.247	10.1536258	0.00200058	160.40485
GH43_10	254141.437	254141.437	6.59980223	0.01192369	130.610391
GH106	159810.392	159810.392	7.23447073	0.0092115	121.146459
GH27	220455.935	220455.935	12.9300258	0.00052763	120.925135
GH43_31	161230.913	161230.913	30.1378256	1.34285679518559e-06	118.704545
GH63	169169.669	169169.669	17.8241333	5.77960800965091e-05	105.929503
GH89	114012.618	114012.618	9.97999637	0.00221633	88.0607992
GH43_28	76112.8388	76112.8388	6.40113251	0.01524277	87.7426345
GH125	112287.582	112287.582	5.90883892	0.01679306	84.9443482
GH115	95229.0626	95229.0626	7.71772451	0.007196	84.1561224
GH154	99589.1298	99589.1298	9.34989061	0.00303204	82.5972941
GH43_3	77329.4168	77329.4168	29.08762	1.88650092856805e-06	82.208279
GH18	99631.3655	99631.3655	6.07015266	0.01562804	81.1795903
GH57	78557.8224	78557.8224	40.1086284	3.59776440702414e-08	77.4922356
GH133	69937.6271	69937.6271	24.8043598	4.29590494107784e-06	70.5412573
GH159	33008.2288	33008.2288	11.8947071	0.00193102	69.3808333
GH141	60864.8194	60864.8194	12.5630072	0.00079372	68.7388684
GH26	53658.7415	53658.7415	4.02602702	0.04939358	66.1506872
GH43_19	33343.9189	33343.9189	5.7163235	0.0203234	53.0695701

GH13_38	35600.6045	35600.6045	20.3343346	3.21609026486974e-05	52.3643338
GH13_8	35582.3299	35582.3299	21.8546101	1.7986662657719e-05	52.3508922
GH5_7	22122.2148	22122.2148	17.5435268	0.00020564	51.8292286
GH13_36	21610.1498	21610.1498	9.55460093	0.00367236	50.4581351
GH30_4	31882.5675	31882.5675	7.16122021	0.00970918	49.7503171
GH13_10	31928.2773	31928.2773	16.7822577	0.0001237	49.6778237
GH43_9	22404.5409	22404.5409	4.6237731	0.0368219	47.5318992
GH116	18652.6342	18652.6342	7.11568788	0.01127056	45.5898714
GH142	17661.8947	17661.8947	9.8028241	0.00278936	37.3411316
GH43_1	10323.7743	10323.7743	7.161034	0.01049978	35.2443663
GH137	13685.465	13685.465	6.52676088	0.01371743	33.6897416
GH139	10460.5987	10460.5987	8.16417959	0.00639575	33.2195077
GH165	3363.87559	3363.87559	7.83422175	0.00809581	21.2693034

Supplementary table 3.3 | 11 known HMO-associated GHs from literature

GHs	**Function**	**Substrates**	**Evidence**
GH2	β-galactosidase	Lactose (Lac), lacto-N-neotetraose (LNnT)	Biochemical assay[19], Transcriptomics[20]
GH20	β-hexosaminidase	Lacto-N-triose II	Biochemical assay[19], Transcriptomics[20]
GH29	Fucosidase	3-FL, LNFP II/III (Fucosylated sugars)	Functional genetics, Transcriptomics[20]
GH33	Sialidase	3'-SL, 6'-SL (Sialylated sugars)	Biochemical assay[19], Transcriptomics[20]
GH42	β-galactosidase	LNT (Lacto-N-tetrose), LNB (Lacto-N-biose)	Biochemical assay[19]

GH95	Fucosidase	2’-FL (2'-fucosyllactose), LNFP I (lacto-N-fucopentaose I)	Biochemical and culture-based assay[19], Transcriptomics[20]
GH112	GNB/LNB Phosphorylase	LNB	Biochemical assay[19]
GH136	Lacto-N-biosidase	LNT	Biochemical assay[19]
GH92	a-mannosidase	2’-FL, DFL, 3’-SL, 6’-SL, and LNT, LNnT (General transcriptional response)	Transcriptomics[29]
GH43	Arabinofuranosidase, xylosidase	2’-FL, DFL, 3’-SL, 6’-SL, and LNT, LNnT (General transcriptional response)	Transcriptomics[29]
GH97	α-Glucosidase	2’-FL, DFL, 3’-SL, 6’-SL, and LNT, LNnT (General transcriptional response)	Transcriptomics[29]

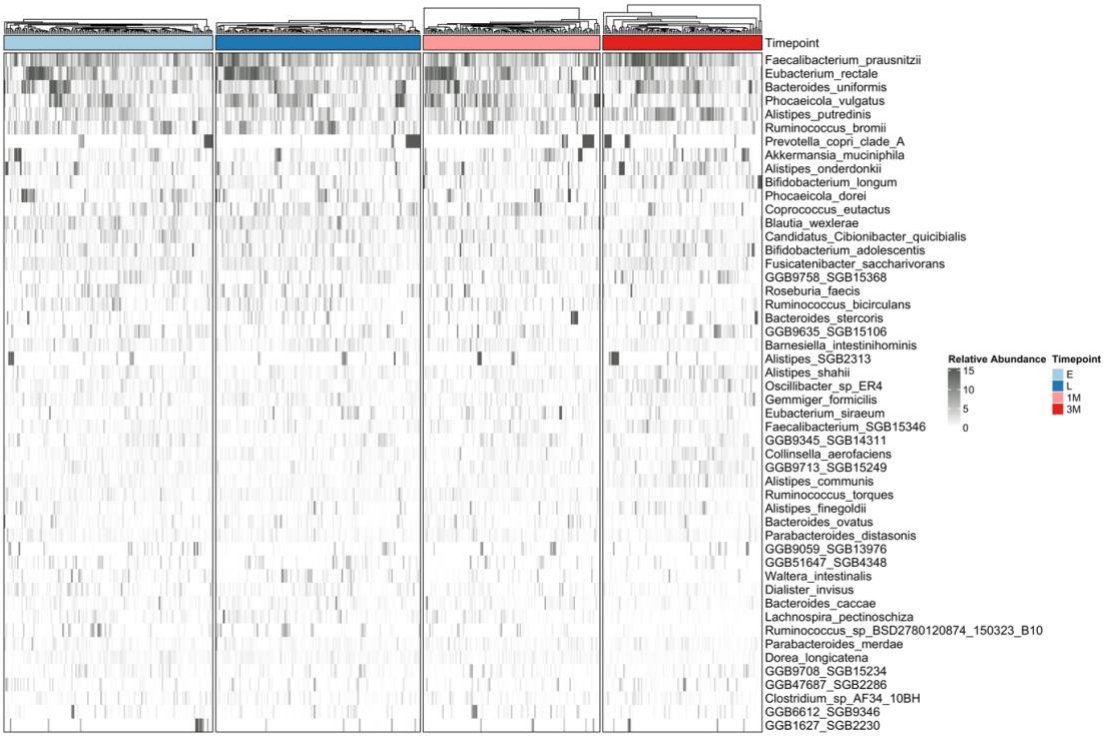
3.3.5 Carbohydrate utilisation gene profiles of frequently transferred bacterial species

We further investigated the metagenome assembled genomes (MAGs) reconstructed from this dataset to predict the presence and diversity of glycosyl hydrolases encoded by species that are frequently shared between mother-infant dyads. We found that the GH2 (Lac and LNnT), GH3 and GH97 (DFL and 6’-SL) and GH92 (DFL and 6’-SL) gene families were abundant in *B. uniformis*, *P. dorei* and *P. vulgatus* compared to *Bifidobacterium* species MAGs (Supplementary Fig 3.5). Members of GH20 (LNT II), GH29 (3-FL and LNFP II/III), GH105, GH43_10 (DFL and 6’-SL), GH28 were particularly abundant in *P. dorei* and *P. vulgatus* compared to *Bifidobacterium* species and *B. uniformis*.

Next, we investigated the presence of 11 known HMO metabolism-associated GHs in MAGs from frequently transferred species. This analysis revealed that these genes were prevalent in *B. uniformis* (9/11), *P. vulgatus* (9/11) and *P. dorei* (10/11) (Fig 4b). *B. bifidum* (8/11) and *B. longum* subsp*. longum* (6/11) species were shown to possess the next most abundant HMO related genes. We looked at the predicted presence of signal peptides in the HMO-associated GHs and found *B. uniformis* (9/9), *P. vulgatus* (9/9) and *P. dorei* (9/10) to have the highest number of signal peptides in their GHs.

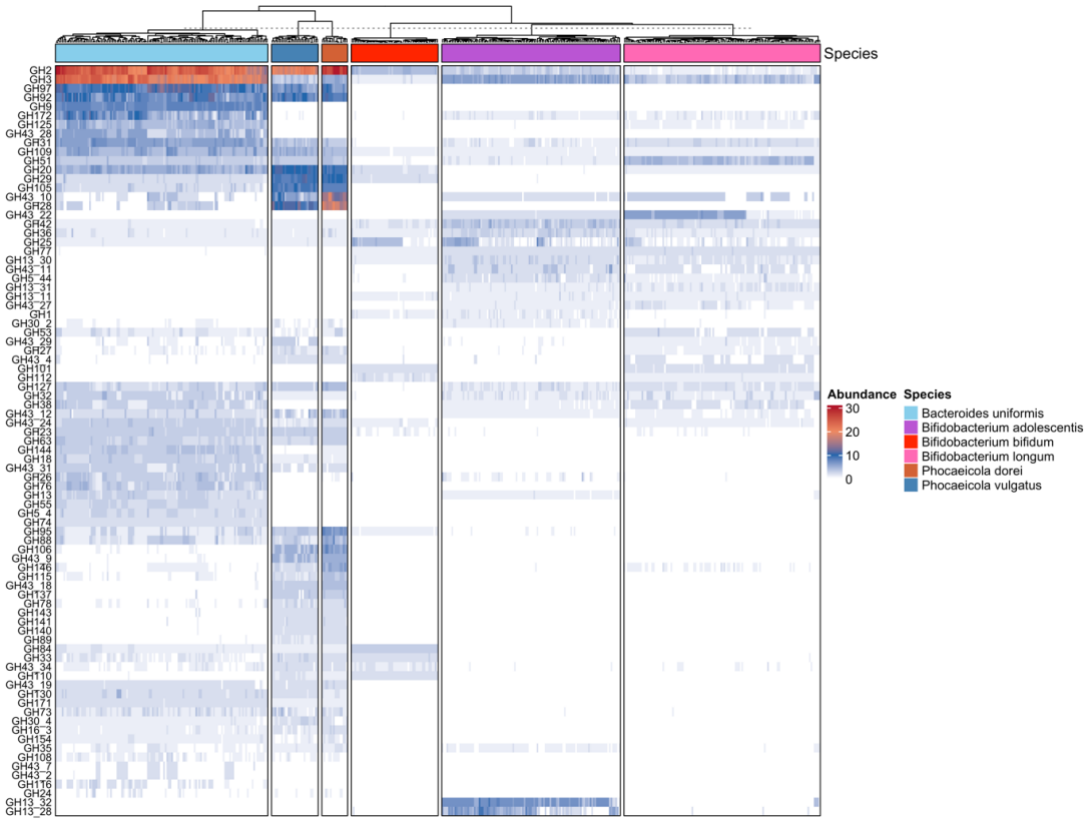
Within the *Bifidobacterium* genus, a higher number of *B. bifidum* (6/8) HMO GHs were predicted to have signal peptides than other species.

For comparison, we analysed the five most abundant species from maternal stool samples in our dataset that were not found to be shared with the infants (Supplementary Fig 3). We found that these species possess fewer HMO-associated GHs compared to the strains described above (Fig 3.4b). Of these, *A. putredinis* was observed to have the highest number of HMO-associated GHs (7/11), followed by *A. rectalis* (4/11) and *F. prausnitzii* (1/11). Most of the HMO-GHs within *A. putredinis* GHs (6/7) contained signal peptides but only one HMO-GH associated with *A. rectalis* (1/4) had a signal peptide and none from *F. prausnitzii* (0/1). Furthermore, Wilcoxon rank-sum test revealed a significant difference in the number of HMO-associated GHs between shared and non-shared species (*W = 14, p = 0.027*).



Supplementary Figure 3.3 | Taxonomic composition of maternal gut microbiome

Heatmap representing taxonomic composition of top 50 species of mother stool samples at various timepoints (E – Early (16 weeks), L- late (34 weeks), 1M- 1 month and 3M – 3 months).



Supplementary Figure 3.5 | Glycosyl hydrolase profiles in MAGs

Heatmap showing the glycosyl hydrolases abundance of frequently shared species MAGs recovered from mother and infant metagenomes.

3.4 Discussion

Bacterial strain transmission from mother to infant is an important phenomenon that shapes the early life infant gut microbiome and potentially influences long term health outcomes[4,34,35]. Understanding the microbial genetic factors involved can help explain the underlying mechanisms and inform targeted intervention strategies. Our study characterises the microbial functions associated with strain sharing, particularly the role of HMO utilisation genes. For breastmilk fed infants, the gut environment is presumed to be rich in HMOs and the ability of species to metabolise HMOs confers a strong competitive advantage. Species belonging to the genera *Bifidobacterium*, *Bacteroides,* and *Phocaeicola* were shown to be the most abundant in the early life gut microbiome. While most of the mother-infant sharing events involved *Bacteroides* and *Phocaeicola* species, their relative abundance was low in infant samples at 1-month and 3-month timepoints compared to *Bifidobacterium* species. We also observed an increase in the number of sharing events increased at later timepoints revealing 30 additional sharing events at 3-month timepoint that had not been detected at earlier timepoints. Previous studies have shown that *Bifidobacterium* species are highly prevalent in the early infant microbiomes due to their HMO utilisation capacity[22]. Recent reports suggest that species from *Phocaeicola* and *Bacteroides* can also metabolise HMOs, although the associated genetic factors require further elucidation[29,30]. Yassour *et al.* found that *Bacteroides* show a generalised transcriptional pattern for HMO utilisation rather than specific genes being upregulated[29], and utilise a more diverse set of genes for HMO utilisation than previously thought[20,30]. The metagenomic analysis presented here reveals a more diverse array of HMO-related glycosyl hydrolases enriched in *Bacteroides* and *Phocaeicola* dominated samples compared to *Bifidobacterium* species, and a higher number of strain sharing events in *Bacteroides* and *Phocaeicola* species compared to *Bifidobacterium*. It is possible that this wide array of HMO-related glycosyl hydrolases along with diverse GHs in *Bacteroides* and *Phocaeicola* is advantageous for these species during transfer. Additionally, the presence of signal peptides in GHs indicates the extracellular secretion of enzymes[36–38] and we found a higher number of signal peptides in HMO-related glycosyl hydrolases in *Bacteroides* and *Phocaeicola* species

indicating their potential for extracellular GH synthesis. This ability may provide a competitive advantage in the infant gut environment and might also be beneficial for other gut commensals that may rely on carbohydrate-dependent cross-feeding[31,39]. Notably, we also found that species with a high number of signal peptide-containing GHs with predicted HMO-degrading function were frequently shared between mother and infants. For example, *B. uniformis* and *P. vulgatus* are two species that are frequently transferred, and we also noted the same species encode the highest number HMO-associated GHs with a signal peptide. Within the *Bifidobacterium* genus, *B*. *bifidum* is a frequently shared species with a much higher number of signal peptide containing HMO-specific GHs compared to *B. longum* subsp*. longum*. In addition to these species, *A. putredinis,* an abundant maternal microbe shared in mother-infant pairs also encodes a relatively high number of HMO-associated GHs compared to other abundant species in the maternal microbiome, suggesting that these species maybe less adapted to the HMO rich infant gut environment. Taken together, our findings suggest that species encoding diverse secreted HMO-specific GHs are well suited to vertical transmission and colonisation in mother-infant pairs.

Our study reveals that *Bacteroides* and *Phocaeicola* species are the most frequently shared in mother-infant pairs and also possess a diverse array of HMO-related GHs, potentially explaining their frequent transmission from mother to infant. These species, along with *Bifidobacterium*, demonstrate a higher abundance of HMO utilisation genes compared to other maternal gut microbes. The differential distribution of GH families and signal peptides among these species points to diverse strategies for HMO utilisation, potentially allowing for niche partitioning in the infant gut ecosystem. Further research is needed to fully elucidate the complex interplay between HMO metabolism, strain sharing, and infant gut colonisation, including longitudinal studies to track the persistence of these shared strains and their impact on long-term health outcomes.

3.5 Methods

3.5.1 Preprocessing and taxonomic analysis

Raw read quality was assessed using fastqc (v 0.12.1) and adapter content and host contamination removal was performed using trimgalore (v 0.6.1) and bowtie2 (v 2.3.4), respectively[40–42]. The quality filtered reads were then used for downstream analysis. Taxonomy assignment was performed using metaphlan (v 4.0.6) with October 2022 database (mpa_vOct22_CHOCOPhlAnSGB_202212)[43]. The alpha and beta diversity metrics were calculated using calculate_diversity Rscript from metaphlan utility scripts. Significant differences between sample types were calculated using Wilcoxon's rank sum test with a significance threshold of 0.05. Data visualization was performed using ggplot2 (v 3.5.1) and Complexheatmap (v 2.20) packages in R (v 4.4.1)[44–47].

3.5.2 Reconstruction of metagenome assembled genomes (MAGs) and strain tracking

Using quality filtered reads, contigs were assembled with MetaWRAP assembly module with metaspades option[48]. Subsequently, MAGs were reconstructed using the binning module with three binning algorithms – MetaBAT2, MaxBin2 and Concoct and the resulting bins with >50% completeness and <10% contamination were refined and reassembled using MetaWRAP refinement module[49–51]. MAG quality was assessed with CheckM (v 1.0.18)[52]. High-quality MAGs were considered for downstream analysis based on the standards recommended by Bowers *et al.* with >90% completeness & <5% contamination. GTDB-TK (v 1.5.0) was used to assign taxonomy to each MAG with reference database (v r202)[53]. Strain sharing events were identified using InStrain (v 1.8.0) with quality filtered metagenomic reads as input and a reference database constructed from the high-quality MAGs dereplicated at 98% identity using dRep (v 3.2.0)[54,55]. Metagenomic reads were mapped against the database using Bowtie2 (v 2.4.4)[42]. The recommended thresholds of 99.999% popANI and 50% reference genome coverage were used to identify strain sharing events in mother-infant pairs.

3.5.3 Profiling of glycosyl hydrolases in metagenomes and MAGs

Assembled contigs from each sample were used to predict genes using prokka (v 1.14)[56]. The resulting genes were used as input for dbcan (v 4.0) to identify GHs in each sample. Signalp (v 4.1) was used along with the dbcan package to identify the presence of a signal peptide in individual predicted GHs[57]. The metagenomic reads were mapped against the predicted GH genes using Burrows-Wheeler Aligner (v 0.7.17)[58]. Seqkit (v 1.4), Samtools (v 1.20) and bedtools (v 2.27) were used for file conversions to process the output files[59–61]. The abundance of GH-encoding genes in metagenomic samples was calculated using dbcan_utils package and the reads were normalized to reads per million for each sample. Samples with *Bacteroides* and *Phocaeicola* genera abundance of >30% were defined as *Bacteroides*/*Phocaeicola* rich samples. The significant GH-encoding genes were identified using analysis of variance model with reads per million and genus abundance from stats package (v 4.4.1) in R. Visualization was performed using ggplot2 (v 3.5.1) and Complexheatmap (v 2.20) packages in R.

High-quality MAGs were used to profile the GHs at strain-level using run_dbcan from dbcan (v 4.0) with Signalp (v 4.1) for signal peptide identification[62]. Dbcan standard database (v 4.0) was used to predict the function of GHs in MAGs. HMO-associated GHs were considered present in a species if half of the MAGs contained those GHs.

3.6 References

1. de Muinck, E. J. & Trosvik, P. Individuality and convergence of the infant gut microbiota during the first year of life. *Nat Commun* 9, 2233 (2018).

2. Coker, M. O. & Madan, J. C. Chapter 3 - The microbiome and immune system development. in *The Developing Microbiome* (ed. Claud, E. C.) 43–66 (Academic Press, 2020). doi:10.1016/B978-0-12-820602-7.00003-9.

3. Sarkar, A., Yoo, J. Y., Valeria Ozorio Dutra, S., Morgan, K. H. & Groer, M. The Association between Early-Life Gut Microbiota and Long-Term Health and Diseases. *J Clin Med* 10, 459 (2021).

4. Enav, H., Bäckhed, F. & Ley, R. E. The developing infant gut microbiome: A strain-level view. *Cell Host and Microbe* 30, 627–638 (2022).

5. Zhang, H., Zhang, Z., Liao, Y., Zhang, W. & Tang, D. The Complex Link and Disease Between the Gut Microbiome and the Immune System in Infants. *Front. Cell. Infect. Microbiol.* 12, (2022).

6. Del Chierico, F. *et al.* Phylogenetic and Metabolic Tracking of Gut Microbiota during Perinatal Development. *PLoS One* 10, e0137347 (2015).

7. Dominguez-Bello, M. G., Godoy-Vitorino, F., Knight, R. & Blaser, M. J. Role of the microbiome in human development. *Gut* 68, 1108–1114 (2019).

8. Stokholm, J. *et al.* Maturation of the gut microbiome and risk of asthma in childhood. *Nat Commun* 9, 141 (2018).

9. Stiemsma, L. T. & Turvey, S. E. Asthma and the microbiome: defining the critical window in early life. *Allergy, Asthma, and Clinical Immunology : Official Journal of the Canadian Society of Allergy and Clinical Immunology* 13, (2017).

10. Jian, C. *et al.* Early-life gut microbiota and its connection to metabolic health in children: Perspective on ecological drivers and need for quantitative approach. *EBioMedicine* 69, 103475 (2021).

11. Hoskinson, C. *et al.* Delayed gut microbiota maturation in the first year of life is a hallmark of pediatric allergic disease. *Nat Commun* 14, 4785 (2023).

12. Cukrowska, B., Bierła, J. B., Zakrzewska, M., Klukowski, M. & Maciorkowska, E. The Relationship between the Infant Gut Microbiota and Allergy. The Role of Bifidobacterium breve and Prebiotic Oligosaccharides in the Activation of Anti-Allergic Mechanisms in Early Life. *Nutrients* 12, 946 (2020).

13. Nayfach, S., Rodriguez-Mueller, B., Garud, N. & Pollard, K. S. An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography. *Genome Res.* 26, 1612–1625 (2016).

14. Feehily, C. *et al.* Detailed mapping of Bifidobacterium strain transmission from mother to infant via a dual culture-based and metagenomic approach. *Nat Commun* 14, 3015 (2023).

15. Wang, S. *et al.* Metagenomic analysis of mother-infant gut microbiome reveals global distinct and shared microbial signatures. *Gut Microbes* 13, 1911571.

16. Browne, H. P., Shao, Y. & Lawley, T. D. Mother–infant transmission of human microbiota. *Current Opinion in Microbiology* 69, 102173 (2022).

17. Mitchell, C. M. *et al.* Delivery Mode Affects Stability of Early Infant Gut Microbiota. *Cell Reports Medicine* 1, 100156 (2020).

18. Moore, R. L. *et al.* Ability of Bifidobacterium breve 702258 to transfer from mother to infant: the MicrobeMom randomized controlled trial. *American Journal of Obstetrics & Gynecology MFM* 5, 100994 (2023).

19. Kiely, L. J., Busca, K., Lane, J. A., van Sinderen, D. & Hickey, R. M. Molecular strategies for the utilisation of human milk oligosaccharides by infant gut-associated bacteria. *FEMS Microbiology Reviews* 47, fuad056 (2023).

20. Marcobal, A. *et al.* Bacteroides in the Infant Gut Consume Milk Oligosaccharides via Mucus-Utilization Pathways. *Cell Host & Microbe* 10, 507–514 (2011).

21. Salli, K. *et al.* Selective Utilization of the Human Milk Oligosaccharides 2'-Fucosyllactose, 3-Fucosyllactose, and Difucosyllactose by Various Probiotic and Pathogenic Bacteria. *J. Agric. Food Chem.* 69, 170–182 (2021).

22. Masi, A. C. & Stewart, C. J. Untangling human milk oligosaccharides and infant gut microbiome. *iScience* 25, 103542 (2022).

23. Ruiz-Moyano, S. *et al.* Variation in Consumption of Human Milk Oligosaccharides by Infant Gut-Associated Strains of Bifidobacterium breve. *Appl Environ Microbiol* 79, 6040–6049 (2013).

24. Walsh, C., Lane, J. A., van Sinderen, D. & Hickey, R. M. Human milk oligosaccharide-sharing by a consortium of infant derived Bifidobacterium species. *Sci Rep* 12, 4143 (2022).

25. Bunesova, V., Lacroix, C. & Schwab, C. Fucosyllactose and L-fucose utilization of infant Bifidobacterium longum and Bifidobacterium kashiwanohense. *BMC Microbiology* 16, 248 (2016).

26. Sakanaka, M. *et al.* Varied Pathways of Infant Gut-Associated Bifidobacterium to Assimilate Human Milk Oligosaccharides: Prevalence of the Gene Set and Its Correlation with Bifidobacteria-Rich Microbiota Formation. *Nutrients* 12, 71 (2020).

27. Kim, J.-H. *et al.* Proteomic Analysis of Bifidobacterium longum subsp. infantis Reveals the Metabolic Insight on Consumption of Prebiotics and Host Glycans. *PLOS ONE* 8, e57535 (2013).

28. Ojima, M. N. *et al.* Priority effects shape the structure of infant-type Bifidobacterium communities on human milk oligosaccharides. *The ISME Journal* 16, 2265–2279 (2022).

29. Kijner, S., Cher, A. & Yassour, M. The Infant Gut Commensal Bacteroides dorei Presents a Generalized Transcriptional Response to Various Human Milk Oligosaccharides. *Front. Cell. Infect. Microbiol.* 12, (2022).

30. Kijner, S., Ennis, D., Shmorak, S., Florentin, A. & Yassour, M. CRISPR-Cas-based identification of a sialylated human milk oligosaccharides utilization cluster in the infant gut commensal Bacteroides dorei. *Nat Commun* 15, 105 (2024).

31. Díaz, R. & Garrido, D. Screening competition and cross-feeding interactions during utilization of human milk oligosaccharides by gut microbes. *Microbiome Res Rep* 3, 12 (2024).

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

33. Flint, H. J., Scott, K. P., Duncan, S. H., Louis, P. & Forano, E. Microbial degradation of complex carbohydrates in the gut. *Gut Microbes* 3, 289–306 (2012).

34. Manara, S. *et al.* Maternal and food microbial sources shape the infant microbiome of a rural Ethiopian population. *Current Biology* (2023) doi:10.1016/j.cub.2023.04.011.

35. Bäckhed, F. *et al.* Dynamics and Stabilization of the Human Gut Microbiome during the First Year of Life. *Cell Host & Microbe* 17, 690–703 (2015).

36. Zhou, F., Chen, H. & Xu, Y. GASdb: a large-scale and comparative exploration database of glycosyl hydrolysis systems. *BMC Microbiology* 10, 69 (2010).

37. Karav, S. *et al.* Oligosaccharides Released from Milk Glycoproteins Are Selective Growth Substrates for Infant-Associated Bifidobacteria. *Applied and Environmental Microbiology* 82, 3622–3630 (2016).

38. Turroni, F. *et al.* Genome analysis of Bifidobacterium bifidum PRL2010 reveals metabolic pathways for host-derived glycan foraging. *Proceedings of the National Academy of Sciences* 107, 19514–19519 (2010).

39. Cho, S. *et al.* Interactions between Bifidobacterium and Bacteroides and human milk oligosaccharides and their associations with infant cognition. *Front. Nutr.* 10, (2023).

40. Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

41. Krueger, F. *et al.* FelixKrueger/TrimGalore: v0.6.10 - add default decompression path. (2023) doi:10.5281/ZENODO.7598955.

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

43. Blanco-Míguez, A. *et al.* Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. *Nat Biotechnol* 41, 1633–1644 (2023).

44. Wickham, H. *Ggplot2*. (Springer International Publishing, Cham, 2016). doi:10.1007/978-3-319-24277-4.

45. Ahlmann-Eltze, C. & Patil, I. ggsignif: R Package for Displaying Significance Brackets for 'ggplot2'. (2024) doi:10.31234/osf.io/7awm6.

46. Gu, Z. Complex heatmap visualization. *iMeta* 1, e43 (2022).

47. R: The R Project for Statistical Computing. https://www.r-project.org/.

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

49. Kang, D. D. *et al.* MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* 7, e7359 (2019).

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

51. Alneberg, J. *et al.* Binning metagenomic contigs by coverage and composition. *Nat Methods* 11, 1144–1146 (2014).

52. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* (2015) doi:10.1101/gr.186072.114.

53. Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* 36, 1925–1927 (2019).

54. Olm, M. R. *et al.* inStrain profiles population microdiversity from metagenomic data and sensitively detects shared microbial strains. *Nat Biotechnol* 39, 727–736 (2021).

55. Olm, M. R., Brown, C. T., Brooks, B. & Banfield, J. F. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J* 11, 2864–2868 (2017).

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

57. dbCAN: a web resource for automated carbohydrate-active enzyme annotation | Nucleic Acids Research | Oxford Academic. https://academic.oup.com/nar/article/40/W1/W445/1076481.

58. Fast and accurate short read alignment with Burrows–Wheeler transform | Bioinformatics | Oxford Academic. https://academic.oup.com/bioinformatics/article/25/14/1754/225615.

59. Shen, W., Le, S., Li, Y. & Hu, F. SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. *PLoS ONE* 11, e0163962 (2016).

60. Sequence Alignment/Map format and SAMtools | Bioinformatics | Oxford Academic. https://academic.oup.com/bioinformatics/article/25/16/2078/204688.

61. BEDTools: a flexible suite of utilities for comparing genomic features | Bioinformatics | Oxford Academic. https://academic.oup.com/bioinformatics/article/26/6/841/244688.

62. Almagro Armenteros, J. J. *et al.* SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nat Biotechnol* 37, 420–423 (2019).

4 Chapter 4

Profiling of vaginal *Lactobacillus jensenii* isolated from preterm and full-term pregnancies reveals strain-specific factors relating to host interaction.

This chapter has been published in Microbial genomics journal.

(doi: 10.1099/mgen.0.001137)

4.1 Abstract

Each year, 15 million infants are born preterm (<37 weeks gestation), representing the leading cause of mortality for children under the age of five. Whilst there is no single cause, factors such as maternal genetics, environmental interactions, and the vaginal microbiome have been associated with an increased risk of preterm birth. Previous studies show that a vaginal microbiota dominated by *Lactobacillus* is, in contrast to communities containing a mixture of genera, associated with full-term birth. However, this binary principle does not fully consider more nuanced interactions between bacterial strains and the host. Here, through a combination of analyses involving genome-sequenced isolates and strain-resolved metagenomics, we identify that *L. jensenii* strains from preterm pregnancies are phylogenetically distinct from strains from full-term pregnancies. Detailed analysis reveals several genetic signatures that distinguish preterm birth strains, including genes predicted to be involved in cell wall synthesis, and lactate and acetate metabolism. Notably, we identify a distinct gene cluster involved in cell surface protein synthesis in our preterm strains, and profiling the prevalence of this gene cluster in publicly available genomes revealed it to be predominantly present in the preterm-associated clade. This study contributes to the ongoing search for molecular biomarkers linked to preterm birth and opens up new avenues for exploring strain-level variations and mechanisms that may contribute to preterm birth.

4.2 Data Summary

Sequence data used in this study can be accessed from European Nucleotide Archive (ENA) under the accession number PRJEB34536. The genomes and metagenome assembled genomes generated in this study can be accessed from ENA under this accession number PRJEB63084.

4.3 Introduction

The vaginal microbiome in humans is distinct, relatively less diverse than other body sites like gastrointestinal tract[1] and often dominated by *Lactobacillus* species[2,3]. By producing lactic acid, *Lactobacillus* spp. reduce vaginal pH and provide protection against growth of pathogens[4–6]. Lactobacilli can also produce antimicrobials such as bacteriocins that benefit the host epithelium by maintaining membrane integrity[7,8]. The species commonly found in the vaginal microbiota of healthy women are *Lactobacillus crispatus*, *Lactobacillus gasseri*, *Lactobacillus iners* and *Lactobacillus jensenii*; most of which produce large amounts of lactic acid[9–11] and are commonly categorised into community state types (CSTs) [2] CST-I, CST-II, CST-III and CST-V, respectively. The remaining CST, CST-IV, represents a community with reduced *Lactobacillus* and an increased collection of obligate and facultative anaerobes, which include *Atopobium*, *Gardnerella*, *Prevotella* and *Megasphaera*[12], and has been linked to bacterial vaginosis (BV)[13]. Interestingly, CST-IV is also observed in a high percentage of the healthy population with asymptomatic BV. This suggests that the relationship between CST and BV is complex and likely involves other factors such as hygiene, ethnicity, and smoking[14]. Various epidemiological studies have associated mixed-species communities with increased risk of poor health outcomes such as sexually transmitted infections[14,15]. However, the mechanistic understanding of these correlations has yet to be explored. Furthermore, various studies have shown an increased risk of preterm birth (PTB) in women with BV[16,17].

Approximately 15 million infants are born preterm (<37 weeks of gestation) annually, which is a risk factor for infant mortality and morbidity[18]. Several factors, such as maternal-fetal genetics, environmental interactions, and the vaginal microbiome, are important contributors to PTB[19]. Furthermore, bacterial infections such as urinary tract infections and periodontal disease have been associated with an increased risk of PTB[20,21]. Previous studies have reported that the taxonomic composition of the vaginal microbiome has a notable influence on PTB risk[22–25]. Some previous studies have associated *Lactobacillus crispatus* with a lower risk of PTB[26]. However, a variety of different taxa have been associated with PTB in some studies[26,27], whereas another study found no association between the vaginal microbiome and PTB[28].

A recent review by Gudnadottir et al.[29] suggested that analysing the vaginal microbiome could aid in predicting PTB. They linked *L. crispatus* dominance to a lower risk of PTB and *L. jensenii* abundance to a higher risk of PTB. Indeed, a study of 133 women, predominantly of the Caucasian population, found a significant association between both *L. jensenii* and decreased lactate levels and PTB[30]. Recent reports also highlight that women with *L. jensenii* dominance over *L. crispatus* had lower lactate and glutamate levels in vaginal fluid and experienced more PTB[31,32]. It is important to note that most studies were observational and based on 16S rRNA sequencing, which limits the taxonomic resolution to the species level, at best, and further limits the potential to understand the functional potential of the community and investigate microbial-host interactions. Recent studies using shotgun metagenomics have begun to identify potential functional signatures associated with PTB, highlighting the power of this approach[26,33]. However, in-depth mechanistic insights relating to strains and their functional differences are needed to identify potential molecular biomarkers that could aid in the early detection of PTB and perhaps reveal genes involved in important host-immune interactions.

In this study, we analyse data from shotgun metagenomics and genome-sequenced isolates to reveal strain-level variation among *Lactobacillus jensenii* from vaginal swabs of pregnant women associated with preterm and full-term birth, respectively.

4.4 Results

4.4.1 Isolation and sequencing of *Lactobacillus jensenii*

Previously we have shown a functional difference in the vaginal microbiota composition of PTB women[26]. Here we used 89 vaginal metagenomic samples from 57 participants of the same cohort to reconstruct 11 high-quality (>90% completeness and <5% contamination) *L. jensenii* metagenome assembled genomes (MAGs). In addition, we isolated, and genome sequenced 9 strains of *L. jensenii* from the same cohort, of which 4 were not paired to any MAG assembly. In total, 15 high-quality genomes, 9 from cultured isolates and 6 MAGs were used for analysis, with an isolate being considered over a MAG when both were recovered. Six of these genomes were from preterm pregnancies and the remaining nine were from full-term pregnancies. Of the six preterm samples, two were late preterm (36.4 and 36.7 weeks), and four were very preterm (28-32 weeks). The mean birthweight for preterm and full-term babies was 2,439 g and 4,061 g, respectively. Four women had a history of previous large loop excision of the transformation zone (LLETZ) surgery with two each in preterm and full-term group.

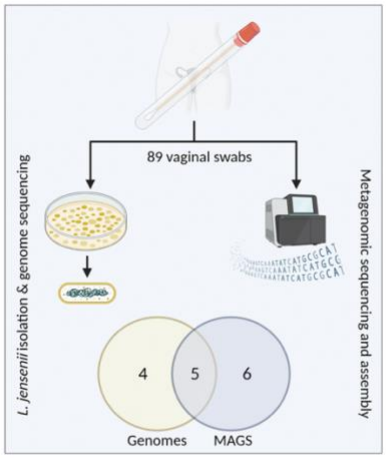
4.4.2 Pangenome and phylogenomic analysis reveals a distinct grouping of preterm and full-term *L. jensenii* genomes.

To facilitate an investigation of the genomic variation between preterm and full-term-associated *L. jensenii* strains, we performed a pangenomic reconstruction of the 15 *L. jensenii* genomes (Fig. 4.1b). A total of 24,928 genes were clustered into 2,669 unique gene families (Fig. 4.1c). Of the 2,669 gene families, 1,163 were universally present (≥ 95% strains) and classified as the core genome. The number of gene clusters in the shell and cloud genomes were 566 and 940, respectively. The pangenome analysis of *L. jensenii* revealed an open pangenome (Supplementary. Fig. 4.1).

Core genome phylogeny, as determined by SNPs in these core genes, revealed that strains from Preterm group were distinct from full-term birth (FTB) (Fig. 4.2). However, three strains from the FTB group (LJ2, LJ3, LJ12) clustered within the PTB

clade. Further investigation revealed that these strains were from women with previous LLETZ surgery and PTB history. Altogether, within the PTB clade, 6/9 strains were directly associated with PTB while the remaining three FTB-linked isolates were associated with a history of PTB or LLETZ surgery. Next, we performed an average nucleotide identity (ANI) comparison between the strains to further understand the genomic variation between these two clades. We found ANI values ranging from 99.30 to 99.91 suggesting a high level of genetic similarity with subtle strain-level variations (Supplementary Fig. 4.2).

(a)



(c)

Pangenome category	*Threshold*	*# of genes*
Core genes	(95% <= strains <= 100%)	1163
Shell genes	(15% <= strains < 95%)	566
Cloud genes	(0% <= strains < 15%)	940

(b)

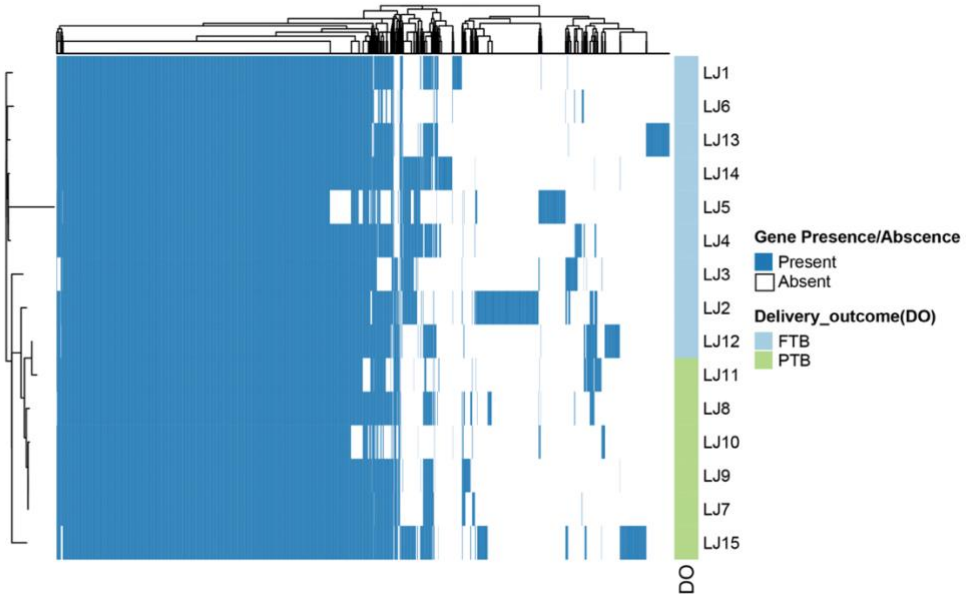
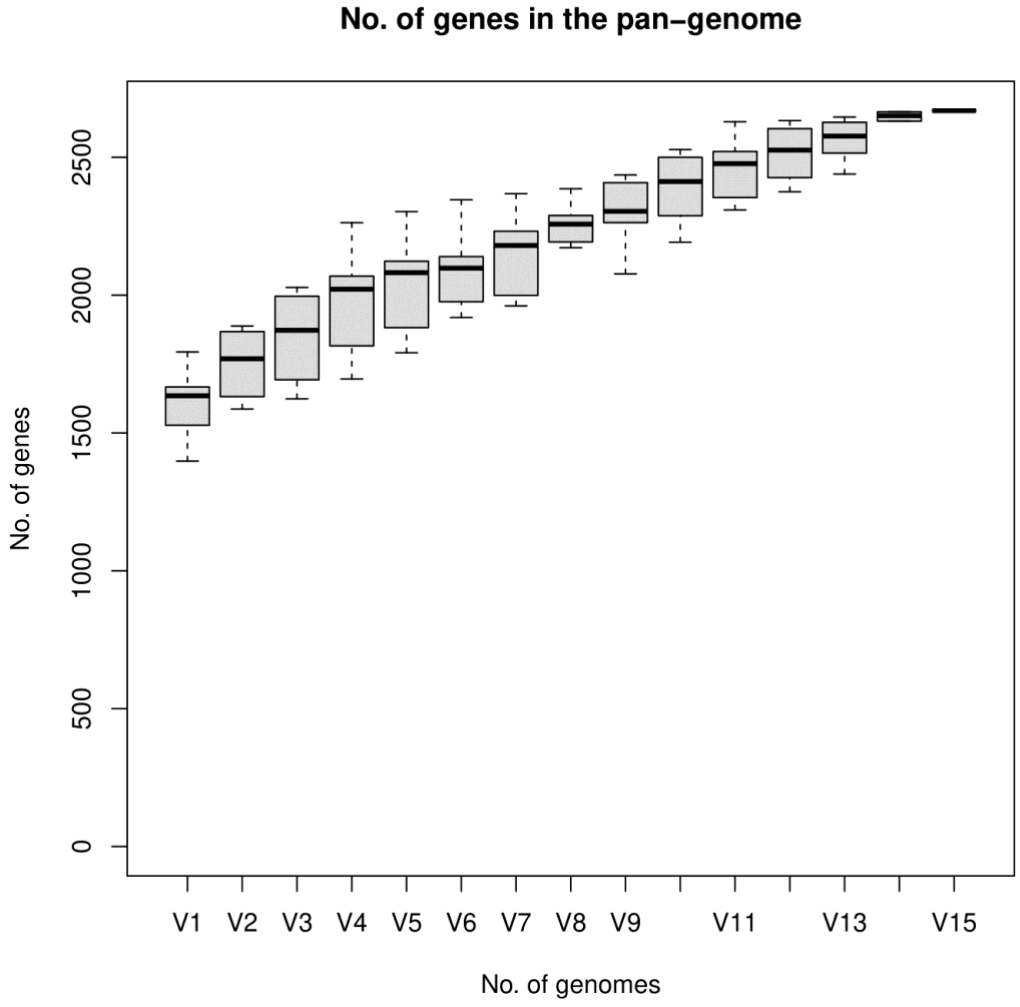


Figure 4.1. (a) Overview of *L. jensenii* sequenced from vaginal metagenomic samples in this study. The Venn diagram represents the number of genomes and MAGs recovered. (b) Pangenome heatmap of *L. jensenii*. Rows are clustered based on the core genome identity, also represented by the phylogenetic tree on the left-hand

side. For each gene, a blue cell indicates its presence, and a white cell indicates its absence in the corresponding genome based on the roary 95% identity threshold. The legend shows the delivery outcome (DO) of each pregnancy associated with the genome. (c) Pangenome statistics of *L. jensenii* genomes in this study.



Supplementary. Fig. 4.1. Plot showing the number of genes (y-axis) in each of the Lactobacillus jensenii genomes (x-axis) generated in our study indicating an open pangenome with size increasing with every new genome added.

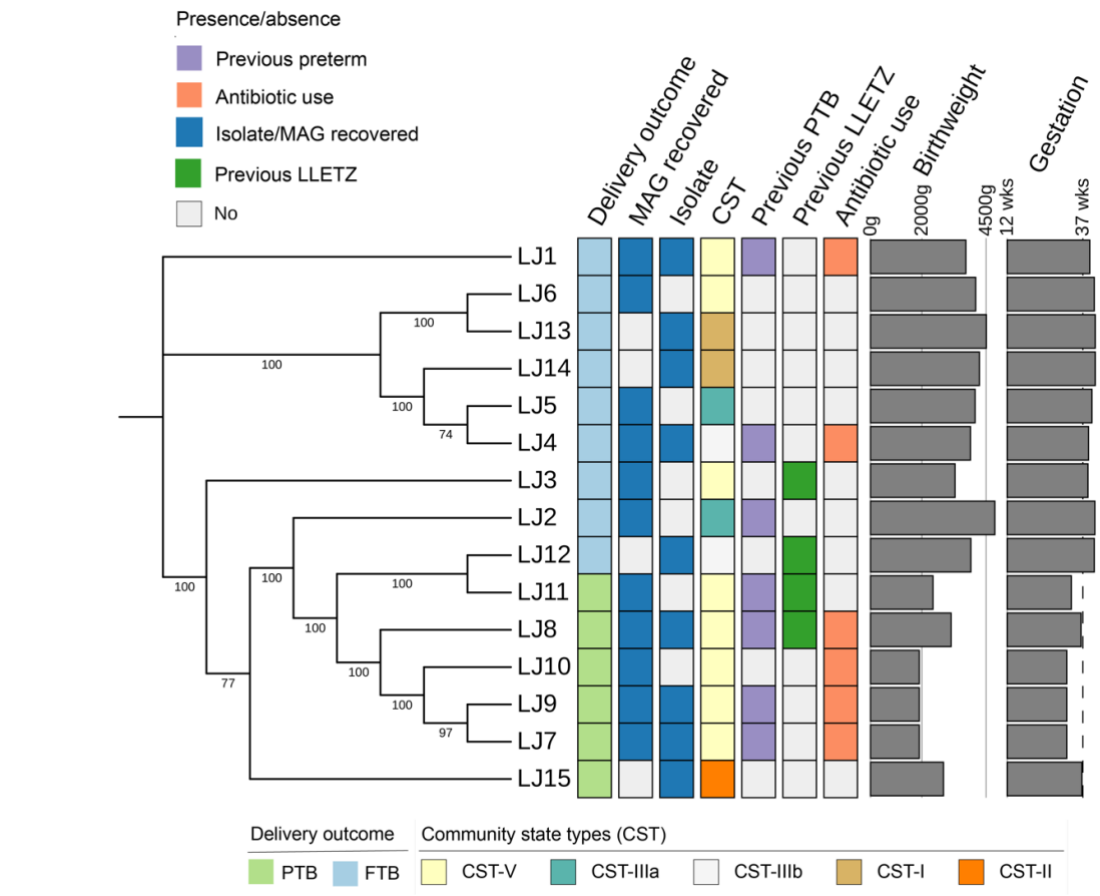
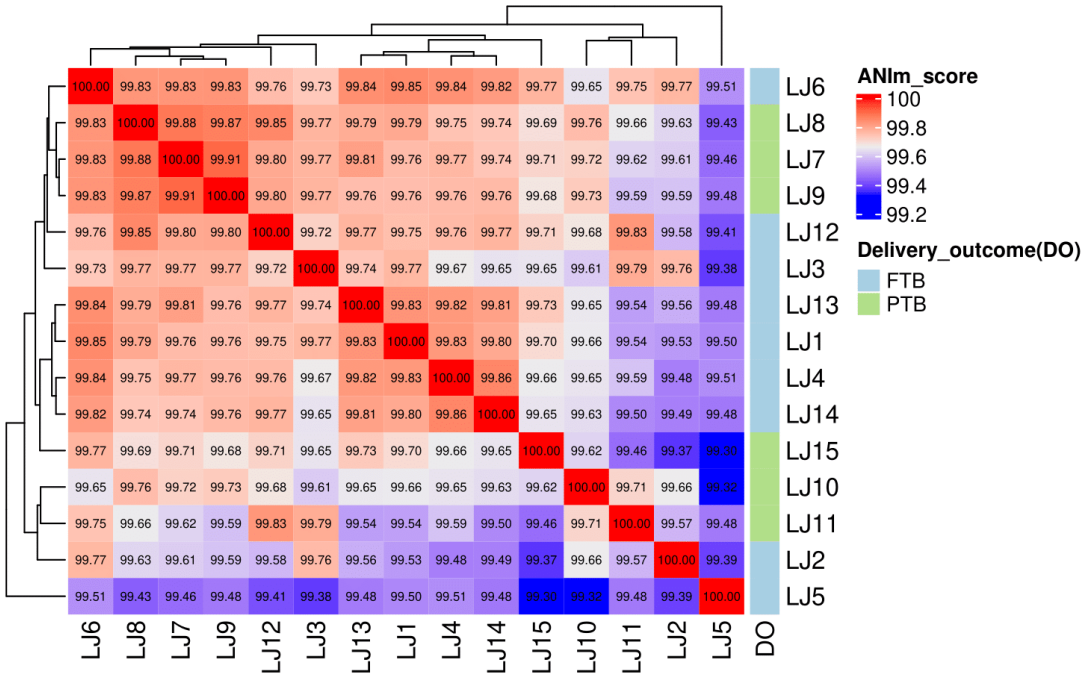


Figure 4.2. Maximum likelihood phylogenetic tree of the core genome of *L. jensenii*. Bootstrap values are represented on the branches. The heatmap on the right-hand side of the cladogram represents the delivery outcome, whether an isolate and/or MAG was recovered from this individual, community state type (CST), previous preterm birth status and LLETZ surgery, and antibiotic use in pregnancy. Bar plots show birthweight and gestation length.



Supplementary. Fig. 4.2. Average Nucleotide Identity (ANI) heat-map showing the ANI scores between 15 L. jensenii in our dataset and side bar representing the delivery outcome (DO) of each pregnancy associated with the genome.

4.4.3 Genetic variation and genes distinct to PTB associated L. jensenii

A total of 1,163 core gene families were extracted from the pangenome to identify variations in genes involved in core functions between the PTB and FTB strains. Following alignment of individual protein sequences from these gene families, 185 non-synonymous mutations specific to PTB genomes were identified in 156 core genes. A notable number of differences between PTB and FTB strains were observed in genes involved in lactate metabolism (*ldhD*), acetate metabolism (*ackA*), and cell wall-related functions (*murK* and *ycbB*) (Fig. 4.3), revealing a genetic signature distinct to PTB associated *L. jensenii*.

Using roary, we analysed the genomes of *L. jensenii* to determine if further genetic differences between strains were present. Genes distinct to preterm/full term birth were identified using the whole pangenome including cloud and shell genes. However, we found significant genes only from the core genome category. In total we observed 14 genes that were distinct to PTB strains and 4 genes distinct to FTB strains. 50% (7/14) of these had an unknown function following automated

annotation with Prokka. Supplemental annotation with Eggnogmapper and BLASTx against the NCBI-NR database identified the functions of 85 % (12/14) of these genes. The PTB specific genes were present within two gene clusters harbouring putative signal peptide, glycosyl transferases, polysaccharide synthesis genes and glycosyl hydrolases.

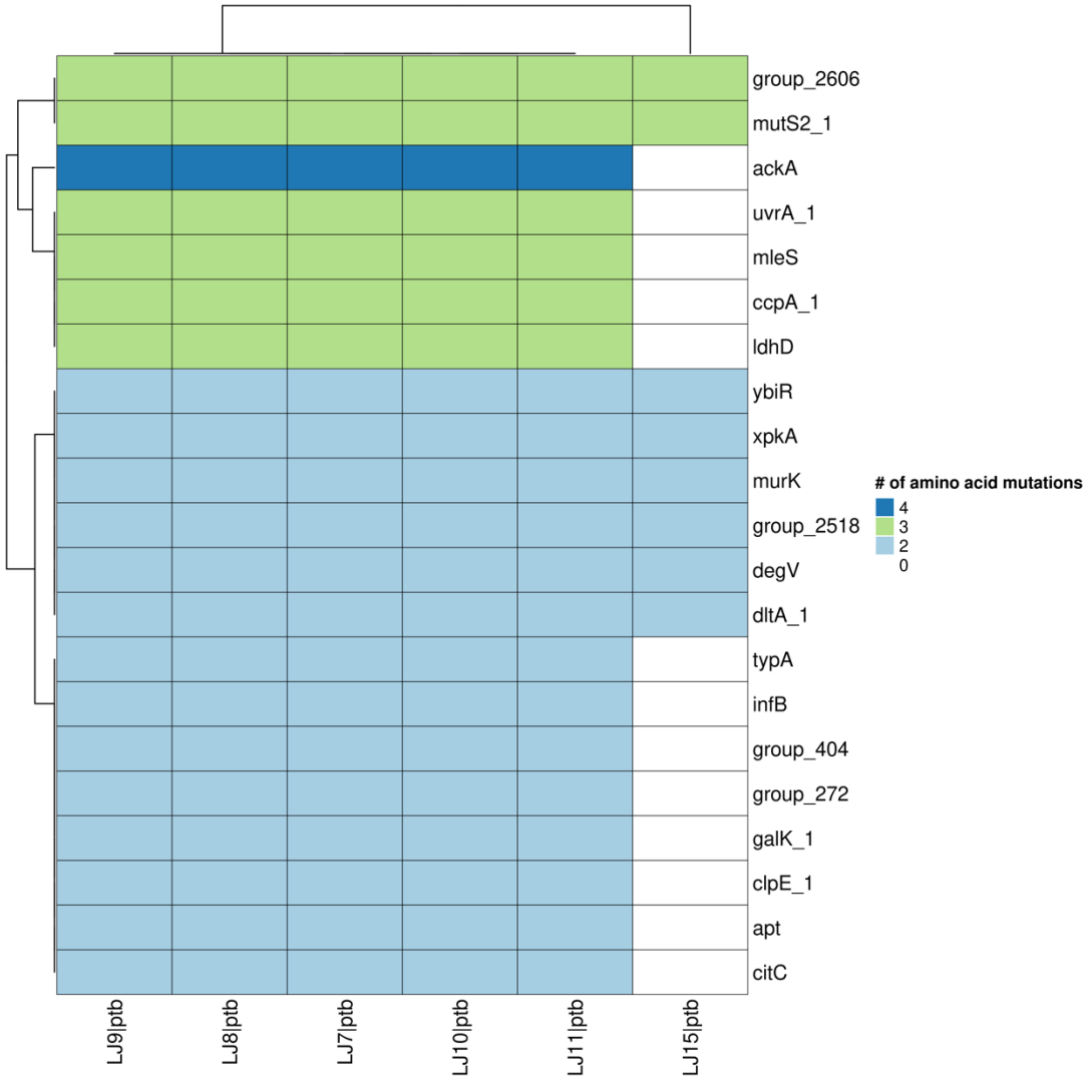


Figure 4.3. Heatmap showing distinct amino acid variations in preterm birth samples. Each colour represents the number of per-gene variations. Only genes with at least two variations were plotted.

4.4.4 Preterm birth associated *L. jensenii* possess a gene cluster with a predicted role in cell surface proteins synthesis

A gene cluster of 10 genes distinct to preterm strains that are identified from the pangenome analysis was observed in a region adjacent to a further 5 genes that were also present in full-term genomes (Fig. 4.4). The gene cluster harbours a large genomic region of ~10 kb length identified as a LPXTG-motif cell wall anchor (CW_MucBP). Within this is a putative single coding sequence for a domain of unknown function (DUF285), a mucin binding domain (MucBP) and a YSIRK signal motif. In addition, this gene cluster contains several putative glycosyl transferase genes (GT), including GT1, GT2 and GT4. Further, a flippase (MurJ), capsular polysaccharide biosynthesis protein (Cap8A), tyrosine-protein kinase (YwqD) and tyrosine-protein phosphatase (YwqE) were also predicted to be encoded within the gene cluster. This gene cluster is predominantly found in genomes from preterm samples. A strong association with the preterm trait was observed for 8 out of the 10 genes. In full-term isolates, the genes required to synthesise polysaccharides, cell surface proteins and mucosal adhesion were absent (Fig 4.4).

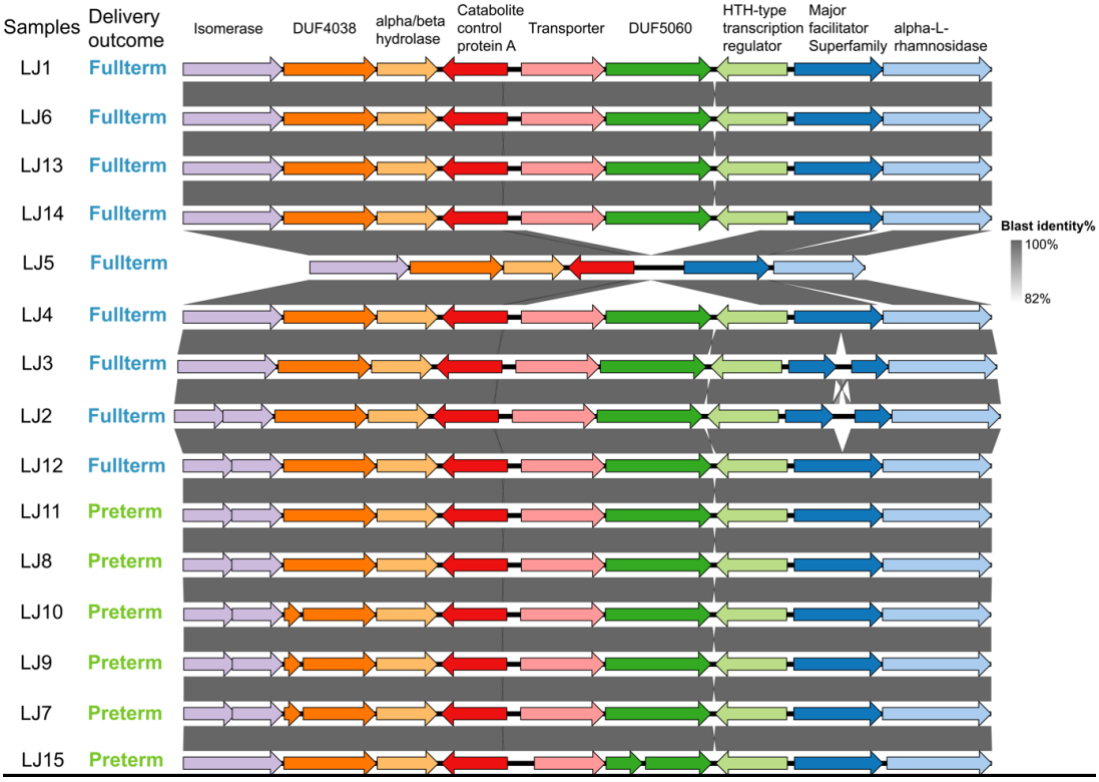


Figure. 4.4. Schematic overview of the organisation of a putative cell surface glycan gene cluster of *L. jensenii.* The length of this gene cluster is ~20kbp. The colour of each box indicates the presence (mint green) and absence (brown) of the corresponding gene annotated across the top.

4.4.5 *L. jensenii* possesses a gene cluster with putative glycan utilisation genes

We identified a gene cluster containing putative glycosyl hydrolase genes from the roary analysis within all genomes of *L. jensenii*. This cluster was approximately 11 kbp in length and comprised of genes predicted to encode an alpha-rhamnosidase, membrane transporters, an isomerase and two other genes of unknown function, designated DUF4038 and DUF5060 (Fig 4.5). Literature searching suggested that DUF4038 and DUF5060 genes are putative glycoside hydrolase and glycoside hydrolase-associated C-terminal with the closest homologs to GH13 and GH14 family, that is, alpha-amylase and beta-amylase [34]. In addition, the alpha-beta hydrolase fold with potential serine protease function and major facilitator superfamily (MFS) transporters along with two potential GHs belonging to GH1 and GH28 were also identified flanking this region. While this cluster was conserved between all genomes, the putative isomerase gene contained a frameshift mutation that was absent in 5/6 PTB genomes and 2/9 FTB genomes, resulting in a predicted change in coding sequence for these strains. Similarly, the DUF4038 gene also has large deletion in 3/6 PTB strains, but none of the FTB genomes.

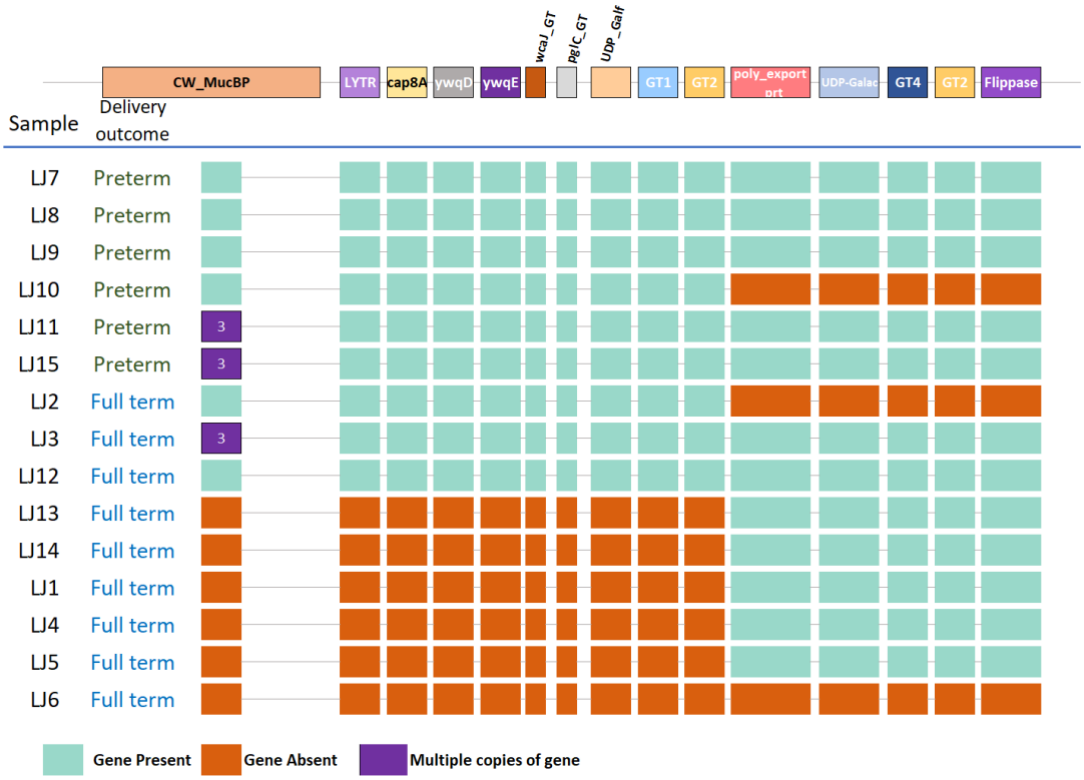


Figure 4.5. Schematic overview of the organisation and sequence relatedness of the glycan utilisation gene cluster of *L. jensenii.* The homology of each gene is represented by different shades of white (low homology) to grey (high homology) between each genome. The colour of the arrows represents each gene.

4.4.6 Profiling of preterm associated gene clusters in publicly available *L. jensenii* genomes

To identify the phylogenetic patterns and presence/absence of preterm associated genomic signature across *L. jensenii* species, we used publicly available genomes along with the genomes from our study. A very high ANI comparison value was observed for all genomes ranging from 98.98 to 100 (Supplementary Fig. 4.4). The core genome phylogeny of these genomes revealed two distinct clades in the *L. jensenii* species (Fig. 4.6). The preterm and full-term related genomes from this study mainly clustered in two different clades, i.e., Clade-I and Clade-II, respectively. Genes from the putative carbohydrate utilisation gene cluster were present in all the isolates from both clades (Supplementary. Fig. 4.3). The putative cell surface protein

encoding gene cluster was found to be predominantly present in the same clade as the preterm-associated genomes (Clade-I) (Fig. 4.7.).

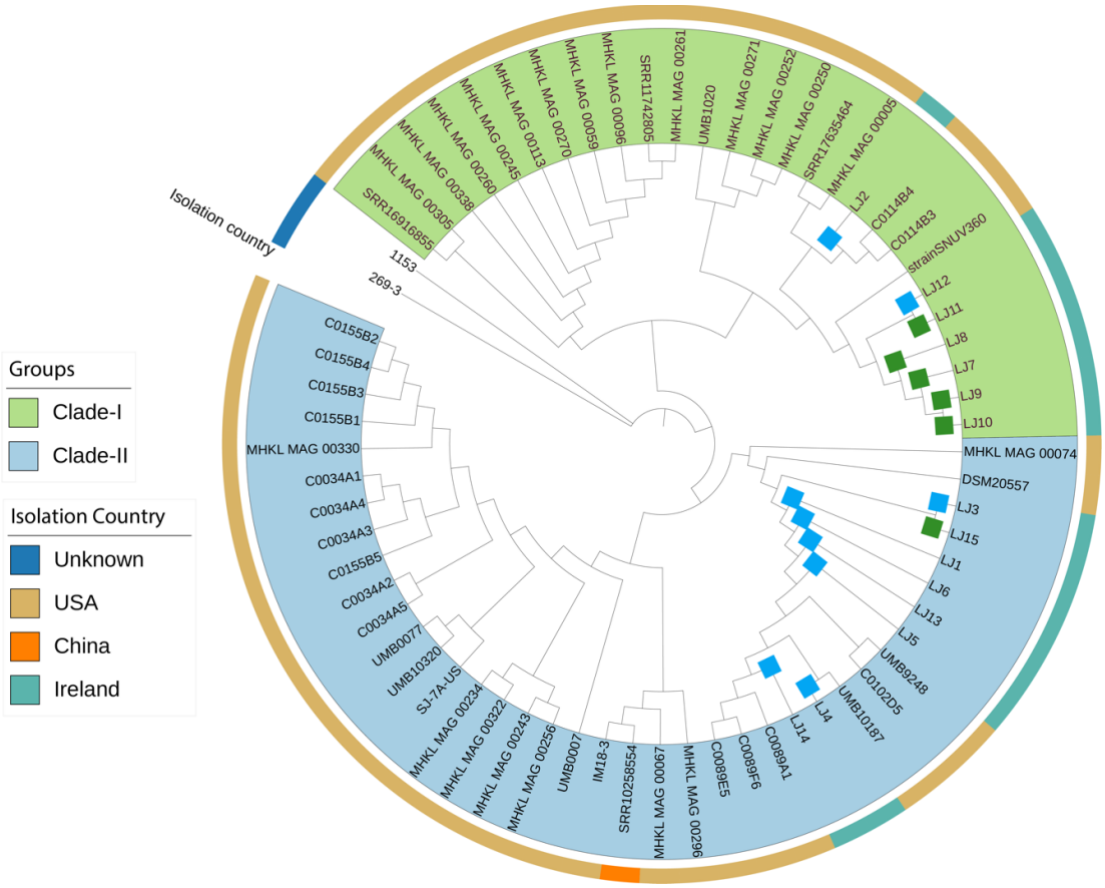


Figure. 4.6. Maximum likelihood phylogenetic tree of *L. jensenii* species reconstructed with publicly available genomes. The outer circle represents the isolation country. The green colour on labels indicates Clade-I and blue indicates Clade-II. The green squares on the branches represent preterm isolates and blue squares indicates full-term strains from our dataset.

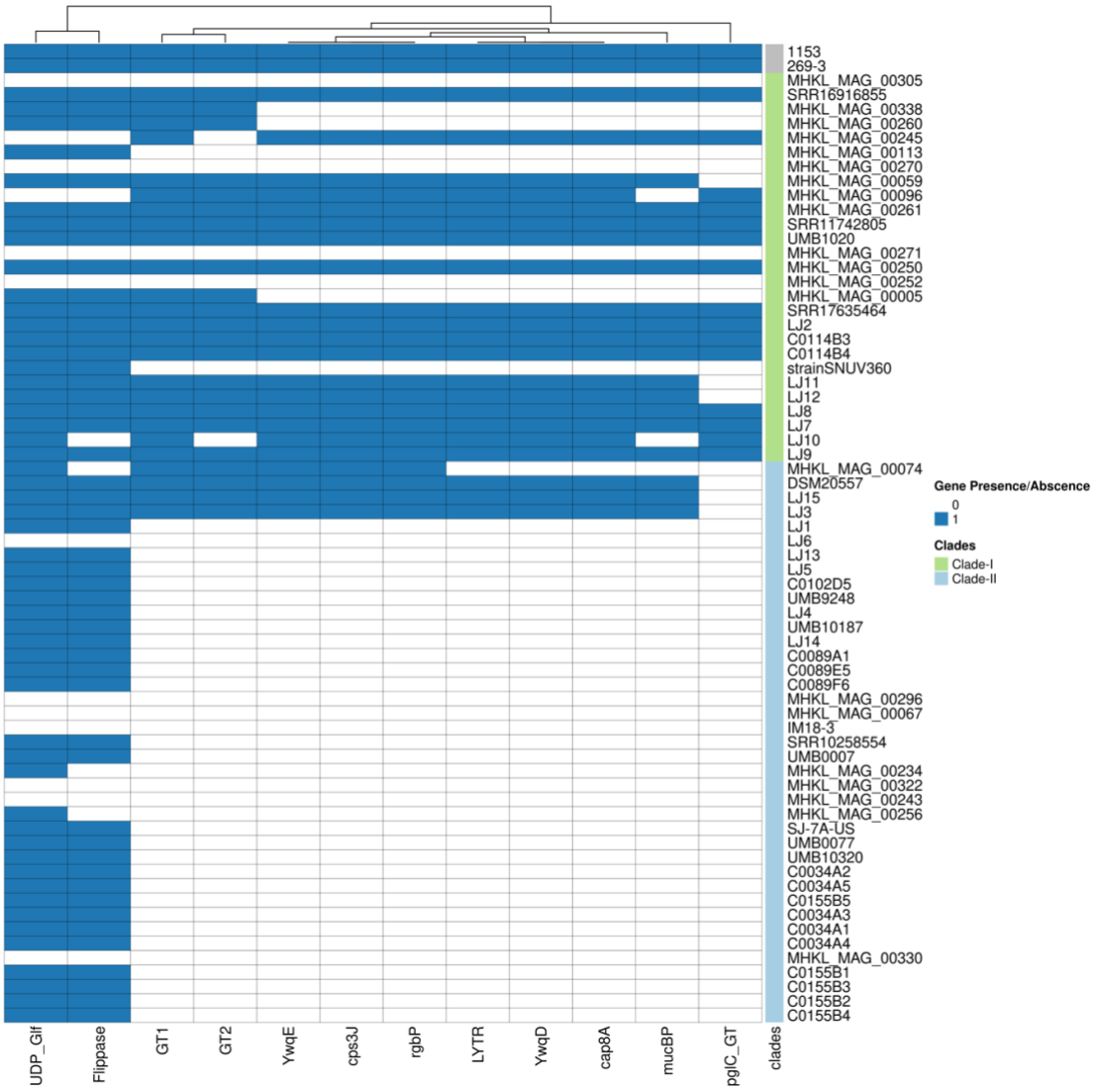
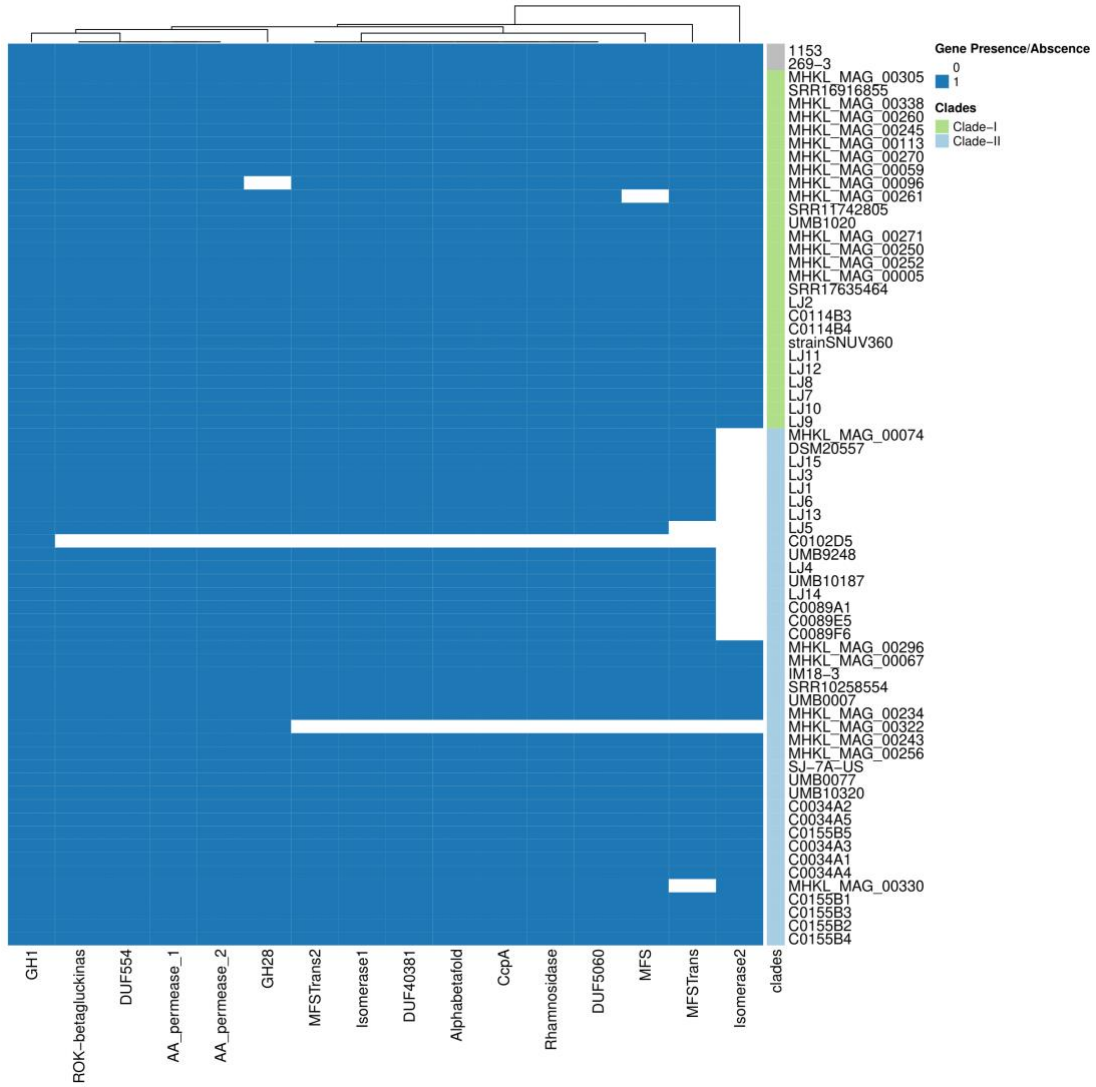
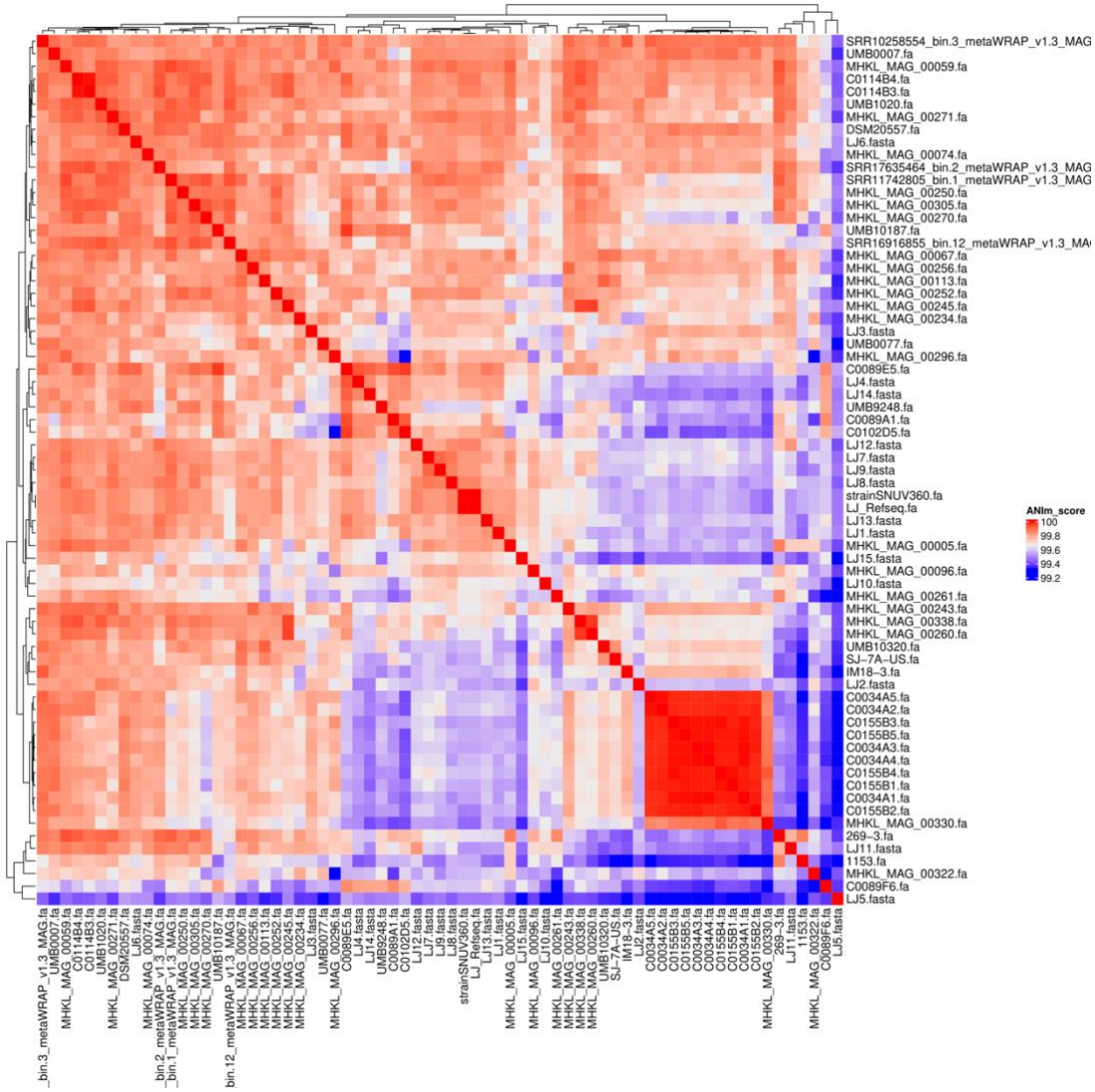


Figure 4.7. Heatmap showing the presence and absence of genes from the putative cell surface protein encoding cluster identified in preterm strains. The x-axis represents the genes from the gene cluster. For each gene, a blue cell indicates its presence, and a white cell indicates its absence in the corresponding genome. The bar on the right indicates the clades respective to the core genome phylogeny.



Supplementary. Fig. 4.3. Heatmap showing the presence and absence of glycan utilisation cluster genes in the publically available genomes. The bar on the right represents the clades respective to core genome phylogeny. The green represents the preterm associated clade, blue represents the full-term clade and grey indicates unknown.



Supplementary. Fig. 4.4. Average Nucleotide Identity (ANI) heat-map showing the ANI scores between 15 L. jensenii in our dataset along with publically available genomes.

4.5 Discussion

Fifteen high-quality *L. jensenii* genomes were analysed from vaginal samples of 6 preterm and 9 full-term pregnancies using a combination of metagenomic sequencing and targeted isolation. In-depth genomic analysis revealed a distinct phylogenetic clade of preterm genomes and identified differences in the core genome involved in cell wall formation and the metabolism of lactate and acetate. Further pangenome analysis identified a gene cluster predicted to be involved in the degradation of glycans typical to vaginal tract including glycogen. Another gene cluster specific to PTB strains was identified with a potential role in the synthesis of cell surface proteins which are crucial for microbial-host interactions. In addition, we profiled the presence of these gene clusters in publicly available genomes, which revealed cell surface protein synthesis genes to be predominantly present in the preterm-associated clade. These results suggest that the preterm *L. jensenii* group carrying these gene clusters may have a functional advantage through an ability to utilise various sugars and potential host-immune interactions within the vaginal niche. The data presented here also makes a significant contribution to vaginal *L. jensenii* reference genomes. Through detailed strain-level analysis, we compared full-term and preterm strains highlighting key genomic variations with a potential role in preterm complication.

The changes in the composition of the vaginal microbiota can be influenced by several factors, such as changes in host physiology, unprotected intercourse, or hormonal fluctuation[35,36]. This makes it challenging to develop biomarkers for preterm/dysbiotic conditions. Recent studies have shown vaginal metabolites and microbiome profiles could be useful in predicting preterm[37,38]. However, the mechanistic understanding of vaginal microbial genomic factors is needed to develop standard molecular based biomarkers that could aid in early PTB detection. A recent study has demonstrated that crosstalk between microbial recognition mediators and host immunity could play a key role in preterm complications highlighting the role of microbial cell surface components in host-immune modulation[39]. However, the role of microbial cell surface proteins, particularly glycans that can interact with host epithelial cells, has been poorly studied in vaginal microbial settings compared to the

gut microbiome[40]. This is due to a majority of studies being limited by a reliance on 16S rRNA sequencing. As such, detailed strain-level genomic insights are lacking in the field. Despite this, recent studies have shown that some *Lactobacillus crispatus* strains with pullulanase type I gene (*pulA*) can directly degrade glycogen[41,42]. It is also reported that the *L. crispatus* strains isolated from dysbiotic vaginal microbiota have a distinct glycosyltransferase gene that could play a role in cell surface glycome [41], further highlighting the potential importance of strain-level differences in vaginal bacteria.

Here we report two distinct gene clusters in *L. jensenii* with one predicted to have a diverse carbohydrate utilisation potential and another cluster to be involved in the biosynthesis of cell surface glycan with a potential role in preterm complication. The ability to selectively use various carbon sources such as glycogen and mucins can be beneficial for *L. jensenii* to survive in the fluctuating vaginal environment. Glycogen produced by vaginal epithelial cells is considered a nutrition source for vaginal bacteria[43], but is thought to be dependent on the initial activity of host amylase to facilitate subsequent use by vaginal commensals[44]. We identify the presence of probable alpha/beta-amylase genes, GH13/GH14 (DUF4038), in glycan utilisation gene cluster along with a gene predicted to encode an alpha-L-rhamnosidase, which may be involved in the hydrolysis of α-linked L-rhamnosides from polysaccharides. This suggests *L. jensenii* might have the ability to utilize distinct carbon sources. Notably, a recent metatranscriptomics study [45] identified GH13 and glycosyl-transferase (GT) 35 as active glycogen-debranching genes across various CSTs. Additionally, the identified variations in DUF4038 (GH13) and *ccpA* genes in the preterm strains relative to the full-term strains indicate a functional difference in metabolic pathways for carbohydrates. Previous studies have shown subtle genetic variations in the glycan utilisation genes can have a strong impact on the phenotype of the strain[59].

In the glycan utilisation gene cluster, we have identified a fucose isomerase gene along with alpha-beta hydrolase fold with potential serine protease flanking this region. It is possible that *L. jensenii* can use this gene machinery for adhesion to cervical mucus present in the vaginal epithelium[46,47]. As the viscoelastic structures of

mucins are the stiff arrangement of carbohydrate side chains with serine-threonine peptide backbone, production of a serine protease may help reduce the viscoelasticity[47] and further adhesion. Interestingly, we have identified an MFS transporter, two amino acid permeases and two GHs belonging to potential GH1 (β-glucosidase/ β-galactosidase) and GH28 (polygalacturonase) groups flanking this gene cluster. As galactose is a common component of mucins, this suggests that *L. jensenii* may utilise sugars from mucin oligosaccharides for nutrition[45]. Critchfield et al.[48] showed that cervical mucus from women at high preterm risk has more extensible and weaker gels than low-risk women, suggesting that alterations to surface mucus may contribute to PTB.

Microbial cell surface proteins are crucial to interactions with the vaginal epithelial cells for various functions from commensalism to invasion[49]. However, the microbial cell surface glycome in the vaginal microbiome is poorly studied particularly in the vaginal commensals like *L. jensenii*. A recent study has identified a distinct glycosyltransferase in a gene cluster in *L. crispatus* from dysbiotic vaginal microbiota suggesting the possibility of distinct microbial cell surface glycome in vaginal bacteria from dysbiotic environments[41]. In this study, we report a distinct gene cluster in vaginal *L. jensenii* containing genes involved in the biosynthesis and export of cell surface proteins with a potential role in microbial-host interactions. We found *cap8A*, *ywqD* and *ywqE* homologues in this cluster, which have been reported to have a direct role in the polysaccharide production in *L. lactis* and *Bacillus subtilis*[50,51]. The presence of potential diverse glycosyl transferases indicate the role of this gene cluster in glycan synthesis. We found a YSIRK signal motif which is associated with the efficient translocation of the surface protein into cellular membranes for anchoring or secretion from the cell[52]. Moreover, the presence of a flippase, known to have a key role in export of lipid bound oligosaccharides in *E. coli*[53], provides further evidence that this gene cluster is involved in the synthesis and translocation of glycans to the cell surface. The fact that these genes are associated predominantly with preterm strains indicates the potential for a difference in the cell surface glycome between preterm and full-term strains, which could have implications for immune evasion.

Further, core genome phylogeny revealed clustering of preterm and full-term strains in two different clades indicating genomic variations distinct to the preterm group. Although the PTB clade did contain three genomes from full-term deliveries, we noted that one of these genomes was from a woman with a previous PTB, and others had previous LLETZ surgery, both risk factors in themselves for PTB. Moreover, we identified variations in the core genes, notably in the acetate, lactate and cell wall related functions in preterm group. Similarly, we have also identified these gene variations in the publicly available *L. jensenii* genomes. Taken together this pattern raises the possibility that host vaginal conditions could select for different strains of *L. jensenii*. The largest number of non-synonymous mutations was present in an essential acetate metabolism gene, acetate kinase (*ackA*). In *Escherichia coli,* it has been shown that deletion of *ackA* led to severe growth impairment by the reduction in glucose uptake rate and concentrations of ATP generated, highlighting the importance of this gene[54]. Notably, a study of 82 pregnant women showed that levels of cervicovaginal fluid acetate were predictive of preterm outcome[55]. Another study with a Caucasian population reported that *L. jensenii* dominance and decreased lactate levels were observed in women who delivered preterm [30]. In preterm strains, we identified three variations in D-lactate dehydrogenase (*ldhD*), a key gene responsible for D-lactic acid production. In addition, other mutations are found in core genes involved in the regulation of glycosyl hydrolases (*ccpA*), such as β-glucosidase and β-galactosidase and bacterial genetic diversity (*mutS2*)[56,57]. Taken together, these results suggest a functional difference between full-term and preterm strains including lactate and acetate metabolism that may play a role in vaginal homeostasis.

The findings of this study provide insights relating to strain-level genomic signatures of *L. jensenii* from preterm and full-term isolates, which could be useful in developing biomarkers of PTB. Possessing the genetic machinery to ferment various carbohydrates along with adhesion genes can be beneficial for *L. jensenii* to survive in conditions associated with preterm birth and the presence of different cell surface synthesis genes may even be indicative of distinct microbial-host immune interactions in preterm strains. The high-quality genomes generated in this study

enabled us to identify the subtle genomic differences in preterm and full-term groups. However, limitations exist due to the relatively small number of vaginal *L. jensenii* genomes and MAGs associated with PTB. Further studies with larger samples coupled with *in vitro* studies are needed to validate the implications of these genomic variations to better understand *L. jensenii*'s role in preterm.

4.6 Methods

4.6.1 Bacterial isolation and genome sequencing

Samples analysed in this study were previously collected as reported[26] with institutional ethics approval by the National Maternity Hospital Research Ethics Committee on 15th December 2015 and maternal written consent. From the same swabs used for metagenomic sequencing, an aliquot of resuspension PBS was serial diluted and plated onto DeMan Rogosa Sharpe (MRS) agar and incubated anaerobically at 37°C for up to 48 h. Colonies were subcultured onto MRS agar to ensure purity. The 16S rRNA gene from isolates was amplified by PCR using universal primers CO1 & CO2[58] and Sanger sequenced to identify the species.

For whole genome sequencing, an overnight culture of an isolate identified by Sanger sequencing as *Lactobacillus jensenii* was used for extraction of genomic DNA using the MoBio PowerFood kit, following manufacturer's instructions. DNA was subsequently quantified using Qubit HS DNA quantification kit and prepares for shotgun sequencing following Illumina Nextera XT protocol with sequencing on the Illumina NextSeq 500 platform.

4.6.2 Genome assembly of cultured isolates

Adapter removal and quality filtering of the raw paired end fastq reads was performed using trim galore (v 0.6.6)[59], cutadapt (v 1.18)[60] and FastQC (v 0.11.9). The resulting filtered reads were assembled using spades (v 3.14.1)[61] with isolate mode and *L. jensenii* RefSeq genome (NZ_CP018809.1) was used as reference for assembly. The draft genomes were further scaffolded using online Medusa with RefSeq genome[62]. The quality of the genomes was determined using QUAST (v 5.1.0)[63].

4.6.3 Reconstruction of metagenome assembled genomes

Metagenomic sequencing reads from the same samples cultured above and previously reported were used in this study[26]. Quality filtering and host contamination removal was done with read_qc module in the metaWRAP (v 1.3.2)[64] using raw paired fastq reads. The assembly of contigs from the filtered reads were assembled using assembly module with metaspades option in the metaWRAP pipeline. The assembled contigs were binned with the binning module with metabat2, maxbin2 and concoct options. The resulting bins were refined and reassembled with the filtered reads to improve the quality of the bins using bin_refinement and reassemble_bins modules in the metaWRAP pipeline. The bins with completeness ≥ 90% and contamination < 5% were considered high quality and retained for downstream analyses. The *L. jensenii* bins were identified with phylophlan (v 0.32)[65] using SGB.Dec20 database.

4.6.4 Pangenome and phylogenetic analysis

The pangenome of *L. jensenii* genomes were reconstructed using roary (v 3.13)[66] with "-e –r –z" parameters. The gene differences between fullterm and preterm genomes were identified using query_pan_genome script in the roary software. Genes associated with preterm birth were identified using the Scoary script. Heap's law was used to determine whether the pangenome is open ($\gamma > 0$) or closed ($\gamma < 0$). The core genome sequences from roary output were aligned with mafft (v 7.475)[67]. The resulting alignment was used for the reconstruction of maximum likelihood phylogeny using IQ-TREE (v 2.1.3)[68] with 3000 bootstraps option. The output trees in the newick format were visualized using iTOL[69].

4.6.5 Core genome variations and gene cluster analysis

Individual core gene families were extracted from the pangenome using roary multi_fasta script. The gene families were individually aligned with mafft (v 7.475)[67]. Using these sequences variations in the core genes between full term and preterm groups were identified with a custom python script. The genes were annotated with prokka[70] within roary program. The function of unannotated genes was further found using Eggnog mapper (v 2.1.7)[71] and online BLAST[72] against NCBI-NR database. The similarity and visualisation of the gene clusters were done using EasyFig (v 2.2.5)[73].

4.6.6 Profiling of gene clusters in publicly available genomes

The publicly available 142 *L. jensenii* genomes were downloaded from the PATRIC database[74] on 25th April 2023. The high quality 63 genomes were filtered with >90% completeness and <5% contamination and only vaginal isolates. Average nucleotide identity was calculated with *L. jensenii* and *L. mulieris* RefSeq genomes using fastANI (v 1.32)[75] and 10 *L. mulieris* genomes were removed resulting in 53 high quality *L. jensenii* genomes. The genes from the carbohydrate utitlisation and cell surface synthesis clusters were used as a database to identify the genes across the 53 *L. jensenii* using BLASTp (v 2.8.1)[72].

4.7 Acknowledgements

This research was funded by a research grant from Science Foundation Ireland (SFI) under [Grant No. 12/RC/2273_P2 and 16/SP/3827] and an Irish Research Council fellowship (EPSPD/2016/25) with Alimentary Health Ltd. S.R.C. Nori has received funding by Science Foundation Ireland through the SFI Centre for Research Training in Genomics Data Science under Grant number 18/CRT/6214 and also supported in part by the EU's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant H2020-MSCA-COFUND-2019-945385.

4.8 References

1. Huttenhower, C. *et al.* Structure, function and diversity of the healthy human microbiome. *Nature* 486, 207–214 (2012).

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

3. Zhou, X. *et al.* Characterization of vaginal microbial communities in adult healthy women using cultivation-independent methods. *Microbiology (Reading, England)* 150, 2565–2573 (2004).

4. Gong, Z., Luna, Y., Yu, P. & Fan, H. Lactobacilli Inactivate Chlamydia trachomatis through Lactic Acid but Not H2O2. *PLoS ONE* 9, e107758 (2014).

5. O'Hanlon, D. E., Moench, T. R. & Cone, R. A. In vaginal fluid, bacteria associated with bacterial vaginosis can be suppressed with lactic acid but not hydrogen peroxide. *BMC Infectious Diseases* 11, 200 (2011).

6. O'Hanlon, D. E., Moench, T. R. & Cone, R. A. Vaginal pH and Microbicidal Lactic Acid When Lactobacilli Dominate the Microbiota. *PLoS ONE* 8, e80074 (2013).

7. Aroutcheva, A. *et al.* Defense factors of vaginal lactobacilli. *American Journal of Obstetrics and Gynecology* 185, 375–379 (2001).

8. Delgado-Diaz, D. J. *et al.* Distinct Immune Responses Elicited From Cervicovaginal Epithelial Cells by Lactic Acid and Short Chain Fatty Acids Associated With Optimal and Non-optimal Vaginal Microbiota. *Frontiers in Cellular and Infection Microbiology* 9, (2020).

9. Witkin, S. S. *et al.* Influence of Vaginal Bacteria and <scp>d</scp> - and <scp>l</scp> -Lactic Acid Isomers on Vaginal Extracellular Matrix Metalloproteinase Inducer: Implications for Protection against Upper Genital Tract Infections. *mBio* 4, (2013).

10. Alonzo Martínez, M. C. *et al.* Study of the Vaginal Microbiota in Healthy Women of Reproductive Age. *Microorganisms* 9, 1069 (2021).

11. Tsementzi, D. *et al.* Comparison of vaginal microbiota in gynecologic cancer patients pre- and post-radiation therapy and healthy women. *Cancer Medicine* 9, 3714–3724 (2020).

12. Srinivasan, S. & Fredricks, D. N. The Human Vaginal Bacterial Biota and Bacterial Vaginosis. *Interdisciplinary Perspectives on Infectious Diseases* 2008, 1–22 (2008).

13. Carter, K. A. *et al.* Associations Between Vaginal Bacteria and Bacterial Vaginosis Signs and Symptoms: A Comparative Study of Kenyan and American Women. *Frontiers in Cellular and Infection Microbiology* 12, (2022).

14. Gosmann, C. *et al.* Lactobacillus-Deficient Cervicovaginal Bacterial Communities Are Associated with Increased HIV Acquisition in Young South African Women. *Immunity* 46, 29–37 (2017).

15. Brotman, R. M. *et al.* Association Between Trichomonas vaginalis and Vaginal Bacterial Community Composition Among Reproductive-Age Women. *Sexually Transmitted Diseases* 39, 807–812 (2012).

16. Leitich, H. & Kiss, H. Asymptomatic bacterial vaginosis and intermediate flora as risk factors for adverse pregnancy outcome. *Best Practice & Research Clinical Obstetrics & Gynaecology* 21, 375–390 (2007).

17. Leitich, H. *et al.* Bacterial vaginosis as a risk factor for preterm delivery: A meta-analysis. *American Journal of Obstetrics and Gynecology* 189, 139–147 (2003).

18. Perin, J. *et al.* Global, regional, and national causes of under-5 mortality in 2000–19: an updated systematic analysis with implications for the Sustainable Development Goals. *The Lancet Child & Adolescent Health* 6, 106–115 (2022).

19. Fettweis, J. M. *et al.* The vaginal microbiome and preterm birth. *Nature Medicine* 25, 1012–1021 (2019).

20. Romero, R., Dey, S. K. & Fisher, S. J. Preterm labor: One syndrome, many causes. *Science* 345, 760–765 (2014).

21. JEFFCOAT, M. K. *et al.* Periodontal infection and preterm birth. *The Journal of the American Dental Association* 132, 875–880 (2001).

22. Hyman, R. W. *et al.* Diversity of the Vaginal Microbiome Correlates With Preterm Birth. *Reproductive Sciences* 21, 32–40 (2014).

23. Subramaniam, A. *et al.* Vaginal Microbiota in Pregnancy: Evaluation Based on Vaginal Flora, Birth Outcome, and Race. *American Journal of Perinatology* 33, 401–408 (2015).

24. Haque, M. M., Merchant, M., Kumar, P. N., Dutta, A. & Mande, S. S. First-trimester vaginal microbiome diversity: A potential indicator of preterm delivery risk. *Scientific Reports* 7, 16145 (2017).

25. Witkin, S. Vaginal microbiome studies in pregnancy must also analyse host factors. *BJOG: An International Journal of Obstetrics & Gynaecology* 126, 359–359 (2019).

26. Feehily, C. *et al.* Shotgun sequencing of the vaginal microbiome reveals both a species and functional potential signature of preterm birth. *npj Biofilms and Microbiomes* 6, 1–9 (2020).

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

28. Romero, R. *et al.* The vaginal microbiota of pregnant women who subsequently have spontaneous preterm labor and delivery and those with a normal delivery at term. *Microbiome* 2, 1–15 (2014).

29. Gudnadottir, U. *et al.* The vaginal microbiome and the risk of preterm birth: a systematic review and network meta-analysis. *Scientific Reports* 12, 7926 (2022).

30. Stafford, G. P. *et al.* Spontaneous preterm birth is associated with differential expression of vaginal metabolites by lactobacilli-dominated microflora. *Frontiers in Physiology* 8, 615 (2017).

31. Amabebe, E., Reynolds, S. & Anumba, D. O. C. Spectral binning of cervicovaginal fluid metabolites improves prediction of spontaneous preterm birth and Lactobacillus species dominance. *Reproduction and Fertility* 2, L4–L6 (2021).

32. de Freitas, A. S. *et al.* Defining microbial biomarkers for risk of preterm labor. *Brazilian Journal of Microbiology* 51, 151–159 (2020).

33. Pace, R. M. *et al.* Complex species and strain ecology of the vaginal microbiome from pregnancy to postpartum and association with preterm birth. *Med* 2, 1027-1049.e7 (2021).

34. Sheydina, A. *et al.* Structural genomics analysis of uncharacterized protein families overrepresented in human gut bacteria identifies a novel glycoside hydrolase. *BMC Bioinformatics* 15, 1–10 (2014).

35. Mehta, S. D. *et al.* The Microbiome Composition of a Man's Penis Predicts Incident Bacterial Vaginosis in His Female Sex Partner With High Accuracy. *Frontiers in cellular and infection microbiology* 10, 433 (2020).

36. Gilbert, J. A. & Lynch, S. V. Community ecology as a framework for human microbiome research. *Nature Medicine 2019 25:6* 25, 884–889 (2019).

37. Kindschuh, W. F. *et al.* Preterm birth is associated with xenobiotics and predicted by the vaginal metabolome. *Nature Microbiology* 8, 246–259 (2023).

38. Lee, J.-Y. *et al.* Development of a New Biomarker Model for Predicting Preterm Birth in Cervicovaginal Fluid. *Metabolites* 12, 734 (2022).

39. Chan, D. *et al.* Microbial-driven preterm labour involves crosstalk between the innate and adaptive immune response. *Nature Communications* 13, 975 (2022).

40. Sanozky-Dawes, R. & Barrangou, R. Lactobacillus, glycans and drivers of health in the vaginal microbiome. *Microbiome Research Reports* 1, 18 (2022).

41. Van Der Veer, C. *et al.* Comparative genomics of human Lactobacillus crispatus isolates reveals genes for glycosylation and glycogen degradation: Implications for in vivo dominance of the vaginal microbiota. *Microbiome* 7, 1–14 (2019).

42. Nunn, K. L. *et al.* Amylases in the Human Vagina. *mSphere* 5, (2020).

43. Tester, R. & Al-Ghazzewi, F. H. Intrinsic and extrinsic carbohydrates in the vagina: A short review on vaginal glycogen. *International Journal of Biological Macromolecules* 112, 203–206 (2018).

44. Spear, G. T. *et al.* Human α-amylase Present in Lower-Genital-Tract Mucosal Fluid Processes Glycogen to Support Vaginal Colonization by Lactobacillus. *The Journal of Infectious Diseases* 210, 1019–1028 (2014).

45. France, M. T. *et al.* Insight into the ecology of vaginal bacteria through integrative analyses of metagenomic and metatranscriptomic data. *Genome Biology* 23, 1–26 (2022).

46. Odeblad, E. & Parade, A. The Discovery of Different Types of Cervical Mucus and the Billings Ovulation Method. (1994).

47. Wiggins, R., Hicks, S. J., Soothill, P. W., Millar, M. R. & Corfield, A. P. Mucinases and sialidases: their role in the pathogenesis of sexually transmitted infections in the female genital tract. *Sexually Transmitted Infections* 77, 402–408 (2001).

48. Critchfield, A. S. *et al.* Cervical Mucus Properties Stratify Risk for Preterm Birth. *PLOS ONE* 8, e69528 (2013).

49. Canny, G. O. & McCormick, B. A. Bacteria in the Intestine, Helpful Residents or Enemies from Within? *Infection and Immunity* 76, 3360–3373 (2008).

50. Nierop Groot, M. N. & Kleerebezem, M. Mutational analysis of the Lactococcus lactis NIZO B40 exopolysaccharide (EPS) gene cluster: EPS biosynthesis correlates with unphosphorylated EpsB. *Journal of Applied Microbiology* 103, 2645–2656 (2007).

51. Mijakovic, I. *et al.* In Vitro Characterization of the *Bacillus subtilis* Protein Tyrosine Phosphatase YwqE. *Journal of Bacteriology* 187, 3384–3390 (2005).

52. Bae, T. & Schneewind, O. The YSIRK-G/S motif of staphylococcal protein A and its role in efficiency of signal peptide processing. *Journal of Bacteriology* 185, 2910–2919 (2003).

53. Butler, E. K., Davis, R. M., Bari, V., Nicholson, P. A. & Ruiz, N. Structure-Function Analysis of MurJ Reveals a Solvent-Exposed Cavity Containing Residues Essential for Peptidoglycan Biogenesis in Escherichia coli. *Journal of Bacteriology* 195, 4639–4649 (2013).

54. Schütze, A., Benndorf, D., Püttker, S., Kohrs, F. & Bettenbrock, K. The Impact of ackA, pta, and ackA-pta Mutations on Growth, Gene Expression and Protein Acetylation in Escherichia coli K-12. *Frontiers in Microbiology* 11, 233 (2020).

55. Amabebe, E. *et al.* Cervicovaginal fluid acetate: A metabolite marker of preterm birth in symptomatic pregnant women. *Frontiers in Medicine* 3, 48 (2016).

56. Molloy, S. MutS2: key to diversity? *Nature Reviews Microbiology 2005 3:3* 3, 191–191 (2005).

57. Muscariello, L., Marasco, R., De Felice, M. & Sacco, M. The Functional ccpA Gene Is Required for Carbon Catabolite Repression in Lactobacillus plantarum. *Applied and Environmental Microbiology* 67, 2903 (2001).

58. Simpson, P. J., Stanton, C., Fitzgerald, G. F. & Ross, R. P. Genomic diversity and relatedness of bifidobacteria isolated from a porcine cecum. *Journal of Bacteriology* 185, 2571–2581 (2003).

59. Krueger, F. *et al.* FelixKrueger/TrimGalore: v0.6.10 - add default decompression path. (2023) doi:10.5281/ZENODO.7598955.

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

61. Prjibelski, A., Antipov, D., Meleshko, D., Lapidus, A. & Korobeynikov, A. Using SPAdes De Novo Assembler. *Current Protocols in Bioinformatics* 70, e102 (2020).

62. Bosi, E. *et al.* MeDuSa: a multi-draft based scaffolder. *Bioinformatics* 31, 2443–2451 (2015).

63. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 29, 1072–1075 (2013).

64. Uritskiy, G. V., Diruggiero, J. & Taylor, J. MetaWRAP - A flexible pipeline for genome-resolved metagenomic data analysis 08 Information and Computing Sciences 0803 Computer Software 08 Information and Computing Sciences 0806 Information Systems. *Microbiome* 6, 1–13 (2018).

65. Asnicar, F. *et al.* Precise phylogenetic analysis of microbial isolates and genomes from metagenomes using PhyloPhlAn 3.0. *Nature Communications 2020 11:1* 11, 1–10 (2020).

66. Page, A. J. *et al.* Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* 31, 3691–3693 (2015).

67. Katoh, K., Misawa, K., Kuma, K. I. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research* 30, 3059–3066 (2002).

68. Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution* 37, 1530–1534 (2020).

69. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Research* 49, W293–W296 (2021).

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

71. Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Molecular Biology and Evolution* 38, 5825–5829 (2021).

72. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *Journal of Molecular Biology* 215, 403–410 (1990).

73. Sullivan, M. J., Petty, N. K. & Beatson, S. A. Easyfig: a genome comparison visualizer. *Bioinformatics* 27, 1009 (2011).

74. Wattam, A. R. *et al.* Improvements to PATRIC, the all-bacterial Bioinformatics Database and Analysis Resource Center. *Nucleic acids research* 45, D535–D542 (2017).

75. Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T. & Aluru, S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nature Communications 2018 9:1* 9, 1–8 (2018).

5 Chapter 5

Strain-level variation among vaginal *Lactobacillus crispatus* and *Lactobacillus iners* as identified by comparative metagenomics

This chapter has been published in NPJ Biofilms and Microbiomes journal.

(doi: 10.1038/s41522-025-00682-1)

5.1 Abstract

The vaginal microbiome, a relatively simple, low diversity ecosystem crucial for female health, is often dominated by Lactobacillus spp. Detailed strain-level data, facilitated by shotgun sequencing, can provide a greater understanding of the mechanisms of colonization and host-microbe interactions. We analysed 354 vaginal metagenomes from pregnant women in Ireland to investigate metagenomic community state types and strain-level variation, focusing on cell surface interfaces. Our analysis revealed multiple subspecies, with *Lactobacillus crispatus* and *Lactobacillus iners* being the most dominant. We found genes, including putative mucin-binding genes, distinct to *L. crispatus* subspecies. Using 337 metagenome-assembled genomes, we observed a higher number of strain-specific genes in *L. crispatus* related to cell wall biogenesis, carbohydrate and amino acid metabolism, many under positive selection. A cell surface glycan gene cluster was predominantly found in *L. crispatus* but absent in *L. iners* and *Gardnerella vaginalis*. These findings highlight strain-specific factors associated with colonisation and host-microbe interactions.

5.2 Introduction

Lactobacilli play a major role in modulating the composition of the human vaginal microbiome. Production of lactic acid by these bacteria is an important trait that results in a lowering of the vaginal pH, preventing colonisation by unfavourable bacteria like those associated with bacterial vaginosis (BV)[1–3]. Additionally, production of hydrogen peroxide[4] and antimicrobial peptides can provide a competitive advantage to certain lactobacilli within this niche[5,6]. The dominance of lactobacilli within the vaginal microbiome is distinct to humans, not being observed in other non-human mammals[7]. The evolutionary pressure driving this phenomenon is unclear, however both dietary and hormonal cycling have been suggested[7,8]. Indeed, the high concentration of glycogen within the vaginal lumen, as well as an abundance of mucin, may be important as *Lactobacillus* species have the capacity to use the former glycan as a carbon source, while mucin is believed to provide a distinct interface for interaction between vaginal microbes and the host[9].

The vaginal microbiome in collective datasets has previously been categorised into community state types (CSTs)[10] primarily based on the dominant bacteria present, such as *Lactobacillus crispatus* (CST-I), *Lactobacillus gasseri* (CST-II), *Lactobacillus iners* (CST-III), and *Lactobacillus jensenii* (CST-V), with a further *Lactobacillus*-reduced community of obligate and facultative anaerobes such as *Gardnerella vaginalis* and *Fannyhessea vaginae* classified as CST-IV[10,11]. Due to the diversity of the dominant species, subgroupings or extensions of CSTs are often required, leading to discordance between reported studies. To address this, a nearest centroid based approach (VALENCIA) was developed to help standardise classification[12]. Ultimately, these stratification approaches are based on taxonomic abundance, whereby taxonomy is often assigned based on limited marker genes (as is the case for MetaPhlAN4)[13]. Therefore, the functional potential of the bacterial communities within the vaginal microbiome is not considered, nor are discrete genetic differences within species.

Recently, Holm and colleagues described metagenomic community state types (mgCSTs) based on both taxonomic and functional composition of the vaginal microbiomes from North American women using the non-redundant gene database, VIRGO[14,15]. Assignment of a microbiome to any mgCST is based on the relative abundances of metagenomic subspecies (mgSs). These mgSs are derived from distinct co-occurring genetic variation within a given species and thus provides more granular resolution of a community structure. Based on this approach a total of 27 mgCSTs are currently recognized for the vaginal microbiome. The mgCST-based approach enables a more comprehensive classification of the vaginal microbiome by integrating taxonomic information and functional potential that are crucial to appreciate subtle differences in the communities. For example, the authors define six distinct *L. crispatus* mgCSTs using the mgCST classification scheme, compared to just one CST using the classical CST method.

Multiple factors including microbial interactions, antibiotic treatment, nutrient availability, hormonal changes, and pregnancy can alter the composition of the vaginal microbiome by acting as selective pressures that further affect microdiversity[16–22]. Previous studies have shown that the bacterial communities of the vagina are stable during pregnancy compared to non-pregnant women[11,16]. However, strain-level studies of the vaginal microbiome are needed particularly in the context of pregnancy. Within other niches, analysis of metagenome assembled genomes (MAGs) has been shown to be a powerful approach with respect to describing community function[23,24]. Given the increasing evidence of the impact of subtle strain-level differences on phenotypic characteristics, it is important to investigate genotypes, especially those potentially involved in host-microbe interactions, at a resolution greater than genus and species[9,25,26]. Notably in this regard, differences with respect to mucin binding genes and vaginal cell adhesion have been reported between strains of *L. gasseri*[27], while Argentini *et al*. have demonstrated that even closely related strains of *L. crispatus* can vary greatly in terms of antimicrobial and competitive abilities[28]. Taken together, there is much to

learn about the microbial interactions in this niche by employing subspecies/strain resolved approaches.

The goal of this study was to perform a high-resolution profiling of vaginal metagenomes from pregnant women by analyses of mgCSTs and MAGs in order to understand observed genetic differences at both mgCST- and individual strain-level for genes predicted to be involved in adaptation to the vaginal niche.

5.3 Results

5.3.1 Metagenomic community state typing reveals multiple sub-species

A total of 354 vaginal metagenomic samples from two Irish cohorts, high-risk preterm (n = 87) (PRJEB34536) and MicrobeMom (n = 267) (PRJEB48251), with a median count of 336,878 (interquartile range of 176,702–796,475 reads) quality trimmed microbial read pairs per sample, were analysed. Using the metagenomic community state type (mgCST) classifier approach, we grouped the metagenomes at subspecies resolution and established that 18 distinct mgCSTs were present across the dataset. A total of five metagenomic subspecies (mgSs) from *Lactobacillus crispatus* (mgCST 1, mgCST 3-6) (Figure 5.1a) were identified, while three and two metagenomic subspecies (mgSs) were shown to correspond to *Lactobacillus iners* (mgCST 10-12 and 27) and *Lactobacillus jensenii* (mgCST 15 and 16), respectively. There were three mgSs identified for *Lactobacillus gasseri* (mgCST 7-9) as well as for *Gardnerella vaginalis* (mgCST 23-25), while only one subspecies was identified for *Bifidobacterium breve* (mgCST 26). In contrast, the conventional CST assignment approach, classified the vaginal metagenomes to only 6 CSTs (CST-1,2,3,4,5 and 8) at species level resolution. To investigate the dynamics and stability of mgCST throughout pregnancy, we compared assignment at both the 2nd and 3rd trimester in a subset of individuals (n = 48). Our analysis revealed that 27% (13/48) of the samples exhibited intra-species mgCST changes (mgCST shifts with the same dominant taxon) compared to the initial sample. We found inter-species mgCST in 17% (8/48) of the samples, indicating shifts to a different dominant taxon. No differences were observed in 56% of the samples (27/48).

Figure 5.1. Metagenomic community state types (mgCSTs) identified in 354 vaginal metagenomes

a. Heatmap showing the relative abundance of metagenomic subspecies in each sample. The barplot on the top represents the alpha diversity within the samples and color strip with gradients indicates the mgCSTs in each species. b, c. Pan-metagenome heatmaps of *L. crispatus* and *L. iners* representing the gene content within each mgCSTs. Heatmaps are clustered based on mgCSTs and indicated by color strip on the top. For each non-redundant gene, a dark green cell indicates its presence (1), and a white cell (0) indicates its absence in the corresponding sample.

5.3.2 Pan-metagenomics reveal mucin-binding genes in *L. crispatus*

Metagenomic community state types dominated by *L. crispatus* (48% of samples (173/354)) and *L. iners* (13% of samples (47/354)) accounted for the majority of all profiled metagenomes (Table 5.1). We therefore reconstructed pan-metagenomes of *L. crispatus* and *L. iners* using non-redundant genes of each species to understand differences in mgCSTs between and within these species (Figure 5.1b, 5.1c). The *L. crispatus* pan-metagenome consisted of 5943 vaginal orthologous groups (VOGs), with 4184 VOGs in the *L. iners* pan-metagenome. Both D- and L- lactate dehydrogenase-encoding genes were present in all *L. crispatus* mgCSTs (5/5 mgCSTs, 173/173 samples) while, *L. iners* mgCSTs lacked the gene encoding a D-lactate dehydrogenase. Similarly, whilst the glycogen debranching gene (*pulA*) was found in all mgCSTs in both *L. crispatus* (5/5 mgCSTs, 168/173 samples) and *L. iners* (2/2 mgCSTs, 47/47 samples), mucin binding genes (*mucBP*) were only present in *L. crispatus* mgCSTs (5/5 mgCSTs, 172/173 samples).

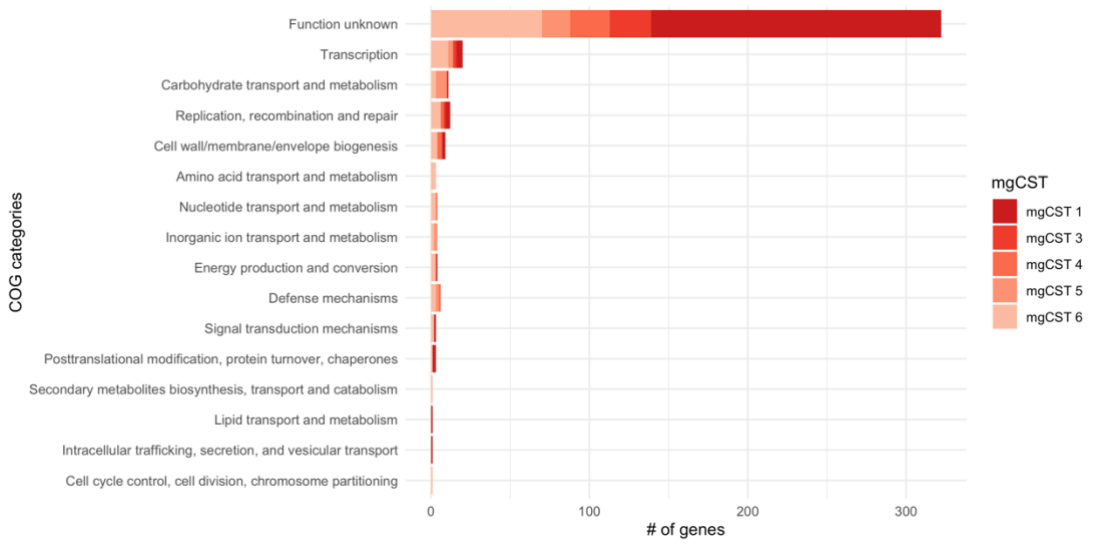
Table 5.1 | Metagenomic community state types identified in the dataset

mgCST	Frequent mgSs (prevalence)	Abundant mgSs	Samples
mgCST 4	*Lactobacillus crispatus 4*	*Lactobacillus crispatus 4*	85
mgCST 1	*Lactobacillus crispatus 1*	*Lactobacillus crispatus 1*	65
mgCST 10	*Lactobacillus iners 1*	*Lactobacillus iners 1*	35
mgCST 16	*Lactobacillus jensenii 2*	*Lactobacillus jensenii 2*	35
mgCST 27	*Bifidobacterium dentium*	*Enterococcus faecalis 3*	18
mgCST 8	*Lactobacillus gasseri 2*	*Lactobacillus gasseri 2*	14
mgCST 26	*Bifidobacterium breve*	*Bifidobacterium breve*	13
mgCST 9	*Lactobacillus gasseri 3*	*Lactobacillus gasseri 3*	13
mgCST 3	*Lactobacillus crispatus 3*	*Lactobacillus crispatus 3*	12

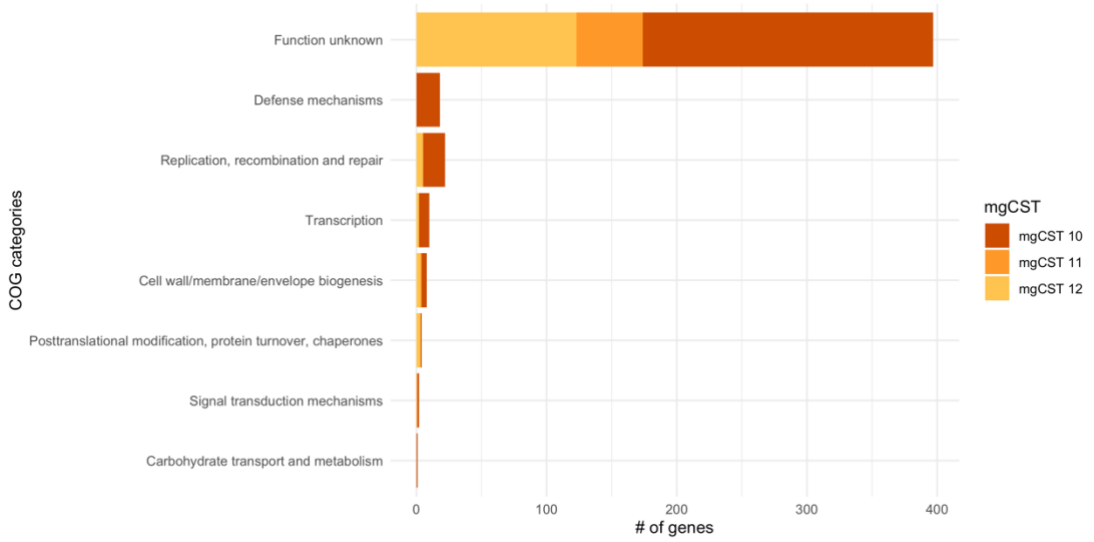
mgCST 25	*Gardnerella vaginalis 5*	*Gardnerella vaginalis 5*	11
mgCST 7	*Lactobacillus gasseri 1*	*Lactobacillus gasseri 1*	11
mgCST 23	*Gardnerella vaginalis 3*	*Gardnerella vaginalis 3*	10
mgCST 12	*Lactobacillus iners 3*	*Lactobacillus iners 3*	9
mgCST 5	*Lactobacillus crispatus 5*	*Lactobacillus crispatus 5*	7
mgCST 15	*Lactobacillus jensenii 1*	*Lactobacillus jensenii 1*	5
mgCST 24	*Gardnerella vaginalis 4*	*Gardnerella vaginalis 4*	4
mgCST 6	*Lactobacillus crispatus 6*	*Lactobacillus crispatus 6*	4
mgCST 11	*Lactobacillus iners 2*	*Lactobacillus iners 2*	3

Within the pan-metagenomes, we identified a total of 405 genes unique to individual *L. crispatus* mgCSTs, with mgCSTs 1 and 6 having the largest number of unique genes, 199/405 and 113/405, respectively. Most of these genes were annotated as 'hypothetical' for mgCST 1 (167/199) and mgCST 6 (53/113). Nonetheless, of those that could be functionally annotated, most of the unique genes assignable to COG functional categories were predicted to have roles in transcription, carbohydrate transport and metabolism, replication and repair and cell wall-related activities (Supplementary Figure 5.1a). A total of 462 unique genes were identified in *L. iners* mgCSTs, corresponding to mgCST 10 (273), mgCST 11 (53) and mgCST 12 (136). Similar to *L. crispatus*, the majority of these genes, i.e., 194 genes in mgCST 10, 45 genes in mgCST 11 and 114 genes in mgCST 12, had no assigned function. The top COG categories in *L. iners* were comparable to *L. crispatus* except for the defense mechanism category where *L. iners* has more mgCST-specific genes (Supplementary Figure 5.1b).

a)



b)



Supplementary Figure 5.1. Distribution of mgCST-Specific Genes in COG Functional Categories

a, b. Barplots showing the distribution of mgCST-specific genes across various functional categories in *L. crispatus* (a) and *L. iners* (b).

5.3.3 The accessory genome shapes the phylogenetic diversity of both *L. crispatus* and *L. iners*

To investigate the vaginal microbiome at strain-level resolution, we reconstructed 337 high-quality metagenome-assembled genomes (MAGs; completeness > 90%, contamination < 5%) from 354 metagenomes. *L. crispatus* (154), *L. iners* (57), and *L. jensenii* (31) represented the most frequently reconstructed species (Figure 5.2a). We found that ~100,000 *L. crispatus* reads were required to recover a high-quality MAG with only ~50,000 reads required for *L. iners* MAG recovery (Figure 5.2b). The median recovery was 1 MAG per sample with a maximum recovery of 8 MAGs in a single sample, reflecting the distribution of species-level diversity expected from the vaginal microbiome. There were no instances where multiple strains/MAGs of the same species were recovered from a single sample.

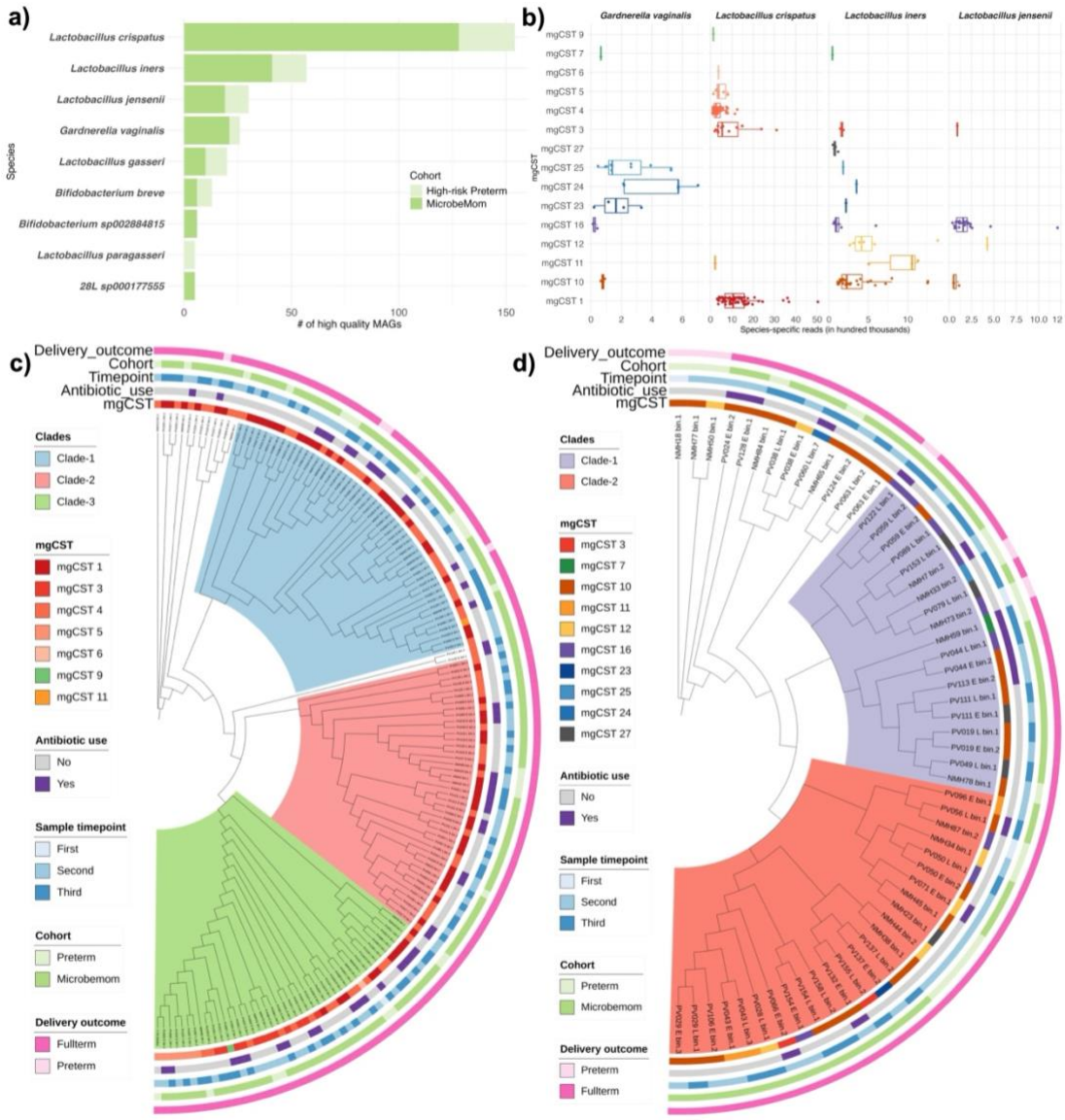


Figure 5.2. Metagenome-assembled genomes (MAGs) from vaginal metagenomes.

a. Stacked barplot showing the number of MAGs recovered from each species from the two cohorts. b. Boxplots are faceted by species and colored by mgCSTs representing the number of species-specific reads needed to recover high-quality MAGs from metagenomic samples. c, d. Maximum likelihood cladograms of *L. crispatus* (c) and *L. iners* (d) reconstructed based on core genomes. Branches are colored based on the phylogenetic clades and color strips from inner to outermost represent mgCSTs, antibiotic usage, sample collection timepoint, cohort and delivery outcome.

High-quality MAGs from *L. crispatus* and *L. iners* were used for phylogenetic analysis. Core genome (genes present in 95-100% of MAGs of that species) phylogenetic comparison revealed the presence of three major clades for *L. crispatus* whilst there were two major clades for *L. iners* (Figure 5.2c and 5.2d). There did not appear to be any relationship between mgCST classification and phylogenetic clades of the core genomes for either *L. crispatus or L. iners.* Furthermore, we reconstructed accessory genome (genes present in <15% of MAGs) phylogenies for *L. crispatus* and *L. iners* to investigate the concordance with the core phylogeny. This analysis revealed a topological discordance between core and accessory genome phylogenies in both *L. crispatus* and *L. iners* with a robinsonfould-distance of 0.58 and 0.70, respectively, indicating accessory genome content might play a role in shaping phylogenetic diversity within these species (Figure 5.3a and 5.3b).

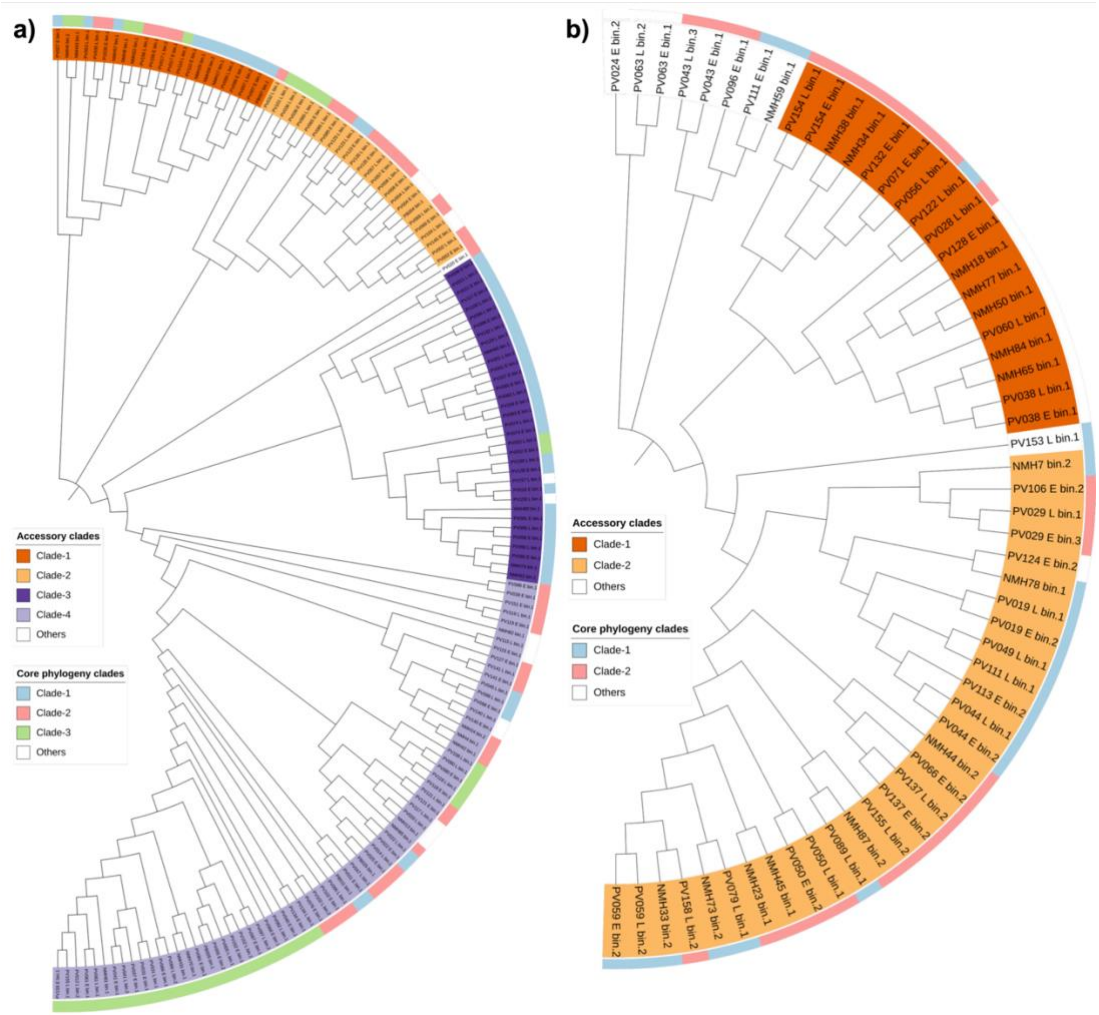


Figure 5.3. Phylogenetic Trees of *L. crispatus* and *L. iners.*

a, b. Maximum likelihood phylogenetic trees of *L. crispatus* (a) and *L. iners* (b) reconstructed based on accessory genome content. Branches are colored based on the phylogenetic clades and the color strip represents the core genome phylogenetic clades.

5.3.4 The intra-species genetic diversity of vaginal *L. crispatus* is greater than that of *L. iners*

We reconstructed the pangenomes for *L. crispatus* and *L. iners*, as these species are supported by at least 50 high-quality metagenome-assembled genomes (MAGs) each to understand the strain-level differences between the clades (Figure 5.4a and 5.4b). The pangenomes of *L. crispatus* and *L. iners* contained a total of 8831 and 3221 gene families, respectively. Within *L. crispatus*, 1109 gene families (12.5% of pangenome) were present in more than 95% of the MAGs (core genome) while 5415 gene families (61.3%) were present in less than 15% (cloud genome). For *L. iners,* the core genome comprised of 835 gene families (25.9%) and had a cloud genome of 1805 (56%). Additionally, we randomly sampled three sets of 57 *L. crispatus* MAGs to match *L. iners* sample size and reconstructed the pangenome and found cloud gene content (mean count of 3322 genes - 53% of pangenome) was always larger than the core (mean count of 1116 genes - 18%) in the *L. crispatus* pangenome (Supplementary Figure 5.2a). Further analysis revealed an open pangenome ($\gamma>0$) for both *L. crispatus* ($\gamma=0.25$) and *L. iners* ($\gamma=0.26$) suggesting high intra-species genetic diversity.

Next, we examined the distribution profiles of core and cloud genes in *L. crispatus* and *L. iners* in various functional categories. Genes putatively involved in cell wall biogenesis (Cloud - 184 vs Core - 35), carbohydrate (160 vs 34) and amino acid transport, and metabolism (151 vs 34) appeared to be more abundant in the cloud of *L. crispatus* compared to the core (Supplementary Figure 5.2b). In contrast, we found a similar number of cloud and core genes in the *L. iners* pangenome associated with cell wall biogenesis (Cloud – 30 vs Core - 22), carbohydrate (36 vs 42) and amino

acid transport and metabolism (30 vs 22) categories. Inter-clade analysis within the *L. crispatus* species revealed 16 genes that were specific to clade-1. No other *L. crispatus* clades had unique genes. Of the 16 genes from clade-1, all were annotated as encoding for hypothetical proteins. There were no clade-specific genes identified in *L. iners*.

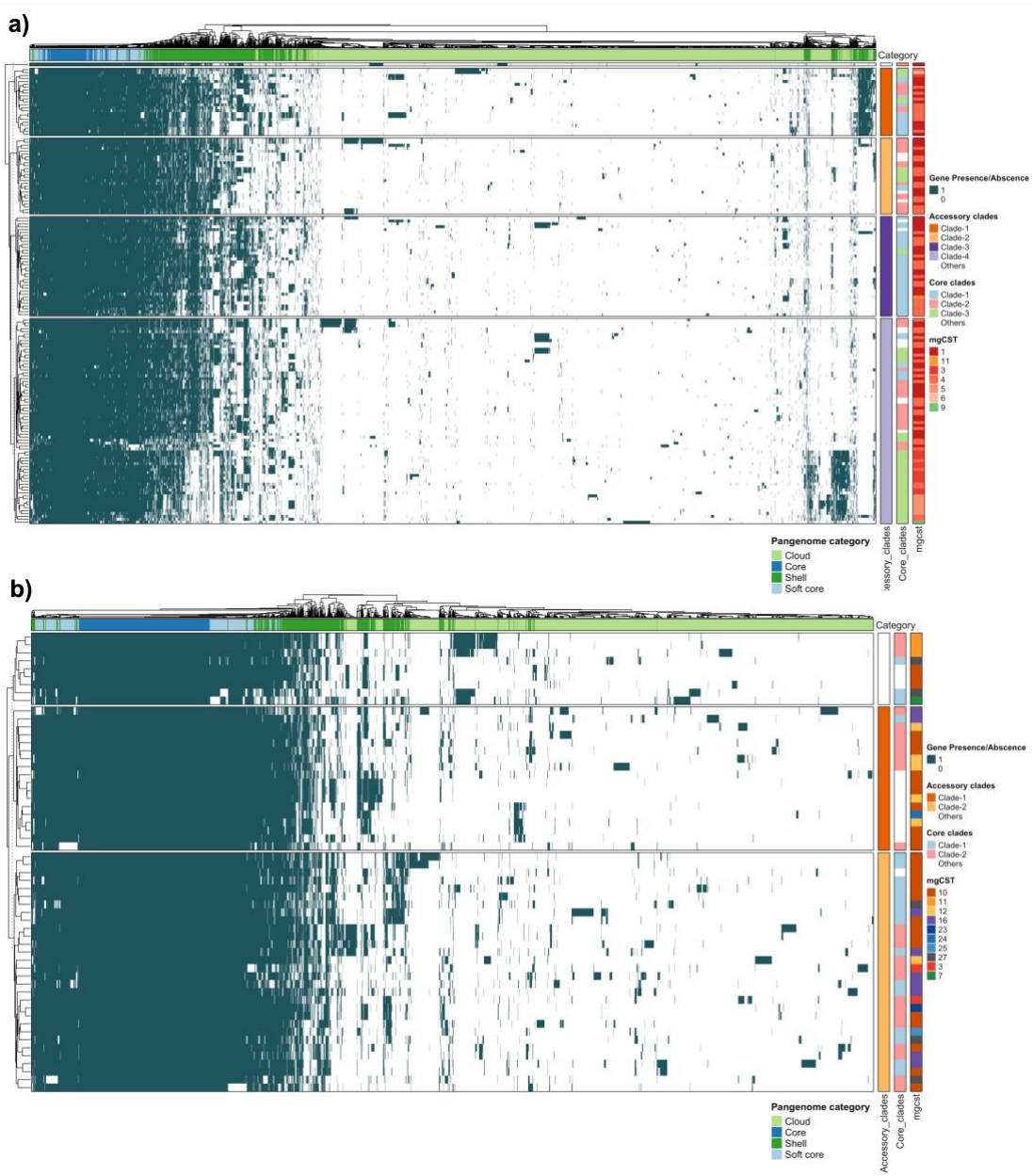
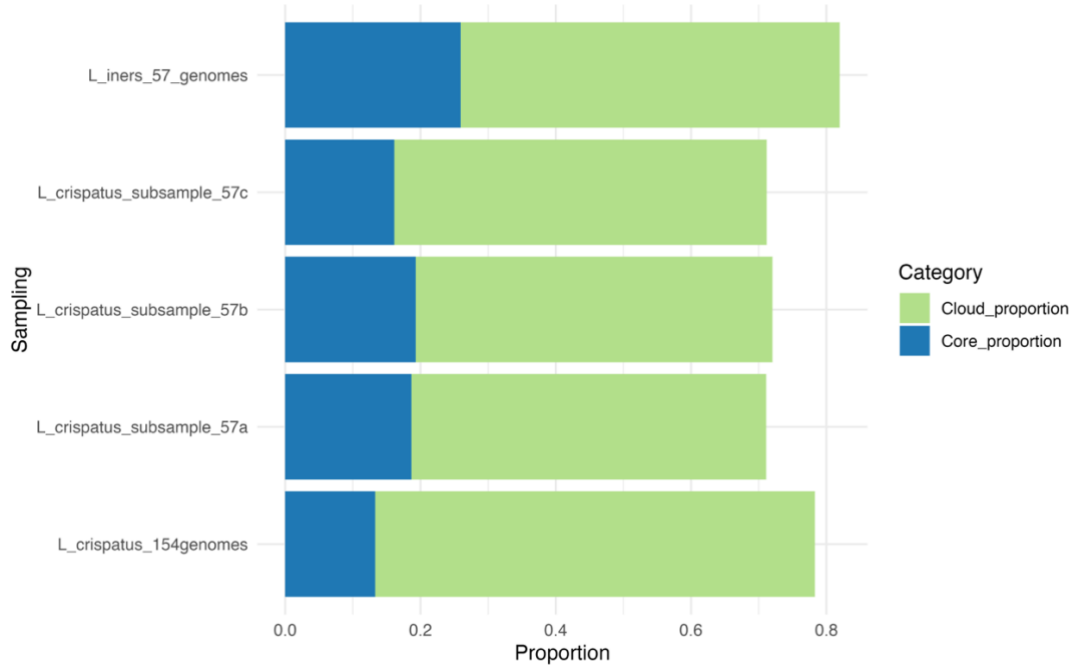


Figure 5.4. Pangenomes of *L. crispatus* and *L. iners* reconstructed from high-quality MAGs.

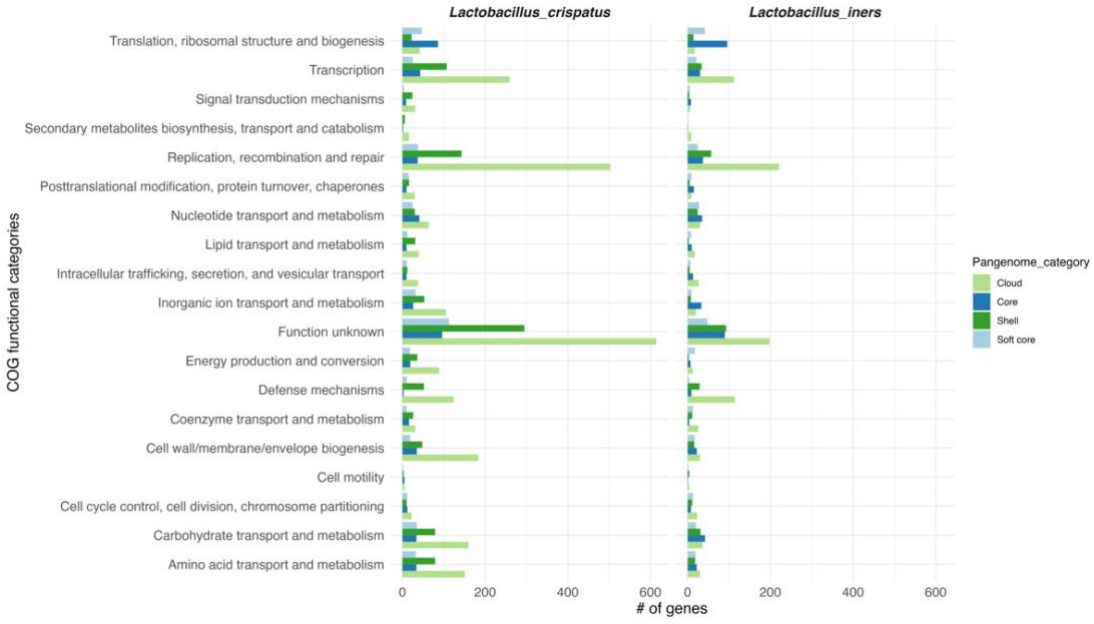
a, b. Pangenome heatmaps of *L. crispatus* (a) and *L. iners* (b) representing the gene content within each MAG. The heatmaps were clustered based on accessory clades

indicated with color strip on the side. For each gene, a dark green cell indicates its presence, and a white cell indicates its absence in the corresponding MAG. The color strip on the top indicates the pangenome category of the corresponding gene and side color strips represents the core genome phylogenetic clades and mgCSTs.

a)



b)



Supplementary figure 5.2. Pangenome content distribution across various functional categories in *L. iners* and *L. crispatus*

a. Stacked barplots showing the comparison of proportion of the cloud and core gene content between *L. iners* and three random samplings of *L*. *crispatus*. b. Barplots of *L. crispatus (n=170)* and *L. iners (n=76)* showing the distribution of the genes of each pangenome category across various functional categories.

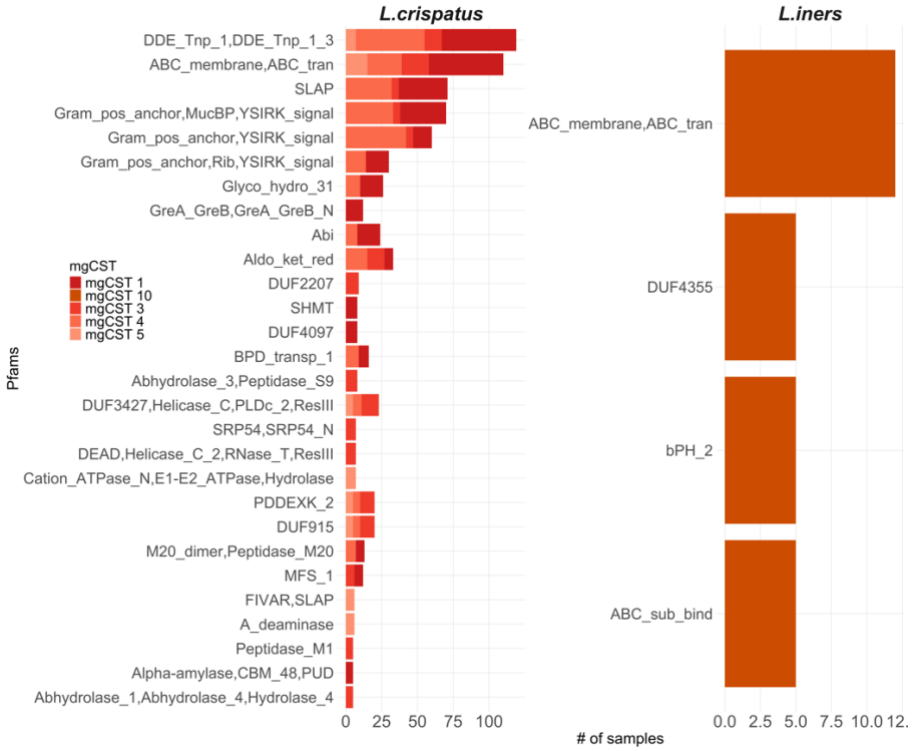
5.3.5 Metagenomic strain profiling reveals potential genes involved in host-microbe interface in *L. crispatus*

Given the marked differences in identified strain-specific genes between *L. crispatus* and *L. iners*, we aimed to identify the genes driving this by searching for genes under positive selection in the metagenomic data. We found 28 genes under positive selection in *L. crispatus* and four genes in *L. iners* (Figure 5.5a). The most prevalent genes under positive selection in *L. crispatus* were predicted to encode transposons (DDE_Tnps), ATP-binding cassette transporters (ABC transporters), S-layer associated protein (SLAP), Gram positive cell wall anchor (Gram_pos_anchor) and mucin-binding protein (MucBP). These genes were found irrespective of the associated mgCSTs assigned to the source metagenome. In *L. iners*, three genes were related to membrane transport (ABC transporters and bPH_2) and the other is annotated as encoding a domain of unknown function (DUF4355).

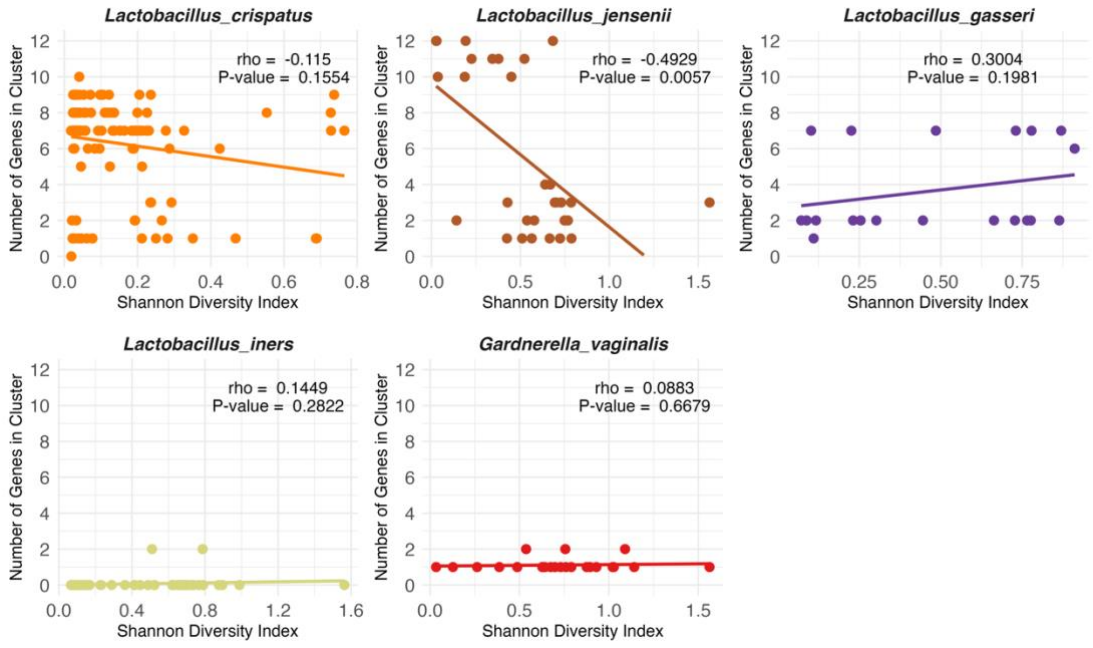
Five species (*L. crispatus, L. gasseri, L iners, L. jensenii,* and *G.* vaginalis) provided greater than 20 high quality MAGs. Within these we observed that a putative cell surface glycan encoding gene cluster containing *mucBP* was predominantly present in *L. crispatus* (86%, 113/154) and not in *L. iners* (0/57) or *G. vaginalis* (0/26) (Figure 5.5c). In *L. jensenii* and *L. gasseri,* we found this gene cluster in 36% (11/30) and 30% (6/20) of the samples, respectively. Moreover, we found that a lower microbiome alpha diversity in a given sample is correlated with a higher probability of detecting this cell surface glycan gene cluster in *L. crispatus* and *L. jensenii* (*p = 0.005*) species (Figure 5.5b).

Finally, we extended this investigation to publicly available genomes of *L. crispatus* (n = 406) and *L. iners* (n = 279) from different countries and sampling sites to survey the presence of this cell surface gene cluster in a more geographically broader context. We found similar results to our dataset, where the gene cluster was present in 63% (256/406) of *L. crispatus* genomes, yet absent in *L. iners* (0/279) (Supplementary Figure 5.3).

a)


L.crispatus
L.iners
DDE_Tnp_1,DDE_Tnp_1_3
ABC_membrane,ABC_tran
SLAP
Gram_pos_anchor,MucBP,YSIRK_signal
Gram_pos_anchor,YSIRK_signal
Gram_pos_anchor,Rib,YSIRK_signal
Glyco_hydro_31
GreA_GreB,GreA_GreB_N
Abi
Aldo_ket_red
DUF2207
SHMT
DUF4097
BPD_transp_1
Abhydrolase_3,Peptidase_S9
DUF3427,Helicase_C,PLDc_2,ResIII
SRP54,SRP54_N
DEAD,Helicase_C_2,RNase_T,ResIII
Cation_ATPase_N,E1-E2_ATPase,Hydrolase
PDDEXK_2
DUF915
M20_dimer,Peptidase_M20
MFS_1
FIVAR,SLAP
A_deaminase
Peptidase_M1
Alpha-amylase,CBM_48,PUD
Abhydrolase_1,Abhydrolase_4,Hydrolase_4
mgCST
mgCST 1
mgCST 10
mgCST 3
mgCST 4
mgCST 5
Pfams
0 25 50 75 100
ABC_membrane,ABC_tran
DUF4355
bPH_2
ABC_sub_bind
0.0 2.5 5.0 7.5 10.0 12.
# of samples

b)


Lactobacillus_crispatus
rho = -0.115
P-value = 0.1554
Lactobacillus_jensenii
rho = -0.4929
P-value = 0.0057
Lactobacillus_gasseri
rho = 0.3004
P-value = 0.1981
Lactobacillus_iners
rho = 0.1449
P-value = 0.2822
Gardnerella_vaginalis
rho = 0.0883
P-value = 0.6679
Number of Genes in Cluster
Shannon Diversity Index

c)

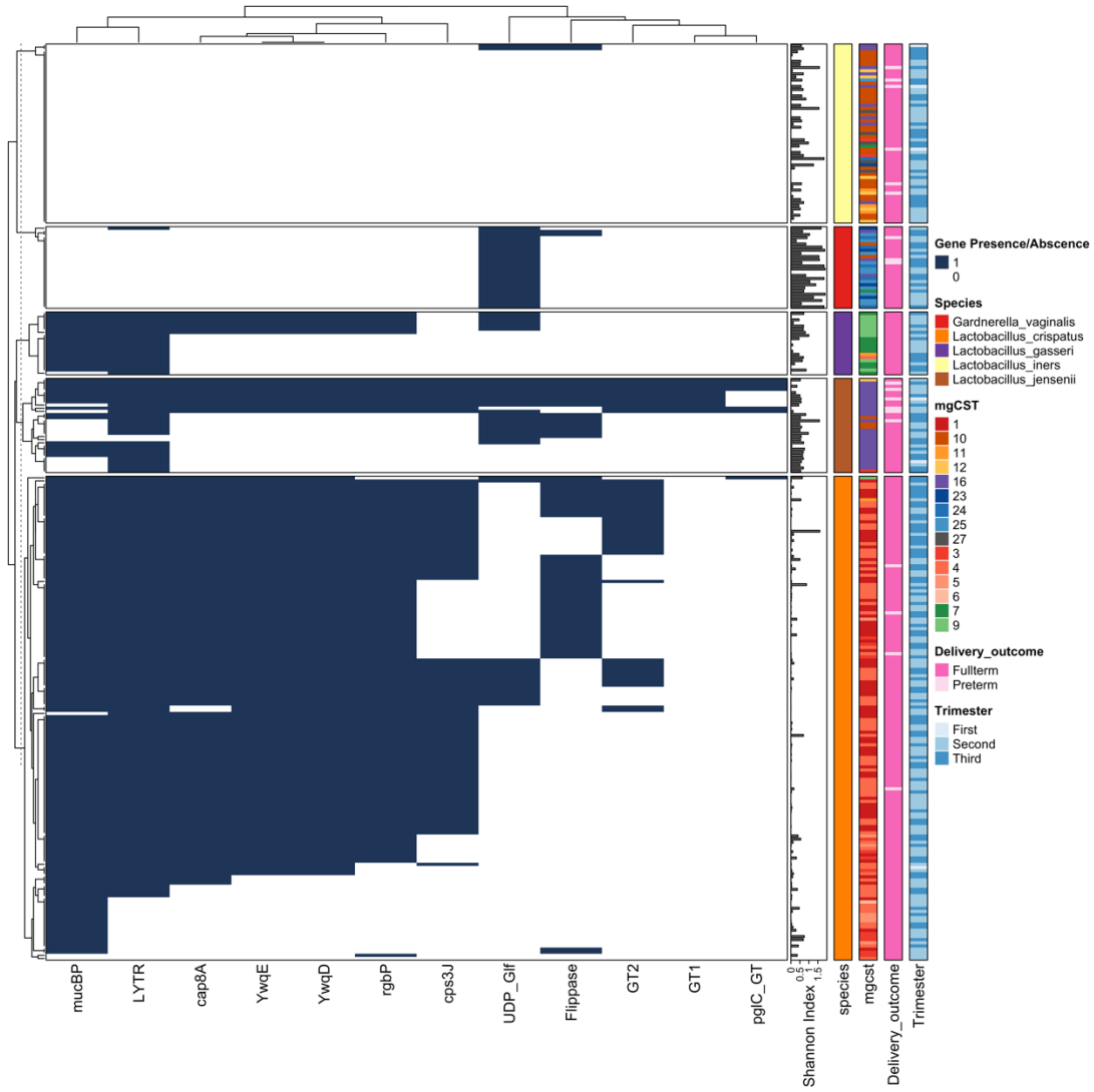
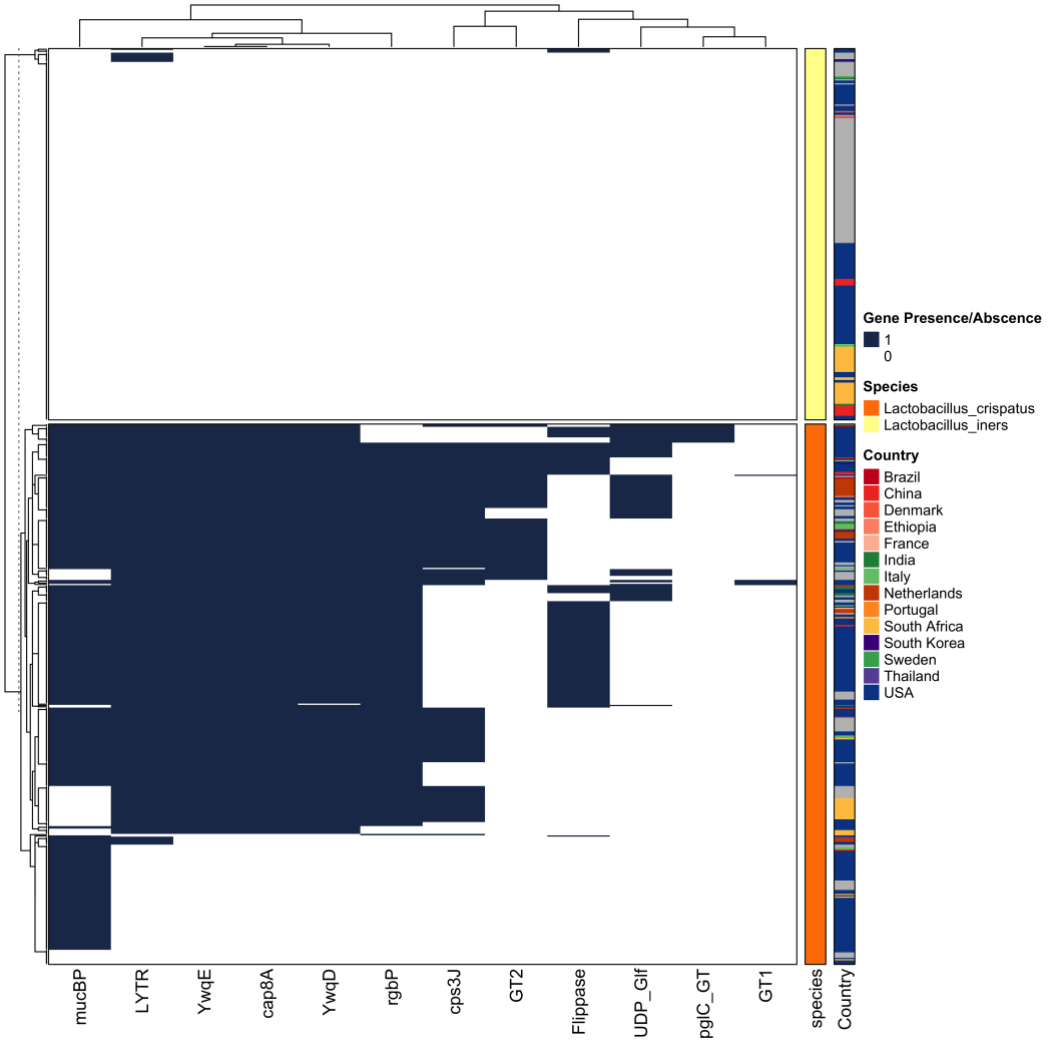


Figure 5.5. Strain level profiling of major vaginal bacteria

a. Stacked barplots faceted by species and colored by mgCSTs representing the number of genes under positive selection in *L. crispatus* and *L. iners* with Pfams annotation. b. Scatterplots faceted by species shows the correlation between Shannon diversity index and the number of cell surface glycan cluster genes detected in the MAGs from all the samples, with Spearman's rho and corresponding p-values indicating the strength and significance of each correlation. c. Heatmap showing the presence and absence of genes from cell surface glycan gene cluster across major vaginal bacteria. The x-axis represents the genes from the gene cluster. For each gene, a blue cell indicates its presence, and a white cell indicates its absence in the corresponding genome. The heatmap was clustered by the similarity of gene content

and separated by species indicated by the color strip on the side. The barplot on the side represents the alpha diversity in the metagenomic sample and color strips represent mgCST, delivery outcome and trimester.



Supplementary Figure 5.3. Presence of cell surface gene cluster in publicly available genomes.

Heatmap showing the presence and absence of genes from cell surface glycan gene cluster in publicly available *L. crispatus* and *L. iners* genomes. The x-axis represents the genes from the gene cluster. For each gene, a blue cell indicates its presence, and a white cell indicates its absence in the corresponding genome. The heatmap was clustered based on the species indicated by the color strip on the side and other color strip represent the isolation country of the corresponding genome.

5.4 Discussion

Understanding the genetic diversity among strains in a microbial community is often very important[29]. In particular, for putatively beneficial bacteria like lactobacilli, there is a requirement to differentiate specific strains that have greater therapeutic potential from those that do not[30–34]. Within the vaginal tract, it has been widely reported that *L. crispatus* is distinctly associated with better health outcomes[35–38]. As such, the general consensus appears to be to promote a microbiome dominated by this species. This however does not consider diversity within the species, or any other species within the vagina. Approaches to study the vaginal microbiome have more often relied on a metataxonomic approach, which does not provide the resolution required to understand functional differences at strain level[39–41]. Within our metagenomic data we revealed the presence of multiple subspecies of bacteria with a diverse gene repertoire in the vaginal microbiome underscoring the importance of studying functional differences even in compositionally similar data. The presence of mgCST-specific genes in *L. crispatus* and *L. iners* related to cell-wall and defence mechanisms suggests different functions of the strains in the vaginal environment. A previous report has shown that *Limosilactobacillus reuteri* strains with mucus adhesins can exert immunoregulatory effects in the gut[42]. Furthermore, mucus-binding proteins from *Lactiplantibacillus plantarum* have shown inhibition towards enterotoxigenic *E*. *coli* cells[43]. The presence of mucin-binding genes in metagenomes dominated by *L. crispatus,* but not in *L. iners,* may have similar functional implications in the vaginal microbiome in terms of bacterial adhesion and pathogen resistance.

The genomic diversity among *L. crispatus* strains has recently been reported and reveals that this results in phenotypic variation, particularly with respect to glycogen metabolism, a major carbon source in the vaginal environment that can determine colonization levels and associated persistence, as well as microbe-host communication ability of the strains[9,26]. Our results show that a gene (*pulA*), encoding a putative glycogen degrading activity, pullulanase, is present in all mgCSTs of both *L. crispatus* and *L. iners,* indicating the importance of this enzyme in glycogen metabolism within the vaginal ecosystem. Separately, previous studies have shown

that D-lactate is more protective against unfavourable vaginal communities than L-lactate and its levels are highest when *L. crispatus* is dominant[44,45]. The presence of D- and L-lactate dehydrogenase-encoding genes in all *L. crispatus* mgCSTs, but the absence of D-lactate dehydrogenase gene in *L. iners,* is in line with previous literature[46]. However, Holm and colleagues have reported that the D-lactate dehydrogenase gene is missing in mgCST 2, one of the subspecies of *L. crispatus,* and that samples dominated by mgCST 2 were compositionally more diverse and had fewer *L. crispatus* strains compared to other mgCSTs[15]. Interestingly, we did not find any samples with mgCST 2 in our dataset, which contains data from samples collected predominantly from healthy women. Altogether, our analysis of taxonomic composition of the vaginal microbiome, coupled with its functional potential, offers insights that are vital for understanding the mechanisms underlying the maintenance of a healthy vaginal environment.

A recent pangenomic survey of major vaginal lactobacilli showed that genes related to adherence functions in *L. crispatus* and *L. iners* are strain specific[47]. In line with this, we report that a high number of cloud genes compared to core genes were found in cell wall biogenesis, carbohydrate and amino acid metabolism functional categories in *L. crispatus* compared to *L. iners*. Furthermore, our pangenome analysis revealed open-pangenomes indicating high intra-species genetic diversity among vaginal *L. crispatus* and *L. iners* strains. This strain-specific genetic diversity, especially in critical functional categories, may underpin differences in colonization efficiency, resilience to environmental stresses, and interactions with the host. Interestingly, we observed a higher positive selection signature among genes related to host-microbe interface in *L. crispatus* suggesting a continuous evolutionary pressure for adaptation. Mainly, we noted cell wall anchor genes with YSIRK signal motifs and mucin-binding genes along with transposons and S-layer associated proteins as major genes under positive selection. A recent study has also shown a positive selection signature in *Lactobacillus* adhesin genes in non-pregnancy cohorts suggesting a ubiquitous selection pressure on these genes[48]. Mucin-binding and cell surface genes are of critical importance as they directly interact with the host epithelial cells and immunity systems[49–51].

Our previous genomics work on *Lactobacillus jensenii* highlighted a cell surface gene cluster which harbours identical cell wall anchor genes and mucin-binding genes along with glycosyltransferases (GTs) and gene machinery required for export of proteins to the cell surface[52]. Zeng and colleagues also showed that MucBP-like domains and a similar cell surface gene cluster in a vaginal *Lactobacillus gasseri* strain plays an important role in adhesion to vaginal epithelium through gene knockout experiments[27]. Interestingly, we observed this gene cluster to be predominantly present in *L. crispatus,* both in our dataset and public genomes, but not in *L. iners* and *Gardnerella vaginalis,* indicating potential species-specific cell surface adaptations. Furthermore, most of the variations within the gene cluster in *L. crispatus* were found in GTases, suggesting a strain-specific microbial cell surface glycome that could be crucial for interactions with vaginal epithelial cells. Moreover, the gene cluster is mainly detected in samples with low alpha diversity indicating the importance of cell surface adaptations in establishing a dominant and stable colonization. Overall, these observations underscore the significance of the cell surface gene cluster and its associated genes in potential host-microbe interactions.

Taken together, our findings highlight the critical role of *Lactobacillus* strain diversity in vaginal microbial ecology, with a specific emphasis on genes predicted to be involved in glycogen metabolism, mucin-binding, and cell surface adaptations. The observed positive selection signature among host-microbe interface genes and the presence of species-specific cell surface gene clusters in *L. crispatus* underline evolutionary adaptations for the vaginal niche. Whilst all these genomic comparisons reveal distinctions with respect to gene content, it still needs to be determined whether a phenotypic difference is also present within strains. Future RNA-seq experiments both *in vitro* and direct-from-swabs would provide a valuable insight into the degree of functional variation within the members of the vaginal microbiota. Nonetheless, our study underscores the importance of strain-level genomic analysis but also opens avenues for novel probiotic developments to support the vaginal microbiome. Further studies with a more diverse dataset are needed to fully understand the implications of strain-level variations in the vaginal microbiome and how this may influence measured health outcomes.

5.5 Methods

5.5.1 Metagenomic community state typing and pan metagenome reconstruction

Raw 354 vaginal metagenomic samples from two Irish pregnancy cohorts, high-risk preterm (n = 87) (PRJEB34536) and MicrobeMom (n = 267) (PRJEB48251) were used for the analysis[53–55]. Inclusion criteria for both cohorts are previously described (52-54). In brief, samples from the high-risk preterm group were pregnant, > 18 years of age, either had previous preterm birth or undergone LLETZ surgery, or no known risk factor (as control group), whilst those from the MicrobeMom group were pregnant, >18 years of age, a BMI between 18.5 and 35 kg/m^2, no gestational diabetes, and singleton pregnancy. The samples were collected from 240 individual participants at different timepoints throughout pregnancy. 94% (225/240) of the participants delivered full-term while 6% (15/240) were preterm. Quality filtering and host contamination removal was performed using Trim Galore (v 0.6.0) and Hostile (v 0.1.0) respectively[56,57]. Alpha diversity of the samples was calculated using VIRGO output with VEGAN r package[58]. The quality filtered reads were annotated against the VIRGO database using built-in scripts[14]. The annotated files were used as input for metagenomic subspecies identification using mgCST classifier[15]. The non-redundant gene profiles of *Lactobacillus crispatus* and *Lactobacillus iners* dominant samples were used for pan-metagenome reconstruction. The Clusters of Orthologous Genes annotation from VIRGO database were used for identifying functional categories.

5.5.2 Reconstruction of metagenome assembled genomes

Metagenome-assembled genomes were reconstructed using quality filtered reads with MetaWRAP (v 1.2.1)[59]. The contigs were assembled using assembly module with metaspades option in the MetaWRAP pipeline. The assembled contigs were binned using binning module with MetaBAT2, Maxbin2 and CONCOCT options. The resulting bins were refined using bin_refinement and reassemble_bins module in the MetaWRAP pipeline. The bins were quality checked with CheckM (v 1.0.18) and only high-quality bins with completeness ≥90 % and contamination <5 % were used for downstream analyses[60]. The taxonomy of the bins was assigned using GTDB-Tk (v

1.5.0)[61]. The number of reads per species needed for MAG recovery was determined based on the number of reads assigned to that species in Kraken2 (v 2.1.1) and Bracken (v 2.2) outputs [62,63].

5.5.3 Pangenome reconstruction and phylogenetic analysis

The pangenomes for *L. crispatus* and *L. iners* were reconstructed using Roary (v3.12) with '-e –n --mafft' options with PROKKA (v 1.14) annotation as input[64,65]. The pangenome categories assigned by Roary were used to define the core genes (99% <= strains <= 100%), Soft core genes (95% <= strains < 99%), Shell genes (15% <= strains < 95%) and Cloud genes (0% <= strains < 15%). Heap's law was used to determine whether the pangenome is open ($\gamma > 0$) or closed ($\gamma < 0$) with Heap_law_for_roary script. The concatenated core genome sequences from Roary output were aligned with MAFFT (v 7.475)[66]. The accessory binary tree was used for tree reconstruction. Maximum likelihood phylogeny was reconstructed using multiple sequence alignment with IQ-TREE2 (v 2.1.3) with 500 bootstraps[67]. Output trees were visualised using iTOL[68]. The distance between core genome phylogeny and accessory phylogeny was computed using Robinson-Foulds (RF) distance with Phangorn R package[69,70]. The COG functional categories for gene families in the pangenomes of *L. crispatus* and *L. iners* were assigned using eggNOG-mapper (v 2.1.7)[71].

5.5.4 Strain profiling of MAGs and publicly available genomes

InStrain (v 1.8.0) was used to identify the genes under positive selection[72]. The high-quality MAGs were dereplicated at 98% identity using dRep (v 3.2.0) and representative MAGs database was built using bowtie2 (v 2.4.4)[73,74]. The quality filtered metagenomic reads were mapped against the MAGs database using bowtie2. The resulting alignment files were used for inStrain gene profiling and further identification of genes under positive selection (dn/ds > 1). Pfam annotations from the eggNOG-mapper output was used to assign the function of the genes.

The publicly available high-quality vaginal *L. crispatus* (n = 406) and *L. iners* (n = 279) genomes were downloaded from PATRIC database on 22nd February 2024[75]. Previously identified cell surface gene cluster genes were used to construct a BLASTp (v 2.8.1) database[52,76]. Protein coding sequences were predicted using Prodigal (v

2.6.3) from high-quality MAGs generated from our dataset and publicly available genomes[77]. The resulting sequences were mapped against the database to identify genes from cell surface gene cluster using BLASTp.

5.6 Data availability

The datasets used in this study can be accessed from ENA using accessions PRJEB34536 (high-risk preterm) and PRJEB48251 (MicrobeMom).

5.7 Acknowledgements

S.R.C. Nori has received funding by Science Foundation Ireland through the SFI Centre for Research Training in Genomics Data Science under Grant number 18/CRT/6214 and supported in part by the EU's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant H2020-MSCA-COFUND-2019-945385. This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under Grant Numbers SFI/12/RC/2273 and 16/SP/3827.

5.8 References

1. Tachedjian, G., Aldunate, M., Bradshaw, C. S. & Cone, R. A. The role of lactic acid production by probiotic *Lactobacillus* species in vaginal health. *Research in Microbiology* 168, 782–792 (2017).

2. Gong, Z., Luna, Y., Yu, P. & Fan, H. Lactobacilli Inactivate Chlamydia trachomatis through Lactic Acid but Not H2O2. *PLoS ONE* 9, e107758 (2014).

3. O'Hanlon, D. E., Moench, T. R. & Cone, R. A. Vaginal pH and Microbicidal Lactic Acid When Lactobacilli Dominate the Microbiota. *PLoS ONE* 8, e80074 (2013).

4. Miko, E. & Barakonyi, A. The Role of Hydrogen-Peroxide (H2O2) Produced by Vaginal Microbiota in Female Reproductive Health. *Antioxidants (Basel)* 12, 1055 (2023).

5. Fuochi, V., Cardile, V., Petronio Petronio, G. & Furneri, P. M. Biological properties and production of bacteriocins-like-inhibitory substances by Lactobacillus sp. strains from human vagina. *Journal of Applied Microbiology* 126, 1541–1550 (2019).

6. Aroutcheva, A. *et al.* Defense factors of vaginal lactobacilli. *American Journal of Obstetrics and Gynecology* 185, 375–379 (2001).

7. Miller, E. A., Beasley, D. E., Dunn, R. R. & Archie, E. A. Lactobacilli Dominance and Vaginal pH: Why is the Human Vaginal Microbiome Unique? *Front. Microbiol.* 7, (2016).

8. Fuochi, V., Li Volti, G. & Furneri, P. M. Commentary: Lactobacilli Dominance and Vaginal pH: Why Is the Human Vaginal Microbiome Unique? *Front. Microbiol.* 8, (2017).

9. Jenkins, D. J. *et al.* Bacterial amylases enable glycogen degradation by the vaginal microbiome. *Nat Microbiol* 8, 1641–1652 (2023).

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

11. Srinivasan, S. & Fredricks, D. N. The Human Vaginal Bacterial Biota and Bacterial Vaginosis. *Interdisciplinary Perspectives on Infectious Diseases* 2008, 1–22 (2008).

12. France, M. T. *et al.* VALENCIA: a nearest centroid classification method for vaginal microbial communities based on composition. *Microbiome* 8, 166 (2020).

13. Blanco-Míguez, A. *et al.* Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. *Nat Biotechnol* 41, 1633–1644 (2023).

14. Ma, B. *et al.* A comprehensive non-redundant gene catalog reveals extensive within-community intraspecies diversity in the human vagina. *Nature Communications* 11, (2020).

15. Holm, J. B. *et al.* Integrating compositional and functional content to describe vaginal microbiomes in health and disease. *Microbiome* 11, 259 (2023).

16. Fettweis, J. M. *et al.* The vaginal microbiome and preterm birth. *Nature Medicine* 25, 1012–1021 (2019).

17. Serrano, M. G. *et al.* Racioethnic diversity in the dynamics of the vaginal microbiome during pregnancy. *Nat Med* 25, 1001–1011 (2019).

18. Liao, J. *et al.* Microdiversity of the vaginal microbiome is associated with preterm birth. *Nat Commun* 14, 4997 (2023).

19. Li, D. *et al.* Vaginal microbiome analysis of healthy women during different periods of gestation. *Bioscience Reports* 40, BSR20201766 (2020).

20. Krog, M. C. *et al.* The healthy female microbiome across body sites: effect of hormonal contraceptives and the menstrual cycle. *Hum Reprod* 37, 1525–1543 (2022).

21. Ahrens, P. *et al.* Changes in the vaginal microbiota following antibiotic treatment for Mycoplasma genitalium, Chlamydia trachomatis and bacterial vaginosis. *PLoS One* 15, e0236036 (2020).

22. O'Neill, I. J. *et al.* Maternal and infant factors that shape neonatal gut colonization by bacteria. *Expert Rev Gastroenterol Hepatol* 14, 651–664 (2020).

23. Saheb Kashaf, S. *et al.* Integrating cultivation and metagenomics for a multi-kingdom view of skin microbiome diversity and functions. *Nat Microbiol* 7, 169–179 (2022).

24. Nayfach, S., Shi, Z. J., Seshadri, R., Pollard, K. S. & Kyrpides, N. C. New insights from uncultivated genomes of the global human gut microbiome. *Nature* 568, 505–510 (2019).

25. Hertzberger, R. *et al.* Genetic Elements Orchestrating Lactobacillus crispatus Glycogen Metabolism in the Vagina. *International Journal of Molecular Sciences* 23, 5590 (2022).

26. Veer, C. V. D. *et al.* Comparative genomics of human Lactobacillus crispatus isolates reveals genes for glycosylation and glycogen degradation: Implications for in vivo dominance of the vaginal microbiota. *Microbiome* 7, 1–14 (2019).

27. Zeng, Z., Zuo, F. & Marcotte, H. Putative Adhesion Factors in Vaginal Lactobacillus gasseri DSM 14869: Functional Characterization. *Applied and Environmental Microbiology* 85, e00800-19 (2019).

28. Argentini, C. *et al.* Evaluation of Modulatory Activities of Lactobacillus crispatus Strains in the Context of the Vaginal Microbiota. *Microbiology Spectrum* 10, e02733-21 (2022).

29. Yan, Y., Nguyen, L. H., Franzosa, E. A. & Huttenhower, C. Strain-level epidemiology of microbial communities and the human microbiome. *Genome Medicine* 12, 71 (2020).

30. Abramov, V. *et al.* Probiotic Properties of Lactobacillus crispatus 2,029: Homeostatic Interaction with Cervicovaginal Epithelial Cells and Antagonistic Activity to Genitourinary Pathogens. *Probiotics & Antimicro. Prot.* 6, 165–176 (2014).

31. Ansari, J. M., Colasacco, C., Emmanouil, E., Kohlhepp, S. & Harriott, O. Strain-level diversity of commercial probiotic isolates of Bacillus, Lactobacillus, and Saccharomyces species illustrated by molecular identification and phenotypic profiling. *PLOS ONE* 14, e0213841 (2019).

32. Azad, M. A. K., Sarker, M., Li, T. & Yin, J. Probiotic Species in the Modulation of Gut Microbiota: An Overview. *BioMed Research International* 2018, (2018).

33. Huang, Z. *et al.* Comparative Genomics and Specific Functional Characteristics Analysis of Lactobacillus acidophilus. *Microorganisms* 9, 1992 (2021).

34. Mu, Q., Tavella, V. J. & Luo, X. M. Role of Lactobacillus reuteri in Human Health and Diseases. *Front Microbiol* 9, 757 (2018).

35. Gudnadottir, U. *et al.* The vaginal microbiome and the risk of preterm birth: a systematic review and network meta-analysis. *Sci Rep* 12, 7926 (2022).

36. Mändar, R. *et al.* Impact of Lactobacillus crispatus-containing oral and vaginal probiotics on vaginal health: a randomised double-blind placebo controlled clinical trial. *Benef Microbes* 14, 143–152 (2023).

37. Armstrong, E. *et al.* Sustained effect of LACTIN-V (Lactobacillus crispatus CTV-05) on genital immunology following standard bacterial vaginosis treatment: results from a randomised, placebo-controlled trial. *Lancet Microbe* 3, e435–e442 (2022).

38. Cohen, C. R. *et al.* Randomized Trial of Lactin-V to Prevent Recurrence of Bacterial Vaginosis. *Obstetrical & Gynecological Survey* 75, 601 (2020).

39. Lebeer, S. *et al.* A citizen-science-enabled catalogue of the vaginal microbiome and associated factors. *Nat Microbiol* 8, 2183–2195 (2023).

40. Wu, M. *et al.* Leveraging 16S rRNA data to uncover vaginal microbial signatures in women with cervical cancer. *Front. Cell. Infect. Microbiol.* 13, (2023).

41. Virtanen, S., Kalliala, I., Nieminen, P. & Salonen, A. Comparative analysis of vaginal microbiota sampling using 16S rRNA gene analysis. *PLOS ONE* 12, e0181477 (2017).

42. Bene, K. P. *et al.* Lactobacillus reuteri Surface Mucus Adhesins Upregulate Inflammatory Responses Through Interactions With Innate C-Type Lectin Receptors. *Front. Microbiol.* 8, (2017).

43. Singh, K. S., Kumar, S., Mohanty, A. K., Grover, S. & Kaushik, J. K. Mechanistic insights into the host-microbe interaction and pathogen exclusion mediated by the Mucus-binding protein of Lactobacillus plantarum. *Sci Rep* 8, 14198 (2018).

44. Nasioudis, D. *et al.* α-Amylase in Vaginal Fluid: Association With Conditions Favorable to Dominance of Lactobacillus. *Reprod. Sci.* 22, 1393–1398 (2015).

45. Witkin, S. S. *et al.* Influence of Vaginal Bacteria and d- and l-Lactic Acid Isomers on Vaginal Extracellular Matrix Metalloproteinase Inducer: Implications for Protection against Upper Genital Tract Infections. *mBio* 4, 10.1128/mbio.00460-13 (2013).

46. France, M. T., Mendes-Soares, H. & Forney, L. J. Genomic Comparisons of Lactobacillus crispatus and Lactobacillus iners Reveal Potential Ecological Drivers of Community Composition in the Vagina. *Applied and Environmental Microbiology* 82, 7063–7073 (2016).

47. Bhattacharya, A., Das, S., Bhattacharjee, M. J., Mukherjee, A. K. & Khan, M. R. Comparative pangenomic analysis of predominant human vaginal lactobacilli strains towards population-specific adaptation: understanding the role in sustaining a balanced and healthy vaginal microenvironment. *BMC Genomics* 24, 1–15 (2023).

48. Wei, X. *et al.* Population genetic analyses of longitudinal vaginal microbiome reveal racioethnic evolutionary dynamics and prevailing positive selection of Lactobacillus adhesins. Preprint at https://doi.org/10.1101/2023.10.11.561685 (2023).

49. Devi, S. M. & Halami, P. M. Diversity and evolutionary aspects of mucin binding (MucBP) domain repeats among Lactobacillus plantarum group strains through comparative genetic analysis. *Syst Appl Microbiol* 40, 237–244 (2017).

50. Dharmani, P., Srivastava, V., Kissoon-Singh, V. & Chadee, K. Role of Intestinal Mucins in Innate Host Defense Mechanisms against Pathogens. *Journal of Innate Immunity* 1, 123–135 (2008).

51. Van Tassell, M. L. & Miller, M. J. Lactobacillus Adhesion to Mucus. *Nutrients* 3, 613–636 (2011).

52. Nori, S. R. C. *et al.* Profiling of vaginal Lactobacillus jensenii isolated from preterm and full-term pregnancies reveals strain-specific factors relating to host interaction. *Microbial Genomics* 9, 001137 (2023).

53. Feehily, C. *et al.* Detailed mapping of Bifidobacterium strain transmission from mother to infant via a dual culture-based and metagenomic approach. *Nat Commun* 14, 3015 (2023).

54. Feehily, C. *et al.* Shotgun sequencing of the vaginal microbiome reveals both a species and functional potential signature of preterm birth. *npj Biofilms and Microbiomes* 6, 1–9 (2020).

55. Moore, R. L. *et al.* Ability of Bifidobacterium breve 702258 to transfer from mother to infant: the MicrobeMom randomized controlled trial. *American Journal of Obstetrics & Gynecology MFM* 5, 100994 (2023).

56. Krueger, F. *et al.* FelixKrueger/TrimGalore: v0.6.10 - add default decompression path. (2023) doi:10.5281/ZENODO.7598955.

57. Constantinides, B., Hunt, M. & Crook, D. W. Hostile: accurate decontamination of microbial host sequences. *Bioinformatics* 39, btad728 (2023).

58. Oksanen, J. *et al.* vegan: Community Ecology Package. (2024).

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

60. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* (2015) doi:10.1101/gr.186072.114.

61. Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* 36, 1925–1927 (2019).

62. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biology* 20, 257 (2019).

63. Lu, J., Breitwieser, F. P., Thielen, P. & Salzberg, S. L. Bracken: estimating species abundance in metagenomics data. *PeerJ Comput. Sci.* 3, e104 (2017).

64. Page, A. J. *et al.* Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* 31, 3691–3693 (2015).

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

66. Katoh, K. & Standley, D. M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol Biol Evol* 30, 772–780 (2013).

67. Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution* 37, 1530–1534 (2020).

68. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Research* 49, W293–W296 (2021).

69. Robinson, D. F. & Foulds, L. R. Comparison of phylogenetic trees. *Mathematical Biosciences* 53, 131–147 (1981).

70. Schliep, K. P. phangorn: phylogenetic analysis in R. *Bioinformatics* 27, 592–593 (2011).

71. Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *bioRxiv* 2021.06.03.446934 (2021) doi:10.1101/2021.06.03.446934.

72. Olm, M. R. *et al.* inStrain profiles population microdiversity from metagenomic data and sensitively detects shared microbial strains. *Nat Biotechnol* 39, 727–736 (2021).

73. Olm, M. R., Brown, C. T., Brooks, B. & Banfield, J. F. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J* 11, 2864–2868 (2017).

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

75. Gillespie, J. J. *et al.* PATRIC: the Comprehensive Bacterial Bioinformatics Resource with a Focus on Human Pathogenic Species ▿. *Infect Immun* 79, 4286–4298 (2011).

76. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* 215, 403–410 (1990).

77. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 11, 119 (2010).

6 Chapter 6

General Discussion

In recent years, the dynamic interplay between maternal and infant microbiomes during the perinatal period has gained recognition as a key determinant of early-life health[1–5]. The work presented in this thesis advances existing research by emphasising the importance of strain-level resolution for understanding these microbial communities. While prior investigations have established that the infant's gut microbiome is partially seeded by maternal bacteria from sources such as the gut, breastmilk, and vaginal tract[1,6–9], much of this earlier research relied on species-level analyses, which can mask important genetic and functional variability within individual taxa. To address this, this thesis employ cutting-edge bioinformatic approaches to reconstruct metagenome-assembled genomes and track minor genomic variations, thereby revealing how even closely related strains may differ profoundly in their capacity for colonisation, their functional gene repertoires, and their potential influences on outcomes such as gestational length.

A strong focus of this thesis is on the mother-infant gut microbiome axis. Chapter 2 illustrates how different binning tools can drastically influence the number and quality of recovered MAGs and, consequently, the observed frequency of mother-infant strain-sharing events. In comparing MetaBAT2 and MetaWRAP, it became apparent that MetaWRAP recovered a larger number of high-quality MAGs, including those with increased completeness and more frequent identification of critical genes like 16S rRNA. A previous report identified the absence of ribosomal genes in a large number of MAGs, attributing this issue to codon usage bias and binning algorithms[10]. This problem can be partially improved by employing multiple binning tools as shown in our work. Furthermore, more refined MAGs from MetaWRAP allowed identification of a higher percentage of shared strains than had been found with the MetaBAT2 MAGs, particularly for genera such as *Bacteroides* and *Phocaeicola*, and it also enabled the detection of more antimicrobial resistance genes. Identifying strain transmissions particularly from these genera are crucial as they can persist even after solid food introduction due to their diverse carbohydrate utilisation gene repertoire and their role in immune maturation[11,12]. Furthermore, previous studies have shown that antibiotics during labour significantly impact the strain transfers[4,13,14], and identification of resistant strains is crucial for understanding the consequences of

antibiotic exposure on infants health[15,16]. These findings highlight the importance of carefully selecting bioinformatic methods and generating data of the highest possible quality to avoid biased or incomplete results when quantifying strain transmission.

Chapter 3 extends this narrative by identifying a potential functional basis for why certain strains, rather than simply the most abundant species, succeed in transmission and colonisation from mother to infant. Specifically, it shows that microbes harbouring genes for human milk oligosaccharide (HMO) utilisation, particularly when these genes encode secreted glycosyl hydrolases with signal peptides, are more frequently shared between mother and infant. This aligns with existing literature that suggests *Bifidobacterium* species can utilise HMOs and are the dominant species in the infant gut microbiome[17,18]. Previous report showed differences in *Bifidobacterium longum* strains between adults and infants[19]. Infant-derived strains have been shown to be capable of utilising human milk oligosaccharides, while adult-derived strains are better adapted to fermenting plant oligosaccharides. This functional advantage suggests that microbial success in establishing in the infant gut is linked not only to their prevalence in the maternal microbiome but also to their ability to utilise the HMO-rich diet of breastfed neonates. Notably, additional strain-sharing events were also identified at later postpartum timepoints.

In Chapters 4 and 5, the focus shifts to the maternal vaginal microbiome, which, although less studied than the gut in terms of strain-level analyses, emerges as a potentially critical factor influencing pregnancy duration and infant health outcomes[20–22]. Chapter 4 examines *Lactobacillus jensenii* isolates from preterm and full-term pregnancies. Even though Lactobacilli are generally considered beneficial for maintaining a low-pH vaginal environment[23–25], detailed phylogenetic and genomic comparisons revealed important differences in their genetic signatures. In particular, preterm-associated strains carried variations in genes linked to cell wall formation, lactate and acetate metabolism, and they also possessed a distinct gene cluster implicated in the synthesis of a cell surface-anchored protein. A previous study by Veer et al. also observed variations in cell surface glycosyltransferases among *Lactobacillus crispatus* strains[26]. Additionally, differences in vaginal cell

adhesion genes were also detected in *Lactobacillus gasseri* strains[27]. These findings raise the possibility that certain *L. jensenii* strains, rather than acting solely as protective commensals, might contribute to or reflect vaginal conditions conducive to preterm birth. This subtle within-species variability illustrates how two strains of the same bacterial species could have markedly different impacts on pregnancy outcomes, highlighting why investigating taxa at a finer resolution is both illuminating and essential.

Chapter 5 continues this exploration of the vaginal microbiome by comparing *Lactobacillus crispatus* and *Lactobacillus iners*, which are often the most dominant Lactobacilli in healthy pregnancies[20,28]. Through comparative metagenomics, it is clear that *L. crispatus* strains exhibit a far broader array of strain-specific genes, many of which are related to mucin-binding, cell wall biogenesis, and glycogen metabolism. A notable discovery is a cell surface glycan gene cluster, present in many *L. crispatus* strains but absent in *L. iners* and *Gardnerella vaginalis*, that may underpin more effective colonisation or stronger host–microbe interactions. Such findings shed light on the different ecological strategies these Lactobacilli employ in the vaginal tract and emphasise the possibility that certain strains with specific gene clusters could confer a protective advantage in terms of preventing bacterial vaginosis or, conversely, be linked to heightened risk if the adaptations somehow undermine vaginal homeostasis. A recent study by Decout *et al.* revealed that the surface layer proteins of *L. crispatus* are involved in anti-inflammatory modulation, while *L. iners* and *G. vaginalis* interact with pro-inflammatory immune receptors[29]. The fact that the vaginal microbiome's strain-level dynamics are comparatively underexplored makes these chapters particularly novel, as they demonstrate how small-scale genetic differences in key *Lactobacillus* species might shape pregnancy length and, by extension, infant outcomes. Future work will include functional characterisation of these strain-level differences through targeted experiments assessing cell wall composition and vaginal epithelial cell models, aimed at elucidating strain-specific immune responses and their implications for vaginal health and pregnancy outcomes.

Taken as a whole, the thesis reveals that the process of maternal-infant microbial transmission is far from uniform or merely dependent on the most numerically dominant species in maternal niches. Instead, a combination of parameters, such as maternal physiology, microbial functional genes, delivery mode, and environmental factors like antibiotic exposure/treatment, determine which strains transfer from mother to infant, and whether they manage to establish and persist. This interplay is especially striking when comparing results across the gut and vaginal ecosystems. While the maternal gut remains the principal reservoir for a large proportion of the infant gut microbiome[3–5], the vaginal microbiome can exert influence on the timing of birth itself potentially affecting which bacterial communities an infant encounters during delivery[3,8] and thus indirectly on the newborn's early microbial composition. Highlighting these parallels and differences adds depth to our understanding of the interconnectedness of these maternal niches.

Nevertheless, several limitations and avenues for future research are evident. In particular, short-read metagenomic sequencing, though widely used, can struggle with highly repetitive genomes or very closely related strains, indicating that integrating long-read or hybrid sequencing may yield even more complete MAGs[30,31]. The work presented here also relied primarily on genomic data, which, while powerful for predicting functional potential, cannot confirm active gene expression or the production of metabolites that might interact with the host. Adding transcriptomic, metabolomic, or culture-based evidence would help bridge the gap between predictive genomics and observable phenotypes. Moreover, direct mechanistic studies could clarify how certain genes, such as those in the novel cell surface clusters identified in *L. jensenii* or *L. crispatus*, influence microbial interactions with host epithelial cells or modulate immune responses that could predispose to preterm delivery.

Overall, this thesis demonstrates that paying close attention to strain-level differences is necessary to capture the inherent complexity of the maternal-infant microbial interface. It highlights how important functional capacities, like carbohydrate metabolism or cell surface protein synthesis, may govern which bacteria succeed in transitioning from mother to infant and perhaps shape crucial

outcomes such as gestational age at birth. From a clinical and public health perspective, these insights open up the prospect of developing targeted probiotic interventions or nutritional strategies that can favourably alter the maternal microbiome, promoting the transfer of beneficial strains and limiting those associated with negative outcomes. At the same time, the findings highlight that "beneficial" species can harbour strains whose effects on pregnancy health are not straightforward. Deepening our grasp of these nuances will require both larger datasets and stronger experimental designs, but the framework established here emphasising robust bioinformatic methodology and functional analysis lays the groundwork for future studies that could ultimately help optimise maternal and infant health by shaping microbial exposures in the perinatal period.

6.1 References

1. Dominguez-Bello, M. G. *et al.* Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. *Proceedings of the National Academy of Sciences* **107**, 11971–11975 (2010).

2. Yatsunenko, T. *et al.* Human gut microbiome viewed across age and geography. *Nature* **486**, 222–227 (2012).

3. Ferretti, P. *et al.* Mother-to-Infant Microbial Transmission from Different Body Sites Shapes the Developing Infant Gut Microbiome. *Cell Host and Microbe* **24**, 133-145.e5 (2018).

4. Feehily, C. *et al.* Detailed mapping of Bifidobacterium strain transmission from mother to infant via a dual culture-based and metagenomic approach. *Nat Commun* **14**, 3015 (2023).

5. Manara, S. *et al.* Maternal and food microbial sources shape the infant microbiome of a rural Ethiopian population. *Current Biology* (2023) doi:10.1016/j.cub.2023.04.011.

6. Benito, D. *et al.* Characterization of Staphylococcus aureus strains isolated from faeces of healthy neonates and potential mother-to-infant microbial transmission through breastfeeding. *FEMS Microbiology Ecology* **91**, fiv007 (2015).

7. Makino, H. *et al.* Multilocus sequence typing of bifidobacterial strains from infant's faeces and human milk: are bifidobacteria being sustainably shared during breastfeeding? *Benef Microbes* **6**, 563–572 (2015).

8. Dos Santos, S. J. *et al.* Maternal vaginal microbiome composition does not affect development of the infant gut microbiome in early life. *Front. Cell. Infect. Microbiol.* **13**, (2023).

9. Mitchell, C. M. *et al.* Delivery Mode Affects Stability of Early Infant Gut Microbiota. *Cell Reports Medicine* **1**, 100156 (2020).

10. Mise, K. & Iwasaki, W. Unexpected absence of ribosomal protein genes from metagenome-assembled genomes. *ISME Communications* **2**, 118 (2022).

11. Browne, H. P., Shao, Y. & Lawley, T. D. Mother–infant transmission of human microbiota. *Current Opinion in Microbiology* **69**, 102173 (2022).

12. Mazmanian, S. K., Liu, C. H., Tzianabos, A. O. & Kasper, D. L. An Immunomodulatory Molecule of Symbiotic Bacteria Directs Maturation of the Host Immune System. *Cell* **122**, 107–118 (2005).

13. Lemas, D. J. *et al.* Exploring the contribution of maternal antibiotics and breastfeeding to development of the infant microbiome and pediatric obesity. *Seminars in Fetal and Neonatal Medicine* **21**, 406–409 (2016).

14. Yassour, M. *et al.* Natural history of the infant gut microbiome and impact of antibiotic treatment on bacterial strain diversity and stability. *Science Translational Medicine* **8**, 343ra81-343ra81 (2016).

15. Patangia, D. V., Ryan, C. A., Dempsey, E., Stanton, C. & Ross, R. P. Vertical transfer of antibiotics and antibiotic resistant strains across the mother/baby axis. *Trends Microbiol* **30**, 47–56 (2022).

16. Azad, M. *et al.* Impact of maternal intrapartum antibiotics, method of birth and breastfeeding on gut microbiota during the first year of life: a prospective cohort study. *BJOG: An International Journal of Obstetrics & Gynaecology* **123**, 983–993 (2016).

17. Kiely, L. J., Busca, K., Lane, J. A., van Sinderen, D. & Hickey, R. M. Molecular strategies for the utilisation of human milk oligosaccharides by infant gut-associated bacteria. *FEMS Microbiology Reviews* **47**, fuad056 (2023).

18. Taft, D. H. *et al.* Bifidobacterial Dominance of the Gut in Early Life and Acquisition of Antimicrobial Resistance. *mSphere* **3**, 10.1128/msphere.00441-18 (2018).

19. Sela, D. A. & Mills, D. A. Nursing our microbiota: molecular linkages between bifidobacteria and milk oligosaccharides. *Trends in Microbiology* **18**, 298–307 (2010).

20. Baud, A. *et al.* Microbial diversity in the vaginal microbiota and its link to pregnancy outcomes. *Sci Rep* **13**, 9061 (2023).

21. Gudnadottir, U. *et al.* The vaginal microbiome and the risk of preterm birth: a systematic review and network meta-analysis. *Sci Rep* **12**, 7926 (2022).

22. Zhou, L. *et al.* Effects of vaginal microbiota transfer on the neurodevelopment and microbiome of cesarean-born infants: A blinded randomized controlled trial. *Cell Host & Microbe* **31**, 1232-1247.e5 (2023).

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

24. Lebeer, S. *et al.* A citizen-science-enabled catalogue of the vaginal microbiome and associated factors. *Nat Microbiol* **8**, 2183–2195 (2023).

25. Amabebe, E. & Anumba, D. O. C. The vaginal microenvironment: The physiologic role of Lactobacilli. *Frontiers in Medicine* **5**, 181 (2018).

26. Veer, C. V. D. *et al.* Comparative genomics of human Lactobacillus crispatus isolates reveals genes for glycosylation and glycogen degradation: Implications for in vivo dominance of the vaginal microbiota. *Microbiome* **7**, 1–14 (2019).

27. Zeng, Z., Zuo, F. & Marcotte, H. Putative Adhesion Factors in Vaginal Lactobacillus gasseri DSM 14869: Functional Characterization. *Applied and Environmental Microbiology* **85**, e00800-19 (2019).

28. MacIntyre, D. A. *et al.* The vaginal microbiome during pregnancy and the postpartum period in a European population. *Sci Rep* **5**, 8988 (2015).

29. Decout, A. *et al.* Lactobacillus crispatus S-layer proteins modulate innate immune response and inflammation in the lower female reproductive tract. *Nat Commun* **15**, 10879 (2024).

30. Orellana, L. H., Krüger, K., Sidhu, C. & Amann, R. Comparing genomes recovered from time-series metagenomes using long- and short-read sequencing technologies. *Microbiome* **11**, 105 (2023).

31. Eisenhofer, R. *et al.* A comparison of short-read, HiFi long-read, and hybrid strategies for genome-resolved metagenomics. *Microbiology Spectrum* **12**, e03590-23 (2024).