UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

Exploring the Gut Microbiome: Metaproteome Analysis of HIV-Exposed Uninfected Infants

Tara Kathleen Miller
(MLLTAR003)

MSc (Med) Chemical Biology

Supervisor: Jonathan Blackburn
Co-investigators: Tariq Ganief; Clive Gray; Heather Jaspan; Anna Happel

Institute of Infectious Disease and Molecular Medicine (IDM)
Division of Chemical and Systems Biology
Department of Integrative Biomedical Sciences
Faculty of Health Sciences
University of Cape Town

This work is based on research supported by the National Research Foundation of South Africa.
(143487)

MLLTAR003@myuct.ac.za

DECLARATION

I, **Tara Kathleen Miller**, hereby declare that the work on which this dissertation/thesis is based is my original work (except where acknowledgements indicate otherwise) and that neither the whole work nor any part of it has been, is being, or is to be submitted for another degree in this or any other university.

I empower the university to reproduce for the purpose of research either the whole or any portion of the contents in any manner whatsoever.

Signature: Signed by candidate

Date: 21/07/2024

Plagiarism Declaration

"This thesis/dissertation has been submitted to the Turnitin module (or equivalent similarity and originality checking software) and I confirm that my supervisor has seen my report and any concerns revealed by such have been resolved with my supervisor."

Name: Tara Kathleen Miller

Student number: MLLTAR003

Signature: Signed by candidate

Date: 21/07/2024

Abstract

Background: HAART has decreased mother-to-child transmission, but it has introduced a new generation of HIV infected and unexposed infants (HEU) with unknown health implications. These infants have been observed to be more vulnerable to disease and have weaker immune responses. The gut microbiome plays a pivotal role in human health by aiding in metabolism, nutrient availability and immune development and function. The mucosal immune system and microbiome work in synergy to maintain host immunity and balance for optimal host health. Gut dysbiosis is linked to diseased states, with gut dysbiosis being well documented amongst HIV infected individuals. HEU infants' altered immunity and health could be linked to gut dysbiosis. Looking into the metaproteome of the infant gut to find any perturbations from healthy counterparts could give insight into their weakened states.

Aim: ***To analyse the metaproteome of the infant gut to identify any alterations in a HIV exposed infant's gut, through analysis of the secreted proteins of the stool microbiome, accounting for their poor health.***

Design: Analysis of infant stool samples for the 1st week of life from a cohort comprised of HIV infected mothers and HIV uninfected mothers. Using data-independent acquisition mass spectrometry (SWATH) to identify human and microbiome proteins in samples using Evosep40-SPD and TTOF 6600. In-silico library in DIA-NN used for protein identification, followed by data clean up and pre-processing in R studio and excel. Analysis of the proteins carried out in R studio and Metaboanalyst 5.0.

Results: Metaproteomic analysis method used allowed for deep metaproteome profiling and gives a greater understanding into the microbiome functioning than genomic analysis. The protein level Random Forest regression model analysis revealed HEU and HU clustering, in which both human and bacterial proteins are contributing. Infant proteins revealed alteration in MUC2 mucin gut barrier and inflammation in HEU infants, differences in immune proteins from breastmilk proteins, and bacterial proteins revealed *Ruminococcus gnavus* as an HEU associated species.

Conclusion: The altered gut barrier paired with gut inflammation is strongly correlated with bacterial dysbiosis seen in HEU guts. The identified differential species, *Ruminococcus gnavus,* has been implicated in various diseased states in the gut (IBD and Crohn's), involved in autoimmune diseases (lupus) and neurological diseases (Parkinson's). This reveals the focus for future research to identify potential targets for therapeutic treatment to improve the health of these infants.

Acknowledgements

I am deeply grateful to Prof Jonathan Blackburn for providing unwavering support and the opportunity to pursue my master's in the field of proteomics. Attending my first conference in 2023, HUPO, and the multitude of opportunities stemming from my work in the Blackburn lab have been invaluable experiences.

Tariq Ganief deserves heartfelt appreciation for his consistent guidance and support throughout this project. His unwavering belief in me, continuous challenge for improvement, and willingness to share his critical thinking process have been instrumental. I'm grateful for his mentorship and for providing a sounding board for my ideas.

To my parents, your unwavering support, love, and guidance have been the foundation for my education and journey through my master's program. Your encouragement to pursue my dreams and ideas has been a source of strength.

Evan, your unwavering support, encouragement, and love have been indispensable throughout this project. Having you by my side made this journey possible.

I extend my heartfelt thanks to the members of the Blackburn lab for their guidance and support during lab meetings, coffee breaks, and office discussions. Special gratitude to Michelle and Nina for their invaluable assistance with the challenging analyses.

To my friends—Kristen, Kellyn, Paige, Bex, and others—who provided unwavering support, lending me their ears, and sharing wine and treats to keep my spirits up during challenging times.

This research was made possible through funding from the NRF for my master's degree, as well as support from the NIH and co-investigators for their contributions to the InFANT project.

A sincere thank you to all the participants of the InFANT study for providing samples and contributing to the efforts aimed at improving the lives of others.

I acknowledge that any opinions, findings, conclusions, or recommendations expressed in publications resulting from NRF-supported research are solely those of the author, and the NRF bears no liability in this regard.

Table of Contents

List of figures

Abbreviations list

HUMAN IMMUNODEFICIENCY VIRUS	HIV
MOTHER TO CHILD TRANSMISSION	MTCT
ANTIRETROVIRAL THERAPY	ARV
ACQUIRED IMMUNODEFICIENCY SYNDROME	AIDS
HIGHLY ACTIVE ANTIRETROVIRAL TREATMENT	HAART
NUCLEOSIDE REVERSE TRANSCRIPTASE INHIBITORS	NRTIs
NONNUCLEOSIDE REVERSE TRANSCRIPTION INHIBITORS	NNRTIs
RITONAVIR	R
PROTEASE INHIBITOR	PI
TENOFOVIR	TDF
EMTRICITABINE	FTC
EFAVIRENZ	EFV
WORLD HEALTH ORGANISATION	WHO
HIV EXPOSED UNINFECTED	HEU
HIV UNEXPOSED UNINFECTED	HU
EXTENDED PROGRAMME OF IMMUNIZATION	EPI
NATURAL KILLER	NK
SOUTH AFRICA	SA
HUMAN MILK OLIGOSACCHARIDES	HMO
ANTI-MICROBIAL PEPTIDES	AMPs
MUCOSAL-ASSOCIATED LYMPHOID TISSUES	MALTs
IMMUNOGLOBULIN	Ig
INTRAEPITHELIAL LYMPHOCYTES	IELs
INTERFERON	IFN
INTERLEUKIN	IL
SEQUENTIAL WINDOW ACQUISITION OF ALL THEORETICAL MASS SPECTRA	SWATH
LIQUID CHROMATOGRAPHY	LC
MASS SPECTROMETRY	MS
DATA DEPENDENT ACQUISITION	DDA
DATA INDEPENDENT ACQUISITION	DIA
SHORT-CHAIN FATTY ACIDS	SCFAs
TRIPLE TIME-OF-FLIGHT	TTOF
RIBONUCLEIC ACID	RNA
RIBOSOMAL RIBONUCLEIC ACID	rRNA
POLYMERAISE CHAIN REACTION	PCR
KYOTO ENCYLEOPEDIA OF GENES AND GENOMES	KEGG
CLUSTER OF ORTHOLOGOUS GROUPS OF PROTEINS	COG
INNATE FACTORS ASSOCIATED WITH NURSING TRANSMISSION	InFANT
PHOSPHATE-BUFFERED SALINE	PBS
TRIS(HYDROXYMETHYL)AMINOMETHANE	Tris
AMMONIUM BICARBONATE	ABC
ACETONITRILE	ACN
MATCH BETWEEN RUNS	MBR
LIMIT OF DETECTION	LOD
VARIABLE INFLUENCE OF PROJECTION	VIP
DITHIOTHREITOL	DTT
FORMIC ACID	FA
OUT-OF-THE-BAG	OOB
NITRIC OXIDE	NO
INDUCIBLE NITRIC OXIDE SYNTHASE	iNOS
TUMOUR NECROSIS FACTOR	TNF
NUCLEAR FACTOR KAPPA-LIGHT-CHAIN-ENHANCER	NFκB
INFLAMMATORY BOWEL DISEASE	IBD

REACTIVE OXYGEN SPECIES	ROS
REACTIVE NITROGEN SPECIES	RNS
MAMMALIAN TARGET OF RAPAMYCIN	mTOR
MANNOSE-BINDING LECTIN	MBL

Tables list

Introduction

Chapter 1: Literature review

HIV & HAART

HIV-1 first appeared in colonial west central Africa and spread unnoticed for 50 to 70 years, with its earliest common ancestor dating back to around 1910 to 1930. The virus diversified early on around Kinshasa, and by 1959 and 1960, different subtypes had already emerged there [3]. HIV-1 and HIV-2 can infect humans because they can counteract the body's natural defences. Factors like large-scale injection campaigns, social instability, rapid urban growth, and a rise in sexually transmitted diseases in west central Africa likely helped the virus spread early on [3]. HIV progresses to AIDS when the immune system is severely weakened, leading to infections and cancers that a healthy immune system would normally fight off. HIV-1 mainly spreads through sexual contact, needle sharing, and from mother to child, with 80% of adults getting it through sexual exposure [3]. The main group of HIV-1 has infected over 60 million people and caused more than 25 million deaths since it was first identified. HIV-2 spreads in similar ways. Most instances of simian immunodeficiency viruses (SIVs) jumping from animals to humans didn't spread widely, except for one event involving chimpanzees in Cameroon, which led to HIV-1 group M, the main cause of the AIDS pandemic [3]. Understanding the genetic changes that allowed these viruses to jump from animals to humans helps us learn more about these risks. The introduction of antiretroviral (ARV) therapy has dramatically reduced AIDS-related deaths and turned HIV/AIDS from a fatal disease into a manageable chronic condition [3]. ARVs have improved the quality of life for many people, reduced HIV transmission rates, and allowed individuals to live longer and healthier lives. However, challenges like access to treatment, drug resistance, and side effects still exist, making ongoing research and development in HIV/AIDS treatment essential [3].

In 2021, globally 38.4 million individuals were living with Human Immunodeficiency virus (HIV) of which 28.7 million had access to antiretroviral therapy [4]. In Sub-Saharan Africa, 63% of all new HIV infections were among women and children, while Southern Africa had 90% of pregnant women accessing treatment in aid to prevent mother to child transmission (MTCT) for the best infant outcomes [4].

However, HIV remains a huge global burden, especially in Sub-Saharan Africa where the number of infections account for two thirds of infections worldwide [5]. The African continent still has large resource constraint areas where access to health care and sufficient nutrition is limited, therefore contributing to further infections and disease progression [5].

The virus invades CD4+ T helper cells, pivotal immune cells in both innate and humoral systems, resulting in the infected individuals being immunocompromised [6, 7]. Additionally, HIV virions infect and manipulate these cells, facilitating their replication for the purpose of virion reproduction [7].

Furthermore, HIV also has the ability to target other immune cells, namely macrophages, immature dendritic cells and some resting T cell subtypes, exacerbating immune vulnerability [7]. Due to HIV's ability to infect and manipulate major players in the human immune system, HIV is able to evade destruction and undergoes rapid replication [7].

In conjunction with immune evasion, HIV consists of retroviral genes as well as regulatory genes, making it a retrovirus with a complex life cycle [7]. This complex life cycle of HIV consists of early and late stage replication, followed by programmed cell death of both the infected immune cells and the bystanders, leaving infected individuals severely immunocompromised [7].

The major steps involved in HIV replication are viral attachment, binding and fusion, reverse transcriptase, uncoating and nuclear entry, integration, transcription, translation, and assembly, and budding and maturation. These essential life cycle steps have been targeted for HIV treatment, known as antiretroviral therapy (ARVs). These ARVs reduce the HIV viral load by inhibiting replication and preventing immune cell destruction, ultimately improving infected individuals' immune systems [7-9].

Optimal results from ARV treatment to recover the immune system and maintain viral loads of individuals below detection levels, require strict adherence to an ARV regime [5, 8-10]. Non-compliance to the treatment regimen results in drug resistance, as well as disease progression into acquired immunodeficiency syndrome (AIDS) which has a poor prognosis [5]. Effective management entails close monitoring of patients' CD4+ count, regimen maintenance and monitoring of regimen to allow the patients' immune systems to recover, aiding in the disease prognosis [5, 8, 10].

The development of ARVs in the mid-1990s transformed HIV diagnosis, converting it from a fatal ailment to a chronic condition [8, 10]. This has been achieved by combination therapy known as HAART (highly active antiretroviral treatment), which suppresses HIV replication and HIV-RNA viral load to levels below detection allowing the patients' immune systems, specifically the CD4+ cells, to recover [8, 10, 11]. HAART typically

combines two or more molecular targets by using three ARV's from various categories for the best possible outcomes [8, 11]. Although HAART cannot eliminate the chronic HIV infection, it allows for increased life expectancy and quality in infected individuals [8-10, 12].

There are six categories of ARV drugs targeting the HIV life cycle, of which twenty-four different drugs have been FDA approved. These include nucleoside reverse transcriptase inhibitors (NRTIs); nonnucleoside reverse transcription inhibitors (NNRTIs); protease inhibitors (PIs) with added ritonavir (r) to increase potency of inhibitor; fusion inhibitors; co-receptor CCR5 antagonists, and integrase inhibitors [8, 10, 11]. HAART is mostly comprised of NRTIs plus either NNRTI or PI/r, with the best combination for pregnant women being Tenofovir (TDF); Emtricitabine (FTC) and Efavirenz (EFV), as recommended by World Health Organisation (WHO) [6, 10].

In HIV burdened countries such as South Africa, pregnant women are advised to undergo HIV testing to prevent MTCT, even throughout the pregnancy, to promptly initiate ARV treatment if needed [11, 13]. To diminish risk of transfer and enhance maternal health, use of HAART through pregnancy, labour, delivery and breastfeeding is highly recommended [14].

Although formula feeding is assumed to be the best option for HIV mothers to reduce MTCT, six months of exclusive breastfeeding is recommended in resource limited settings for the best outcome for infant health [14]. This initiative is known as Option B+, which encourages life-long HAART treatment for pregnant and breastfeeding mothers, to both decrease MTCT and improve disease progression in the mothers' and infants' health [6, 12].

With the combination of maternal HAART treatment and infant prophylaxis (daily nevirapine), MTCT during breastfeeding is reduced significantly and provides optimal protection for both infant and mother [14]. Nevirapine is a cost-effective prophylaxis which would benefit infants through direct administration, specifically in low resource areas where HIV and other opportunistic diseases are rife [14]. Further research into pregnancy and infant care for HIV+ mothers and HIV exposed uninfected infants (HEU) is critically required for the best possible outcomes for both mother and child [12].

Nonetheless, a limitation lies in the fact that the use of ARVs during pregnancy has been observed to impact birth outcomes and give rise to metabolic complications in infants. Furthermore, these medications are associated with reduced birth weights among neonates [6]. Although HAART has been assessed in numerous studies to have a potentially increase the risk of preterm birth (which would have a big impact in resource limited settings due to health care for preterm infants), along with previously mentioned side effects, the benefits of HAART still outweigh the side effects [12]. Women living with HIV also have increased risk of co-infections in resource-limited settings, such as TB and malaria, which could also contribute to adverse pregnancy outcomes [12].

In sub-Saharan Africa millions of women give birth annually, contributing to the higher incidence of HIV transmission to children [4, 13]. The three most prominent modes of HIV transmission for infants are

antepartum, intrapartum and postpartum periods, with each one being mitigated effectively through HAART treatment, as high viral loads have been associated with increased MTCT [6]. The best combination of preventative strategies to decrease MTCT is HAART; elective caesarean birth and exclusive formula feeding [13]. However, in resource limited settings Option B+ (exclusive breastfeeding on HAART) is recommended by WHO for 4-6 months for the best infant outcome, alongside strict adherence to ARV treatment and regular clinic check-ups [6, 13]. Neonates born to mothers living with HIV are exposed *in utero* to ARVs; antibiotics, as well as anti-tuberculosis therapy, which could all influence the growth of the neonate and the outcome of the pregnancy [6]. Sadly, maternal HIV infection has been associated with stillbirths, growth retardation, infant mortality, low birth weights, and spontaneous abortion [13]. Therefore, a strict regimen of HAART, can decrease the mother's viral load to below detection, allowing the immune system to recover and prevent transfer to the infant [8, 10, 11, 13]. However to note, one ARV (Efavirenz) has been deemed unsafe for use in early pregnancy as it results in neural tube defects in the infant [13].

One major contributor to fetal growth is placenta health, which depends on placental sufficiency and the ability for nutrients to cross the placenta to the fetus [6]. It is thought that HIV infection may have negative impacts on the placental size- causing changes to the physiology and health of the placenta which results in deficiencies and poor health of the fetus [6]. This could potentially also account for the health of HEU infants in addition to drug exposure i.e., ARVs and antibiotics.

Another factor affecting HEU infants is their vulnerable immune systems, however, the source of this issue remains unclear. However, it may, be attributed to either or both the *in-utero* HIV exposure or ARV exposure [6]. When comparing HEU infants with HIV unexposed uninfected (HU) infants, there are reduced levels of specific maternal antibodies found in HEU infants [6]. In addition to the decreased passive immunity, the HIV exposure could also be directly affecting the T cell differentiation and homeostasis, and may even extend to the thymic functioning into early childhood [6]. Further research into the health of the HEU infant population is needed to provide the best possible outcomes for HEU infants due to HIV exposure, ARV exposure and altered immune system.

The extended programme of immunization (EPI) has globally demonstrated its efficacy in preventing mortality from preventable childhood diseases [15]. However, the immune system of HEU infants, due to its immature alterations, compromises vaccine responses, resulting in notably lower antibody concentrations towards vaccines. This vulnerability places HEU infants at risk for preventable, potentially fatal diseases[15, 16]. One immune factor that is being comprised by the HIV exposure is the infants' lymphoproliferative responses, as well as there being a reduction in neutrophil functioning. This leaves such infants more susceptible to severe infections [15, 17]. Another contributing factor to the comprised immune system of HEU infants is the use of ARVs [18]. These are believed to both lower neutrophil levels (around 2-fold lower) due to myeloid cell development being suppressed, as well as hemopoiesis being negatively affected by ARV

use, specifically Zidovudine [18]. Furthermore, there is a growing population of HEU infants with immature immune systems, who are susceptible to opportunistic diseases [19]. This susceptibility increases the mortality rate for HEU infants due to the disproportionate burden of disease [19]. In addition, HEU infants' growth is impaired with links to either maternal HIV status or to the targeting of Natural Killer (NK) cells, causing the altered immune system of HEU infants [19]. These NK cells are activated early on *in utero* due to a proinflammatory environment, which is caused by maternal HIV infection and results in decreased cytolytic activity. Together with the changes to the innate immune system, the adaptive immune system could also be altered, resulting in HEU infants being left susceptible to disease [19].

Due to the delayed immune system maturation of HEU infants, BCG vaccine responses have resulted in decreased frequency of proliferating T cells [20]. However, other studies reported by *Manssor et al.*, investigating the difference in response for BCG vaccination between HEU and HUU infants observed no differences between the groups [21]. Yet, HIV infected infants had an impaired BCG vaccination response [21]. One suggestion to obtain better responses to BCG vaccination was to delay the vaccine by up to 8 weeks to allow for robust T-cell responses [22]. This implementation could provide a modified strategy for HEU infants' vaccination for optimal protection against infectious diseases [22].

HIV exposed uninfected infants

Annually, one million women infected with HIV deliver infants which are either HIV infected, or HIV exposed [23]. In South Africa (SA), 80% of women HIV-infected are on HAART, with 30% of pregnant women in SA being HIV positive [24]. Due to education and research around HIV, ARV regimens have been implemented into most countries effectively to reduce transmission and infection severity. As a result of the ramped-up HAART treatments during pregnancy and breastfeeding, MTCT have successfully been reduced to 4% in countries like SA, resulting in fewer HIV infected children [25]. In resource rich areas, the risk of MTCT is currently 0% due to health care access, but resource limited areas still have MTCT risk accompanied by the higher risk of opportunistic diseases due to the poor living conditions [26]. Nonetheless, the HIV pandemic has given rise to a distinct group of individuals: HIV-exposed but uninfected infants (HEU). These infants manage to evade HIV infection, yet unfortunately, they experience a heightened morbidity and mortality rate in comparison to their unaffected counterparts [23].

Studies conducted in Botswana have highlighted the growth challenges faced by HEU infants. These infants tend to exhibit growth stunting and typically have lower birth weights in comparison to HU infants [27, 28]. Additionally, HEU infants often face elevated hospital admission rates due to conditions like fever, diarrhea, and vomiting [25, 29-31]. Notably, despite the risk of HIV transmission, breastfeeding emerges as a protective factor for infant health. Exclusively breastfed infants exhibit improved outcomes, emphasising the positive impact of breastfeeding on their well-being [23, 25, 31].

The HEU infants have been shown to have a 3-fold higher rate of death, with respiratory tract infections being the main cause, as well as a 7-fold higher mortality rate if not breastfed [29, 32]. These adverse effects are

not limited to resource limited areas only as shown by *C. Adler* who found HEU infants in a study conducted in Belgium, had a 4-fold higher rate of invasive pneumonia than their healthy counterparts, despite their accessibility to necessary resources [33].

Maternal HIV infection has been demonstrated to make HEU infants more vulnerable to infectious diseases, including conditions like pneumonia and tuberculosis, even in regions that are not heavily burdened by these diseases [33]. Notably, African cohorts, such as those in South Africa, where prevalent conditions like TB and malaria contribute to a challenging health landscape, experience heightened rates of infectious disease transmission among HEU infants, consequently leading to an rise in infant mortality [18, 34].

In addition, immune responses by HEU infants are seemingly lower than HUU infants. This disparity could be due to factors like possible placental dysfunction, *in utero* HIV exposure, or attenuated maternal antibodies resulting from ARV use [35]. Research has pointed out that HEU infants exhibit decreased cytokine production by mononuclear cells, alongside a difference in B cell characteristics observed as early as weeks of age [36]. This deviation in innate immunity might indeed contribute to the vulnerability of HEU infants to disease and evaluated mortality rates [36].

As previously mentioned, the Option B+ strategy, involving lifelong HAART, has successfully reduced mother-to-child transmission of HIV, yet infants born to HIV-positive mothers still face challenges [23, 25, 27]. A recommended ARV regimen, TDF-FTC-EFV, shows positive outcomes, especially when initiated before conception [23, 37]. Alternate regimens have led to adverse effects like low birth weights and neonatal mortality, emphasizing the importance of proper regimen selection. ART use has been linked to preterm delivery and lower birth weights in HEU infants, while maternal malnutrition in resource-limited areas impacts breast milk composition and infant health [23, 37]. Another theory suggests placental health disruption from ART use might contribute to adverse outcomes in HEU infants [37]. Despite the benefits of ARV use in HIV infections to reduce MTCT, decreased cognitive and motor functioning in HEU infants has been observed. A study investigating the brain structures of infants, the HEU infants have found to have less grey matter when maternal CD4+ counts are low, as well as decreased brain volume, which is associated with higher maternal immunosuppression [38]. Another study assessing ARV presence in maternal blood showed the presence of ARVs at 32 weeks of pregnancy in maternal blood resulted in decreased cognitive functioning, for examples learning disabilities and memory issues. Therefore, it is believed HEU infants have delayed neurodevelopment compared to their healthy counterparts either due to HIV exposure or ARV exposure. However, other factors in these resource-limited settings such as malnutrition, maternal health, partner violence, and stress could also attribute to these neurodevelopment delays [24, 27, 38-40].

One of the significant contributors to brain development in humans is the gut microbiome, due to the importance of the gut-brain axis. The colonization of the infant gut is a highly orchestrated, complex process

which results in the microbiome which has a major role in digestion and immune capabilities of the infant [41]. This intricate microbial community plays an integral part in human health as these resident microbes support bodily functions, nutrient absorption, and is a major contributing factor in immune system functioning [42, 43]. The gut microbiota is thought to play a role in major pathways including neural, immunological and the development of the brain [44]. Inadequate colonization during infancy could result in a dysbiosis, leading to adverse effects on the infant's immune functioning, rending them more susceptible to pathogens and impeding neural development [42] [40, 45].

Microbiome dysbiosis has been observed in HIV-infected individuals, possibly from ARV treatment and the direct impact of viral particles on the microbiome [44, 45]. In HIV-positive mothers, this altered microbiome has the potential to be transferred to the infant, resulting in severe side effects on the infants' health, predominantly compromising their immune function [45].

After a metagenomics analysis by *P. Ferretti*, the maternal gut microbiome showed the highest transmission of strains to the infant [42]. In addition to the gut microbiome, the breastmilk microbes are a main source of microbes for the infants, namely *Lactobacillus*; *Bifidobacterium longum* and *Bifidobacterium bifidum* [44] [42]. These two maternal sources (gut and breastmilk) are crucial for developing the infant microbiome [40, 42]. The microbes that colonize the HIV immunosuppressed human body are different to healthy counterparts, with a higher diversity when there is low *Lactobacillus* colonization [44]. *Lactobacillus* is beneficial due to its protective properties against infections, in addition to its potential as a prebiotic [43, 44].

Breast milk bacteria play a pivotal role in initializing the gut microbiome development of infants, a crucial factor in their overall growth and immune system maturation. These bacteria colonize the infant's gut, while human milk oligosaccharides (HMOs) serve as prebiotics for these microorganisms. An alteration in HMOs has been observed between HIV-infected and uninfected women, characterized by elevated sialyllactose levels in breast milk from HIV-infected individuals, likely arising from HIV's impact on the microbiome [45-47].

HMOs hold not only prebiotic roles but also possess immunomodulatory functions, with highly frucosylated HMOs providing robust protection for HEU infants [48]. This composition of HMOs in breast milk may significantly contribute to the growth and developmental advantages that breast milk offers to infants [48].

Furthermore, breastfeeding proves to be the optimal form of nutrition for infants in resource-limited areas, regardless of maternal HIV status [25, 31, 49, 50]. It provides infants with the necessary nutrients for growth and protection against active pathogens [25, 26, 28, 31, 50]. Notably, non-breastfed infants have a heightened risk of mortality due to an increased incidence of malnutrition, diarrhoea, and vomiting [31, 49]. For the most robust defence against disease and malnutrition, exclusive breastfeeding for 5-6 months is recommended. This practice is particularly encouraged for HIV-positive women undergoing ARV treatment, yielding the best outcomes for infants [31, 49]. The heightened infant survival rate linked to breastfeeding

occurs with the combination of maternal transfer of antibodies against pathogens, supply of optimal infant nutrition, and the establishment of colonizing bacteria in the infant's gut microbiome [49] [44, 45].

Beyond microbes present in breast milk, infants acquire microbiota through interactions with maternal skin, vaginal and oral microbes, and their surrounding environment [42]. Thus, advocating for exclusive breastfeeding, irrespective of HIV status, remains paramount for ensuring the best survival rate among HEU infants.

Gut microbiome

The Human Gut Microbiome: A Dynamic Microbial Organ

Within the mucosal tissues of humans exists a microbiome, which has gained the nickname “the microbial organ” due to its significance to human health [51, 52]. In the human gut, this microbiome represents the largest symbiotic community, comprising of millions of microbial species that reside within or on the human host [51-53]. Analysis through stool sampling has revealed a microbial genome approximately 150-fold larger than that of humans [54, 55]. Over the past few decades, the study of the human microbiome has gained significant interest, shedding light on the mutually beneficial relationship between millions of microbes and their host, crucially impacting the host's health [51, 56, 57].

These microbes thrive in the warm and nutrient-rich environment of the mucosal tissue, playing vital roles in the host's metabolism, enhancing nutrient bioavailability, supporting immune system maturation, and safeguarding against pathogenic invasions [51, 56-58]. The composition of this microbial organ is influenced by a myriad of factors, including host genes, environmental exposures, the birth process, and growth patterns [52].

Previously, infants were believed to be born sterile, but recent studies examining the placenta, meconium, and umbilical cords have provided evidence of microbiome presence starting *in utero* [51, 59-61]. As the fetus reaches the third trimester and becomes neurologically developed, the swallowing process commences, potentially initiating the colonization of the infant's gut by microbes present in the amniotic fluid [51]. During the first two years of life, the gut microbiome undergoes dynamic changes until it reaches a mature and stable state, which significantly influences the host's future health and susceptibility to diseases [51, 53]. Hence, the early stages of life likely play a critical role in fine-tuning the microbiome [51]. Despite its importance, the process of gut microbiome maturation remains incompletely understood, calling for further research [53].

The microbiome exhibits a remarkable ability to adapt and respond to various factors [51, 58]. Even minor supplementation of formula during breastfeeding can shift the gut microbiome towards a formula-associated profile [51]. However, dysbiosis, wherein the symbiotic relationship is altered, has been linked to inflammation and disease [62]. Dysbiosis in the infant's gut could disrupt development by affecting crucial metabolic functions and nutrient absorption, which are typically aided by specific microbial species [63].

Therefore, understanding the intricate interplay of factors that shape the microbiome and its consequences for human health is of utmost importance.

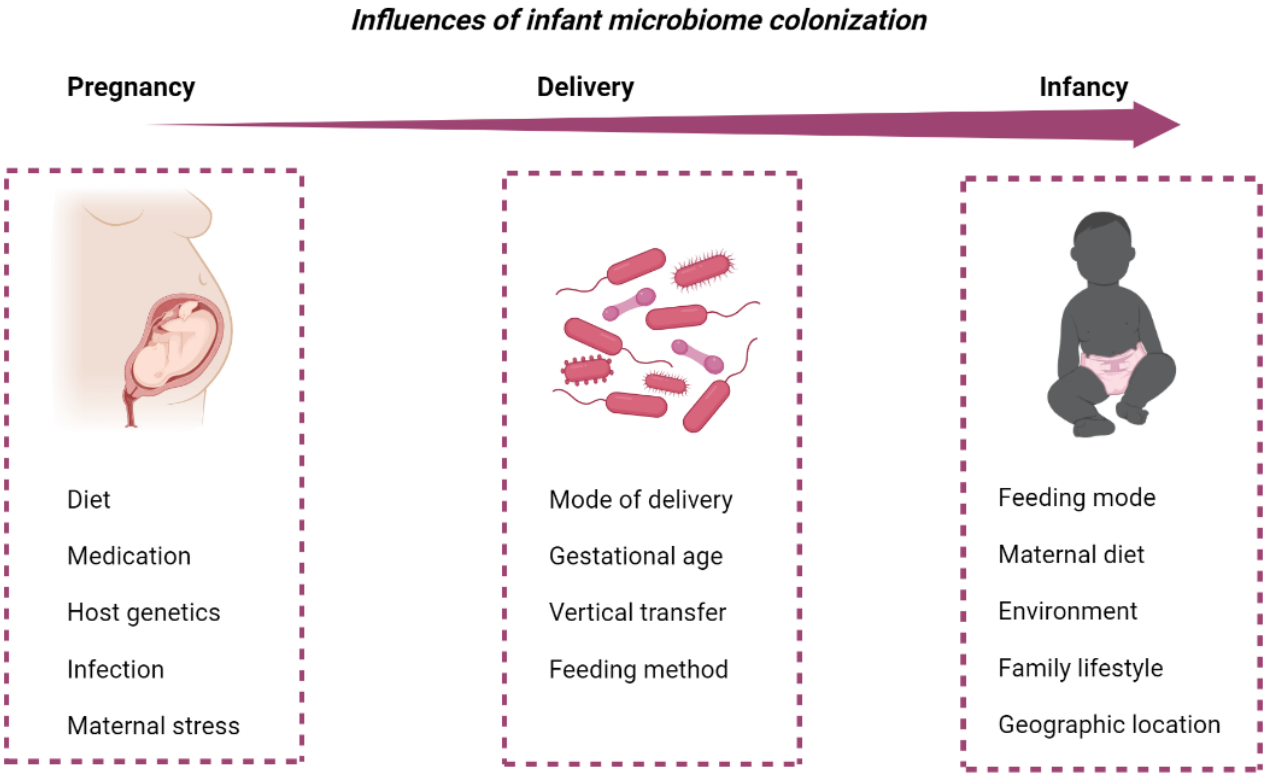


Figure 1: Influences on the microbiome through the gestational period from pregnancy to infancy. *Created in BioRender.com.*

Composition: Influences, Associations, and Interventions

Four main phyla have been identified in the gut microbiota, namely Firmicutes, Actinobacteria, Proteobacteria, and Bacteroides, which display variations influenced by socioeconomic factors such as geographical location, diet, lifestyle, and age [52, 53, 56, 59, 64]. Studies have associated a higher abundance of Actinobacteria (Corynebacterium and Rothia) and Firmicutes (Lachnospiraceae and Ruminococcacaea) with an enhanced immune response to vaccinations, while higher levels of Proteobacteria and Bacteroides (Bacteroides and Prevotella) have been linked to lower vaccination responses [57]. Moreover, both Actinobacteria and Firmicutes, as prominent gut microbiomes, have been associated with metabolic diseases, indicating a potential link between certain bacteria and susceptibility to inflammatory diseases [53, 57].

Factors influencing gut colonization encompass delivery mode, feeding methods, antibiotic usage, environmental factors, and breastmilk weaning [53, 58]. The colonization of the infant gut is significantly influenced by maternal vertical transfer, which includes microbial transmission from mucosal surfaces, skin, oral tract, and urogenital areas [51, 65]. Notably, the mode of delivery (either vaginal tract or caesarean section) has been found to impact the gut microbiome, with vaginally delivered infants showing a dominance in Lactobacillus and Prevotella species, while caesarean-delivered infants tend to harbour mainly skin-

originating microbes, such as Staphylococcus, Corynebacterium, and Propionibacterium [51, 58]. Moreover, antibiotic use has a substantial effect on the gut microbiome, leading to a shift towards an Actinobacteria-dominant community and decreased overall microbial diversity [58]. This highlights the significant impact of early antibiotic exposure on the infant's gut microbiome and the potential for dysbiosis.

In individuals infected with HIV, depleted Bacteroides and enriched Prevotella have been observed, along with increased Proteobacteria and diminished Firmicutes, indicating reduced diversity and potential microbiome dysbiosis, leading to decreased gut epithelial integrity [56, 64, 66, 67]. Such reduced diversity in HIV-infected individuals could be linked to microbiome dysbiosis and its effects on gut epithelial integrity [56, 64, 66]. The Prevotella-enriched community has been observed in healthy Agrarian cultures. In contrast, Western diets (rich in fat and protein) have shown a high Bacteroides/Prevotella ratio compared to Agrarian diets (mainly unprocessed food with grains and lean protein), which exhibit a low ratio [56]. Prevotella has been associated with an increased immune response involving T cells (activation and depletion) and dendritic cells, resulting in inflammation [56, 64]. Geographic location and sexual behaviour play important roles in gut microbiome composition during HIV infection. In the United States, men who engage in homosexual sex show increased Prevotella species, while in Uganda, individuals with HIV have significantly decreased Prevotella species [68]. The gut microbiome is dynamic and can be influenced by various interventions. For instance, analysis of 16S rRNA on the use of ARVs has shown a slight shift in the gut community, but it does not fully revert to the microbiome state prior to HIV infection [67]. Another form of microbiome manipulation involves the use of pre- and probiotics. Prebiotics supply food ingredients that favour the growth of beneficial bacteria like Lactobacillus, while probiotics contain live cultures of favourable gut bacteria, such as Lactobacillus and Bifidobacterium [51].

The infant's diet is another influential factor, with breastfed infants possessing a gut microbiome distinct from that of formula-fed infants. Breastfed infants exhibit higher levels of Bifidobacterium, while formula-fed infants have increased abundances of *Escherichia coli*, Clostridium, Prevotella, Lactobacillus, and Bacteroides [51]. Notably, breast milk contains Lactoferrin, one of its main ingredients, which has immunomodulatory and immunostimulatory effects, promoting the growth of beneficial bacteria like Lactobacillus and Bifidobacterium in the infant's gut [51]. Additionally, breast milk contains proteins, fats, carbohydrates, immunoglobulins, and immune cells [69]. Carbohydrates (oligosaccharides) in breast milk are only partially digested in the infant gut until reaching the colon, where Bifidobacterium ferments them, producing short-chain fatty acids [69]. The composition of breast milk shapes the gut microbiome, as certain bacteria prefer specific compounds for growth. For example, Bacteroides avidly consume fucosylated oligosaccharides [69]. Moreover, breast milk contains its own microbiome, mainly populated by Staphylococcus, Streptococcus, Bifidobacterium, and Lactobacillus [51]. Maternal health and diet influence the breast milk microbiome, underscoring the importance of maternal diet variety and a focus on a healthy

gut microbiome [65]. In woman with HIV-infection ART drugs in the breast milk are associated with lower Bifidobacterium abundance in HEU infants in the gut microbiota over the first 6 months postpartum [70].

Furthermore, it is believed that the maternal gut microbiome influences the breast milk microbiome through the "entero-mammary" pathway, involving bacteria from the gut travelling into the lymphoid tissue system to reach the breast milk [51].

In addition to colonization, commensal bacteria play a crucial role in maturing the mucosal gut immune system [52, 57]. The interplay between the microbiome and the immune system is essential for gut homeostasis, as they work together in intricate feedback loops to support a healthy host immune system [52, 71]. Germ-free mice studies have provided compelling evidence for the vital role of the microbiome in immune development and function, as the absence of a gut microbiome resulted in immune deficiencies and inflammation [52, 55, 58, 72]. These findings underscore the significance of a healthy and symbiotic gut microbiome for promoting optimal immune development and functioning [58, 73].

Interplay between the microbiome and immune system

The interplay between the gut microbiome and the immune system is a complex and critical relationship that influences various aspects of host's health. The gut's mucosal immune system acts as a crucial barrier against the translocation of bacteria into the bloodstream, which could lead to systemic infections in other organs [64]. Maintaining a healthy gut microbiome is essential for immune homeostasis, and any disruptions to the microbiome can result in dysbiosis, inflammation, and compromised immune system functioning, making the host susceptible to diseases [65, 74-76].

One population where the impact of the gut microbiome on the immune system has been extensively studied is HIV-infected individuals. HIV infection weakens the epithelial barrier integrity in the gut, leading to increased bacterial translocation and chronic inflammation, which further impairs immune function [56, 74]. Even with ARV treatment, chronic inflammation due to gut dysbiosis persists [74]. Hence, gaining insight into the functioning of the gut, encompassing both the microbiome and the immune system, could help identify crucial intervention targets to enhance host health [52, 74].

The gut mucosal immune system is comprised of various immune organs, cells, cytokines, and receptors [52, 64, 67, 76]. These components work together to maintain the integrity of the gut lining and prevent bacterial translocation and systemic inflammation. The mucosal immune system serves as the first line of defence against pathogens in the gastrointestinal tract [52]. Epithelial cells play a critical role in immune surveillance and communicate with the immune system through cytokine and chemokine production [72]. Additionally, mucosal cells secrete proteoglycans, forming mucus that contains anti-microbial peptides (AMPs) [72]. These secretory cells in the small intestine are Paneth cells and are an essential part of the innate gut immune

system, as they secrete alpha-defensin peptides in response to microbes and their antigens [52, 67, 76]. These AMPs play a crucial role in the mucosal defence of the gut, directly affecting the gut microbiome [52].

Lymphoid aggregates, known as mucosal-associated lymphoid tissues (MALTs), are present in all mucosal sites and facilitate B and T cell maturation upon antigen exposure [72]. In the gut, these MALTs are referred to as Peyer's patches [72]. In germ-free mice, MALTs are hypoplastic and have reduced immune cells, highlighting the importance of the microbiome in effective immune function [72]. Peyer's patches also facilitate immunoglobulin A (IgA) B cell maturation and the production of plasma cells, which are crucial for T-cell responses in the gut [52, 72]. These cells then migrate to the lamina propria to carry out specific immune responses [52].

The lamina propria houses innate lymphoid cells that secrete cytokines and IgA upon activation, playing a role in inflammation [72, 76]. Commensal bacteria play a vital role in the development and activation of lamina propria T cells. The microbiome also affects the production of various cytokines and the proliferation and differentiation of immune cells, including T cells (regulatory and helper) and B cells (IgA or IgG-secreting B cells) [58, 73]. Some bacteria, like Lactobacillus, Bifidobacterium, Bacteroides, Streptococcus, and Clostridium, have been linked to inducing T regulatory cell differentiation, which aids in immune surveillance and maintaining tolerance to commensal intestinal microbes while eliminating pathogens [58, 69, 73]. (IELs) are critical for gut homeostasis alongside the microbiome, as they produce IFN-y, IL-22, and IL-17 [52, 64].

To facilitate cytokine and chemokine communications, innate immune cells await antigen exposure beneath the epithelial layer [72]. Once exposed to breached bacteria, macrophages phagocytose the antigen and activate immune cell recruitment, including neutrophils, B cells, and T cells [72]. Microbe presence also influences mature B cell class switching, highlighting the significance of the microbiome in immune functioning and development [72]. Germ-free mice, which lack a microbiome, exhibit reduced Peyer's patches, IgA plasma cells, and decreased lamina propria CD4+ cells, further emphasizing the critical role of the microbiome in immune functioning and development [55].

The gut microbiome also has a significant impact on nutrient processing and metabolism. Microbes assist in carbohydrate fermentation, production of short-chain fatty acids (SCFAs), and vitamin synthesis, which in turn benefit the host [52]. SCFAs, such as acetate, have the ability to move through the lamina propria and enter the circulatory system, further illustrating the microbiome's regulation of the immune system [64]. The gut mucosal lining serves as the primary site for microbe-host interactions, and the influence of the microbiome extends beyond the gut, impacting overall host health [52]. The host immune system is not solely influenced by internal factors, but it can also be impacted by external factors such as the maternal gut microbiome. During pregnancy, the maternal gut microbiome plays a crucial role in shaping the in-utero infant's immune development. However, if the maternal microbiome experiences dysbiosis, it can potentially lead to compromised immune development and negatively affect the infant's health [51, 65].

The interplay between the gut microbiome and the immune system is a dynamic and intricate relationship that plays a pivotal role in maintaining host health. The gut's mucosal immune system acts as a frontline defence, while the microbiome contributes to immune homeostasis and influences the functioning and development of various immune cells. Dysbiosis in the gut microbiome can lead to inflammation and compromise the immune system, making the host vulnerable to diseases. Understanding the intricate connections between the microbiome and the immune system is essential for identifying potential intervention targets to enhance host health.

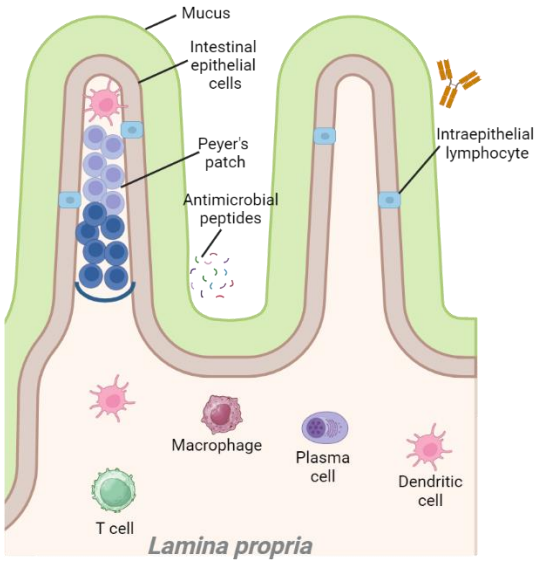


Figure 2: The cellular and structural composition of the small intestine highlighting the gut immune system. Intestinal epithelial cells with a mucosal layer separate the lumen from the underlying tissue. Immune active sites, such as the Peyer's patch, response to microbial stimulation via lymphocyte maturation and migration to effectors sites (lamina propria). Adapted from McDermott [2]. *Created in BioRender.com.*

Analysing the complex microbiome

The gut microbiome has been a subject of extensive research, initially relying on limited culturable organisms from stool samples and mucosal gut tissue. As technology advanced, more sophisticated techniques such as PCR, 16S rRNA gene sequencing, Human Intestinal Tract chip microarray, and deep shotgun sequencing were employed [51, 56, 57, 60]. The advent of 'omics' technologies, particularly metaproteomics, has ushered in a deeper understanding of the complex microbial community within the human host [54, 56, 77].

The early focus of gut microbiome research was simply identifying the different microbial species present, but a significant paradigm shift has occurred. Researchers now emphasize understanding the functional status of the microbiome - "what is active and interacting?" [54, 56, 77]. This change in approach has been pivotal in advancing our knowledge about the intricate interactions between the gut microbiome and its host. Proteomics has emerged as a powerful tool to bridge knowledge gaps, enabling researchers to gain insights into the active genes, proteins, and pathways of the gut microbiome, a comprehensive analysis referred to as the metaproteome [56, 77]. Through metaproteomics, researchers explore the interactions between the

gut microbiome and the host, shedding light on how the microbiome exerts its effects within the host's gut [77].

Metaproteomics fills the gap in understanding gut microbiota functionality that other microbiome analysis methods miss [78]. Metaproteomic analysis of fecal samples offers valuable insights into the functioning of gut microbes involved in digestion, metabolite production, and homeostasis maintenance [54, 56, 77]. The proteins and peptides present in these samples, which can be utilized for disease biomarkers or targeted treatments, result from secretion, leakage, or exfoliation within the gut [77].

Mass spectrometry

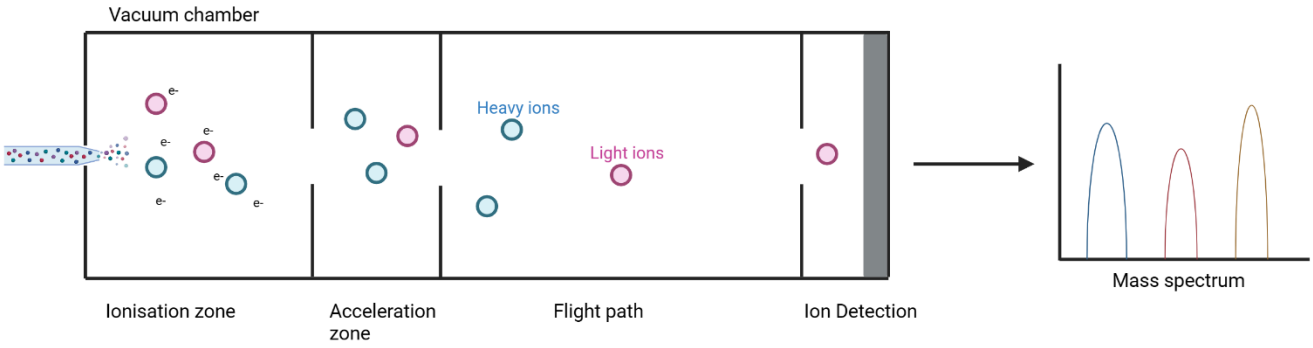


Figure 3: Time of Flight mass spectrometer. Samples are vapourised and injected into vacuum. Electron impact ionisation occurs by firing electrons at ions to positively charge them in the ionisation zone. The now positive ions are now accelerated by an electric field in the acceleration zone. Lighter ions accelerate faster than heavier ones. Accelerated ions then enter flight path without an electric field and their time to reach the detector is measured. Lighter ions will hit earlier than heavier due to speed, allowing for ion detection. The mass spectrum is then produced and ready for analysis.

Proteomic analysis methods have been developed and utilized to study the functioning of present genes and systems, with a shift in focus towards analysing complex samples for disease identification. Techniques such as gel-based and gel-free approaches are coupled with mass spectrometry and array-based methods [77].

Mass Spectrometry Separation Methods:

For complex samples, bottom-up mass proteomic methods involving enzymatic or chemical digestion prior to peptide analysis via MS/MS fragmentation are favored over top-down methods to obtain sequence data [74]. The gel-based analysis involves HPLC and LC to separate proteins based on size, properties, and binding affinities. While liquid chromatography coupled with tandem mass spectrometry (LC-MS/MS) is a basic tool for proteome analysis, challenges such as reproducibility, sensitivity, and accuracy persist [76, 77].

Mass spectrometry (MS) is a pivotal analytical technique that involves the isolation and characterization of gas-phase ions based on their mass-to-charge ratio [78]. Conventionally, electron impact was employed by colliding electron beams within a vacuum, yet this method posed limitations, especially for analyzing large

molecules or biological samples due to instability in forming radical ions. To overcome these challenges, more robust and versatile techniques were developed, such as fast atom bombardment (FAB), liquid secondary ion mass spectrometry (LSIMS), matrix-assisted laser desorption ionization (MALDI), and electrospray ionization (ESI). These advancements have transformed MS by providing alternative pathways for generating ions, allowing for the precise analysis of diverse molecular structures and complex biological samples, thus expanding the applications of mass spectrometry across various scientific disciplines [78].

The primary ionization methods employed in mass spectrometry, notably MALDI and ESI, are popular methods used to analyze a wide range of compounds, especially complex biological samples. ESI involves dissolving samples in a polar solvent, which are then introduced into the mass spectrometer as a spray via a charged needle under high electric potential, resulting in nebulization into charged droplets [78, 79]. With the introduction of nitrogen and the application of electric forces, coulombic interactions generate ions to move towards the detector. This method has found extensive application due to its compatibility with diverse samples and complex biological molecules. It can also be coupled with liquid chromatography (LCMS), expanding its analytical capabilities. Conversely, MALDI relies on desorbing ions from a solid state. MALDI is more capable in analyzing DNA, lipids, and glycoconjugates by preparing the sample in a solvent with excess MALDI matrix, spotting it onto a MALDI plate, and air-drying it for co-crystallization. MALDI's ionization occurs upon absorption of laser energy, but ESI often provides superior reproducibility in data [78, 79]. However, ESI tends to have higher levels of contaminants due to its sample preparation methods. ESI's advantage lies in its broader mass range and ability to produce multiple charged species, particularly beneficial for peptide analysis. Both MALDI and ESI represent soft ionization methods, making them preferable for clinical samples compared to electron impact (EI), which is too harsh for clinical settings. Overall, the choice between MALDI and ESI often depends on the specific sample type and the desired analytical outcomes in clinical and research applications [78, 79].

Mass analyzers play a crucial role in separating ions based on their mass-to-charge ratios (m/z), primarily driven by electrical forces. Four major analyzers - quadrupole (Q), quadrupole ion trap (QIT), time of flight (TOF), and Fourier transform ion cyclotron resonance (FT-ICR) - offer distinct advantages in mass spectrometry. The quadrupole analyzer utilizes four parallel rods with applied electric and radiofrequency fields, causing ion oscillation towards the detector. It's cost-effective, robust, and relatively easy to maintain, yet it has limitations in mass range and MS/MS capabilities. TOF operates on the principle of ions travelling through a tube to reach the detector. Its advantage lies in the fact that all ions eventually reach the detector (*see Figure 3).* Longer tubes enhance resolution and accuracy by increasing flight time. Serial connection of multiple analyzers allows ion isolation and fragmentation, enabling MS/MS analysis. Orbitrap, a newer mass analyzer, traps ions within an electrostatic field, offering high resolution, mass accuracy, and sensitivity. It's commonly paired with electrospray ionization and provides superior performance in many aspects. Quadrupole mass filters isolate ions of specific m/z ratios, TOF calculates ion flight times for detection, and

Orbitrap combines high resolution and sensitivity, particularly advantageous for complex sample analyses and precise measurements in various scientific domains. Each analyzer type has specific strengths, and the choice often depends on the desired analytical goals and sample characteristics in mass spectrometry applications [78, 80, 81].

DIA and DDA

As previously discussed, mass spectrometry (MS) analysis encompasses various approaches, prominently Data-Independent Acquisition (DIA) and Data-Dependent Acquisition (DDA), each offering unique strengths and limitations. DIA stands out for its exceptional reproducibility in proteome analysis, leveraging advanced data analysis software for intricate spectral interpretation and rapid mass scanning, resulting in heightened sensitivity and extensive coverage [79, 80]. In contrast, DDA relies on selecting peptide ions with robust intensities in full scans to generate fragments for database searches, albeit suffering from poorer run-to-run reproducibility and limited detection of low abundant ions. DIA strategically isolates peptide ions within specific m/z windows, subjecting them to iterative fragmentation until the full m/z range is covered, often requiring prior sample knowledge for library-based matching (*see Figure 4*) [79, 80]. While DIA demands specialized computation tools, high identification coverage, accuracy, and reproducibility, DDA, with its easier data analysis pipelines, presents medium identification coverage, lower reproducibility, accuracy, and dynamic range. Additionally, tandem mass spectrometry (MS/MS) techniques, integral to these approaches, enable protein and peptide fragmentation, providing insights into amino acid sequences for both top-down and bottom-up proteomics [79, 80]. DIA's single precursor ion selection contrasts with DDA's multiple precursor ions, impacting their suitability for complex samples and diverse ion selections. Moreover, targeted LC-MS/MS-based proteomics leverages prior DDA-derived information to detect specific peptides, while untargeted methods aim for broader peptide discovery within precursor windows [79, 80]. These tandem MS methods hold vast potential across biological and clinical research, enabling qualitative and quantitative analysis of proteins, peptides, their modifications, and interactions in various biological samples, whether normal or pathological. [79, 80]

Mass spectrometry techniques, particularly SWATH (sequential window acquisition of all theoretical mass spectra), are increasingly used in metaproteomics [77]. Data-Dependent Acquisition Mass Spectrometry (DDA MS) has limited reproducibility with complex samples [78]. SWATH is a Data-Independent Acquisition (DIA) method that overcomes DDA's limitations, as shown by its successful analysis of lab-assembled microbial mixes and human samples [78, 81, 82]. In shotgun proteomics (discovery proteomics), the widely used method operates in DDA, where MS2 spectra are obtained for selected precursor ions detected in the

MS1 scan. Peptide sequences are then assigned through sequence database searching [81]. On the other hand, targeted proteomics involves MS in selected reaction monitoring, which requires querying the sample for the presence and quantification of a limited set of peptides specified prior to data acquisition [81]. SWATH has higher reproducibility and consistency of data compared to DDA, making it suitable for large-scale proteome characterization [81, 82]. However, SWATH data analysis requires more sophisticated software and significant informatics resources than DDA [82]. Overall, SWATH is a powerful approach for deep proteome coverage with accuracy and reproducibility, making it ideal for high-throughput analysis of proteomes in complex samples [78, 81, 82].

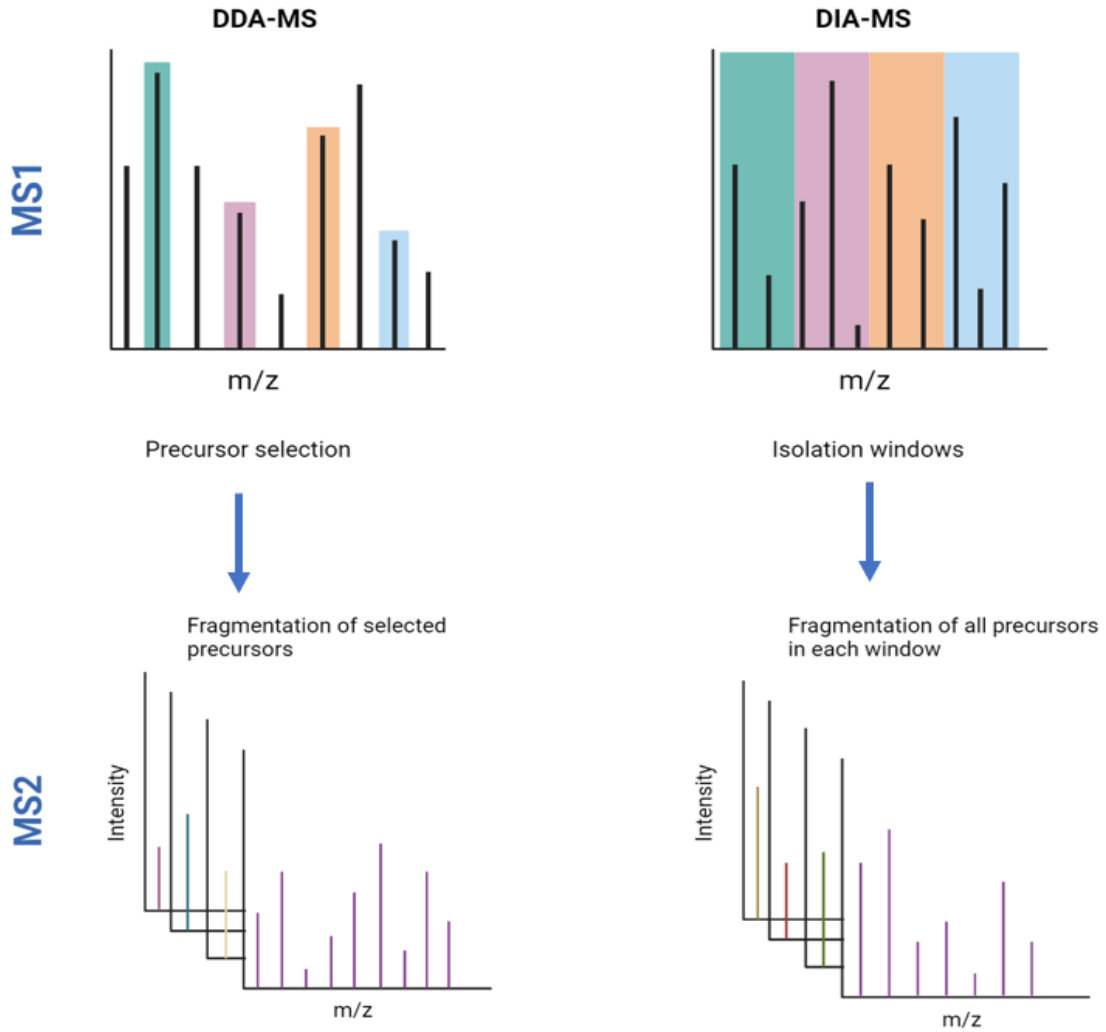


Figure 4: Comparison between two main mass spectrometry methods, data dependent vs data independent acquisition. DDA-MS showing precursor selection for fragmentation and analysis of intensities verses DIA-MS dividing mass range into isolation windows for fragmentation. DIA-MS analyses all ions regardless of abundance whereas DDA-MS selected most abundant precursors. *Created in BioRender.com.*

Despite the high instrument cost, mass spectrometry methods, both data dependent acquisition and data independent acquisition, are proving useful for analysis due to their sensitivity and coverage. DIA methods, like SWATH, offer advantages in terms of coverage, speed, and resolution, and are employed for complex sample analysis and biomarker identification [77]. After mass spectrometry analysis, data is coupled with bioinformatic tools for processing, statistical, and functional analysis. Grouping proteins into functional

categories using databases like KEGG and COG through software like eggnog is a common method of analysis [54, 83] . However, limitations exist in functional analysis due to gaps in knowledge, especially regarding proteins with unknown functions, highlighting the need for further research into the gut microbiome and its functioning [54, 84].

Metaproteomics and mass spectrometry

Metaproteomics is complex due to several reasons: the vast diversity of microorganisms within a community, each with their own unique set of proteins and the dynamic nature of microbial communities, with changes in composition and protein expression influenced by various factors such as diet, host health, and environmental conditions [85]. And finally, the technical challenges in extracting and analysing proteins from complex microbial samples, including the need for efficient protein extraction methods and advanced mass spectrometry techniques. SWATH is useful in metaproteomics analysis as SWATH allows for comprehensive and unbiased sampling of the metaproteome by systematically fragmenting all precursor ions within defined mass windows. The generation of a spectral library for a targeted approach for analysing the metaproteome is particularly useful as it allows for the detection and quantification of low-abundance proteins from complex microbial communities [85]. SWATH also provides better reproducibility and allows for retrospective data analysis, as all precursor ions are fragmented and can be re-interrogated for specific proteins. Older methods for analysing the metaproteome, such as DDA, have limitations compared to newer methods like DIA SWATH. Some of the older analysis method for analysing the metaproteome via DDA include MetaNovo, Maxquant, Proteome discoverer and OpenMS [85]. These analysis methods still retain the DDA limitations into investigating the metaproteome and new methods following DIA-SWATH have been developed. These include programs such as DIA-NN, Spectronaut, Skyline and MaxDIA [86].

DIA-NN is an integrated software suite designed for the processing of data-independent acquisition (DIA) proteomics experiments [87]. It utilizes deep neural networks (DNNs) and new quantification and signal correction strategies to improve identification and quantification performance in DIA proteomic applications. DIA-NN employs a peptide-centric approach, where a collection of precursor ions is used to generate a library of negative controls and extract chromatograms for target and decoy precursors [87]. Elution peaks are described by scores reflecting peak characteristics. DIA-NN uses DNNs to distinguish real signals from noise, with each trained network providing a likelihood score for a set of scores originating from a target precursor. These scores are used as discriminant scores to calculate q-values for identification [87]. Additionally, DIA-NN includes an algorithm for detecting and removing interferences from tandem-MS spectra, improving quantification accuracy [87]. DIA-NN has been benchmarked and demonstrated to achieve better identification performance compared to other software tools in metaproteomic analysis.

A limitation in metaproteomic analysis is the ‘missingness’ in the datasets due to the large variation of microbial species between individuals in clinical samples. SWATH data generally has missingness due to

sample complexity and the absence of various bacterial communities and their proteins, and not due to technical error or solely due to chance [88]. Missingness in metaproteomics data can pose challenges for accurate downstream analysis and interpretation of the results, which could affect the quantification and identification of proteins [88]. Strategies for handling missingness in metaproteomics data include imputation methods to estimate missing values, regression models, statistical approaches to account for missing data in downstream analysis, and quality control measures to assess the impact of missingness on the overall results [88]. However, it is important to carefully consider the missingness patterns and choose appropriate imputation methods to ensure valid and accurate analysis of SWATH data.

In conclusion, metaproteomic studies with the powerful analysis tools that exist offer valuable insights into the intricate functions of the gut microbiome and its interactions with the host. The continuous development of proteomic techniques and bioinformatic tools promises to deepen our understanding of this complex ecosystem and its potential applications in medicine and human health.

The focus of this project involves the analysis of intricate biological samples employing the Electrospray Ionization (ESI) source coupled with the Time-of-Flight (TOF) mass analyser, adopting the Data-Independent Acquisition (DIA) approach through tandem MS/MS. Additionally, SWATH was utilised with DIA-NN for protein identification. This methodology has been developed to prioritize sensitivity, comprehensive coverage, and accuracy in analysis.

Aims and objectives

Aim: To analyse the metaproteome of the infant gut through the secreted proteins to identify any alterations in a HIV exposed infant's gut, through stool microbiome analysis, accounting for their poor health.

The objectives of this study are:

1. Develop and optimize a SWATH-based mass spectrometry methodology for a metaproteomic analysis of infant stool samples (using the TTOF 6600).
2. To identify and quantify the metaproteome of secreted proteins in the infant gut microbiome. Carry out a statistically powered metaproteomic analysis on the protein data obtained from mass spectrometry.
3. Determine analysis methods for the metaproteome to help identify any potential causative relationships existing between the microbiome and infant's health.

Chapter 2: Methods

Clinical data

Retrospective study design

This study forms part of the parent study: *Innate, Adaptive and Mucosal Immune Responses in HIV-1 Exposed Uninfected Infants: A Human Model to Understand Correlates of Immune Protection* with **HREC/REF:285/2012**, referred to as the "**InFANT study**" (Innate Factors Associated with Nursing Transmission). This case-control study makes use of samples collected from an ongoing cohort established in Nigeria and South Africa for the InFANT study. The cohort enrolled 500 HIV-exposed mother-child pairs and 200 HIV uninfected controls. The infants were dominantly exclusively breastfed.

Ethics statement

Ethical clearance given to parent study "InFANT study" for infant stool samples **HREC/REF:285/2012**. Sub-study protocol for ethical clearance to do this research was obtained from The UCT Human Research Ethics Council (**HREC/REF:219/2022**). The "InFANT" study has all participating individuals signed consent forms.

Sample collection and recruitment

No new participants were recruited under this sub-study; it was a retrospective analysis of banked infant stool samples already collected and stored under **(HREC REF: 285/2012),** hereafter referred to as the "**Infant study"** by Dr Clive Gray. A subset of these infant stool samples was obtained and used for this study from the South African cohort.

Study site and population

Infant study (**HREC/REF:285/2012)** enrolled women from Cape Town, Khayelitsha from 2013-2021. We analysed a subset of 96 samples previously collected in the InFANT cohort to analyse the metaproteome of the gut during the first week of life (see Appendix A for full manifest metadata).

Inclusion criteria included mothers > 18 of age; new-born infant (> 2 kg); gestational period > 35 weeks, have chosen to breastfeed the infant and have known HIV status. Prior obstetrical history was gained (such as previous number of pregnancies and outcomes), last child feeding, menstrual period history, and delivery date. In terms of HIV history, their HIV status, ARV history and CD4+ cell count was collected.

Additionally, the infant ARV and HIV history was collected. The infant health outcomes were obtained: HIV status, weight, length, vaccination history, use of medications, and any health issues. Any complications such as feeding related issues from mother (breast milk production, nipple health) and infant adverse outcomes (rash, diarrhoea) were documented.

Table 1: InFANT study metadata summary from participants.

InFANT sub-study clinical data	
HIV status:	
HIV infected	52
HIV uninfected	42
Age of mother (years):	
Average	27
Range	23
Age of infant	Day 4-7
Gender of infant:	
Female	28
Male	60
Infant weight (grams):	
Average	3156
Range	2160
Gestational period (weeks):	
Average	39
Range	7
First pregnancy:	
Yes	18
No	62
Delivery mode:	
Vaginal	81
Caesarean section	7

Table 2: HEU vs HU metadata

	HEU	***HU***
Maternal age		
Average	29.02	26.23
Range	19	23
Number of samples	48	40
Gender		
Male	33	27
Female	15	13
Weight of baby		
Average	3146.46	3168.75
Range	1760	1960
Gestational age (weeks)		
Average	39.08	38.9
Range	5	7
1st pregnancy		
Yes	6	13
No	42	27
Delivery mode		
Vaginal	44	37
Caesarean	4	3

*Some metadata is missing for samples due to clinician technical errors

Laboratory methods

Sample preparation

Each stool sample was catalogued with the InFANT study manifest, and then analysed and grouped in a low quantity or normal quantity category due to the variability in stool volume received. The samples were defrosted and placed in clean Eppendorf tubes for suspension with phosphate-buffered saline (PBS). The amount of PBS required to suspend the sample was based off this volume categorisation low and normal which varied between 50 µL to 1 mL of PBS, respectively. Once PBS was added to each sample, all samples were vortexed to ensure all sample were in liquid. 200 µL of sample was pipetted into a 96 well plate and each location was catalogued. Samples were then centrifuged for 10 minutes at 4000 xg and 20 µL of supernatant from each sample was removed and placed in a clean 96 well plate for future metabolic analyses. The pellets were resuspended in 180 µL of a 20 °C 8:1 acetone methanol solvent. After samples were placed in the -20 °C freezer overnight. Samples were then placed in the centrifuge and spun at 4000 xg for 10 minutes. 100 µL of supernatant was removed and placed into a new 96 well plate for further metabolomic

and lipidomic analysis. The remaining pellets were washed with 80 % acetone and centrifuged for another 10 minutes at 4000 xg. The supernatant was then discarded, and the pellets were air dried briefly before adding denaturation buffer. 100 µL of denaturation buffer (6 M Urea, 2 M thiourea in 10 mM Tris pH 8.0) was added to each sample and left for 2 days on a shaker in the -20 °C freezer, ready for protein quantification and digestion.

Protein quantification and digestion

Samples were split over 2 plates to add standards onto each plate with the samples. Standards of Bovine serum albumin (BSA) (0 – 2000 mM) were prepared by serial dilution. 2.5 µL of each sample and standard was pipetted into designated wells and 75 µL Coomassie reagent was added to each well. After shaking on plate shaker for 30 seconds, samples were incubated for 10 minutes in the dark. Protein concentration was read on the spectrophotometer at 595 nm using the standard curve based off the standards.

Total protein was calculated for each sample and 0.5 µg/µL protein was added to the total volume of 20 µL using PBS for each sample. However, the low quantity samples with less than 10 µg of protein, were made up to 5 ug protein up to 10 µL with PBS. Samples were reduced with 3 mM DTT (dithiothreitol) from 1 M stock solution (0.06 µL each well) for 20 minutes. The samples with 5 µg protein had 0.03 µL DTT added (3 mM). Samples were alkylated with 15 mM IAA (iodoacetamide) for 20 minutes in the dark. Ensuring pH is adjusted to 8 with Tris where necessary. Diluted samples in ammonium bicarbonate (ABC) in a 6 X dilution to ensure Urea is below 1 M. Samples digested with trypsin (1:100, Promega), then vortexed and incubated overnight at 30 °C. All samples were used to create a pool (1 µL of each sample) for control during the mass spectrometry run. See Appendix B for concentrations loaded and Appendix C for samples of concern and removed or lost samples.

De-salting and purification

To stop the trypsin digestion, samples were acidified with 0.5 % formic acid. Evosep tips were rinsed with 20 µL Solvent B (ACN, 0.1 % formic acid). The ends of tips containing the resin were soaked in propanol until resin turned pale white. Tips were equilibrated with 20 µL Solvent A (milliQ H_2O, 0.1 % formic acid) and then the sample was loaded to each Evotip. The loaded tips with sample in were washed with 20 µL Solvent A. After each solution was added, the loaded Evotips were centrifuged at 800 g for 60 seconds to ensure solution had passed through the resin in the tips. To ensure the tips and samples were preserved before analysis, 100 µL of Solvent A was transferred to the tips and centrifuge at 800 g for 10 seconds. Solvent A loaded tips were stored at 4 °C until analysis.

Mass Spectrometry

Sample in Evotips containing 1 µg peptides were analysed using an Evosep One Autosampler (Evosep, Odense, Denmark) LC system coupled to quadrupole time-of-flight (QTOF) mass spectrometer, TripleTOF 6600 SCIEX, with the optiflow source from SCIEX.

The Evosep40-SPD method on the autosampler was used with Chronos to load samples on recommended C18 column (EV112 Performance column – 15 cm x 75 µL, 1.9 µm) from Evosep and chromatic separation performed on C18 column (Aurora Elite column) on a water and acetonitrile gradient, 0.1 % FA.H_20 / 0.1 % ACN. Gradient flow of 100 nl/min with 31 minutes method duration at 50 °C for all samples.
SWATH acquisition was performed on the samples, each cycle started with MS-spectrum (5 – 1500 m/z) followed by MS/MS with 120 variable SWATH windows (variable scan range: 344 m/z - 900 m/x,). The MS1 injection time was 50 msec and was followed by MS2 with injection time of 20 msec (with MS2 fragment scan range at 50 – 1500). Analyst software (SCIEX) was used to acquire all MS and MS/MS data.

*5 samples had technical difficulty on the mass spectrometer and were lost, see Appendix C.

Data analysis

Data pre-analysis and clean up

DIA-NN

A spectral library for DIA-NN 1.8 was made from the 16S data for day 4-7 infants from the same cohort (previously run by Saori, PhD Heather Jaspan) [89]. This was done by selecting the taxa who were present in many samples or had a high abundance in few samples (done by multiplying frequency and abundance and determining a cut off). DIA-NN was chosen as the software of choice due to its in-depth proteomic profiling and accurate and reproducible profiling [86]. Initially, the top 55 taxa from the 16S dataset were chosen and used in DIA-NN to make a spectral library. Following the assessment of this output, additional taxa were added, as well as a few taxa with low frequency and abundance to assess mass spectrometry sensitivity. This finalised 108 taxa proteomes were then used for the spectral library. These taxa's proteomes were downloaded from UNIPROT (if proteome unavailable, UniProtKB file used), along with the human proteome. These fasta files were used in DIA-NN to create a spectral library for which our 95 SWATH files could be searched against.
Initially, the top 55 taxa were used to create an in-silico-predicted spectral library from their respective fasta files from UNIPROT, with the FASTA digest library-free search/generation and deep learning based-prediction mode enabled. This was then repeated with the additional fasta files to make up the top 108 taxa, which was then followed by the final run which made use of a merged in-silico spectral library of the 2 smaller libraries generated (55 taxa run and 108 taxa run) was performed with reannotate on and match between runs (MBR) selected.
DIA-NN was run with reannotation on, FASTA digest enabled, deep-learning on enabled, the neural network classifier set to double-pass mode, with protein inference set to genes (specific-specific), quantification

method selected as robust LC, RT-dependent, smart profiling enabled, optimal results selected, heuristic protein inference enabled, no shared spectra, use isotopologues on, FDR set to 1 % and threads at 32 for the final sample run. The MBR in the final run helps create a new library and reanalyse the data from this, with the same settings in DIA-NN as described above (Output from DIA-NN can be seen in Appendix D).

The final DIA-NN library search matched 6558 proteins from the in-silico libraries. Instead of using the protein groups (pg)_matrix, we made use of the main report matrix due to the stringency applied to the pg matrix in DIA-NN. The PG.Max.LFQ values were used with protein group, protein name, and file names used to create a protein group matrix after 0.01 filtering on Lib.PG.Q.Value and Lib.Q Value. This filtering was performed in R studio using tidyverse followed by data clean up and processing (see protein quantification per sample in Appendix E).

Data processing

Main data report was imported into R studio using tidyverse package for filtering. Both Lib.Q and Lib.PG.Q values filtered to exclude values > 0.1.

Pools were analysed in R studio using ggplot and Pearson coefficients (> 0.97) to ensure mass spectrometry runs were consistent throughout the run and any differences found are biological and not due to technical error (see pool analysis in Appendix F).

Due to limitations of DIA-NN's machine learning algorithm, we accounted for misidentifications for proteins by assessing the ratio of the protein abundance sum to the protein median. Any value that fell outside of our decided 400 ratio limit was further researched and excluded if deemed misidentified. This resulted in 6 proteins being excluded from different taxa (Appendix I).

To remove samples with low protein content, any sample with a protein count that was lower than 2 standard deviations of the total samples' mean was excluded (following the Empirical 95 % Rule). This resulted in 4 samples being excluded, see Appendix G and H for cut offs and samples removed.

A 10 % observational filtering was done on the proteins in Perseus. The 10 % cut off was applied to ensure filtering is not overly stringent and avoids removing potential biomarkers. Proteins were needed to be detected in > 10 % of total samples (in at least one group) to be kept. This filtering removed proteomic outliers in the samples (which could account for individual heterogenicity) from analysis. Missingness has been shown to be reproducible as a biological observation, rather than due to limit of detection (LOD) and is not random, with SWATH data. Therefore, imputing in SWATH data would remove biological differences and significances, as well as stringent filtering on SWATH data, and therefore was not done. This method of dealing with missingness in SWATH data was acquired from *McGurk et al* [88].

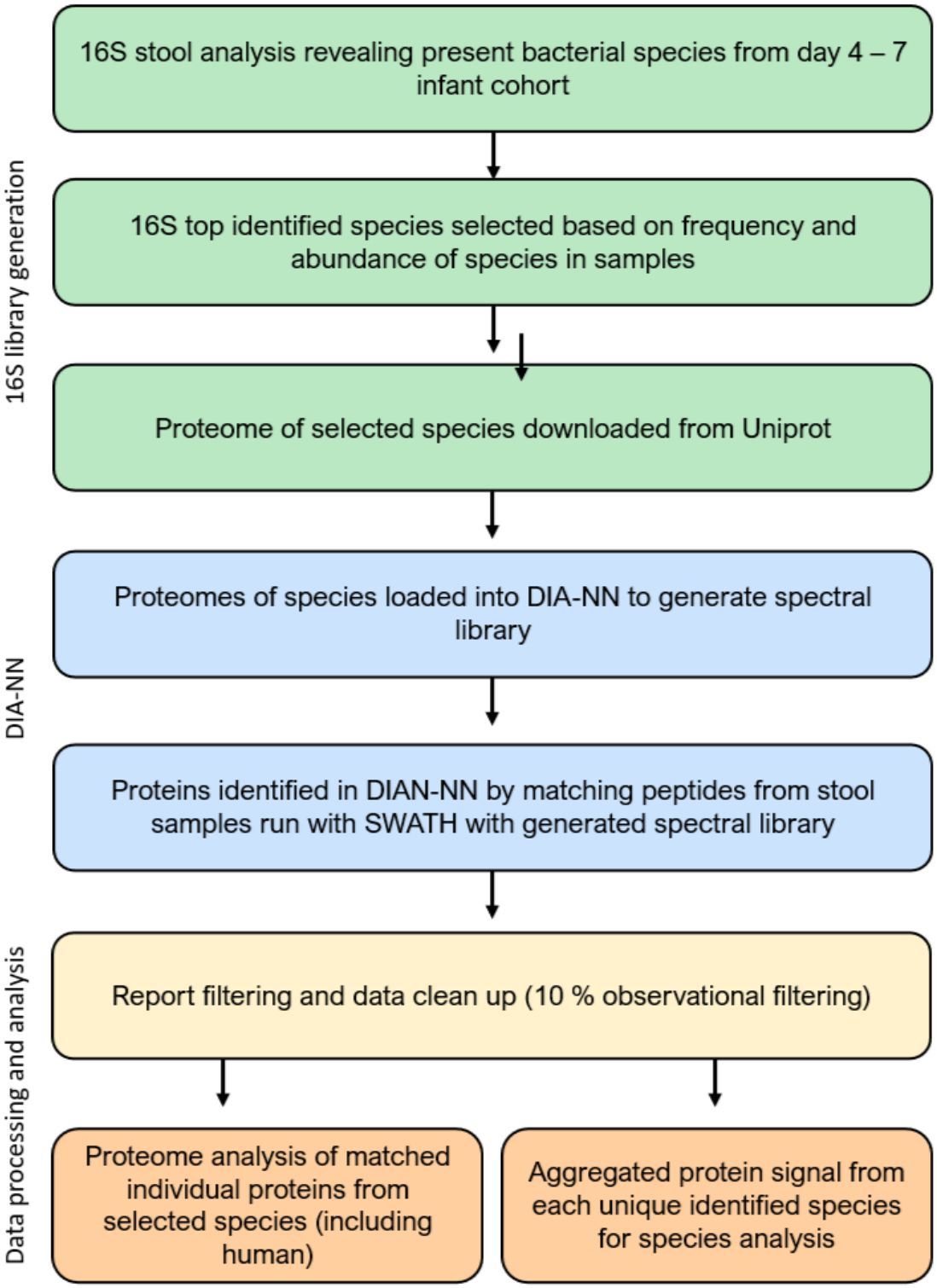


Figure 5: Flow diagram of data analysis pipeline from library generation based off the 16S data to proteome and species analysis from DIA-NN analysis.

Data analysis

Unipept and phylum analysis

The main report file from DIA-NN was filtered in R studio using tidyverse for peptide sequences was done under same conditions as the filtering for the proteins, Lib.Q and Lib.PG.Q > 0.1. However, using peptide

sequences, instead of protein groups.

Peptide list was run in the metaproteomics analysis tab on Unipept (Unipept (ugent.be)) to assign species to peptide sequences. This allowed insight into main phyla, genera, and families, all the way down to species level. The identified main phyla were used to group the proteins and compare phyla between the 2 groups (HEU and HU), as well as the total taxa composition of samples. Relative bacterial composition of each sample, from phyla data, was loaded into Graph pad prism 10.0 and Welch (Mann-Whitney) 0.05 p-value t-test with column scatter plots was created showing means for the two groups in each phyla (Figure 11).

Database assignment

Protein data was split into different datasets using R studio and Microsoft excel human, bacterial, species, phyla, breastmilk, and infant derived proteins. The breastmilk database was created from literature to try account for maternal origin proteins separately to infant derived proteins from all the human proteins [90-92]. R studio packages used to create these databases: readxl, openxlsx, writexl, dplyr and tidyverse.

Using R studio, the relevant taxa were assigned to one of the four main phyla (Actinobacteria, Firmicutes, Proteobacteria and Bacteroides), see Figure 10B.

Metadata

Readxl and openxlsx was used to filter out the subset of samples used in this research project from the total InFANT manifest. The subset of samples were assigned to their signals in excel and missing metadata for samples was identified. Selected metadata was chosen included age of mother, mother's housing, running water in house, living with a man, gender of baby, weight of baby (grams), gestational ages (weeks), pregnancy number, age of baby (days) and delivery method (see appendix A). Metadata known to influence the microbiome (delivery and gender) were used to assess clustering in a heatmap, however this metadata was underpowered, so focus remained on HEU vs HU infants (see Appendix N for this clustering).

Diversity analysis

In R studio vegan, ggplot2 and readxl packages were used for the diversity analyses. The Shannon index was applied to the taxa database to assess the alpha diversity of the samples and box plots were used to visualise these results. Following this, Mann-Whitney/ Welch's t-test was performed between the two groups. The Bray method was applied for the beta diversity assessment, with a PCA graph used to visualise these results. Following this, Mann-Whitney/ Welch's t-test was performed between the two groups for the beta diversity.

Metaboanalyst

All the protein databases for the samples were loaded into metaboanalyst 5.0 in the generic format using the statistical analysis (one factor) tool. Peak intensities were loaded in samples in columns (unpaired). All missing values were replaced with 0 to highlight absence of protein and no other imputations performed to prevent showing a presence. No further data filtering was done as to avoid removal of heterogenicity of the samples and biologically significant proteins. Data was normalised through being log transformed and Pareto

scaling was performed on the data (see effects of transforming and scaling in Appendix J).
Unsupervised clustering via k-means was first used to identify any clusters in the data which could account for differences between HEU and HU clusters. This produced 3 distinct clusters, which were unaccounted for from the current clinical metadata (see Appendix K). A larger cohort would be needed for statistical power to analyse these clusters in this manner. Following this analysis, supervised clustering (Random Forest model) was undertaken to identified distinguishable features between the two groups pf interest, HEU and HU infants.

To distinguish between the HEU and HU group, a Random Forest regression model was used to generate classifiers based on the dataset provided. This method aided in the issues that missingness brought into the analysis and keeping sample individuality.

A Random Forest model was then applied to the top 10 % of proteins identified as classifiers as the new basis for the model. This repeat of the model on a subset of the data improved the classification accuracy, out-of-the-bag (OOB) error. The table below shows the OOB errors for each Random Forest model from the different datasets. A medium level of classification is defined by an OOB error of around 0.5, with a stronger classification accuracy associated with OOB error below 0.5 tending towards 0.
For the human datasets, the top 45 proteins were then mapped in STRING DB for an enrichment analysis on the protein interactions. The top 25 proteins from the repeated Random Forest were then used in Enrichr and Appyter for GO Biological process analysis and functional analysis. See Appendix M for top 10 % of human proteins from the Random Forest model.

Table 3: Random Forest model out-of-the-bag errors.

Dataset	**All proteins OOB error**	**10 % of proteins OOB error**
All proteins	0.506	0.253
Human all proteins	0.501	0.304
Infant derived proteins	0.430	0.253
Breastmilk proteins	0.506	0.354
Bacteria	0.519	0.266
Species	0.481	0.241
Firmicutes	0.457	0.265
KEGG pathways	0.564	0.504

Kyoto Encyclopedia of Genes and Genomes assignment (KEGG)

To help allow for the missingness in the microbial data, bacterial gene ontology and functional analysis was performed. The chosen method was to use the diamond algorithm within eggNOG-mapper and the EggNOG database used to assign functional groups to the proteins, *see Figure 6* [93]. Protein names were loaded into Uniprot list option and the compressed fasta file was made for all identified proteins. This was then loaded into EggNOG mapper v5.0 and the output file with all matching KEGGs downloaded and matched to protein

intensities in excel. The two selected outputs used from EggNOG were KO terms for bacterial gene ontology to account for bacterial species performing the same function in a community setting, and KEGG pathway terms, which accounts for pathways the proteins are involved in, allowing for bacterial pathway analysis.

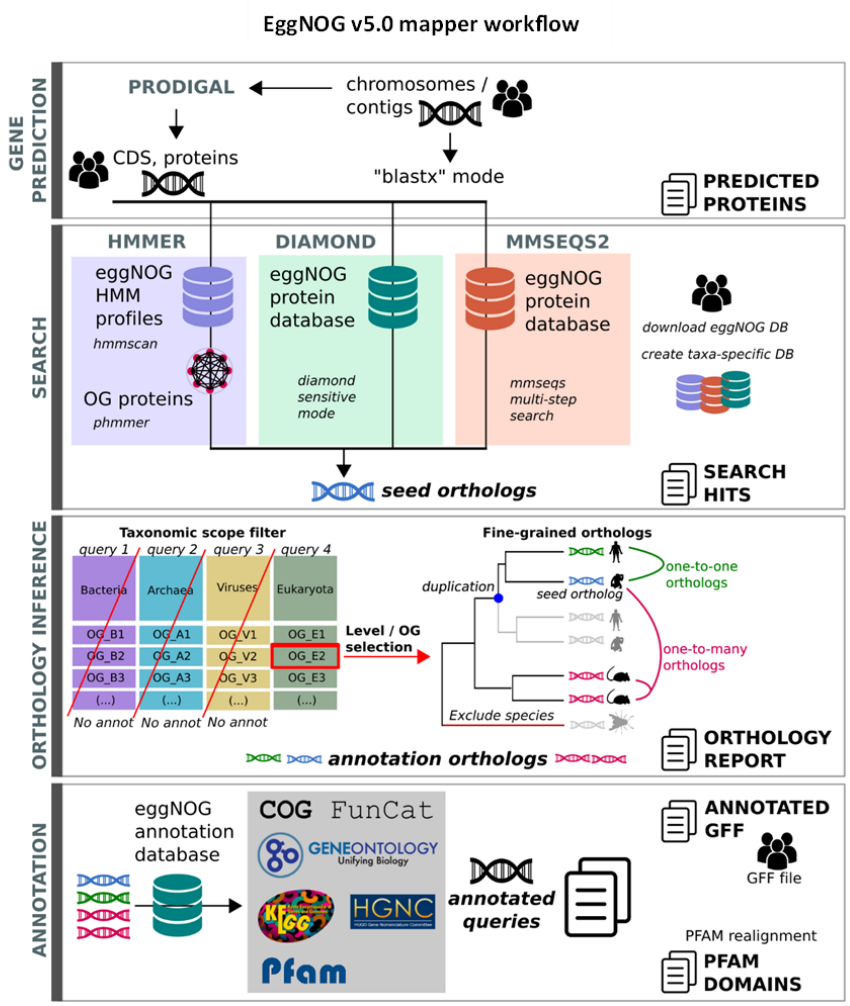


Figure 6: EggNOG v5.0 mapper overview [1].

Statement of Contribution

Clinical samples (including infant stool and matched maternal breast milk samples from HEU and HUU participants) were previously collected and banked by clinical collaborators as part of the parent 'INFANT' study, as described above.

I carried out all sample preparation (protein preparation, tryptic digest, peptide clean-up and quantification) for mass spectrometry analysis. Mass spectrometry on the prepared peptides to generate raw data files was performed by Dr Tariq Ganief. I then carried out all analysis of the resultant raw and processed mass spectrometry data, as well as analysis of the clinical metadata.

Chapter 3: Results

Chapter 3.1 – Mass spectometry summary statistics

Following DIA-NN spectral matching, 6557 proteins were identified with 13828 peptides identified, with medians of 980.5 proteins and 987.5 peptides per sample. See Appendix E and G for protein distribution and protein counts of each sample. Proteins were identified with a median of 1 unique peptide per protein group, with a total of 3349 unique protein groups being identified with 1 peptide, and subsequently the remaining proteins having more than 1 peptide per protein identification. Reproducibility was high, > 97 % between pool replicates, indicating this method is highly reproducible (see Appendix F). Median presence of total identified protein per sample was 14.95 %, highlighting the missingness present in the SWATH data.

Table 4: Mass spectrometry statistics.

Summary statistics	Count
Proteins identified	6557
Peptides identified	13828
Proteins per sample	980.5
Median peptides per sample	987.5
Median peptides per unique protein group	1
Proteins with 1 peptide	3349

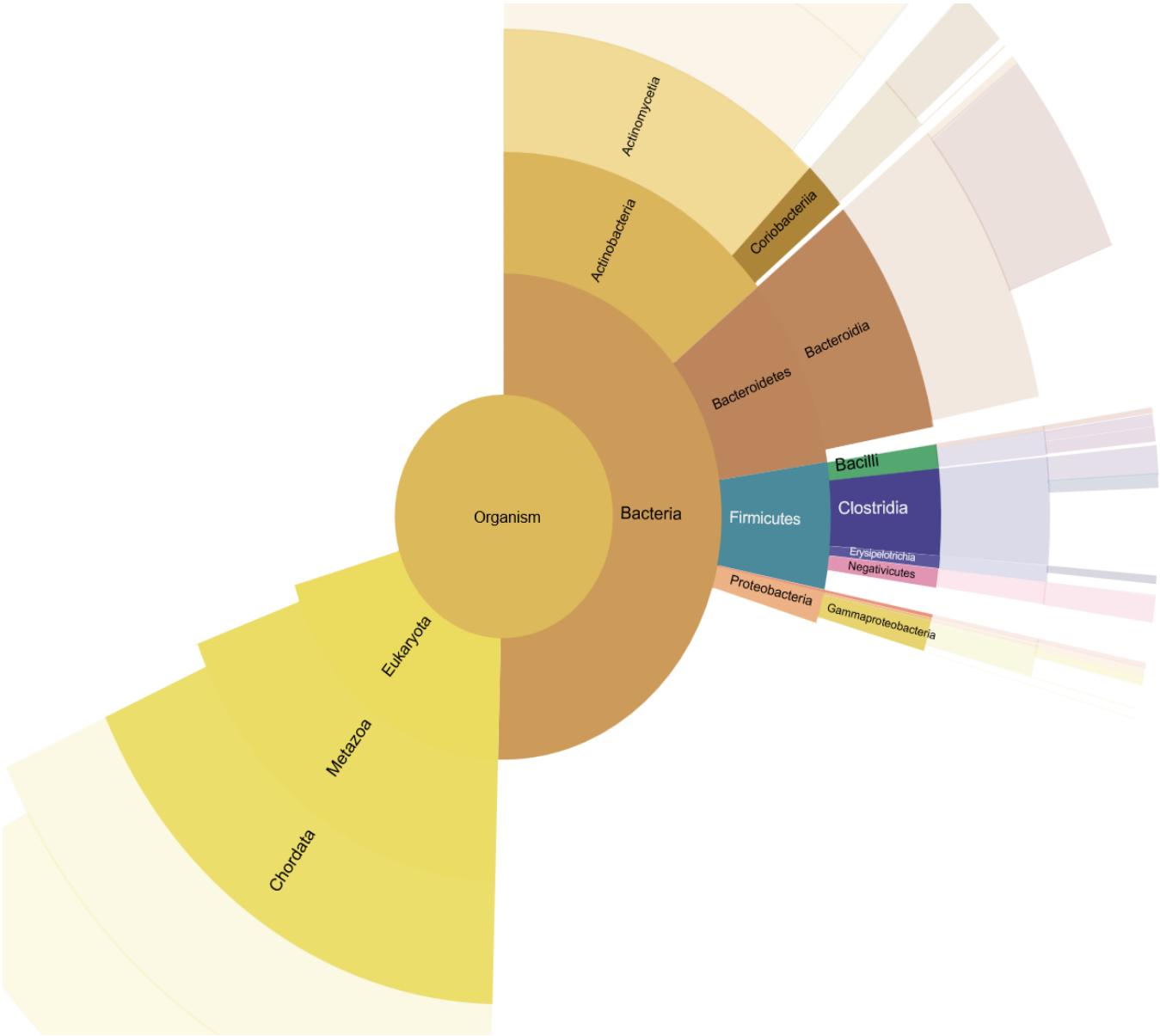
Chapter 3.2 – Composition

To gain a deeper understanding of the samples' composition, Unipept was utilised to classify the metaproteome using the identified peptides. This classification revealed the prominent phyla present in the samples, shedding light on their composition. The primary identified phyla included Actinobacteria, Proteobacteria, Bacteroides, Firmicutes, and human components (Metazoa), as displayed in *Figure 7A and 7B*.

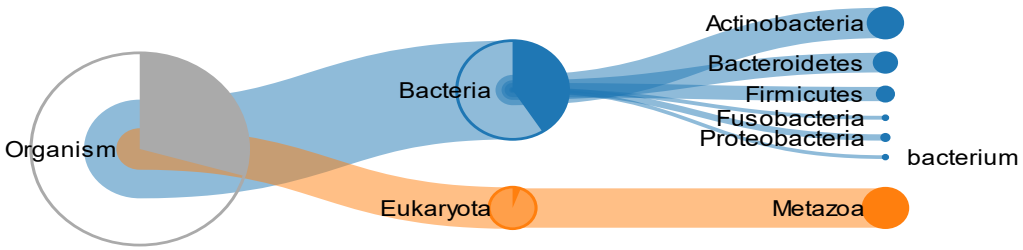
The classified peptides were then categorized into their respective domain, kingdom, phylum, class, order, and family, as shown in *Figure 7B and 7C*. This analysis unveiled the predominant families and their relative representation within the samples, visually represented by the size of the slices. The key families identified were Bifidobacteriales, Coriobacteriales, Bacteroidales, Lactobacillus, Eubacteriales, and Enterobacterales, as shown in *Figure 7C.* This compositional sunburst analysis of identified peptides shows how many bacterial

peptides compared to human peptides were identified. This Unipept analysis does not account for the abundance, rather the identified peptides, so following this analysis the species were assigned to their genus and phylum for quantitative analysis.

A



B



C

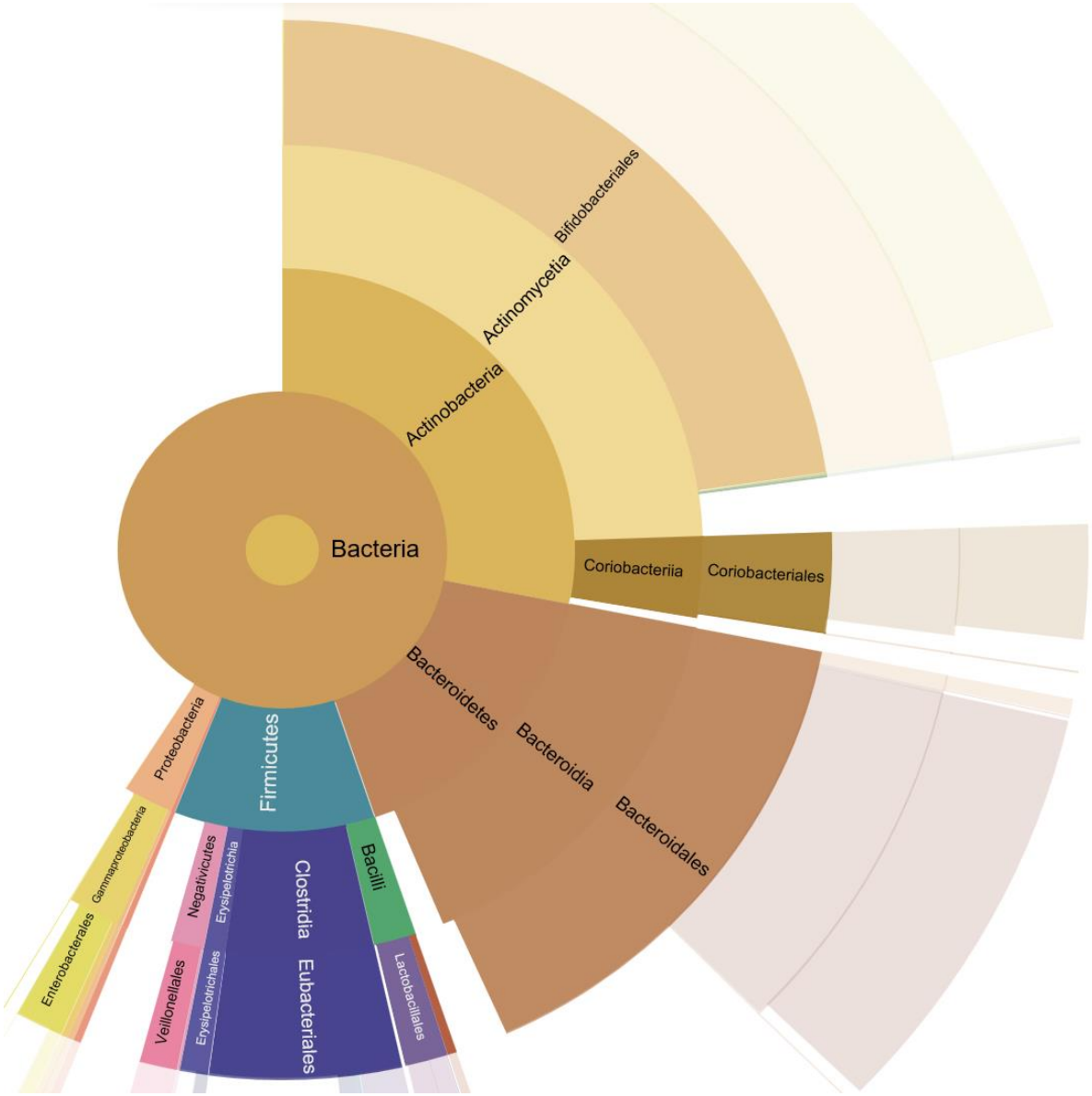


Figure 7: A- *Sunburst figure of peptides identified in Unipept showing the domain, kingdom, phylum and class.*
B- Unipept tree view of peptides origins. Two domains of life Bacteria and Eukaryota. Within bacteria are the main phylum identified.
C- Sunburst figure of bacterial peptides identified in Unipept showing the domain, phylum, class, order and family.

Following these identifications, an overall analysis of bacterial composition was conducted after assigning species to their corresponding phyla, as depicted in *Figure 8A,* based off Unipept analysis. The top phyla identified from the overall bacterial signal were Firmicutes (50%), Actinobacteria (26%), Bacteroides (15%), and Proteobacteria (9%), as shown in *Figure 8A.* The total protein intensity and protein count was determined for both human and bacterial origin, shown in *Figure 8B.* This demonstrates the comparison between the active bacterial proteins from various species showcased in *Figure 8A* and the active proteins in humans

(which represent a single species). It highlights how the developed metaproteomic analysis method in this study not only identifies host proteins but also reliably detects a wide array of bacterial species' proteins with significant sensitivity.

The percentage of signal within each phylum depicted in *Figure 8A* was proportionate to the number of species, as evident in the % composition of signal shown in *Figure 8C*. This correlation allows us to draw the conclusion that the level of bacterial presence can be linked to their activity, as indicated by the protein intensity.

The data was then stratified into the two groups under investigation, HEU and HU, and the relative abundance of the total bacterial signal was visually represented at the phylum level between the two groups, as shown in *Figure 8D*. A relative Actinobacteria total signal in the HEU group was lower than in the HU group, 25 % of the total HEU samples' signal compared to 37 % from the HU group. The Proteobacteria and Firmicutes phyla had similar relative abundances between the two groups, around 4 % and 33 % respectively. While Bacteroides relative signal was much higher in the HEU samples compared to the HU samples, 37 % HEU compared to 25 % HU.

To further explore this, a heatmap of the phylum and human composition was created to assess the abundance of the signal from each sample in the two groups to try identifying any clusters, as seen in *Figure 8E*. This heatmap representing the individual sample signal for each phylum and its human origin displayed a distinct clustering of samples on the far left, characterized by notably high levels of bacterial signal. However, these samples were not lacking in human signal, as seen in *Figure 8F.* While this cluster was intriguing, the reason for this clustering remained unknown. To prevent potential skewing of future sample analysis, a decision was made to exclude this cluster from further analysis.

Statistical analysis conducted in GraphPad Prism aimed to explore the composition of phyla between the two groups by examining individual samples' signals, as shown in *Figure 8F*. This approach helped assess the sample distribution within each group and assess any statistically significant differences in Phyla composition. Despite observing differences in relative abundance for Actinobacteria and Bacteroides levels in *Figure 8D*, the subsequent statistical analysis using GraphPad Prism did not reveal any significant disparities (p-values > 0.05) between the two groups for these phyla, as depicted in *Figure 8F*. Notably, Actinobacteria exhibited a distinct distribution between the groups and the lowest p-value (0.2332), yet it remained insignificant in differentiating the two groups.

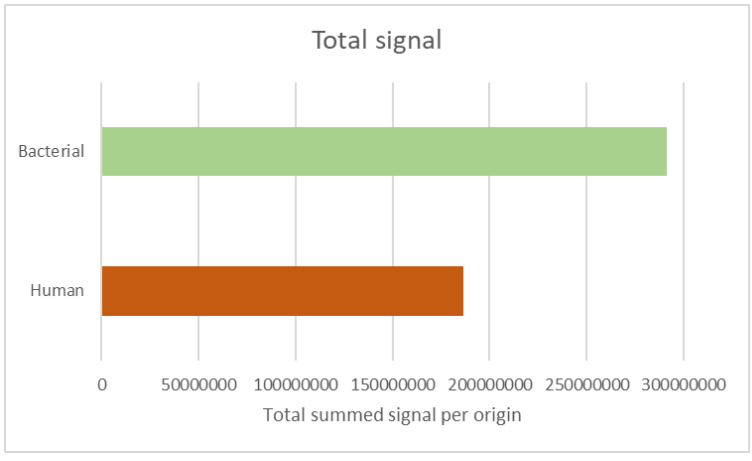
Confirming earlier observations in *Figure 8D*, Proteobacteria displayed a minimal difference with a p-value of 0.9969, indicating a similar composition in both groups. Conversely, Firmicutes stood out by having the second lowest p-value between the two groups, contrasting the observations in *Figure 8D* where similar relative abundances were noted. Finally, Bacteroides demonstrated a comparable sample distribution between the groups, with no significant difference observed and a p-value of 0.658.

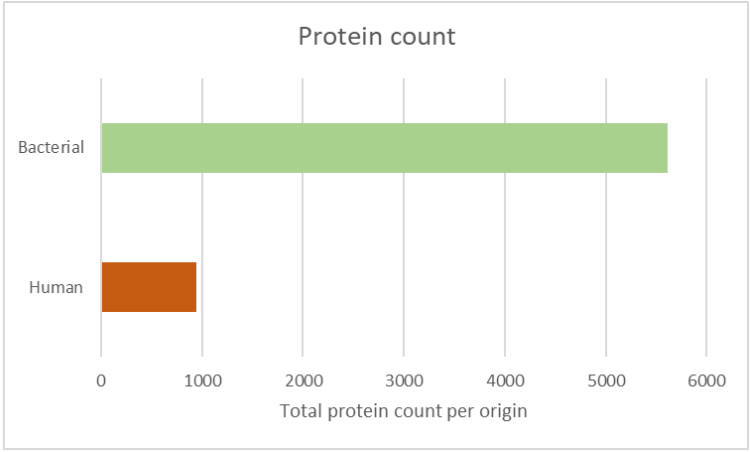
Consequently, the assessment of compositional differences using the total protein intensity per phyla did not yield significant distinctions between the two groups. This merits the need for further analysis of individual protein intensities to identify any potential differences among the infant groups.

A

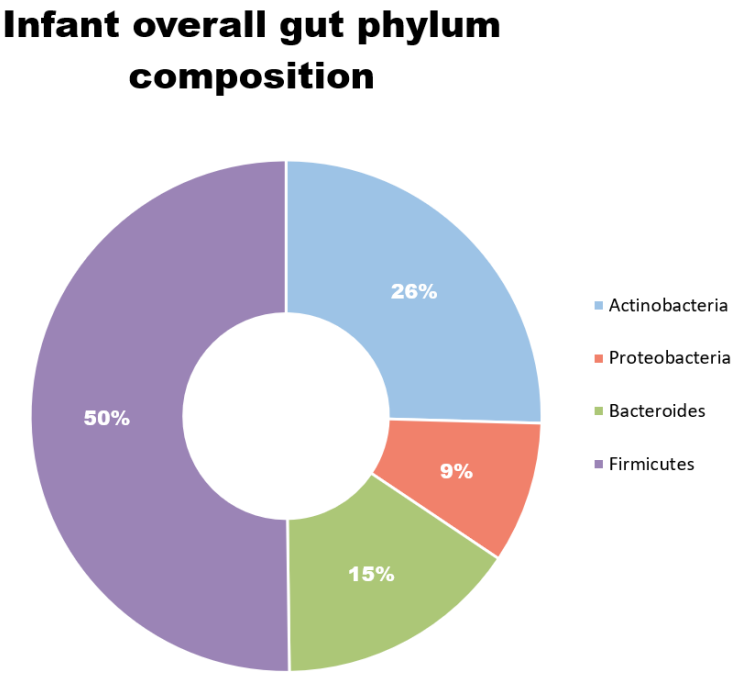
Firmicutes			Actinobacteria	Bacteroides	Proteobacteria
Staphylococcus aureus	Lactococcus lactis	Gemella sp	Bifidobacterium adolescentis	Bacteroides caccae	Escherichia coli
Staphylococcus caprae	Limosilactobacillus	Flavonifractor sp	Bifidobacterium animalis	Bacteroides fragilis	Escherichia marmotae
Staphylococcus epidermidis	Limosilactobacillus fermentum	Ruminococcus gnavus	Bifidobacterium bifidum	Bacteroides ovatus	Succinivibrio dextrinosolvens
Staphylococcus hominis	Limosilactobacillus oris	Ruminococcus torques	Bifidobacterium breve	Bacteroides stercoris	Klebsiella aerogenes
Staphylococcus lugdunensis	Limosilactobacillus vaginalis	Agathobacter ruminis	Bifidobacterium catenulatum	Bacteroides thetaiotaomicron	Klebsiella pneumoniae
Clostridium innocuum	Enterococcus faecalis	Dialister pneumosintes	Bifidobacterium dentium	Bacteroides uniformis	Klebsiella quasipneumoniae
Clostridium butyricum	Blautia obeum	Dolosigranulum pigrum	Bifidobacterium kashiwanohense	Prevotella bivia	Klebsiella variicola
Clostridium perfringens	Blautia pseudococcoides		Bifidobacterium longum	Prevotella copri	Haemophilus paracuniculus
Erysipelatoclostridium ramosum	Megamonas hypermegale		Collinsella aerofaciens	Prevotella veroralis	Haemophilus parainfluenzae
Faecalibacterium prausnitzii	Megasphaera elsdenii		Collinsella sp.	Parabacteroides distasonis	Haemophilus pittmaniae
Faecalibacterium sp	Megasphaera hexanoica		Corynebacterium amycolatum	Parabacteroides merdae	Shigella flexneri
Streptococcus agalactiae	Phascolarctobacterium faecium		Corynebacterium propinquum	Phocaeicola dorei	
Streptococcus anginosus	Veillonella atypica		Corynebacterium tuberculostearicum	Phocaeicola vulgatus	
Streptococcus iniae	Veillonella criceti		Olsenella sp.	Alistipes onderdonkii	
Streptococcus lutetiensis	Veillonella dispar		Olsenella uli		
Streptococcus oralis	Veillonella parvula		Slackia isoflavoniconvertens		
Streptococcus parasanguinis	Catenibacterium sp.		Parolsenella catena		
Streptococcus peroris	Subdoligranulum variabile		Actinomyces ruminicola		
Streptococcus salivarius	Solobacterium moorei		Senegalimassilia anaerobia		
Lachnoclostridium sp.	Romboutsia ilealis		Rothia mucilaginosa		
Lactobacillus gasseri	Holdemanella porci		Cutibacterium acnes		
Lactobacillus iners	Granulicatella sp		Enterorhabdus sp.		

B

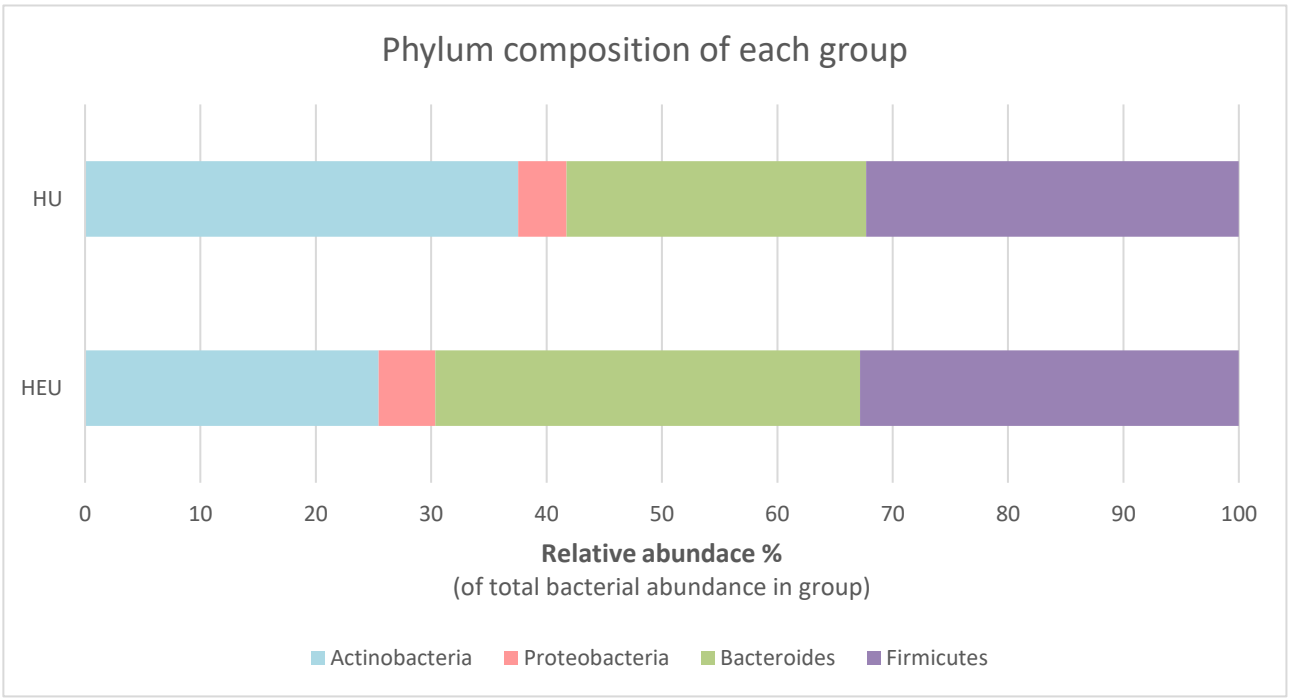




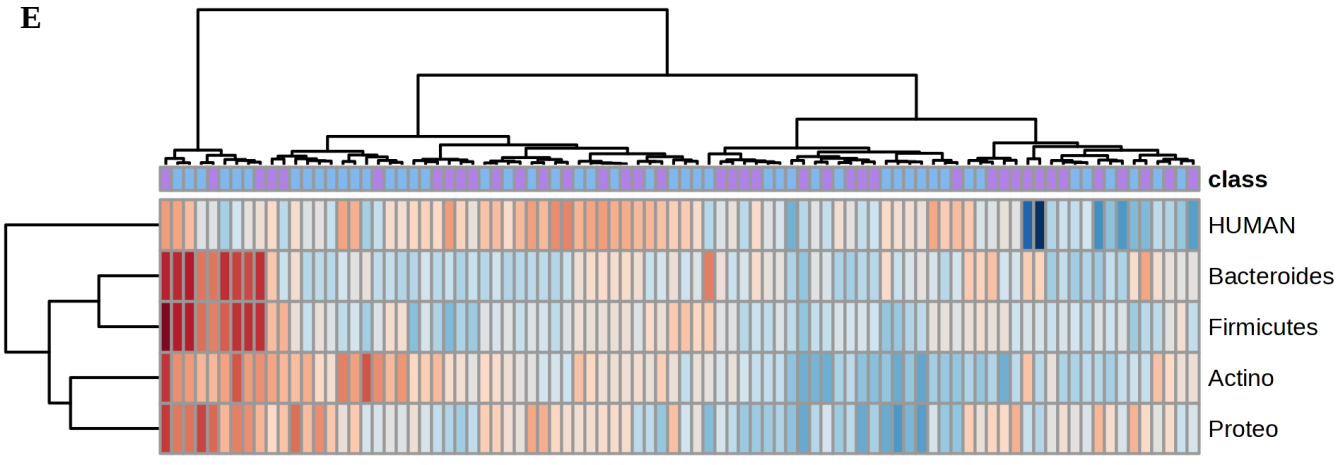
C


Infant overall gut phylum composition
26%
9%
15%
50%
Actinobacteria
Proteobacteria
Bacteroides
Firmicutes

D


Phylum composition of each group
HU
HEU
0
10
20
30
40
50
60
70
80
90
100
Relative abundace %
(of total bacterial abundance in group)
Actinobacteria
Proteobacteria
Bacteroides
Firmicutes

E


class
HUMAN
Bacteroides
Firmicutes
Actino
Proteo

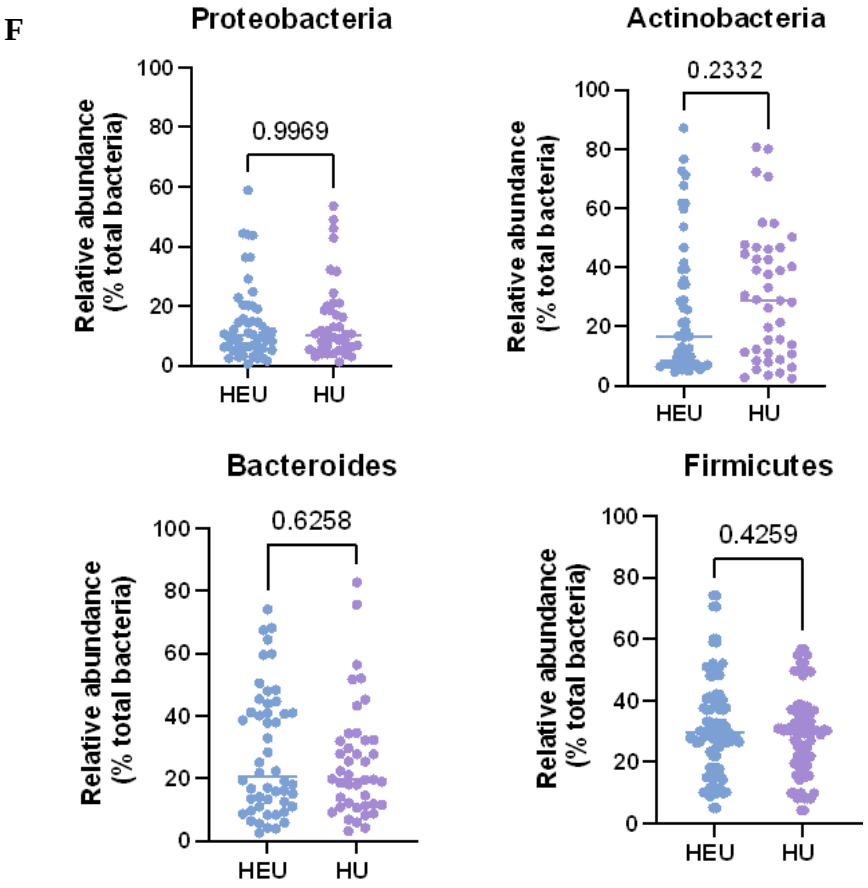


Figure 8: Compositional analysis of infant stool samples.
A- A list of the species identified in each phylum in all the samples.
B- Bar charts showing the total protein intensity and protein count for human and bacterial proteins.
C- Donut pie chart showing the phylum composition of the total signal of all the samples' compositions.
D- Stacked bar chart comparing the relative abundance of the phylum from the total bacterial signal between the two groups, HEU and HU.
E- Heatmap showing the phylum abundance in each sample. Highlighting a distinct clustering with high abundances.
F- Mann-Whitney two tailed t-test performed on each phylum's relative abundance of the phylum % of the total bacteria signal between the two groups, HEU and HU. No significance was seen in any phylum ($p > 0.05$), with similar distribution of samples seen in each phylum between the groups. Made in *GraphPad prism 10.0.*

Following the phyla compositional analysis, a diversity assessment between the 2 groups on a species level was conducted, Figure 9. The successful identification of all taxa used to create the in-silico DIA-NN library based off the provided 16S data validates this comprehensive metaproteomic method, showcasing its ability to provide an effective analysis of the microbial composition within the gut microbiome (see identified species in Figure 8A). After the taxa identification was assessed, the 1st diversity assessment was conducted between the two groups, evaluating the diversity within each group through alpha diversity measurements, using the Shannon index. The comparison of alpha diversity between the two groups was performed using a Mann-Whitney t-test, as depicted in Figure 9A. It is worth noting that alpha diversity quantifies the species

richness within the host, while beta diversity compares the community structures between hosts. The analysis revealed no statistically significant difference in alpha diversity between the two groups.

Subsequently, beta diversity was examined to assess the diversity of samples in the two groups, using the Bray index, as illustrated in *Figure 8B.* Once again, this analysis did not show any significant differences between the two groups. These results prompted further investigation into the protein composition of the samples.

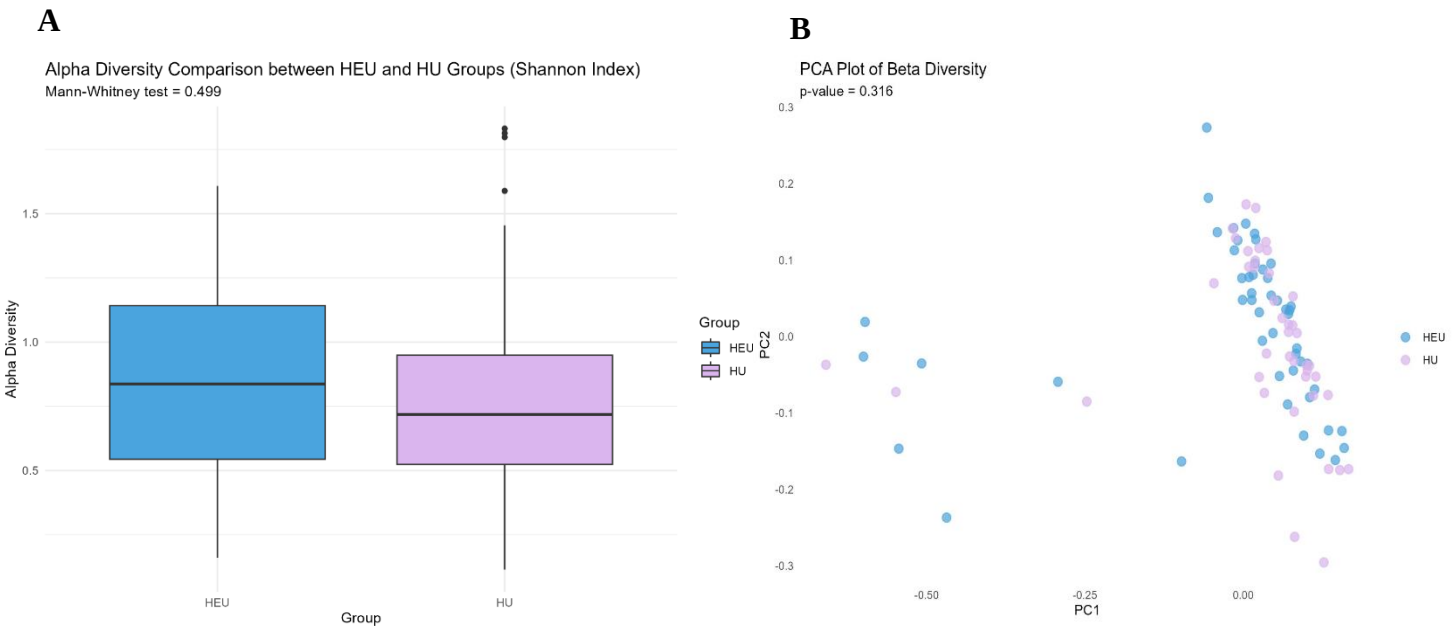


Figure 9: Diversity analysis of sample composition between the 2 groups, HEU and HU.
A- Box and whisker of Alpha diversity of samples between the two groups, HEU and HU, using the Shannon Index comparing the two groups. Wilcox test performed between the 2 diversities resulting in insignificant difference between the two groups as p-value > 0.05.
B- PCA plot of Bray index beta diversity of the samples comparing the sample diversity between the two groups, HEU and HU. Beta diversity between the two groups was insignificant as p > 0.05.

3.2.1 Composition discussion

There is no one-size-fits-all gut microbiome, as each person's unique genetic and environmental factors dictate the specific needs of their gut microbiome [94]. A balanced microbiome is essential for optimal metabolic and immune functions, contributing to overall host health and disease prevention [94]. Commensal bacteria play a pivotal role in disease prevention, nutrient processing (including the digestion of insoluble fibres), nutrient breakdown, and the production of metabolites essential for optimal nutrition, energy efficiency, and storage [51, 56, 94]. This balance is maintained when symbiotic and commensal bacteria outcompete pathogenic strains for resources [56]. Conversely, dysbiosis in the microbiome can lead to pathogenic strains gaining the upper hand, potentially resulting in inflammation and disease [56]. Importantly, dysbiosis during infancy and early development can have lasting health implications. To assess

dysbiosis compared to a healthy microbiome diversity, we conducted alpha and beta diversity analyses with results presented in *Figure 9,* showing no perturbations in gut diversity. This finding is further supported by Jackson et al, who conducted a genomic study on 240 mother-infant pairs with HEU and HU infants, revealing no differences in the alpha or beta diversity of the gut microbiomes [95].

The composition of the gut microbiome, as depicted in *Figures 7*, highlights the presence of key phyla and families. These top-identified phyla align with the existing literature, with Actinobacteria, Proteobacteria, Firmicutes, and Bacteroides being the predominant gut phyla [51, 56, 94]. Contrary to *Figure 8*, which indicates no significant difference in phylum composition between the groups, *Jackson et al* found higher Firmicutes abundances and lower Proteobacteria in HEU infants [95]. Additionally, the top families in the gut within these four main phyla are as follows: Bifidobacterium from Actinobacteria, Lactobacillus, Ruminococcus, Clostridiaceae, and Staphylococcus from Firmicutes, Bacteroides, Prevotella, and Alistipes from Bacteroides, and Escherichia, Bilophila, and Helicobacter from Proteobacteria [56]. These genera were positively identified, as seen in *Figure 7*.

Various factors influence the gut microbiome, including BMI, antibiotic use, exercise, and disease states [51, 94]. Additionally, during infancy, significant influences on microbiome composition arise from factors such as delivery mode, gestational age, feeding methods, and maternal influence, including microbial transfer from the mother [94-96]. Different genera have been associated with various delivery modes and feeding methods. Vaginally born infants tend to have higher proportions of Lactobacillus and Prevotella, while caesarean births are dominated by Clostridium, Staphylococcus, Streptococcus, and Corynebacterium (many to be known skin microbes) [51, 96]. As the primary focus of our study was the effect of HIV on infants, the data on delivery modes was insufficient for a proper analysis compared to the existing literature (see appendix N). Live bacteria in breast milk include Bifidobacterium, Staphylococcus, Streptococcus, and Lactobacillus, all of which play a vital role in seeding the infant gut for optimal health [96]. These were positively identified in the infants' stool samples*, Figure 7*, and suggest the influence of breastmilk composition on the infants' gut microbiomes. Further to this, breastfed infants tend to be rich in Bifidobacterium, which assists in human milk oligosaccharide digestion [96].

Another factor affecting the gut microbiome is HIV infection, as it is associated with gut dysbiosis in HIV-positive individuals [94, 97]. *Liu et al* found a greater abundance of Prevotella and a decrease in Bacteroides at the genus level in the guts of HIV-positive individuals [94]. *Ishizaka et al* also observed increased Prevotella abundance in the gut, with lower CD4 counts correlating with Actinobacteria levels and decreased alpha diversity associated with low CD4 counts [97, 98]. HIV-positive individuals tend to have a Prevotella-rich enterotype, characterized by increased inflammatory cytokines and decreased anti-inflammatory cytokines [97]. Unfortunately, our study lacked the metadata for samples, preventing us from drawing associations related to maternal CD4 counts and HEU infants' gut microbial composition.

Moore et al referred to the TEDDY study, which assessed 903 stool samples from infants aged 3 months and older, highlighting the lack of data on 1-week-old infants, and its effects on the microbiome from early infancy [95]. While *Figure 8* did not show any differences in Firmicutes between HEU and HU groups, further exploration of Firmicutes is warranted, as *Jordan et al* demonstrated that Firmicutes have systemic immune control without activating host inflammation [96]. Bacilli and Clostridia from the Firmicutes phylum were associated with the ability to induce a cytokine response to fight disease and aid in immune responses [98]. The assessment in *Figure 7, 8 and 9* focused on what is present in the gut rather than what is active and contributing to gut functioning, prompting us to perform a proteomic analysis to gain functional insights into the host and microbiome.

Chapter 3.3 – Proteins

Both host and microbiome contribute to the differences seen in HEU infants.

A protein-level analysis was conducted using Metaboanalyst, following data filtering and processing in R Studio and Perseus. To ensure accurate results, all identified proteins were filtered for a minimum sample presence of 10 %, thereby removing any individuality of the samples while keeping proteins that could be biologically significant from the metaproteome.

Due to the missing data in the samples, a supervised learning algorithm was applied to determine the most distinct proteins between the two groups. The out-of-bag (OOB) error for the top 10% of proteins was found to be 0.304 (*Table 2*), indicating a strong classification accuracy of the model. This result instilled confidence in the classifiers identified between the two groups.

The top most discriminating proteins identified by the Random Forest regression model analysis included both human and bacterial proteins, contributing to the differentiation between the two groups. These top 20 classifiers from the model are detailed in *Table 4*, with the top 10 classifiers visually represented in a heatmap in *Figure 10*. The human and bacterial proteins are almost equally represented in the top 20 classifying proteins between the groups, 55 % from human proteins and 45 % from bacterial proteins (*Figure 10*). The identification of both human and bacterial classifiers in near equal abundances highlights the synergistic relationship that exists between the host and the microbiome.

In *Figure 10*, the heatmap reveals the presence of two distinct clusters: the left cluster follows the trend of the HU group, while the right cluster predominantly consists of HEU samples. Notably, the HEU cluster exhibits a higher protein abundance compared to the HU cluster. This observation may reflect the intricate interplay between the host and the microbiome, with both bacterial and human proteins increasing together (*Figure 10*).

This interesting finding warrants a deeper assessment into the contributions of the host and the microbiome, separately, towards the difference between HEU and HU infants' proteomes. This separation will aid our

understanding of the distinct host and microbial responses. It is important to note that the microbial data contains a substantial amount of missingness when compared to the human data (see Appendix L showing the bacterial protein missingness). To avoid skewing the analysis, it is important to conduct separate investigations. An example of this disparity in missingness can be seen in the heatmap, *Figure 10*, with reference to proteins A0A1G9WT85 and A0A0F2CTR5.

Table 5: List of top discriminating proteins between the two groups following Random Forest model analysis, highlighting the proteins' origin.

UNIPROT Protein ID	***Protein***	***Mean decreased accuracy***	***Origin***	***Organism***
Q6UWV6	ENPP7	0.0024	Human	Homo sapiens
A0A1G9WT85	purD	0.0022	Bacteria	Actinomyces ruminicola
A0A0F2CTR5	PGAM	0.0021	Bacteria	Streptococcus oralis
P20848	SERPINA2	0.0019	Human	Homo sapiens
A0A822WI55	ompA	0.0017	Bacteria	Enterobacter hormaechei
Q13939	CCIN	0.0016	Human	Homo sapiens
A0A0E2AW47	PpiC	0.0011	Bacteria	Bacteroides fragilis
A0A075B7B6	IGHV	0.0010	Human	Homo sapiens
P08697	SERPINF2	0.0010	Human	Homo sapiens
Q8G5N0	lacZ	0.0010	Bacteria	Bifidobacterium longum
P0DP23	CALM1	0.0010	Human	Homo sapiens
A6L4M4	atpA	0.0010	Bacteria	Phocaeicola vulgatus
H3KH51	FbP	0.0009	Bacteria	Sutterella parvirubra
Q9UHL4	DDP7	0.0009	Human	Homo sapiens
P05543	SERPINA7	0.0009	Human	Homo sapiens
B8DVV6	Idh	0.0009	Bacteria	Bifidobacterium animalis lactis
A8K7I4	CLCA1	0.0009	Human	Homo sapiens
A0A0B4J1V0	IGHV	0.0009	Human	Homo sapiens
A0A075B6P5	IGKV	0.0009	Human	Homo sapiens
P0A953	fabB	0.0008	Bacteria	Escherichia coli

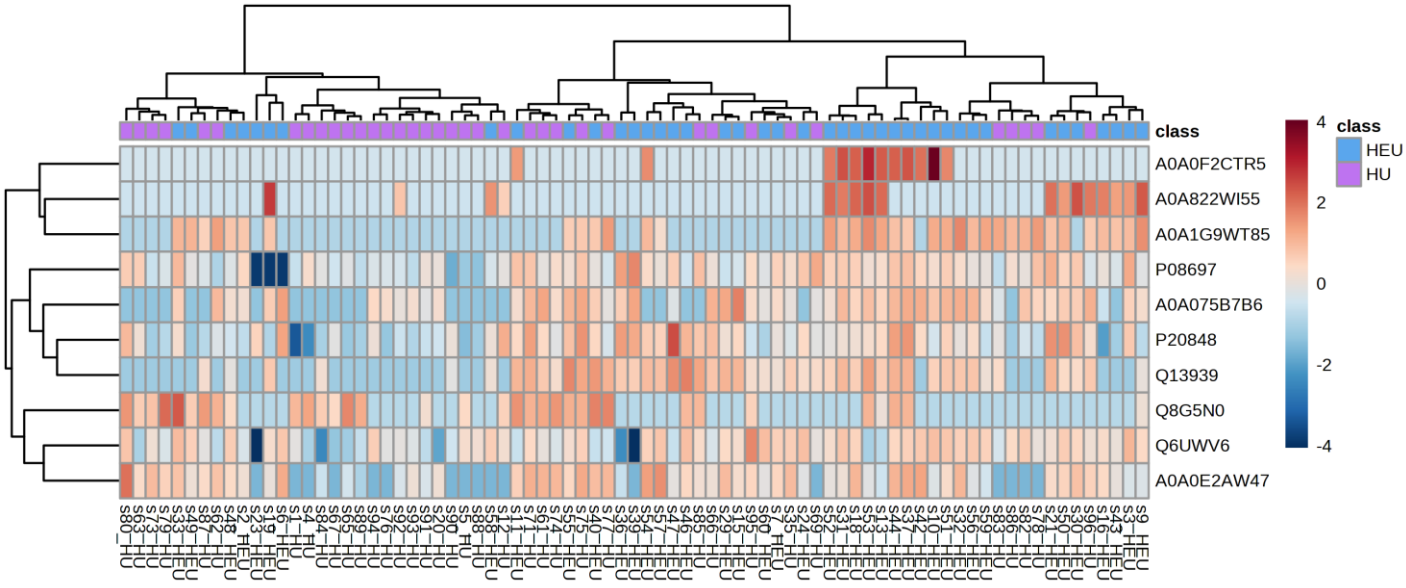


Figure 10: Top 10 classifying proteins from Figure 12 using the Random Forest model. Heatmap created in Metaboanalyst 5.0.

3.3.1 - Protein discussion

The gut microbiota is critically involvement in host physiology, metabolic activities, and immune system development, as well as its role in protecting against pathogenic bacteria, has been extensively discussed [99]. This dynamic interaction is characterized by the bidirectional influence between the microbiome and host gene expression, regulating pathways crucial for disease, including immune development and energy metabolism [100]. The evidence presented in *Table 4* and *Figure 10* points out the joint contribution of both the host and bacteria to the observed health differences in HEU infants, aligning with existing literature on this symbiotic relationship.

An imbalance of the normal gut microbiota has been linked to gastrointestinal conditions like IBD and IBS [101]. Additionally, the gut microbiome has been linked to other systemic disease states such as allergies, obesity, type 2 diabetes, and brain development [99, 101]. Highlighting the involvement of the gut microbiome with disease, within gut and beyond the gut.

Nicholas et al's discusses the profound impact of the microbiome on the host genome, elucidating mechanisms such as chromatin remodelling, differential splicing, alteration of the epigenetic landscape, and interruption of host signalling cascades [100]. The genetic alterations induced by these processes may lead to unfavourable functions in the host, compelling a proteomic investigation to assess the active functions happening in the gut and their impact on the host. The effected genome, believed to result from microbial-derived stimuli and small molecules, highlights the importance of studying the microbiome separately to account for distinct bacterial functional communities, providing crucial insights into the functional capabilities of the gut microbiome [100]. The results of human and bacterial proteins being equally important as classifiers between the two groups, supports the discussion of *Nicholas et al* in the bacterial synergistic dynamic with the host .

Notably, various bacterial species within these microbiotas contribute to the production of Short-chain fatty acids (acetate, propionate, butyrate, etc.), which play pivotal roles in maintaining a healthy gut environment. These functions include pH regulation, pathogen inhibition, water and sodium absorption stimulation, cholesterol synthesis participation, and energy provision to colonic epithelial cells [102]. Particularly relevant to this study is the impact of Bifidobacterium in the infant gut, producing Choline metabolites that modulate lipid metabolism and glucose homeostasis [102]. Examining the proteomes of both the host and microbiome separately becomes necessary to assess perturbations in host gene expressions and gain a comprehensive understanding of the pathways in which the microbiome is involved.

Another example of microbiome and host interaction from *Fu et al* highlights the microbiome's role in regulating detoxification pathways within the host, demonstrated by the decreased expression of intestinal Cytochrome P450 enzymes and transporter genes in germ-free mice [103]. This substantiates the importance of exploring the intricate interplay between the microbiome and host at a proteomic level, providing valuable insights into the molecular mechanisms underlying disease etiology and designing microbiome-targeted therapeutics [103].

As demonstrated by numerous studies on genomic effects and alterations from microbial influence, it is evident the microbiome plays a pivotal role in influencing host gene expressions at a proteomic level, and this level of analysis could help reveal any molecular mechanisms underlying disease etiology. Therefore, understanding the balance and synergy that exists between the host and microbiome could provide valuable insight to aid in the understanding of the complexities of human health. This holistic approach in analysing the functionality of the gut through the host metaproteome could provide evidence for potential therapeutic targets for health interventions.

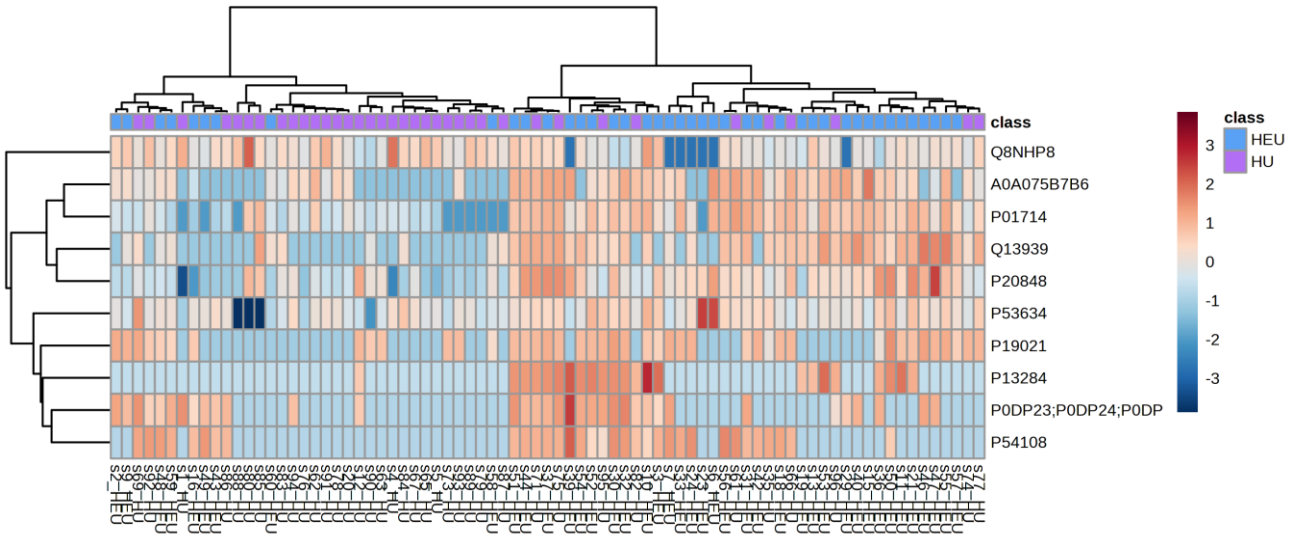
Chapter 3.4 – Human proteome

Both infant gut proteins and maternal immune proteins contribute to differences seen in HEU infants.

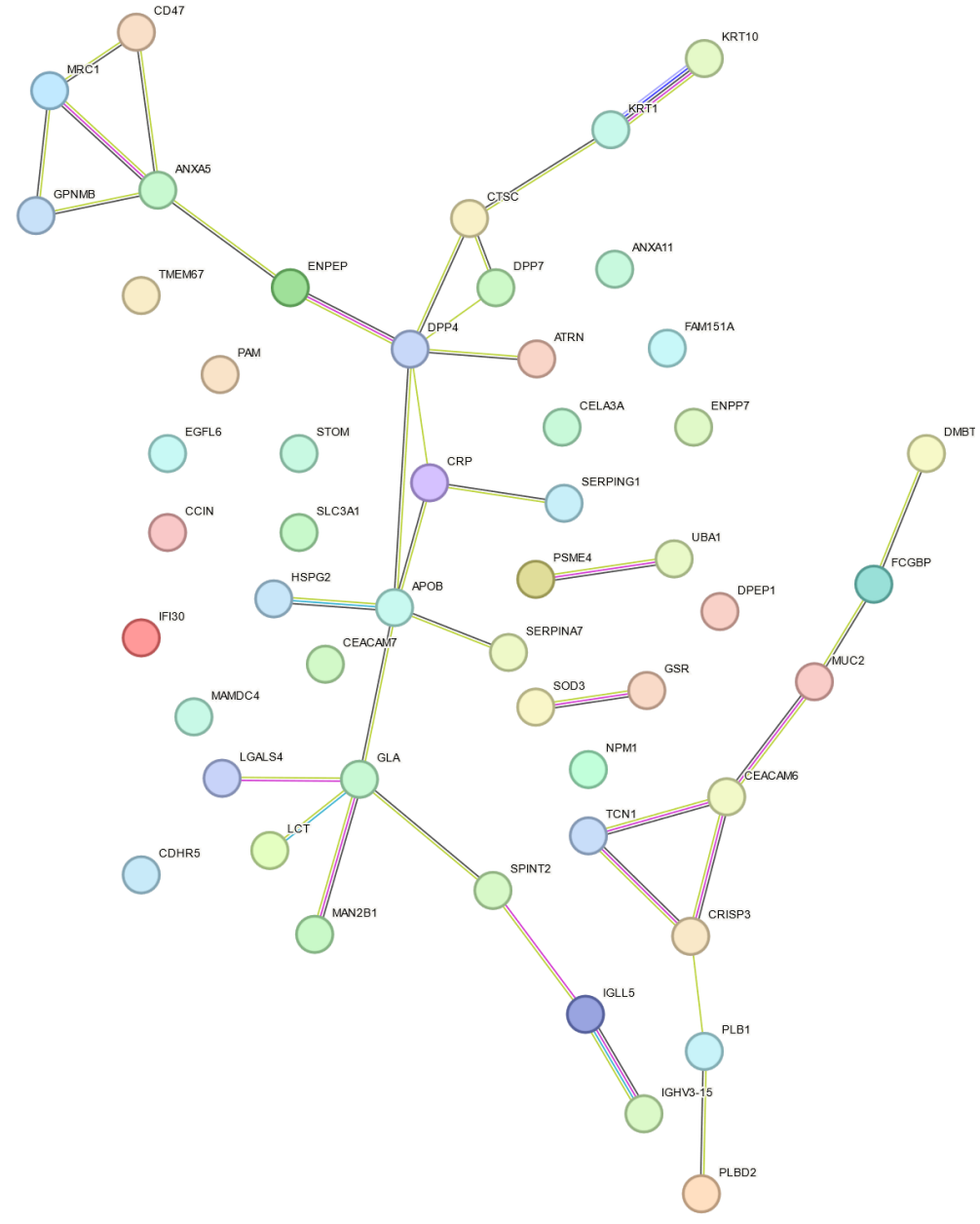
The application of the random forest regression model to human proteins yielded an Out-of-Bag (OOB) error of 0.304 when applied to the top 10% of classifiers from the dataset, as outlined in *Table 2*. In *Figure 11A*, the top 10 human proteins from the 10 % top classifiers revealed two clusters in the heatmap, with the left cluster showing lower protein abundance following a HU trend and the right cluster indicating higher protein abundance and HEU dominance. The proteins were mapped in String DB, revealing their Reactome and tissue expressions which are visualised in *Figure 11B and 11C*. The identified proteins were associated with gut origin (infant-derived proteins) and immune system processes (maternal origin), considering the dependence of 1st-week infants on maternal immunity. These proteins identified and their respective system associations are presented in *Figure 11C*. Notably, tissues with significant enrichment included colon, large intestine, plasma cell, GI tract, immune system, and innate immune system tissues and reactomes. The immune associated Reactomes identified had low False discovery rates as seen in *Figure 11C*, emphasizing the reliability of the assignments. The lowest False discovery rates seen for Tissues identification was among the Digestive Gland, Bone marrow, Plasma cell, Intestine and Colon, showcasing the GI tract and immune associated protein differences seen between the HEU and HU groups (*Figure 11C).*

These results revealed the deep insight a stool proteomics analysis can reveal from maternal origin as well as infant response. Thus, it was decided to split the human dataset into infant derived proteins to assess the host response, as well as into breastmilk proteins to gain insight into maternal perturbations. To achieve this protein division, a comprehensive list from literature of identified proteins from breastmilk samples across various geographical locations was compiled [90-92]. Subsequently, any human proteins from this literature list that matched our identified list were categorized as human breastmilk proteins, while the remaining proteins were assigned as infant derived.

A


class
Q8NHP8
A0A075B7B6
P01714
Q13939
P20848
P53634
P19021
P13284
P0DP23;P0DP24;P0DP
P54108
class
HEU
HU
3
2
1
0
-1
-2
-3

B


CD47
MRC1
KRT10
KRT1
ANXA5
GPNMB
CTSC
ANXA11
TMEM67
ENPEP
DPP7
DPP4
ATRN
FAM151A
PAM
CELA3A
ENPP7
EGFL6
STOM
DMBT
CRP
SERPING1
CCIN
SLC3A1
UBA1
PSME4
FCGBP
HSPG2
APOB
DPEP1
IFI30
SERPINA7
CEACAM7
MUC2
GSR
SOD3
MAMDC4
NPM1
LGALS4
GLA
CEACAM6
TCN1
LCT
CDHR5
SPINT2
MAN2B1
CRISP3
IGLL5
PLB1
IGHV3-15
PLBD2

C

Category	Term ID	Term description	False discovery rate
Reactome	*HSA-168249*	Innate Immune System	8.95E-05
Reactome	*HSA-6798695*	Neutrophil degranulation	0.00044
Reactome	*HSA-168256*	Immune System	0.00091
TISSUES	*BTO:0000345*	Digestive gland	2.28E-07
TISSUES	*BTO:0000511*	Gastrointestinal tract	2.80E-05
TISSUES	*BTO:0000141*	Bone marrow	3.47E-05
TISSUES	*BTO:0000392*	Plasma cell	0.00011
TISSUES	*BTO:0000648*	Intestine	0.00026
TISSUES	*BTO:0000706*	Large intestine	0.00056
TISSUES	*BTO:0000269*	Colon	0.00079
TISSUES	*BTO:0000570*	Hematopoietic system	0.0033
TISSUES	*BTO:0000574*	Hematopoietic cell	0.0194
TISSUES	*BTO:0000886*	Mucosa	0.0194

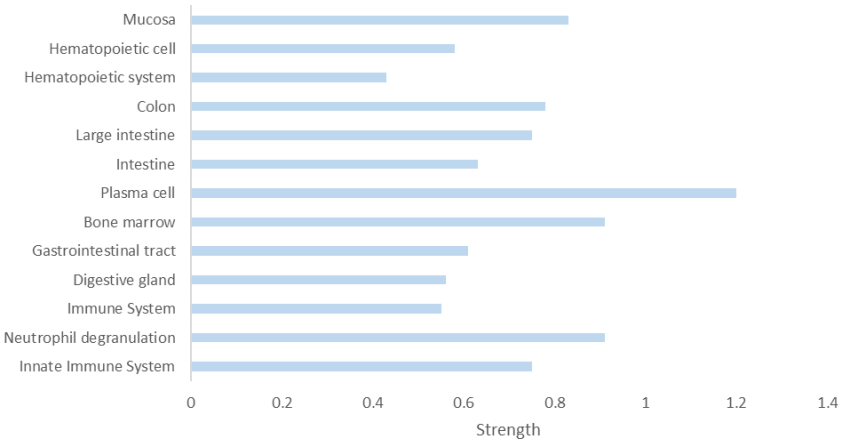


Figure 11: Human proteome analysis from the Random Forest model.

A- Top 10 distinguishing human proteins between HEU and HU from Random Forest model. Heatmap generated in Metaboanalyst 5.0.

B- STRING DB enriched proteins from top 10 % of Random Forest proteins for all human proteins.

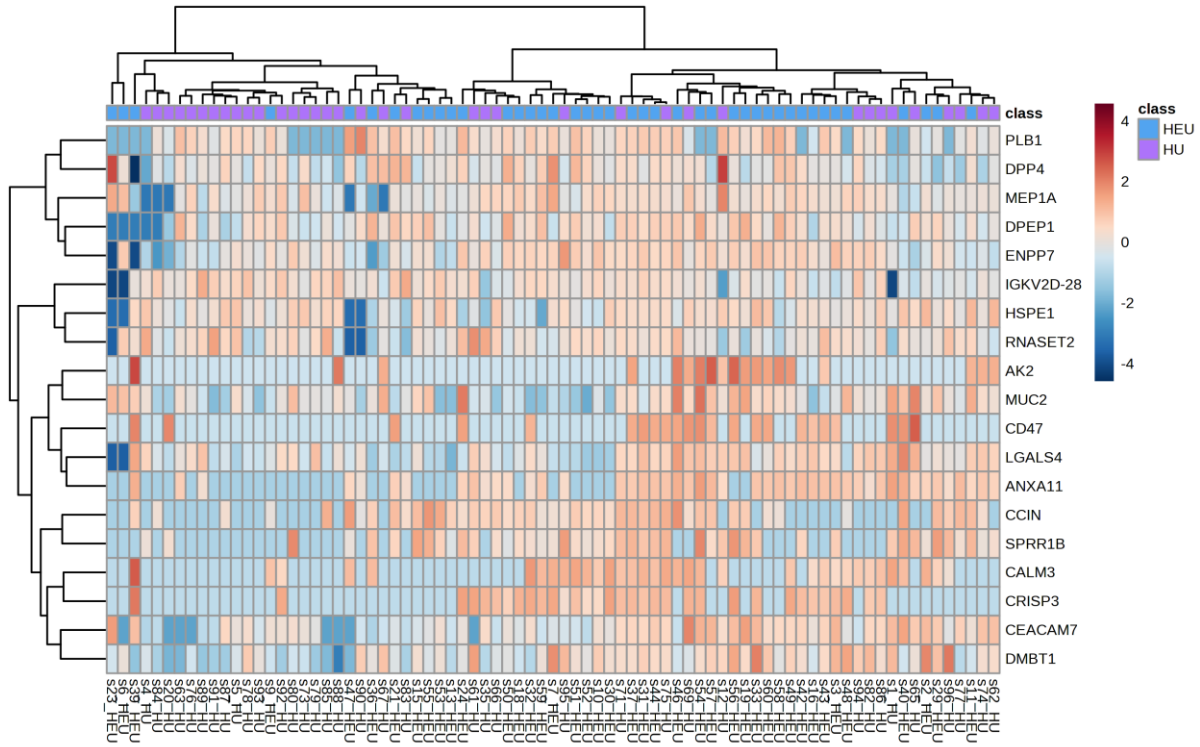
C- String DB enriched terms from top 10 % random forest human proteins. Each enriched term from the human proteins with the strength of enrichment shown. Log10 (observed/ expected) to show strength of enrichment.

Gut barrier MUC2 protein and inflammation is altered in HEU infants.

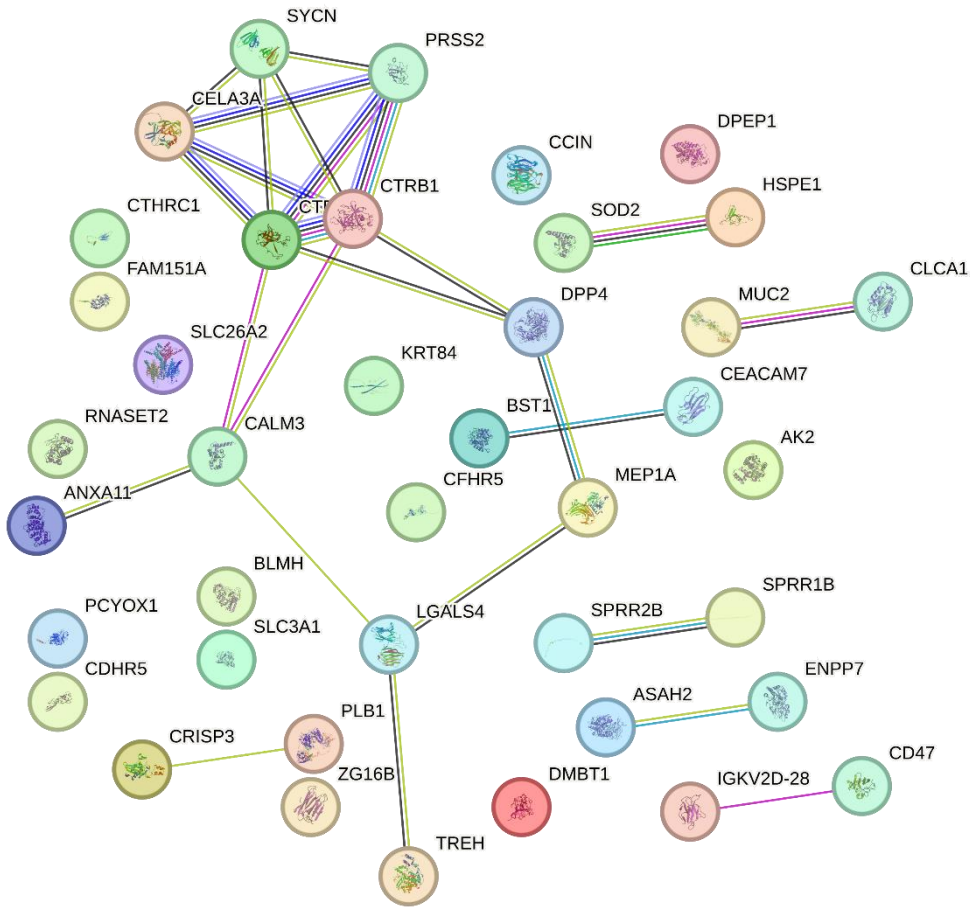
The top infant-derived proteins, presented in the heatmap in Figure 12A, displayed two clusters, with the HEU-dominant cluster showing higher protein abundance than the HU-dominant cluster. These top 10% protein classifiers were mapped in String DB and analysed in Enrichr and Appyter, with results shown in *Figure 12. Figure 12B* demonstrated a strong network with many associations existing between the top proteins and String enrichment analysis in gut processes and metabolism. The Appyter analysis, illustrated in *Figure 12C and D*, revealed significant GO Biology terms, with the most interesting being the involvement of MUC2 in the regulation of gut microbiota, particularly upregulated in the HEU-dominant cluster *(Figure 12A*). Three enriched processes related to inflammation were identified, involving genes DPEP1, CD47, and DPP4 (*Figure 12C and 12D).* The upregulation of DPP4 and CD47 in the HEU cluster, when assessing the heatmap in *Figure 12A,* indicated increased inflammation from TNF cytokine production in these infants' guts. Furthermore, DPEP1's stronger association with HEU infants suggested a potential role in gut inflammation via response to Nitric Oxide and Reactive Nitrogen Species.

Another interesting GO term enriched in the infant proteins, was regulation of lipid transport via ENP77, which plays a role in breastmilk metabolism from the lipid component (*Figure 12D*).

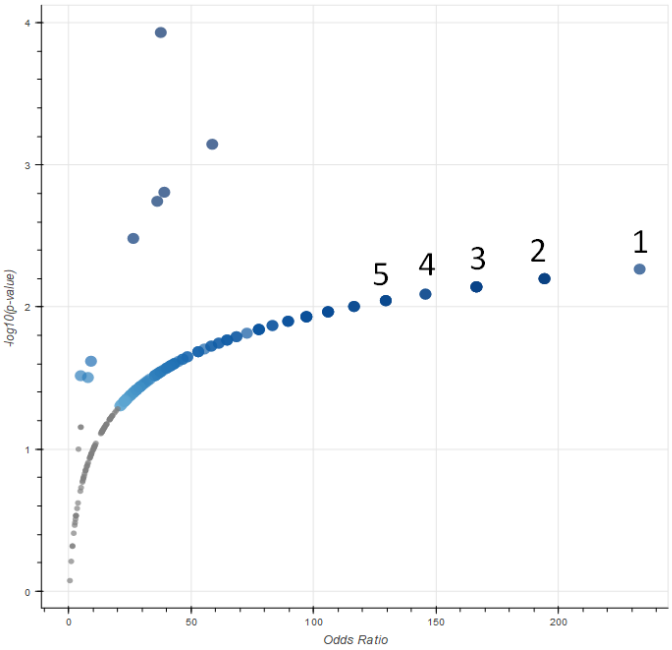
A



B



C



	GO Biological Process	Gene
1	Heterocycle Catabolic Process (Go:0046700)	DPEP1
2	Host Mediated Regulation of Intestinal Microbiota Composition (GO:0048874) Receptor-Mediated Virion Attachment to Host Cell (GO:0046813) Regulation Of Intestinal Cholesterol Absorption (GO:0030300)	MUC2 DPP4 ENPP7
3	Cellular Response to Nitric Oxide (GO:1902170)	DPEP1
4	Regulation Of Tumour Necrosis Factor Superfamily Cytokine Production (GO:1903555)	CD47 & DPP4
5	Homocysteine Metabolic Process (GO:0050667) Regulation Of Cell-Cell Adhesion Mediated by Integrin (GO:0033632)	DPP4 CD47

D

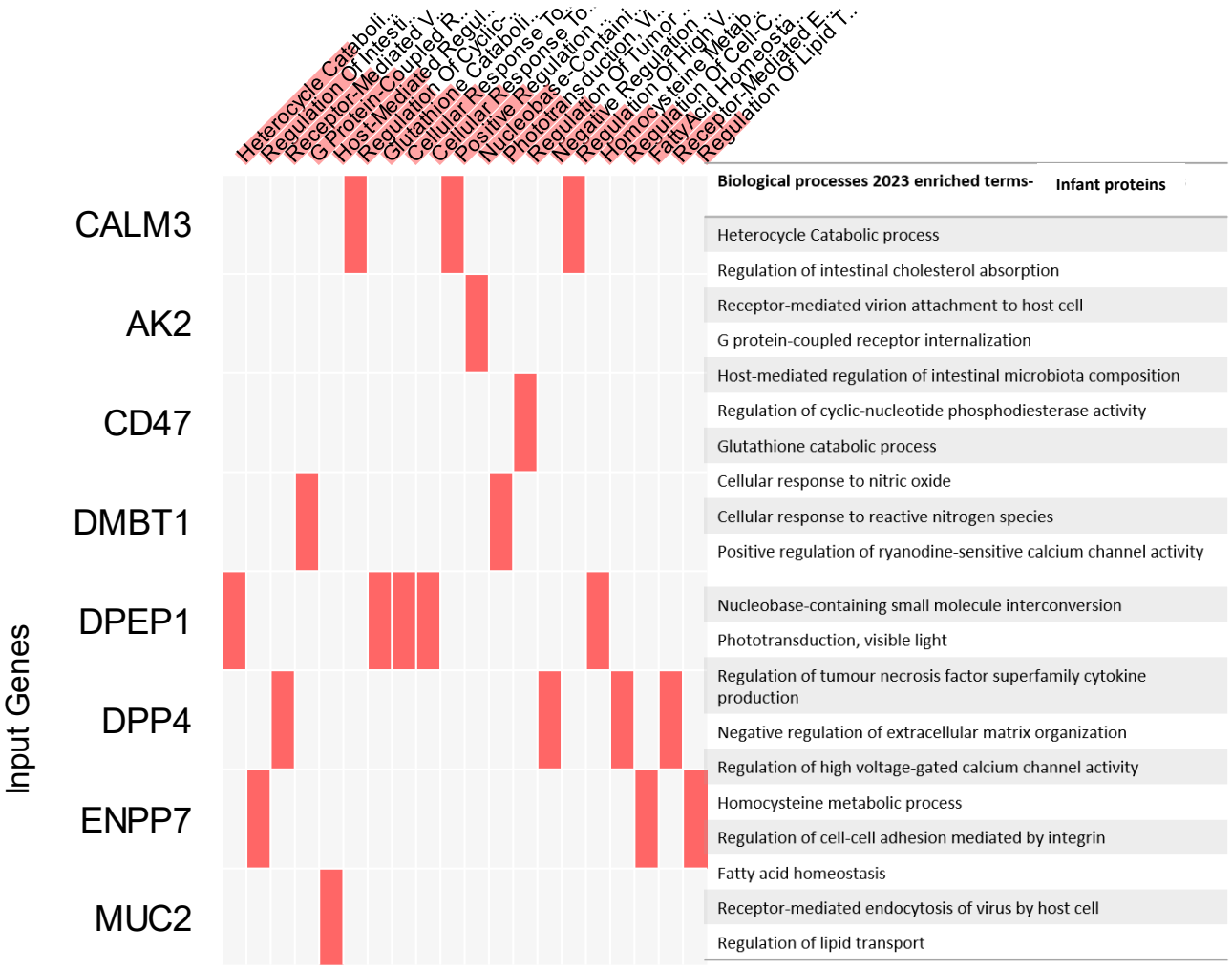


Figure 12: Infant protein analysis from the Random Forest model.

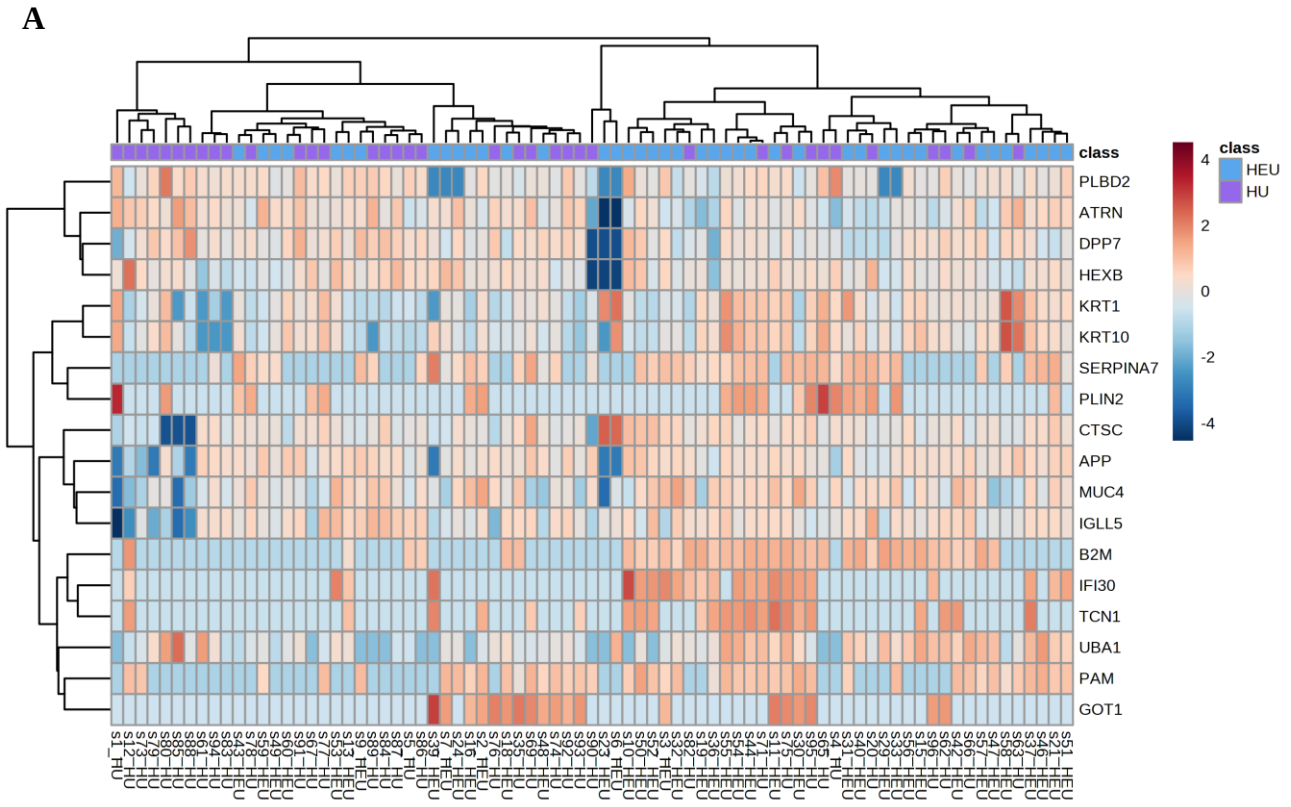
A- Heatmap of Random Forest top infant-derived proteins from 10 % model. Created in Metaboanalyst 5.0. B- STRING DB enriched proteins from top 10 % of Random Forest proteins for infant-derived proteins. *C-* Volcano plot showing the enriched terms for the infant proteins. Significance of gene set plotted with odds ratios and -log(p-values, with the darker the dot representing more significance.

D- Clustergram of genes enriched in the infant proteins under different GO Biological pathways. GO Biological pathways enriched from infant proteins on the right.

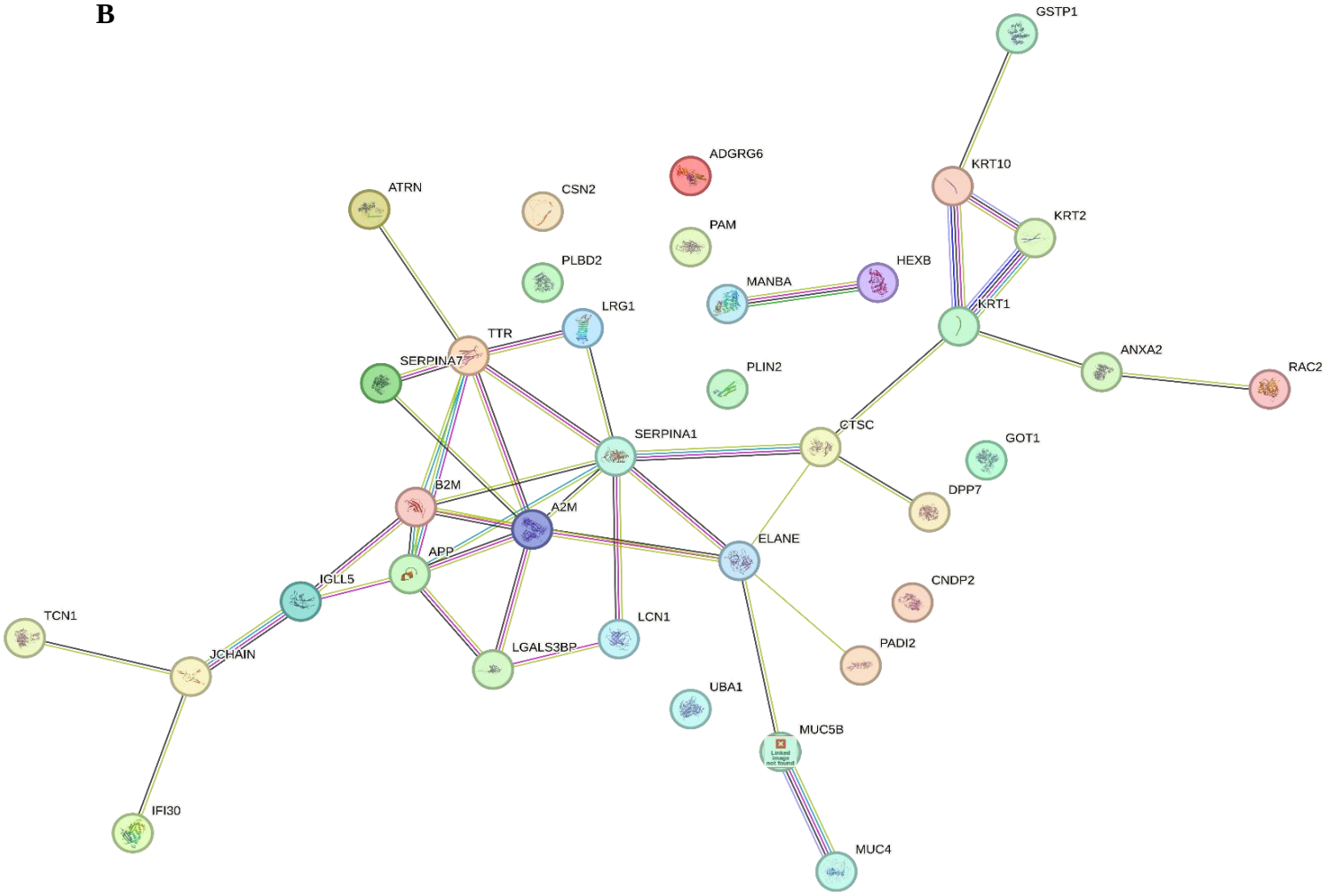
Breastmilk proteins show altered immunity from HIV exposed mothers.

Following Random Forest model application, the top classifiers from the breastmilk proteins are illustrated in a heatmap between the two groups HEU and HU in *Figure 13A*, revealing two distinct clusters: HEU dominant on the right and HU dominant on the left. Further analysis into the GO biological processes involved in these proteins was conducted in String DB, showing strong correlations between the identified proteins (*Figure 13B*). These mapped proteins were then analyzed in Enrichr and Appyter, following the same analysis as the infant-derived proteins (*Figure 13C and 13D*).

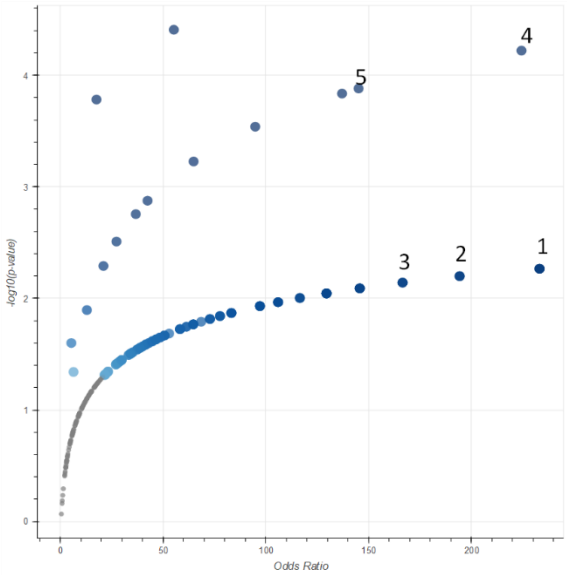
Figure 13C presents the volcano plot of the GO biological processes enriched in top-classifying breastmilk proteins. The most significant process was Aspartate metabolism from GOT1 in the GO biological terms Aspartate Metabolic Process and Aspartate Family Amino Acid Biosynthetic Process (*Figure 13C).* The next enriched GO term of interest was involved in the Adaptive immune system, specifically antigen presenting and processing via MHC class I for peptides and exogenous peptides, involving proteins IFI30 and B2M (*Figure 13C and 13D*). These proteins were upregulated in the HEU-dominant cluster when analysing the heatmap in *Figure 13A*. Lastly, KRT1 was enriched and involved in complement activation via the lectin pathway, suggesting innate and adaptive immune responses were upregulated in the HEU group from Antigen presentation and lectin pathway activation (*Figure 13A and 13D).* Therefore, a proposed altered immunity is seen in from HIV exposed mothers. An independent analysis of the breastmilk could be undertaken to assess the affect HIV has on the vertical transfer of immunity to the breastfed infants through breastfeeding.



B



C



	GO Biological Process	Gene
1	Aspartate Metabolic Process (Go:0006531)	GOT1
2	Antigen Processing and Presentation of Peptide Antigen Via MHC Class I (GO:0002474)	IFI30 & B2M
3	Aspartate Family Amino Acid Biosynthetic Process (Go:0009067)	GOT 1
4	Protein Heterotetramerization (Go:0051290)	KRT1 & KRT10
5	Antigen Processing and Presentation of Exogenous Peptide Antigen Via MHC Class I (GO:0002474)	IFI30

D

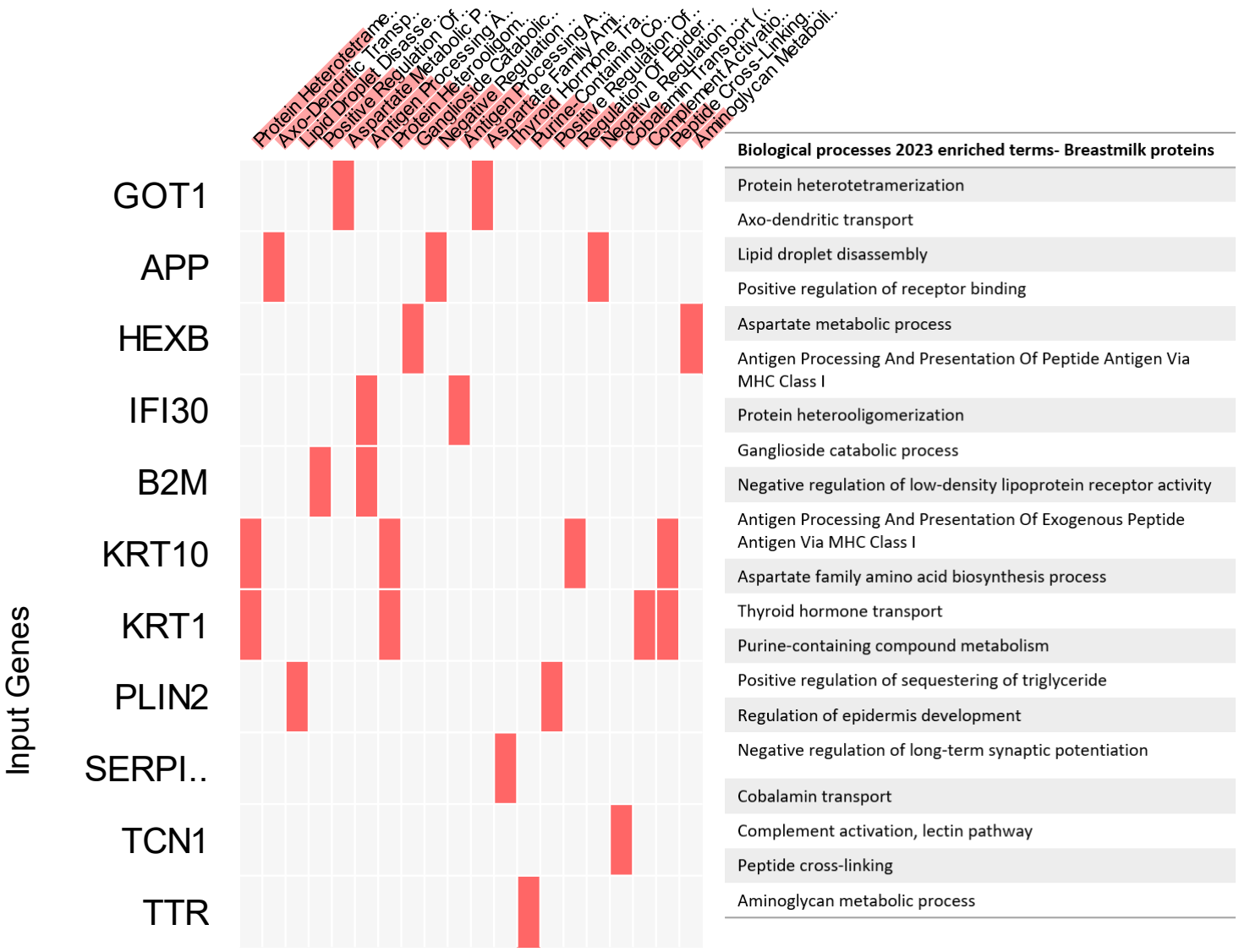


Figure 13: Breast milk protein analysis from the Random Forest model.

A- Heatmap of Random Forest top breastmilk proteins from 10 % model. Created in Metaboanalyst 5.0.

B- STRING DB enriched proteins from top 10 % of Random Forest proteins for breastmilk proteins.

C- Volcano plot showing the enriched terms for the breastmilk proteins. Significance of protein set plotted with odds ratios and -log(p-values, with the darker the dot representing more significance.

D- Clustergram of genes enriched in the breastmilk proteins under different GO Biological pathways. GO Biological processes enriched from breastmilk proteins on the right.

3.4.1 Human proteome discussion

Following the same methodology, the random forest regression model was applied to both infant protein and breastmilk datasets. The resulting OOB errors for the 10% classifiers datasets were 0.253 and 0.354, respectively (*Table 2*). These OOB errors revealed the infant proteins exhibited a stronger accuracy in predicting classifiers compared to the human and breastmilk protein datasets. A possible explanation for the non-distinct clustering in the breast milk proteins could be the contribution of mixed feeding, despite all participants stating they were exclusive breastfeeding. This hypothesis could be tested using known proteins identified in rice milk, cow's milk, and formula to reanalyse samples in DIA-NN for possible matches.

The intestinal mucosal barrier consists of epithelial cells coated in a mucosal layer which provides protection against pathogens and hosts beneficial microbiota colonization [101]. This mucosal layer is made up of mucins which interact and control gut microbes [104]. This protective effect can be achieved through the mucosal gut barrier itself (with anti-microbial peptides on it) or competitive-exclusion effect (good bacteria outcompete the bad for binding sites and nutrients) [101]. Additionally, commensal bacteria produce their own antimicrobial substances (bacteriocins) which aid in inhibiting pathogens and other bacteria growth [101]. MUC2 is a pivotal player in the gut barrier as it regulates mucin for the mucosal layer [105]. Defects in MUC2 can lead to unfavourable microbial growth which can be resistant to antimicrobial peptides, such as β-defensin 2 [105]. MUC 2 plays a role in stimulating β-defensin-2 production when proinflammatory cytokines are present, while also protecting bacteria (both beneficial and pathogenic) against antimicrobial effects. Defects in the MUC2 mucin barrier have been linked to inflammatory bowel disease and thus lead to deficient production of β-defensin-2 and result in a microbial gut dysbiosis [105]. An influencing factor for MUC2 production is inflammation, likely this increase is a mechanism to facilitate the expulsion of pathogens and maintain the integrity of the mucus layer [106]. During pathogenic infection increased intestinal inflammation is observed, with upregulated MUC2 mucin secretion observed through the NFkB pathway [106]. This aligns with our findings where the random forest model revealed a cluster of proteins, including MUC2, that exhibited higher abundance in HEU infants, potentially influencing gut microbiota regulation. MUC2's upregulation in HEU-dominant cluster (*Figure 11*) supports literature findings on MUC2's role in regulating gut microbiota. GO Biological enrichment terms highlight MUC2's involvement in gut microbiota regulation. This is consistent with literature indicating MUC2's role in maintaining a balanced microbiome.

Inflammation in the gut can be a driving factor for gut dysbiosis by creating a favourable environment for bacterial species that utilize the nutrients in the inflamed gut environment [106]. Resulting in reduction in diversity and the presence of non-beneficial bacteria such as Enterobacteriaceae, *Helicobacter hepaticus* and *Citrobacter rodentium,* can induce further intestinal inflammation and alter gut microbiome [106]. The relationship between gut microbiome dysbiosis and inflammation is a complex, tightly linked one, in which further investigation into it is needed for improved understanding for therapeutic interventions [106].

One contributing factor to inflammation is nitric oxide (NO), generated by nitric oxide synthetase (iNOS),

which can cause oxidative stress and damage to the intestinal gut barrier [106, 107]. The balance of RONS in the gut is important for maintaining normal and physiological cell functions, and disruption of this balance can lead to cytotoxic concentrations and damage to cellular components [107]. NO, generated by inducible nitric oxide synthase (iNOS), can have pro-inflammatory effects and disrupt the balance of the gut microbiota [107]. Excessive NO production can lead to oxidative stress and damage to the intestinal epithelial cells, compromising the gut barrier function [107]. One proposed mechanism of action in inflammation of NO, is its interaction with reactive oxygen species to form peroxynitrite which causes tissue damage and gut inflammation [106, 107]. The inflammatory conditions can result in increased oxygen conditions (with reactive oxygen species (ROS) and reactive nitrogen species levels increased (RNS)), favouring the growth of facultative anaerobes like Enterobacteriaceae while inhibiting the growth of obligate anaerobes like Bacteroidia and Clostridia [106]. This was supported by *Maziouridou et al* who performed *in vivo* mice imaging to show how ROS is dependent on NO to form peroxynitrite as an effector molecule to regulate gut microbes [104]. Further to imaging, knockout studies of NO showed the reduction of ROS, resulting in decreased beneficial bacteria, such as Lactobacilli [104]. Our findings support this, as infants with higher inflammation markers, including DPP4 and CD47, exhibited increased abundance of proteins associated with NO-related responses, indicating a potential link between NO, inflammation, and gut dysbiosis (*Figure 12*).

Further to NO dependent inflammation, TNF has been identified as a proinflammatory molecule in inflammatory bowel disease (IBD) and gut inflammation. To prove this, antibody studies against TNF were performed for IBD patients yielding promising results, concluding that TNF is a major contributor to gut inflammation [108]. TNF promotes proinflammatory responses through proinflammatory cytokines [108]. The TNF is believed to cause gut barrier defects and increase inflammation [108]. TNF presence can disrupt the mucus layer and weaken intercellular junctions, contributing to the entry of microorganisms into the lamina propria, further compromising the gut barrier causing gut microbiome dysbiosis [108]. In IBD, increased inflammation from TNF and NO is seen alongside MUC2 mucin barrier defects, resulting in deficient production of β-defensin-2 and gut dysbiosis [106, 108]. Highlighting the complex relationship existing between inflammation, gut barrier integrity and microbiome balance. Our research reveals a cluster of proteins associated with increased inflammation in HEU infants, suggesting a potential role of TNF in disrupting the mucus layer and weakening intercellular junctions, contributing to gut dysbiosis. These identified genes (DPEP1, CD47, DPP4) associated with inflammation, supports the literature on TNF-induced inflammation and its potential impact on the gut microbiome (*Figure 12*) [6].

Machiavelli et al investigated the inflammatory state and microbiome of HIV infected infants' guts through blood cytokines and stool samples, HEU and HU infants' guts [109]. No compositional difference form 16S was observed, however the alterations in the gut microbiome of HEU children may occur not only in microbial abundance but also at the functional level [109]. It is known that the microbiota of mothers has an impact on the development of newborns' microbiota and considering that HIV-infected women may have altered

microbiota, it is reasonable to hypothesize that HIV-exposed children also present dysbiosis [109]. Our research, employing a regression model on infant-derived proteins, identified clusters associated with HEU dominance, suggesting potential functional differences in gut microbiome regulation. The heatmap displays clusters of proteins, with the HEU-dominant cluster showing higher abundance. This aligns with the literature, emphasizing dysbiosis in HEU infant.

Research has shown the pivotal role of breastmilk in shaping the infant's immune system and gut development through its rich components, including growth factors, IgA, immune cells, mucins, chemokines, antioxidants, anti-inflammatory agents, and nutrients [51, 110-112]. IgA within breastmilk exhibits innate-like immune activities, activating pathogen phagocytosis and aiding in innate immune defence [51, 110]. The maternal entero-mammary circulation transfers IgA from the mother's gut to the infant, proposing how dysbiosis in the mother's gut might impact IgA production and functionality, potentially rendering the infant more susceptible to unfavourable bacterial colonization and altered immune tolerance [111, 112]. Moreover, breastmilk provides maternal major histocompatibility complex class I (MHC I) presenting immune cells, such as macrophages, T cells, and dendritic cells, which are transferred to train the infant's adaptive cells to produce their own IgA while the immune system and microbiome grows [111, 112]. HIV infection in the mother may impact the number of presenting cells produced, especially CD4+ cells, which could affect the infant's immune system development [111]. Dysregulation in these systems could contribute to acute conditions like diarrhoea and respiratory disease, as well as chronic conditions such as inflammatory bowel disease in HEU infants [112]. Our results align with these findings as HEU infants are seen to have altered health compared to their unexposed counterparts. The investigation into breastmilk associated proteins agreed with how HEU infants could be receiving altered immunity from their mother’s breastmilk, which could be resulting in immature infant immune system. The GO biology enrichment revealed both innate and adaptive immunity being different between the HEU and HU infant, with upregulated immunity seen in the HEU cluster (*Figure 13*). Altered IgA levels and antigen presentation from the adaptive immunity from dysbiosis in the maternal gut, could be allowing tolerance to pathogenic bacteria to be trained and promoting growth of unfavourable bacteria in the HEU infants’ gut microbiomes.

Furthermore, breastmilk contains live bacteria, antimicrobial peptides, and human milk oligosaccharides that nourish bacterial growth, aid in immune tolerance and digestion of non-digestible carbohydrates in the colon [51, 111, 112]. These bacteria, including Bifidobacterium, Staphylococcus, Streptococcus, and Lactobacillus, whose colonization is stimulated by glycan production, oligosaccharide availability, and immune tolerance, play a crucial role in training the infant's immune system to tolerate beneficial bacteria for optimal gut health [51, 111, 112].

The amino acids in breastmilk, notably aspartate and leucine, serve as essential building blocks for infant growth, gut function, mucosal secretion, and immune response [113, 114]. From our data, aspartate

metabolism was observed as an enriched GO biological pathway associated with the breastmilk proteins from the random forest model (*Figure 13*). Aspartate supplementation in a low-energy diet has been shown to improve intestinal barrier integrity and stimulate gut cell proliferation by activating the mTOR signalling pathway [113]. Aspartate not only aids in cell growth and protein synthesis but also serves as a substrate for gut bacteria to synthesize other amino acids crucial for gut function [113, 114]. Additionally, bacteria can use aspartate to produce histamine, exerting anti-inflammatory effects by suppressing TNF [114]. However, when assessing the abundance of the associated protein, GOT1, in *Figure 13*, a less distinct clustering is seen between the two groups, with a weak higher association linked to HU cluster. Leading to the hypothesis that aspartate metabolism and content could be altered from HIV infected mothers and could be causing altered gut barrier health in HEU infants. This hypothesis warrants further investigation into the breastmilk composition as a separate study to assess these effects on the HEU infants' health.

The relationship between gut epithelium integrity, breastmilk composition, immunity, and the gut microbiome form a fundamental link crucial for overall host health, particularly in vulnerable populations like HEU infants. The integrity of the gut barrier, by factors such as the MUC2 mucin barrier, is essential in maintaining a balanced gut microbiome. Disruptions in this barrier, often linked with inflammation, can lead to dysbiosis, influencing microbial diversity and composition. Simultaneously, breastmilk, rich in immune components like IgA and growth factors, aids in maturing immunity and shaping the developing gut environment. The components in breastmilk, including amino acids and antimicrobial agents, not only nourishes the microbiome but also potentially modulates its composition and function. The delicate interplay between breastmilk-derived factors, gut integrity, and immune responses orchestrates an environment that primes the gut microbiome, influencing its diversity, resilience, and ultimately, the host's health. Understanding these interconnected mechanisms offers insights into how nurturing the gut ecosystem through optimal breastmilk composition and preserving gut integrity can foster a healthier microbial environment, fortifying immune defences and supporting overall well-being in HEU infants.

After this analysis into the host proteome differences influencing gut health and immunity in HIV-exposed uninfected (HEU) infants, we proceeded to perform an analysis their gut microbiomes. As stated previously, the gut microbiome plays an essential role in shaping overall health, particularly in early life stages, interacting intricately with factors such as intestinal barrier integrity, inflammatory responses, and the influence of breastmilk components. Understanding the composition and functional implications of the gut microbiome in HEU infants is paramount, as it impacts host immune system development, growth, and overall health. Our focus on microbiome analysis aims to uncover specific microbial signatures and their potential associations with the discussed host factors, offering insights for potential interventions to optimize gut health and fortify immune development in this vulnerable population.

Chapter 3.5 – Microbial proteome

Species level analysis cannot be extrapolated to protein level differences.

In this investigation, the aggregation of all proteins per identified species provided a signal per species, enabling an assessment of total species proteome activity within the gut microbiome (*Figure 14*). However, this cannot distinguish between specific proteins of interest being different between the two groups. While the composition assessed in an unsupervised manner in Chapter 3.2 found no differences between groups, HEU and HU, subsequent supervised analysis was used to try identifying any classifiers.

Initially, PLSDA was used to observe any clustering for the 2 groups, HEU and HU. *Figure 14A* illustrates the top 30 species contributing to the clustering from PLSDA, showcasing a higher abundance of bacterial signal in the HEU cluster compared to the HU cluster. To assess if the signal from bacterial origin was greater in one of the groups of interest (HEU or HU) accounting for any differences between the groups, a comparison between the total protein intensity from the top 30 bacterial species and total human protein intensity was assessed. This yielded no significant disparity between HEU and HU signal, and therefore the microbiome, as well host response, was not distinctly more active in either group (see appendix O for this signal comparison).

To extend this analysis, the Random Forest model was applied to the species data. While this method identified top classifying species, it proved less distinct in clustering of the 2 groups. The top identified classifying species from this random forest model were *Prevotella bivia, Parabacteroides merdae, Streptococcus parasanguinis, Holdemanella porci, Faecalibacterium sp, Enterobacter hormaechei, Solobacterium moorei, Sutterella parvirubra, and Bacteroides thetaiotaomicron (Figure 13B)*. Notably, *Bacteroides thetaiotaomicron*, previously associated with lower vaccine response, emerged as one of the significant species in HEU [89, 115]. However, when further assessing the proteins identified from *Bacteroides thetaiotaomicron* no significant proteins were identified, due to the missingness in protein data, which was not due to chance but rather the proteins not being present (See Appendix P). From this, we concluded that summing the protein intensity per taxa was not effective in identifying proteins of interest from the top classifying species, and their involvement in microbiome functioning could not be assessed. To try achieving this, a pathway assessment and protein level analysis was therefore needed to identify proteins which could distinguish between the two groups, HEU and HU. This level of assessment of functional pathways can be used to identify any functions different in HEU infants accounting for health issues associated with these infants.

Of note, previous 16S analysis did not reveal any taxonomic differences between the two groups, which contrasts this proteomic analysis at a species level when summing the proteins' signal from each species [89]. This highlights the deeper insight a proteomic analysis can achieve into the microbiome functionality compared to genomic analysis by assessing the active genes from the species, possibly accounting for these differences.

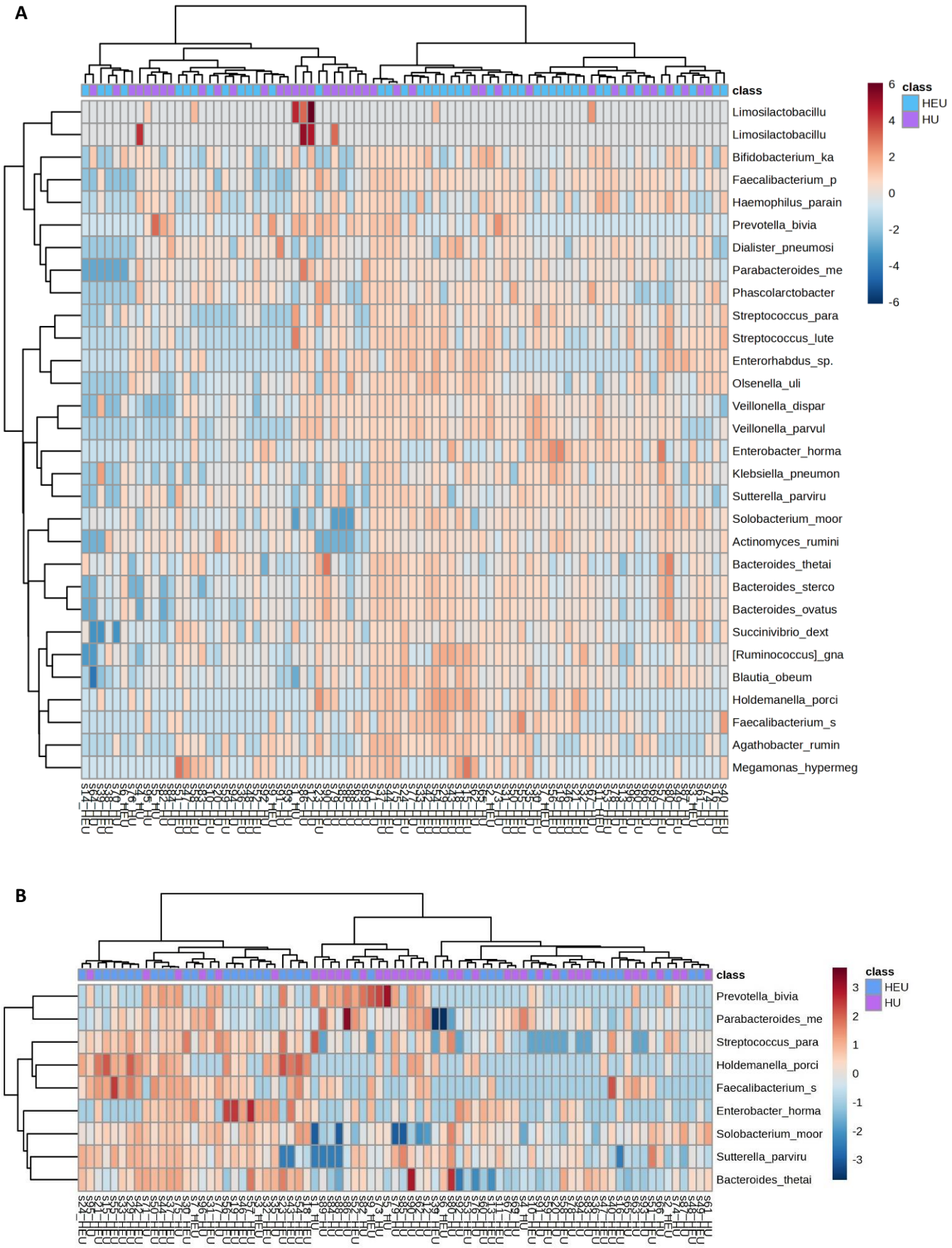


Figure 14: Species analysis of cumulated protein intensity in the infant samples between HEU and HU.

A- Top 30 species PLSDA vip scores between HEU and HU.

B- Top species identified from data between HEU and HU from Random Forest model.

Bacterial pathways are under annotated.

The subsequent pathway analysis involved in the identified microbiome was done to assess if any pathway differences could account for the differences in HEU and HU health issues. Proteins involved in the same pathways were assigned KO terms in EggNOG mapper v.5 and analysed from matching signal data. This was aimed at analysing microbial communities performing the same function in the gut and assess any differences occurring at this level between the two groups.

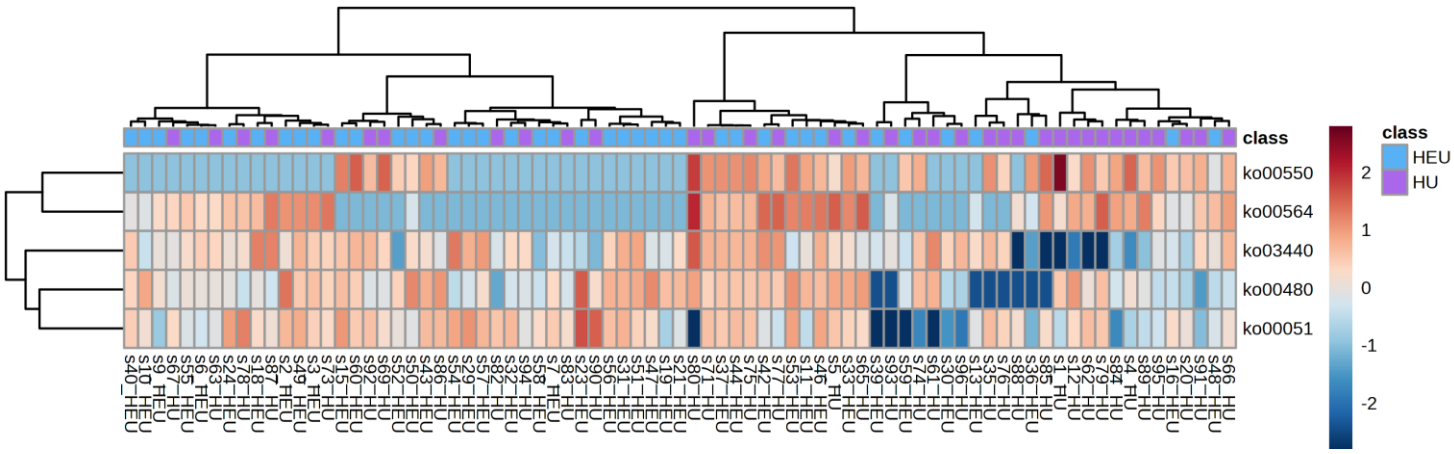
The pathway analysis from EggNOG mapper's annotated KO pathways was performed using the Random Forest regression model (*Figure 15A*). This analysis was done on a subset of proteins which had known functions from EggNOG mapper, unfortunately this only included 65 % of the identified total bacterial proteins. This highlights the fact that bacterial functions are not as well annotated as the human database, and that the need for further studies into the microbiome are needed for in depth analysis for therapeutic treatments which could involve the microbiome. *Figure 15A* shows the random forest results in a heatmap representing the top 5 discriminating pathways between the two groups. Of the known identified pathways, the top 5 revealed Firmicutes contributing the most to the functions, with Actinobacteria being secondary (*Figure 15B*). This was further elaborated on in *Figure 15C* with the number of proteins contributing to this function from each phylum and their total signal towards the function.

The top pathways upregulated in the HU cluster were peptidoglycan biosynthesis and glycerophospholipid metabolism. Furthermore, the top upregulated pathways in the HEU cluster were homologous recombination, glutathione metabolism and fructose and mannose metabolism. Of Interest out of these pathways was fructose and mannose metabolism with Bacteroides having the most activity from 16 proteins, followed by Firmicutes with 16 proteins, Actinobacteria with 14 proteins and Proteobacterium with 6 proteins.

Further analysis into the metabolites produced by bacterial pathways would be beneficial for insight into the microbiome functionality and active pathways as these are the effector molecules involved in host interaction. Since all the identified bacterial proteins were not annotated for the pathway analysis, a protein level analysis was required to identify any proteins different between the two group, and thus identify any genus or species that could be discriminant.

The prior methods aimed to address missing data in microbial protein analysis. However, insufficient annotations of protein pathways and results from species analysis, which couldn't be extrapolated to the protein level due to missing data, necessitated a focus on protein-level analysis. Initially, the approach involved aggregating redundant proteins across various taxa to reduce missing data in the protein analysis. This process, performed at the KO level using EggNOG mapper, resulted in unsatisfactory clustering and outcomes, as well as unannotated proteins from KO level (refer to the appendix Q). Consequently, the focus shifted to analysing all identified proteins from bacterial species to account for specific species or genera of interest.

A



B

KO	PATHWAY	UPREGULATED GROUP	DOMINANT PHLYUM CONTRIBUTING PROTEIN
KO00550	peptidoglycan biosynthesis	HU	Actinobacteria
KO03440	homologous recombination	HEU	Firmicutes
KO00564	glycerophospholipid metabolism	HU	Firmicutes
KO00480	glutathione metabolism	HEU	Actinobacteria
KO00051	fructose and mannose metabolism	HEU	Firmicutes

C

ko00550

Peptidoglycan biosynthesis

Phylum	**Number of proteins**	**% phylum to KO**
Proteobacterium	0	0
Bacteroides	3	1.214
Firmicutes	1	1.174
Actinobacteria	2	97.614

ko03440

Homologous recombination

Phylum	**Number of proteins**	**% phylum to KO**
Proteobacterium	0	0
Bacteroides	1	0.107
Firmicutes	4	99.893
Actinobacteria	1	0

ko00564

Glycerophospholipid metabolism

Phylum	**Number of proteins**	**% phylum to KO**
Proteobacterium	1	0.325
Bacteroides	1	3.219
Firmicutes	2	87.823
Actinobacteria	3	8.633

ko00480

Glutathione metabolism		
Phylum	**Number of proteins**	**% phylum to KO**
Proteobacterium	3	2.777
Bacteroides	9	29.233
Firmicutes	5	10.505
Actinobacteria	4	57.485

ko00051

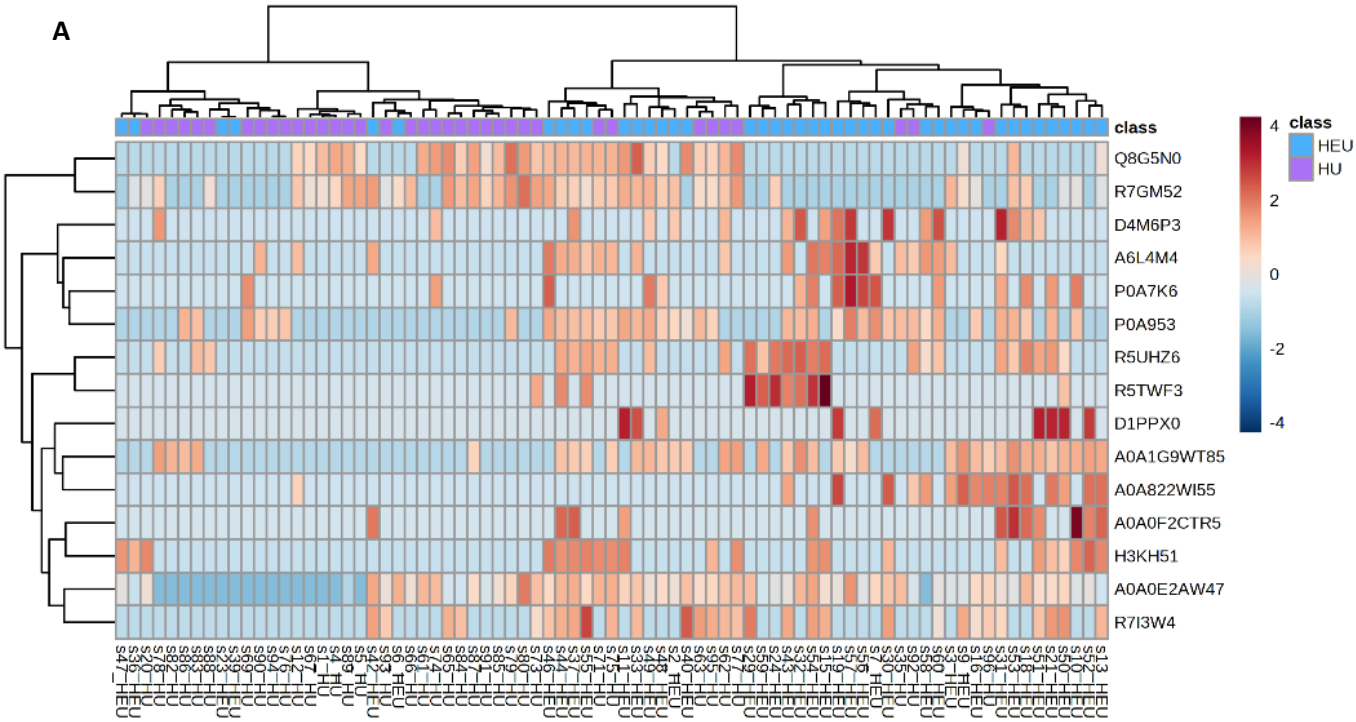
Fructose and mannose metabolism		
Phylum	**Number of proteins**	**% phylum to KO**
Proteobacterium	6	25.758
Bacteroides	16	33.974
Firmicutes	16	26.159
Actinobacteria	14	13.518

Figure 15: Bacterial pathway analysis between HEU and HU groups using Random Forest model.
A- Top 5 KEGG pathways identified as different between the two groups from Random Forest model. Heatmap generated in Metaboanalyst 5.0.
B- Top identified KEGG pathways from Random Forest model with pathway involvement, upregulated group and dominant phylum contributing to the pathway.
*C- Table s*ummary of top 5 KEGG pathways different between HEU and HU. Each KEGG table showing phylum summary, contributing signal to pathway and number of proteins contributing form each phylum.

Microbial proteins can be used to identify a classifying species from HEU infants.

In contrast to the species analysis in *Figure 14*, the random forest performed on bacterial proteins identified proteins from the different genus being Ruminococcus, Bifidobacterium, Bacteroides, Enterobacter, Escherichia, and Streptococcus. As mentioned previously, analysing the proteins offers insights into the active functionality of bacteria, providing a more detailed understanding compared to assessing the entirety of species presence. This protein level analysis yielded clearer clustering when compared to the species and pathway analysis (*Figure 16*). *Figure 16A* illustrates the absence of most bacterial proteins in the samples, contributing to the complexity of the sample analysis. Employing supervised clustering and machine learning enabled the analysis of samples despite this missing data, avoiding imputations while also retaining biologically significant findings. This indicated a trend among HEU infants, suggesting a more active bacterial functional community with a higher abundance of bacterial proteins. Notably, these prominent proteins weren't confined to a single upregulated phylum; instead, proteins from all phyla contributed to this observed pattern, as depicted in *Figure 16B*. The clustering of the two groups is evident among the top 15 proteins, primarily originating from the Firmicutes phylum, followed by Proteobacteria, Bacteroides, and Actinobacteria, in that order.

A



B

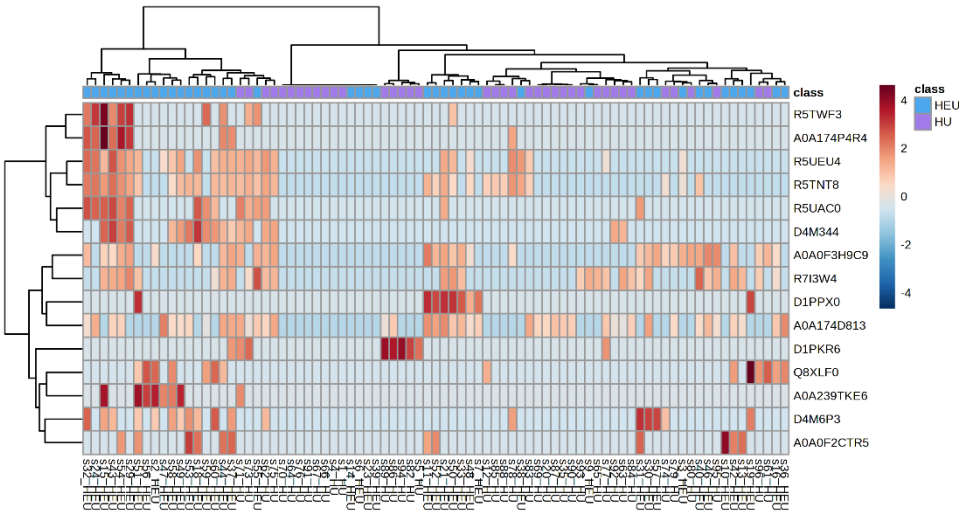
Protein	***Genus***	***Phylum***
R5TWF3	*Ruminococcus*	*Firmicutes*
P09373	*Escherichia*	*Proteobacteria*
R5UHZ6	*Ruminococcus*	*Firmicutes*
Q8G5N0	*Bifidobacterium*	*Actinobacteria*
D4M6P3	*Ruminococcus*	*Firmicutes*
A0A822WI55	*Enterobacter*	*Proteobacteria*
A0A0F2CTR5	*Streptococcus*	*Firmicutes*
A0A0E2AW47	*Bacteroides*	*Bacteroidetes*
P0A953	*Escherichia*	*Proteobacteria*
R7GM52	*Catenibacterium*	*Firmicutes*
A6L4M4	*Phocaeicola*	*Proteobacteria*
A0A1G9WT85	*Actinomyces*	*Actinobacteria*
D1PPX0	*Subdoligranulum*	*Firmicutes*
R7I3W4	*Faecalibacterium*	*Firmicutes*
H3KH51	*Sutterella*	*Actinobacteria*

Figure 16: Random Forest model results of bacterial proteome.
A- Top 15 bacterial proteins from Random Forest model shown in Heatmap.
B-Table of top proteins from Random Forest model showing their Genus and Phylum.

Firmicutes has the most classifying proteins between HEU and HU infants.

From this analysis, Firmicutes was identified as a phylum of interest due to almost 50 % of the discriminating proteins being from Firmicutes, see *Figure 16B*. Consequently, a focused analysis of proteins within Firmicutes was conducted to identify discriminating genera or species. The results depicted in *Figure 17* revealed the predominant genera, Ruminococcus, within Firmicutes responsible for classifying proteins between the two groups. Ruminococcus stood out, contributing to nearly 50 % of the top proteins, while Blautia and Streptococcus each featured 2 proteins as classifiers within their respective genera (as detailed in *Figure 17B).* This focused analysis on Firmicutes highlighted Ruminococcus as a noteworthy genus and species of interest (as indicated in *Figure 17).* Moreover, Ruminococcus emerged as a classifier in the total bacterial proteome analysis depicted in *Figure 16*. Specifically, within the top 5 bacterial proteins contributing to the disparity between HEU and HU guts, the Ruminococcus genus was linked to three proteins—R5TWF3, R5UHZ6, and D4M6P3. Hence, it was imperative to conduct further analysis on Ruminococcus.

A



B

Protein	Genus
Q8XLF0	Clostridium
R7I3W4	Faecalibacterium
A0A0F2CTR5	Streptococcus
A0A0F3H9C9	Streptococcus
A0A174P4R4	Blautia
A0A174D813	Blautia
A0A239TKE6	Megamonas
D1PKR6	Subdoligranulum
D1PPX0	Ruminococcus
D4M344	Ruminococcus
R5TNT8	Ruminococcus
D4M6P3	Ruminococcus
R5UEU4	Ruminococcus
R5UAC0	Ruminococcus

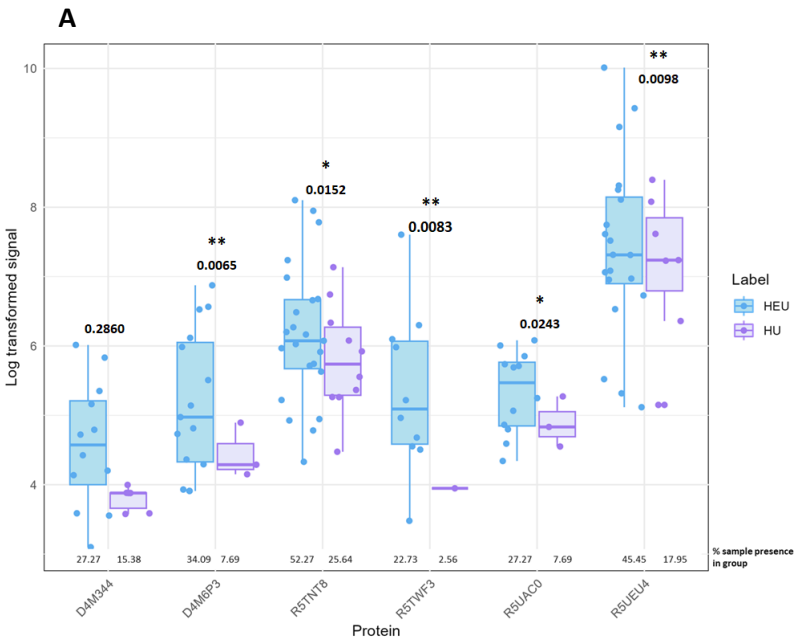
Figure 17: Firmicutes proteome analysis using Random Forest model.
A- Firmicutes top 15 discriminating proteins from Random Forest model analysis. Heatmap made in Metaboanalyst 5.0.
B- Table showing the top discriminating proteins from Random Forest model between the two groups from the Firmicutes phylum, showing the protein's name and genus.

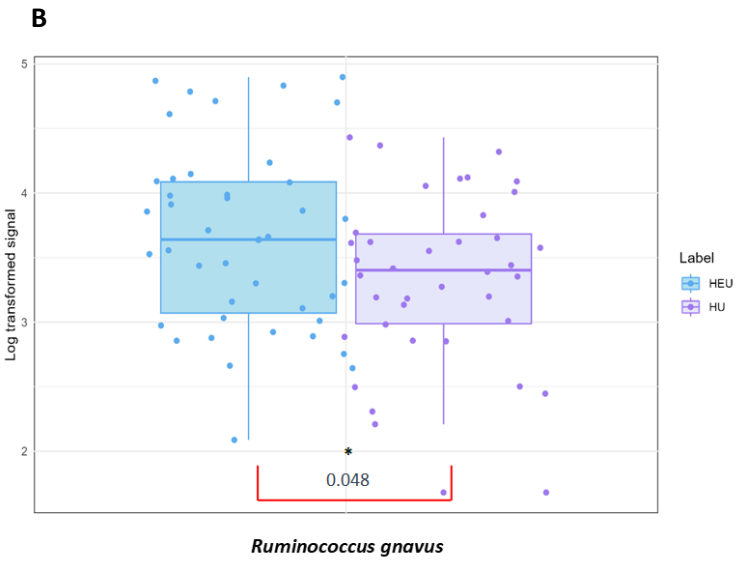
Ruminococcus Gnavus has an altered proteome between HEU and HU infants.

In *Figure 18A*, the results of Fischer's exact test for the presence of the top Ruminococcus genus proteins are displayed. Among these proteins, 5 out of 6 showed a significant presence in the HEU group compared to the HU group. The Fischer's exact test yielded p-values lower than 0.05 for proteins R5TNT8 and R5UAC0, and values lower than 0.01 for proteins D4M6P3, R5TWF3, and R5UEU4. Notably, proteins R5TNT8, R5UAC0, R5TWF3, and R5UEU4 were all from the species *Ruminococcus gnavus* as identified in the Firmicutes analysis (*Figure 17)*, along with classifying proteins R5TWF3 (also present in *Figure 18A*) and R5UHZ6 identified in the total bacterial proteome in *Figure 16*. This then supported the idea to further analyse *Ruminococcus gnavus* as a species to assess its proteome and functionality in the infants' guts.
Figure 18B involved an analysis of the total protein intensity from the *Ruminococcus gnavus* species between the two groups using a Fischer's exact test. The resulting value of < 0.05 indicated a significant difference in *Ruminococcus gnavus* levels between the two groups. This supports the advantage of conducting a protein-level analysis compared to a species-level analysis, as this now proven significant species was overlooked in the species analysis, in *Figure 14.*

For a more in-depth investigation of this species, all proteins from *Ruminococcus gnavus* underwent a Fischer exact T-test, revealing 17 proteins that exhibited significant differences between the HEU and HU infants. These findings are detailed in *Figure 18C* (Box plots provided in the Appendix R). Within *Figure 18C*, the known annotated KO pathways and their pathway involvement are shown, highlighting the unknown bacterial functions by the absence of many pathway annotations.

This highlights the extensive unknowns surrounding microbiomes and their impact on host effects, emphasizing the need for research in this field to facilitate therapeutic treatments for improving host health. Notably, among the known annotated pathways of these differential proteins, glycolysis emerged as the most involved pathway, indicating the pivotal role of bacterial metabolism in maintaining a healthy gut function. These results leave us interested in *Ruminococcus gnavus* as an associated bacteria with HEU infants' gut microbiome, and how its possible involvement in the gut leads to perturbated host health.





C

Protein	Fischer's p value	KO pathway	Pathway
R5TL56	0.016341	ko00261	Monobactam biosynthesis/ glycine, serine, and threonine
R5TLH5	0.027236	-	-
R5TNT8	0.037976	-	-
R5TQK2	0.031869	-	-
R5TRC7	0.006004	-	-
R5TRJ0	0.031869	ko00220	Arginine biosynthesis
R5TSJ6	0.036916	-	-
R5TSL4	0.012859	-	-
R5TVA6	0.027236	ko00250	Alanine, aspartate, and glutamate metabolism
R5TWF3	0.008212	ko00010	Glycolysis/ gluconeogenesis
R5TYD4	0.012859	-	-
R5U8M1	0.036916	ko00770	Pantothenate and CoA Biosynthesis
R5UAC0	0.024032	ko00260	Glycolysis/ gluconeogenesis
R5UAN2	0.027236	ko00010	Glycolysis/ gluconeogenesis
R5UDE1	0.006004	-	-
R5UEU4	0.016813	ko02010	ABC transporters
R5UHZ6	0.016813	ko00010	Glycolysis/ gluconeogenesis

Figure 18: Ruminococcus gnavus proteome analysis using Random Forest model.

A- Ruminococcus gnavus species signa boxplot between the two groups. Welch independent t-test with p-value < 0.05 showing significant difference in Ruminococcus signal between the 2 groups (HEU and HU). *B-* Firmicutes top significant proteins from Ruminococcus. Sample presence in each group below. Fischer's exact test for protein presence with p-value < 0.05 * and < 0.01 **.

C- Table summarizing the top Ruminococcus proteins. Showing the proteins Fischer's p-value, KO pathway.

3.5.1 - Bacterial proteome discussion

This proteomic analysis on the microbiome has given insight into the secreted proteins from the active bacteria present in the gut. This allows assessment into function and interactions in the gut rather than what genomics offers about the presence alone [116]. To further expand this analysis for host and microbiome interactions, a metabolomic analysis could aid in giving an overall view of the metabolic states of the microorganisms in the gut. These 3 "-omics" techniques used in this study can be used together to gain a comprehensive understanding of the gut microbiome system in effect in these infants' guts [116].

From the literature, microbiome studies to date have largely focused on genomic analysis using 16S sequencing analysis of gut communities [97, 109, 117]. The results from the present study prove proteomics can be used to detect, identify, and quantify the microbial communities and their functions in the gut, enabling connections to be found between the microbiome and the host through present proteins [116]. *Figures 14* highlights the potential in identifying species and quantifying their activity in the microbiome. While the extrapolation from a protein level to a species of interest is proven by assessment of the bacterial proteins in *Figure 16,* leading to the findings of *Ruminococcus gnavus* as a species of interest from its classifying proteins for the 2 groups in *Figure 16.*

Figure 15 highlights the upregulation of fructose and mannose metabolism in the HEU cluster, a significant finding from the pathway analysis that reveals distinct differences in bacterial metabolism within the gut of these HEU infants. Mohammed et al. discuss how bacterial enzymes catalyze fructose into substrates that intersect with human metabolic pathways, such as Phosphofructokinase, ultimately producing Glyceraldehyde-3P—a critical component of the glycolysis pathway for energy generation [118]. This intricate interplay between bacterial metabolism, specifically involving fructose and mannose, and the host's metabolic processes highlights the vital role of these pathways in energy production necessary for cellular functions, growth, and repair.

Alterations in the gut microbiome's metabolic pathways can lead to dysregulated energy production, affecting nutrient breakdown and immune functions. This disruption can contribute to the development of inflammatory bowel disease, metabolic disorders, altered disease susceptibility, and autoimmune diseases [118]. Understanding these metabolic changes at the bacterial level can provide crucial insights into host responses, potentially opening new possibilities for therapeutic targets. Further metabolomic studies, as previously discussed, could offer valuable insights into the microbiome's metabolic state.

An intriguing aspect of mannose metabolism involves its interaction with the host's innate immune system, particularly through the Mannose-binding lectin (MBL)—a pattern recognition molecule [119]. MBL can bind to bacteria bearing mannose residues on their surface, activating the complement system and triggering downstream processes such as opsonization and phagocytosis of pathogens [119]. Studies suggest that MBL might enhance the release of pro-inflammatory cytokines from monocytes in response to bacteria.

When bacterial mannose metabolism is upregulated, more mannose residues are presented on the bacterial surface. This increases the likelihood of MBL binding to these bacteria. The binding of MBL to mannose residues activates the complement system, enhancing the opsonization and subsequent phagocytosis of these bacteria by immune cells. The increased phagocytosis of bacteria, facilitated by MBL binding, can lead to an enhanced release of pro-inflammatory cytokines. This inflammatory response, while crucial for clearing pathogens, can become dysregulated and lead to a chronic inflammatory state. The chronic inflammatory response can disrupt the gut microbiome, potentially reducing commensal bacteria and favouring pathogenic bacteria growth. Thus, in the HEU infants an upregulated mannose metabolism could be contributing to the inflammatory state of their guts through this mechanism.

It is important to note that while increased mannose metabolism might enhance microbial clearance, it can also lead to a paradoxical effect where inflammation promotes the growth of pathogenic bacteria by creating a hostile environment for commensal bacteria. This contrast highlights the complexity of the host-microbiome interaction and the delicate balance required to maintain gut health, even in infants so young [119].

Our discovery of increased Ruminococcus in the gut microbiomes of HEU infants resonates with *Wang et al*.'s 16S study, which also noted increased Ruminococcus levels in the gut microbiomes of HIV-infected individuals [117]. This finding suggests a potential transmission of gut microbiomes from HIV-infected mothers to their infants, regardless of the infants' HIV status [117]. Further supporting this, *Xu et al*. confirmed elevated Ruminococcus levels in the guts of HIV-infected individuals, whilst also observing increased Proteobacteria, linked to mediating inflammatory responses, contributing to inflamed guts in HIV individuals [118].

Ruminococcus gnavus, a robust gram-positive anaerobe from the Firmicutes phylum (supported by its association with Peptidoglycan synthesis in *Figure 15*), emerged as a significant bacterium of interest from the Firmicutes phylum in our analysis of HEU stools (*Figure 16*) [119].

Ruminococcus gnavus exhibits the capability to metabolize human milk oligosaccharides (HMO) and mucins, influencing the colonization of other gut bacteria and impacting the gut barrier. It competes with other microbes by producing bacteriocins, actively hydrolysing host carbohydrates (HMO), and employing adhesins to mucus for outcompeting microbes [119].

The genomic makeup of *Ruminococcus gnavus* includes genes associated with adapting to disease-related oxidative stress in the gut and the ability to stimulate TLR4 activity, potentially inducing inflammatory states in the gut, contributing to pathogenesis in diseases like IBD and Crohn’s [119]. Additionally, *Ruminococcus gnavus* expresses a VH3 B cell receptor targeted B cell superantigen, implicated in the pathogenesis of lupus by activating B cells [119]. As stated, *Ruminococcus gnavus* has been implicated in various gut diseases including Crohn's, IBD, and is also associated with lung and neurological diseases such as Parkinson's, all tied

to inflammation [119-121]. This highlights *Ruminococcus gnavus'* tendency to induce inflammatory responses in the host [119-121].

Studies by *Chua et al*. have linked *Ruminococcus gnavus* to infant allergy development through early-life microbiome colonization [121]. Infants extensively colonized by *Ruminococcus gnavus* were found to be at increased risk of allergies, particularly respiratory issues. This effect is believed to extend to the lungs through immune cell interactions, such as T helper cells, eosinophils, and inflammatory colonic cytokines [121]. *Ruminococcus gnavus* can degrade the mucin layer, allowing other bacteria to inhabit the mucus without an inflammatory response. However, when *Ruminococcus gnavus* breaches the mucin layer, it triggers an inflammatory response upon recognition by host dendritic cells [121]. The impact of *Ruminococcus gnavus* is thought to be bidirectional, contributing to a vicious cycle of microbiome dysbiosis and gut inflammation [119].

Another disease associated with increased bacterial growth of *Ruminococcus gnavus* is Systemic Lupus Erythematosus [122]. This autoimmune disease's etiology remains unknown, but one related health issue involves a compromised gut epithelial barrier, leading to leaky gut syndrome and bacterial migration to systemic organs [122]. Research indicates that *Ruminococcus gnavus* produces enzymes that degrade the gut's mucin layer, thereby increasing the permeability of the intestinal barrier. Additionally, this bacterium generates metabolites that promote inflammation and disrupt the epithelial barrier [122]. Pro-inflammatory cytokines such as TNF and IL-6 are linked to the presence of *Ruminococcus gnavus* in the gut, suggesting that targeting *Ruminococcus gnavus* could offer a potential therapeutic approach for leaky gut syndrome [122].

Chapter 3.6: Metaproteome

Distinct clustering has been observed in the gut microbiome proteome of infants from their stool samples within their first week of life. This was interesting as the complexity and diversity among the samples has been evident throughout the analysis, showcasing unique microbiomes that may stem from differing environments and exposures being influenced by HIV exposure, starting during their first week of life.

To try tie these clustering observed in *Figures 12, 13 and 16,* the relationship was analysed in a heatmap between the gut proteins, breastmilk proteins, and bacterial proteins, as depicted in *Figure 19*. Notably, our analysis revealed a strong correlation between the gut proteins of HEU infants and bacterial proteins from the same cluster, emphasizing the pivotal bi-directional association between bacterial dysbiosis and gut inflammation [116]. This correlation was assessed by overlaying the samples annotated clusters, with the gut and bacterial protein clusters having a correlation of 76.92 % (*Table 5*).

While the breastmilk protein clusters displayed weaker correlations with both gut and bacterial proteins (61.54 % and 65.38 % respectively, *Table 4*), it became evident that the differences in health outcomes between HEU and HU groups were primarily influenced by gut and bacterial proteins rather than maternal

breastmilk proteins. This insight emphasizes the significance of conducting a separate analysis on the breastmilk to investigate potential differences at this level independently.

Table 6: The comparison between the 3 different clusters shown in Figure 31 showing % overlap for correlation between clusters.

Comparison between RF identified clusters	Overlay in clusters
Gut vs Bacterial	76.92%
Gut vs Breastmilk	61.54%
Bacterial vs Breastmilk	65.38%

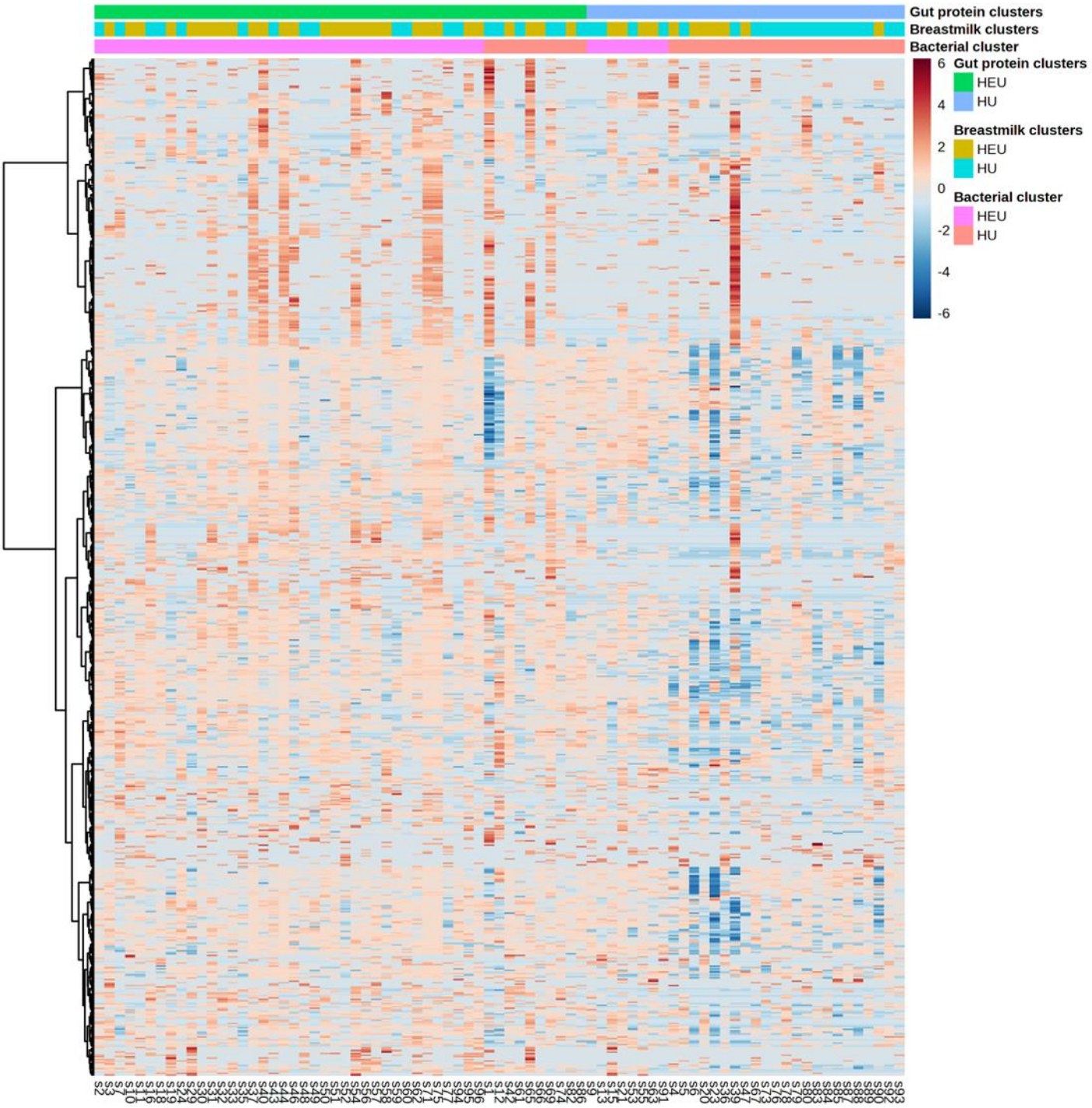


Figure 19: Overlay of clusters from gut, breastmilk, and bacterial clusters.

Chapter 4: Conclusion

Our proteomic investigation into the gut microbiome provided a unique window into the functional dynamics of the microbiome's activity [104]. Our study further supports the emerging shift from solely relying on genomic analyses, such as 16S sequencing, to explore gut communities [97, 109, 117]. The comprehensive SWATH-based metaproteomic analysis method developed here can both quantify and identify the proteome of both human and bacterial proteins simultaneously. This enables the in-depth analysis of the metaproteome to identify the interactions happening in the gut between the host and inhabiting

microbiome. This demonstrates the efficacy of proteomics in identifying, quantifying, and understanding microbial communities and their functions in the host, revealing the potential for a deeper analysis between the microbiome and host through the presence of specific proteins [116, 117]. Furthermore, this metaproteomic analysis of the gut revealed proteins from bacteria, infants, as well as some information about the mothers, in the form of breastmilk proteins.

Of particular interest, the identification of *Ruminococcus gnavus* emerged as a significant enriched species within the HEU gut microbiome. Its multifaceted involvement in metabolizing human milk oligosaccharides and mucins, as well as its potential role in various diseases like Crohn's, IBD, and even autoimmune diseases in infants, emphasizes its substantial impact on gut health and immune responses [119-121].

Moreover, our investigation into metabolic pathways highlighted crucial insights into bacterial metabolism within the gut ecosystem, and how these are linked to gut inflammation when perturbated [106]. The alterations in these metabolic pathways could lead to dysregulated energy production, affecting nutrient breakdown and immune functions, ultimately contributing to the development of inflammatory bowel disease, metabolic disorders, and autoimmune diseases [123].

The complex relationship between inflammation and gut dysbiosis became evident through our findings [106, 107]. The impact of inflammatory molecules like NO, TNF, and their influence on the gut barrier integrity and microbial balance underscores the complex interplay that can contribute to gut dysbiosis [106, 108]. Specifically, our research indicates a potential role of these inflammatory molecules in disrupting the mucus layer, weakening intercellular junctions, and influencing the gut microbiome composition [106, 108]. Together the gut inflammation, effected gut barrier and bacterial dysbiosis could be responsible for the weakened immune development and altered overall health in the HEU infants.

In addition to the findings detailed earlier, it's crucial to consider the transmission of the maternal gut microbiome to the infant. This transfer might be facilitated by the immune cells tolerant to the maternal gut microbes present in breastmilk, shaping the initial microbial colonization and potentially influencing the infant's gut microbiota development. Therefore, a dysbiosis in the maternal microbiome due to HIV infection could be influencing the infant via vertical transmission, predisposing them to similar health effects despite being HIV-free. Additionally, the presence of ART metabolites in HIV infected mothers' breastmilk could be influencing the gut microbiome of the infants [70]. Furthermore, our study adds to the knowledge surrounding differences in microbiome associated health issues among HIV-exposed infants [109].

While previous research showed no compositional differences due to HIV exposure using 16S analysis, our investigation into the functional level of the microbiome unveils potential dysbiosis in these infants, and subsequently gut inflammation associated with HIV exposure. This not only deepens our understanding of gut microbiome dynamics in infancy but also allows for potential therapeutic interventions aimed at modulating the gut ecosystem for improved health outcomes in infancy and future development. By

addressing and possibly rectifying dysbiosis in the microbiome, these interventions could pave the way for improved health trajectories and resilience against associated health complications.

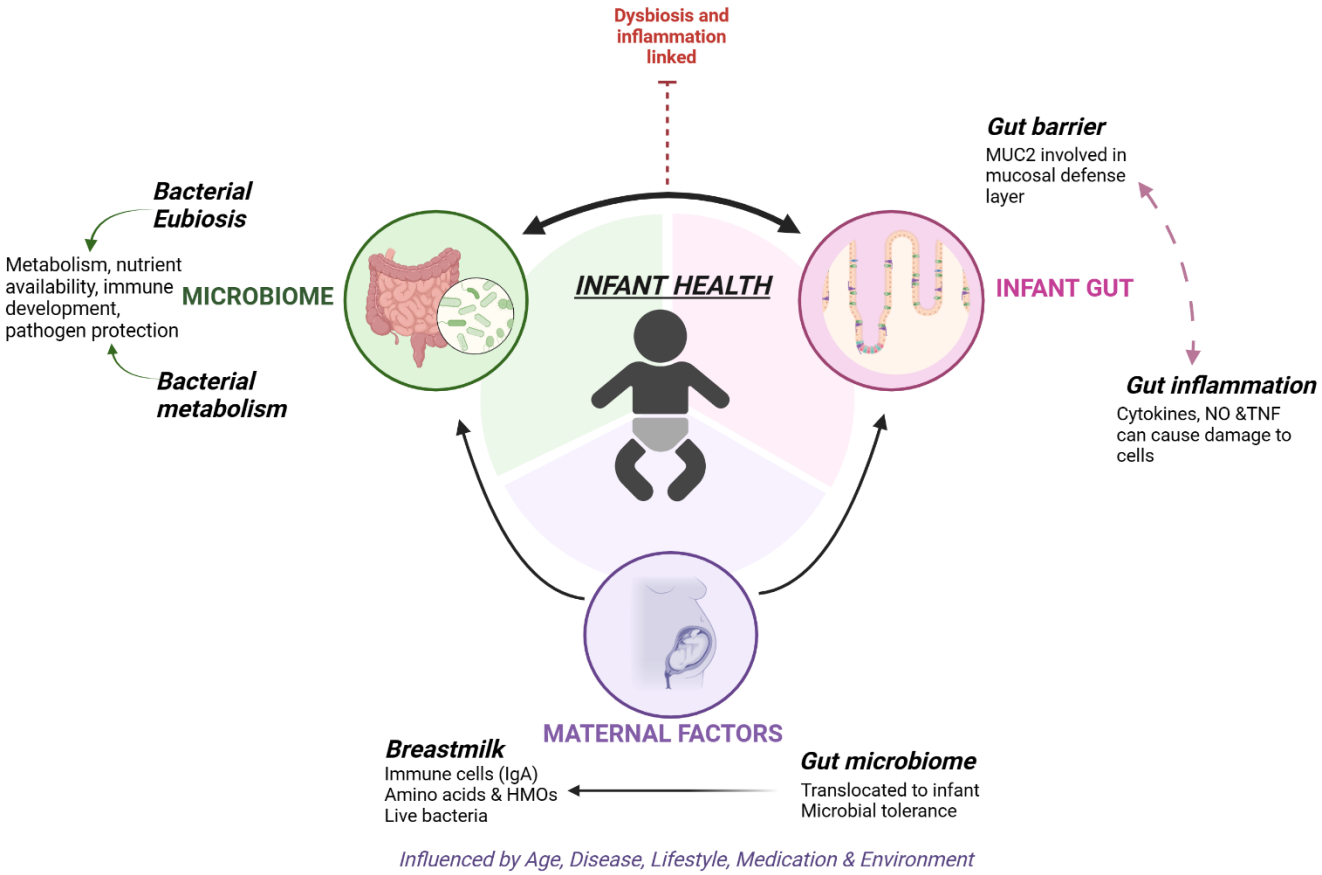


Figure 20: Overview of the intricate network of factors effecting infant's health. Created in Biorender.com

Chapter 5: Challenges and future work

Challenges

Metaproteome analysis is an emerging field of study that enables researchers to unravel the interactions between the microbiome and the host. However, this undertaking presents several challenges that need to be addressed.

One of the initial hurdles is the complexity of the samples due to their high microbial diversity. Potential solutions to mitigate this complexity include the use of enrichment strategies and sample fractionation. However, it's important to note that these approaches come with increased costs and time requirements [124].

Another significant challenge is the size of the human gut database used for searching samples for microbial protein identification. The database, such as the one found in MDB (Microbiome Database for sequencing, research, and project) contains a staggering 9.9 million genes in the microbiome metagenome. Conducting

searches of this magnitude demands substantial computational resources and necessitates access to computers with multiple cores. This, in turn, incurs additional costs and extends the time needed for analysis.

Nonetheless, we implemented a targeted search approach, utilising 16S data to identify the species of interest and construct an *in-silico* library. This approach helped reduce both analysis time and computational demands.

Following protein identification, the heterogeneity between samples introduces the challenge of missing data (see Appendix L). Statistical analysis becomes complicated due to the presence of missing values, and imputation is not feasible since the missing data have biological significance [88]. To address this issue, we employed a 10% presence filtering approach, based on *McGurk*'s research on dealing with missing values in Data-Independent Acquisition (DIA) studies [88].

Furthermore, the absence or insufficient metadata that could account for unknown clustering is not available either due to missed follow up in clinics or human error from clinicians (see Appendix K). Improved metadata collection in clinics could strengthen the analysis as well as increasing the sample size for subgroup frequencies.

Technical limitations also come into play, as there is often a lack of standardized protocols for sample preparation and analysis. This lack of standardization can result in variations between research groups, making it difficult to compare research findings [124].

Furthermore, as metaproteomics is a new growing field, software development and analysis tools are still in their pilot stages. The initial versions of these tools may have limitations in terms of computational capabilities and efficiency. Consequently, our method could not be supported by these early versions, leading us to explore alternative analysis methods.

Lastly, the microbiome is composed of thousands of species, with a substantial portion of them still uncharacterized and uncultivable [83, 125]. As evidenced by the KEGG analysis, our understanding of the functional aspects of microbial proteins remains limited. Therefore, further research is essential to achieve a comprehensive understanding of microbial functions within the microbiome.

Future work

In this study, we focused on assessing secreted soluble human and bacterial proteins. Future research undertakings could extend to evaluating intracellular proteins following the deliberate lysis of recovered bacterial cells present in stool samples. Given that the gut microbiome is a dynamic entity continually influenced by environmental factors, lifestyle, and health, exploring multiple time points in longitudinal studies would provide valuable insights into the developmental trajectory of the microbiome in infants and their development [126].

Moreover, investigating the maternal microbiome, encompassing both breastmilk and gut microbiota, could shed light on the vertical transfer effects from mothers to infants and their impact on the development of the infant's gut microbiome.

It is worth noting that the microbiome resides within the mucosal layers of the gastrointestinal (GI) tract, with distinct sections of microbiome interest found in the small and large intestines. Instead of using fecal samples, collecting mucosal biopsies, or performing fecal aspiration sampling could give a more accurate representation of the microbial communities residing in different intestinal regions [55, 56]. This approach would provide a more in-depth understanding of the microbiota present in specific gut segments. Additionally, to gain deeper insights into the functioning of the mucosal immune system within the gut, it is essential to focus on the mucosa-adherent bacteria (through mucosal biopsy sampling) [55, 56]. Investigating these bacterial communities will help us better understand their role and impact on the overall gut ecosystem and immune responses.

To extend this research, conducting follow-up studies on these infants to track the development of metabolic disorders, autoimmune diseases, and other common health complications would be invaluable. This longitudinal approach aims to assess if any of these conditions can be predicted based on the metaproteomes of infants at one week old. However, to achieve robust and reliable results, a study with a larger sample size, adequately powered to account for the frequency of health issues, would be essential. This necessitates a larger cohort, involving multiple follow-ups over several years, aimed at identifying potential predictors of childhood diseases and distinct subgroups within these conditions.

As previously highlighted, complementing this research with a metabolomic study of the gut microbiome would offer a comprehensive understanding of the microbiome's functionality and its interactions with host cells. Metabolomic pathways tend to have more comprehensive annotations compared to bacterial pathways derived from proteins, providing deeper insights into the functional aspects of the microbiome.

Conflict of Interest

I declare no conflict of interest.

Funding

The research is funded by National Institute of Health (NIH) and student funded by National Research Fund (NRF).

References

1. Cantalapiedra, C.P., et al., *eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale.* Mol Biol Evol, 2021. **38**(12): p. 5825-5829.
2. McDermott, A.J. and G.B. Huffnagle, *The microbiome and regulation of mucosal immunity.* Immunology, 2014. **142**(1): p. 24-31.
3. Sharp, P.M. and B.H. Hahn, *Origins of HIV and the AIDS pandemic.* Cold Spring Harb Perspect Med, 2011. **1**(1): p. a006841.
4. UNAIDS, *UNAIDS Fact Sheet 2022.* 2022.
5. Klimas, N., A.O. Koneru, and M.A. Fletcher, *Overview of HIV.* Psychosom Med, 2008. **70**(5): p. 523-30.
6. Singh, R., *Care and Management of a Neonate Born to an HIV Positive Mother*, in *HIV Infection in Children and Adolescents*. 2020. p. 203-221.
7. Kirchhoff, F., *HIV Life Cycle: Overview*, in *Encyclopedia of AIDS*. 2013. p. 1-9.
8. Arts, E.J. and D.J. Hazuda, *HIV-1 antiretroviral drug therapy.* Cold Spring Harb Perspect Med, 2012. **2**(4): p. a007161.
9. Services, D.o.H.a.H., *Panel on Antiretroviral Guidelines for Adults and Adolescents. Guidelines for the Use of Antiretroviral Agents in Adults and Adolescents with HIV.* 2021.
10. Yeni, P., *Update on HAART in HIV.* J Hepatol, 2006. **44**(1 Suppl): p. S100-3.
11. Research, A.W.G.o.t.O.o.A. and A.C. (OARAC), *Panel on Antiretroviral Guidelines for Adults and Adolescents. Guidelines for the Use of Antiretroviral Agents in Adults and Adolescents with HIV. Department of Health and Human Services*.
12. Watts, D.H. and L.M. Mofenson, *Antiretrovirals in pregnancy: a note of caution.* J Infect Dis, 2012. **206**(11): p. 1639-41.
13. Gray, G.E. and J.A. McIntyre, *HIV and pregnancy.* BMJ, 2007. **334**(7600): p. 950-3.
14. Charles S. Chasela, M.G.H., Denise J. Jamieson, Dumbani Kayira, Mina C. Hosseinipour, Athena P. Kourtis, Francis Martinson, *Maternal or infant antiretroviral drugs to reduce HIV-1 transmission.* The New England journal of medicine, 2010.
15. Lohman-Payne, B., et al., *In-utero infection with HIV-1 associated with suppressed lymphoproliferative responses at birth.* Clin Exp Immunol, 2014. **178**(1): p. 86-93.
16. Tejiokem, M.C., et al., *HIV-infected children living in Central Africa have low persistence of antibodies to vaccines used in the Expanded Program on Immunization.* PLoS One, 2007. **2**(12): p. e1260.
17. Slogrove, A., et al., *HIV-exposed uninfected infants are at increased risk for severe infections in the first year of life.* J Trop Pediatr, 2012. **58**(6): p. 505-8.
18. Taron-Brocard, C., et al., *Increased risk of serious bacterial infections due to maternal immunosuppression in HIV-exposed uninfected infants in a European country.* Clin Infect Dis, 2014. **59**(9): p. 1332-45.
19. Dzanibe, S., et al., *Impact of maternal HIV exposure, feeding status, and microbiome on infant cellular immunity.* J Leukoc Biol, 2019. **105**(2): p. 281-289.
20. Mazzola, T.N., et al., *Impaired Bacillus Calmette-Guerin cellular immune response in HIV-exposed, uninfected infants.* AIDS, 2011. **25**(17): p. 2079-87.
21. Mansoor, N., et al., *HIV-1 infection in infants severely impairs the immune response induced by Bacille Calmette-Guerin vaccine.* J Infect Dis, 2009. **199**(7): p. 982-90.
22. Tchakoute, C.T., et al., *Delaying BCG vaccination until 8 weeks of age results in robust BCG-specific T-cell responses in HIV-exposed infants.* J Infect Dis, 2015. **211**(3): p. 338-46.
23. Nyemba, D.C., et al., *Lower birth weight-for-age and length-for-age z-scores in infants with in-utero HIV and ART exposure: a prospective study in Cape Town, South Africa.* BMC Pregnancy Childbirth, 2021. **21**(1): p. 354.
24. White, M., et al., *Does in utero HIV exposure and the early nutritional environment influence infant development and immune outcomes? Findings from a pilot study in Pretoria, South Africa.* Pilot Feasibility Stud, 2020. **6**(1): p. 192.

25. Ruperez, M., et al., *Maternal HIV infection is an important health determinant in non-HIV-infected infants.* AIDS, 2017. **31**(11): p. 1545-1553.
26. Arikawa, S., et al., *Mortality risk and associated factors in HIV-exposed, uninfected children.* Trop Med Int Health, 2016. **21**(6): p. 720-34.
27. Ramokolo, V., et al., *In Utero ART Exposure and Birth and Early Growth Outcomes Among HIV-Exposed Uninfected Infants Attending Immunization Services: Results From National PMTCT Surveillance, South Africa.* Open Forum Infect Dis, 2017. **4**(4): p. ofx187.
28. Chalashika, P., et al., *Birthweight, HIV exposure and infant feeding as predictors of malnutrition in Botswanan infants.* J Hum Nutr Diet, 2017. **30**(6): p. 779-790.
29. Bork, K.A., et al., *Morbidity in relation to feeding mode in African HIV-exposed, uninfected infants during the first 6 mo of life: the Kesho Bora study.* Am J Clin Nutr, 2014. **100**(6): p. 1559-68.
30. Munyagwa, M., et al., *Mortality of HIV-infected and uninfected children in a longitudinal cohort in rural south-west Uganda during 8 years of follow-up.* Trop Med Int Health, 2012. **17**(7): p. 836-43.
31. Rollins, N.C., et al., *Exclusive breastfeeding, diarrhoeal morbidity and all-cause mortality in infants of HIV-infected and HIV uninfected mothers: an intervention cohort study in KwaZulu Natal, South Africa.* PLoS One, 2013. **8**(12): p. e81307.
32. Koyanagi, A., et al., *Morbidity among human immunodeficiency virus-exposed but uninfected, human immunodeficiency virus-infected, and human immunodeficiency virus-unexposed infants in Zimbabwe before availability of highly active antiretroviral therapy.* Pediatr Infect Dis J, 2011. **30**(1): p. 45-51.
33. Adler, C., et al., *Severe Infections in HIV-Exposed Uninfected Infants Born in a European Country.* PLoS One, 2015. **10**(8): p. e0135375.
34. Dauby, N., et al., *Uninfected but not unaffected: chronic maternal infections during pregnancy, fetal immunity, and susceptibility to postnatal infections.* The Lancet Infectious Diseases, 2012. **12**(4): p. 330-340.
35. Reikie, B.A., et al., *Antibody responses to vaccination among South African HIV-exposed and unexposed uninfected infants during the first 2 years of life.* Clin Vaccine Immunol, 2013. **20**(1): p. 33-8.
36. Reikie, B.A., et al., *Altered innate immune development in HIV-exposed uninfected infants.* J Acquir Immune Defic Syndr, 2014. **66**(3): p. 245-255.
37. Zash, R., et al., *Comparative Safety of Antiretroviral Treatment Regimens in Pregnancy.* JAMA Pediatr, 2017. **171**(10): p. e172222.
38. Wedderburn, C.J., et al., *Early structural brain development in infants exposed to HIV and antiretroviral therapy in utero in a South African birth cohort.* J Int AIDS Soc, 2022. **25**(1): p. e25863.
39. Alcaide, M.L., et al., *Maternal Factors Associated With Infant Neurodevelopment in HIV-Exposed Uninfected Infants.* Open Forum Infect Dis, 2019. **6**(10): p. ofz351.
40. Yang, I., et al., *The Infant Microbiome: Implications for Infant Health and Neurocognitive Development.* Nurs Res, 2016. **65**(1): p. 76-88.
41. Boudry, G., et al., *The Relationship Between Breast Milk Components and the Infant Gut Microbiota.* Front Nutr, 2021. **8**: p. 629740.
42. Ferretti, P., et al., *Mother-to-Infant Microbial Transmission from Different Body Sites Shapes the Developing Infant Gut Microbiome.* Cell Host Microbe, 2018. **24**(1): p. 133-145 e5.
43. Guarner, F. and J.-R. Malagelada, *Gut flora in health and disease.* The Lancet, 2003. **361**(9356): p. 512-519.
44. Gonzalez, R., et al., *Breast milk and gut microbiota in African mothers and infants from an area of high HIV prevalence.* PLoS One, 2013. **8**(11): p. e80299.
45. Bender, J.M., et al., *Maternal HIV infection influences the microbiome of HIV-uninfected infants.* Sci Transl Med, 2016. **8**(349): p. 349ra100.
46. Pannaraj, P.S., et al., *Association Between Breast Milk Bacterial Communities and Establishment and Development of the Infant Gut Microbiome.* JAMA Pediatr, 2017. **171**(7): p. 647-654.
47. Van Niekerk, E., et al., *Human milk oligosaccharides differ between HIV-infected and HIV-uninfected mothers and are related to necrotizing enterocolitis incidence in their preterm very-low-birth-weight infants.* J Nutr, 2014. **144**(8): p. 1227-33.

48. Kuhn, L., et al., *Oligosaccharide composition of breast milk influences survival of uninfected children born to HIV-infected mothers in Lusaka, Zambia.* J Nutr, 2015. **145**(1): p. 66-72.
49. Ciaranello, A.L., et al., *Individualizing the WHO HIV and infant feeding guidelines: optimal breastfeeding duration to maximize infant HIV-free survival.* AIDS, 2014. **28 Suppl 3**: p. S287-99.
50. Desmonde, S., et al., *Health and survival of HIV perinatally exposed but uninfected children born to HIV-infected mothers.* Curr Opin HIV AIDS, 2016. **11**(5): p. 465-476.
51. Gritz, E.C. and V. Bhandari, *The human neonatal gut microbiome: a brief review.* Front Pediatr, 2015. **3**: p. 17.
52. Shi, N., et al., *Interaction between the gut microbiome and mucosal immune system.* Mil Med Res, 2017. **4**: p. 14.
53. Praveen, P., et al., *The role of breast-feeding in infant immune system: a systems perspective on the intestinal microbiome.* Microbiome, 2015. **3**: p. 41.
54. Qin, J., et al., *A human gut microbial gene catalogue established by metagenomic sequencing.* Nature, 2010. **464**(7285): p. 59-65.
55. Romano-Keeler, J., et al., *Early life establishment of site-specific microbial communities in the gut.* Gut Microbes, 2014. **5**(2): p. 192-201.
56. Liu, J., et al., *Inside Out: HIV, the Gut Microbiome, and the Mucosal Immune System.* J Immunol, 2017. **198**(2): p. 605-614.
57. Zimmermann, P. and N. Curtis, *The influence of the intestinal microbiome on vaccine responses.* Vaccine, 2018. **36**(30): p. 4433-4439.
58. Tanaka, M. and J. Nakayama, *Development of the gut microbiota in infancy and its impact on health in later life.* Allergol Int, 2017. **66**(4): p. 515-522.
59. Aagaard, K., et al., *The placenta harbors a unique microbiome.* Sci Transl Med, 2014. **6**(237): p. 237ra65.
60. Ardissone, A.N., et al., *Meconium microbiome analysis identifies bacteria correlated with premature birth.* PLoS One, 2014. **9**(3): p. e90784.
61. Jimenez, E., et al., *Is meconium from healthy newborns actually sterile?* Res Microbiol, 2008. **159**(3): p. 187-93.
62. Wang, Z.K., et al., *Intestinal microbiota pathogenesis and fecal microbiota transplantation for inflammatory bowel disease.* World J Gastroenterol, 2014. **20**(40): p. 14805-20.
63. Kau, A.L., et al., *Human nutrition, the gut microbiome and the immune system.* Nature, 2011. **474**(7351): p. 327-36.
64. Dillon, S.M., et al., *An altered intestinal mucosal microbiome in HIV-1 infection is associated with mucosal and systemic immune activation and endotoxemia.* Mucosal Immunol, 2014. **7**(4): p. 983-94.
65. Gough, E.K., et al., *Maternal fecal microbiome predicts gestational age, birth weight and neonatal growth in rural Zimbabwe.* EBioMedicine, 2021. **68**: p. 103421.
66. Vazquez-Castellanos, J.F., et al., *Altered metabolism of gut microbiota contributes to chronic immune activation in HIV-infected individuals.* Mucosal Immunol, 2015. **8**(4): p. 760-72.
67. Lozupone, C.A., et al., *Alterations in the gut microbiota associated with HIV-1 infection.* Cell Host Microbe, 2013. **14**(3): p. 329-39.
68. Rocafort, M., et al., *HIV-associated gut microbial alterations are dependent on host and geographic context.* Nat Commun, 2024. **15**(1): p. 1055.
69. Marcobal, A., et al., *Consumption of human milk oligosaccharides by gut-related microbes.* J Agric Food Chem, 2010. **58**(9): p. 5334-40.
70. Grant-Beurmann, S., et al., *Dynamics of the infant gut microbiota in the first 18 months of life: the impact of maternal HIV infection and breastfeeding.* Microbiome, 2022. **10**(1): p. 61.
71. Belkaid, Y. and T.W. Hand, *Role of the microbiota in immunity and inflammation.* Cell, 2014. **157**(1): p. 121-41.
72. Kuhn, K.A., I. Pedraza, and M.K. Demoruelle, *Mucosal immune responses to microbiota in the development of autoimmune disease.* Rheum Dis Clin North Am, 2014. **40**(4): p. 711-25.
73. Atarashi, K., et al., *ATP drives lamina propria T(H)17 cell differentiation.* Nature, 2008. **455**(7214): p. 808-12.

74. Cebula, A., et al., *Thymus-derived regulatory T cells contribute to tolerance to commensal microbiota.* Nature, 2013. **497**(7448): p. 258-62.
75. Ayabe, T., et al., *The role of Paneth cells and their antimicrobial peptides in innate host defense.* Trends Microbiol, 2004. **12**(8): p. 394-8.
76. Furusawa, Y., et al., *Commensal microbe-derived butyrate induces the differentiation of colonic regulatory T cells.* Nature, 2013. **504**(7480): p. 446-50.
77. Jin, P., et al., *Mining the fecal proteome: from biomarkers to personalised medicine.* Expert Rev Proteomics, 2017. **14**(5): p. 445-459.
78. Aakko, J., et al., *Data-Independent Acquisition Mass Spectrometry in Metaproteomics of Gut Microbiota-Implementation and Computational Analysis.* J Proteome Res, 2020. **19**(1): p. 432-436.
79. Kitata, R.B., J.C. Yang, and Y.J. Chen, *Advances in data-independent acquisition mass spectrometry towards comprehensive digital proteome landscape.* Mass Spectrom Rev, 2023. **42**(6): p. 2324-2348.
80. Neagu, A.N., et al., *Applications of Tandem Mass Spectrometry (MS/MS) in Protein Analysis for Biomedical Research.* Molecules, 2022. **27**(8).
81. Gillet, L.C., et al., *Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis.* Mol Cell Proteomics, 2012. **11**(6): p. O111 016717.
82. Ludwig, C., et al., *Data-independent acquisition-based SWATH-MS for quantitative proteomics: a tutorial.* Mol Syst Biol, 2018. **14**(8): p. e8126.
83. Kanehisa, M., et al., *KEGG as a reference resource for gene and protein annotation.* Nucleic Acids Res, 2016. **44**(D1): p. D457-62.
84. Sajulga, R., et al., *Survey of metaproteomics software tools for functional microbiome analysis.* PLoS One, 2020. **15**(11): p. e0241503.
85. Duan, H., et al., *Assessing the Dark Field of Metaproteome.* Anal Chem, 2022. **94**(45): p. 15648-15654.
86. Lou, R., et al., *Benchmarking commonly used software suites and analysis workflows for DIA proteomics and phosphoproteomics.* Nat Commun, 2023. **14**(1): p. 94.
87. Demichev, V., et al., *DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput.* Nat Methods, 2020. **17**(1): p. 41-44.
88. McGurk, K.A., et al., *The use of missing values in proteomic data-independent acquisition mass spectrometry to enable disease activity discrimination.* Bioinformatics, 2020. **36**(7): p. 2217-2223.
89. Iwase, C., et al., *Longitudinal gut microbiota composition of South African and Nigerian infants in relation to tetanus vaccine responses* Microbiol Spectrum, 2024. **0**(0).
90. Sari, R.N., et al., *Comparative Proteomics of Human Milk From Eight Cities in China During Six Months of Lactation in the Chinese Human Milk Project Study.* Front Nutr, 2021. **8**: p. 682429.
91. van Herwijnen, M.J., et al., *Comprehensive Proteomic Analysis of Human Milk-derived Extracellular Vesicles Unveils a Novel Functional Proteome Distinct from Other Milk Components.* Mol Cell Proteomics, 2016. **15**(11): p. 3412-3423.
92. Hettinga, K.A., et al., *Difference in the breast milk proteome between allergic and non-allergic mothers.* PLoS One, 2015. **10**(3): p. e0122234.
93. Huerta-Cepas, J., et al., *eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses.* Nucleic Acids Res, 2019. **47**(D1): p. D309-D314.
94. Rinninella, E., et al., *What is the Healthy Gut Microbiota Composition? A Changing Ecosystem across Age, Environment, Diet, and Diseases.* Microorganisms, 2019. **7**(1).
95. Moore, R.E. and S.D. Townsend, *Temporal development of the infant gut microbiome.* Open Biol, 2019. **9**(9): p. 190128.
96. Jordan, C.K.I., et al., *Symbiotic Firmicutes establish mutualism with the host via innate tolerance and resistance to control systemic immunity.* Cell Host Microbe, 2023. **31**(9): p. 1433-1449 e9.
97. Jackson, C.L., et al., *Evolution of the Gut Microbiome in HIV-Exposed Uninfected*

and Unexposed Infants during the First Year of Life. American Society For Microbiology, 2022. **13**(5): p. 17.

98. Ishizaka, A., et al., *Unique Gut Microbiome in HIV Patients on Antiretroviral Therapy (ART) Suggests Association with Chronic Inflammation.* Microbiol Spectr, 2021. **9**(1): p. e0070821.
99. Turroni, F., et al., *The infant gut microbiome as a microbial organ influencing host well-being.* Ital J Pediatr, 2020. **46**(1): p. 16.
100. Nichols, R.G. and E.R. Davenport, *The relationship between the gut microbiome and host gene expression: a review.* Hum Genet, 2021. **140**(5): p. 747-760.
101. Bull, M.J. and N.T. Plummer, *The Human Gut Microbiome in Health and Disease* Integrative Medicine, 2014. **13**(6): p. 6.
102. Nicholson, J.K., E. Holmes, and J. Kinross, *Host-Gut Microbiota Metabolic Interactions.* Science, 2012. **336**.
103. Fu, Z.D., et al., *RNA-Seq Profiling of Intestinal Expression of Xenobiotic Processing Genes in Germ-Free Mice.* Drug Metab Dispos, 2017. **45**(12): p. 1225-1238.
104. Matziouridou, C., et al., *iNOS- and NOX1-dependent ROS production maintains bacterial homeostasis in the ileum of mice.* Mucosal Immunol, 2018. **11**(3): p. 774-784.
105. Cobo, E.R., et al., *Colonic MUC2 mucin regulates the expression and antimicrobial activity of beta-defensin 2.* Mucosal Immunol, 2015. **8**(6): p. 1360-72.
106. Zeng, M.Y., N. Inohara, and G. Nunez, *Mechanisms of inflammation-driven bacterial dysbiosis in the gut.* Mucosal Immunol, 2017. **10**(1): p. 18-26.
107. Al-Shehri, S.S., *Reactive oxygen and nitrogen species and innate immune response.* Biochimie, 2021. **181**: p. 52-64.
108. Souza, R.F., et al., *Study of tumor necrosis factor receptor in the inflammatory bowel disease.* World J Gastroenterol, 2023. **29**(18): p. 2733-2746.
109. Machiavelli, A., et al., *The impact of in utero HIV exposure on gut microbiota, inflammation, and microbial translocation.* Gut Microbes, 2019. **10**(5): p. 599-614.
110. Cacho, N.T. and R.M. Lawrence, *Innate Immunity and Breast Milk.* Front Immunol, 2017. **8**: p. 584.
111. Field, C.J., *The immunological components of human milk and their effect on immune development in infants.* J Nutr, 2005. **135**(1): p. 1-4.
112. Newburg, D.S. and W.A. Walker, *Protection of the neonate by the innate immune system of developing gut and of human milk.* Pediatr Res, 2007. **61**(1): p. 2-8.
113. Deng, Y., et al., *Effects of glutamine, glutamate, and aspartate on intestinal barrier integrity and amino acid pool of the small intestine in piglets with normal or low energy diet.* Front Vet Sci, 2023. **10**: p. 1202369.
114. Portune, K.J., et al., *Gut microbiota role in dietary protein metabolism and health-related outcomes: The two sides of the coin.* Trends in Food Science & Technology, 2016. **57**: p. 213-232.
115. Nyangahu, D.D., et al., *Bifidobacterium infantis assoicates with T cell immunity in human infants and is sufficient to enhance antigen-specific T cells in mice.* Science Advances, 2023. **9**.
116. Horgan, R.P. and L.C. Kenny, *'Omic' technologies: genomics, transcriptomes, proteomics and metabolomics.* The Obstetrician & Gynaecologist, 2011. **13**: p. 189-195.
117. Wang, Z., et al., *Altered Gut Microbiota and Host Metabolite Profiles in Women With Human Immunodeficiency Virus.* Clinical Infectious Diseases, 2020. **71**(9): p. 2345-2353.
118. Xu, H., et al., *Intestinal mucosal microbiota composition of patients with acquired immune deficiency syndrome in Guangzhou, China.* Experimental and Therapeutic Medicine, 2021. **21**(4).
119. Silverman, G.J., D.F. Azzouz, and A.V. Alekseyenko, *Systemic Lupus Erythematosus and dysbiosis in the microbiome: cause or effect or both?* Current Opinion in Immunology, 2019. **61**: p. 80-85.
120. Crost, E.H., et al., *Ruminococcus gnavus: friend or foe for human health.* FEMS Microbiology Reviews, 2023. **47**(2).
121. Chua, H.-H., et al., *Intestinal Dysbiosis Featuring Abundance of Ruminococcus gnavus Associates With Allergic Diseases in Infants.* Gastroenterology, 2018. **154**(1): p. 154-167.
122. Ma, L. and L. Morel, *Loss of Gut Barrier Integrity In Lupus.* Front Immunol, 2022. **13**: p. 919792.
123. Mohammed, A. and C. Guda, *Application of a hierarchical enzyme classification method reveals the role of gut microbiome in human metabolism.* BMC Genomics, 2015. **16**(S7).
124. Lee, P.Y., et al., *Metaproteomic analysis of human gut microbiota: where are we heading?* J Biomed Sci, 2017. **24**(1): p. 36.

125. Dias, C.K., et al., *Database limitations for studying the human gut microbiome.* PeerJ Comput Sci, 2020. **6**: p. e289.
126. Gagne, D., et al., *Proteomics Profiling of Stool Samples from Preterm Neonates with SWATH/DIA Mass Spectrometry for Predicting Necrotizing Enterocolitis.* Int J Mol Sci, 2022. **23**(19).

Chapter 6: Appendix

A- Metadata table for InFANT samples

METADATA table for samples										
Sample number	HIV status	Age of mother (years)	Mother's housing. 0= House, 1= Shack	Running water in house. 0= No 1= Yes	Gender of baby. 0= Male 1= Female	Weight of baby (grams)	Gestational age (weeks)	1st pregnancy? 0= No 1= Yes	Mother deliver vaginally? 0= No 1= Yes	Baby age
1	Negative	24	0	0	0	3130	37	0	1	7
2	Positive	28	0	1	0	2550	39	0	1	5
3	Positive	32	1	0	0	3200	39	0	1	7
4	Negative	23	0	0	0	3100	38	1	1	7
5	Negative	23	1	1	0	3830	41	0	1	6
6	Positive	31	1	1	0	3200	40	0	1	9
7	Positive	28	1	1	0	2880	38	0	1	7
8	Positive	30	1	0	0	3220	38	0	1	5
9	Positive	29	0	1	0	2650	39	0	1	7
10	Positive	19	0	1	1	2660	38	1	1	7
11	Positive	25	0	1	0	3400	39	0	1	6
12	Negative	21	0	0	0	2710	36	1	1	8
13	Positive	25	0	1	0	3800	39	0	1	7
14	Positive	24	1	0	1	3080	39	0	1	8
15	Positive	26	1	0	0	3200	41	0	1	7
16	Positive	23	1	1	1	3600	39	0	1	6
17	Positive	30	0	1	0	3200	39	0	1	7
18	Positive	19	0	1	0	3560	40	1	1	6
19	Positive	24	0	1	0	2900	38	0	1	5
20	Negative	31	0	0	0	3380	37	0	1	8
21	Positive	21	1	0	1	2240	37	1	1	7
22	Positive	25	1	0	0	2460	38	0	1	6
23	Positive	29	1	0	1	2800	39	0	1	8
24	Positive	35	0	1	1	2700	38	0	1	7

29	Positive	26	0	1	0	3400	41	1	1	7
30	Positive	27	1	1	0	3200	38	1	1	8
31	Positive	34	0	1	1	2400	39	0	1	7
32	Positive	28	0	1	1	3270	38	0	1	7
33	Positive	26	0	1	0	3900	40	0	1	6
34	Positive	38	1	0	1	3000	37	0	1	5
35	Negative	22	1	0	1	2910	38	0	1	8
36	Positive	21	0	1	0	3480	39	0	1	5
37	Positive	34	1	0	0	3520	39	0	1	7
38	Positive	36	1	0	0	3600	42	0	1	5
39	Positive	27	1	0	0	3410	40	0	1	9
40	Positive	29	1	0	0	2720	38	0	1	10
41	Positive	31	1	1	1	3280	40	0	1	8
42	Positive	38	0	1	0	3200	41	0	1	8
43	Positive	28	1	1	1	3120	40	0	1	5
44	Positive	30	0	1	0	2800	38	0	1	6
47	Positive	35	1	0	0	3460	39	0	1	7
48	Positive	32	1	1	1	3000	39	0	1	6
49	Positive	24	0	1	1	4000	39	0	1	8
50	Positive	37	0	1	0	2950	39	0	1	5
51	Positive	32	0	1	1	2340	39	0	1	6
52	Positive	34	1	1	0	3080	39	1	1	6
53	Positive	25	1	1	0	3720	40	0	1	6
54	Positive	38	1	1	0	3030	40	0	0	8
55	Positive	32	1	1	0	3700	39	0	1	5
56	Positive	27	1	1	0	3200	40	0	0	6
57	Positive	26	1	0	0	3390	39	0	0	8
58	Positive	26	1	1	0	3100	39	0	1	6
59	Positive	31	1	0	1	3580	40	0	1	6
60	Positive	38	0	1	0	2880	38	0	0	15

61	Negative	26	1	1	0	3200	39	0	1	6
62	Negative	20	1	0	0	3360	40	0	1	7
63	Negative	31	0	1	0	3120	39	0	1	8
64	Negative	26	0	1	1	2440	37	1	1	8
65	Negative	35	0	1	1	3200	39	0	1	-
66	Negative	33	0	1	1	3180	42	0	1	9
67	Negative	22	0	1	1	3100	38	1	1	7
68	Negative	21	0	1	0	3380	39	1	1	
69	Negative	26	1	1	0	2500	40	0	1	
70	Negative	28	0	1	0	3600	40	0	1	
71	Negative	26	0	1	1	3000	40	0	1	
72	Negative	26	0	1	0	3400	40	0	1	
73	Negative	25	1	0	1	4400	41	0	1	
74	Negative	21	1	0	0	3200	39	0	1	
75	Negative	24	1	1	0	3100	39	0	1	
76	Negative	20	1	1	1	2480	37	1	1	
77	Negative	35	1	1	1	3400	39	0	1	
79	Negative	18	1	1	0	3000	38	1	1	
80	Negative	20	1	1	0	3400	40	1	1	
81	Negative	28	0	1	0	3260	37	0	1	
82	Negative	30	0	1	0	3120	35	0	1	
84	Negative	22	1	1	0	2920	39	0	1	
85	Negative	21	1	1	1	3350	39	1	1	
86	Negative	27	0	1	0	2800	38	0	1	
87	Negative	24	0	1	0	3200	38	1	1	
88	Negative	28	1	1	1	3720	39	0	1	
89	Negative	20	0	1	0	2820	38	1	1	
90	Negative	37	1	1	0	3300	40	0	1	
91	Negative	37	1	1	0	3500	42	0	1	
92	Negative	26	1	0	1	2980	40	0	1	

93	Negative	26	1	1	1	3100	39	1	0	
94	Negative	26	1	1	0	3600	41	0	1	
95	Negative	41	0	1	0	3080	40	0	0	
96	Negative	29	1	1	0	2480	38	1	0	

B- Concentration of protein loaded for digestion

Concentration of protein loaded				
Sample	Conc ug/ml	Vt (mL)	Volume sample for 10 ug	DB added 20 uL Final volume (0.5 ug/uL)
S01	580.917	0.0175	17.21416	2.785837
S02	1549.096	0.0175	6.455376	13.54462
S03	785.745	0.0175	12.72678	7.273225
S04	198.624	0.0175	50.34638	-
S05	1130.295	0.0175	8.847248	11.15275
S06	(-)	0.0175	20	0
S07	1347.915	0.0175	7.418865	12.58114
S08	1260.736	0.0175	7.931872	12.06813
S09	373.416	0.0175	26.77978	-
S10	1208.31	0.0175	8.276022	11.72398
S11	526.538	0.0175	18.99198	1.008018
S12	925.602	0.0175	10.80378	9.19622
S13	1388.151	0.0175	7.203825	12.79617
S14	356.42	0.0175	28.05679	-
S15	1051.527	0.0175	9.509979	10.49002
S16	776.792	0.0175	12.87346	7.126541
S17	1891.105	0.0175	5.287914	14.71209
S18	1354.621	0.0175	7.382138	12.61786
S19	107.916	0.0175	92.66467	-
S20	1388.151	0.0175	7.203825	12.79617
S21	1037.384	0.0175	9.639632	10.36037
S22	1803.926	0.0175	5.543464	14.45654
S23	1148.167	0.0175	8.709534	11.29047
S24	1602.745	0.0175	6.239296	13.7607
S25	692.709	0.0175	14.43608	5.563924
S26	1508.86	0.0175	6.627519	13.37248
S27	1515.566	0.0175	6.598194	13.40181
S28	1381.445	0.0175	7.238795	12.7612
S29	1081.246	0.0175	9.248589	10.75141
S30	290.825	0.0175	34.38494	-
S31	533.168	0.0175	18.75581	1.244186
S32	1274.149	0.0175	7.848378	12.15162
S33	1300.973	0.0175	7.686556	12.31344
S34	1542.39	0.0175	6.483443	13.51656
S35	1113.251	0.0175	8.9827	11.0173
S36	1482.036	0.0175	6.747474	13.25253

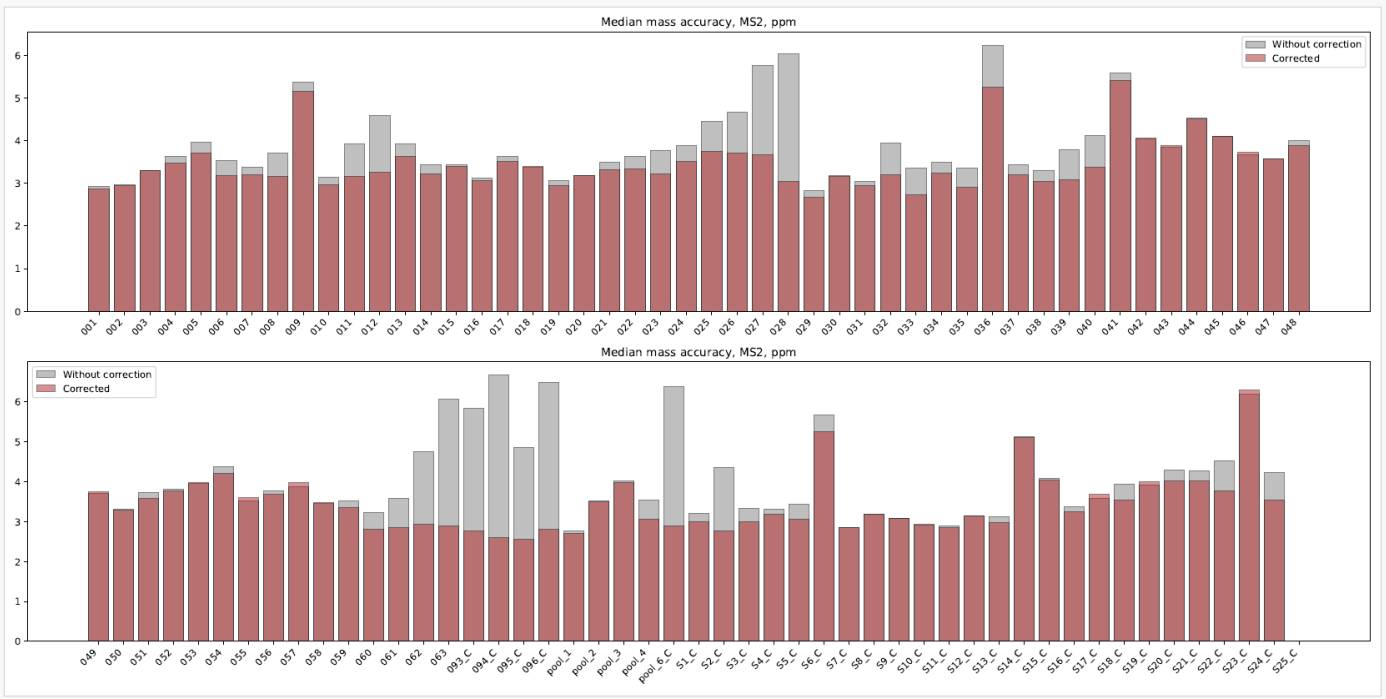
S37	1096.93	0.0175	9.116352	10.88365
S38	(-)	0.0175	20	0
S39	1186.97	0.0175	8.424813	11.57519
S40	692.709	0.0175	14.43608	5.563924
S41	1502.154	0.0175	6.657107	13.34289
S42	1240.618	0.0175	8.060497	11.9395
S43	1904.517	0.0175	5.250676	14.74932
S44	960.277	0.0175	10.41366	9.586338
S45	1441.8	0.0175	6.935776	13.06422
S46	1267.443	0.0175	7.889904	12.1101
S47	1502.154	0.0175	6.657107	13.34289
S48	573.945	0.0175	17.42327	2.576728
S49	1130.27	0.0175	8.847444	11.15256
S50	1083.929	0.0175	9.225697	10.7743
S51	1373.837	0.0175	7.278884	12.72112
S52	1040.377	0.0175	9.6119	10.3881
S53	960.614	0.0175	10.41001	9.589991
S54	1496.707	0.0175	6.681334	13.31867
S55	1553.454	0.0175	6.437267	13.56273
S56	951.217	0.0175	10.51285	9.487152
S57	708.654	0.0175	14.11126	5.888741
S58	1652.762	0.0175	6.050479	13.94952
S59	1373.837	0.0175	7.278884	12.72112
S60	1009.434	0.0175	9.906542	10.09346
S61	1317.229	0.0175	7.591694	12.40831
S62	1222.542	0.0175	8.179678	11.82032
S63	1252.025	0.0175	7.987061	12.01294
S64	(-)	0.0175	20	0
S65	1801.723	0.0175	5.550242	14.44976
S66	1752.069	0.0175	5.707537	14.29246
S67	1354.025	0.0175	7.385388	12.61461
S68	1787.536	0.0175	5.594292	14.40571
S69	278.681	0.0175	35.88332	-
S70	(-)	0.0175	20	0
S71	1237.064	0.0175	8.083656	11.91634
S72	1155.391	0.0175	8.655079	11.34492
S73	821.82	0.0175	12.16811	7.831885
S74	694.497	0.0175	14.39891	5.60109
S75	1624.388	0.0175	6.156164	13.84384
S76	871.241	0.0175	11.47788	8.52212
S77	1009.434	0.0175	9.906542	10.09346
S78	646.401	0.0175	15.47027	4.529727
S79	1702.416	0.0175	5.874007	14.12599
S80	1222.542	0.0175	8.179678	11.82032
S81	525.481	0.0175	19.03018	0.969816
S82	454.996	0.0175	21.97822	-
S83	1794.63	0.0175	5.57218	14.42782
S84	979.762	0.0175	10.20656	9.79344

S85	673.617	0.0175	14.84523	5.154769
S86	1441.262	0.0175	6.938364	13.06164
S87	495.624	0.0175	20.17659	-
S88	1610.201	0.0175	6.210403	13.7896
S89	1237.064	0.0175	8.083656	11.91634
S90	1040.377	0.0175	9.6119	10.3881
S91	1051.004	0.0175	9.514712	10.48529
S92	2028.712	0.0175	4.929236	15.07076
S93	1467.528	0.0175	6.81418	13.18582
S94	970.128	0.0175	10.30792	9.692082
S95	914.705	0.0175	10.93249	9.067514
S96	1417.179	0.0175	7.056272	12.94373
Samples too low so added 20 uL sample to load 5 ug protein instead of 10 ug				

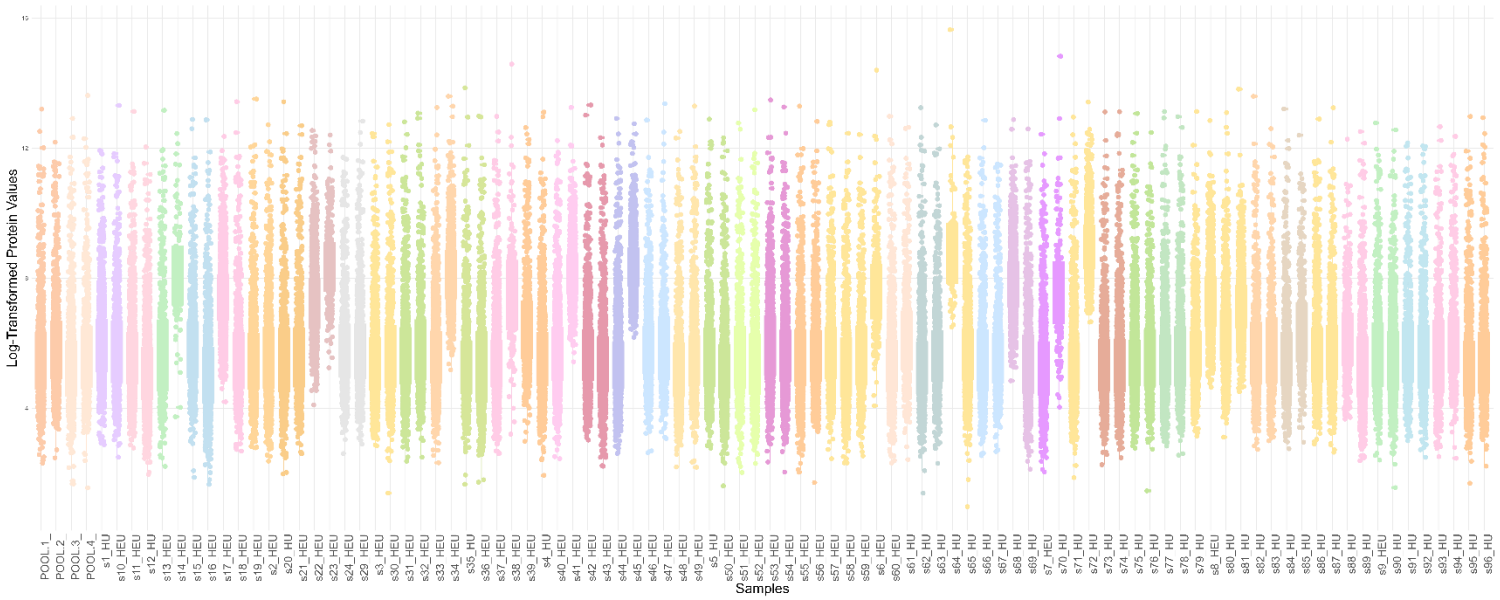
C- Samples of interest

Samples with little sample provided	Samples lost to contamination	Samples with low protein content	Samples lost on Mass Spectrometer
1	70	9	25
3	6	11	26
4	38	19	27
6		81	28
14		82	
15		87	
25			
30			
31			
34			
69			
73			
89			
90			
56			

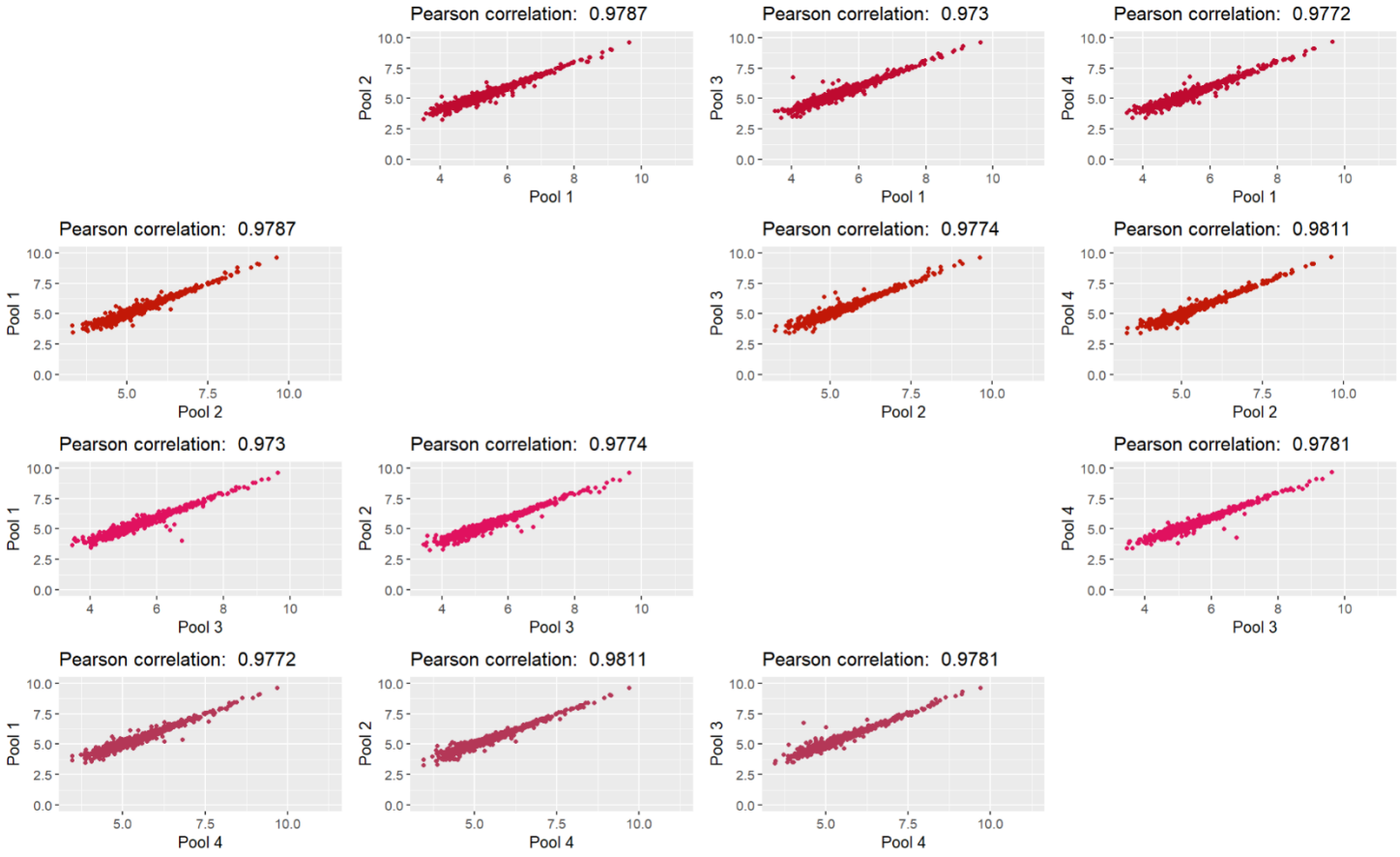
D- DIA-NN results of mass accuracy



E- Protein abundance and distribution per sample

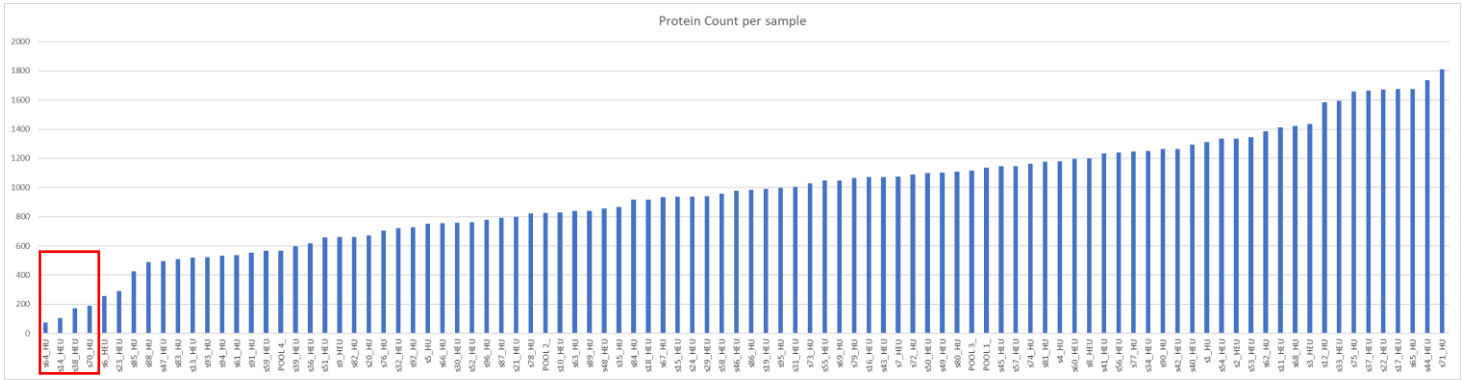


F- Pool analysis



G- Sample protein count and low protein cut off

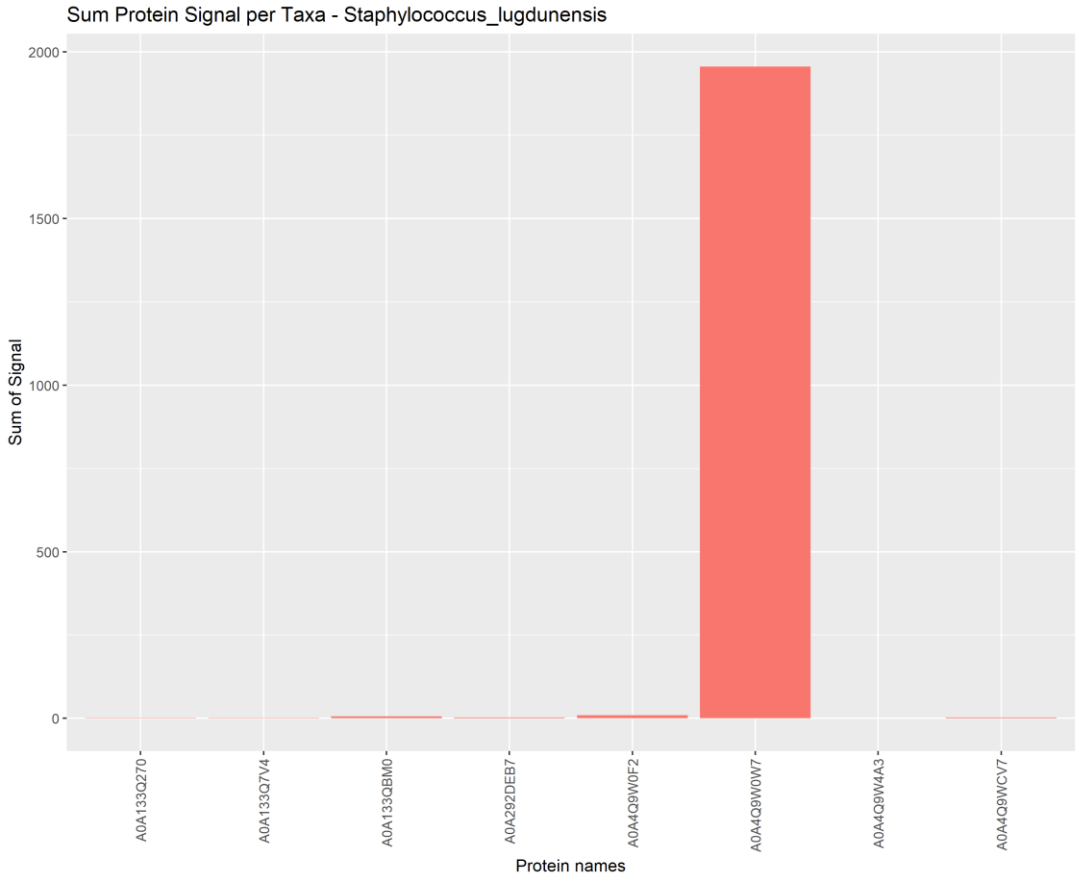
Average protein count	965.40
Standard deviation	385.40
95% rule cut off lower counts	194.59

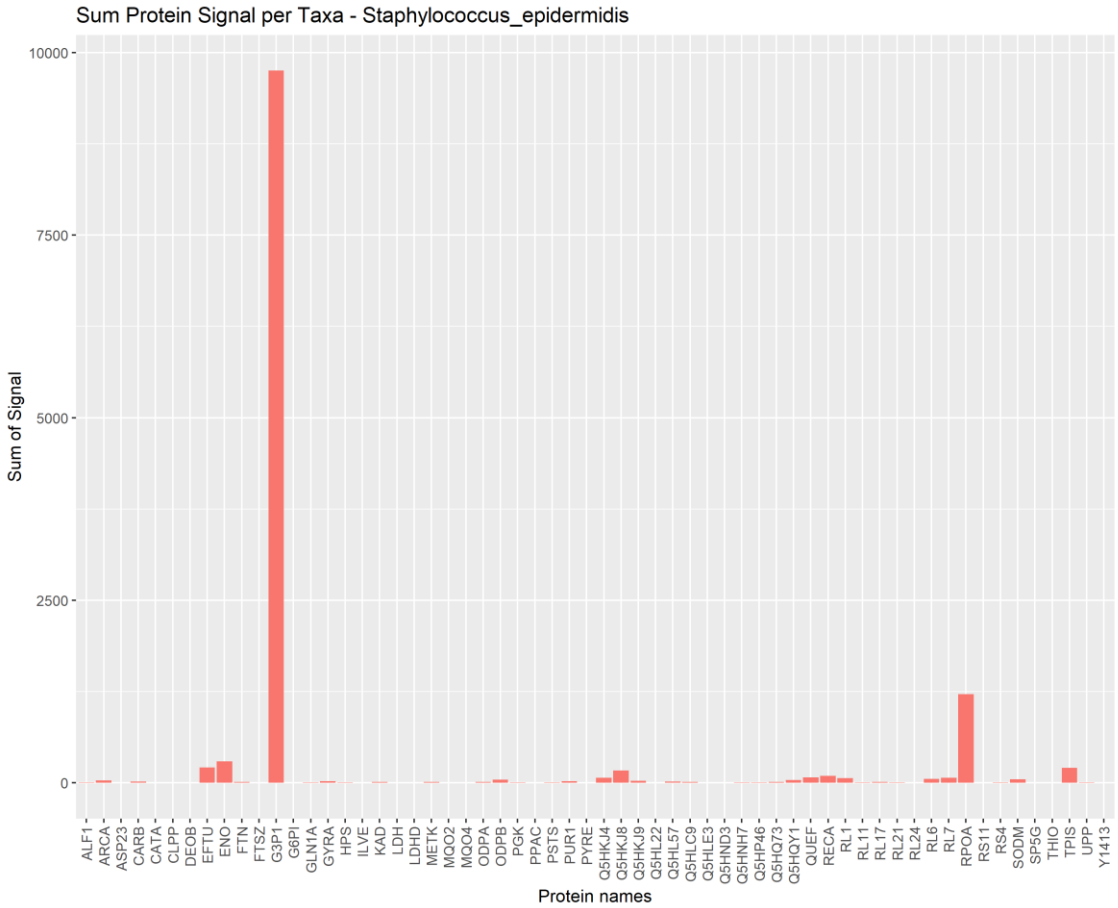


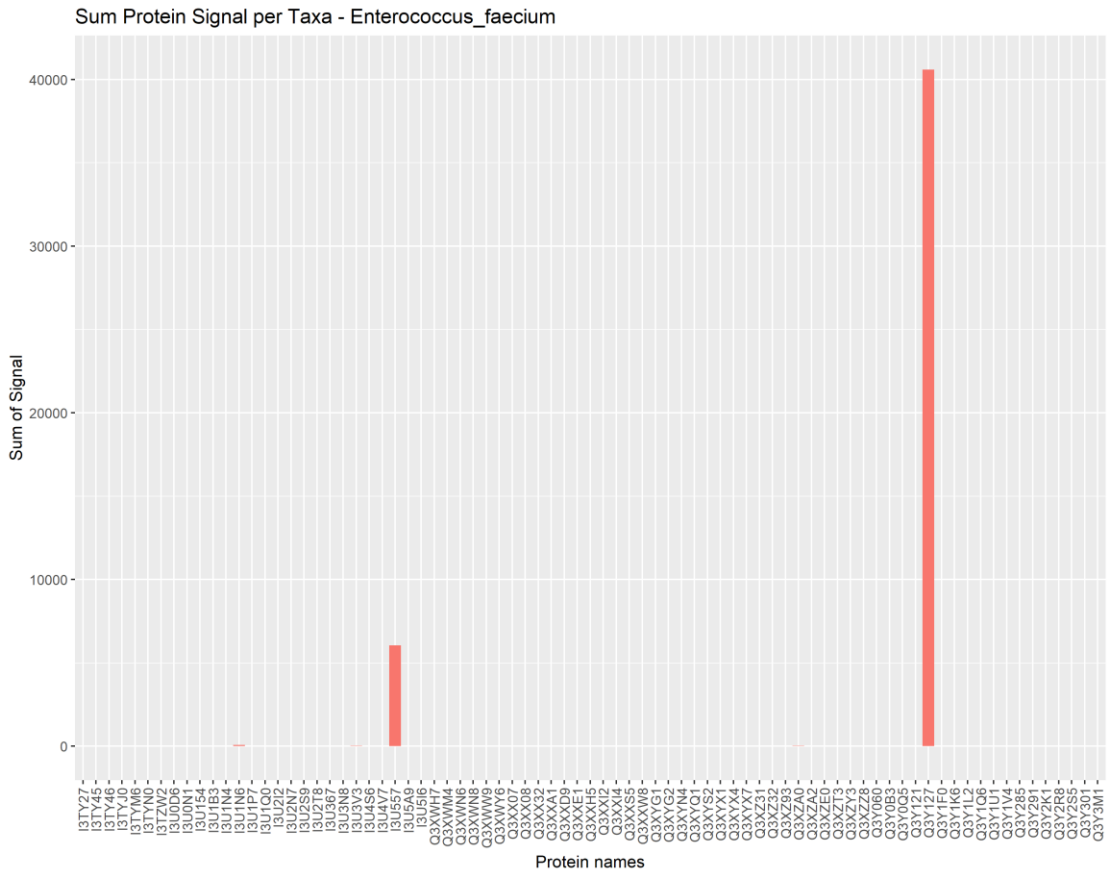
H- Identified proteins from ratio cut off for DIA-NN miss identifications

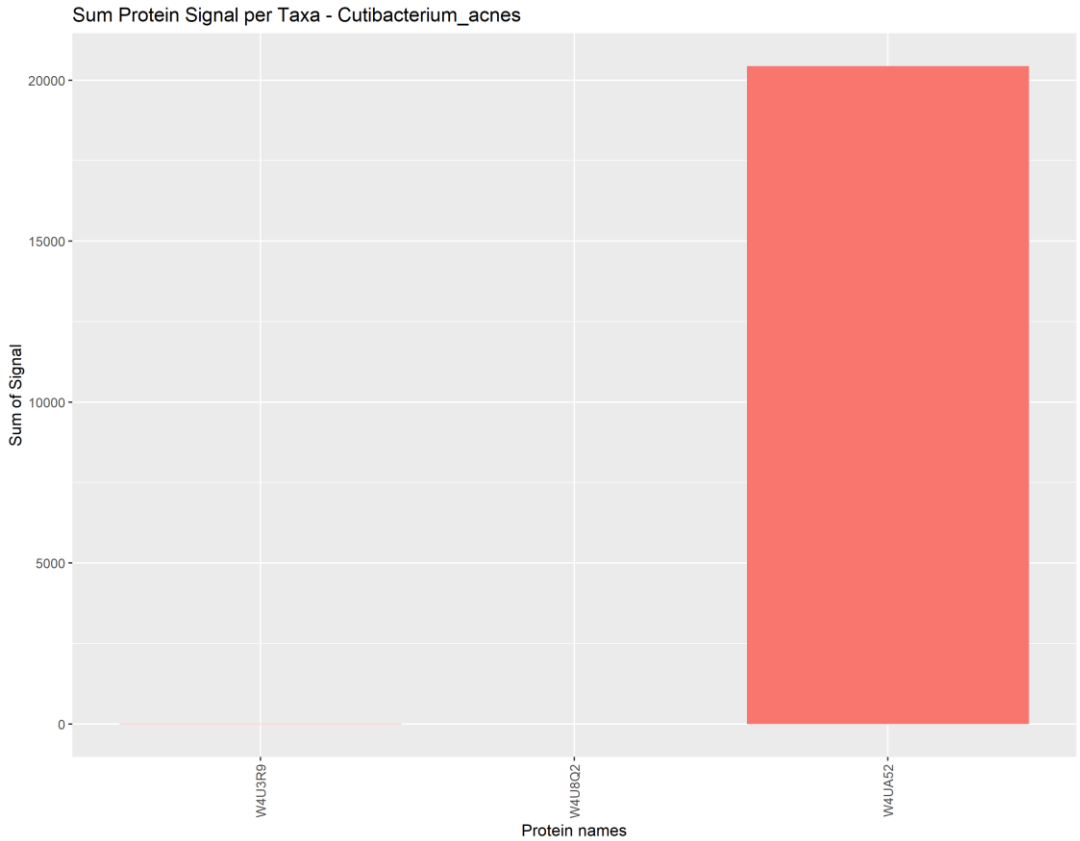
Taxa	Protein names	Sum	Taxa Median	Ratio Sum / Median
Staphylococcus_lugdunensis 1 peptide – 5 taxa on blast	A0A4Q9W0W7	708088.89	1011.60	699.97
Staphylococcus_epidermidis	G3P1	1649287.55	1465.08	1125.73
Enterococcus_faecium 1 peptide – multiple taxa	Q3Y127	4638941.87	135.86	34144.51
Enterococcus_faecium 1 peptide – multiple taxa	I3U557	691230.056	135.86	5087.74
Cutibacterium_acnes 1 peptide – multiple taxa	W4UA52	14604582.2	6433.17	2270.20
Bifidobacterium_breve Longum and breve (multiple peptides)	S2ZLX6	346803.58	732.87	473.21

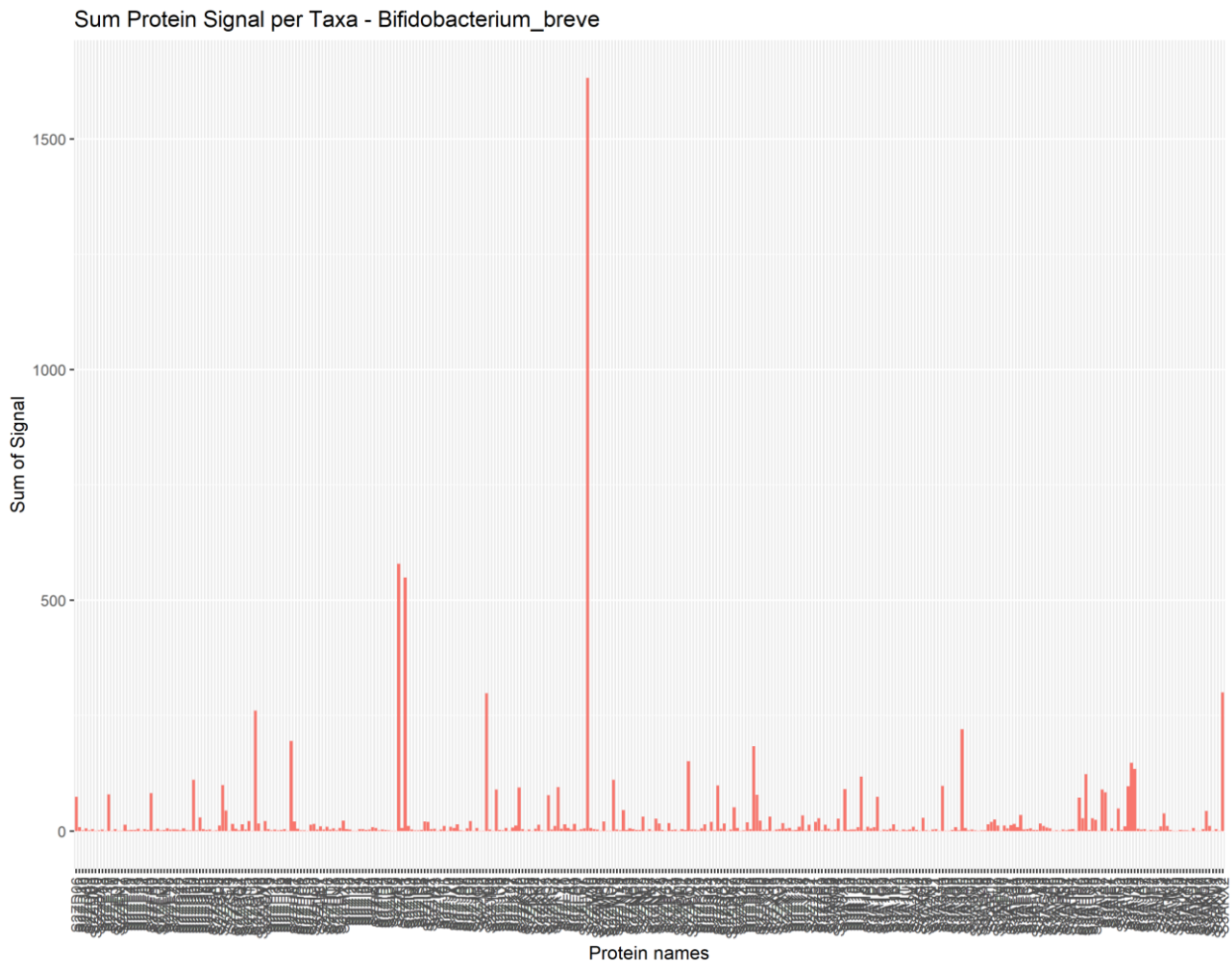
I- Species proteins signals with misidentified protein and with it removed



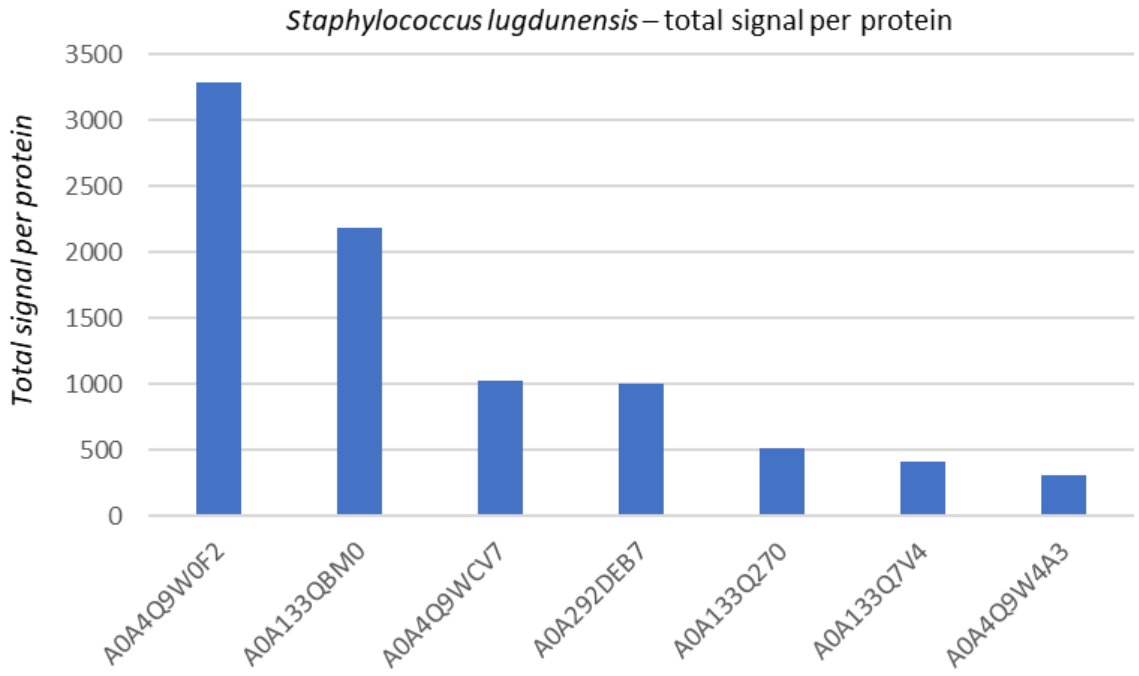


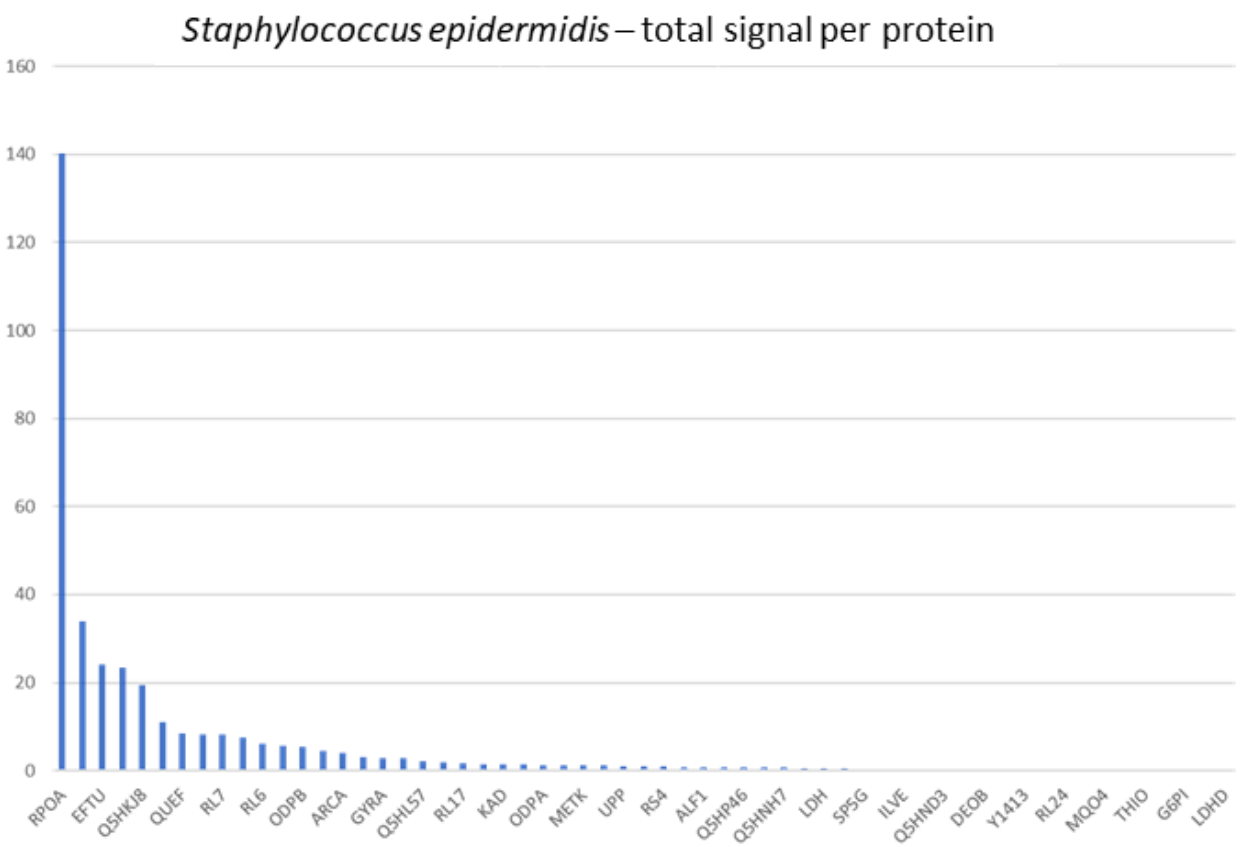

Sum Protein Signal per Taxa - Enterococcus_faecium
Sum of Signal
40000
30000
20000
10000
0
Protein names

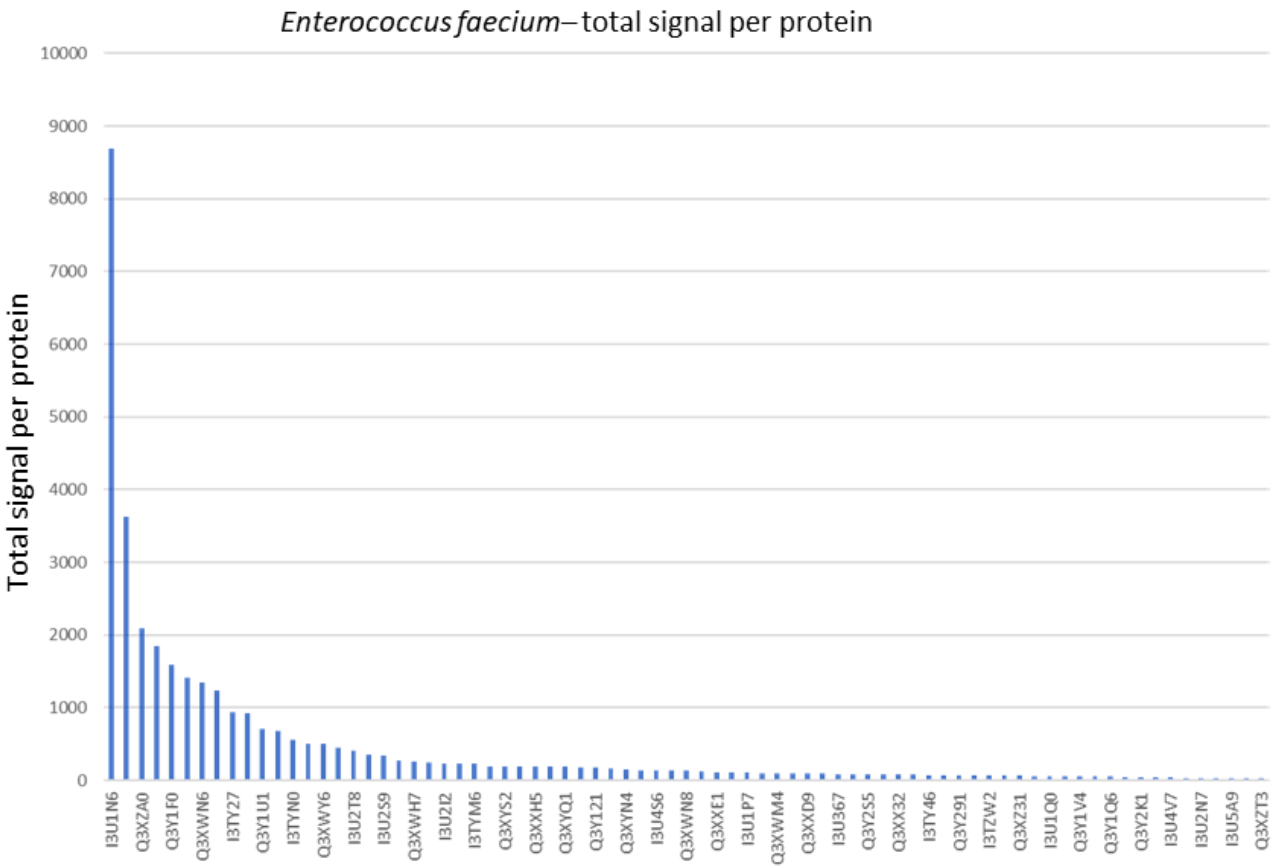

Sum Protein Signal per Taxa - Cutibacterium_acnes
Sum of Signal
20000
15000
10000
5000
0
W4U3R9
W4U8Q2
W4UA52
Protein names

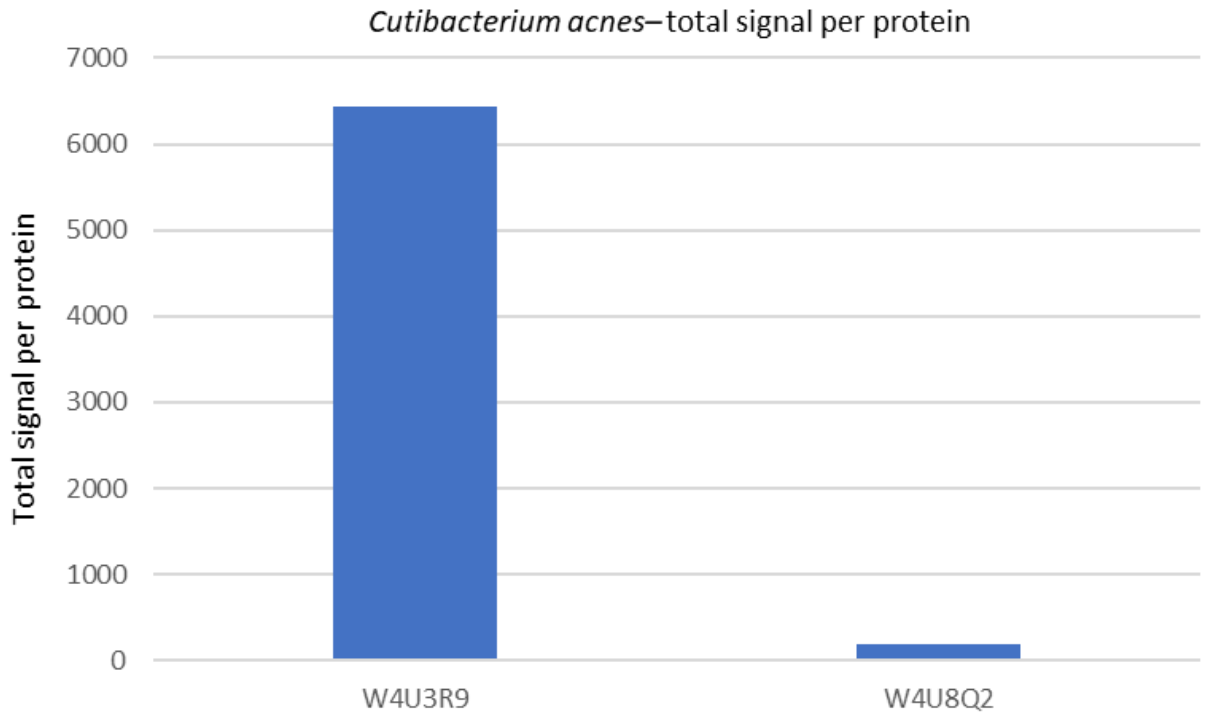


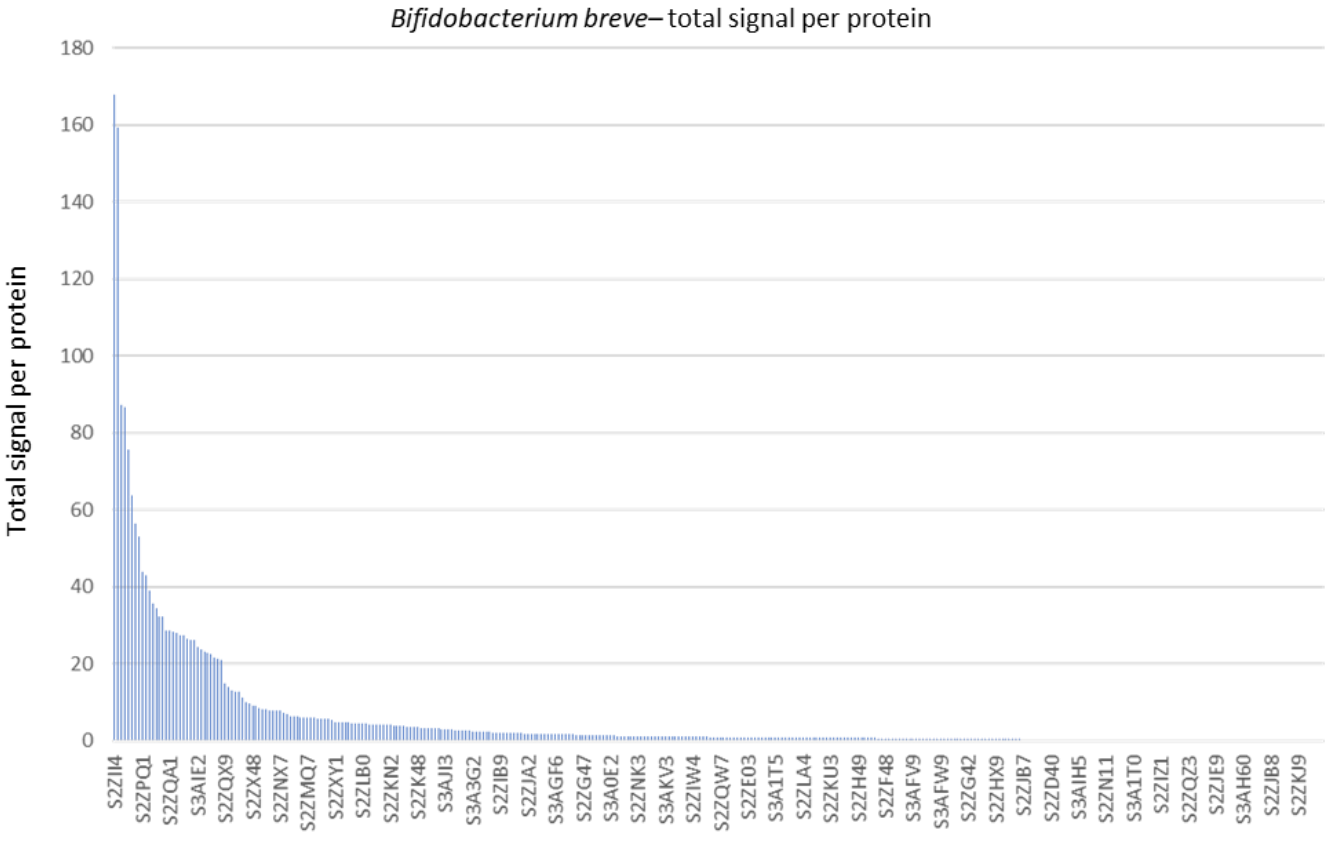
Taxa proteins signals after removed misidentified proteins



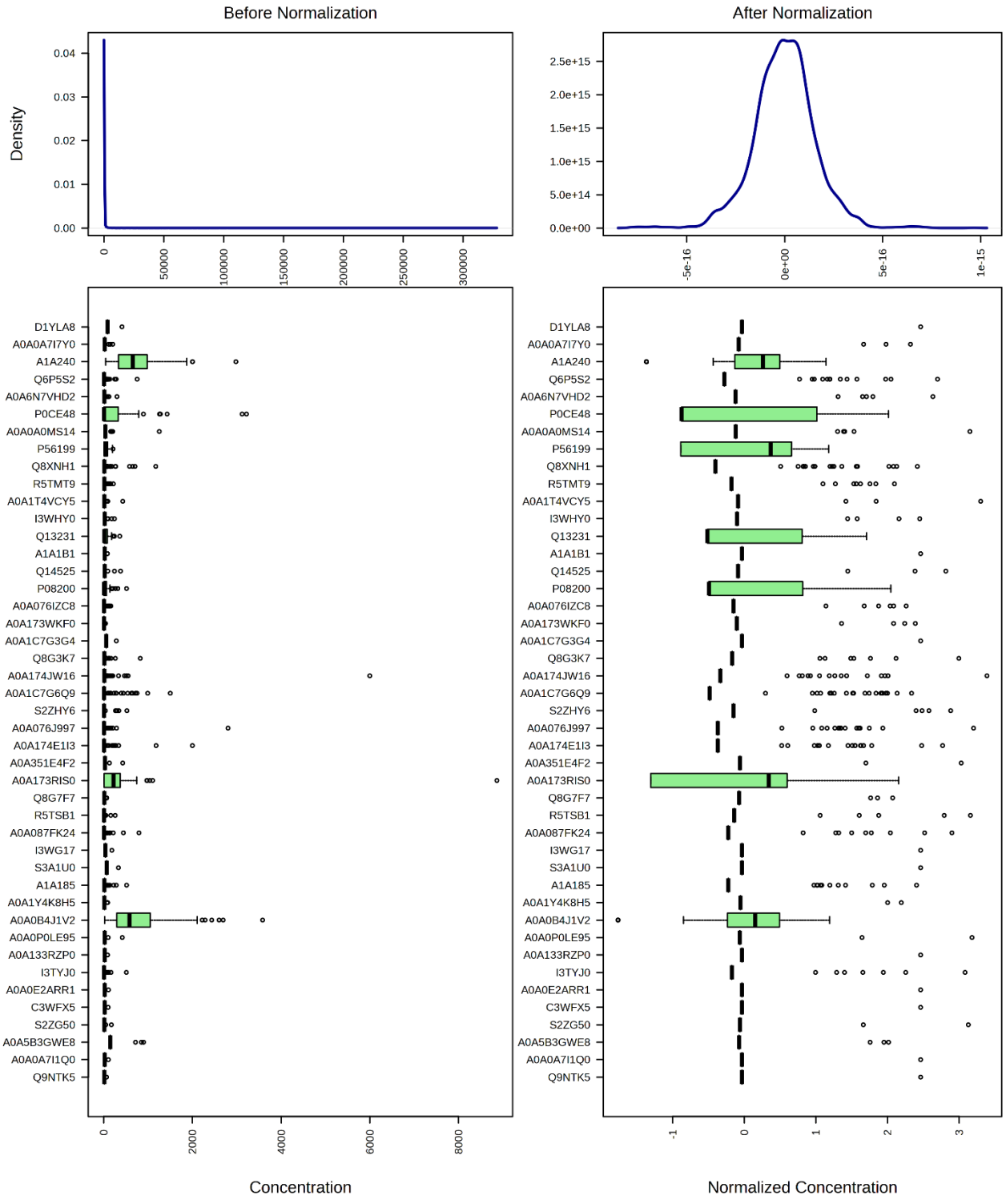

Staphylococcus epidermidis – total signal per protein
160
140
120
100
80
60
40
20
0
RPOA
EFTU
Q5HKI8
QUEF
RL7
RL6
ODPB
ARCA
GYRA
Q5HL57
RL17
KAD
ODPA
METK
UPP
RS4
ALF1
Q5HP46
Q5HNH7
LDH
SP5G
ILVE
Q5HND3
DEOB
Y1413
RL24
MQO4
THIO
G6PI
LDHD


Enterococcus faecium – total signal per protein
Total signal per protein
10000
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
I3U1N6
Q3XZA0
Q3Y1F0
Q3XWN6
I3TY27
Q3Y1U1
I3TYN0
Q3XWY6
I3U2T8
I3U2S9
Q3XWH7
I3U2I2
I3TYM6
Q3XYS2
Q3XXH5
Q3XYQ1
Q3Y121
Q3XYN4
I3U456
Q3XWN8
Q3XXE1
I3U1P7
Q3XWM4
Q3XXD9
I3U367
Q3Y255
Q3XX32
I3TY46
Q3Y291
I3TZW2
Q3XZ31
I3U1Q0
Q3Y1V4
Q3Y1Q6
Q3Y2K1
I3U4V7
I3U2N7
I3U5A9
Q3XZT3

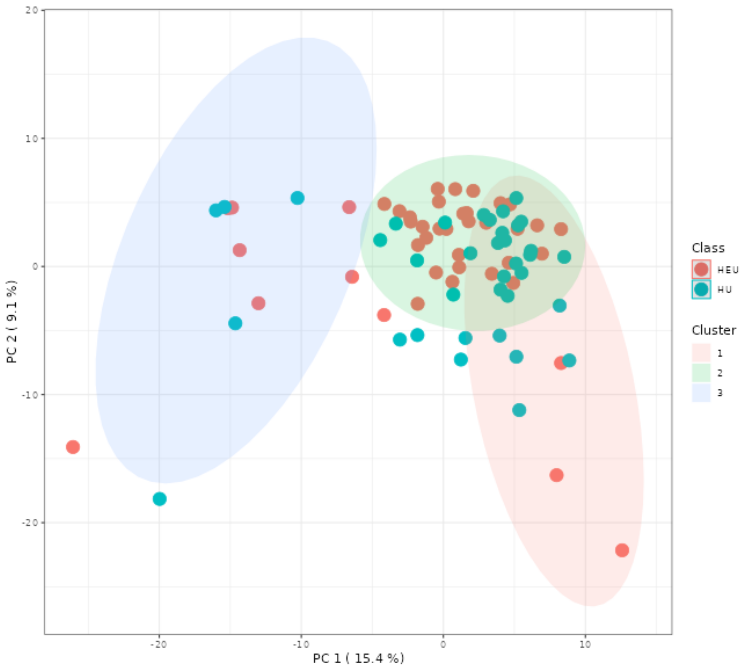

Cutibacterium acnes– total signal per protein
Total signal per protein
7000
6000
5000
4000
3000
2000
1000
0
W4U3R9
W4U8Q2


Bifidobacterium breve– total signal per protein
Total signal per protein
180
160
140
120
100
80
60
40
20
0
S2ZII4
S2ZPQ1
S2ZQA1
S3AIE2
S2ZQX9
S2ZX48
S2ZNX7
S2ZMQ7
S2ZXY1
S2ZLB0
S2ZKN2
S2ZK48
S3AJI3
S3A3G2
S2ZIB9
S2ZJA2
S3AGF6
S2ZG47
S3A0E2
S2ZNK3
S3AKV3
S2ZIW4
S2ZQW7
S2ZE03
S3A1T5
S2ZLA4
S2ZKU3
S2ZH49
S2ZF48
S3AFV9
S3AFW9
S2ZG42
S2ZHX9
S2ZJB7
S2ZD40
S3AIH5
S2ZN11
S3A1T0
S2ZIZ1
S2ZQZ3
S2ZJE9
S3AH60
S2ZJB8
S2ZKJ9

J- Metaboanalyst normalisation results after transforming and Pareto scaling

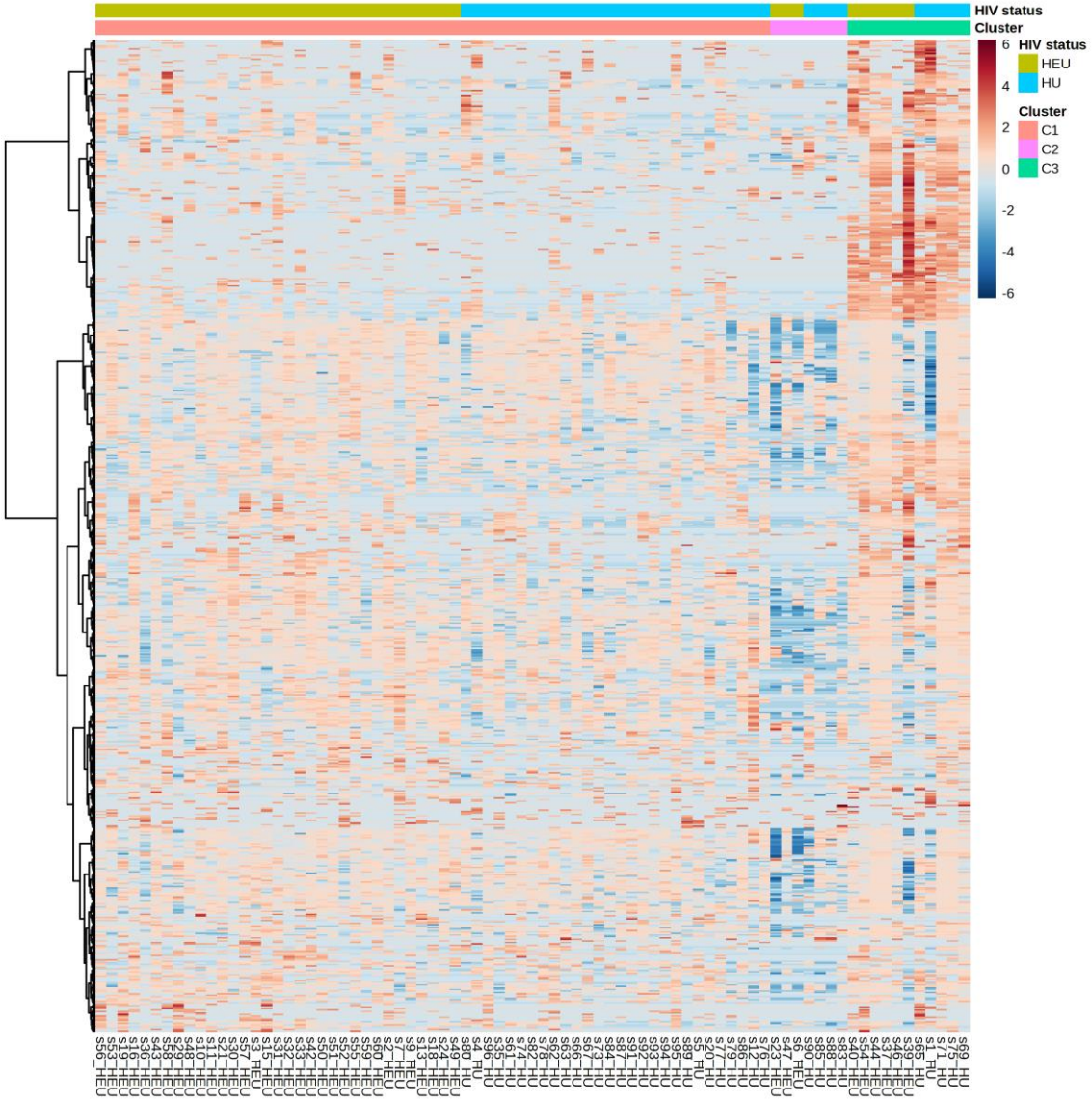


K- Unsupervised K-means clustering results

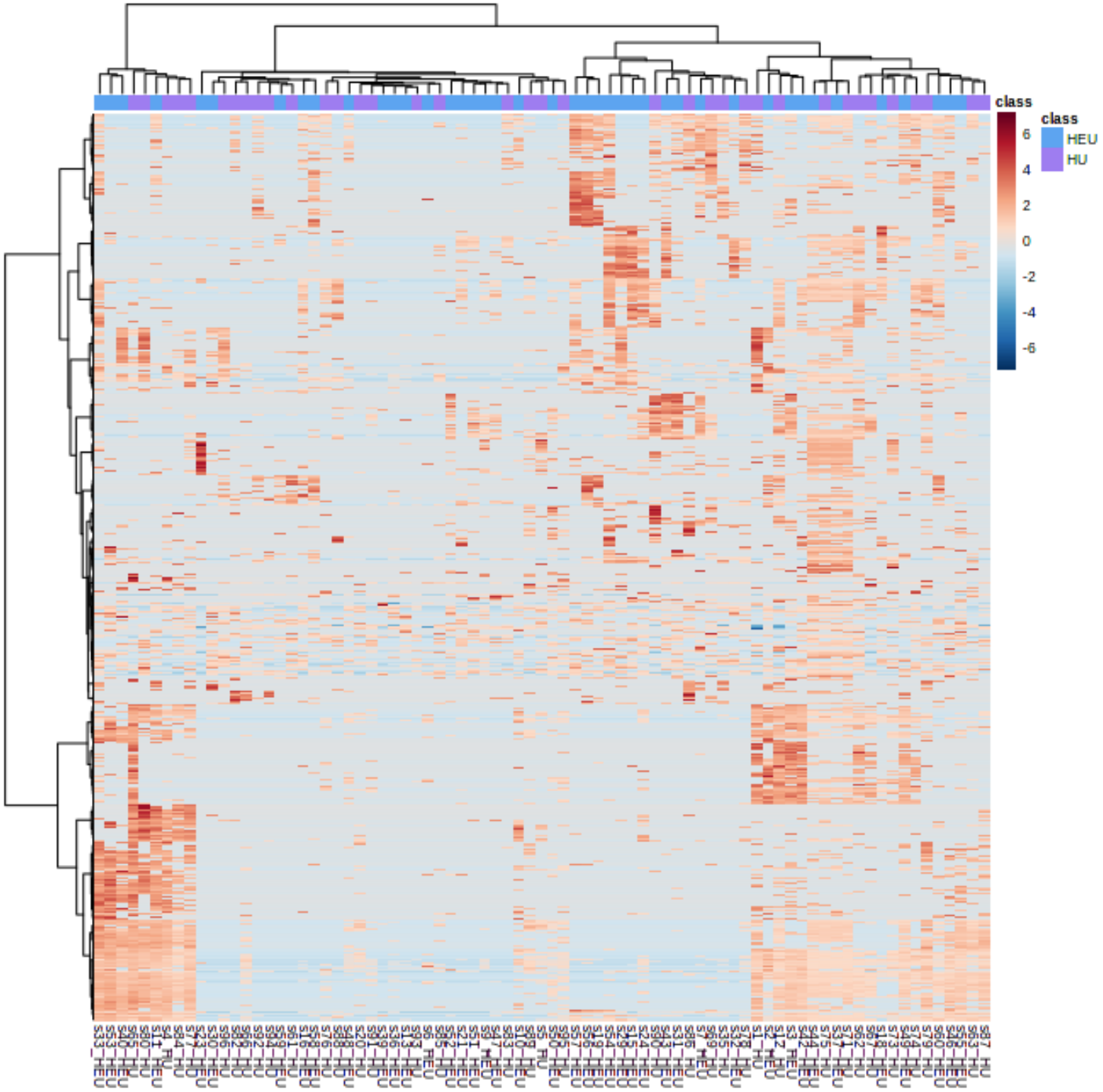


K-means clustering details:

Cluster	Members
Cluster 1	s56_HEU , s53_HEU , s19_HEU , s16_HEU , s36_HEU , s43_HEU , s58_HEU , s29_HEU , s48_HEU , s10_HEU , s11_HEU , s21_HEU , s30_HEU , s57_HEU , s3_HEU , s15_HEU , s31_HEU , s32_HEU , s33_HEU , s42_HEU , s50_HEU , s51_HEU , s52_HEU , s55_HEU , s59_HEU , s60_HEU , s2_HEU , s7_HEU , s9_HEU , s13_HEU , s18_HEU , s24_HEU , s49_HEU , s80_HU , s4_HU , s96_HU , s35_HU , s61_HU , s74_HU , s82_HU , s78_HU , s62_HU , s63_HU , s66_HU , s67_HU , s73_HU , s84_HU , s87_HU , s91_HU , s92_HU , s93_HU , s94_HU , s95_HU , s89_HU , s5_HU , s20_HU , s77_HU , s79_HU , s86_HU , s12_HU , s76_HU
Cluster 2	s23_HEU , s47_HEU , s6_HEU , s90_HU , s85_HU , s88_HU , s83_HU
Cluster 3	s40_HEU , s54_HEU , s44_HEU , s37_HEU , s46_HEU , s39_HEU , s65_HU , s1_HU , s71_HU , s75_HU , s69_HU



L- Heatmap of all bacterial proteins highlighting missingness in data

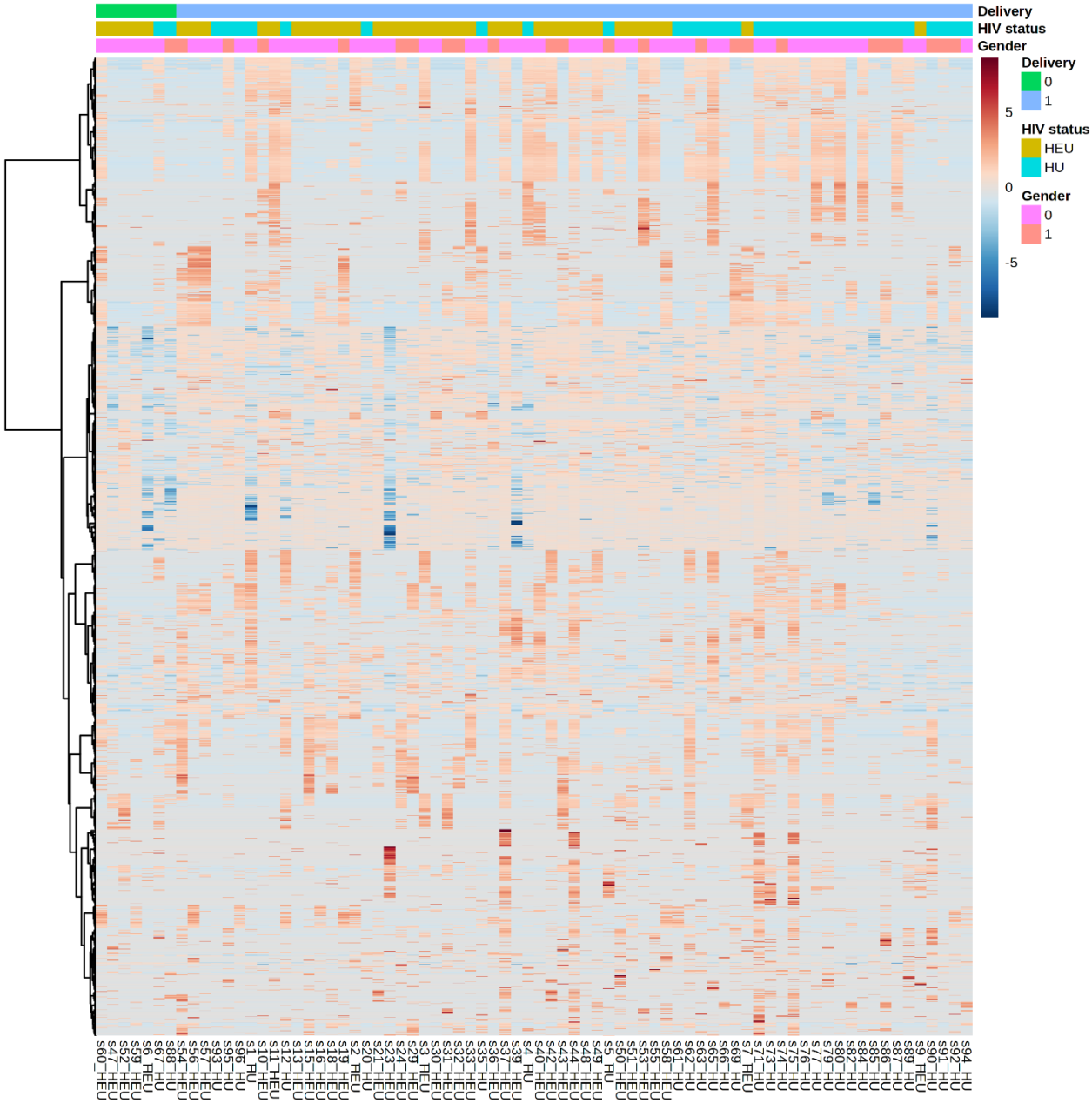


M- Top 10 % of human proteins from Random Forest model

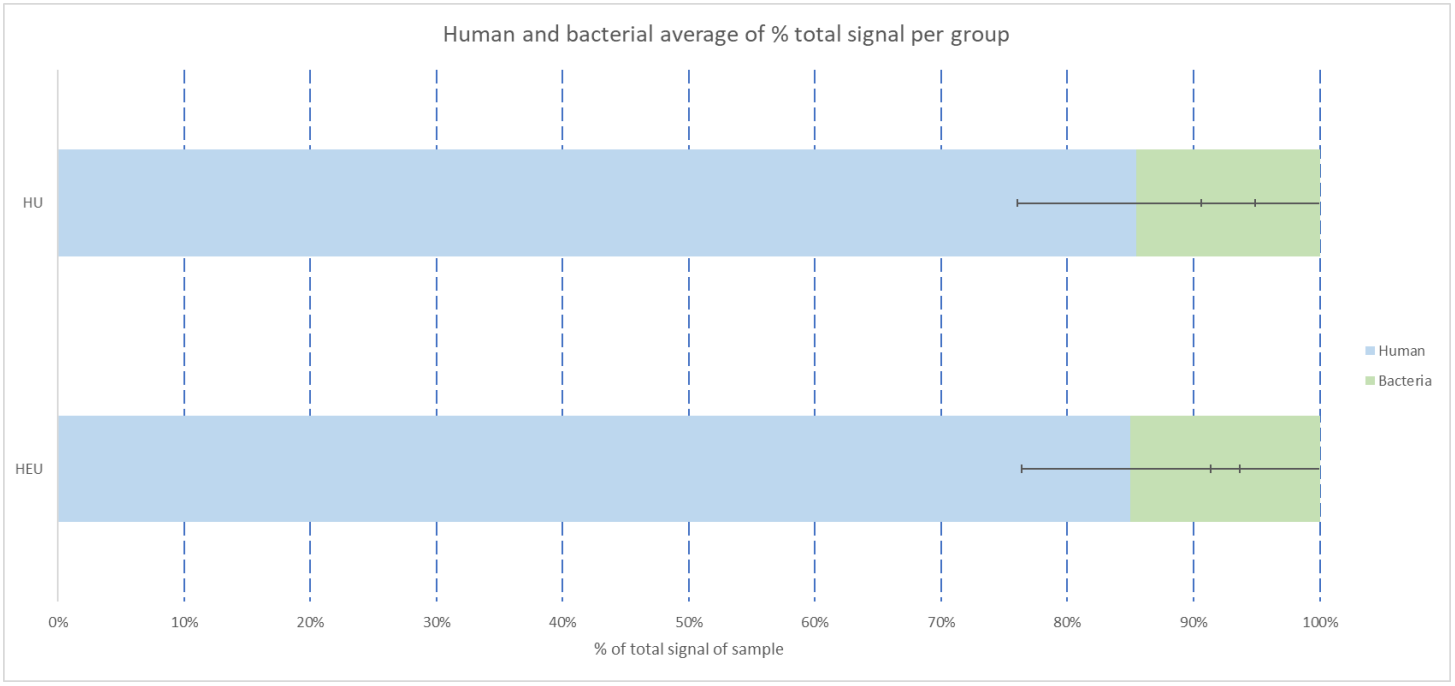
BREASTMILK RANDOM FOREST TOP 10%	
Protein	Mean Decrease Accuracy
Q6UWV6	0.004717
P20848	0.002434
Q13939	0.002287
P22528	0.001665
A0A075B7B6	0.001439
P0DP23	0.00141
Q08722	0.001314
A0A0A0MRZ8	0.001296
P35325	0.001206
P11117	0.001192
Q6UWP2	0.00111
Q02817	0.001032
Q6P1J6	0.001015
A0A0A0MTA3	0.001002
P16444	0.000854
A0A075B6P5	0.000832
Q9NR71	0.000817
A8K7I4	0.000816
Q8WW52	0.000808
O95497	0.000772
P06280	0.000758
P01705	0.000739
P03973	0.000722
P27487	0.000706
A0A075B6R9	0.000697
P04179	0.000664
Q16820	0.000622
P08133	0.000607
P56470	0.000592
Q9BXR6	0.000584
Q9UHG3	0.000537
Q6GPI1	0.000519
Q9HBB8	0.000508
O43280	0.000489
Q9NSB2	0.000462
P54108	0.000429
Q86UP6	0.000425
P06731	0.000411
Q8IUX8	0.000393
P09622	0.000384
P54819	0.000384
O00584	0.000377
Q13867	0.000376
P12532	0.000375
P48052	0.00037
P40879	0.00037
Q9UGM3	0.00036
P04118	0.00036

INFANT RANDOM FOREST TOP 10%	
Protein	Mean Decrease Accuracy
Q6UWV6	0.015612
P20848	0.011255
Q13939	0.008541
P11117	0.006631
P56470	0.005098
Q6P1J6	0.004224
P22528	0.00405
P0DP23	0.003896
Q02817	0.003848
Q08722	0.003675
P54108	0.003581
A0A075B7B6	0.003521
Q9NR71	0.00323
O00584	0.00292
P27487	0.002696
P50995	0.002594
P16444	0.002373
P61604	0.002078
Q14002	0.001954
A0A075B6P5	0.001871
Q9UGM3	0.001842
Q16819	0.001774
A0A0A0MRZ8	0.00176
P54819	0.00167
P01705	0.001624
P17538	0.001499
Q07837	0.001444
A0A0A0MTA3	0.001277
Q8WW52	0.001204
Q9UHG3	0.001151
Q9HBB8	0.001135
P04179	0.001032
Q96DA0	0.000992
Q9NSB2	0.000857
P35325	0.000818
Q9BXR6	0.0008
Q96CG8	0.000764
P09093	0.000512
A0A0J9YXX1	0.000459
O43280	0.00043
Q13867	0.000352
A8K7I4	0.000316
P50443	0.000265
P07478	4.37E-05
A0A075B6R9	-4.49E-05
Q6GPI1	-0.00018
Q0VAF6	-0.00035
Q10588	-0.0005

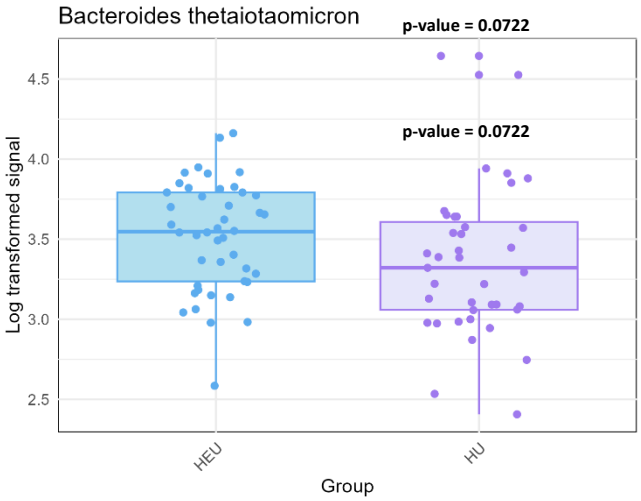
N- Metadata overlay with delivery and gender data

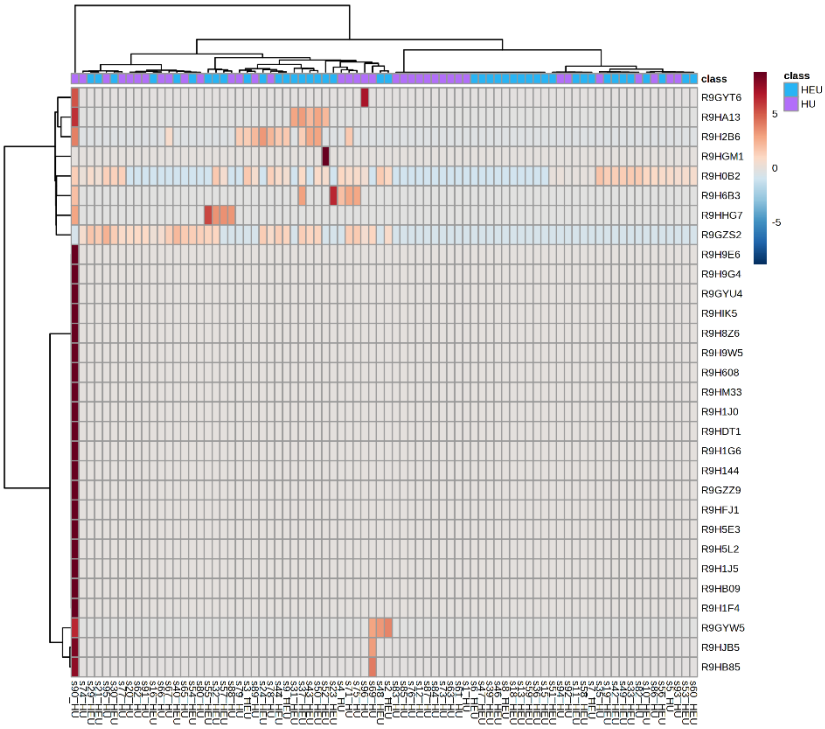


O- Average compositional comparison between HEU and HU groups in human and bacterial contributions to the total samples. Error bars showing variation between the groups.

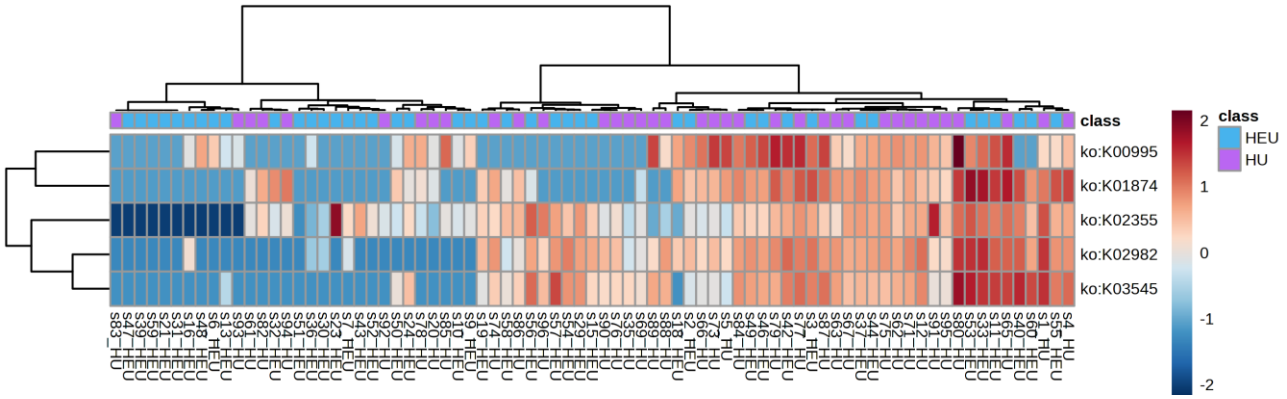


P- *Bacteroides thetaiotaomicron* boxplot with p-value between the species signal as 0.0722, and a corresponding heatmap showing the missingness in protein data.



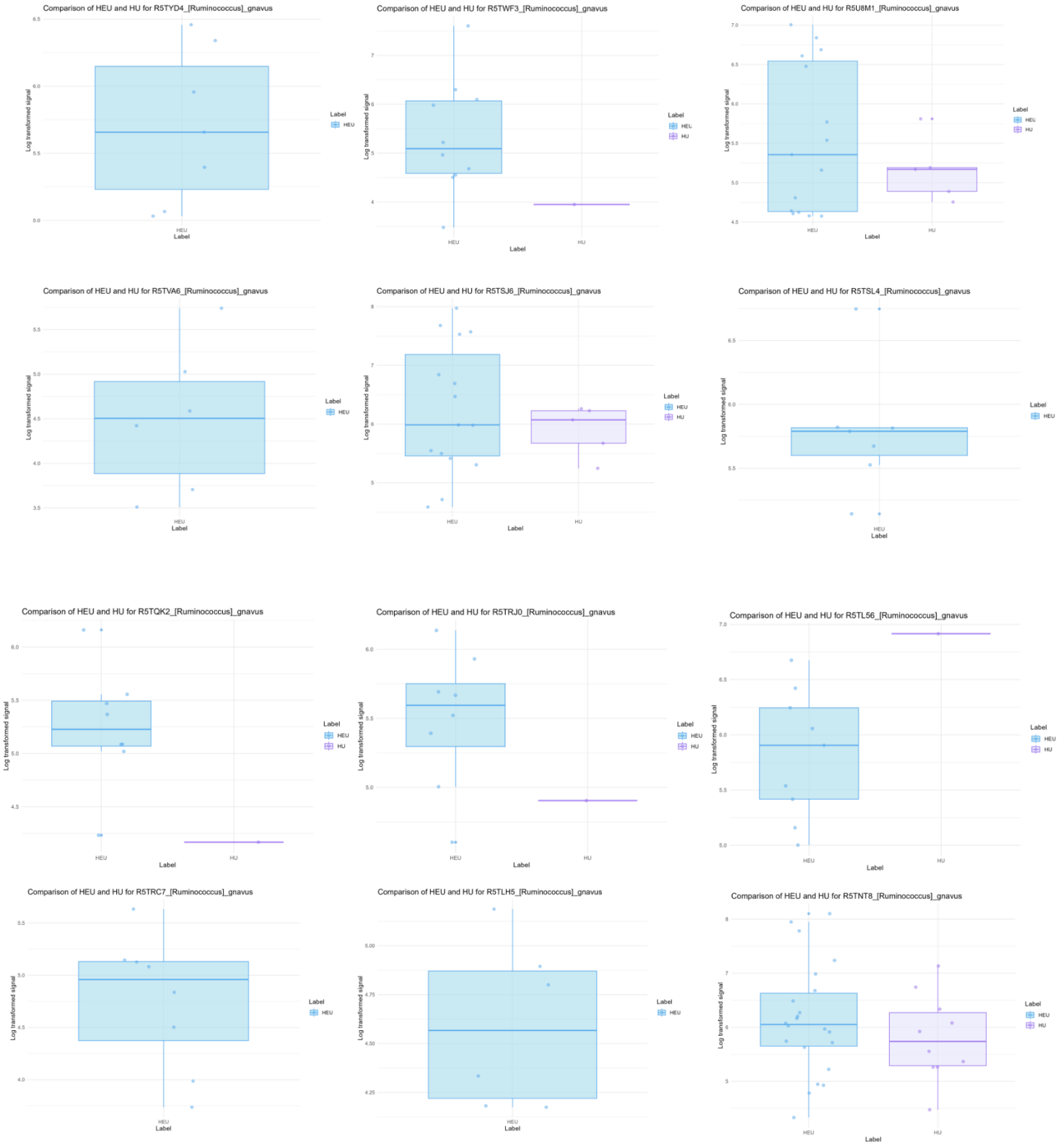


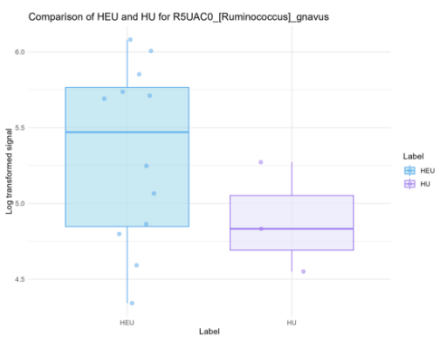
Q- Microbial protein ontology. Redundant proteins from different taxa accounted for by EggNOG mapper v5.

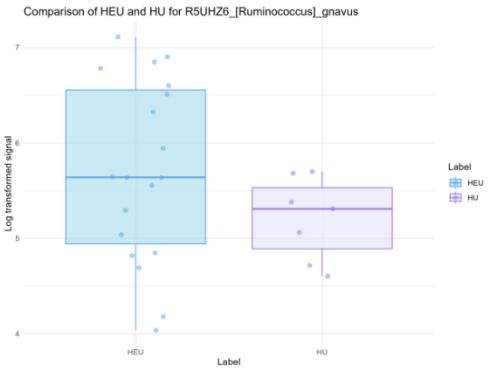


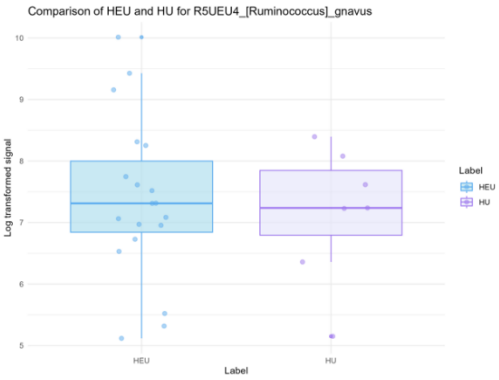
KO	Protein name	Pathway	Number proteins	Protein phylum
ko:K00995	g3p 3-phosphatidyltransferase	Glycerophospholipid metabolism	1	Firmicutes
ko:K01874	methionyl-tRNA synthetase	Selenocompound metabolism	6	3 Firmicutes, 2 Bacteroides, 1 Actinobacteria
ko:K02355	elongation factor G	Translation factors	33	16 Firmicutes, 8 Bacteroides, 5 Actinobacteria, 4 Proteobacteria
ko:K02982	small subunit ribosomal protein S3	Ribosome	12	6 Firmicutes, 4 Actinobacteria, 1 Proteobacteria, 1 Bacteroides
ko:K03545	trigger factor	no pathway	13	6 Actinobacteria, 3 Firmicutes, 2 Proteobacteria

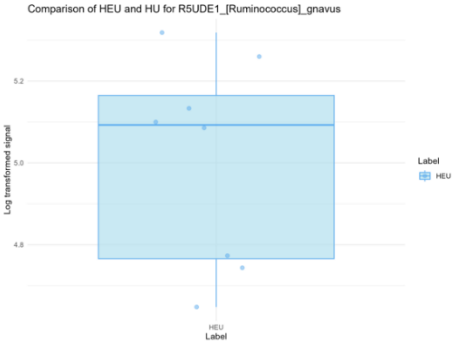
R- Significantly different *Ruminococcus gnavus* proteins' boxplots between HEU and HU

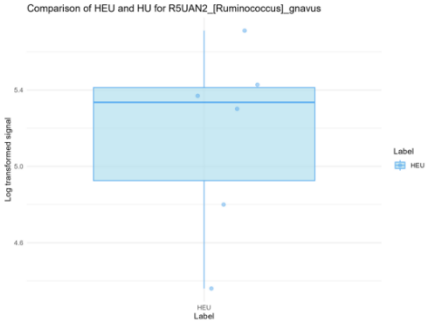



Comparison of HEU and HU for R5UAC0_[Ruminococcus]_gnavus
Log transformed signal
Label
HEU
HU


Comparison of HEU and HU for R5UHZ6_[Ruminococcus]_gnavus
Log transformed signal
Label
HEU
HU


Comparison of HEU and HU for R5UEU4_[Ruminococcus]_gnavus
Log transformed signal
Label
HEU
HU


Comparison of HEU and HU for R5UDE1_[Ruminococcus]_gnavus
Log transformed signal
Label
HEU


Comparison of HEU and HU for R5UAN2_[Ruminococcus]_gnavus
Log transformed signal
Label
HEU