The gut microbiota throughout paediatric haematopoietic stem cell transplantation

Gintare Vaitkute

April 2020

A thesis submitted to University College London for the degree of Doctor of Philosophy

Declaration

I confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Gintare Vaitkute

April, 2020

Abstract

Haematopoietic stem cell transplantation is a curative procedure for a variety of underlying diagnoses including haematological, immunological and metabolic conditions such as acute myeloid leukaemia, Wiskott-Aldrich syndrome and Hurler's syndrome, respectively.

The procedure, which includes conditioning and a period of neutropaenia prior to the transplant engraftment, can lead to significant morbidity including viraemia, bacteraemia and Graft-*versus*-host disease. Gut microbiota has been extensively studied in adult haematopoietic stem cell transplantation. It is known to affect stem cell reconstitution and has been frequently linked to various clinical outcomes; however, paediatric studies are scarce.

This work profiles the gut microbiota of paediatric patients undergoing haematopoietic stem cell transplantation at Great Ormond Street Hospital using both 16S rRNA sequencing and nuclear magnetic resonance spectroscopy. The initial chapter focuses on optimising the 16S rRNA methodology for this project. Subsequent chapters investigate the longitudinal gut microbiota and its dynamics throughout hospitalisation, as well as searches for clinical biomarkers at several time points throughout haematopoietic stem cell transplantation. We find that gut domination with a single taxon, specifically *Enterococcus*, *Enterobacteriaceae*, *Streptococcus* and *Staphylococcus* is common in this population. Additionally, we observe a loss of diversity around the time of the transplantation and that most patient microbiota profiles are unlike those of healthy individuals, even upon admission. We also identify three clusters within the data, revealing interesting cluster-switching patterns throughout transplantation and find that one cluster is linked to a higher risk of developing viraemia. We also find several biomarkers of viraemia and Graft-*versus*-host disease at baseline and around the time of engraftment, which may be indicative of overall gut health at that point. The final chapter profiles the faecal metabolome throughout haematopoietic stem cell transplantation and aims to broadly link the metabolome to the 16S data. We observe a loss of short chain fatty acids such as butyrate and acetate and increases in lactate and glucose throughout, which may be indicative of damaged gut epithelium and a loss of beneficial obligate anaerobes. Overall, these findings are indicative of a disruption of the host-microbe cross-talk.

Impact statement

The last few decades have seen significant increases in our understanding of the microbiome and its role in human health and disease. Despite this, longitudinal paediatric studies have been infrequent. This thesis aims to profile the gut microbiota and its functionality by way of metabolites throughout paediatric haematopoietic stem cell transplantation and the findings are likely of both clinical and scientific importance. The initial chapter provides an important example of the difficulties in optimising a complex microbiota sequencing workflow. As new technologies and reagents emerge, optimisation remains necessary.

The second chapter reveals intricate patterns in the dynamics of the paediatric gut microbiota throughout transplantation, the samples separating into three community types. The links between one of the community types and viraemia and the findings that metabolite trajectories may be different between them may allow us to better predict risk in this patient cohort.

Once validated, the biomarkers at baseline could be used to stratify patients according to their risk of viraemia and Graft-*versus*-host disease, which may allow for early clinical measures such as isolation or more timely clinical responses upon suspicion of an adverse event. Observed metabolite patterns reinforce findings from earlier chapters and provide a greater insight into the potential underlying mechanisms. Overall, the findings also provide ideas as to specific interventions that may or may not be useful to this specific cohort of patients.

To our knowledge this work represents the largest longitudinal paediatric cohort of this kind and is a foundation for further mechanistic work into the role of microbiota in haematopoietic stem cell transplantation.

Acknowledgements

First and foremost, I would like to thank my supervisors Dr Mona Bajaj-Elliott, Professor Nigel Klein and Dr Elaine Cloutman-Green for the opportunity to undertake work under their supervision and for their advice and encouragement throughout the project.

This work benefited from the expertise of many individuals. A heartfelt thanks to Professor Paul Veys, Dr Reem Elfeky, Dr Giovanna Lucchini and Dr Elizabeth Rivers, who put time aside from their busy clinical schedules to explain the complex world of stem cell transplantation and who always answered my queries, however simple they may be. I would also like to acknowledge Dr Lisa Dawson, who shared her phenol-based DNA extraction method and Dr Kathryn Harris, who provided us with the mock community that was evaluated in the optimisation chapter. Dr Gordana Panic and Dr Jonathan Swann have been kind enough to assist the project and to introduce me to the world of NMR, in particular I thank Dr Gordana Panic for helping me with the more challenging aspects of NMR data pre-processing, such as manual spectra alignment and peak extraction and helping to write the NMR section of the methods. She has given up many hours of her time to explain the more intricate details of NMR data processing and analysis to me, as well as supervising me in the lab to ensure we get the best quality data.

I would also like to thank everyone who helped me with both lab work and bioinformatics over the course of the project: Dr Joe Standing, Dr Silke Gastine, Dr Patricia Hunter, Dr Grace Logan, Dr Athina Gkazi, Dr Ronan Doyle, Dr Sarah Watters, Dr Ben Margetts, Dr Liam Shaw, Dr Stuart Adams and John Booth. I am especially thankful to Dr Dagmar Alber, who has taken me under her wing, supporting me with both lab work and bioinformatics. More importantly, thank you for the encouragement whenever I had doubts and for the stimulating scientific discussions, which always encouraged me to think about both problems and findings from different points of view.

Additionally, I would like to thank the Great Ormond Street Children's Charity and the BRC for supporting this research. It is a privilege to work in such a fascinating field. I owe a debt of gratitude to the patients and their guardians who agreed to participate in this study- without them this work would not have been possible.

Finally, I would like to thank my family, who have been a great support throughout this PhD and to Chad, who always made a great cup of chai when times were tough.

Table of Contents

List of Figures

List of Tables

Table A10 Details of contaminants found within each sequencing plate as well as taxa removed from each dataset

Table A11 Univariate and multivariable Cox models with Graft-*versus*-host disease as the dependent variable

Table A12 Univariate Cox models with viraemia as the dependent variable

Table A13 Optimal cut-offs for significant taxa for the baseline sub-cohort

Table A14 Differentially abundant taxa for Graft-*versus*-host disease and viraemia in pre-engraftment samples

Table A15 Optimal cut-offs for significant taxa for pre-engraftment sub-cohort

Table A16 Models run in this cohort

Acronyms

3-IS	3-indoxyl sulfate
5-ASA	5-aminosalicylic acid
AA	amino acids
ADV	adenovirus
AhR	aryl hydrocarbon receptor
aGvHD	acute graft *versus* host disease
ALL	acute lymphoblastic leukaemia
Allo-FMT	allogeneic faecal microbiota transplantation
Allo-HSCT	allogeneic haematopoietic stem cell transplantation
AML	acute myeloid leukaemia
APCs	antigen presenting cells
AUC	area under the curve
Auto-FMT	autologous faecal microbiota transplantation
Auto-HSCT	autologous haematopoietic stem cell transplantations
BM	bone marrow
BP	base pairs
BSI	bloodstream infections
CB	cord blood
CDI	*Clostridium difficile* infection
CI	confidence interval
CMV	cytomegalovirus
CRP	C-reactive protein
CST(s)	community state type(s)
DAMPs	danger-associated molecular patterns
DNA	deoxyribonucleic acid

EBV	Epstein-Barr virus
EN	enteral nutrition
FMT	faecal microbiota transplantation
GF	germ free
GI	gastrointestinal
GOSH	Great Ormond Street Hospital
GvHD	graft *versus* host disease
GvL	graft *versus* leukaemia
HC	healthy control
HLA	human leukocyte antigen
HMP	Human Microbiome Project
HR	hazard ratio
HSCT	haematopoietic stem cell transplantation
ICI	immune checkpoint inhibitors
IEC	intestinal epithelial cells
LDA	linear discriminant analysis
LPS	lipopolysaccharide
MAMPs	microbial associated molecular patterns
MC	mock community
NEC	negative extraction control
NMDS	non-metric multidimensional space
NK	natural killer cells
NT	nucleotide
OPLS	orthogonal partial least-square
OPLS-DA	orthogonal partial least-square discriminant analysis
OR	odds ratio
OTU	operational taxonomic unit

PAM	partitioning around medoids
PAMPs	pathogen-associated molecular patterns
PB	peripheral blood
PCA	principal component analysis
PCR	polymerase chain reaction
PPAR-γ	peroxisome proliferator activated receptor γ
PPM	parts per million
ROC	receiver operator curve
rRNA	ribosomal ribonucleic acid
SCID	severe combined immunodeficiency
SCFA(s)	short chain fatty acid(s)
sIgA	secreted immunoglobulin A
TCA	tricarboxylic acid cycle
TMS	tetramethylsilane
TPN	total parenteral nutrition
TRM	transplant-related mortality
t-SNE	T-distributed stochastic neighbour embedding

Chapter 1- Introduction

1.1 Haematopoietic stem cell transplantation

Haematopoietic stem cell transplantation (HSCT) is a potentially curative procedure for a variety of underlying haematological, immunological and metabolic conditions. HSCT was first performed in the 1950s and since then the development of more accurate tissue typing methods and less cytotoxic conditioning regimes has allowed for this procedure to be used in a wider patient population. Despite this, HSCT can result in transplant-related complications and carry an increased risk of infections and Graft *versus* Host Disease (GvHD).

1.1.1 HSCT procedure

Indications for HSCT in the paediatric population are varied. It presents as the only treatment option for a variety of conditions including primary haemophagocytic lymphohistiocytosis, chronic myeloid leukaemia and Fanconi anaemia. Severe combined immune deficiency (SCID) is one of the immunodeficiencies that requires an HSCT immediately post-diagnosis. It is also a potential treatment for more common diagnoses including acute myeloid leukaemia (AML) and acute lymphoblastic leukaemia (ALL). The European Society for Blood and Marrow Transplantation report that in 2012 the main indications for paediatric allogeneic HSCT were ALL (26%) and primary immune deficiencies (16%). For autologous HSCT, on the other hand, it was solid tumours (66%) and lymphomas (15%)[1].

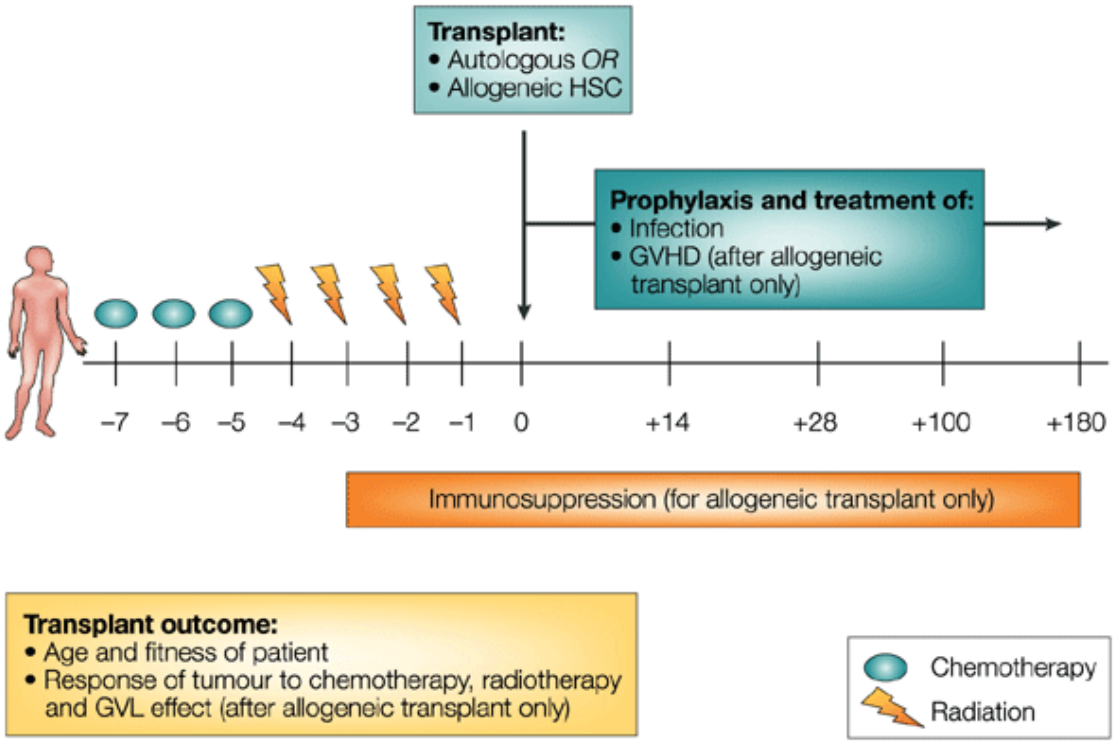


Figure 1.1 A schematic of the HSCT procedure. Patients receiving stem-cell transplants for haematological malignancies are first treated by a regimen of chemotherapy and/or chemoradiation (conditioning). The patients then receive their own cells (autologous transplant) or related/unrelated donor cells (allogeneic transplant). Patients also receive immunosuppressive drugs to prevent rejection of the transplanted cells. Figure adapted from Barker *et al*[2].

HSCT can be thought of as a rescue therapy and the procedure is illustrated in Figure 1.1 above[2]. That is, it is always preceded by a preparative regime, which is tailored to each individual depending on, among other factors, the underlying disease and the type of graft. It aims to eliminate the recipient's bone marrow cells (the origin of haematopoietic cells) in order to provide space for the donor graft, to suppress the immune system so that the graft is not rejected and in the case of haematological conditions, to eliminate any underlying disease. A donor stem cell graft is then given, in effect replacing the host's depleted immune system. Additionally, both allogeneic transplantation- receiving stem cells from an unrelated or related sibling donor; and autologous transplantation (receiving one's own cells after they have been genetically modified in the laboratory) can be undertaken. The main contributing factors of transplantation outcomes include the susceptibility of the cancer to the chemoradiation therapy, the Graft-*versus*

leukaemia (GvL) effect by the transplanted graft and the overall health of the patient.

HLA- matching

Once the decision to perform an HSCT has been taken, a suitable donor is identified, which is followed by matching or tissue typing between the donor and the recipient to ensure that the recipient does not reject the graft. The donors can either be a matched related/unrelated donor or a mismatched unrelated donor, although haploidentical (i.e. half-matched) transplantation has also become common. Matching is done using the human leukocyte antigen (HLA) genes at several loci. Allorecognition of HLA allelic differences by T lymphocytes increases the risk of acute Graft-*versus*-host disease (aGvHD) and mortality[3]. At most transplant centres the gold standard 'match' is a 10/10 allele match across the HLA-A, -B, -C, -DRB1, and -DQB1 loci. The addition of the HLA-DPB1 typing means a 12/12 match can also be sought[3]. Any disparities increase the risk of transplant-related morbidity and mortality and transplant rejection and disparities in certain regions, such as HLA-DQB1, seem to be better tolerated than others[3]. An ideal match for transplantation would be an HLA-identical sibling, however this is not feasible for most individuals. The graft recipients in this cohort were typed for either 10 or 12 alleles. 57% of patients in this cohort were matched to 9/10;10/10 or 11/12;12/12 alleles.

Advances in preventing transplant-related morbidity and mortality mean that grafts from mismatched unrelated or haploidentical individuals is more feasible and haploidentical transplants are on the rise[4]. Veys *et al* found that graft manipulation using TCRαβ/CD19 depletion is associated with a lower incidence of GvHD in mismatched HSCT in paediatric patients with primary immunodeficiencies[5].

Stem cell source

Bone marrow (BM) was predominately used for paediatric stem cell transplantation prior to 1977. Since then, transplantation using hematopoietic stem cells from peripheral blood (PB) has increased significantly as they were used in 30% of all allogeneic HSCT (allo-HSCT) and 85% of autologous HSCT (auto-HSCT) between 1999 and 2002[6]. More recently, the British Society for

Blood and Marrow Transplantation and Cellular Therapy recorded a total of 427 first time paediatric stem cell transplants in 2018, 75% of these being allogeneic. Of those, 61% used bone marrow as the cell source, 30% were peripheral blood and 9% were from cord blood (CB)[7].

More recently, stem cells from CB have also been used. In the 1980s Broxmeyer *et al* demonstrated that CB contains a significant number of progenitor cells and suggested that it could be used for HSCT[8]. CB has several advantages to other stem cell types including lower risk of GvHD and viral infections in comparison to BM, which is primarily due to the relative naivety of the donor CB cells[9]. CB can be an ideal match for paediatric AML patients when a matched sibling donor cannot be identified. Keating *et al* found that paediatric AML patients receiving a CB had better leukaemia-free survival and less GvHD compared to matched unrelated graft recipients[10].

Peripheral blood is also frequently used as a haematopoietic stem cell source in paediatric HSCT. The major advantage of PB is the shorter engraftment times for platelets and neutrophils. Despite this, PB HSCT is associated with higher rates of GvHD and lower rates of overall survival compared to BM as well as potential risks for the donor including difficulties in venous access and side effects of the mobilisation drugs given (to stimulate the production of stem cells)[11].

Conditioning

Once a transplant match is found, the preparative conditioning regime and prophylaxis are decided upon. The regimes can be broadly split into myeloablative, reduced intensity and minimal intensity and are chosen based on a variety of factors including the underlying condition, the patient's age and the source of the stem cells. Myeloablative and reduced intensity regimens often contain both a chemotherapy and an immunosuppressive agent (with or without radiotherapy), whereas minimal intensity regimens can contain only the immunosuppressive agent.

The early conditioning for allo-HSCT was based on total body irradiation, which was followed by the use of cyclophosphamide and busulfan to increase both the anti-tumour effect and immunosuppression. Despite this, the toxicity and transplant-related mortality (TRM) limited the use of these regimens to fitter and

younger patients with malignant indications. This was followed by findings that reduced intensity conditioning could provide sustained engraftment and eradicate hematologic malignancy and patient's haematopoiesis, which allowed for the expansion of HSCT to older and less fit patients, as well as its broader application to non-malignant indications. Reduced intensity conditioning is based on immunosuppression to prevent graft rejection, in combination with a reduced dose of anti-tumour chemotherapy in order to reduce the toxicity and TRM[12].

Reduced-intensity conditioning is known to have several advantages compared to myeloablative conditioning, including reduced-frequency of mucositis, faster platelet engraftment, the need for fewer transfusions and less total parenteral nutrition (TPN) as well as fewer cytomegalovirus (CMV) infections[13]. As well as the type, the combinations of conditioning regimens are known to affect post-transplantation outcomes. Among paediatric patients receiving HSCT for AML, the use of busulfan/cyclophosphamide/melphalan conditioning is superior to total body irradiation/cyclophosphamide and busulfan/cyclophosphamide in reducing relapse and improving leukaemia-free survival[14]. Therefore overall, the most appropriate conditioning regimen is chosen with a trade-off between toxicity, whilst maintaining necessary immunosuppression and anti-tumour effect.

Although there is a broad consensus between tertiary transplant centres, differences in prophylaxis, conditioning regimes and transplant manipulation approaches exist. Between 2008 and 2014 for example, a Centre for International Blood and Marrow Transplant Research report found that the majority of allogeneic paediatric HSCT recipients in the United States of America (USA) had myeloablative conditioning (97%) in contrast to ~60% in our cohort[15]. Similarly, *in vivo* T cell depletion and reduced-intensity conditioning, commonly used at Great Ormond Street Hospital (GOSH), is used less frequently in paediatric allo-HSCT recipients in the USA, however the proportion of graft sources (BM/PB/CB) was similar to our cohort. Overall, these differences can make comparisons between studies with heterogeneous patient cohorts more difficult[15].

Immune reconstitution

Once the graft is transplanted, immune reconstitution begins. As a result of conditioning, most patients experience a period of neutropaenia, which persists

for around 2 weeks (Figure 1.2). Neutrophils are the first to reconstitute followed by natural killer cells (NK), although it is thought that they remain functionally underdeveloped several months post-HSCT[16]. Adaptive immune cell subsets- B and T cells, recover much more slowly post-HSCT. CD8 T cells recover within around 100 days in a process driven by exposure to alloantigens or viral infections, whereas CD4 T cells display a more prolonged recovery, which results in an inverted CD8:CD4 ratio that may persist for many years post-HSCT. The immune system of patients receiving an auto-HSCT reconstitute more rapidly than those receiving an allo-HSCT[17]. Neutrophils and platelets for example reconstituted in a median of 10 and 11 days respectively in autologous and in 19 and 21 days in allogeneic patients[18].

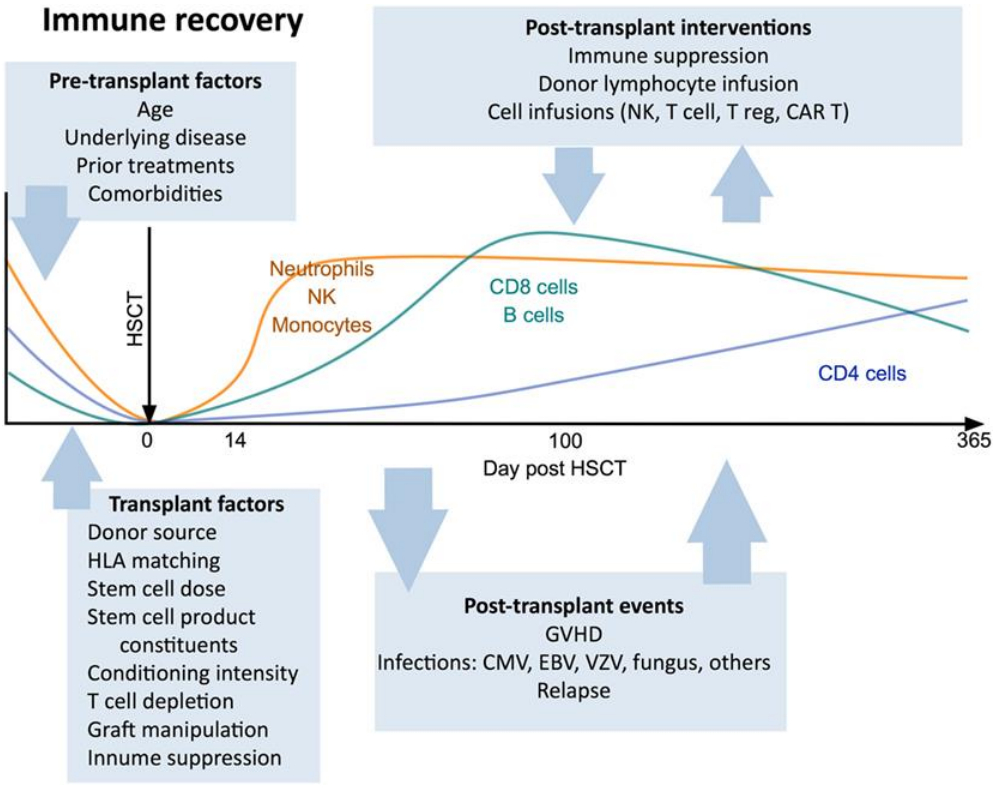


Figure 1.2 An overview of immune reconstitution following allo-HSCT. In the first year post allo-HSCT the major cell subsets follow a similar pattern of recovery under the influence of numerous factors including pre-transplant patient characteristics such as age, underlying disease and comorbidities; Transplant factors include the source of the graft; Post-transplant events such as infections and GvHD and post-transplant interventions including immune suppression. CAR T- chimeric antigen receptor T cells, GvHD- Graft *versus* Host Disease, CMV- cytomegalovirus, VZV-varicella zoster virus, EBV- Epstein-Barr virus. Figure adapted from Stern *et al*[16].

Delayed or inadequate immune reconstitution influences morbidity and mortality post-HSCT and multiple factors such as patient factors, transplant factors and post-transplant interventions can impact immune reconstitution[19]. Individuals receiving grafts from cord blood, which contain fewer CD34+ hematopoietic stem cells than BM or mobilised PB grafts exhibit slower engraftment kinetics and a delayed T cell reconstitution. Veys *et al* have demonstrated that omission of *in vivo* T cell depletion in paediatric CB HSCT recipients however, results in rapid thymic-independent T cell expansion[8]. The naïve T lymphocytes, particularly CD4+ T cells, can then differentiate into viral-specific T cells, which aid in clearing infections. Mobilised PB grafts generally contain the highest numbers of mature lymphocytes and committed progenitors and therefore lead to faster rates of immune reconstitution, yet greater rates of GvHD.

1.1.2 Prognosis and outcomes post-transplantation

Advances in reduced-toxicity conditioning and prophylactic regimes, among other factors, have reduced transplant-related morbidity and mortality. Nevertheless, HSCT remains associated with significant risks including bacterial and viral infections, GvHD and relapse. Mortality rates in adult patients at 100 days post-transplantation vary from 7% in those receiving a matched sibling graft to 27% in those with refractory acute leukaemia receiving a matched unrelated graft[12]. TRM varies in paediatric cohorts and rates were found to be on average 13% three years post-HSCT in paediatric patients undergoing allo-HSCT for underlying malignancies. The most frequent causes were organ failure, infections, GvHD and post-transplant lymphoproliferative disorders[20]. In 2018, GOSH reported an overall mortality rate of 21% a year post-HSCT and the overall HSCT population mortality rates can remain 5 to 9-fold higher than those expected in the general population at least 30 years after transplantation[21].

Infectious complications

The conditioning regimen renders the patient aplastic until haematopoiesis commences, leaving the individual particularly susceptible to infections. Infections constitute a major cause of TRM and the estimated incidence is 20-44%, with a mortality rate of 10-50%, mostly due to antibiotic-resistant strains[22,23]. The incidence and severity are highly dependent upon the use of immunosuppression and immune reconstitution following transplantation. The

risk of infection is higher, for example, in patients with delayed immune reconstitution following a cord blood transplant[12]. S*treptococcus, Staphylococcus, Escherichia* and *Klebsiella* taxa constitute the most common bacteraemias[12]. For this reason, fluoroquinolones, specifically ciprofloxacin, are commonly used for prophylaxis during the neutropenic phase[24].

The most common causes of viraemia include adenovirus (ADV), CMV and Epstein-Barr virus (EBV) and reactivation of latent viruses is common during the neutropenic phase. It has been found that pre-existing ADV, CMV or EBV seropositivity and *in vivo* T cell depletion are linked to higher risk of viral reactivation in blood[25]. Similarly, transplantation from a seropositive donor to a seronegative recipient is generally not recommended, however with the use of viral prophylaxis mismatched viral status is acceptable. Reactivation and acquisition of these viruses can lead to significant morbidity and mortality in this population; therefore each patient is screened for viral as well as bacterial and fungal infections weekly and aciclovir prophylaxis is common. Additionally, immunotherapeutic approaches including donor lymphocyte infusions and virus-specific cytotoxic T cells can also be used in treatment.

Viral gastroenteritis, predominately caused by norovirus, ADV and rotavirus is also common post-HSCT, which results in diarrhoea, vomiting and nausea. The cumulative incidence of norovirus infection was found to be 12.9% in one paediatric cohort 2 years post-HSCT and 19% all-cause viral gastroenteritis in another paediatric cohort[26].

GvHD

Alongside infectious complications, GvHD is another common cause of TRM in this population. GvHD can typically affect the skin, gastrointestinal (GI) tract, liver and eyes and can present with rashes, abdominal cramps, diarrhoea and abnormal liver function tests. GvHD severity varies from clinically insignificant grades 0-I, to clinically significant grades II-IV and is broadly classed into acute GvHD (up to 100 days post-HSCT) and chronic GvHD (100+ days post-HSCT).

GvHD incidence varies at each transplantation centre, with rates of paediatric chronic GvHD as low as 6% with matched sibling cord transplants to as high as 65% with matched unrelated PB transplants[27,28]. Whilst skin GvHD is generally

treatable, GI GvHD presents with more significant rates of mortality. The probability of survival was found to be the worst (42.5-66.9%) with most severe grades of GvHD (III-IV) in paediatric leukaemia patients 3 years post-HSCT[29,30]. Skin is generally the most affected organ in both acute and chronic GvHD followed by the GI tract and liver. GOSH reports a grade III-IV GvHD incidence of 18% in 2018 based on Glucksberg criteria (Tables 2.1/2.2).

Figure 1.3 details the development and persistence of acute GvHD. Acute GvHD can be broadly divided into several stages, whereby the host tissues, such as the gut epithelium, are damaged and disrupted by conditioning, leading to the release of inflammatory cytokines, which in turn activate antigen presenting cells (APCs). The APCs then activate mature donor T cells present in the graft, which differentiate and activate CD4, CD8 and NK cells and can in turn mediate tissue damage to the skin, lung, liver and/or the GI tract. Additionally, lipopolysaccharide (LPS) that leaks through the damaged GI epithelium recruits myeloid cells, which enhance the cytokine storm[31]. In the gut, immune dysregulation observed during acute GvHD is thought to lead to further perturbations to the intestinal epithelium, Paneth cells, stem cells and goblet cells. GvHD prophylaxis, such as mycophenolate mofetil, an immunosuppressant, is commonly given to patients considered at higher risk for GvHD including those receiving a graft from a mismatched unrelated donor, which can include *in vitro* and/or *in vivo* T cell depletion.

Importantly, the gut microbiota and associated metabolites appear to have a role in GvHD development and prevention. Several studies have found that changes to the gut microbiota can lead to an increased risk of GvHD. Domination (>30%) with *Enterococcus* post-HSCT was found to increase the risk of GvHD in an adult cohort, whereas depletion of anti-inflammatory *Clostridia* preceded the development of GvHD in a paediatric cohort[32,33]. Similarly, changes in microbiota-derived metabolites, known to protect the intestinal barrier, such as short chain fatty acids (SCFA) and indole derivatives have been observed during HSCT[34]. As a result, administration of an indole derivative to a murine HSCT model alleviated the severity of GvHD, the damage to the intestinal epithelium and transepithelial bacterial translocation[35].

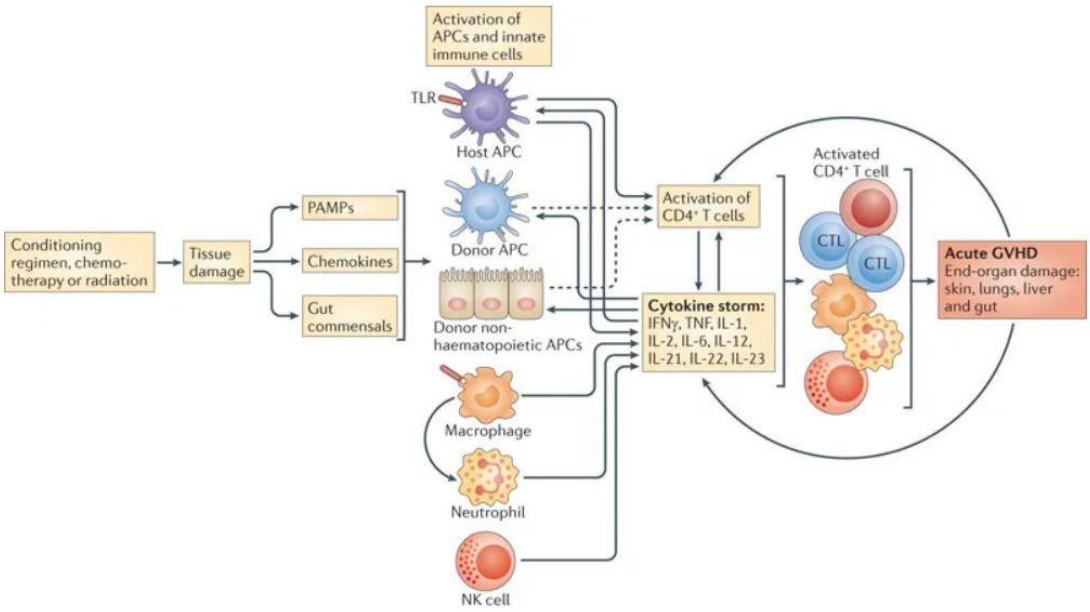


Figure 1.3 Development and persistence of acute GvHD consists of several phases. The conditioning regimen leads to the activation of antigen-presenting cells via tissue destruction. Through the release of gut bacteria, chemokines and PAMPs, conditioning can also lead to the activation in immune cell types that participate in direct tissue damage. Following antigen presentation, a cytokine response is initiated, which promote the recruitment of effector T cells, which in turn augment the pro-inflammatory milieu. NK cells, T effector cells, macrophages and pro-inflammatory cytokines result in end-organ damage. The resulting damage can then amplify the process to more sever stages of GvHD. CTL, cytotoxic T lymphocyte; IFNγ, interferon-γ; TLR, Toll-like receptor. Figure adapted from Blazar *et al*[31].

1.1.3 Summary

HSCT is a curative procedure for a variety of conditions spanning malignant and non-malignant haematology, primary immunodeficiencies and inborn errors of metabolism. The field has advanced considerably since the 1950s, with a significant increase in survival rates due to reduced-toxicity conditioning regimes, infection monitoring and prophylaxis and treatment of post-transplant complications. Despite this, the procedure is associated with numerous adverse side effects including GvHD and an increased risk of infection and is therefore not without risk.

1.2 The human microbiome

The human microbiome is the collection of all microorganisms that live in and on the human body. The term collectively includes all microorganisms- bacteria, viruses, archaea, fungi- however from here on microbiome/microbiota will be used to solely refer to the bacterial component. The virome, mycobiome and archaeome components are not investigated in this work.

Although the field has received considerable interest in the past few decades, the first concepts of the microbiome came as early as 1853, in Joseph Leidy's book 'A Flora and Fauna within Living Animals' and the use of faecal transplantation has been mentioned as early as the 4th century in traditional Chinese medicine as reviewed by Stripling *et al* [36].

The use of anaerobic culture and the realisation that germ-free mice lack normal physiology which could be reconstituted by faecal transplantation, propelled the research forward. The field became particularly mainstream in part as a result of the seminal work undertaken by the Gordon group in the USA, which linked gut microbiota to obesity and for the first time demonstrated phenotype transfer *via* the transfer of the gut microbiota[37].

Since then, the decrease in cost of next generation sequencing as well as the use of 'omics' technologies, such as metabolomics and transcriptomics, has given us new insights into the complex composition of the microbiome at high resolution. The NIH Human Microbiome Project (HMP) was the seminal body of work that aimed to characterise the microbiome of healthy individuals at several sites. As a result of the project, several papers reported on the composition and diversity of the microbiome at the gut, oral, throat and other sites[38,39]. One of the key findings was that taxonomic composition was more variable between individuals, than metabolic pathways, which suggested that multiple taxa may perform the same functions, ensuring homeostatic fitness[38].

1.2.1 The gut microbiota

Development

Upon birth, neonates are exposed to a variety of bacteria that begin to colonise the body. The finding of a 'placental microbiota' has caused some controversy about whether the placenta, previously assumed to be sterile, is in fact colonised by bacteria[40]. Several papers have found bacteria within the placenta, however most recently, de Goffau *et al* have linked these findings to contamination, which is known to be more significant in low biomass studies[41]. Amniotic fluid however, does appear to have at least a simple microbiome, although this is also controversial[42]. It is likely therefore, that there may be some bacterial exposure prior to birth, but that colonisation mostly happens during vaginal delivery or Caesarean section[43].

For the first few years of life, the gut microbiota is relatively immature with few taxa in comparison to what is observed in adults. Figure 1.4 details gut microbiota development and maturation. One of the largest studies to date (903 infants, 3 countries) classified gut microbiota development into three phases, based on the dynamics of the most abundant bacterial phyla (*Actinobacteria*, *Bacteroidetes*, *Firmicutes*, *Proteobacteria* and *Verrucomicrobia*)[44]. The phases include a developmental phase (months 3-14), in which the phyla detected start to gradually change with *Bacteroidaceae* and *Enterobacteriaceae* being predominant; a transitional phase (months 15-30), in which an increase in the *Bacteroides* and a decrease in *Enterobacteriaceae* are observed; and a stable phase (≥31 months) in which the phyla stabilise. *Bifidobacterium* species are found to dominate the developmental phase, whereas *Firmicutes* predominate the stable phase. Most of the gut microbiota development is thought to occur within three years, therefore the gut microbiota children below the age of 3 is both taxonomically and functionally different to adults[45] (Figure 1.4).

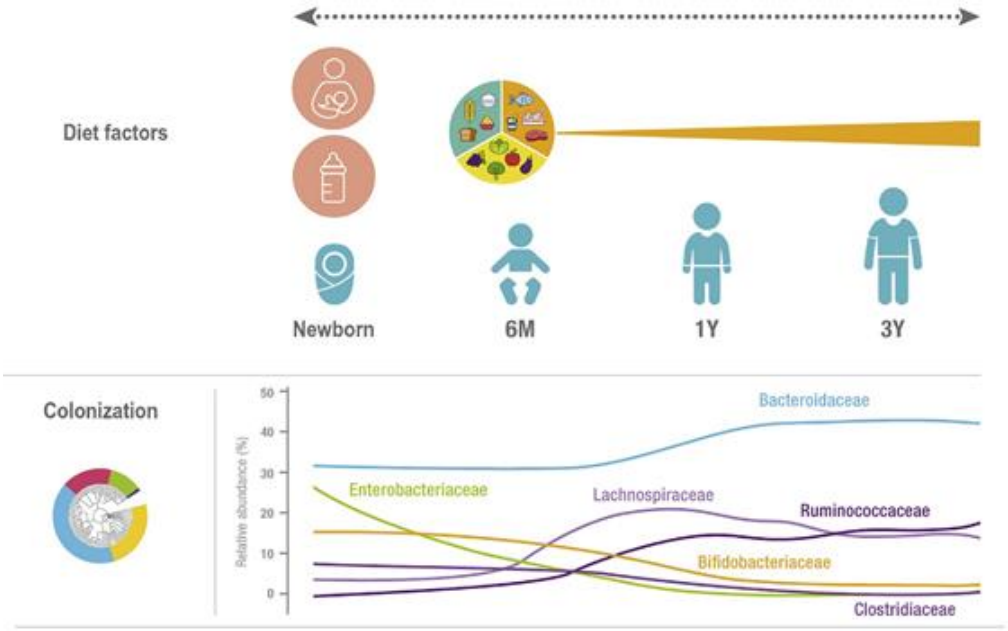


Figure 1.4 The development of gut microbiota. External factors such as mode of delivery and feeding patterns shape the development of the gut microbiota. Figure adapted from Derrien *et al*[45].

The establishment of the gut microbiota is now considered a key determinant in promoting maturation of the gut, its associated immune system and the maturation of distal organs, which affect systemic homeostasis [46,47].

Following on from birth, a variety of factors shape the gut microbiota including mode of delivery (Caesarean *vs* vaginal delivery) and subsequent nutrient supply (formula *vs* breast milk)[48]. Extensive research indicates that breastfeeding (associated with greater levels of *Bifidobacterium)* promotes greater maturation of the gut microbiota when compared to infants who are bottle-fed[44]. Interestingly, vaginal delivery favours the presence of *Bacteroides*, which is independently associated with gut microbiota maturation[44].

More recent studies suggest that a decrease in *Bifidobacterium* and other anaerobes and an increase in pro-inflammatory faecal metabolites in infants can lead to the development of cow's milk allergy, atopy or asthma[49]. In a murine model, a faecal microbiota transplant (FMT) from an infant with cow's milk allergy oriented the murine immune system towards an atopic profile with enhanced clinical symptoms of allergy[50]. Additionally, epidemiological studies suggest that

early-life exposures to antibiotics may increase one's risk of obesity[51]. Taken together, there is unequivocal evidence that the gut microbiota and its establishment is crucial for host homeostasis.

Gut microbiota and the immune system

Upon birth, a baby is colonised by an array of bacteria, which signal the start of the microbiota development. The initial encounters between the host immune system and the microbiota have profound and long-term effects on human health and the microbiome is known to be critical in the appropriate development of the human immune system.

Both colostrum and milk are known to contain live bacteria, IgA and cytokines which are thought to shape the infant microbiome and the immune responses to it. Maternal IgA for example, binds bacterial antigens in the presence of oligosaccharides and thus restricts immune activation[52]. Early exposure is also known to repress invariant natural killer T cells, which are involved in pro-inflammatory responses and can have long-term consequences for one's capacity to develop inflammatory conditions[53].

The importance of early-life exposures to the microbiome become clearer with germ free (GF) animals. Hansen *et al,* found that a short GF postnatal period caused permanent changes in the levels of systemic regulatory T cells and NK, in comparison to colonisation from birth[46]. Inoculating GF mice with caecal content 3 weeks after birth induced a pro-inflammatory immune response[46]. GF mice are also known to present with smaller lymphoid structures[52].

Similarly, early disturbances to the microbiome have been associated with a variety of allergic diseases including asthma. Russell *et al* administered vancomycin and streptomycin to mice during the neonatal period and found that vancomycin depleted the gut microbiota and led to indicators of allergic asthma and a decrease in cells expressing the $CD4^{+}CD25^{+}Foxp3^{+}$ T_{reg} phenotype[54]. Similarly, Ni *et al* found that the use of antibiotics during the first year of life is associated with lifetime asthma and the microbiome is thought to play a role in this[55].

It is likely that many bacterial strains co-evolve with the immune system to ensure stable long-term colonisation. There are several well-described examples of species-specific immunomodulation of the host immune system. *B. fragilis* and *B. breve, for* example, have evolved mechanisms to promote their own immunotolerance by the host. *B. fragilis* uses polysaccharide A, a capsule component, to interact with antigen presenting cells (APCs), such as dendritic cells, to induce IL-10 production by T_{reg} cells, which allows it to colonise the mucus layer[56]. *B. breve* uses its exopolysaccharides to induce IL-10 production, as well as preventing a B cell response and decreasing pro-inflammatory cytokines[56].

Clostridia are also well known to induce regulatory T cells and a combination of strains appear to be more effective than a single strain[57]. Bacterial coating with secreted immunoglobulin A (sIgA) is also thought to promote immunological tolerance. Rag-1 knockout mice, which effectively have no adaptive immune system, demonstrated that coating *B. thetaiotaomicron* with sIgA reduces inflammatory signalling and bacterial epitope expression[58].

1.2.2 Factors shaping the gut microbiota

Core gut microbiota

The gut microbiota has been extensively profiled and multiple studies have attempted to define its core composition. Once established, the gut microbiota is dominated by five major phyla: *Firmicutes, Bacteroidetes, Actinobacteria, Proteobacteria* and *Verrucomicrobia* in decreasing abundance[59,60]. Although a core microbiota has been defined, a great deal of inter-personal variation is common and this is greater than intra-personal variation over time. David *et al* found that the gut microbiota is relatively stable long-term, however lifestyle choices such as diet can affect the composition on daily timescales[61]. Significant events such as antimicrobial administration or travel were also conversely associated with profound disturbances to the microbiota community structure[62]. In contrast to this, Martinson *et al* observed a population-level turnover of *Enterobacteriaceae* resident clones over short periods, which underlie stability at higher taxonomic levels[63].

Diet

Diet-derived polysaccharides shape the intestinal microbiota. Whilst the composition of the luminal gut microbiota appears to fluctuate with dietary changes, the composition of the mucosal microbiota is likely to be more stable. Schluter *et al* propose that host selection is a key mechanism in the stabilisation of mutualism between a host and its microbiome[64]. The epithelium-host interface acts as a selectivity amplifier via nutrient secretion and this mechanism is more powerful than the negative selection of antimicrobials[64]. *A. municiphila* for example, is an efficient mucin-degrader that is commonly found in the mucus layer and consumption of the glycans in the mucus allows it to colonise the mucus regardless of the diet. Other species are also able to colonise the mucus layer as they can digest specific, limited nutrients.

In an HSCT setting, diet is of profound importance for both nutritional support and for supporting the gut during severe GvHD. Historically, parenteral nutrition (intravenous) has been favoured over enteral nutrition for nutritional support. Despite this, prolonged periods of parenteral nutrition with no oral intake are linked to an increased risk of acute GvHD, whereas enteral nutrition is associated with an improved rate of overall survival, reduced risk of GvHD and reduced central line complications[65,66]. The mode of feeding has been linked to the changes in the gut microbiota. In adult HSCT populations patients predominately on parenteral nutrition exhibit increases in the *Enterococcus* genus in comparison to healthy adults, whereas those on enteral nutrition exhibit increases in *Clostridium cluster XIV_a* in comparison to patients on parenteral nutrition[67].

Antibiotics and the gut microbiota

Despite the tremendous benefits that antibiotics have provided for healthcare, they constitute a strong selective pressure on the microbiota and the effects of this are starting to be realised.

As a result of chemotherapy and prolonged immunosuppression, HSCT patients are at a heightened risk of infections. Due to this, they receive prophylactic fluoroquinolones (in most cases ciprofloxacin) until haematopoiesis commences and, in the event of an infection, β-lactams therapeutically. A 2012 Cochrane analysis found that prophylactic use of fluoroquinolones in haematological HSCT

patients reduces the relative risk of mortality, bloodstream infections (BSI) and fever during neutropaenia[68]. Despite this, numerous studies indicate that antimicrobials affect the gut microbiota and their use is in turn linked to an increased risk of non-relapse mortality and GvHD[69,70].

Due to the use of multiple antimicrobials throughout treatment, patients undergoing HSCT have been considered as potential reservoirs of antimicrobial resistance. The gut microbiota of the patients was found to act as a dynamic reservoir of antimicrobial resistance genes[71]. Patients with acute GvHD exhibit distinct profiles of resistance genes including an expansion of those present prior to HSCT as well as acquisition of antimicrobial resistance genes, at times not directly related to the antibiotics given[71].

Additionally, recent research has highlighted the unintended effects of antibiotic administration on the efficacy of anti-tumour immunotherapy[72]. Immune checkpoint inhibitors (ICI) induce clinical responses against a variety of cancers in a minority of patients. The study found that administration of antibiotics to mice inhibited clinical benefits of ICI, however an FMT from patients that responded to ICI into mice exhibited ameliorated anti-tumour effects of blockades[72].

Additionally, HSCT patients represent a unique cohort which receives repeated antibiotic administrations. Several studies have investigated the effects of antibiotics on the composition and the stability of the gut microbiota in healthy individuals. Jakobsson *et al* found that a week-long administration of clarithromycin and metronidazole led to a profound change in the gut microbiota composition[73]. Although the diversity of the microbiome returned to its initial state, the composition remained disturbed up to four years post treatment[73]. The findings of early work have been reproduced by more recent work. Dethlefsen *et al* found that the administration of two courses of ciprofloxacin led to a decrease in diversity, which returned within weeks, however this return was often incomplete[74]. Finally, Palleja *et al* administered a 4-day course of gentamicin, vancomycin and meropenem to healthy individuals[75]. They observed blooms of *Enterobacteria* and a loss of butyrate-producing anaerobes as well as a decrease in diversity. Although diversity recovered within 1.5 months, certain taxa remained undetectable six months after treatment and antibiotic resistance carriage is thought to modulate the recovery process[75].

Overall, antibiotic administration provides a negative selection pressure on the gut microbiota and is one of the most dramatic disruptions that it can experience. Repeated administration of antibiotics, as is common with prophylaxis in HSCT patients, is likely to have unintended consequences on the gut microbiota.

1.2.3 Summary

Overall, children are distinct to adults in their gut microbiota composition and function. In a healthy host, the gut microbiota interacts with and contributes to the host metabolism, resulting in immune tolerance and a mutually beneficial relationship. As a result of these associations, changes to the microbiome early in life from numerous factors such as one's diet and antimicrobial usage can lead to unintended consequences such as higher risks of allergic diseases.

1.3 Microbiome in haematopoietic stem cell transplantation

The interest in the role of the microbiome in HSCT began with the finding that GF mice undergoing HSCT experienced less GvHD and had better survival than conventional mice[76,77]. Although gut decontamination, whereby the gut is decontaminated of any bacteria, is no longer undertaken at most paediatric centres including GOSH, as clinical studies have failed to show consistent benefit, the development of next-generation sequencing has led to a renewed interest in investigating the role of the microbiome in HSCT[78].

Multiple studies to date have profiled gut microbiota during HSCT, predominately in adult patients. Overall trends point to significant changes throughout the procedure with domination of a single taxon, such as *Enterococcus* and *Staphylococcus,* common post-transplantation and bacteraemia by these organisms is often found to be preceded by their expansion[79]. Additionally, a depletion of obligate anaerobes and diversity post-HSCT is frequent[79-81].

Paediatric studies remain limited and Biagi *et al* were one of the first to examine alterations to the paediatric gut microbiota throughout HSCT[82]. Broadly, they found that individuals appeared healthy in terms of SCFA production and gut microbiota composition prior to HSCT, however transplantation induced significant changes to the composition with the loss of anaerobes and thus SCFA levels. SCFA are produced by anaerobic fermentation of complex carbohydrates

and are the primary source of fuel for intestinal colonocytes, which utilise it for energy production. SCFA are thought to exert multiple beneficial effects on the human host including the maintenance of the epithelial barrier and its effects on the immune system[83]. Within two months individuals recovered the loss in diversity and SCFA levels, returning to their initial state[82]. Others also observe decreases in diversity and temporal variation in phyla and genera in acute lymphoblastic leukaemia patients undergoing HSCT[84].

Thus overall, the procedure and associated treatment, such as the use of antibiotics, disrupt both the composition and the functionality of the gut microbiota as well as affecting the host, which allows for the expansion of pathogenic facultative bacteria.

1.3.1 Association with clinical outcomes

Studies in both mice and humans suggest important associations between the gut microbiota and clinical outcomes post-HSCT. The section below summarises existing evidence for these associations in both paediatric and adult patients.

As previously mentioned, gut microbiota diversity decreases throughout HSCT. These changes can be rapid, likely as a result of conditioning, the use of antibiotics and dietary changes. A retrospective study on 80 adults receiving an allogeneic transplant found that low diversity at the time of engraftment and higher abundance of *Proteobacteria* is associated with a higher risk of mortality[69]. Gut microbiota changes have also been linked to relapse post-HSCT. Peled *et al* found that higher abundances of *Eubacterium limosum* were associated with a lesser risk of relapse although the mechanism has not yet been explored[85].

Near-domination of a single taxon is also common in both populations and domination by *Enterococcus* or *Proteobacteria* has been linked to a higher risk of developing bacteraemia with vancomycin-resistant *Enterococcus* in an adult cohort[79,80]. In paediatric ALL patients, domination with *Enterococcaceae* or *Streptococcaceae* during chemotherapy was found to precede BSI in the subsequent phase of chemotherapy[86]. Similarly, strains causing BSI as a result of mucosal barrier injury were found to dominate the gut of paediatric patients and appeared to precede the episodes by a median of 17 days[87].

Finally, gut microbiota has been found to play an important part in the development of GvHD. A study of adult HSCT patients found a decrease in diversity and *Blautia* at day 12 post-transplantation, which correlated to increased GvHD-mortality[88]. Another study found an increased abundance of *Lactobacillales* and a decreased abundance of *Clostridiales* in mice with GvHD, which mirrored findings in adult HSCT patients[89]. Eliminating *Lactobacillales* prior to HSCT in mice aggravated GvHD and re-introducing *Lactobacillus johnsonii* improved it, suggesting that *Lactobacillales* induce a protective effect.

Further mechanistic work has found that in mice, the gut microbiota-derived metabolite butyrate is decreased in intestinal epithelial cells (IEC) post-HSCT, which results in decreased histone acetylation[34]. Butyrate restoration mitigated GvHD, improved IEC junctional integrity and decreased apoptosis, demonstrating a role of gut microbiota-derived metabolites in GvHD[34]. Additionally, donor-derived IL-17A appears necessary for immune tolerance and can mediate late immunopathology of GvHD. IL-17A-deficient mice developed acute intestinal GvHD and a microbiota transfer from IL17A-deficient mice induce a GvHD phenotype in wild type co-housed mice[90]. It is likely that GvHD, among other clinical outcomes post-HSCT, is a result of complex ecological interactions between the taxa as well as their direct or indirect interactions with the host.

These findings indicate the importance of maintaining host-microbiome gut homeostasis, including the epithelial barrier, throughout HSCT.

1.3.2 Gut microbiota and predictive biomarkers

Identifying biomarkers with a high prognostic value is crucial for stratifying patients at risk, as well as for selecting timely therapeutic approaches. The links between the gut microbiota and clinical outcomes post-HSCT mean that several studies have investigated it for potential biomarkers.

Low urinary 3-indoxyl sulfate (3-IS) levels (a major conjugate of indole and a by-product of the gut microbiota) 10 days post-HSCT were found to be associated with higher TRM and lower overall survival a year after HSCT[91]. Furthermore, risk factors of low 3-IS included gut decontamination and early antibiotic use and higher levels were associated with *Lachnospiraceae* and *Ruminococcaceae* families, which are linked to lower gut inflammation[91]. Although a potentially

useful biomarker it has not been widely utilised, in part potentially due to difficulties in using mass spectrometry in clinical diagnostics. Furthermore, a genetic modifier of the gut microbiota fucosyl transferase 3, was found to modify the risk of GvHD and bacteraemia post-HSCT[92]. The proposed mechanism is through the glycosylation of the intestinal surface proteins.

Other work has identified microbial taxa predictive of various outcomes including GvHD and mortality post-HSCT. Prior to conditioning, ≥5% *Enterobacteriaceae* associated with a higher risk of sepsis, whereas ≤10% *Lachnospiraceae* associated with a higher risk of overall mortality in adult patients[93]. Similarly, there was a decrease in diversity and lower abundances of various taxa including *Fecaelibacterium* and *Sutterella* prior to BSI development in patients undergoing HSCT for non-Hodgkin's lymphoma[94]. At neutrophil engraftment (approximately 15 days post-HSCT) lower diversity and the presence of oral *Actinobacteria* as well as oral *Firmicutes* were positively correlated to GvHD development, whereas *Lachnospiraceae* were negatively correlated[95].

1.3.3 Gut microbiota and immune reconstitution

Timely immune reconstitution is a key factor to a successful recovery post-HSCT and delays can have an impact on complications and clinical outcomes post-transplantation. Immune reconstitution involves both innate immunity components (neutrophils, monocytes etc) and adaptive immunity (T and B cells). Evidence that certain taxa can influence immune cell differentiation makes it likely that the gut microbiota can, directly or indirectly, influence immune reconstitution post-transplantation. One of the few observational studies found that paediatric patients with high abundances of obligate anaerobes such as *Ruminococcaceae* showed rapid NK and B cell reconstitution, no or mild GvHD and low mortality in comparison to paediatric patients grouped into other clusters with high abundances of *Staphylococcus, Enterobacteriaceae* and *Lactobacillaceae*[96]. A trial which looks to investigate a similar hypothesis in adults is currently ongoing (NCT03616015).

1.3.4 Summary

In summary, there is increasing evidence that the gut microbiota influences clinical outcomes and the clinical course post-HSCT. Additionally, it has proven

to be a potential source of biomarkers for predicting clinical outcomes. Although the adult HSCT population has been repeatedly profiled, there is currently an apparent lack of studies exploring the gut microbiota throughout HSCT in paediatric cohorts.

1.4 Hypothesis and aims

We hypothesise that the gut microbiota will be profoundly affected by the HSCT procedure, both taxonomically and functionally.

The following work broadly aims to explore the gut microbiota of patients undergoing HSCT both taxonomically (using 16S ribosomal ribonucleic acid (rRNA) sequencing) and functionally (using nuclear magnetic resonance spectroscopy) with the following specific aims.

- We aim to optimise the 16S rRNA workflow
- We aim to explore the dynamics of the paediatric gut microbiota during HSCT
- We aim to explore the paediatric gut microbiota for clinical biomarkers at several timepoints.
- We aim to explore faecal metabolites during paediatric HSCT

Chapter 2- Materials and Methods

2.1 Ethical approval and consent

Written, informed consent was obtained from patients and/or their legal guardians. The study was approved by an NHS Health Research Authority ethics committee (17/SW/0061; 14/LO/0364).

The following sections detail the methods used for chapters 4, 5 and 6. Methods for chapter 3 are detailed within the chapter itself.

2.2 Cohort

The cohort studied includes 64 paediatric patients admitted to GOSH for haematopoietic stem cell transplantation between July 2017 – January 2018. Patient characteristics are detailed in Table 2.3.

Most patients underwent allogeneic HSCT, but several autologous transplantations including 5 gene therapies and 2 CART cell infusions were also performed. Underlying diagnoses for transplantation were mostly haematological such as ALL and AML. Immunodeficiencies were mostly a mixture of Wiskott-Aldrich syndrome, SCID and chronic granulomatous disease (CGD) patients and metabolic patients all had Hurler's syndrome. Preparative regimens were defined as follows: reduced-intensity conditioning (RIC) included treosulfan/fludarabine or fludarabine/melphalan or reduced-intensity busulfan/fludarabine (targeted busulfan area under the curve of 45 to 65 mg*h/L). Myeloablative (MAC) protocols included myeloablative busulfan/fludarabine (targeted busulfan area under the curve, >70 mg*h/L) or treosulfan/fludarabine /thiotepa. Within our cohort, 59% of patients had MAC.

In terms of clinical outcomes, the mortality rate within the cohort was 23% and TRM consisted of transplant complications, GvHD and infectious complications (Table 2.3). 38% of patients had clinically relevant grades of GvHD (≥II). GvHD grading was done by clinicians according to the Glucksberg criteria[97]. Briefly, the stage of the GvHD is graded within each organ- skin, liver or gut, based on clinical parameters such as the levels of diarrhoea or liver bilirubin (Table 2.1). This can then be translated into GvHD grades, for example grade I GvHD relates to skin involvement only at stages + or ++, meaning a maculopapular rash on up to 50% of body surface (Table 2.2).

Table 2.1 The organ grading system for the Glucksberg acute GvHD classification. Table adapted from Hatzimichael *et al*

Stage	Skin/maculopapular rash	Liver/bilirubin, µmol/L	Gastrointestinal/diarrhoea
+	<25% of body surface	34-50	>500mL
++	25-50% of body surface	51-102	>1000mL
+++	Generalised erythroderma	103-255	>1500mL
++++	Generalised erythroderma with bullae formation and desquamation	>255	Severe abdominal pain with or without ileus

Table 2.2 The overall Glucksberg grading system for acute GvHD. Table adapted from Hatzimichael *et al*

Grade of aGvHD	Degree of organ involvement
I	Skin + to ++
II	Skin + to +++ Gut and/or liver + Mild decrease in clinical performance
III	Skin ++ to +++ Gut and/or liver ++ to +++ Marked decrease in clinical performance
IV	Skin ++ to ++++ Gut and/or liver ++ to ++++ Extreme decrease in clinical performance

Viraemia, such as infections with ADV and CMV, was particularly prevalent within the population. Bacteraemia, however, was less frequent with several cases of *Staphylococcus* (coagulase-negative *Staphylococcus*; *Staphylococcus hominis; Staphylococcus epidermidis*) and *Streptococcus* species. Bacteraemia was defined as an isolation of a non-commensal organism on at least two consecutive occasions that was eventually treated according to the US National Healthcare Safety Network, whereas viraemia was defined as a viral blood isolate in the blood, regardless of viral load[98].

We collected an average of 8 faecal samples (range; 2-32) per person. Additionally, we sequenced single samples from 8 healthy controls. The controls were not matched to the patients and consisted of stool samples from 5 females and 3 males with the age range of 4-14 years (median- 8 years). The healthy controls did not receive antibiotics 6 months prior to sample collection.

Table 2.3 Cohort characteristics table

Patient Characteristics (N =64)	No (%)
Age at transplant, median (range)	5.3(0.4-14)
Age at transplantation, years	
<2	16(25)
>2	48(75)
Sex	
Male	40(62)
Female	24(38)
Underlying diagnosis	
Haematological malignancy	33(51)
Haematological non-malignancy/Other	9(14)
Immunodeficiency	19(30)
Metabolic	3(5)
Conditioning regimen	
Myeloablative	38(59)
Reduced intensity	26(41)
Treatment[1]	
HSCT	57(89)
CART cells	2(3)
Gene therapy	5(8)
Stem cell source	
Bone Marrow	33(58)
Cord	6(11)

Peripheral blood	18(31)
Donor type	
Matched Sibling/Family[2]	9(16)
Matched unrelated[2]	24(42)
Mismatched/Haplo[3]	24(42)
More than one transplant over lifetime	
Yes	9(14)
No	55(86)
Mortality, No. (%)	15(23)
Transplant-related[4]	11(17)
Relapse	4(6)
Acute GvHD	
Grade 0- I	34(61)
Grade II -III	21(37)
Patients with infection outcomes, No. (%)	
Bacteraemia	6(9)
Viraemia	42(66)

[1]Two patients had 2 transplants. A single patient had a peripheral blood and a cord transplant, whereas 1 had a bone marrow and a peripheral blood transplant. A single patient had 2 CART cell infusions and an HSCT.[2]Matched refers to a full HLA match (10/10; 12/12); [3]Mismatched refers to a lesser HLA match. Haplo refers to a half HLA match to the patient. [4]Transplant-related mortality is defined as mortality due to a complication other than a relapse following an HSCT.

2.3 Faecal sample collection and initial processing

Faecal samples utilised in this study were collected in 30ml universal containers by the nurses on the ward and stored at 4ºC until receipt (typically within 4 hours). Upon receipt, samples were thoroughly mixed, aliquoted to 200mg and frozen at -80°C until extraction.

2.4 DNA extractions

Faecal samples (200mg) were extracted using a Qiagen QIAamp DNA stool kit (Qiagen, Germany) as per manufacturer's instructions with the following modifications. Once buffer ASL was added, the sample was heated to 95ºC, in order to aid bacterial cell lysis. Homogenisation was done with the aid of Lysing Matrix E beads (MP Biomedicals, USA) using the TissueLyser LT (Qiagen, Germany) with 8 bead-beating steps (60s, 50osc) and 60s rest on ice in between each step. Negative extraction controls were run alongside each batch of

extractions. Deoxyribonucleic acid (DNA) was eluted in 200µl of AE buffer and stored at -20ºC until further processing. DNA obtained was quantified using a dsDNA HS Assay Kit as per manufacturer's instructions (1µl DNA + 199µl master mix, ThermoFisher Scientific, USA) using the Qubit (Version 2.0, ThermoFisher Scientific, USA).

2.5 PCR amplification

Extracted DNA was diluted 1:10 in nuclease-free water and amplified in a single-step polymerase chain reaction (PCR). Samples that failed to amplify above 1ng/µl were re-amplified in neat. Primers spanning the V3-4 region of the 16S gene were used (Appendix; Table A1. Adapted from Kozich *et al* [99]). As the extracted DNA is amplified, two different index sequences are added to each sequence, flanking the specific 16S rRNA primers and Illumina-specific adapters on either side. This means that each sample contains 2 unique barcodes, allowing sequencing of 96+ samples at any one time.

The PCR mixture (50µl total volume) consisted of 5µl DNA template, 1x PCR Buffer, 1x Q Solution, 2mM $MgCl_2$, 0.5µM of an appropriate forward and reverse primer, 200µM of each dNTP, 2.5U Taq DNA polymerase/reaction and nuclease-free water (Taq PCR Core Kit, Qiagen, Germany). Cycling conditions were as follows: 95°C for 3 minutes followed by 30 cycles of 95°C for 30s; 54°C for 30s and 72°C for 1 minute. This was followed by a final extension of 72°C for 10 minutes.

In addition to the extracted DNA, a mock community consisting of DNA from 8 bacterial species was amplified and sequenced (D6305, Zymo, USA). Mock community composition is detailed in the Appendix (Table A2). Negative extraction controls (NEC) consisting of buffer and a PCR control consisting of PCR reagents and water were also amplified and sequenced alongside the study samples.

2.6 DNA library construction and sequencing

PCR products following each amplification were purified using Agencourt AMPure XP beads as per manufacturer's instructions (0.7x beads/reaction; Beckman Coulter, UK). DNA was then quantified using the Qubit dsDNA HS kit

(1µl DNA + 199µl master mix), normalised to the lowest concentration (generally ≥1ng/µl) in nuclease-free water and pooled to produce a library. To confirm primer amplification and library composition the pooled library was analysed using the Agilent Tapestation (4200, Agilent, USA) and quantified using the NEBNext Library Quant Kit for Illumina (NEB, USA) according to the manufacturer's protocol.

For sequencing, pooled libraries were denatured using freshly prepared sodium hydroxide (0.2M; 1:1) and inactivated using Tris-HCL (200mM; 1:1) followed by the addition of HT1 buffer (Illumina, USA) to make a 4pmol library. It was then spiked with 10% PhiX (Illumina, USA) and sequenced using a V2 kit (500-cycle) on the Miseq sequencing platform (Illumina, USA). Libraries and PhiX were loaded at a concentration of 3.6pmol and 2pmol respectively. Read 1, read 2 and the index primers were spiked into the cartridge at 0.5µM.

2.7 Bioinformatics and dataset pre-processing

Data was demultiplexed (barcodes removed) with version 2.1.12 of the Illumina instrument control software. Analysis of sequence data was performed using Mothur (version 1.35.1)[100]. Modifications to the standard Mothur protocol are as follows: the contigs were made using standard settings. The assembled sequences were screened using the 'screen.seqs' command to remove ambiguous sequences and sequences containing homopolymers longer than 8 base pairs (bp). In addition, any sequences longer than 460bp were removed. Unique reads were aligned using a region-specific Silva bacterial database (release 128). Any sequences outside the expected alignment coordinates were removed. The correctly aligned sequences were subsequently filtered ('filter.seqs') with 'vertical=T' and 'trump=.'. The filtered sequences were de-noised by allowing six mismatches in the "pre.clustering" step and chimeras were removed using vsearch with the dereplicate option set to 'true'. Chimera-free sequences were classified using the Mothur-formatted Bayesian RPD database and a cut-off value of 80. Mitochondria, archaea, chloroplast, eukaryota and unknown sequences were removed. A distance file was generated with a 0.03 cut-off and the resulting matrix was split and clustered with large option set to 'true'. A shared file was then made and operational taxonomic units (OTU) were

classified. The resulting shared, taxonomy and metadata files were combined into a biom file and were exported for further analysis.

Each dataset was initially filtered at a level that was chosen from the mock community sequenced with each run (Chapter 3). The mock community was first assessed for any OTUs that would not be expected to be there. An abundance level was then chosen to filter out OTUs that were not expected to be present, whilst maintaining the expected composition of the mock community. Following this, taxa not classified at the phylum level were removed, as these are unlikely to give much insight into sample composition. Finally, extraction controls were inspected individually and OTUs or taxa deemed to be contaminants were removed from the datasets. OTUs with a higher abundance in samples than in extraction controls were not considered contaminants, whereas those with a higher abundance in extraction controls than in samples were considered contaminants and were removed. Table 2.4 details an example of this. Contaminant OTUs were further inspected and those that have been previously published as contaminants and were not of human origin (e.g. only reported to be found in lake sediment) were removed in their entirety.

Table 2.4 Contamination removal strategy

Hypothetical OTU*	Extraction control	Patient Sample	Outcome
OTU 001	Higher	Lower	Contaminant
OTU 002	Lower	Higher	Not a contaminant

*For any single OTU found in an extraction control a decision is made based on its abundance in patient samples. If they are more abundant in patient samples than in extraction controls they are not considered to be a contaminant. If they are more abundant in the extraction control, they are considered a contaminant and are subsequently removed.

The datasets were then agglomerated to genus level and OTUs were appended with the genus name prior to combining into a single dataset. There was a wide range of reads within samples (184-349993) and the dataset was further filtered to only keep samples with >2000 reads, resulting in the loss of 28 samples. This resulted in a dataset with 540 samples, including 8 healthy controls.

2.8 Further analysis

The dataset was subsampled without replacement to an equal depth of 2100 reads per sample in R for the comparisons of alpha diversity. Alpha diversity and beta diversity at this level were representative of diversity in all samples, however diversity did collapse below 1000 reads (data not shown). Shannon effective entropy was calculated using Rhea in R by exponentiating Shannon entropy[101]. Further data analysis was conducted using vegan, phyloseq and microbiome packages in R (version 3.5.1).

T-distributed stochastic neighbour embedding (t-SNE) plots were generated using tsnemicrobiota package in R with perplexity set at 25 and using 'bray' distance. t-SNE is a dimensionality-reduction technique, particularly suited to visualising high-dimensional datasets, which uses local relationships between points to create a low-dimensional mapping[102].

2.8.1 Clustering into community state types

In order to partition the data into state types based on composition, the raw, genus-aggregated dataset was imported into the package Deseq2 and transformed using the function 'varianceStabilizingTransformation'. This accounts for both library size differences and biological variability.

Samples were assigned to community state type(s) (CST(s)) by partitioning around medoid (PAM) clustering using the function 'pam' in package cluster based on a Jensen-Shannon distance. The number of clusters was determined by using the gap statistic evaluation and silhouette width quality validation. R code used in this part of the analysis was adapted from Ingham *et al*[96]. To identify the best number of clusters, k, we employed the Gap statistic. It assesses a metric of error (the within cluster sum of squares) with regard to the choice of k. Generally, the error decreases steadily as k increases. The point at which the error stops decreasing substantially is considered the best number of clusters within the data. The results were then evaluated in an NMDS ordination. We also used a silhouette plot, which evaluates cluster separation distance. The higher the cluster, the better the classification within a cluster i.e. negative values indicate that samples may have been assigned into the wrong cluster or could fit into more than one cluster (Appendix; Figure A5).

In individuals with two allogeneic transplants, only samples from the first transplant were used for producing CST assignment, transition and Cox models. In the individual with three transplantations, only samples relating to the allogeneic HSCT were utilised.

2.8.2 Transition models

We further assessed patients' movement between CSTs throughout transplantation by using transition models. We were initially interested in how often a cluster was preceded by another cluster, and thus all samples were used for this analysis. To investigate the dynamics of cluster progression over time in more detail, we subset the dataset to samples collected in the first 5 weeks, starting at day -7 relative to transplantation, as the number of patients remaining in hospital steadily decreased after week 5. In the case of more than one sample being collected within a week, the first sample was retained. For both analyses, a transition matrix containing frequencies of transitions between respective CSTs was generated and plotted using the qgraph package in R. The function 'sojourn.msm' was used to estimate the mean sojourn times for each CST. R code used in this analysis was adapted from Stewart *et al*[44].

2.8.3 Time-dependent Cox models

In order to evaluate associations between CSTs and clinical outcomes, right-censored time-dependent Cox models were used. We focused on GvHD and viraemia here, as there were not enough bacteraemia nor mortality cases in the cohort for this type of analysis.

The following formula was used: coxph(surv(start_time, end_time, dependent variable)~ independent variables+ cluster(patient_id)). Using GvHD (≥II) or viraemia as the dependent variable, we initially performed univariate analysis with the following independent variables: sex, age (>2), diagnosis (malignant), multiple transplants, conditioning (myeloablative), Shannon diversity, stability and CST. Stability was determined as the number of different CSTs that the individual has been in throughout hospitalisation. Models for GvHD were also adjusted for serotherapy (in vivo) and graft manipulation (in vitro). Multivariable analysis with significant variables was then performed.

A univariate model was then run using viraemia as a dependent variable and taxon dominance (>30%), predominantly found in CST3, as an independent variable (vs non-dominant samples in other CSTs). Finally, *Enterococcus* dominance (>30%) was used as a dependent variable in a univariate model with sex, age, diagnosis, multiple transplants, conditioning, cell source, quinolones (ciprofloxacin, moxifloxacin), broad-spectrum beta lactams (ceftazidime, co-amoxiclav, imipenem, meropenem, piperacillin-tazobactam) and macrolides (azithromycin, clarithromycin, erythromycin) as independent variables, followed by a multivariable model adjusted for significant variables from the univariate model. Covariates included in the models were selected *a priori* from domain knowledge before running the model. To reduce confounders, only samples from individuals receiving an allogeneic HSCT were used in time-dependent Cox models.

A p-value of ≤0.05 was considered significant. Repeated measures were adjusted using robust sandwich estimator (cluster()). Cox models were run using survival package in R[103].

2.8.4 Biomarker discovery

The dataset was filtered to keep the taxa with at least 5 reads in at least 3 samples, to decrease false discovery rates. This led to a total of 74 genera and 42 families. In individuals with two transplants, only samples from the first transplant were used in this analysis and in the individual with three transplantations, only samples relating to the allogeneic HSCT were utilised. To reduce confounders, only samples from individuals receiving an allogeneic HSCT were used in this analysis. A single sample for each individual was used at each time point. The first sample collected for each individual was used as the 'baseline' sample, on average 10 days prior to transplantation (range; -44;-1); whereas the first sample prior to engraftment, on average 16 days post transplantation (0;60), was used as the 'pre-engraftment' sample for each patient.

We investigated viraemia, GvHD (acute i.e. within 100 days) and mortality (any cause) as clinical outcomes. Differences between baseline/engraftment samples with differing outcomes were interrogated using linear discriminant analysis (LDA) effect size (LEfSe)[104]. LEfSe uses the two-tailed nonparametric Kruskal-

Wallis test and LDA, which estimates the effect size of each differentially abundant OTU. For stringency an LDA score (log10) of > 2.5 was considered significant.

Receiver operator curves (ROC) of all differentially abundant taxa were made using the pROC package and optimal cut-offs were determined using cutpointr package in R based on optimal sensitivity and specificity. To aid selection of the most predictive taxa, we calculated the area under the curve (AUC) as well as a p value (in comparison to an AUC of 0.5, based on the ratio of standard error of the area under the curve). Taxa with an AUC >0.6 and a p value of ≤0.05 were explored further.

Logistic regression models were then employed to investigate associations between clinical outcomes (GvHD and viraemia) and differentially abundant taxa at specific cut-offs with an AUC >0.6. The following formula was used: glm(dependent variable ~ independent variables + family= 'binomial').

Initially, baseline characteristics including age, sex, diagnosis, multiple transplants, any antifungals and any antivirals were used as independent variables. The full model was then subjected to stepwise reduction using the Akaike criterion (AIC). Following this, specific taxa were used as independent predictors alongside any significant baseline characteristics and the model was again subjected to stepwise reduction using AIC.

In terms of mortality, differentially abundant taxa with an AUC of >0.6 and a p value of <0.05 were further examined using Kaplan-Meier (KM) curves and the log-rank test in Prism (version 5, Graphpad, USA).

2.9 Statistical analysis

Comparisons between DNA yields in the samples were performed using the Kruskal-Wallis test with a Dunn's multiple comparison test. Comparisons between alpha diversities were performed using the Kruskal-Wallis test with Benjamini-Hochberg correction. Comparisons between cohort characteristics in Tables 5.1 and 5.5 were performed using a T-test for numerical and Fisher's exact test for categorical variables.

2.10 Nuclear magnetic resonance

Sample processing and ^{1}H NMR spectroscopy

Faecal samples were randomized to prevent run order effects and thawed to room temperature. Approximately 100mg of a faecal sample was weighed and added to 700µL of dH_20. The samples were then homogenized using a bead beater (Precellys 24, Bertin Technologies, UK) with lysing matrix E beads, for 45 seconds at 6500rpm twice and subsequently centrifuged at 13 000rpm for 20 minutes (Eppendorf 5417 R, UK). The aqueous portion was removed and centrifuged again under the same settings and 630µL of the aqueous portion was mixed with 70µL of pre-prepared phosphate buffer containing 1.5M KH_2PO_4, 2mM NaN_3 and 1% trimethylsilypropanoic acid in D_2O (Sigma-Aldrich, Germany). The mixture was then vortexed and centrifuged at 13000g for 10 minutes and the resulting supernatant (600µL) was passed to a 5mm NMR tube. Quality control samples were inserted in each run, comprising of a pool of each sample, to ensure instrument performance. A 600MHz Bruker Avance III spectrometer was used to perform the spectroscopy. Standard one-dimensional (1D) pulse sequence with the NOE pulse sequence for water suppression was applied to obtain the NMR spectra. The raw spectra were calibrated to trimethylsilypropanoic acid using Topspin 3.2 (Bruker Biospin). The spectra were then imported into the MATLAB (Version R2018b; Mathworks Inc., USA) and the redundant regions- corresponding to water, trimethylsilypropanoic acid and urea- were removed. Manual alignment to the median spectrum and to quality control spectra was done using Recursive Segment-Wise Peak Alignment, followed by normalisation of spectra to the probabilistic quotient[105],[106].

Spectral processing and data analysis

Unsupervised principal components analysis (PCA) was performed to summarise the overall structures of the whole dataset and identify and inspect any outliers. Of the 439 samples, 14 outliers were identified and removed for the following reasons: visual outlier, no spectra identified.

Supervised orthogonal partial least-square (OPLS) and OPLS discriminant analysis (OPLS-DA) were utilized to reveal variations between groups with respect to continuous and categorical Y variables, respectively. Only models with

positive Q^2Y values, which are indicative of a model's predictive ability, were further tested for significance using permutation testing (100 permutations, p-value threshold: $p \leq 0.05$). To identify which metabolites significantly contributed to the model, a cut-off of $p \leq 0.05$ for the calculated p-values of each peak. The corresponding metabolites were then identified using an in-house database, Statistical Total Correlation Spectroscopy using MATLAB[107], Chenomx NMR Suite 7.0 (Chenomx, Edmonton, Canada) and the Human Metabolome Database (Edmonton, Canada).

To conduct univariate tests on the individual metabolites, peaks from the raw spectra were first integrated using trapezoidal numerical integration. Wilcoxon rank sum test was used to compare metabolite levels between healthy controls and patients at baseline. A $p \leq 0.05$ was considered significant. Pearson correlation coefficient was used to correlate metabolite and alpha diversity levels at baseline. SantaR package in R was used to compare longitudinal metabolite trajectories. Trajectories were adjusted using Benjamini-Hochberg correction. Finally, logistic regression was utilised to investigate association between clinical outcomes and metabolites as described in the biomarker discovery section.

Chapter 3- Optimisation of the 16S rRNA sequencing workflow

3.1 Introduction

Culture-independent methodologies, such as 16S ribosomal ribonucleic acid (rRNA) gene amplicon sequencing, have revolutionised our ability to gain insight into microbial composition of complex environments. As the intestinal microbiota is extremely diverse and complex, attempting to decipher its composition presents numerous difficulties.

The standard 16S rRNA workflow involves nucleic acid extraction, amplification of a chosen region of the bacterial 16S rRNA gene, followed by sequencing on a preferred platform and bioinformatics (Figure 3.1). There are nine variable regions within the ubiquitous prokaryote 16S rRNA gene, each flanked by conserved regions, which allow for the design of numerous primers[108].

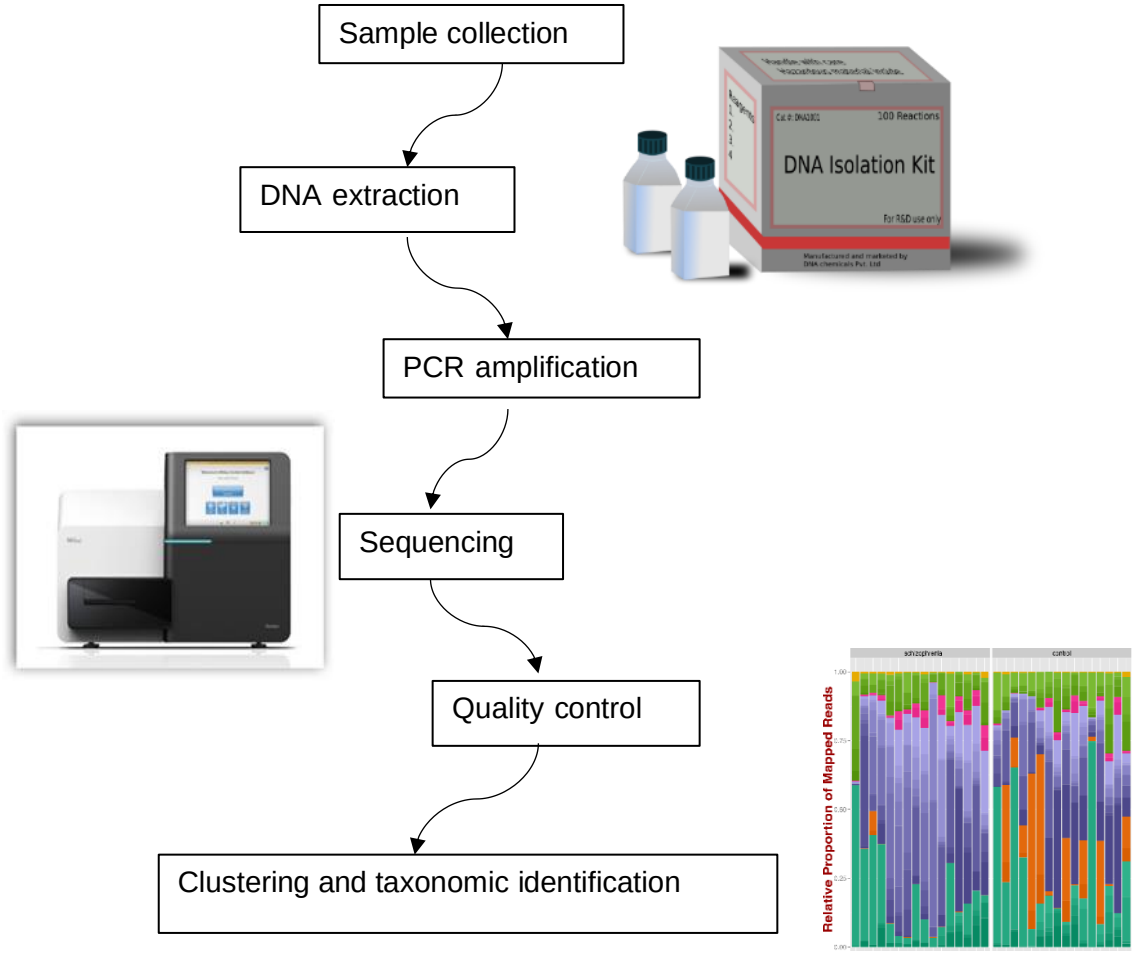


Figure 3.1 A standard 16S rRNA microbiome profiling workflow

3.1.1 DNA extraction method

Nucleic acid extraction from faecal samples is the initial step in the workflow to interrogate microbiota composition. Currently, numerous commercial DNA extraction methods exist, all column-based, notably differing in their disruption/lysis approach (mechanical, enzymatic, heat or chemical) and composition of buffers and proprietary reagents.

Several studies have explored the effects of DNA extraction methods on the composition of the microbiota[109-114]. Wu *et al* compared five extraction kits and recommend the use of bead beating in hot phenol to aid in cell lysis of the *Firmicutes* phylum[110]. Maukonen *et al* found that extraction of both *Clostridia* and *Actinobacteria* were highly dependent upon the extraction method[111]. Yuan *et al* found the addition of bead-beating and/or mutanolysin, to aid cell lysis[109]. Although extraction methods vary, the consensus is that mechanical disruption by means of bead beating is necessary to aid the lysis of Gram-positive organisms[115,116].

3.1.2 Region choice

The process of 16S rRNA amplification itself can lend to significant bias. Studies have evaluated different 16S rRNA regions (V2/V4, V2/V3 and V1/3, V4/7) [108,117-119]. Claeeson *et al* found the V3/4 and V4/5 regions would provide the highest classification accuracies[120]. Bukin *et al* on the other hand, found that V2-3 region provided the highest accuracy for lower-rank taxa[121]. In addition, longer regions can improve the accuracy of taxonomic assignment[122].

It is worth noting that within a body site, technical variation due to primer choice can outweigh biological variation between individuals and data can cluster primarily by study[122,123]. The finding that the data from the HMP, sequenced using two different primer sets, clustered separately, highlights the importance of choosing appropriate 16S rRNA primers[123]. More recently, Fuks *et al* found that combining results from several 16S rRNA regions shows improvements in identification accuracy in comparison to using a single region[124].

3.1.3 PCR steps

PCR itself, such as the cycle number can be a source of bias. For example, the addition of selected primers together with sample-specific barcodes and Illumina specific-adapters can be done in either a single-step, or a double-step PCR reaction. In our laboratory, the double-step reaction was found to be at times useful in amplifying samples with very low bacterial biomass. Sinclair *et al* have found that beta diversity was significantly different between single-step and double-step PCR reactions, although their methods differed to those employed within our laboratories[125]. Worryingly, PCR amplification has also been linked to increases in chimeras, which can lead to incorrect OTU assignments[126]. Additionally, a paper by Glassing *et al* identified bacterial contamination within commonly used PCR and extraction kits, which may be more problematic with methods including additional PCR amplification cycles[127].

Although not evaluated here, the way that the reads are processed can have a significant impact on the microbiota composition. Read trimming, alignment quality and error filtering significantly affect the results[126,128]. In order to control for the bias introduced by bioinformatics, we will use identical analysis pipelines for both sequencing runs.

3.1.4 Sample contamination

Contamination has been found to be commonplace in microbiome studies and several studies have attempted to define extraction kit-specific contamination[129,130]. Others have identified contamination in PCR reagents[127,131]. Contamination can influence the results of the study and is particularly critical for low biomass samples, such as meconium, blood and placental tissue, where it can make up a large proportion of the sample and may indicate a microbiome, where there is none. A recent study for example, indicates that the placenta has no functional microbiome, and the majority of the previously found taxa are contaminants[41]. Despite this, it is critical to investigate and appropriately handle contaminants regardless of sample origin.

3.1.5 Aims

The optimisation work presented herein aims to identify the degree to which technical variation impacts upon microbiota profiles in order to establish the most robust and suitable workflow for the forthcoming study. We investigate:

- Extraction methods
- PCR amplification steps
- 16S rRNA regions
- Mechanical homogenisation
- Contamination and dataset error

3.2 Methods

Faecal sample collection and initial processing

Faecal samples utilised were residual samples obtained from the Immunology Department at GOSH. As the samples were anonymised, no clinical action was planned based upon results and the samples were due to be discarded, no ethical approval was required. Faecal samples were kept at 4°C for a week prior to receipt. Upon receipt, samples were thoroughly mixed, aliquoted and frozen at -80°C. A visual representation of the study design is shown in Figure 3.2.

DNA extractions

Faecal samples from three individuals were extracted in triplicate using a phenol chloroform method (50mg) and three extraction kits (200mg). Commercial kits (QIAamp Fast DNA stool kit, Qiagen, Germany; FastDNA SPIN Kit for faeces, MP Biomedicals, USA; MoBio Powerfecal DNA isolation kit; MoBio Laboratories, USA) were used as per manufacturer's instructions with the following modifications. All protocols, bar phenol, included one step of mechanical homogenisation. Homogenisation for the Qiagen and MPBio protocols was done using the TissueLyser LT (Qiagen, Germany) and Lysing Matrix E beads (MPBiomedicals, USA) for 60s and 50osc, once the initial buffer is added. An extraction control consisting of buffer was extracted with each extraction batch.

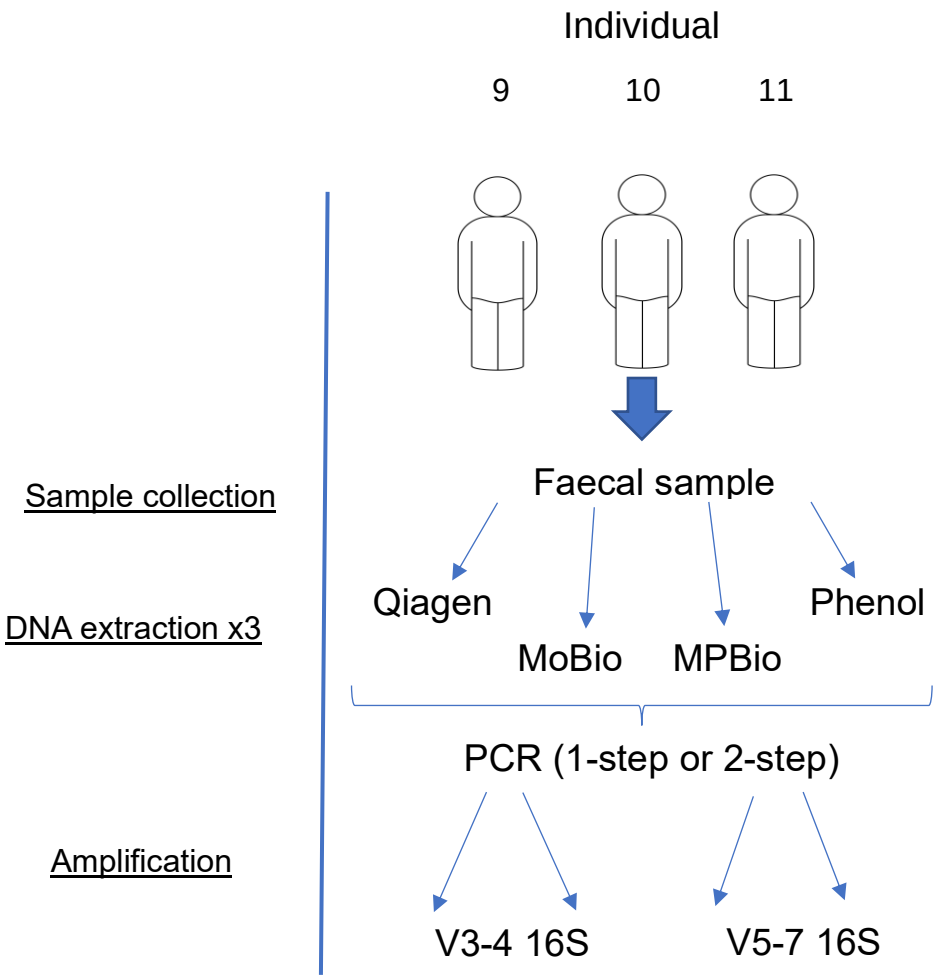


Figure 3.2: A schematic representation of the optimisation study design

Modified phenol chloroform DNA extraction

Frozen faecal aliquots were incubated in a Lysing matrix E tube (MP Biomedicals, USA) with 300µl of buffer (10mM Tris, 100mM NaCl, 50mM EDTA, pH 7.5) at 37°C for 20 minutes prior to homogenisation (60s). Following this, 40µl of lysozyme (20mg/ml) and 10µl RNase A (10mg/ml) were added and sample incubated for 30 minutes at 37°C. Next, 70µl of proteinase K and 100µl of 20% SDS were added and the sample was incubated for a further 30 minutes. 500µl of phenol: chloroform: isoamyl alcohol (25:24:1 v/v) was then added and the mixture was homogenised (60s, 50osc) and centrifuged at 4°C (10000g, 10 minutes). The upper aqueous phase was transferred into a fresh tube containing 300µl of chloroform, incubated for 5 minutes at room temperature and vortexed for 15s followed by centrifugation at 4°C (10000g, 10 minutes). The upper phase was transferred into a fresh tube containing 600µl of ice-cold ethanol (100%) and

stored at -20°C overnight. Next, the sample was centrifuged (10000g, 30 minutes, 4°C) and the resultant DNA pellet washed with 900µl of ethanol (70%) three times. The pellet was centrifuged (10000g, 15 minutes, 4°C), the supernatant discarded and the pellet air-dried. Finally, the pellet was re-suspended in 50µl of nuclease-free water and cleaned using the MPBio FastDNA SPIN Kit (starting at step 4 of the protocol). An extraction control consisting of buffer was extracted alongside each batch of extractions. DNA was quantified using the dsDNA HS Assay Kit as per manufacturer's instructions (1µl DNA + 199µl master mix, ThermoFisher Scientific, USA) using the Qubit (Version 2.0, ThermoFisher Scientific, USA).

DNA extraction for the bead-beating investigation

DNA was re-extracted from the same stool samples as above (single samples) using the Qiagen extraction kit. Homogenisation was done using the TissueLyser LT (Qiagen, Germany) with 1, 3 or 8 steps (60s, 50osc) with 60s rest on ice in between each step. The resulting DNA was amplified using a single PCR reaction and processed and sequenced as detailed below.

Single- and double-step PCR amplification

Extracted DNA (1:10 dilution) was amplified either in a single-step or in a double-step PCR reaction.

Single-step PCR: The V3-4/V5-7 region was amplified with long primers (Appendix; Table A3). The V5-7 primers were previously described by Doyle *et al*, whilst the V3-4 primers were adapted from Kozich *et al* and Joss *et al*[99,132,133]. The PCR mixture (50µl) consisted of 5µl DNA template, 1x PCR Buffer, 1x Q Solution, 2mM $MgCl_2$, 0.5µM of each primer, 200µM of each dNTP, 2.5U Taq DNA polymerase and nuclease-free water (Taq PCR Core Kit, Qiagen, Germany). Cycling conditions were as follows: 95°C for 3 minutes followed by 30 cycles of 95°C for 30s; 54°C for 30s and 72°C for 1 minute. This was followed by a final extension of 72°C for 10 minutes.

Double-step PCR: The V3-4/V5-7 region was first amplified using region-specific primers only using the same reaction components as for the single-step PCR reaction with no additional $MgCl_2$ or Q solution (341F,

CCTACGGGNGGCWGCAG/785R, GGACTACHVGGGTWTCTAAT; 785F, GGATTAGATACCCBRGTAGTC/1175R, ACGTCRTCCCCDCCTTCCTC)[132,134]. Cycling conditions were as follows: 95°C for 3 minutes followed by 15 cycles of 95°C for 30s; 55°C for 30s and 72°C for 30 seconds. This was followed by a final extension of 72°C for 10 minutes. The resulting DNA was then amplified a second time using the single-step PCR protocol described above with 15 cycles.

A mock community (MC) was amplified using the either single or double PCR steps with the V3-4 or V5-7 primer sets and sequenced (HM-783D, BEI Resources, USA). Mock community composition is detailed in the Appendix (Table A4).

DNA library construction and sequencing

PCR products generated were purified (Agencourt AMPure XP beads; 0.7x beads for V3-4 and 0.8x beads V5-7; Beckman Coulter, UK) and quantified using the Qubit dsDNA HS kit (1µl DNA + 199µl master mix), normalised to the same concentration in nuclease-free water and pooled to produce a library. To confirm fragment size, the pooled library was analysed using the Agilent Bioanalyser (2100, Agilent, USA) and quantified using the NEBNext Library Quant Kit for Illumina.

For sequencing, pooled libraries were denatured using freshly prepared sodium hydroxide (0.2M; 1:1) and inactivated using Tris-HCL (200mM; 1:1) followed by the addition of HT1 buffer to make a 4pmol library. It was then spiked with 10% PhiX (Illumina, USA) and sequenced using a V2 kit (500-cycle) on the Miseq sequencing platform. Libraries and PhiX were loaded at a concentration of 3.6pmol and 2pmol respectively. Read 1, read 2 and the index primers were spiked into the cartridge at 0.5µM.

Bioinformatics

Data was demultiplexed (barcodes removed) with version 2.1.12 of the Illumina instrument control software. Sequences obtained were analysed using Mothur (version 1.35.1)[100]. Modifications to the standard Mothur protocol were as follows: the contigs were made with standard settings, the assembled sequences were screened using the ‘screen.seqs’ command to remove ambiguous sequences,

those containing homopolymers > 8 base pairs and sequences > 460 base pairs (> 430 in V5-7) were removed. Unique reads were aligned using a region-specific Silva bacterial database (release 128). Any sequences outside the expected alignment coordinates were removed. The correctly aligned sequences were subsequently filtered ('filter.seqs') with 'vertical=T' and 'trump=.'. The filtered sequences were de-noised by allowing six mismatches in the 'pre.clustering' step and chimeras were removed using vsearch with the dereplicate option set to 'true'. The chimera-free sequences were classified using the Mothur-formatted Bayesian RPD database and a cut-off value of 80. Mitochondria, archaea, chloroplast, eukaryota and unknown sequences were removed. A distance file was generated with a 0.03 cut-off and the resulting matrix was split and clustered with large option set to 'true'. A shared file was generated and OTUs classified. The resulting shared taxonomy and metadata files were combined into a biom file.

The V3-4 and V5-7 datasets were subsampled to an equal depth of 8309 and 10983 respectively for the comparison of alpha diversity and richness between PCR steps and the kits. For comparisons of the 16S regions and the bead-beating comparisons, the datasets were subsampled to an even depth of 8309 and 48477 reads respectively. Shannon effective entropy was calculated using Rhea in R[101]. Further data analysis was conducted using vegan and phyloseq packages in R (version 3.5.1). Indicator species analysis was done using the 'indicspecies' package in R with 199 permutations.

Statistical analysis

Comparisons between alpha diversities were performed using the Kruskal-Wallis test with Benjamini-Hochberg correction.

Dataset-specific error and contamination assessment

Assessment of dataset-specific error and contamination was done on mock communities and NEC sequenced alongside patient samples collected for the main study, after deciding on the optimal 16S rRNA workflow. The mock community used differed from those used in the initial optimisation (Appendix; Table A2). Primers are detailed in chapter 2. NEC consisting of 200µl ATL buffer,

were extracted with each extraction batch and sequenced alongside patient samples.

3.3 Results

Several studies to date have found that the choice of methodology preceding next-generation sequencing of gut microbiota can significantly affect the outcome[113,122,134]. We therefore evaluated several steps of the 16S rRNA workflow including four DNA extraction methods, sample amplification using single and double PCR steps and two primer sets targeting different areas of the 16S rRNA gene and sample homogenisation by a way of bead-beating.

3.3.1 Extraction kits

In the present study, three extraction kits plus a phenol-based method were tested. Notable differences between the methods are highlighted in Table 3.1

Table 3.1 Characteristics of the DNA extraction methods investigated.

DNA extraction method (abbreviation)	Recommended starting amount (mg)	Cell lysis method	Price £/sample	Time taken
QIAamp fast DNA stool kit (Qiagen)	≤220	BB,CLB,T	2.6	1.15hr
FastDNA SPIN Kit for faeces (MPBio)	≤500	BB,CLB	6.6	2hr
MoBio Powerfecal DNA isolation kit (MoBio)	≤250	BB,CLB,T	3.9	2hr
Phenol:Chloroform extraction (phenol)	NA	BB,T,SDS,L	6.9	24hr*

Completion time was calculated from the start of faecal sample processing to the elution of DNA *Inclusive of the MPBio clean-up. BB= bead beating; CLB= cell lysis buffer; T= temperature; SDS= sodium dodecyl sulphate, L= lysozyme.

DNA extraction methods do not impact DNA yield

All methods yielded DNA for 16S rRNA sequencing (1.86-125ng/µl). Overall, the yield appeared uniform between the methods (Figure 3.3), although the phenol method resulted in significantly lower yields than the MPBio method, which had variable yields ($p \leq 0.05$).

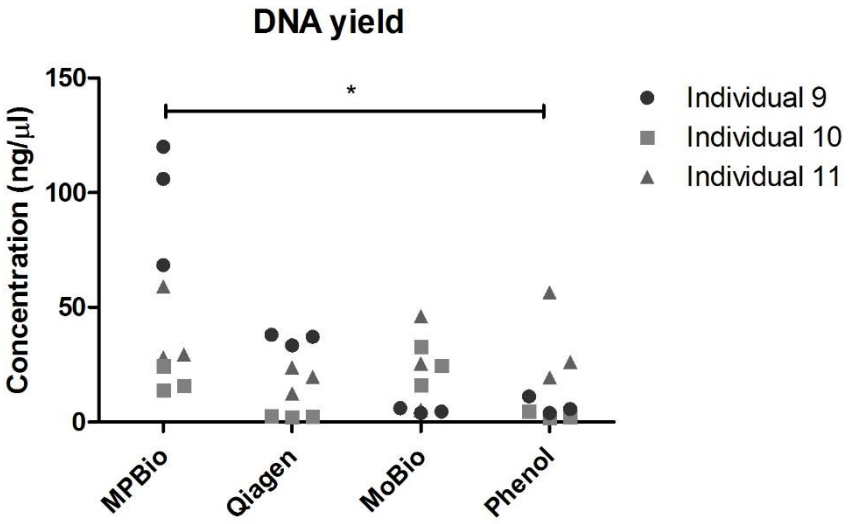


Figure 3.3 The mean yield of DNA (ng/µl) extracted by the four methods. Each extraction was performed in triplicate Statistical significance was determined using a Kruskal-Wallis with a Dunn's post hoc test. (*p≤0.05)

Variation in alpha diversity

We then attempted to ascertain whether DNA extraction methods influenced the composition of the gut microbiota. Observed richness (number of OTUs observed), was lower in samples extracted by the phenol method in the V3-4 region but higher than Qiagen and MoBio methods within the V5-7 region (Figure 3.4 A/B).

Effective Shannon entropy (alpha diversity) was higher in samples extracted by the phenol method in comparison to Qiagen within the V3-4 region and between phenol and Qiagen/Mobio methods within the V5-7 region (Figure 3.4 C/D). It was interesting to note greater variability in alpha diversity within the MPBio extraction method compared to the other methods.

DNA extraction methods contribute to the variation observed within the study

We were also interested in whether DNA extraction methods would affect beta diversity. PCA plots of both primer sets revealed that the samples primarily clustered by an individual, rather than by the extraction method (Figure 3.5 A/B). Anosim implies that there are significant differences between the extraction

methods (p=0.001), however that the variable has a small effect in both models (R=0.218; R=0.207)

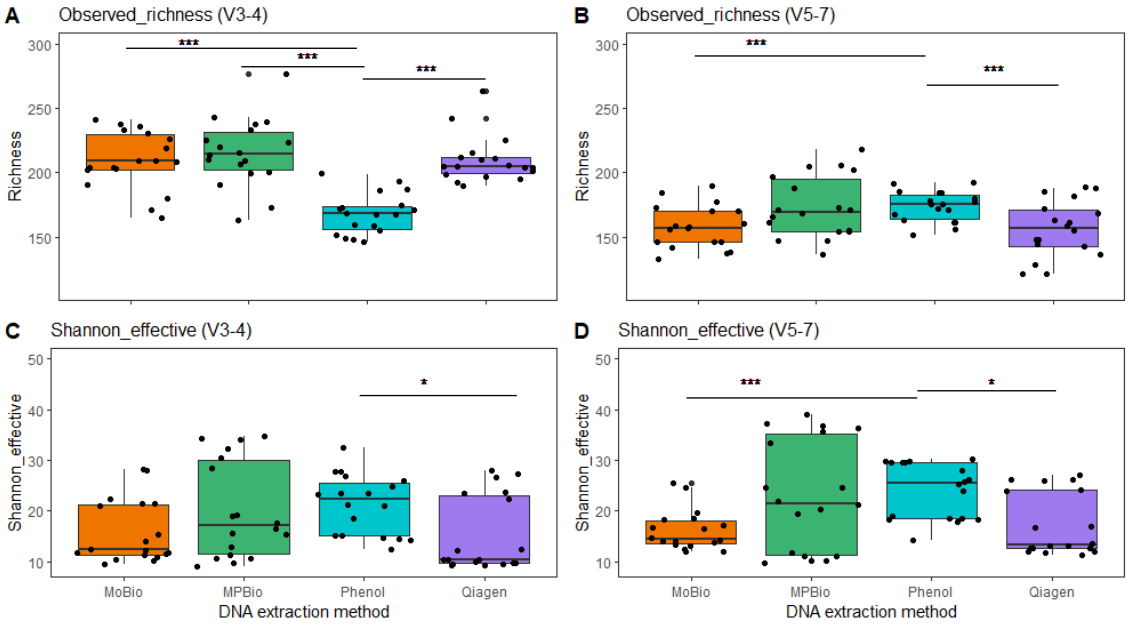


Figure 3.4 Comparison of the observed richness and effective Shannon entropy for **A)** V3-4 and **B)** V5-7 region. Boxplots detail the median and the interquartile range. Statistical significance was determined using Kruskal-Wallis with a Benjamini-Hochberg correction (*p≤0.05; **p≤0.01; ***p≤0.001).

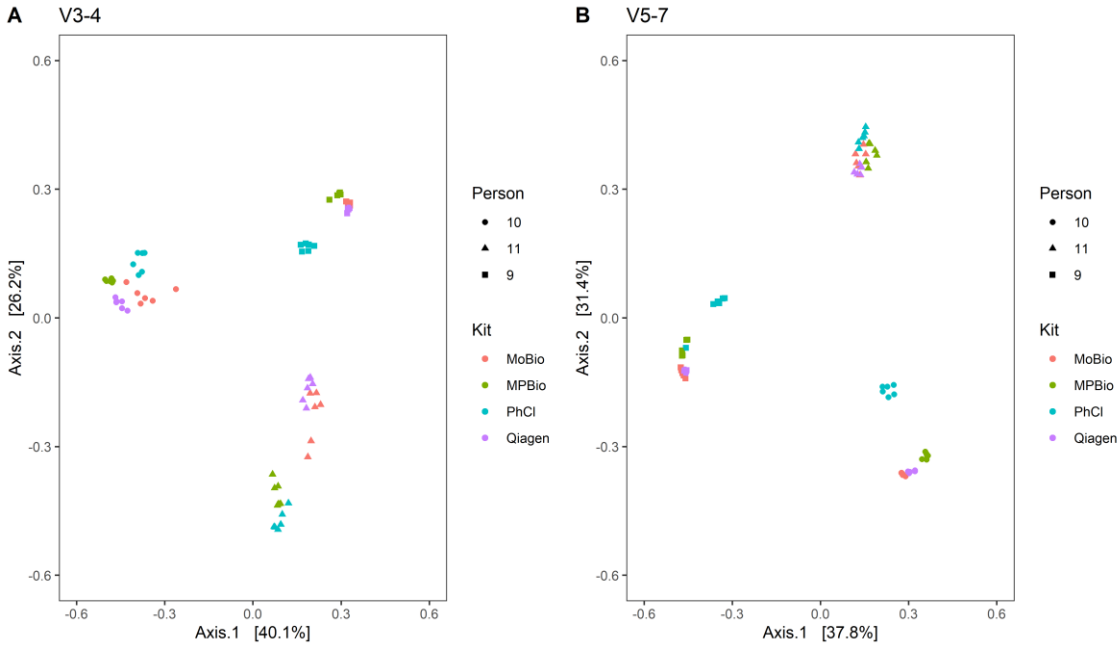


Figure 3.5 PCA based on Bray-Curtis dissimilarities between samples extracted by four DNA extraction methods for the **A)** V3-4 and **B)** V5-7 primer sets.

Effects of DNA extraction kits on the taxonomic composition of the faecal microbiota

We then visualised taxonomic composition of samples extracted by the four extraction methods. Figure 3.6 details taxonomic composition at the phylum level. It is evident that there are marked differences between samples extracted by the phenol method *versus* the other methods. Phenol method yielded a greater proportion of *Firmicutes* and *Actinobacteria* in comparison to the other methods, whereas the MPBio method yielded a greater proportion of *Proteobacteria*. Qiagen and MoBio methods appear similar at the phylum level.

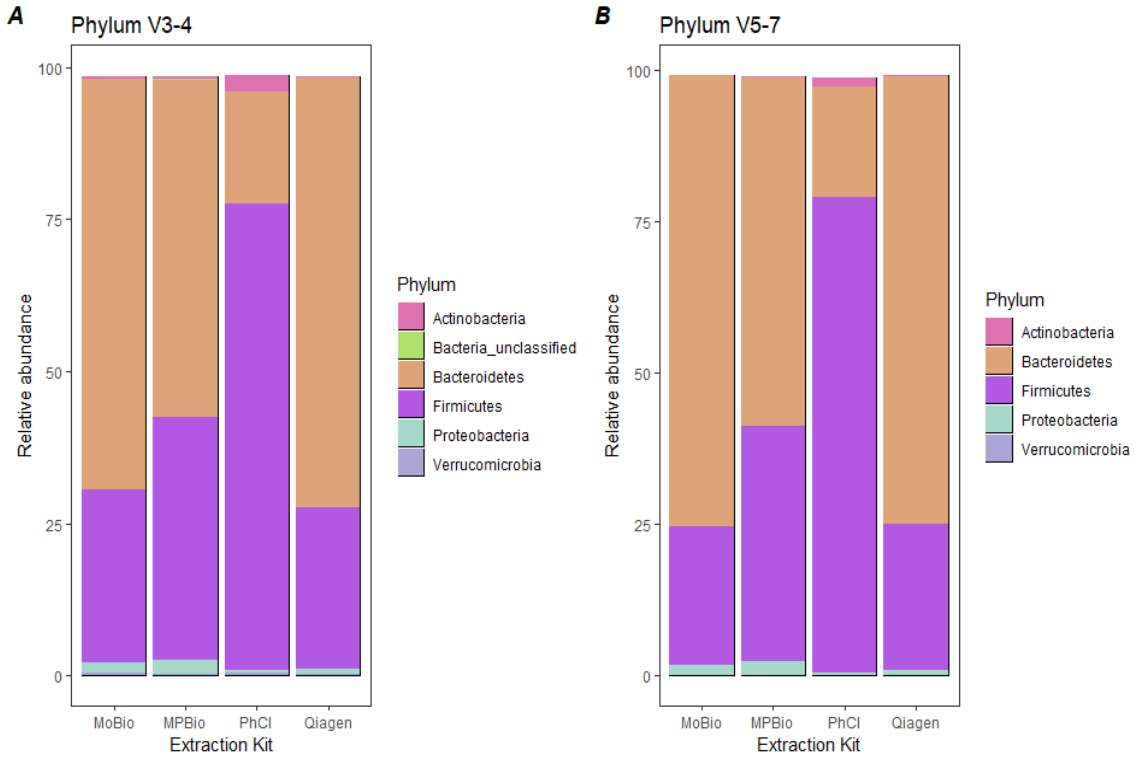


Figure 3.6 Taxonomic composition at phylum level for samples extracted by the four DNA extraction methods for the **A)** V3-4 and **B)** V5-7 primer sets.

To determine which taxa were associated with each of the extraction methods we used Indicator Species Analysis (ISA). We found a total of 38 and 42 taxa that were associated with the phenol extraction method within the V3-4 and the V5-7 regions respectively (Appendix; Tables A5 and A6). Several taxa were indicative of the phenol method such as the Gram-positive genera *Blautia*, *Streptococcus, Clostridium* and *Bifidobacteriaceae* among others. These differences are in concordance with the phyla in Figure 3.6. There were few taxa consistently predictive of the other extraction methods, except for *Enterococcus and Gemella* for the MPBio and the MoBio extraction methods respectively.

3.3.2 PCR amplification

PCR steps do not influence microbiota composition

Additionally, we sought to ascertain whether the use of single *versus* double PCR steps would alter the microbiota composition. There were no differences in observed richness or effective Shannon entropy between samples sequenced in a single or double PCR steps (Figure 3.7).

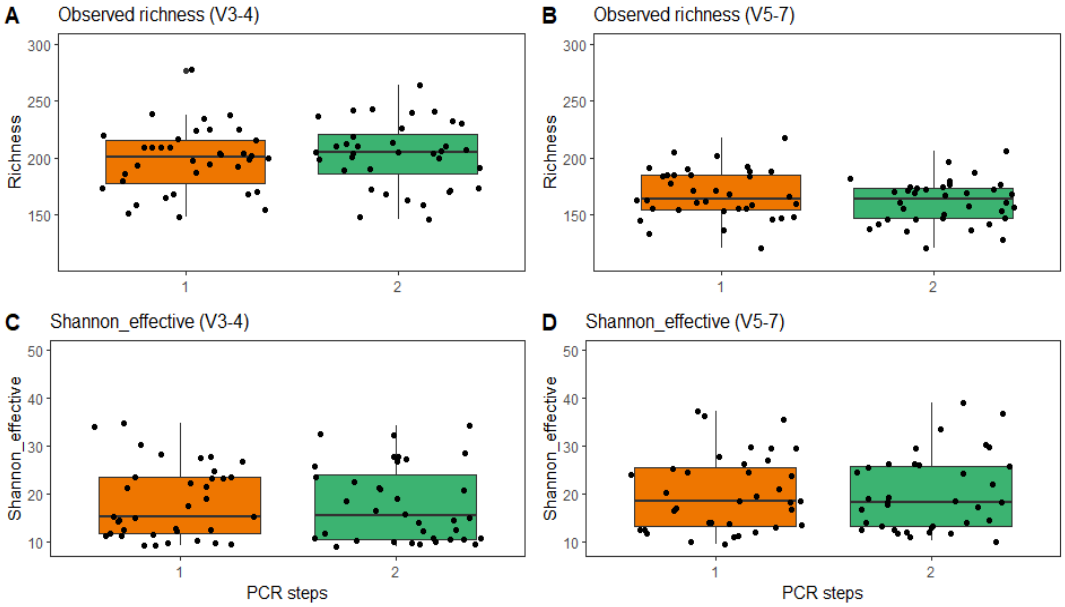


Figure 3.7 Comparison of the observed richness and effective Shannon entropy for **A)** V3-4 and **B)** V5-7 primer sets. Boxplots detail the median and the interquartile range.

Next, we compared beta diversity of samples amplified using single or double PCR steps within both primer sets. There was no obvious clustering in relation to the PCR steps, indicating that they do not contribute to the model (Figure 3.8), which is further confirmed by non-significant Anosim results (p=0.94; p=0.92).

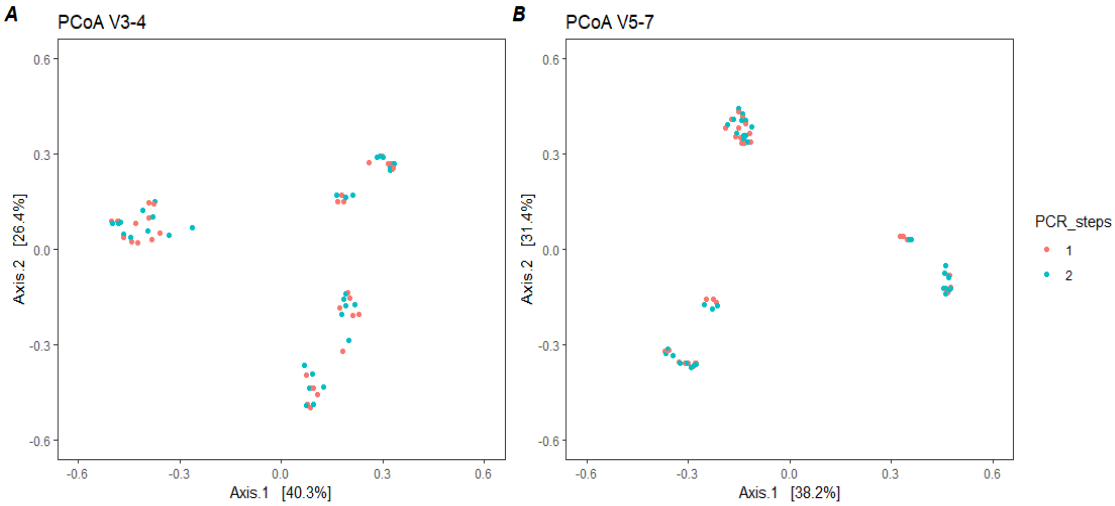


Figure 3.8 PCA comparison based on Bray-Curtis dissimilarities between samples amplified by single (1) or double (2) PCR steps for both primer sets.

3.3.3 Primer sets

Primer choice affects microbiota composition

We were also interested in whether targeting different 16S rRNA regions would result in variation in the microbiota composition. In total, post quality checking, approximately 4.4 million sequences from the V3-4 dataset and approximately 3.7 million sequences from the V5-7 dataset were obtained. We found that 0.7% and 1% of sequences flagged up as chimeric for the V3-4 and the V5-7 regions respectively.

Within the V3-4 region, 90% of sequences were classified at the genus level, whereas only around 75% of sequences were classified using the V5-V7 primers (Figure 3.9).

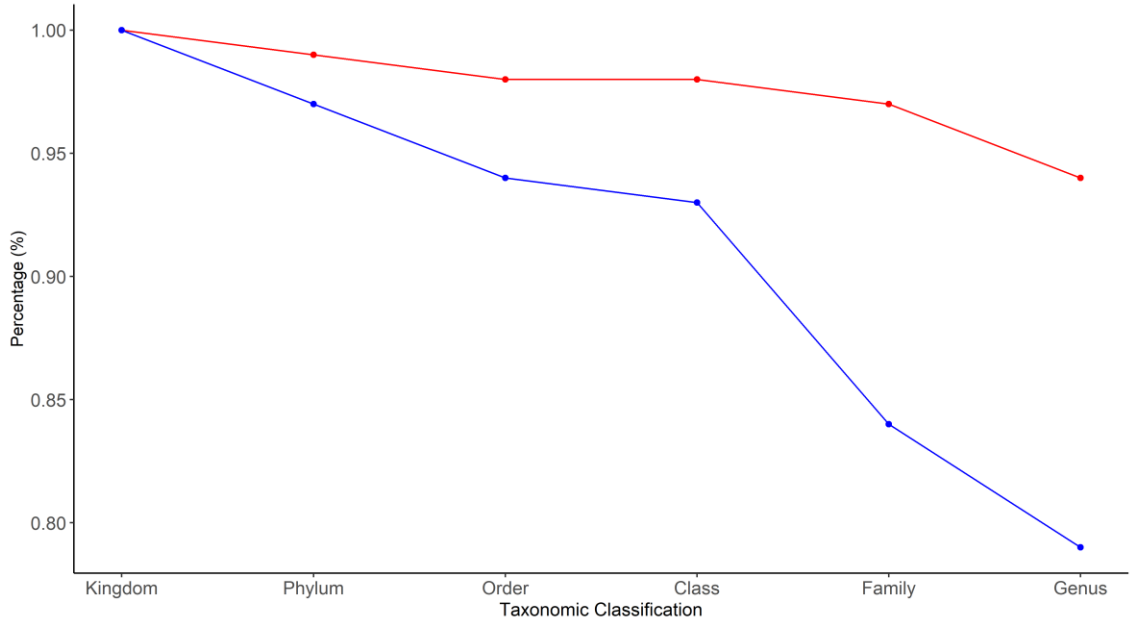


Figure 3.9 A mean proportion of the 16S sequences successfully classified and assigned at different taxonomic levels. Red- V3-4; Blue- V5-7 16S rRNA regions

To evaluate PCR bias and sequencing error in further detail, we additionally sequenced a mock community (Appendix; Table A4).

The V3-4 dataset identified 22 and 25 OTUs for the single and double PCR steps respectively, whereas the V5-7 dataset identified 103 and 53 OTUs (Appendix; Table A7). The V3-4 dataset was unable to pick up the *Actinomyces* and *Propionibacterium* taxa and appeared to overinflate the OTU number for the *Escherichia* and the *Staphylococcus* genus (Appendix; A7). In terms of abundance, both the *Streptococcus* and the *Escherichia* genera were overinflated (Figure 3.10). Despite this, the primer set performed relatively well as it identified 85% and 90% of taxa by single and double PCR steps respectively.

In comparison, the V5-7 dataset performed considerably worse. It was unable to pick up several taxa, namely *Deinococcus* and *Enterococcus* by single-step PCR and *Helicobacter* and *Propionibacterium* by double-step PCR. It overinflated the OTU number for the *Escherichia* and the *Staphylococcus* genus, as well as the *Bacteroides* genus. In terms of abundance, both the *Streptococcus* and the *Staphylococcus* genera were greatly overinflated (Figure 3.10). More worryingly, the V5-7 dataset identified several taxa not found in the mock community sample including *Alistipes*, *Enhydrobacter* and *Corynebacterium*.

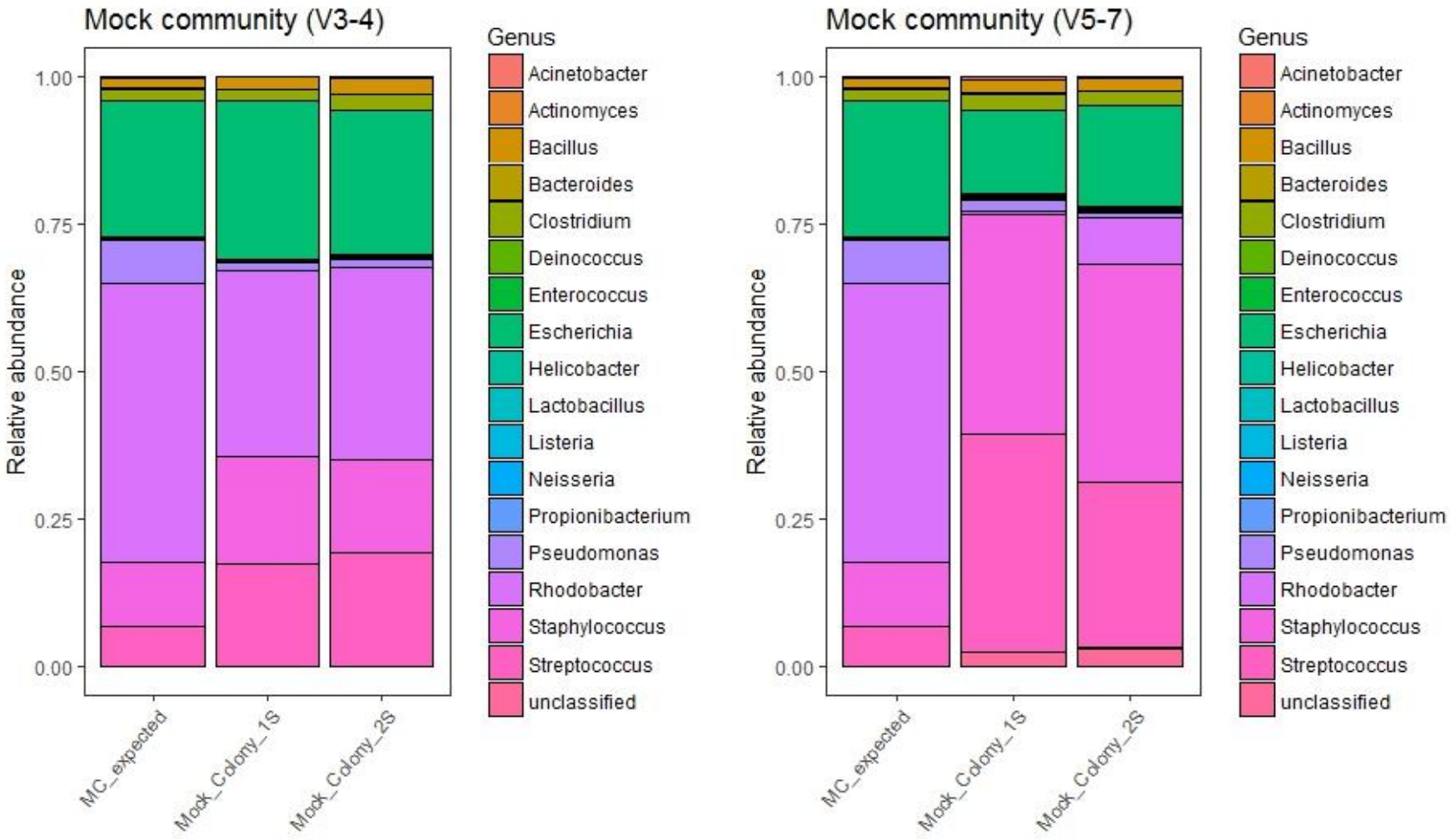


Figure 3.10 Taxonomic composition (genus level) of the mock community for the V3-4 and the V5-7 primer sets amplified by single (1S) or double (2S) PCR steps. Unclassified category is composed of taxa unclassified at the genus level as well as taxa not found within the mock community sample. MC_expected sample details the expected composition of the mock community sample.

3.3.4 Sample homogenisation

Having chosen the most suitable workflow for the 16S rRNA sequencing, including the Qiagen extraction kit, a single PCR step and the V3-4 16S rRNA region, we finally sought to investigate the optimal number of bead-beating steps for the workflow.

Overall, we extracted 15.7ng/µl, 21.5ng/µl, 27ng/µl and 28ng/µl of DNA from samples homogenised with no, 1, 3 and 8 bead-beating steps respectively, although the yield differences were not significantly different.

There was an increase in *Actinobacteria* and *Firmicutes*, both Gram-positive phyla, with increasing bead-beating steps (Figure 3.11 A). Despite this, when we interrogated taxonomic profiles for each individual (Figure 3.11 B), the patterns appeared less clear. Whilst the relative abundance of *Actinobacteria* increased consistently with increasing bead-beating steps, the relative abundance of *Firmicutes* and *Proteobacteria* was more variable. Thus, increasing bead-beating steps appears to have a different effect within each individual (Figure 3.11 B).

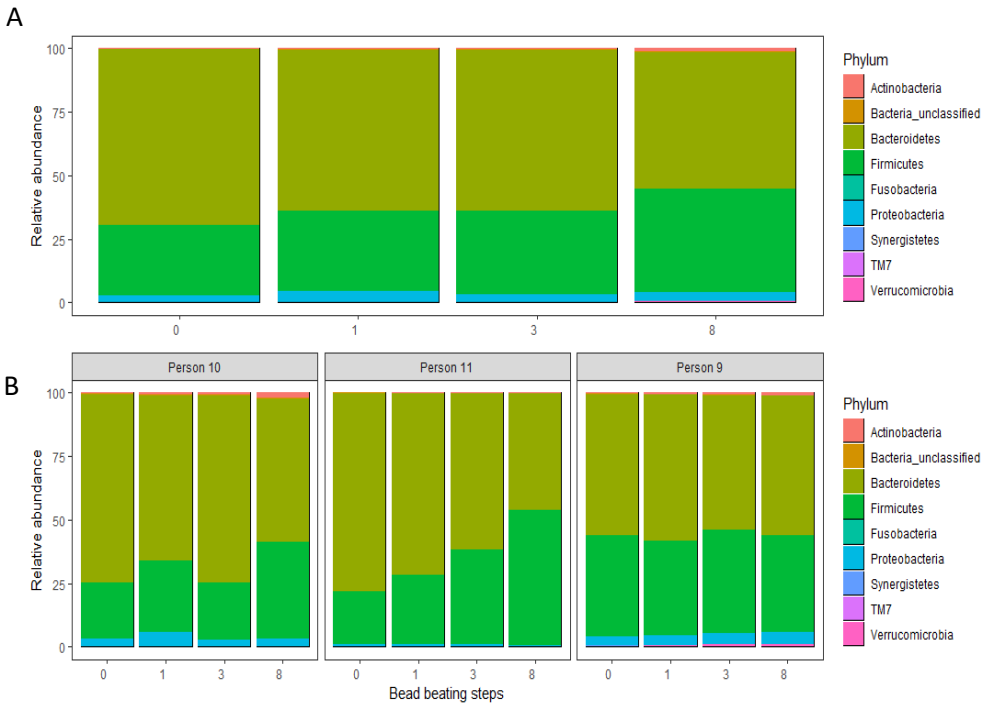


Figure 3.11 Taxonomic composition at the phylum level for the samples homogenised using a varying number of bead-beating steps **A)** grouped by bead beating steps and **B)** grouped for each individual

Additionally, we investigated alpha and beta diversities between the samples. Although alpha diversity is broadly similar with 0, 1 and 3 bead-beating steps and increases with 8 bead-beating steps, these changes were not significant (Figure 3.12 A). Similarly, beta diversity (Figure 3.12 B) reflects the high inter-personal diversity and a lesser effect of homogenisation (Anosim; R=0.03).

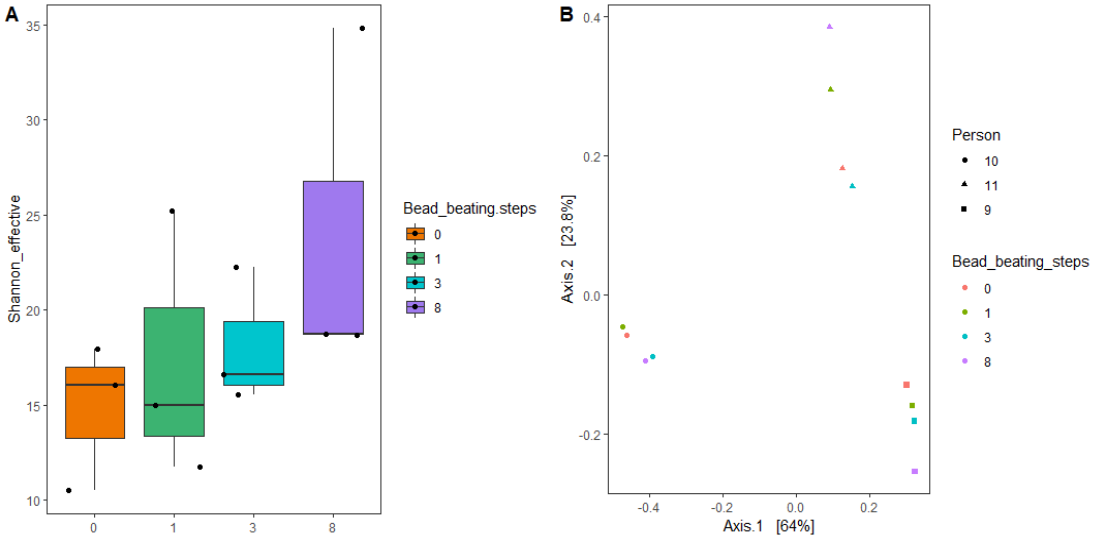


Figure 3.12 A) Alpha diversity (Shannon effective) and **B)** PCA based on Bray-Curtis dissimilarities between samples homogenised using a number of bead-beating steps.

3.3.5 Dataset-specific error

Having decided on using 8 bead-beating steps, as it was likely to increase Gram-positive yield from the more complex samples without undue effect on the less complex samples, all study samples were sequenced.

Mock communities (Appendix; Table A2) were sequenced alongside the samples in this study. Taxonomically, the mock communities looked uniform across the sequencing runs (Figure 3.13). Several taxa, namely *Escherichia* and *Streptococcus*, as well as *Bacillus*, were overinflated, whereas *Lactobacillus* was consistently underrepresented.

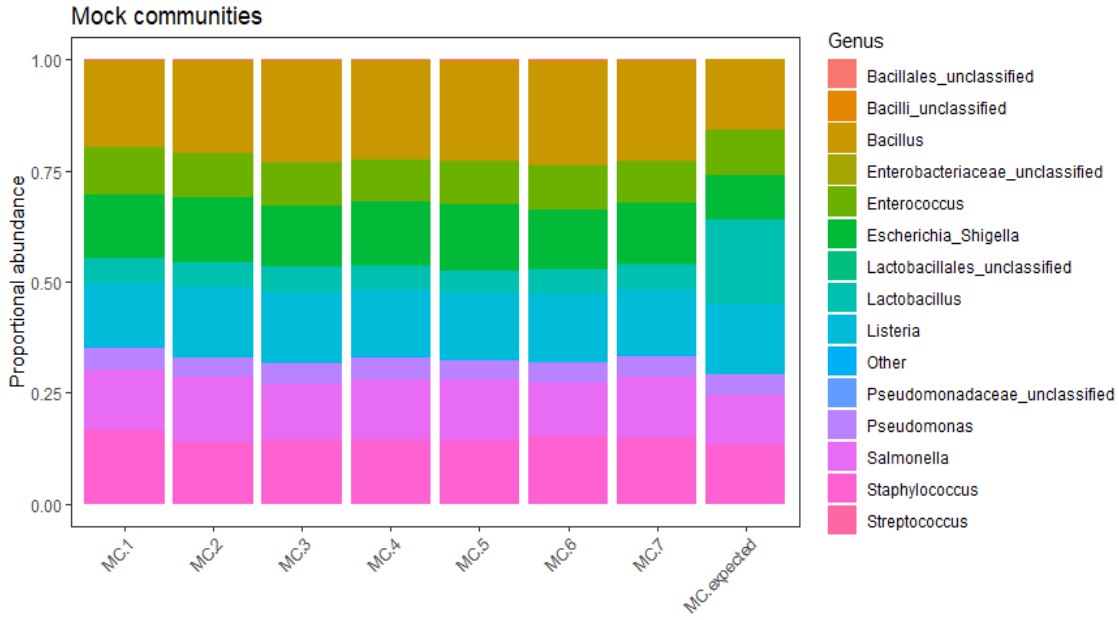


Figure 3.13 Taxonomic composition at the genus level for the mock communities sequenced with each sequencing run. MC.expected details the expected composition of the mock community

A representative table of composition for one of the communities is shown (Table 3.2). The number of OTUs in the mock community with no filtering (867) was higher than expected. A large proportion of these are OTUs have not been classified enough i.e. to genus level and 41 genera come from taxa that should not be in the mock community.

The most abundant taxa that was not expected to appear, *Klebsiella*, had a relative abundance of 2.49x 10^{-4}, whilst the second most abundant, *Citrobacter*, had the relative abundance of 1.19x 10^{-4}. Filtering at the relative abundance of the first erroneous OTU eliminated all taxa that were not fully classified as well as all of the erroneous taxa, which reduced the overall OTU number to what was expected. Filtering at the 2nd most abundant erroneous OTU left a single false OTU, whereas filtering at the 3rd left two OTUs in (Table 3.2). All of the mock communities were therefore individually assessed, and the respective dataset filtered at an abundance level that eliminates as many erroneous OTUs as possible, without affecting the expected composition. Filtering levels ranged from 9.5x10^{-4} to 7.1x10^{-6} (Appendix; Table A8).

Table 3.2 Composition of an example mock community prior to- and as a result of varying levels of filtering.

	Not filtered	Most abundant erroneous OTU ($2.49x10^{-4}$)	2nd most abundant erroneous OTU ($1.19x10^{-4}$)	3rd most abundant erroneous OTU ($7.11x10^{-5}$)	4th most abundant erroneous OTU ($5.93x10^{-5}$)
Overall number of OTUs	867	8	9	13	16
Pseudomonas aeruginosa	2	1	1	1	1
Escherichia coli	115	1	1	1	1
Salmonella enterica	3	1	1	1	1
Lactobacillus fermentum	23	1	1	2	3
Staphylococcus aureus	55	1	1	1	1
Enterococcus faecalis	13	1	1	2	2
Listeria monocytogenes	80	1	1	1	1
Bacillus subtilis	6	1	1	1	1
Other- not expected	41	0	1	2	3
Other (not classified enough)	529	0	0	1	2

3.3.6 Contamination in the sequenced negative controls

Contamination is a concern in microbiota studies[127,129,130,135,136]. We assessed the extraction controls sequenced alongside the study samples, in order to decide on the most appropriate approach for controlling contamination.

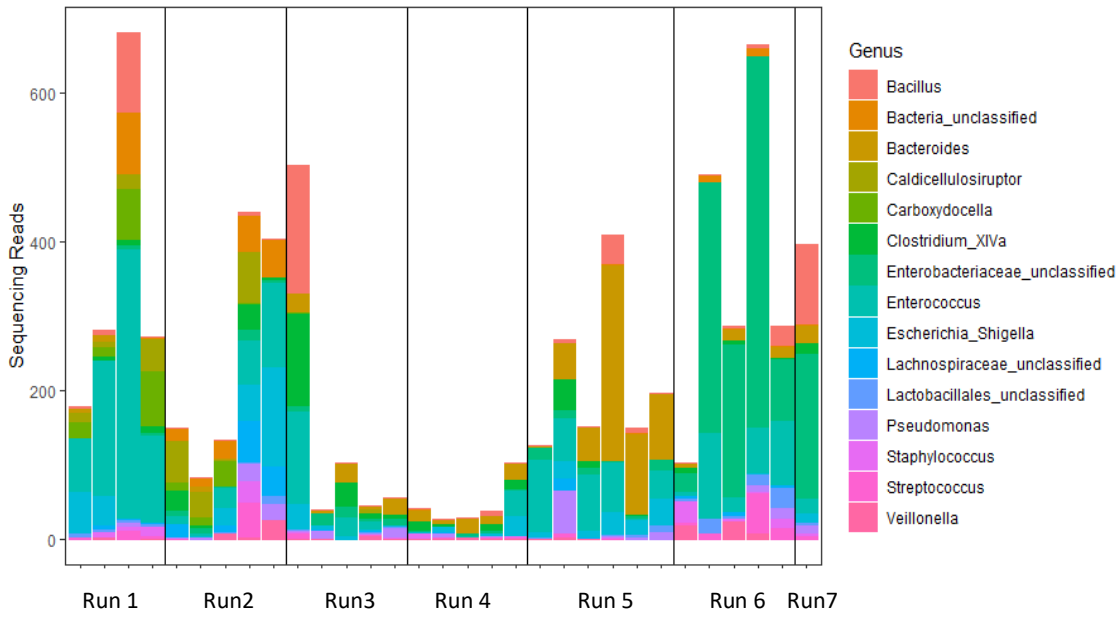


Figure 3.14 Taxonomic composition at the genus level of the NEC sequenced alongside the samples of the study. OTUs not observed at least twice in at least 1/5 of the samples were removed.

NEC showed raw reads ranging from 41 to 1978 and the taxonomic composition was diverse (Figure 3.14). Although contamination varied with each extraction, there were some general trends such as *Veillonella* and *Bacteroides* being primarily present in selected subsets of NEC, which is suggestive of reagent contamination.

We then compared all the NEC to find common contaminants. Twelve taxa were found in all NEC and seven taxa in 6 out of 7 sequencing runs (Appendix; Table A9). Interestingly, the most common contaminants such as *Enterococcus* and *Bacteroides* are known commensals in the human gut. Although less frequent, contaminants such as *Bradyrhizobium* and *Pelomonas*, reported by others were also observed in certain NEC.

3.3.7 Contamination removal strategy

As sequenced NEC seemed to have both external contamination and potential contamination from sample-to-control cross-over, we decided to inspect each OTU more closely. Certain OTUs were present in lesser quantities in the controls *versus* samples (Figure 3.15 A), whereas others were more abundant (Figure 3.15 B). We concluded that higher abundance in an NEC than a sample was indicative of a 'contaminant' and the OTU was subsequently removed, whereas lower abundance was indicative of 'unintended cross-over' and the OTU was left in the dataset.

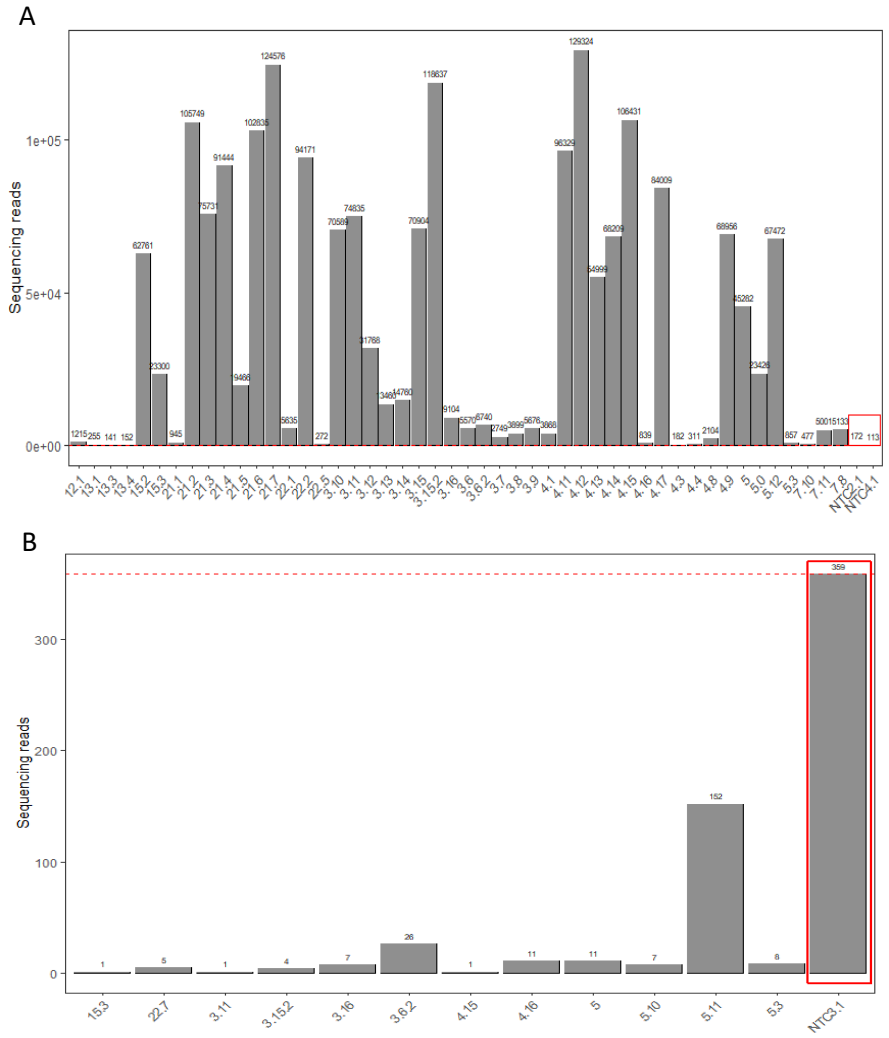


Figure 3.15 Examples of OTUs in samples that were classed as **A)** a non-contaminant, which was kept in the dataset and **B)** a contaminant which was subsequently removed from the dataset

A filtering strategy for the datasets was then decided upon (Figure 3.16). Each sequencing run would require initial filtering at a specific level as indicated by a unique mock community sequenced with each sequencing run. Post manual inspection, OTUs present at higher abundance in NEC alongside any taxa that remained unclassified at the phylum level and any OTUs that fit the criteria of a contaminant OTU would be removed. Taxa that are a published contaminant and not of human origin would be removed from the dataset in their entirety. Table A10 (Appendix) details such taxa.

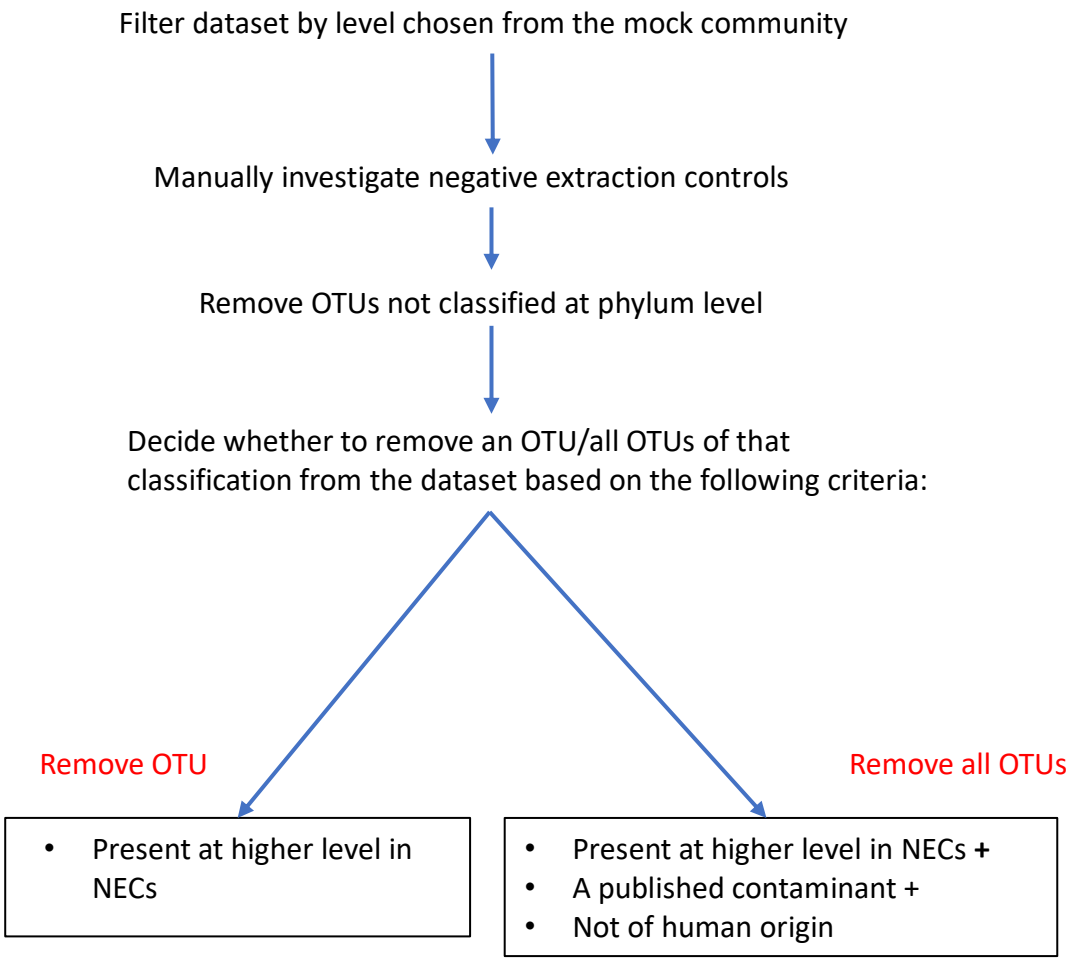


Figure 3.16 Dataset processing strategy

3.4 Discussion

In order to minimise technical variation in microbiota studies it is important to understand the sources of and the degree of technical variation introduced by the steps within the processing workflow. We thus evaluated several components of the workflow including DNA extraction methods, sample homogenisation, single or double PCR steps and two primer sets. Additionally, we investigated the mock communities and negative extraction controls for dataset-specific errors and contamination.

3.4.1 DNA extraction method

Here we evaluated and compared four extraction methods. Table 3.3 details the evaluation criteria used in selecting an optimal kit.

Table 3.3 DNA extraction method evaluation criteria

Extraction method	Reproducibility	Time needed	Frequently used in literature	Discontinued
Qiagen	Yes	1.15hr	Yes	No
MPBio	No	2hr	Yes	Yes
MoBio	Yes	2hr	Yes	Yes
Phenol	Yes	24hr	No	No

Generally, the three commercial extraction methods were comparable to each other. Of note is that whilst the Qiagen, MPBio and phenol methods utilise the same method for bead-beating in tubes containing glass, ceramic and silica spheres, the MoBio method uses garnet beads. Phenol on the other hand, utilises additional cell lysis methods including lysozyme and an additional bead-beating step.

DNA yield was comparable between the methods (Figure 3.3), although the MPBio method shows higher intrapersonal variation than the other methods, which is undesirable in a longitudinal study. As a result, we decided not to use the MPBio kit.

In concordance with another study, we found that samples primarily clustered by an individual and thus inter-individual differences are the primary source of

variation observed (Figure 3.5)[113]. Despite this, the phenol method extracted a greater proportion of Gram-positive bacteria, which could be due to its use of use of phenol, lysozyme and an additional bead-beating step [109,112]. This hypothesis is also supported by our findings that the phenol method extracted a Gram-positive bacterium spiked-into stool samples (*S. pneumoniae*) more efficiently than the other methods (data not shown). Furthermore, the three commercial methods extracted lower proportional abundance of *Firmicutes* meaning that they are less optimal for lysing Gram-positive organisms.

Despite this, the time taken to extract DNA using this method as well as the use of hazardous chemicals makes it less practical than the commercial methods. Furthermore, phenol method is less commonly used in the literature, making it more difficult to compare findings between studies. Finally, higher hands-on time needed for the extraction and human error may make this extraction method less reproducible. As a result, we decided against using the phenol method.

This left a choice between the Qiagen and the MoBio kits and as the MoBio kit was recently discontinued we decided to use the Qiagen kit. It shows yield reproducibility between the sample aliquots, is easy to use, is frequently used in the literature and extracts Gram-positive organisms with the addition of a single homogenisation step.

Despite these findings, without knowing the true composition of the sample it remains difficult to ascertain which method yields results closer to the truth. Consequently, these results highlight the importance of using the same DNA extraction method throughout microbiota profiling studies, as well as urges caution when comparing work done using dissimilar methods. It may be useful to extract bacterial communities with a defined composition and to spike-in a known concentration of bacteria post-extraction, which can aid quantification in future work[137].

3.4.3 16S rRNA primer sets

Table 3.4 details the evaluation criteria used in selecting a primer set. In terms of quality control, the sequencing quality of the V5-7 dataset was inferior to that of the V3-4 dataset as there were more chimeras, which may be indicative of PCR-introduced error.

Table 3.4 16S rRNA region evaluation criteria

16S rRNA region	Chimeras	% Genus identified	No. taxa not identified	Additional taxa identified	Abundance over inflation
V3-4	0.7%	>95	2-3	0	Some
V5-7	1%	<80	2-3	4-6	Significant

The V5-7 dataset was also significantly worse at taxonomic classification as fewer than 80% of sequences were classified to the genus level (Figure 3.9). These differences may be due to the length of the respective regions as a correlation between the length of the amplified region and classification accuracy has been previously observed[120]. Martinez-Porchas *et al* found that the shorter the length of the amplified region, the greater the probability that a higher proportion of the classifications may be erroneous, due to the lower specificity and sensitivity in the taxonomic classification of the shorter regions[138].

Both primer sets did not identify certain genera including *Actinomyces* and *Propionibacterium* by the V3-4 primer set and a variety of taxa by V5-7 primer set (Appendix; Tables A7). As the differences within the V5-7 dataset were not consistent between the PCR steps, it is likely that this is due to PCR bias. It is unclear however, why V3-4 primers consistently did not detect these specific genera. Nevertheless, this was unlikely to be due to low copy numbers, as genera with similar copy numbers were detected by the primer set. This could either be due to PCR bias for the specific taxa or the short overlap within the V3-4 16S rRNA region. This could however be addressed by utilising a cartridge with a longer read length in the future.

More worryingly, the V5-7 primer set detected several taxa which were not present in the original mock community. This may be due to the shorter length of the region and the inability to differentiate between taxa, resulting in bioinformatics error, which makes the region less suitable for microbiota profiling, as well as due to contamination. Additionally, the V5-7 dataset exhibited an over-inflation of OTUs in the mock community, potentially due to both PCR bias and bioinformatics error.

The data presented here was not initially filtered for low abundance OTUs, however the V5-7 dataset may benefit from this strategy, should it be used in the future. Based on the overall findings, we believe that it is advisable to include a mock community in every 16S rRNA run to assess the error introduced by PCR and sequencing. This would also allow for the decision regarding filtering to be made. Whilst not in the scope of this study, it is also important to remember that the sequencing of short 16S rRNA gene fragments is not entirely representative of the true diversity of the full 16S rRNA gene[117].

Thus, as the V3-4 primer set gave the most reliable representation of the mock community samples, fewer spurious OTUs and led to a more in-depth taxonomy, we found it preferable to the V5-7 primer set.

3.4.2 PCR amplification steps

PCR amplification is also known to be a source of bias and has thus prompted concerns over reproducibility and accuracy of microbiome studies. Previous studies have identified PCR-related errors and bias relating to amplification parameters, amplification of contamination, chimera introduction and primer mismatches/degeneracies[125,139-143]. We were interested in whether our single and double PCR step amplification protocols differ from each other.

We did not observe any differences in observed richness or alpha diversity between the two PCR methods (Figure 3.7). Similarly, there was no difference between the two PCR methods in terms of beta diversity, as samples preferentially clustered by individual (Figure 3.8). It is therefore reassuring that there does not appear to be a difference between the two protocols and that neither contribute towards variation observed in the β-diversity model (Figure 3.8).

Despite this, a closer look at the mock communities we sequenced reveals some differences between the two methods (Figure 3.10). These appear to be most pronounced with the V5-7 primer set, where the double-step PCR amplified a greater proportion of the *Escherichia* and *Staphylococcus* genera, yet fewer *Streptococcus* in comparison to the single-step PCR. Within the V3-4 primer set the PCR amplifications appeared similar. It has been previously found that PCR

primers may preferentially amplify certain taxa over others. Suzuki *et al* found that despite differences in initial sample DNA concentration, the primers were biased towards a 1:1 ratio due to annealing bias, whereas others have found that this may be due to homoduplex formation and gene G+C content[144,145]. This could explain why certain taxa were overinflated and others underinflated in comparison to the expected mock sample composition.

This does not explain the differences observed between the two PCR amplification methods with the V5-7 primer set, however apparent differences could be a result of differing amplification efficiencies. It may be hypothesised that amplification of shorter fragments (V5-7) would result in greater amplification efficiency, and thus increased variability[144]. As the single-step and double-step PCR within the V3-4 region were similar, we chose to use the single-step protocol as it requires less hands-on time and a single-step PCR is less likely to introduce variability.

Although PCR steps did not introduce significant variation into this study, the subtle changes in abundance may be undesirable when intra-individual variation is of interest, for example in longitudinal studies. Additionally, although less of a problem in this study, PCR amplification may introduce greater changes within samples with low bacterial biomass[146]. Unfortunately, there is no apparent solution to PCR bias. Despite this, using standardised protocols throughout a study would ensure that the samples are exposed to similar bias.

3.4.4 Sample homogenisation

As previously reported, increasing a number of bead-beating steps leads to a higher proportion of Gram-positive organisms (Figure 3.11)[111,116,134]. Interestingly, this was not seen uniformly across all individuals. It seems that increasing the number of bead-beating steps may be more useful for certain sample types than others. More complex, solid samples such as those from individual 11 appear to benefit from increased bead-beating, whereas for less solid samples, such as that from individual 10, increased bead-beating does not seem to be necessary, at least from a taxonomic perspective and could lead to DNA degradation. As we do not know the exact composition of the samples, it is difficult to ascertain the specific changes induced by additional homogenisation.

Despite this, it is critical to ensure that all samples are exposed to the same bias throughout the workflow and therefore we chose to use 8 bead-beating steps for the study. This will ensure that the more complex samples are thoroughly homogenised and should not be detrimental to samples with a simpler composition. In addition, 8 minutes of bead-beating is recommended by the International Human Microbiome Standards guidelines[147].

3.4.5 Dataset-specific error and contamination removal strategy

Mock communities sequenced alongside each sequencing run reflected our observations from those sequenced for the main optimisation work (Figure 3.13). We observed an over-inflation in the number of OTUs found in the initial unfiltered mock communities (Table 3.2), likely as a result of both PCR bias and bioinformatics error. This was true for all the sequenced MCs, giving an indication of systematic bias. Additionally, erroneous taxa were also present, likely due to bioinformatics error as contamination is less likely.

We decided to filter the MCs, as well as the associated datasets, by the abundance of an erroneous OTU appearing in the MC. We hoped to strike a balance whereby as many erroneous OTUs as possible are removed, whilst maintaining the original composition of the MC. This will filter the datasets at a low level, which should remove some of the OTUs created as a result of bioinformatics error and is generally recommended for 16S rRNA datasets. In addition to MCs, spike-ins may be useful to monitor error on a sample-per-sample basis.

In terms of contamination, we observed that most are gut commensals such as *Enterococcus* and *Bacillus*. It is therefore likely that at least some proportion of these is due to sample-to-control cross-over during extraction or well-to-well contamination prior to and after PCR. It has been previously reported that well-to-well contamination is more common in low-biomass samples [127,135,136].

Previously published contaminants of environmental origin were also observed, however these were not uniform across all NEC, and thus could be from extraction kits, PCR reagents or other consumables[129,130]. Although considerable care was taken in preparing samples for sequencing, contamination is

unavoidable, therefore it may be useful to track consumable lot numbers in the future, so that we are able to identify consumable-specific contaminants. Additionally, we sequenced a negative PCR control, however as only a single control was sequenced, we drew no conclusions from the findings. It would be useful to sequence repeated PCR controls in future studies, in order to identify PCR-specific contaminants. Further steps to reduce the impact of contamination could also include sample randomisation and reagent treatment with dsDNase [130,131].

We developed a filtering workflow as detailed in Figure 3.16. Although no consensus for contamination removal exists, several strategies have emerged. Some studies subtract contamination reads from sequenced samples, others choose an arbitrary prevalence cut-off, which determines a level at which an OTU is considered a contaminant and is filtered out of the dataset[148]. We felt that an OTU's distribution may not be a useful measure to indicate its origin, whereas a simple subtraction of all OTUs may not be sufficient and may adversely skew the data. Although it is possible, though unlikely, for a true contaminant to be introduced into the dataset at higher quantities than into the NEC, we felt that using abundance is appropriate to identify the key contaminants.

It is less likely that contamination would have a great impact on the results in studies utilising high-biomass samples, however it remains important to monitor and control for contamination so that we become more aware of common contaminants and can effectively control for them, ensuring the dataset is as free from bias as possible.

3.5 Conclusion

As a result of this optimisation work our proposed 16S rRNA workflow is as follows. Upon collection, samples will be thoroughly mixed and aliquoted into smaller sub-samples which will be kept at -80°C until further processing. Samples will be extracted using the Qiagen method (QIAmp fast DNA stool kit) with 8 bead-beating steps. The V3-4 primer set will be used to amplify the sample in a single-step PCR and the resulting DNA will be sequenced. Sequenced data will be filtered at an appropriate mock community level and contaminants identified and removed as detailed in the contamination removal workflow.

Chapter 4-Investigating longitudinal gut microbiota changes in children undergoing hematopoietic stem cell transplantation

4.1 Introduction

Few studies to date have explored the dynamics of the gut microbiota of paediatric HSCT patients, but perhaps the most in depth exploration to date comes from Ingham *et al,* who observed a decrease in alpha diversity around the time of transplantation, as well as a decrease in *Lachnospiraceae* and *Ruminococcaceae* and an increase in *Enterococcaceae* families between one and five weeks post-transplant[96]. They observed three community state types within their 16S rRNA data; a cluster with a high abundance of *Lactobacillaceae* and low abundance of *Ruminococcaceae,* which was common in patients with high human beta-defensin 2 levels and who went on to develop mild to severe GvHD and had higher levels of overall mortality; a cluster with a high abundance of *Ruminococcaceae* and *Lachnospiraceae*, a transient cluster common in patients who went on to have a successful recovery and a cluster with high abundance of *Enterococcaceae* associated to high levels of inflammation as measured by high CRP levels[96].

Bekker *et al* have also investigated the dynamics of the paediatric gut microbiota throughout total gut decontamination *versus* selective gut decontamination prior to HSCT[149]. They observed that total decontamination rendered the microbiome less stable and domination by *Enterococcus* and *Staphylococcus* was common, whereas composition remained relatively stable and dominated by *Bacteroides* throughout selective decontamination[149].

Finally, a recent study from Schluter *et al* used longitudinal samples from more than 2000 individuals to explore the gut microbiota and peripheral immunity[150]. The study revealed consistent associations between intestinal bacteria and peripheral immune cell dynamics[150]. Higher abundances of *Faecalibacterium* and *Akkermansia* were associated with increases in neutrophil levels, whereas *Ruminococcus* was associated with both neutrophil and lymphocyte increases. These associations support the idea that haematopoiesis responds to the composition of the gut microbiota.

As briefly alluded to in the introduction, the gut microbiota is dynamic and can shift from day to day as a result of dietary changes, drug use and other environmental factors. Although multiple studies have focused on the role of gut

microbiota in HSCT, cross-sectional studies, which focus on pre-specified timepoints throughout the procedure are unlikely to capture the true variability and dynamics of the gut microbiota. In this chapter I aim to explore the dynamics of gut microbiota throughout paediatric HSCT.

4.1.1 Aims

The overall aim of this chapter was to characterise the longitudinal changes in the gut microbiota of children undergoing HSCT.

4.2 Results

4.2.1 Patient cohort

As previously described (Chapter 2), 64 individuals (4 months to 14 years) undergoing HSCT at GOSH were consented to participate in our study. We collected a total of 648 samples; an average of 8 samples (range; 2-32) per person. We were unable to sequence 89 of these, as they did not result in enough DNA for sequencing. Most of the samples which could not be sequenced were collected around two weeks post-HSCT, therefore this could be due to antibiotic administration that begins prior to transplantation (Figure 4.1).

Antibiotic administration can be patient specific, however the majority start on ciprofloxacin as prophylaxis (until central line removal; Figure 4.1) and cord blood HSCT recipients receive vancomycin until neutrophil engraftment (neutrophils $>0.5x10^9$/L for three consecutive days). Additionally, co-trimoxazole is given to all patients until their CD4 counts are ≥300 cells/μl and they are off immunosuppression. Penicillin or azithromycin are administered prophylactically to cover Gram-positive organisms around day 28 post-HSCT; co-trimoxazole/pentamidine is used to cover *Pneumocystis jirovecii* pneumonia. Overall, the use of antibiotics with anaerobic coverage (e.g. ciprofloxacin, metronidazole, vancomycin, meropenem, piperacillin-tazobactam) is high within this population.

Similarly, acyclovir is used as antiviral prophylaxis from the start of conditioning (Appendix Figure A1 A). Antifungal prophylaxis with itraconazole is generally used until neutrophil engraftment, CD4 counts are ≥300 cells/μl and they are off immunosuppression (Appendix Figure A1 B).

In addition to antimicrobial usage, we investigated corresponding changes in immunological status (neutrophil counts and C-reactive protein (CRP) levels) during transplantation. Focusing on the first 100 days post-HSCT, neutrophil counts decreased following conditioning, which was followed by a gradual increase (neutrophil engraftment) approximately 2 weeks post-transplantation (Figure 4.2 A). CRP levels were variable (Figure 4.2 B), however there was an overall decline in CRP post-HSCT, higher levels being indicative of systemic inflammation.

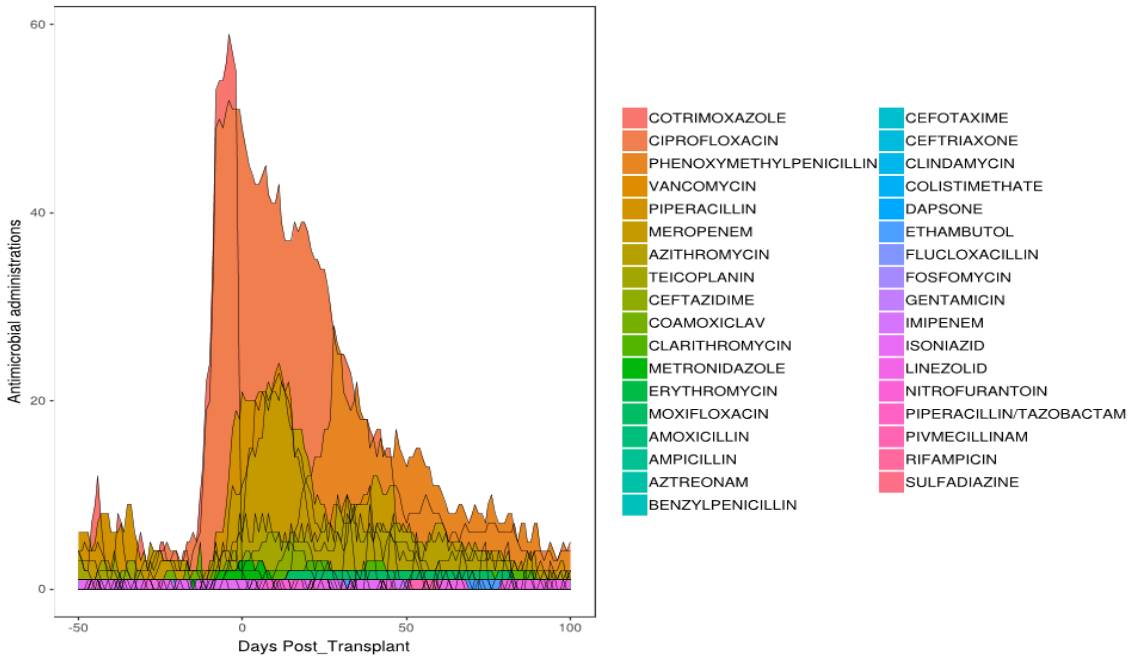


Figure 4.1 Antibiotic administrations during the first 100 days of treatment.

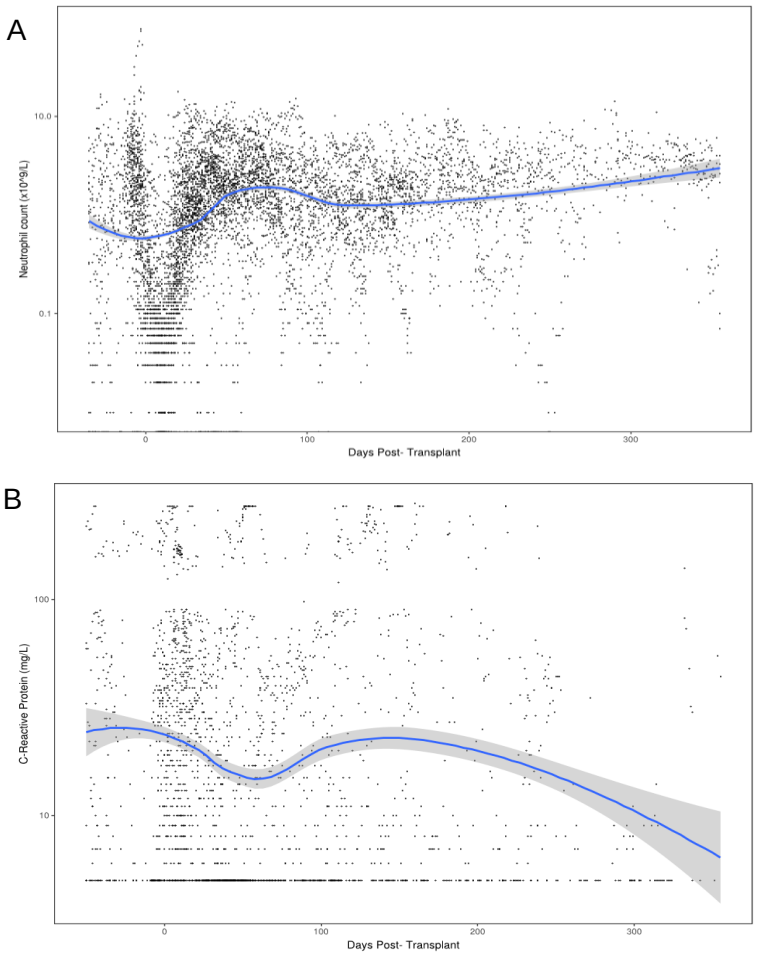


Figure 4.2 Log10 transformed **A)** neutrophil counts (x10^9/L) and **B)** CRP (mg/L) trends in response to treatment. The fitted line shows a local polynomial regression fit calculated using loess in R, with the grey region indicating the 95% CI.

4.2.2 Changes in faecal microbiota alpha diversity

A decrease in alpha diversity (effective Shannon entropy) was seen in patients post-HSCT (healthy range 7-21) (Figure 4.3 A). The decrease was particularly evident when samples taken at the start of treatment (at admission) were compared with those obtained post-HSCT. Importantly on average, alpha diversity did not recover within the observed period, although note there was a considerable decrease in the number of patients remaining in hospital beyond 100 days post-HSCT.

Next, we found that the dynamics of alpha diversity differed in those under the age of 2 when compared to those older than 2 years as well as between allogeneic and autologous recipients (Figure 4.3 B/C). Alpha diversity appeared less stable in those under 2 years of age, with a decrease early post-HSCT. Interestingly, this decrease was eventually followed by an increase to the levels seen in those older than 2 years of age. Similarly, there was a steady decrease in alpha diversity throughout the first 100 days post-allogeneic transplantation (Figure 4.3 C). In contrast Shannon effective entropy improved greatly (following the initial decrease) in autologous transplantation.

We then investigated longitudinal alpha diversity in more detail by looking at the changes from the initial sample for each patient throughout their treatment (Figure 4.4 A).

Irrespective of duration of hospitalisation, most individuals did not recover their diversity during their stay, which is an important finding (Figure 4.4 B/C). Figure 4.4 C reflects the microbiota of an individual who was re-admitted and we found their alpha diversity did not recover even almost a year post-HSCT. Interestingly, a few patients did recover within a relatively short amount of time (Figure 4.4 D). In a subset of individuals, uncharacteristic spikes in diversity were detected, generally a few weeks after transplantation, these spikes most likely reflected colonisation with a diverse ‘hospital’ microbiota (Figure 4.4 E).

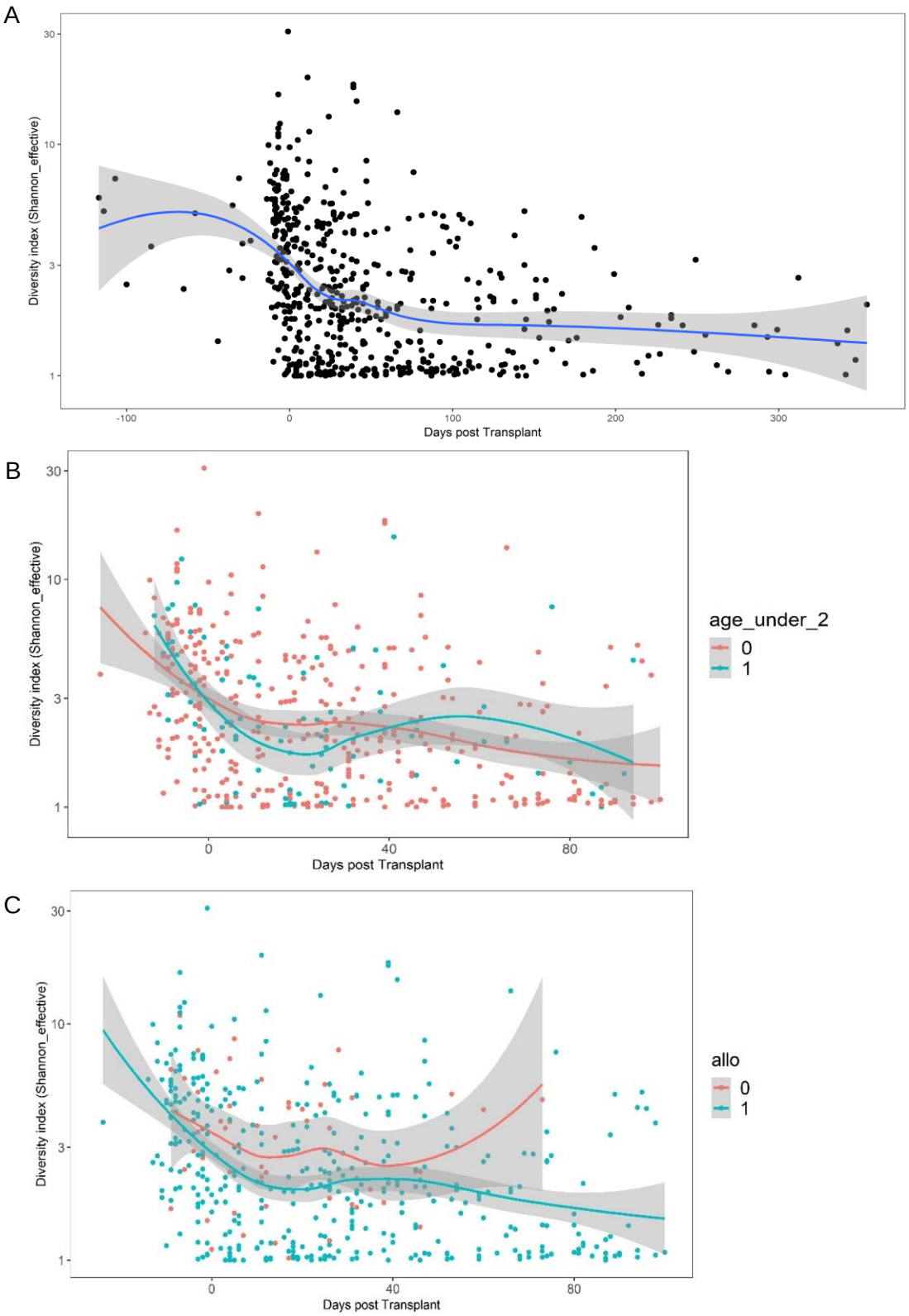


Figure 4.3 A) Log transformed Shannon effective entropy throughout transplantation; broken down by whether samples came from patients **B)** >2 (0) vs <2 years of age (1) and **C)** allogeneic transplantation (1) vs autologous transplantation (0) within the first 100 days of inpatient stay. The fitted line shows a local polynomial regression fit calculated using loess in R, with the grey region indicating the 95% CI.

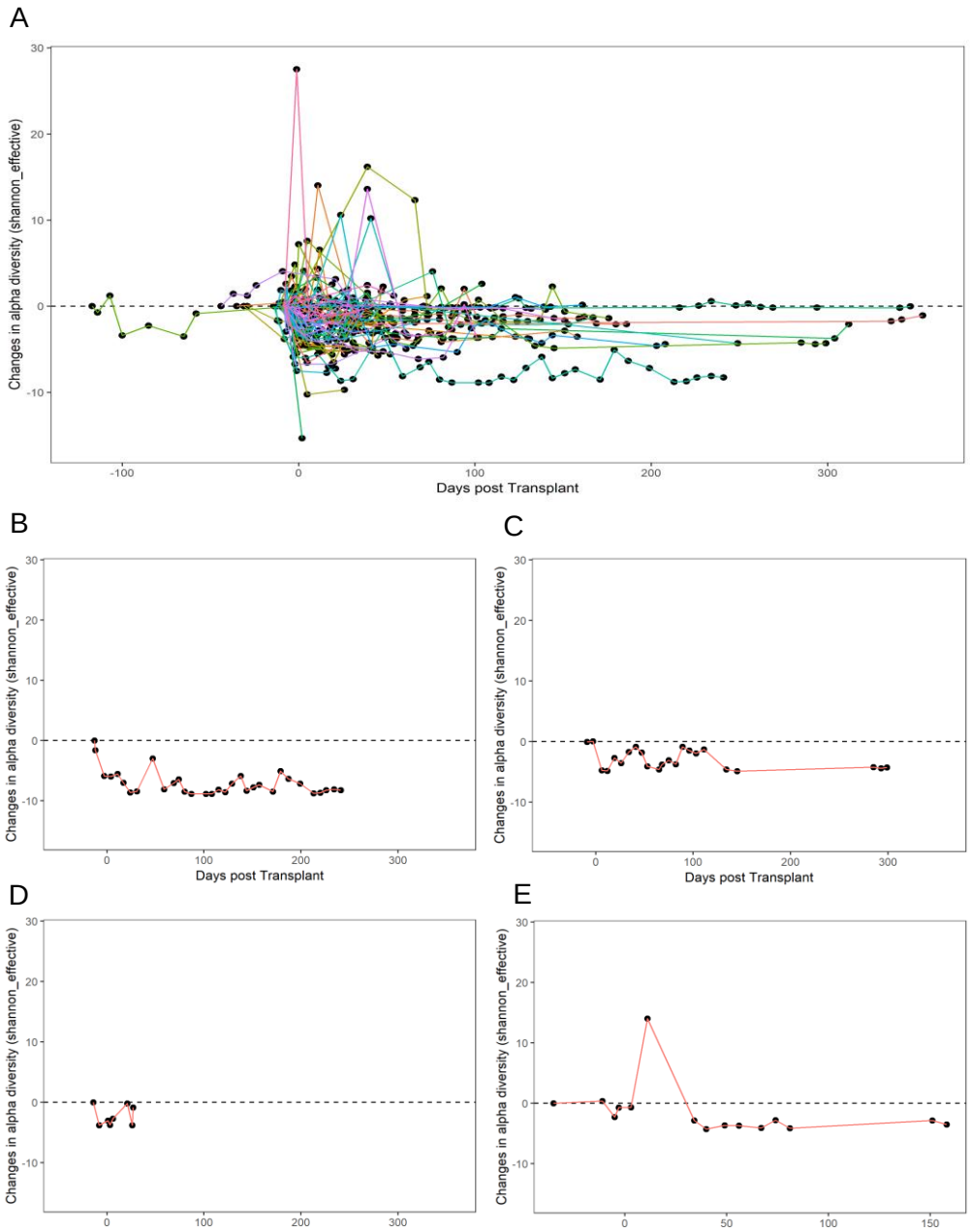


Figure 4.4 A) Deviation in Shannon effective entropy from the initial sample for each individual for the cohort, as well as several examples of individual profiles **(B, C, D, E)**.

4.2.3 Development of microbiota domination

As expected, the microbiota changed dynamically and frequently throughout treatment in children undergoing HSCT. Figures 4.5 and 4.6 detail microbiota profiles for several individuals at the family level.

Domination of specific taxa (>30% relative abundance) was evident post-HSCT, with *Enterococcacceae*, *Enterobacteriaceae*, *Streptococcaceae* and *Staphylococcaceae* being the most dominant (Figures 4.5/4.6). Domination has been previously reported in other cohorts, however certain dominant taxa appeared to be specific to this cohort[79,96]. Patients had varying microbiota dynamics. For example, patient 31 (Figure 4.5), appeared to go through several different dominations, reflecting a highly dynamic microbiota. In contrast, patient 26 showed the presence of a stable microbiota with an approximate return to the initial composition (Figure 4.6). Further taxonomic plots are detailed for both allogeneic and autologous HSCT recipients (Appendix; Figures A2/A3).

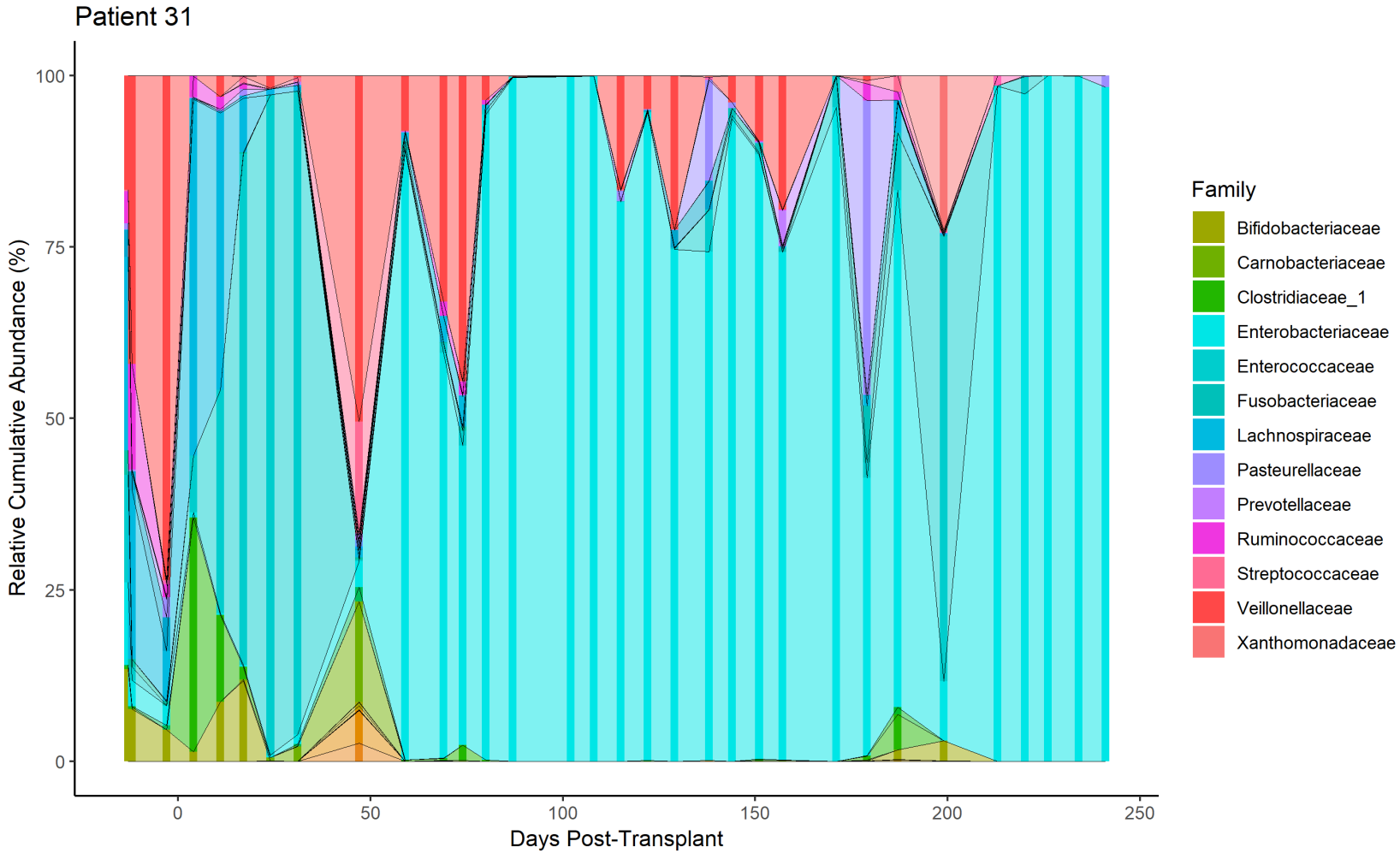


Figure 4.5 A taxonomic plot for patient 31 (allogeneic HSCT) throughout hospitalisation at a family level. Only families with an overall abundance of >10% are labelled for clarity. Darker vertical bars reflect sample collection points, whereas abundances between these are inferred.

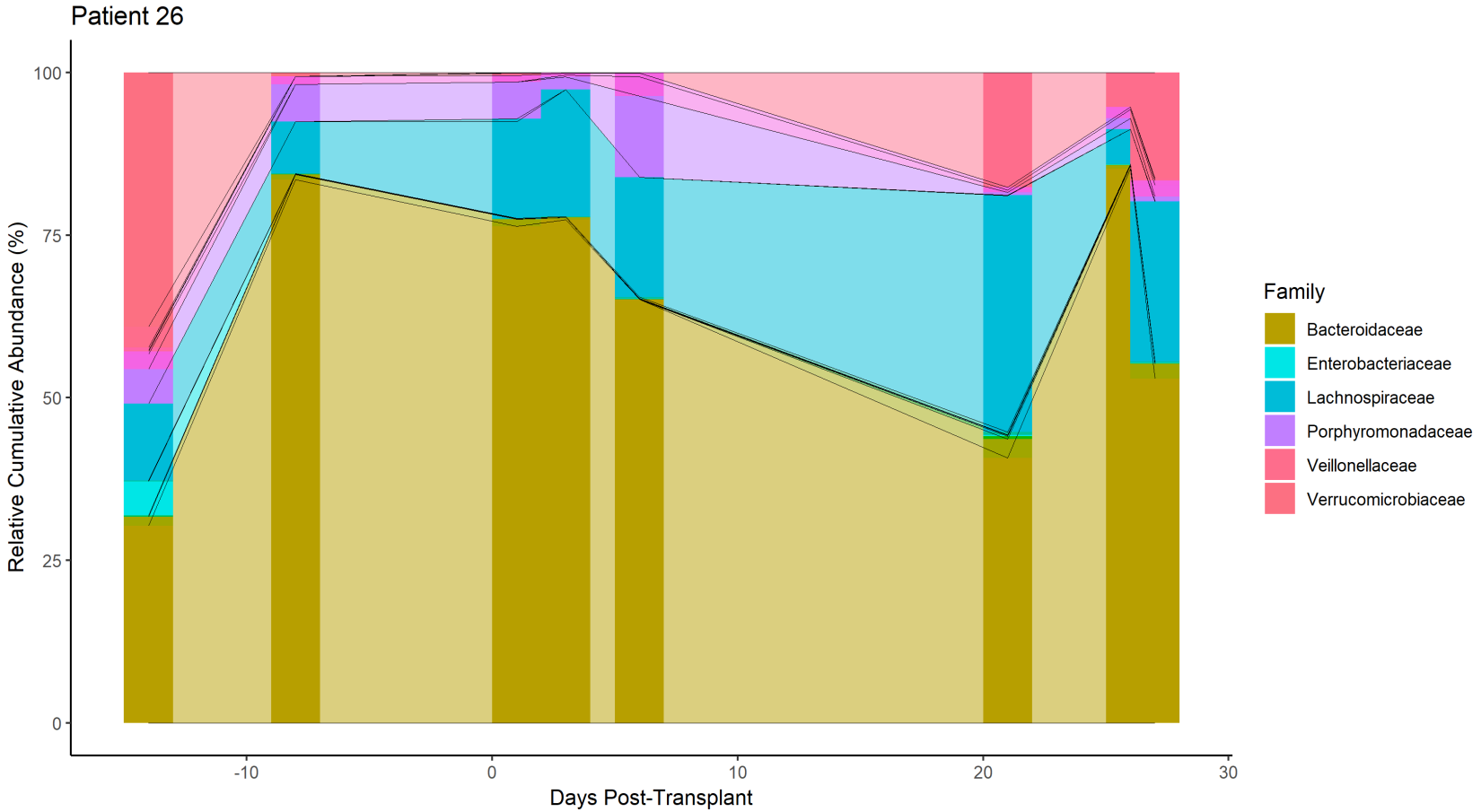


Figure 4.6 A taxonomic plot for patient 26 (allogeneic HSCT) throughout hospitalisation at a family level. Only families with an overall abundance of >10% are labelled for clarity. Darker vertical bars reflect sample collection points, whereas abundances between these are inferred.

4.2.4 Individual taxa dynamics

Investigation of individual taxonomic profiles revealed several patterns of domination in the first 100 days post-HSCT in the GOSH cohort. Some of the taxa found to be dominant throughout hospitalisation are detailed in Figure 4.7.

Time-dependent changes in patterns of domination post-HSCT were observed. Both *Bacteroidaceae* and *Lachnospiraceae* taxa decreased post-transplantation. *Enterococcaceae* dominance for example was most common early post-transplantation, whereas *Enterobacteriaceae* showed a steady increase in relative abundance over time and high levels of dominance (>30%) occurred later post-HSCT (Figure 4.7 D). Both *Staphylococcaceae* and *Streptococcaceae* relative abundances increased with time, however *Streptococcaceae* dominance was more common earlier post-transplantation.

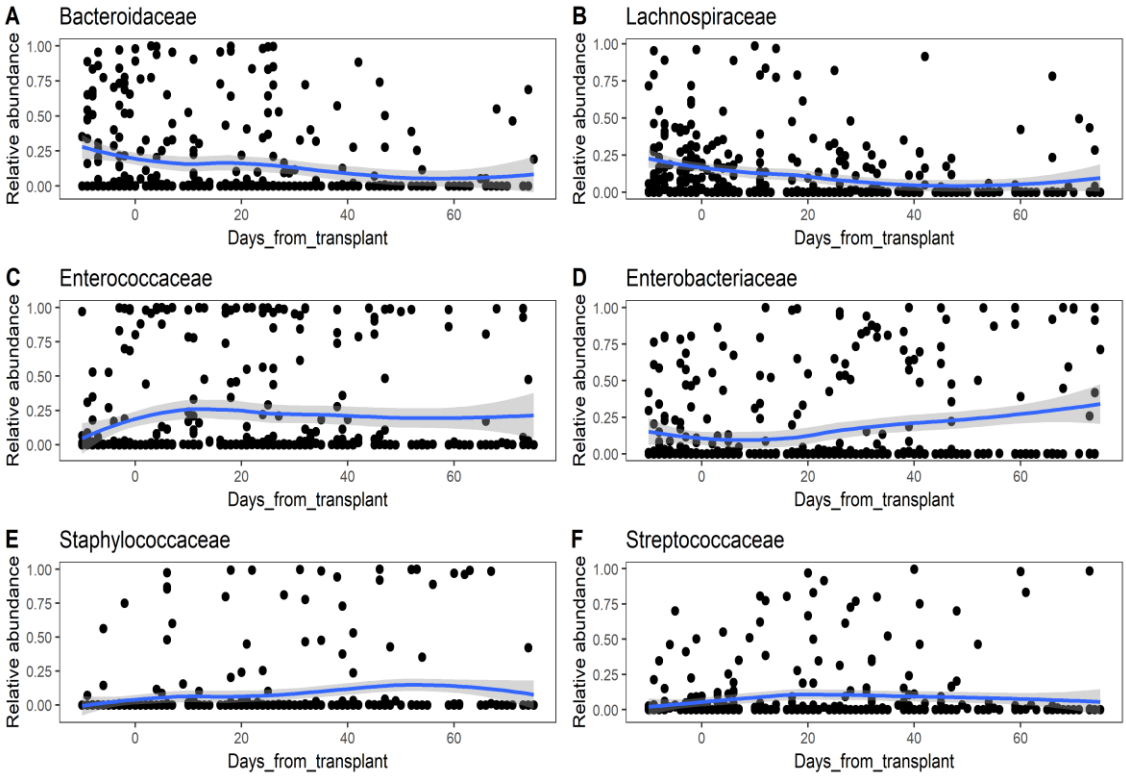


Figure 4.7 Relative abundance of several taxa at family level found to be dominant in the cohort throughout the first 100 days of inpatient stay. The fitted line shows a local polynomial regression fit calculated using loess in R, with the grey region indicating the 95% CI.

4.2.5 Microbiota trajectories throughout HSCT

Having broadly investigated taxonomic profiles of each individual throughout transplantation we employed t-SNE as an ordination method to visualise all samples and their local and global relationships. Although other methods, such as PCA, were initially utilised, they were unable to resolve the complex community structures within the large dataset and so a t-SNE was preferable.

Samples did not cluster by any baseline variable such as age, sex or by individual. The plot was then coloured by the predominant taxon within a particular sample, which revealed a dynamic landscape of gut microbiota composition throughout HSCT with several regions of domination within (Figures 4.8/4.9). We highlighted individual trajectories through this landscape. Figure 4.8 details the trajectory of patient 31, whose initial sample places them in a *Lachnospiraceae* area of the landscape, followed by progression to *Enterococcaceae* predominance and finally settling in an *Enterobacteriaceae* predominant landscape. The last sample (black square) was dissimilar to the composition of the initial sample (black triangle). This individual's trajectory reflects those seen in majority in our cohort, i.e. showing *(a)* stochastic and manifold movements and *(b)* lack of recovery to the initial microbiota status whilst hospitalised.

Figure 4.9 details the trajectory of patient 26, who started within the *Bacteroidaceae* area of the landscape and remained within the same area post-HSCT. This reflects the relative microbiota stability in this individual, compared to patient 31 who moved stochastically throughout the landscape.

Although most trajectories did not recover to their initial state within the hospitalisation period, several patients did return for further follow up after discharge. Patient 1 for example, returned for GvHD treatment a year post-transplantation, however their microbiota remained *Enterococcaceae/Enterobacteriaceae* predominant, suggesting ongoing complexity in host immune-microbial interactions.

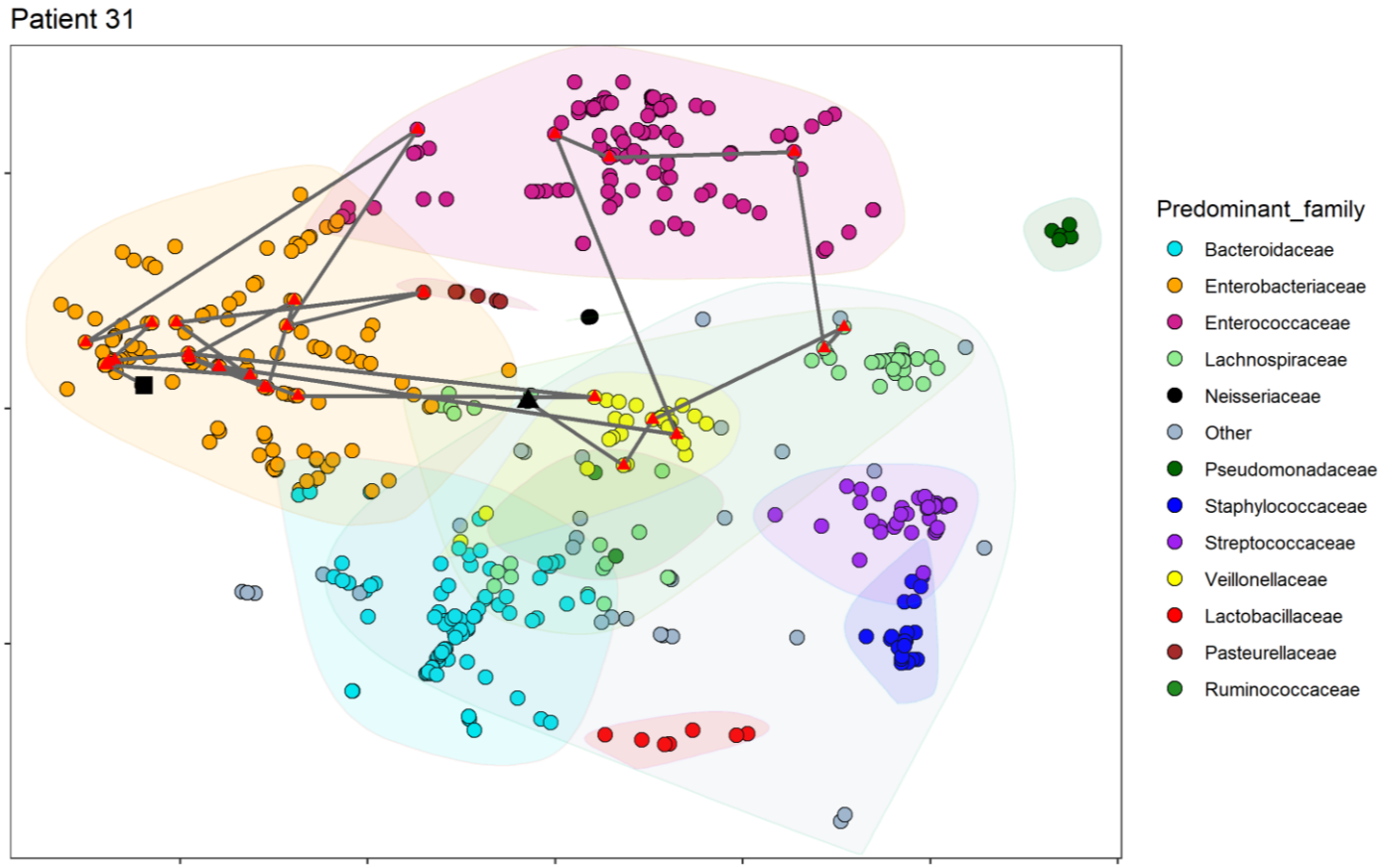


Figure 4.8 A t-SNE plot of all samples collected in the study. Each sample is coloured by its predominant taxa (>30%) The lines represent a trajectory for patient 31 with a black triangle signifying the first collected sample and the black square signifying the last collected sample. Red triangles indicate all samples collected for an individual. Other is composed of several infrequently dominant taxa.

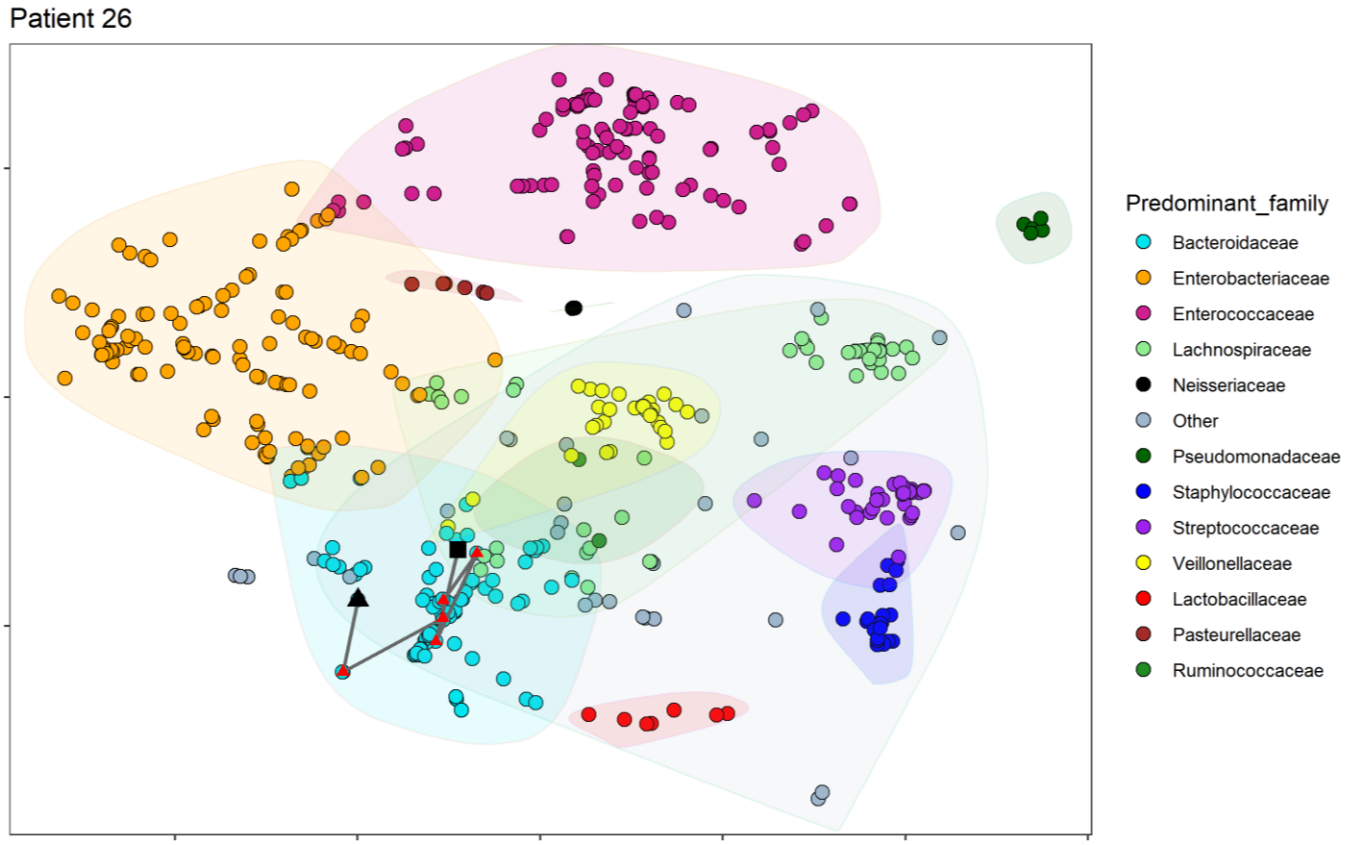


Figure 4.9 A t-SNE plot of all samples collected in the study. Each sample is coloured by its predominant taxa (>30%) The lines represent a trajectory for patient 26 with a black triangle signifying the first collected sample and the black square signifying the last collected sample. Red triangles indicate all samples collected for an individual. Other is composed of several infrequently dominant taxa.

4.2.6 Microbiota of HSCT patients *versus* healthy controls

Having found that some patients appear more stable than others and recover from microbiome changes throughout transplantation, we wondered which area of the landscape could be considered more healthy-like. For this, we re-plotted the t-SNE plot together with the healthy controls (Figure 4.10 A).

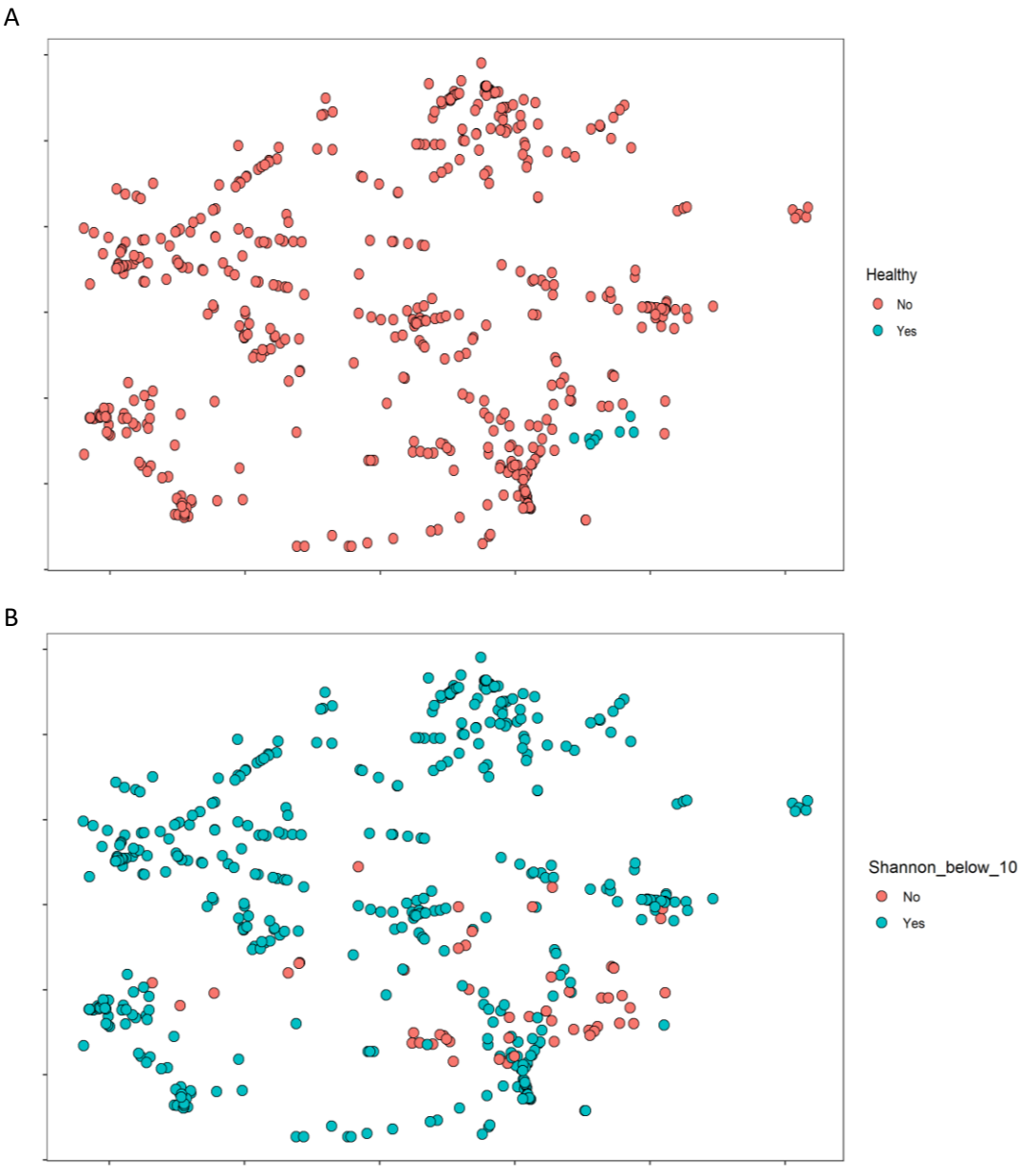


Figure 4.10 A t-SNE plot of all samples collected in the study. Samples are coloured by **A)** whether they came from a healthy control or a patient and **B)** by whether the sample's Shannon effective entropy is <10.

Healthy individuals clustered together closest to the *Bacteroidaceae* area of the landscape. We also observed a higher diversity area (around the *Bacteroidaceae*

predominant area) within the landscape (Figure 4.10 B). It seems therefore, that whilst some patients may start out in a relatively heathy area (Figure 4.9), a large proportion of patients do not (Figure 4.8). In summary, many of the samples collected in the patient cohort can be considered dissimilar from healthy controls.

4.2.7 Microbiota community state types in HSCT

As domination was a common occurrence within the cohort, we employed PAM clustering, in order to identify distinct bacterial community patterns. The data clustered best into three CSTs. CST diagnostic plots are detailed in the Appendix (Figure A5).

Figure 4.11 details the family level taxonomic composition of all samples within their CST. *Bacteroides, Clostridium_XIVa*, *Blautia* and *Lachnospiraceae_unclassified* genera appear to be predominant in CST1, with the majority of the early samples (and healthy controls) classifying as CST1 (n=128). *Staphylococcus* and *Streptococcus* predominated in CST2 (n=279), whilst *Enterococcus*, *Escherichia_Shigella*, *Veillonella*, *Klebsiella* and *Enterobacteriaceae_unclassified* were predominant in CST3 (n=125). Taxa were not exclusive to their cluster, for example *Enterococcus* mean abundance was higher in CST3, even though it was present in all three CSTs.

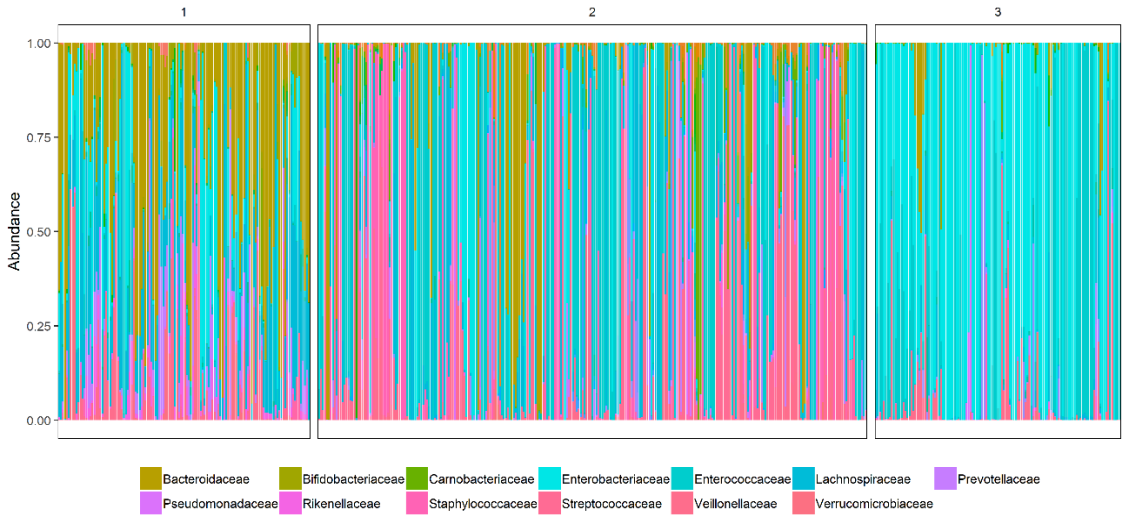


Figure 4.11 Taxonomic composition (at family level) of all collected samples broken down by the community state type they belong to.

Having identified three CSTs in our data set, we wanted to visualise CST patterns throughout HSCT. Figure 4.12 details a timeline of all samples, coloured by their respective CST. A small proportion of individuals remained in their respective CST (1 or 2) throughout transplantation. Having said this, we were not able to sequence a sample prior to transplantation for all individuals that remained in CST2 throughout, therefore their CST status prior to HSCT is unclear.

A large proportion followed a similar pattern, whereby their initial sample was classified into CST1, followed by a transition into CST2 or CST3 post-transplantation (Figure 4.12). A small proportion of individuals however showed an initial CST2 phenotype or, less commonly, CST3. Finally, a proportion of individuals exhibited complex CST patterns, switching continuously between the states. This appeared as common among both individuals that start out in CST1 and those that do not. In summary, individuals moved dynamically through the CSTs throughout transplantation, CST1 was more common pre-transplant, whereas CSTs 2 and 3 dominated post-transplantation.

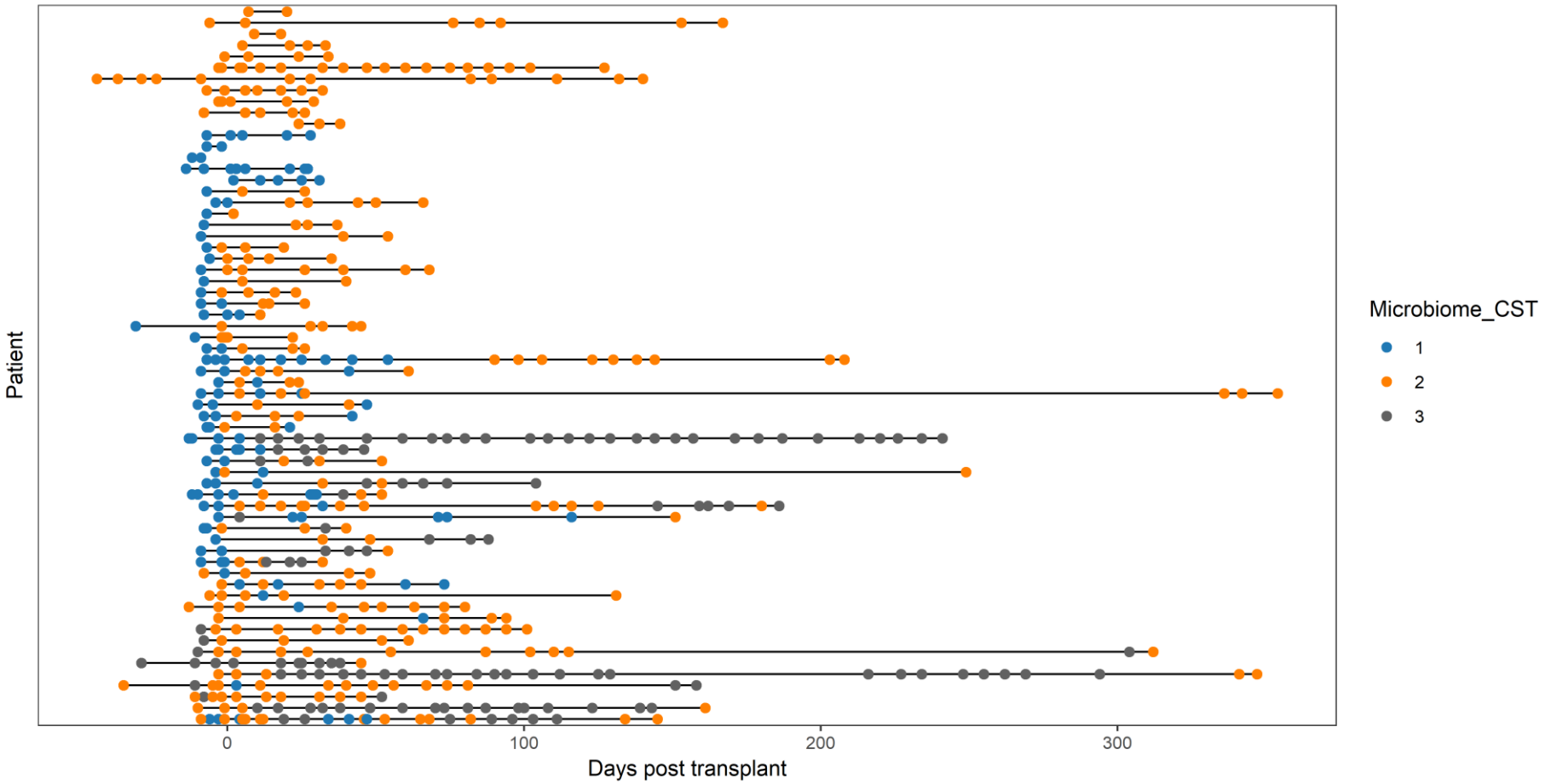


Figure 4.12 A timeline of all samples collected in the study. Samples are coloured by the CST they have been classified into.

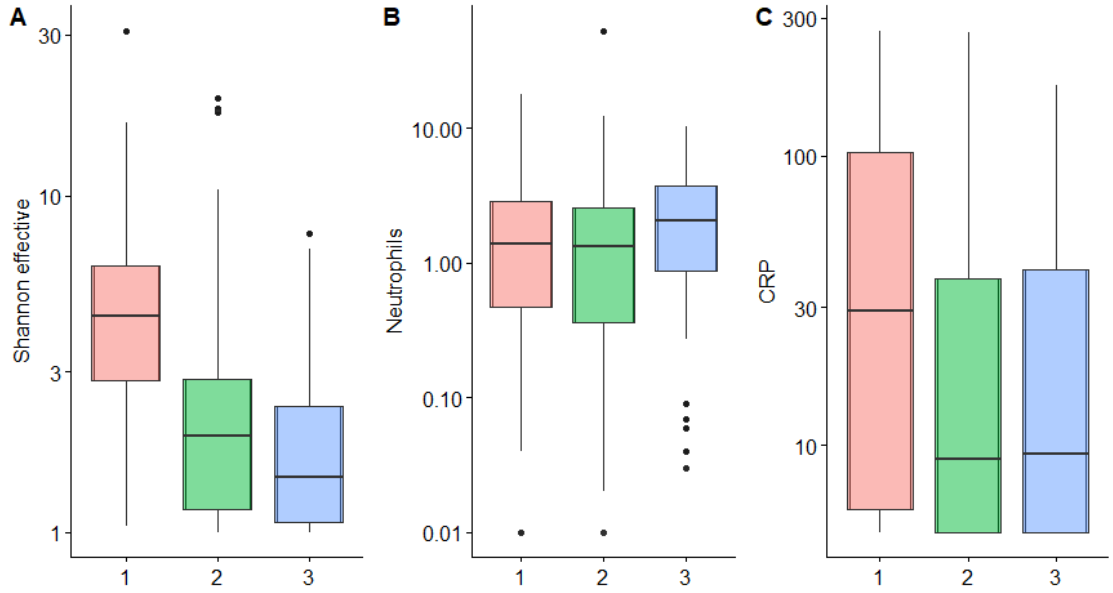


Figure 4.13 Characteristics of CSTs. Differences in **A)** Shannon effective entropy **B)** absolute neutrophil count (x10^9/L) and **C)** CRP (mg/L) between CSTs.

We then investigated CST patterns further by plotting them against alpha diversity and several immunological markers (Figure 4.13). CST1, as expected, reflects greater diversity (Shannon effective entropy) as it does trend towards a healthier composition. Similarly, neutrophil counts were higher for CST3, in comparison to CST1, potentially as CST3 was more predominant post-transplant. Finally, CRP appeared to be very variable, but appears to be higher in CST1 in comparison to CSTs 2 and 3, which is likely reflective of greater inflammation in patients prior to treatment. It is likely that the immunological markers were an indication of time post-transplantation, rather than CST characteristics.

4.2.8 CST dynamics

The probabilities of remaining in the current CST or switching to another was then examined in the cohort. Figure 4.14 details a transition model generated from transitional probabilities of all samples. Overall, CST2 and CST3 showed high self-transition probabilities of 0.78 and 0.83 respectively. That is both CST2 and CST3 are quite stable, and once a patient is in one of these CSTs, it is likely that they will remain there. In contrast, CST1 showed a self-transition probability of 0.4.

The probability of CST switching was variable (Figure 4.14). As observed in the timeline (Figure 4.12) a switch from CST1 to CST2 (0.49) was more likely than a direct transition from CST1 to CST3 (0.11). It is unlikely that one would return to CST1 from either CST2 (0.13) or CST3 (0.06). Similarly, the probabilities of switching between CST2 and CST3 were also low (0.09/0.11).

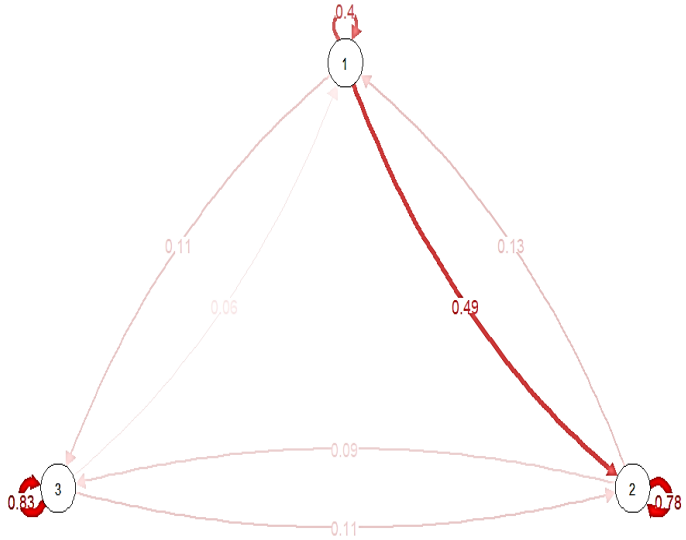


Figure 4.14 Transition model showing the progression of samples through each CST. Colour intensity is indicative of transition frequencies.

We then subset the samples to only those within the first five weeks as most patients remain in hospital for at least a month and we were interested in exploring CST dynamics immediately post-transplantation, when the patients are more susceptible to complications (Appendix; Figure A6). Figure 4.15 details transition frequencies over the first five weeks of treatment, with the colour intensity of the lines reflecting transition frequencies. It is interesting to note that whilst CST1 was the most abundant in week 1 (prior to transplantation) and its abundance decreases with time, CST3 showed an opposite trend of an increase in abundance with time, rarely being present in week 1. CST2 on the other hand, appeared most prevalent in week 2 (the week following transplantation), however there was no evident time-dependent trend in terms of its prevalence. CST3 was almost non-existent at weeks 1 and 2, however there was an increased frequency of transitions beginning at week 3.

The majority of CST1 to CST2 transitions occurred immediately following transplantation, which may reflect the start of the development of the dominations

observed in the taxonomic plots. There were some transitions back to CST1, however this was a rare event and decreased in frequency with time. There were no transitions to CST1 after week 4. The average time individuals remained in each CST was variable, with CST1 being 8 days, CST2- 26 days, and CST3- 37 days.

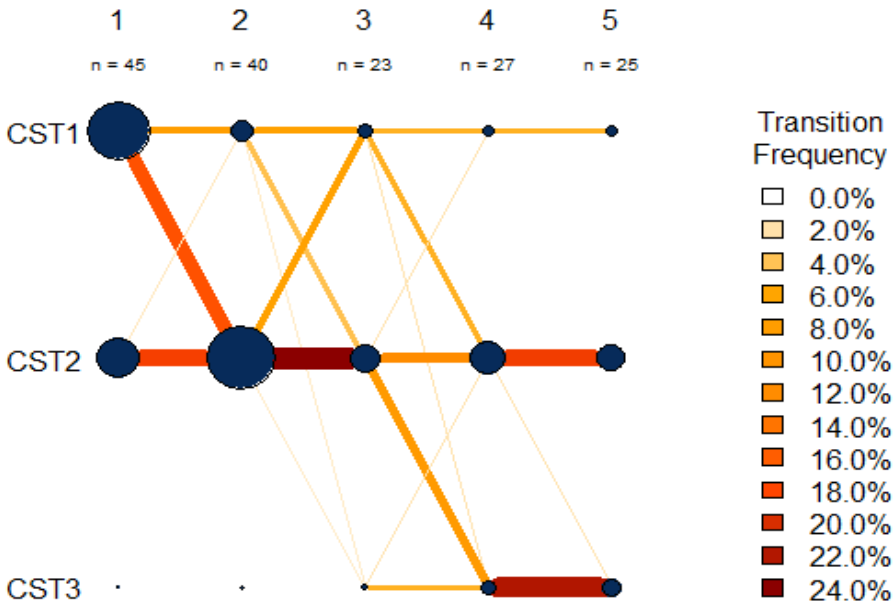


Figure 4.15 Transition model showing the progression of samples through each CST per each time point, from week 1 (starting at day -7 relative to transplantation) to week 5. Line colour intensity is indicative of transition frequencies and the node sizes correspond to sample size within each node.

4.2.9 Community state types and clinical outcomes

Finally, we were interested in whether a particular CST is linked to adverse clinical outcomes. We focused on GvHD and viraemia using either outcome as a dependent variable in a time-dependent Cox model. No significant associations between GvHD (>grade II) and CST or other clinical variables (Appendix; Table A12) were found.

Univariate Cox model indicated that *(a)* more than one transplant and *(b)* microbiota CSTs associated with a higher risk of viraemia (Appendix; Table A13). Adjusting for more than one transplant within one's lifetime, CST3 was associated with an increased the risk of viraemia in comparison to CST1 (Figure 4.16; $p<0.01$; HR-2.07).

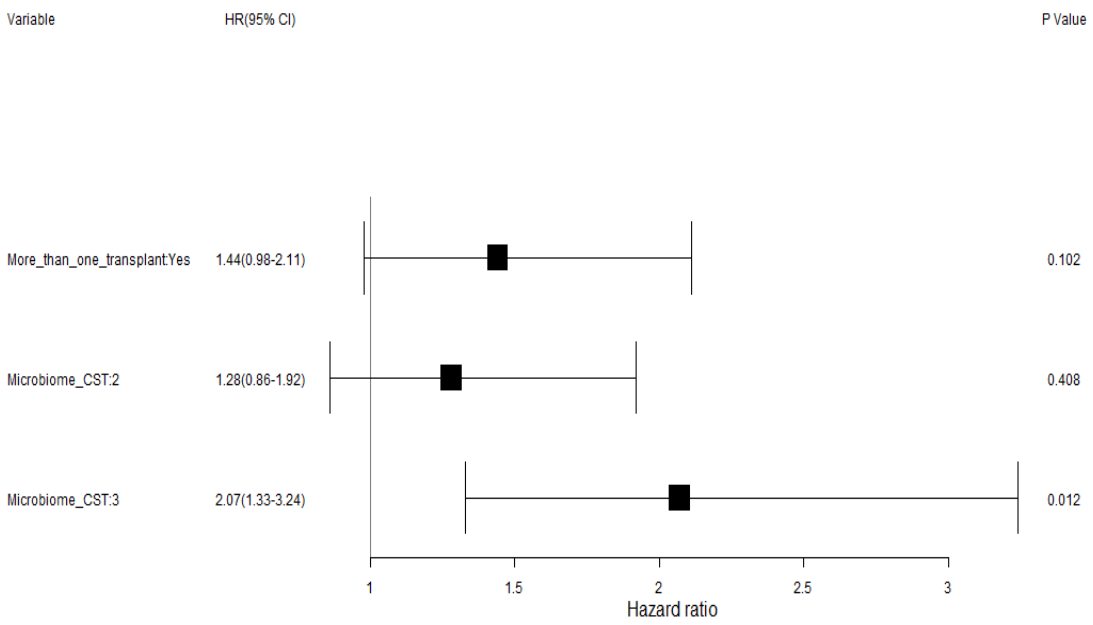


Figure 4.16 Forest plot of hazard ratios (HR) with 95% confidence intervals (CIs) for viraemia associated with microbiota CSTs from a multivariable time-dependent Cox model. The reported HR were adjusted for more than one transplant and the 95% CI and P values were estimated using the robust sandwich estimator.

Having found an association between CST3 and an increased risk of viraemia we wished to know if this potential association was driven by a particular taxon. Dominance with any of the five taxa with the highest mean relative abundance within CST3 (*Enterococcus, Veillonella, Escherichia_Shigella, Enterobacteriaceae_unclassified, and Klebsiella*) was used as the independent variable *versus* samples not dominant in any taxa that are predominant in CST3 (Figure 4.1). No significant associations between the taxa and viraemia were found although *Enterobacteriaceae*_unclassified and *Enterococcus* were borderline significant.

Table 4.1 Univariate time-dependent Cox models with taxa domination (>30%) as an independent variable and viraemia as the dependent variable

Variable	HR	95% CI	P value
Enterococcus_dominant vs others	1.45	0.94-2.26	0.09
Enterobacteriaceae_unclassified_dominant vs others	1.73	0.94-3.18	0.08
Escherichia_Shigella_dominant vs others	1.02	0.63-1.66	0.92
Klebsiella_dominant vs others	0.57	0.19-1.71	0.32
Veillonella_dominant vs others	1.44	0.88-2.36	0.14

Others (samples not dominant in the particular taxon and not in CST3). The 95% CI and P values were estimated using the robust sandwich estimator. P value of <0.5 was considered significant. Abbreviations: CI, confidence interval; HR, hazard ratio.

Finally, we investigated if any clinical and baseline variables may be associated with developing *Enterococcus* domination. Other dominant taxa within CST3 were also of interest, however *Enterococcus* domination was the most common within CST3 and the only one with sufficient numbers for further investigation.

In the univariate Cox model (Table 4.2) both myeloablative conditioning (in respect to reduced intensity conditioning) and the use of macrolides were associated with an increased risk of *Enterococcus* domination with hazard ratios (HR) of 2.14 and 2.04 (p=0.03; 0.05), respectively. In a multivariable Cox model, neither of these variables were significant, although myeloablative conditioning was borderline significant (p=0.06).

Table 4.2 Univariate and multivariable Cox models with enterococcus domination (>30%) as the dependent variable. The 95% CI and P values were estimated using the robust sandwich estimator. P value of <0.5 was considered significant.

	Univariate			Multivariable		
Variable	HR	95% CI	P value	HR	95% CI	P value
Sex_Female:Yes	0.61	0.23-1.62	0.32	-	-	-
Age_under_2:Yes	0.61	0.21-1.77	0.36	-	-	-
Diagnosis:Malignant	0.63	0.32-1.23	0.30	-	-	-
More than 1 transplant:Yes	1.71	0.85-3.42	0.13	-	-	-
Conditioning:Myeloablative vs Reduced/Minimal intensity	2.14	1.08-4.23	0.03	1.89	0.97-3.66	0.06
Cell source: Cord vs Others	0.45	0.10-1.91	0.28	-	-	-
Quinolones:Yes	0.13	0.58-2.25	0.70	-	-	-
BS_beta_lactams:Yes	1.63	0.85-3.14	0.145	-	-	-
Macrolides:Yes	2.04	1.00-4.15	0.05	1.61	0.81-3.19	0.18

Abbreviations: CI, confidence interval; HR, hazard ratio; -, not significant.

4.3 Discussion

The interactions between the paediatric gut microbiota and its impact on clinical outcomes in patients undergoing HSCT is unclear. To gain better understanding of the impact of this axis, we initiated a study exploring longitudinal changes in the composition of gut microbiota in paediatric patients undergoing HSCT. The present study is the largest investigation in paediatric allo-HSCT conducted in the UK to date.

4.3.1 Alpha diversity

Alpha diversity measures the species diversity within a sample and Shannon entropy takes into account the evenness and the overall numbers of the taxa within each sample to calculate this value. The mean Shannon effective entropy for patients was approximately 3 (range; 1-30.8), whereas the healthy controls had an effective Shannon entropy of ~13 (range; 7.7-21.4).

Our results indicate that alpha diversity decreases throughout transplantation, most notably immediately following the HSCT procedure. Additionally, it does not, in most cases, return to its initial value within the timeframe of the investigation. This is in line with a previous publication in adults, which found that diversity does not re-establish within the hospitalisation period[79]. Several patients came back for further treatment a year post-HSCT, however even at this point their diversity had not re-established, which is possibly due to continuing antibiotic and other drug administrations. Interestingly, the diversity we observed appears lower than that reported for another paediatric HSCT cohort, perhaps reflecting the more complex clinical cases treated at GOSH, such as a greater number of mismatched transplantations undertaken[96]. Our preliminary data suggests that auto-HSCT patients show an improved diversity trajectory compared to allo-HSCT patients, which is in line with observations in adults[151]. The subtleties of the trends observed with age and allogeneic *versus* autologous transplantation would be of interest in the future.

4.3.2 Microbiome domination

A key finding was the observation of microbiome domination throughout hospitalisation. The overgrowth of several taxa notably *Enterococcaceae*,

Enterobacteriaceae, *Staphylococcaceae* and *Streptococcaceae* was routinely recorded, whilst domination by *Bacteroidaceae, Veillonellaceae* as well as other taxa was less frequent. Several adult studies have reported this phenomenon. Pamer *et al* have done extensive work in adult allo-HSCT and observed domination of *Enterococcus* and less commonly *Streptococcaceae* and *Proteobacteria*, containing the *Enterobacteriaceae* family among others[69,79].

The few paediatric studies suggest domination of *Enterobacteriaceae* and *Lactobacillaceae* as well as *Enterococcaceae*[96,84,152]. Biagi *et al* highlight taxonomic differences between three paediatric transplant centres included in their study[153]. Collectively, the data suggests that the driver for domination by specific taxa is likely to be multifactorial and that some of the dominant taxa we observe may be specific to our cohort. These differences could be partly due to dissimilar antimicrobial and other drug regimens used in each transplant centre as well as the hospital environment.

Intestinal microbiota is in a complex relationship with the mucosal epithelium. Epithelial cells and the microbiota establish a state of equilibrium, which facilitates optimal nutrient absorption as well as resistance to infections[154]. Chemotherapy-induced epithelial cell damage, the dampening of the immune system as well as administration of antibiotics together are likely to disrupt this equilibrium, leading to the phenomenon we observe[154]. Given the role of microbiota in immune homeostasis, it would be interesting to find out whether extreme microbiota shifts overall, or specific taxon dominance impacts on immune reconstitution post-HSCT. Indeed, several murine studies indicate that germ-free mice have defects in haematopoiesis and antibiotic-treated mice showed multilineage repression of haematopoiesis[155,156]. Additionally, Staffas *et al* recently demonstrated that intestinal microbiota contributes to haematopoiesis post-HSCT via improved dietary energy uptake in mice [157].

Antibiotic administration plays an important role in the development of the domination of certain taxa. Researchers have found links between the use of antimicrobials with anaerobic coverage such as metronidazole and vancomycin, and *Enterococcus* domination[79,80]. Morjaria *et al* found that both meropenem and piperacillin-tazobactam led to a loss of obligate anaerobes[158]. In contrast, we did

not observe any potential association(s) between any of the antimicrobials and *Enterococcus* domination. This lack of association is most likely due to small sample size and the ubiquitous use of certain antimicrobials such as quinolones, which can obscure their effect. It is likely that antibiotic use, as well as the epithelial damage as a result of conditioning, leads to a change in the gut environment such as a pH change and the loss of beneficial taxa and associated metabolites, which leads to an overgrowth of already present pathogenic taxa.

Although antibiotic administration is likely to play a major role, other non-antimicrobial medication such as proton-pump inhibitors, antivirals and antifungals also have unintended antibacterial effects and so it is unclear whether and to which extent they play a part in taxa domination[159]. To clarify the impact of antimicrobials and other drugs on taxa domination, a larger study in our patient cohort is warranted.

4.3.3 Microbiome dynamics

In order to visualise each individual's trajectory throughout hospitalisation, we plotted the samples in a t-SNE space. This revealed a dynamic HSCT landscape, with an area of relative health and areas of taxa domination. Most individuals showed frequent stochastic movements across the space, with multiple dominations throughout their inpatient stay and no return to their initial state, whereas others showed relatively few movements and some individuals returned close to their initial microbiome composition.

Many questions remain, such as why certain individuals return to their initial profile within the monitoring timeframe, whereas others end up far from their initial profile. Could there be a more intrinsic element to this such as prior clinical status upon arrival, or is it solely due to treatment and antimicrobial administration? It may be that a dominant microbiome at the beginning of the treatment may be more susceptible to further modulation. Lavelle *et al* also suggest an individualised response to antimicrobials by germ-free mice colonised with different donor microbiotas and Raymond *et al* found a similar effect in adults[160,161].

Assuming that there is a single equilibrium value with a functionally optimal microbiome for each individual, an insult, such as a course of antibiotics, may lead to a perturbation and thus a shift away from this optimum. Studies have found that such shifts are often transient and although adults broadly return to their original state after a single perturbation such controlled studies have not been undertaken in children[75]. It is likely that transitioning to a different state with reduced diversity may increase the risk of colonization and overgrowth of pathogenic species. At present, no controlled studies replicating continuous perturbations in human subjects, akin to those experienced by HSCT patients, have been undertaken.

It may be pertinent to hypothesise that with the number of continuing insults during HSCT, this single equilibrium value is gradually shifted away to a new equilibrium value and eventually the landscape itself is altered[74,162]. It is unclear whether this could lead to irreversible change and whether this would be detrimental to host homeostasis, however it would be of interest to follow these individuals several years post-HSCT for this purpose. Palleja *et al* showed that healthy adults are resilient to a short course of broad-spectrum antibiotics as they returned to near-baseline composition within 1.5 months[75]. Despite this, certain taxa remained undetectable 6 months post-treatment. It is therefore not surprising that the majority of patients in this study did not return to their initial microbiota status given the observation period and repeated antimicrobial administrations throughout this time. Additionally, as long-term stability of the gut microbiota is thought to be maintained by the action of restoring forces that maintain its state within a certain range, it is possible that HSCT recipients may be lacking these forces in some way.

Although the landscape provides a useful visualisation of the shifts in the intestinal microbiota, it should not be generalised to the whole of the paediatric HSCT population. It would be of great interest to collectively profile patients from different paediatric transplant centres worldwide in order to document detailed differences in microbial changes, highlighting the most useful intervention for the whole paediatric HSCT population.

4.3.4 Community state types

In order to identify distinct bacterial community patterns within the population the data was partitioned into three CSTs, each with a varied microbial composition. Transition models revealed time-dependent patterns of the CSTs such as transition between CST2 and CST3 to CST1 becoming less common with time, with no such transitions observed after week 4. This suggests that there may be a critical time-period after transplantation, during which the microbiome is able to return towards a healthier state. The microbiome around this period may potentially be more amenable to intervention.

Both CST2 and CST3 appear to have high self-transition probabilities, meaning both CSTs are more stable than CST1. This is similar to another study in adults, which suggests that the resilience of a biodiverse state is low during the post-HSCT period, with an observed self-transition probability of 49%, in comparison to our observation of 40%[79]. Although certain dominations are specific to the current cohort, the dynamics of the microbiome appear broadly similar to that seen in adults. The recovery of gut microbiota and the impact of the procedure on paediatric patients however is likely to be different to that in adults. It is possible that age and immaturity of the gut microbiota may play a part in one's resilience and ability to regain at least some of the diversity seen prior to transplantation.

Importantly, we found that CST3 was associated with a higher risk of viraemia (HR = 2.07; p value = 0.012). CST3 is a state with a complex composition, and on further analysis, we were unable to identify any specific taxa responsible for this association, although both *Enterococcus* and *Enterobacteriaceae* were close to significance. This is potentially due to small sample size as well as the complexity of the state types, i.e. none of the taxa are exclusive to one state type, e.g. *Enterococcus* is present in both CST3 and CST2. Thus, a higher risk of viraemia with CST3 could be a result of a cumulative effect of several taxa or it may be unresolvable with such small numbers. Similarly, a proportion of *Enterobacteriaceae* remain unresolved at the genus level, making clearer elucidation difficult[163].

The composition of CST3 is complex and includes several *Proteobacteria* including *Klebsiella, Escherichia* and *Enterococcus.* Previous work has linked domination of certain taxa with adverse clinical outcomes, such as domination by *Enterococcus* and an increased risk of bacteraemia with vancomycin-resistant *Enterococcus* or Gram-negative bacteria[79,93]. It is unclear how *Enterococcus* and/or other taxa may contribute towards an increased risk of viraemia; this could be due to indirect action via delayed immune re-constitution and/or through re-activation and/or delayed/impaired response to re-activation. The commensal microbiota (through MAMPs engaging pattern-recognition receptors) as well as derived metabolites are also known to impact and generate optimal innate and adaptive immune responses important for controlling systemic viral infections as well as have an impact on viral-specific CD8 T cell memory in a murine model infected with CMV[164,165]. Additionally, CST3 could simply be a marker of bad gut health such as damage to the colonic mucosa and domination. It would be of great interest to study viral reactivation in respect to the intestinal microbiota shifts in future studies.

4.3.5 Limitations

This work has several limitations. The population studied here is heterogeneous in terms of the underlying disease and treatment. Thus, findings may not be easily generalizable to the entire paediatric HSCT community. As a result, it may be challenging to find a generic interventional approach to modulating the microbiome prior to assessing the wider paediatric HSCT population.

Although the largest study for this population to date, the sample size, given the heterogeneity of the patients, is still relatively small, which prevents us from undertaking more in depth analysis including stratification of the patients by their underlying condition or assessment of the effects of antimicrobial administration on the gut microbiota. It would be of interest to collect further information on patient antimicrobial use prior to admission and to also investigate the effects of other medication on the microbiota.

Similarly, we aimed to collect samples approximately weekly therefore the microbiota data is limited in this respect. This means that we do not know the state of the intestinal microbiota between the specimens and therefore it is

plausible that we were unable to fully capture all dominations and thus all CSTs, especially as we could not sequence all collected samples. Additionally, the cut-off of 30% used to determine dominance was based on an adult cohort and could be proportionally adjusted to reflect the high dominance levels observed in our cohort[79,81].

In terms of methodology, generating clusters is not entirely objective as one eventually has some input in choosing the optimal number. Additionally, although three clusters appeared to be best for the dataset, they are nonetheless not optimal and therefore some sample misclassification is possible. Other methods, such as Dirichlet multinomial mixtures and dendrograms may be useful in confirming the clustering findings[166].

Given the link between antibiotic administration and intestinal microbiota, it would be of interest to assess their impact on the gut microbiota in this cohort. Despite this, the study was not powered for this as all patients were on varying combinations of antimicrobials throughout the inpatient stay and the sample numbers were relatively small for this type of analysis.

Additionally, the models utilising drug data did so on the day of their administration. It may be more useful to incorporate a lag time into the models in the future. Some of the commonly administered antimicrobials such as ciprofloxacin have short half-lives, and therefore lag time may not be necessary. Additionally, patients do receive the drugs continuously, therefore incorporating useful lag times may be particularly challenging. Modelling the long-term impact of antibiotics on the microbiome could, however, help to influence the use of antibiotics in clinical management.

Finally, 16S rRNA sequencing is known not to be entirely representative of the 16S gene and is unable to reliably resolve the taxa down to species level. As previously mentioned (Chapter 3), it is also biased as a result of PCR and due to varying 16S rRNA copy numbers. Additionally, we focus on abundances, which does not give us an idea of absolute numbers. Finally, stool may not accurately reflect the overall gut microbiota composition, as the mucosal microbiota, which reflects the taxa that interact with the immune system more closely, is known to

be different to the luminal, i.e. stool, microbiota which likely reflects the non-adherent and the shed mucosal bacteria[56].

Using other omics approaches as well as other methods including qPCR and functional assays will be essential to recapitulate the current findings and take the work forward.

4.4 Summary

In this chapter, the dynamics of the intestinal microbiota throughout paediatric HSCT were investigated. Taxonomic profiling of the cohort revealed shifts in the intestinal microbiota throughout transplantation, with taxa dominance, at times as high as 95% relative abundance. The majority of patients had highly variable taxonomic profiles throughout their inpatient stay with varying dominations and did not recover their pre-transplant diversity before discharge. A small proportion of patients did broadly recover, which highlights the potentially individual responses to treatment and drug administrations throughout transplantation.

Transitional models revealed time-specific transitions between CSTs, with the majority of patients transitioning to seemingly less healthy CSTs in the first few weeks post-transplantation, which emphasises the severity of the procedure on the intestinal microbiota. CST3 associated with a higher risk of viraemia, however, we were unable to identify particular taxa driving this association.

In terms of gut microbiota dynamics, the paediatric HSCT cohort was similar to that recorded in adult HSCT; however, some differences such as domination by particular taxa may be specifically relevant to this cohort. These findings warrant further paediatric studies in order to bring our understanding of the role of microbiota throughout HSCT in paediatric patients in line with an ever-growing amount of information on the topic in adult patients. A better understanding of the dynamics that render a complex gut microbiota permissive to pathogens and the dynamics of its recovery has the potential to shape antibiotic management or identify useful periods for intervention.

Chapter 5- Intestinal microbiota and predictive biomarkers in paediatric HSCT

5.1 Introduction

A biomarker is defined as a molecule by which a physiological or pathological process or a disease can be identified. Prognostic biomarkers are particularly useful as they can be used as an early indication of the event of interest, which allows one to make clinically relevant decisions early. The non-invasive and the non-limiting nature of the gut microbiota sampling using stool samples and the feasibility of sequencing makes the sample matrix ideal for biomarker discovery.

Although survival has improved significantly, HSCT still carries significant risks. A considerable proportion of patients suffer from adverse clinical outcomes including viral, bacterial and fungal infections as well as GvHD. Growing evidence suggests that the gut microbiota is involved in the clinical course of HSCT in both adult and paediatric populations, thus studies have focused on the gut microbiota as a potential source of biomarkers.

Despite these advances, biomarker studies in a paediatric HSCT setting are rare. Hakim *et al* investigated an ALL cohort undergoing chemotherapy[86]. They stratified gut microbiota samples into five clusters and found that a cluster with a high abundance of *Enterococcaceae* and *Streptococcaceae* predicted a higher risk of subsequent diarrhoeal illness. Additionally, an abundance of more than 0.01% of *Proteobacteria* was associated with an increased incidence of febrile neutropenia. Nearing *et al* also found that lower phylogenetic diversity is associated with higher prevalence of subsequent infection and that *Faecalibacterium*, *among* other taxa, is a strong classifier of infection[167]. Finally, Biagi *et al* found that individuals who went on to develop GvHD had lower *Blautia* and higher *Fusobacterium* abundances at baseline than those who did not[153].

5.1.1 Aims

In this chapter I investigate the gut microbiota at several timepoints throughout HSCT, namely baseline and pre-engraftment, with a view to identifying biomarkers for adverse clinical outcomes in this population.

5.2 Methods

Methods for this chapter are detailed in Chapter 2.

5.3 Results

5.3.1 Baseline microbiota of patients and healthy controls

Prior to investigating whether the gut microbiota at baseline (a baseline sample was the first sample collected for an individual upon admission, providing it was collected prior to transplantation (range -44 to -1 days)) may be a useful biomarker of clinical outcomes, we explored the taxonomic composition of the initial samples and compared them to healthy controls (HC; unmatched) (Figure 5.1).

HC (median age 10 years) exhibited a similar composition with obligate anaerobes *Bacteroidaceae, Lachnospiraceae* and *Ruminococcaceae* comprising the main taxa. The patient baseline samples in contrast showed greater variation in composition. Whilst a few samples were similar to HC, some were already dominated by a single taxon, with *Enterobacteriaceae, Enterococcaceae* and *Bacteroidaceae* being the most predominant. There was also a notable absence of *Ruminococcaceae* in most patient samples. Even at baseline, patient samples had a dissimilar taxonomic composition to that of HC.

Although patients undergoing auto-HSCT were not included in the biomarker exploration in order to reduce the complexity of the cohort, for completeness we include a taxonomic plot of the baseline samples for the auto-HSCT patients (Figure 5.3). Baseline samples showed the presence of taxa similar to those seen in HC, it was however interesting to observe a lack of *Ruminococcaceae* and taxa domination in certain samples, features similar to allo-HSCT (Figure 5.1).

Alpha diversity was significantly greater in HC compared to patient samples in both allogeneic and autologous HSCT ($p \leq 0.001$, Mann-Whitney; Figures 5.2/5.3).

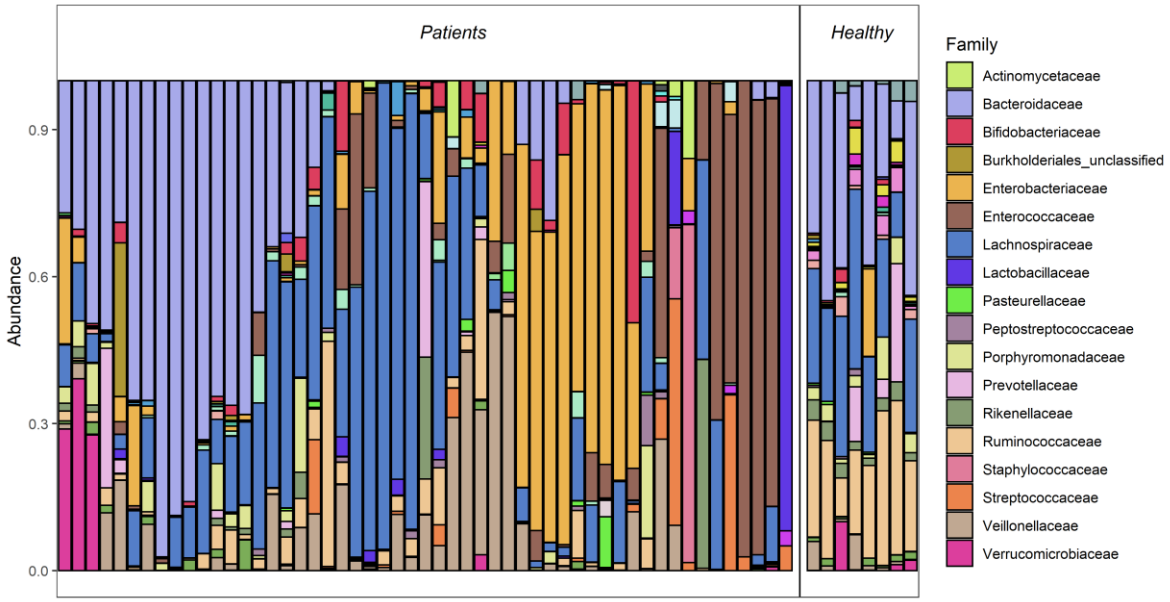


Figure 5.1 Family level taxonomic plot of patient baseline samples and unmatched HC. Patients refers to patients undergoing allogeneic transplantation. Only taxa with relative abundance of >10% are labelled.

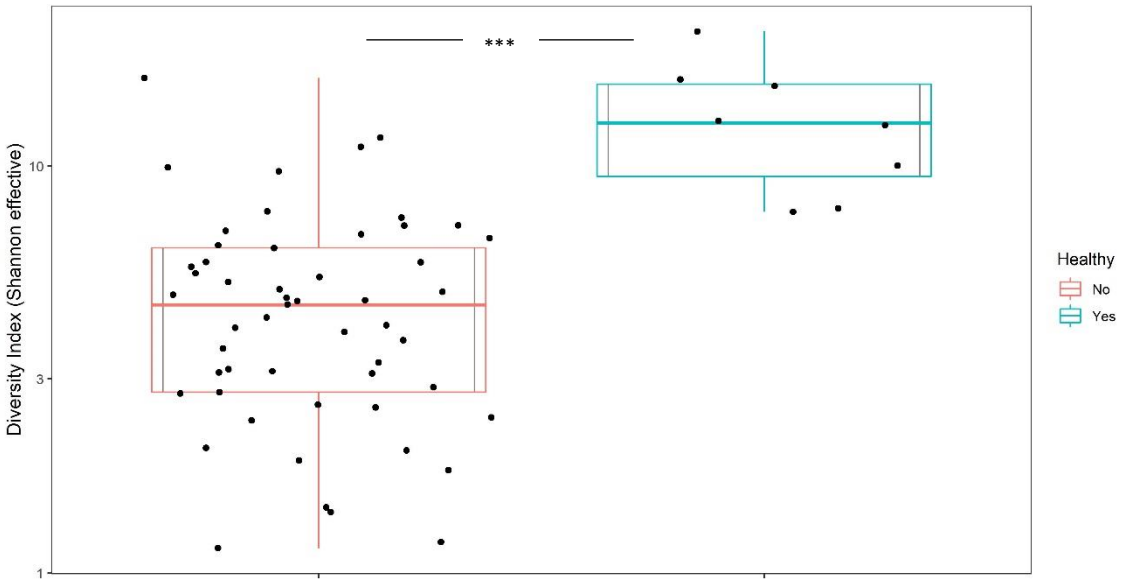


Figure 5.2 Log10 transformed alpha diversity (Shannon effective) of allogeneic-HSCT patient baseline samples and unmatched healthy controls. ***<0.001 Mann-Whitney test

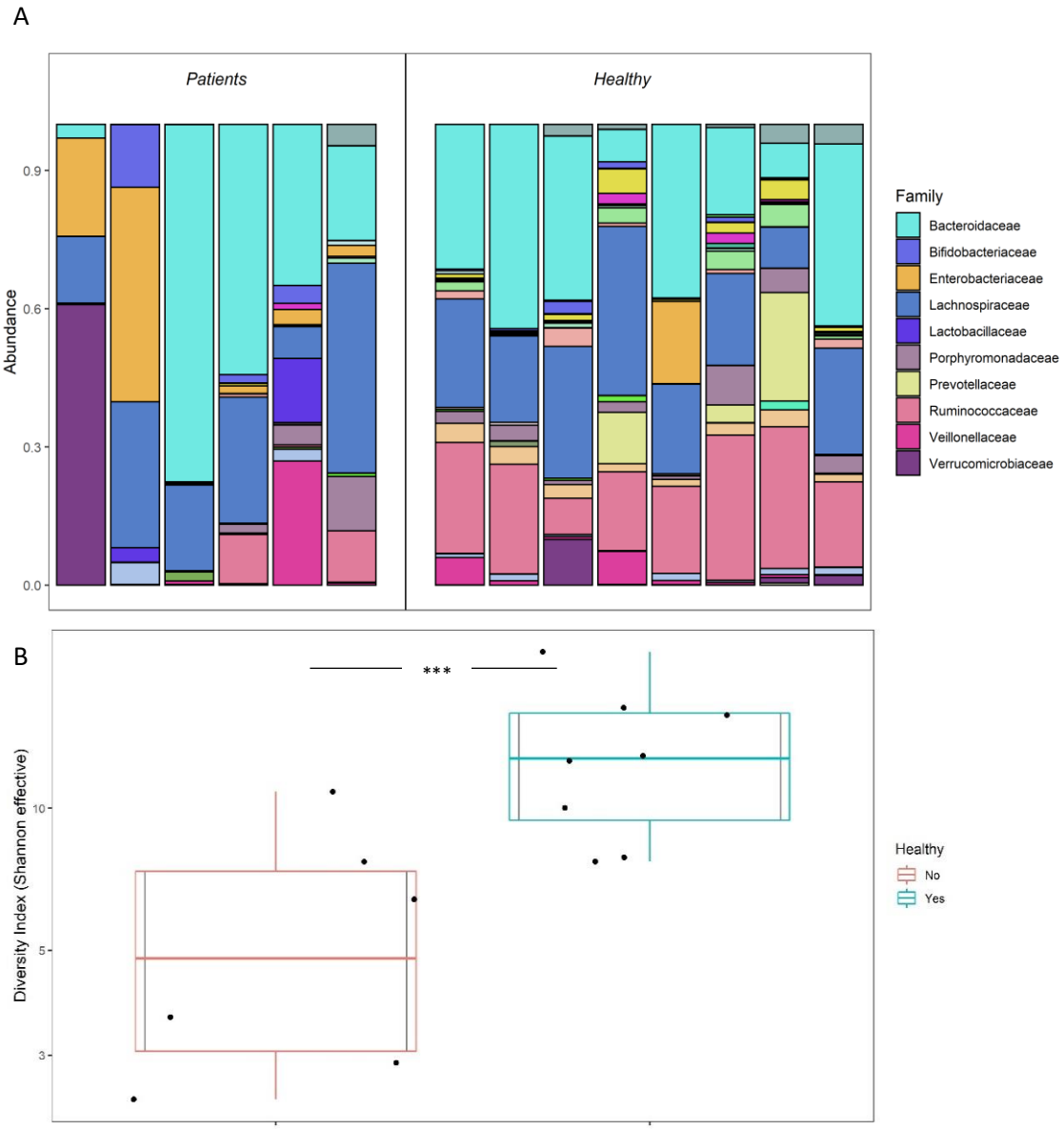


Figure 5.3 A) Family level taxonomic plot of patient baseline samples and unmatched healthy controls. Unhealthy refers to patients undergoing autologous transplantation. Only taxa with the relative abundance of >10% are labelled. **B)** Log10 transformed alpha diversity (Shannon effective) of autologous-HSCT patient baseline samples and unmatched healthy controls. ***<0.001 Mann-Whitney test

5.3.2 Baseline microbiota and clinical outcomes

Demographics of the baseline cohort

Table 5.1 details the demographics of the baseline cohort sub-classified by the clinical outcomes: overall survival, GvHD (≥ grade II) and viraemia. We collected 53 baseline samples in this cohort of which 51 samples were used for the GvHD comparison (two patients died immediately after transplantation).

Table 5.1 Demographics of the baseline cohort

Outcome: Survival	n = 53	Survival: Yes (N = 40)	Survival: No (N = 13)	P-value
Shannon_effective				0.09
min	1.15	1.45	1.15	
max	16.46	16.46	6.93	
mean (sd)	4.88 ±2.96	5.28 ±3.14	3.66 ±1.95	
Age_under_2 n(%)				0.71
Yes	11 (21)	9 (22)	2 (15)	
No	42 (79)	31 (78)	11 (85)	
Diagnosis n(%)[1]				1
Yes	30 (57)	23 (58)	7 (54)	
No	23 (43)	17 (42)	6 (46)	
Sex_Female n(%)				0.19
Yes	20 (38)	13 (32)	7 (54)	
No	33 (62)	27 (68)	6 (46)	
Outcome: Viraemia	n = 53	Viraemia:No (N = 12)	Viraemia: Yes (N = 41)	P-value
Shannon_effective				0.04
min	1.15	2.55	1.15	
max	16.46	16.46	11.14	
mean (sd)	4.88 ±2.96	6.42 ± 4.27	4.43 ±2.33	
Age_under_2 n(%)				0.69
Yes	11 (21)	3 (25)	8 (20)	
No	42 (79)	9 (75)	33 (80)	
Diagnosis n(%)				1
Yes	30 (57)	7 (58)	23 (56)	
No	23 (43)	5 (42)	18 (44)	
Sex_Female n(%)				
Yes	20 (38)	4 (33)	16 (39)	1
No	33 (62)	8 (67)	25 (61)	
Outcome: GvHD	n=51	GvHD: No (N = 30)	GvHD: Yes[2] (N = 21)	P-value
Shannon_effective				0.07
min	1.15	1.19	1.15	
max	16.46	16.46	9.93	
mean (sd)	4.87±2.99	5.07 ± 3.42	4.57 ± 2.22	
Age_under_2 n(%)				1
Yes	10 (20)	5 (16)	5 (25)	
No	41 (80)	26 (84)	15 (75)	
Diagnosis n(%)				1
Yes	29 (57)	18 (58)	11 (55)	
No	22 (43)	13 (42)	9 (45)	
Sex_Female n(%)				1
Yes	19 (37)	10 (32)	9 (45)	
No	32 (63)	21 (68)	11 (55)	

[1]Diagnosis refers to malignant and non-malignant.[2]GvHD is classed as grades II or above. sd-standard deviation. Fisher's exact test and a T-test were used for categorical and numerical variables respectively.

Taxonomies by clinical endpoints

We initially explored taxonomies of baseline samples by clinical outcomes at the genus level (Figure 5.4). Family level taxonomic plots are detailed in the Appendix (Figure A7). The samples appear highly variable and no clear taxonomic patterns between the clinical outcomes emerge.

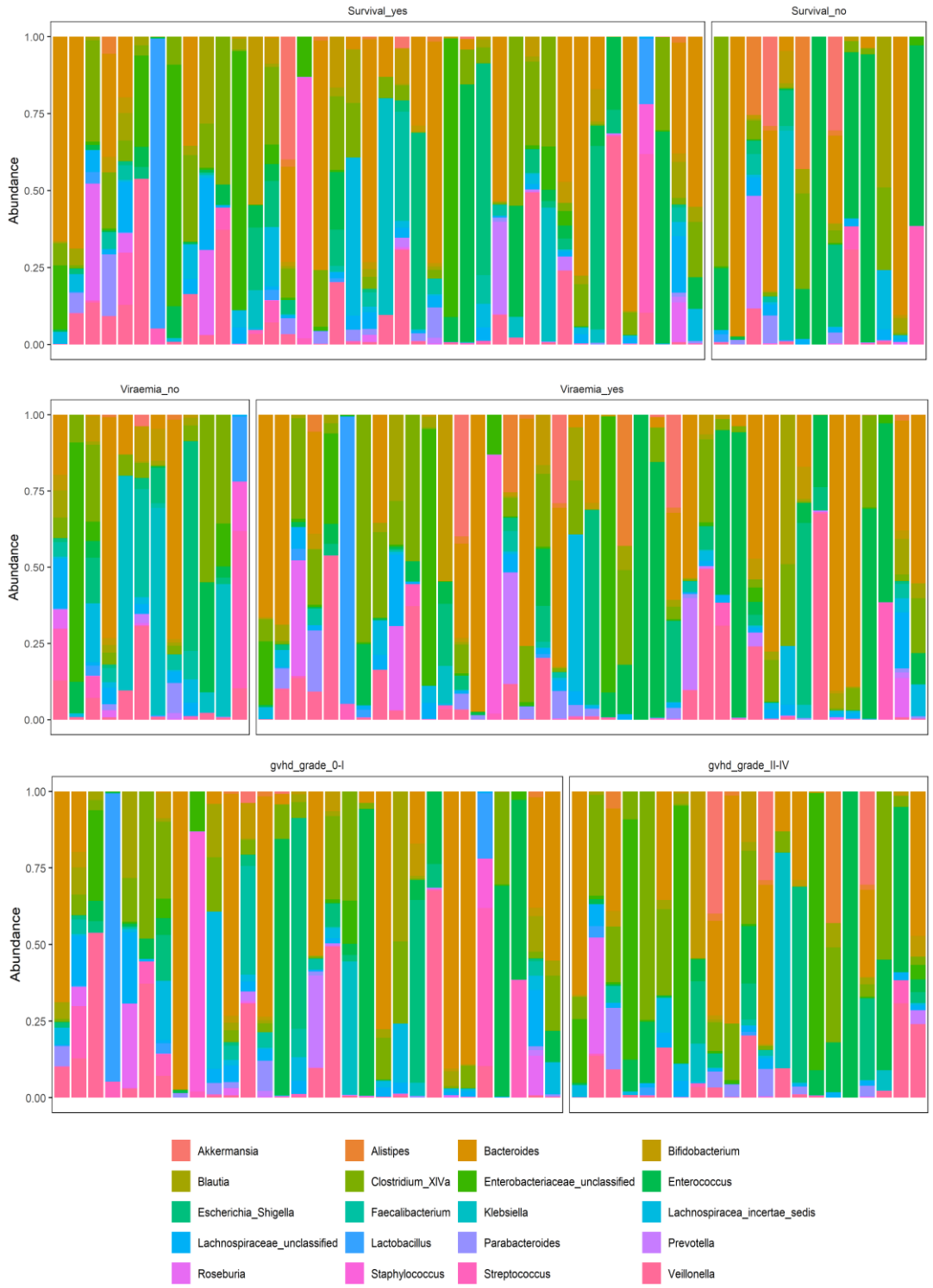


Figure 5.4 Genus level taxonomic plot comparing baseline samples of patients with varying clinical outcomes

Alpha diversity at baseline

Following this, we investigated alpha diversity between clinical outcomes (Figure 5.5). Baseline alpha diversity was lower in patients who did not survive in comparison to those who did, however this trend did not reach statistical

significance (Figure 5.5 A). Similarly, alpha diversity was variable at baseline and no major differences were observed between patients who developed clinically significant GvHD (grade ≥II) and those who did not (grade ≤I) (Figure 5.5 C). Alpha diversity displayed varying trends when the grades were plotted separately (Figure 5.5 D). Alpha diversity was significantly lower in baseline samples of patients who did go on to develop viraemia, in comparison to those who did not (Figure 5.5 B).

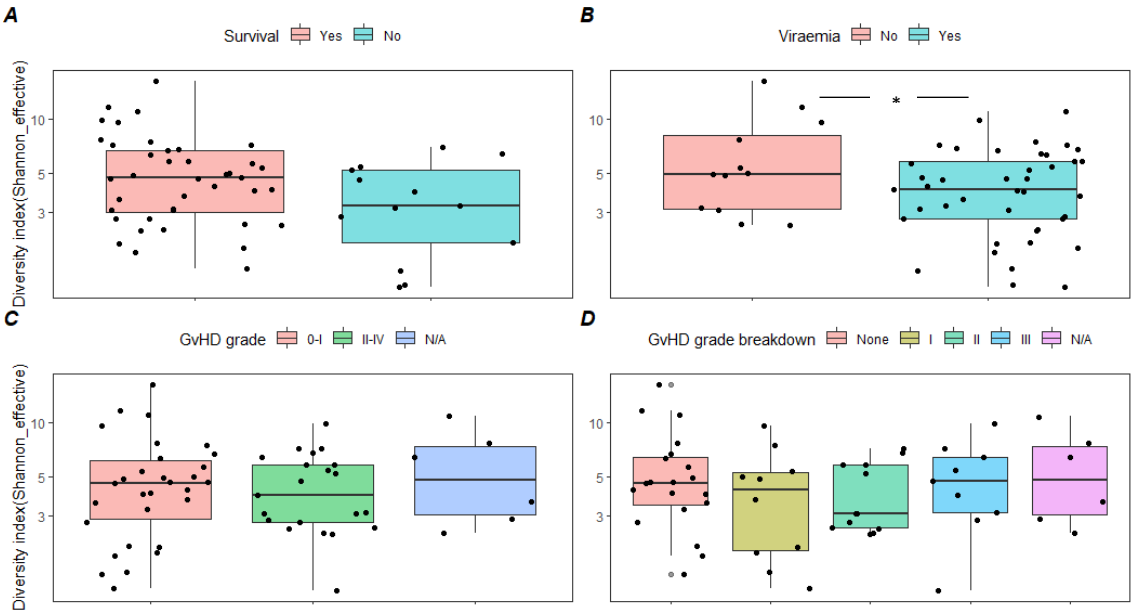


Figure 5.5 Log10 transformed alpha diversity (Shannon effective) between baseline samples with varying clinical outcomes **A)** Survival **B)** Viraemia **C)** GvHD (clinically significant vs clinically insignificant) **D)** GvHD (all grades). *<0.05; Mann-Whitney test. N/A- autologous HSCT recipients.

Differentially abundant features at baseline

Next, we compared baseline samples in order to see if there were any discriminating taxa between the clinical outcomes. Table 5.2 details the taxa (genera and families) found to be differentially abundant between the samples with different outcomes. Most taxa appeared to be found in samples with no eventual adverse outcome, such as *Veillonella* in samples of patients who survived and *Clostridium_XVIII* in samples of patients with no viraemia throughout their inpatient stay.

Table 5.2 Differentially abundant taxa at baseline

Feature	Enriched in	LDA* effect size	P-value
Outcome: Survival			
Veillonella	Yes	4.74	0.02
Enterobacteriaceae	Yes	4.82	0.02
Veillonellaceae	Yes	4.61	0.02
Peptostreptococcaceae	Yes	4.78	0.04
Outcome: Viraemia			
Dysgonomonas	No	3.99	0.0001
Clostridium_XVIII	No	3.65	0.004
Robinsoniella	No	3.40	0.01
Holdemania	No	4.16	0.01
Faecalibacterium	No	4.28	0.01
Neisseria	Yes	3.18	0.02
Turicibacter	No	3.18	0.04
Outcome: GvHD			
Klebsiella	Yes	4.29	0.03

*Linear discriminant analysis

Selecting optimal cut-offs

Having found certain taxa to be differentially abundant in samples with and without eventual adverse clinical outcomes, we next wished to find optimal cut-offs, using ROC. Table 5.3 details optimal cut-offs for each taxon, which maximise the sensitivity and specificity of detection. Only significant taxa (in comparison to an AUC of 0.5 and thus of no predictive value) are shown. Taxa with an AUC of >0.65 and a p-value of ≤0.05 were considered to be of predictive discriminative value[93]. Details of all taxa are noted in the appendix (Table A14). ROC curves of alpha diversity (Shannon effective) for all outcomes were created; however, they were not of any predictive value. Analyses revealed four potentially predictive taxa for overall survival, one for GvHD and two for viraemia.

Table 5.3 Optimal cut-offs for significant taxa at baseline

Taxa	AUC	Cut-off	Sensitivity	Specificity	P- value
Outcome: Survival					
Veillonella	0.72	0.0005	0.62	0.8	0.02
Veillonellaceae	0.72	0.01	0.87	0.58	0.02
Enterobacteriaceae	0.72	0.002	0.62	0.83	0.02
Peptostreptococcaceae	0.69	0.001	0.92	0.53	0.04
Outcome: Viraemia					
Clostridium_XVIII	0.7693	0.0024	0.88	0.73	0.007
Feacalibacterium	0.7136	0.0409	0.98	0.46	0.03
Outcome: GvHD					
Klebsiella	0.6685	0.000034	0.65	0.74	0.04

Taxa associations with clinical outcomes

Finally, we explored if the taxa (at an optimal cut-off chosen above) correlated with clinical outcomes (GvHD and viraemia) using logistic regression. Clinical characteristics such as age, sex or underlying diagnosis were not associated with any of the clinical outcomes as the null models were preferable (based on AIC criterion).

Table 5.4 details the final regression models for both a) viraemia and b) GvHD. *Klebsiella* was associated with a higher risk of developing grade II or higher GvHD with an odds ratio (OR) of 3.9 and a probability of 0.79 (95% CI- 1.2-13.4). Following stepwise elimination, *Clostridium_XVIII* was associated with a lesser risk of developing a viraemia (OR-0.05; 95% CI- 0.009-0.25; Probability-0.05). ROC curves for these taxa are detailed in the Appendix (Figure A8).

Table 5.4 Logistic regression models to predict and a) viraemia and b) GvHD at baseline

a) **Viraemia**	Estimate (Standard error)	Pr(>|z|)
(Intercept)	2.46(0.60)	<0.001
Clostridium_XVIII	-2.93(0.83)	<0.001
b) **GvHD**	Estimate (Standard error)	Pr(>|z|)
(Intercept)	-1.1(0.44)	0.01
Klebsiella	1.36(0.61)	0.03

We additionally investigated the taxa that were differentially abundant for survival using Kaplan-Meier (KM) curves, however none were significant (Appendix, Figure A9).

Figure 5.6 details the distribution of relative abundances of the taxa found to be associated with either GvHD or viraemia in all baseline samples. In both cases, especially with *Klebsiella,* the range of relative abundances among the patient samples was highly variable. Although a higher proportion of *Klebsiella* at baseline associated with a higher probability of GvHD, a proportion of patients did not show the presence of *Klebsiella* (Table 5.4 and Figure 5.6 B).

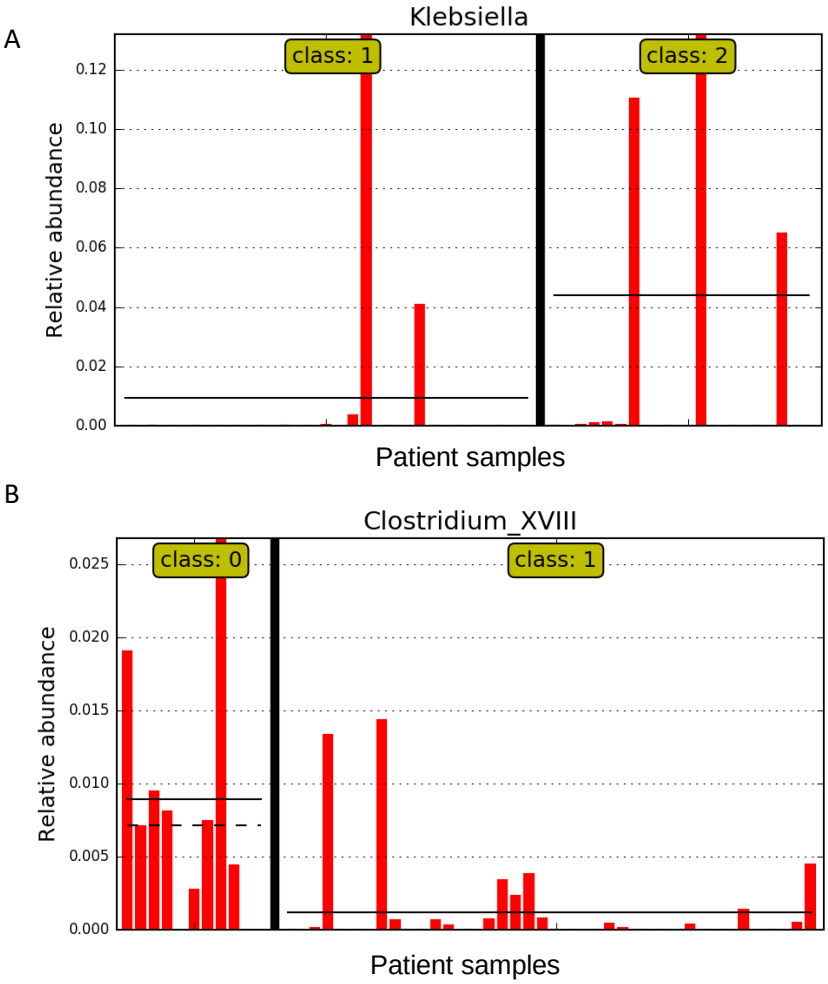


Figure 5.6 Relative abundance of taxa found to be linked to **A)** GvHD (class 1: grade ≤ I GvHD; class 2: grade ≥ II GvHD) and **B)** Viraemia (class 0: no viraemia). Vertical lines indicate the median relative abundance.

5.3.3 Pre-engraftment microbiota and clinical outcomes

Having identified potential associations between taxa and clinical outcomes at baseline, we wondered whether a timepoint prior to engraftment may serve as a useful timepoint for biomarker discovery for predicting future clinical outcomes.

Demographics of the pre-engraftment cohort

Table 5.5 details the demographics of the pre-engraftment cohort sub-classified based on clinical outcome- overall survival, GvHD (≥ grade II) and viraemia. A pre-engraftment sample was a sample closest to neutrophil engraftment post-transplantation for each patient (range 0-60 days). We collected 48 pre-engraftment samples in this cohort.

Table 5.5 Demographics of the pre-engraftment cohort

Outcome:Survival	n = 48	Survival:Yes (n=37)	Survival:No (n=11)	P-value
Shannon_effective				0.69
min	1	1	1.09	
max	19.48	13.19	19.48	
mean (sd)	3.32 ± 3.68	3.20 ± 3.10	3.72 ± 5.35	
Age_under_2 n(%)				0.25
Yes	12 (25)	11 (30)	1 (9)	
No	36 (75)	26 (70)	10 (91)	
Diagnosis n(%)				0.51
Yes	26 (54)	19 (51)	7 (64)	
No	22 (46)	18 (49)	4 (36)	
Sex_Female n(%)				0.30
Yes	19 (40)	13 (35)	6 (55)	
No	29 (60)	24 (65)	5 (45)	
Outcome:Viraemia	n = 48	Viraemia:No (n=14)	Viraemia:Yes (n=34)	P-value
Shannon_effective				0.67
min	1	1	1	
max	19.48	13.19	19.48	
mean (sd)	3.32 ± 3.68	3.67 ± 4.30	3.17 ± 3.45	
Age_under_2 n(%)				0.29
Yes	12 (25)	5 (36)	7 (21)	
No	36 (75)	9 (64)	27 (79)	
Diagnosis n(%)				0.53
Yes	26 (54)	9 (64)	17 (50)	
No	22 (46)	5 (36)	17 (50)	
Sex_Female n(%)				1
Yes	19 (40)	5 (36)	14 (41)	
No	29 (60)	9 (64)	20 (59)	
Outcome:GvHD	n = 48	GvHD: No (n=28)	GvHD: Yes[1] (n=20)	P-value
Shannon_effective				0.69
min	1	1.01	1	
max	19.48	11.36	19.48	
mean (sd)	3.32 ± 3.68	2.59 ± 2.39	4.34 ± 4.84	
Age_under_2 n(%)				1
Yes	12 (25)	7 (25)	5 (25)	

No	36 (75)	21 (75)	15 (75)	
Diagnosis n(%)				1
Yes	26 (54)	15 (54)	11 (55)	
No	22 (46)	13 (46)	9 (45)	
Sex_Female n(%)				1
Yes	19 (40)	9 (32)	10 (50)	
No	29 (60)	19 (68)	10 (50)	

[1]GvHD is classed as grades II or above. sd -standard deviation. Fisher's exact test and a T-test were used for categorical and numerical variables respectively.

Taxonomies by clinical endpoints

As observed with baseline samples, taxonomic composition at engraftment was highly variable (Figure 5.7). Domination by a single taxon was more common at this timepoint in comparison to baseline (Figure 5.4). Samples from patients with viraemia showed greater proportions of *Enterobacteriaceae* and *Enterococcaceae.* Genus level taxonomic plots are detailed in the Appendix (Figure A10).

Alpha diversity at pre-engraftment

Following this, we investigated alpha diversity between clinical outcomes (Figure 5.8). Alpha diversity was very variable at the pre-engraftment timepoint. No significant differences in alpha diversity were observed, although those with a viraemia outcome showed a trend for higher alpha diversity, an opposing trend to that observed with baseline samples.

Differentially abundant features at pre-engraftment

We again compared pre-engraftment samples in order to see if there are any discriminating taxa between the clinical outcomes. Figure 5.9 details the taxa (genera and families) found to be differentially abundant between the samples with different outcomes. Unlike at baseline, taxa predictive of both adverse clinical outcome and no adverse outcome were found. Table A15 (Appendix) details the p-values and LDA effect size for all differentially abundant taxa.

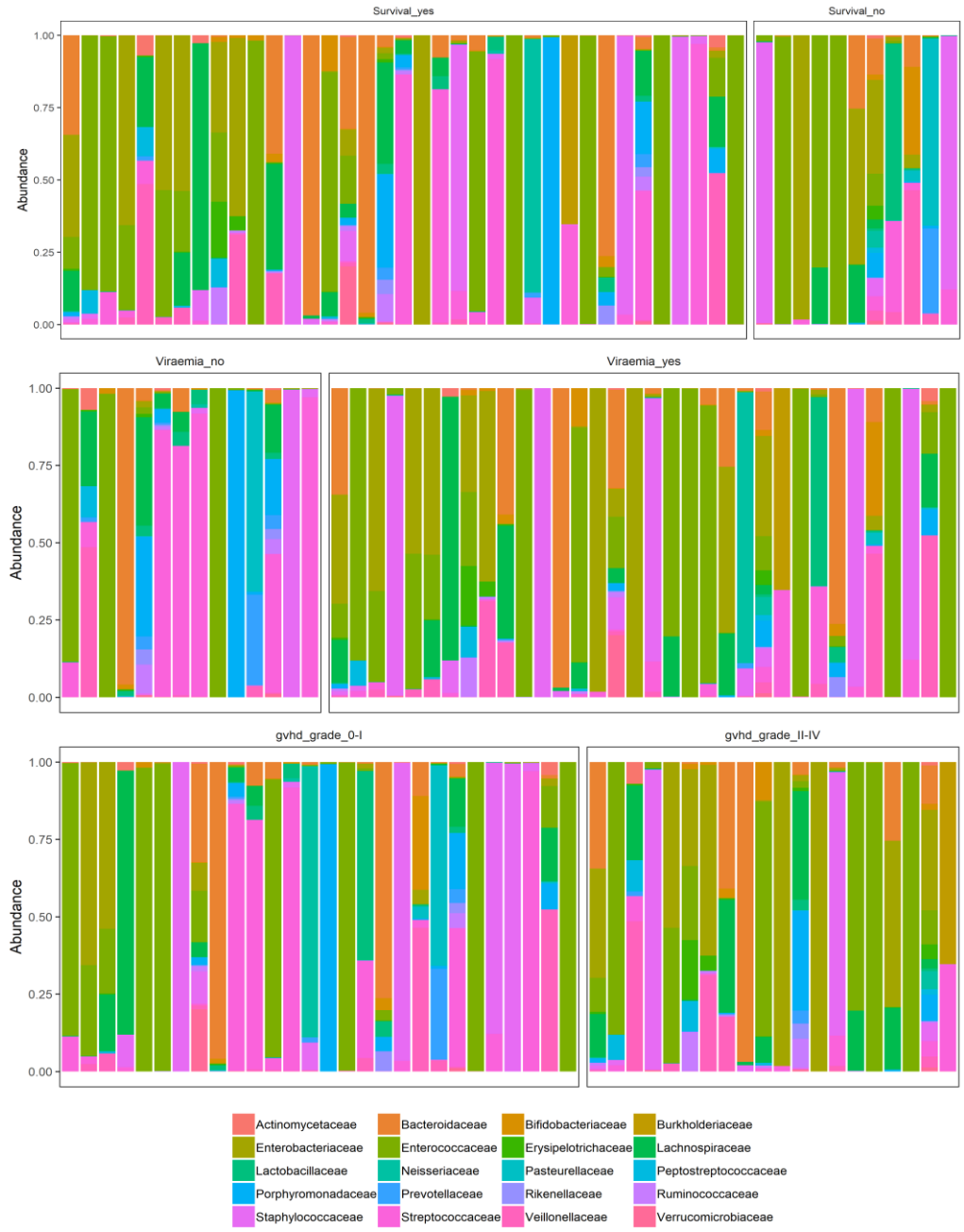


Figure 5.7 Family level taxonomic plots comparing pre-engraftment samples of patients with varying clinical outcomes

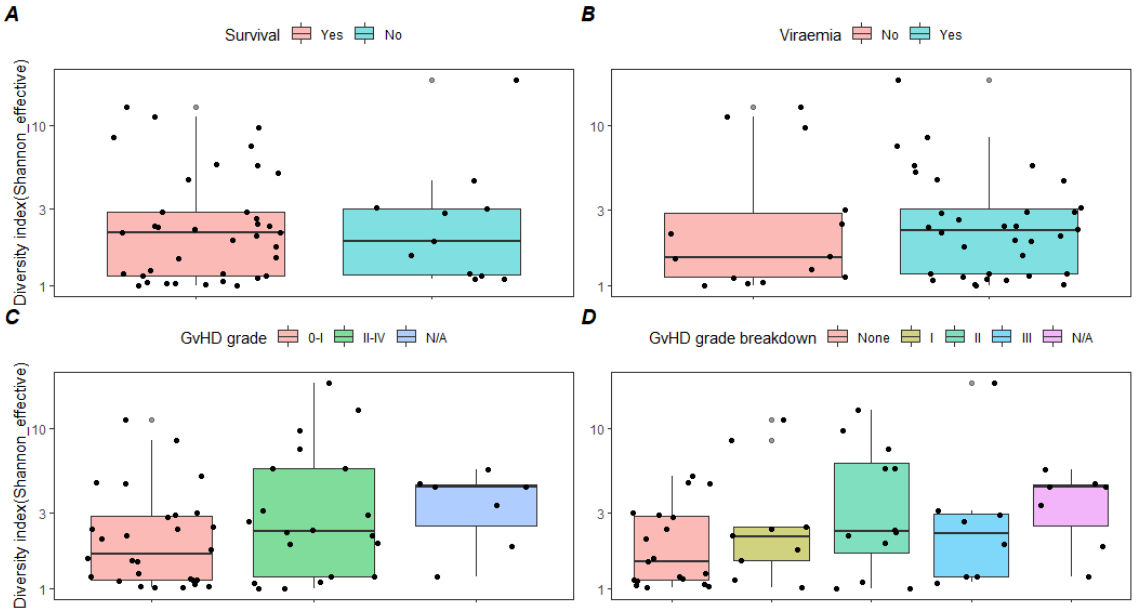


Figure 5.8 Alpha diversity (Shannon effective) between pre-engraftment samples with varying clinical outcomes

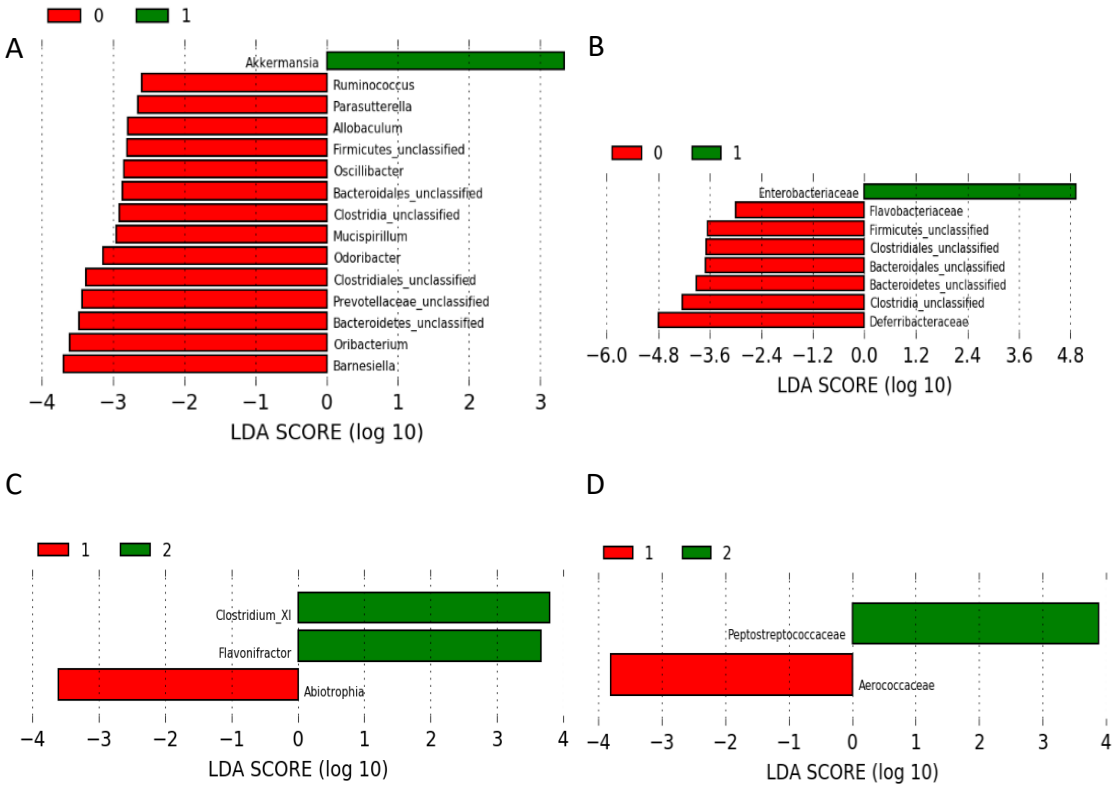


Figure 5.9 Differentially abundant **A)** genera and **B)** families for viraemia and **C)** genera and **D)** families for GvHD in pre-engraftment samples. **A/B** – 0/1- no viraemia/viraemia; **C/D** – 1/2- no GvHD/GvHD.

Selecting optimal cut-offs

Having found certain taxa to be differentially abundant in samples with and without adverse clinical outcomes, we again aimed to find optimal cut-offs, using ROC curves. Only a single taxon, *Enterobacteriaceae*, was found to be of predictive value (AUC=0.74; p-value=0.01). Details of all taxa and the ROC curve for *Enterobacteriaceae* are noted in the appendix (Table A16; Figure A11).

Taxa associations with clinical outcomes

Finally, we wanted to see if *Enterobacteriaceae*, at an optimal cut-off chosen above, correlated to viraemia using logistic regression. Clinical characteristics such as age, sex or underlying diagnosis as well as conditioning were not associated with viraemia as the null models were preferable.

Table 5.6 details the final regression model for viraemia. *Enterobacteriaceae* was associated with a higher risk of developing viraemia with an OR of 8.57 and a probability of 0.89 (95 CI- 1.95-60.89). Additionally, we were interested in any associations between antibiotics administered in the 10 days prior to the sample collection and viraemia. As ciprofloxacin is administered ubiquitously in this population it was not possible to include it in the model. Other antibiotic groups such as broad/narrow-spectrum beta-lactams, macrolides, aminoglycosides or glycopeptides were not significantly associated with the outcome, as the null model was preferable after stepwise regression (data not shown).

Table 5.6 Logistic regression models to predict viraemia at pre-engraftment

Viraemia	Estimate (Standard error)	Pr(>|z|)
(Intercept)	0.15(0.39)	0.70
Enterobacteriaceae	2.15(0.84)	0.01

Figure 5.10 details the distribution of relative abundances of *Enterobacteriaceae* in patient samples. Unlike with certain taxa at baseline, *Enterobacteriaceae* appears in most of the pre-engraftment samples.

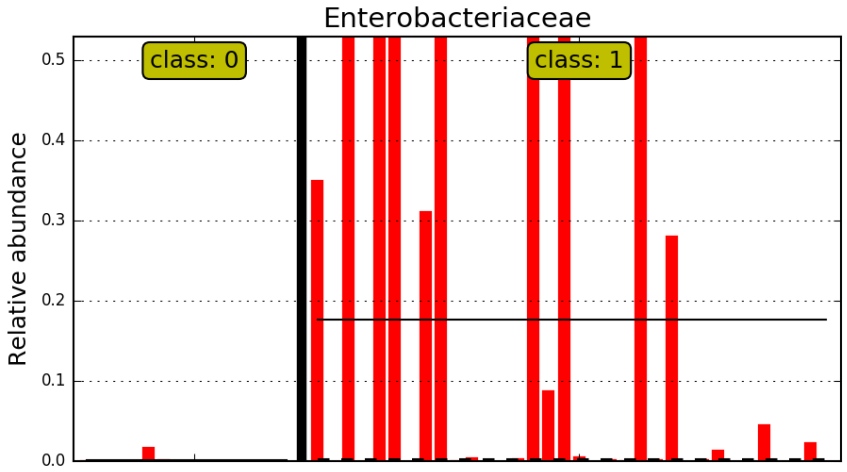


Figure 5.10 Relative abundance of *Enterobacteriaceae* in pre-engraftment samples of those patients who did not (class: 0) and did (class: 1) get viraemia throughout their inpatient stay. Vertical lines indicate the median relative abundance.

5.4 Discussion

In this chapter the composition of the gut microbiota at baseline and prior to engraftment was investigated with a view to finding microbial biomarkers for clinical outcomes including overall survival, GvHD and viraemia.

It was apparent that most baseline samples were dissimilar to HC. These initial samples were of variable composition, and their alpha diversity was significantly lower in comparison to HC. The loss of obligate anaerobes including *Ruminococcaceae* and *Lachnospiraceae* in comparison to HC as well as the observed domination by a single taxon was most likely due to previous exposure to antimicrobials, other drugs and prolonged exposure to a healthcare setting. This is an attractive hypothesis but requires confirmation in future studies. The heterogeneity of this patient group makes biomarker discovery in this population potentially challenging and an intervention relying on an initial healthy-like gut microbiota composition, such as auto-HSCT, less feasible.

The baseline timepoint was chosen as the ability to predict clinical outcomes at this point would be particularly useful in a clinical setting. Studies in adult allo-

HSCT have highlighted the relevance of the baseline microbiota in relation to clinical outcomes[153,168,169].

Firstly, we did not find any biomarkers for overall survival at either of the timepoints. This is perhaps not surprising as mortality has a complex aetiology, the outcome numbers were low, and we were unable to stratify the outcome. Several studies have found biomarkers such as <10% *Lachnospiraceae* at baseline to be associated with a higher risk of mortality post-HSCT as well as *Blautia* and higher diversity at engraftment with a lower risk of GvHD-related mortality, which appears to highlight gut health as beneficial for survival[88,93].

We were additionally interested in GvHD, as alterations in the intestinal microbiota are known to be associated with GvHD incidence and severity[170]. Despite this, it is not known whether the composition of the initial gut microbiota may be predictive of GvHD development.

One of the only studies that has investigated potential GvHD biomarkers at baseline in children found no differences in alpha diversity between patients who will go on to develop GvHD to those who will not, which is in line with our observations[153]. Additionally, they found an increase in *Fusobacterium*, a decrease in *Blautia* and a disturbed microbiome prior to HSCT in patients who will develop GvHD. Although we did not observe these genera to be differentially abundant in our cohort, we found *Klebsiella* to be predictive of GvHD development at baseline (Table 5.4). *Enterobacteriaceae,* including *Klebsiella,* are gut commensals known to cause BSI in the HSCT population and their expansion post-HSCT is often observed[87]. Although it is unclear whether *Klebsiella* could have a causative role in GvHD development, its presence at high levels prior to HSCT may contribute towards pathogen expansion post-transplantation. Several patients did also present with *Klebsiella* domination at baseline, which may be due to prior antimicrobial administration. We briefly looked into the distribution of *Klebsiella* in a recently published cohort of paediatric HSCT patients[96]. Although a potential biomarker in our cohort, patients in the other paediatric cohort appeared to have no *Klebsiella* in their stool, thus reinforcing the notion that biomarkers and cut-offs may be hospital-specific and not as easily generalizable to the whole paediatric population[96].

Despite this, multiple studies to date have investigated potential GvHD biomarkers in adult cohorts around the time of engraftment. Mancini *et al* report an increased risk of GvHD with a decreased diversity, whilst Golob *et al* find that oral *Firmicutes* and *Actinobacteria* positively associate with GvHD at engraftment[93,95]. The only two studies in children reported a decrease in anti-inflammatory *Clostridia* in patients with GvHD and Biagi *et al* observed an increase in *Bacteroides,* but no specific biomarkers were identified[33,153]. We did not find any biomarkers for GvHD at engraftment. This could be potentially due to *(a)* variability in the cohort and/or *(b)* sampling points and/or *(c)* sample size.

Finally, we were interested in investigating infectious outcomes such as bacteraemia and viraemia in this cohort, however there were too few cases of bacteraemia in this cohort to further investigate this outcome. We did however find *Clostridium_XVIII* at baseline to be associated with a lower risk of viraemia. Although no studies to date link *Clostridium_XVIII* to viraemia, the result is instinctive and is likely a reflection of good gut health, as *Firmicutes* are known for their beneficial effects in the gut such as via metabolite production[171]. BSIs are one of many transplant-related complications and several studies have linked gut microbiota and BSI development. Mancini *et al* found a link between >5% *Enterobacteriaceae* at baseline and an increased risk of sepsis[93]. Montassier *et al* found that adult patients who developed subsequent BSIs whilst undergoing chemotherapy had decreased diversity and a decreased abundance of several taxa such as *Barnesiellaceae, Faecalibacterium* and *Suterella* at baseline*,* all of which were found to be protective against the development of BSI[94]. In agreement to another paediatric study and the studies mentioned above, we also found certain beneficial taxa such as *Faecalibacterium* to be enriched at baseline in those who did not go on to develop viraemia[167]. It may therefore be useful to design a panel of taxa, reflective of good overall gut health in the future, which could be used to assess patients upon admission and may ultimately be informative for clinical management[152]. This has, for example, been previously achieved for colorectal cancer, whereby five taxa, predominantly from the *Bacteroidaceae, Lachnospiraceae* and *Clostridiaceae* families*,* improved the ability to predict adenomas in comparison to the commonly used faecal occult blood testing[172]

We also found *Enterobacteriaceae* at engraftment to be associated with a higher risk of viraemia. This is somewhat intuitive, as we found a link between CST3, which had a high relative abundance of *Enterobacteriaceae*, and viraemia (Chapter 4). It is unclear if the high relative abundance of *Enterobacteriaceae* at this time-point is an indication of an unhealthy-like microbiota and the development of domination, since *Enterobacteriaceae* domination is a common event in our cohort, or whether this taxon is causally linked to viraemia development. Overall however, the baseline timepoint or ideally a timepoint prior to admission (e.g. during an outpatient appointment) or two timepoints such as before and immediately after HSCT, may be a more clinically useful timepoint than pre-engraftment.

It was interesting that most of the taxa that were differentially abundant between those who did or did not develop viraemia are known indicators of good health. Those who did eventually develop viraemia showed variable microbiome compositions. Despite this, absence of health-promoting taxa was a common theme amongst patients, highlighting conditions that promote pathogen colonisation and expansion.

We included alpha diversity as a potential predictor at both timepoints, however unlike in adult cohorts it was not a useful predictor of any clinical outcomes. These observations were surprising and one likely explanation may be the heterogeneity of the cohort studied. More specifically, age was variable and approximately 20% of the cohort was under the age of 2 and therefore could represent a developing microbiota compared to an adult population. It would be of interest to work out the 'microbiota-for-age *Z*-score' for these patients in comparison to healthy controls in the future, in order to assess their development at admission and whether this may have an impact on clinical outcomes[173]. Additionally, it would be interesting to assess the impact of antibiotics on alpha diversity and microbiome stability in this population[174].

Our findings reported herein are preliminary and to the best of our knowledge the first to be conducted in the UK. Further validation is the next necessary step, and it would be important to validate them in another cohort. Initially, a larger cohort from GOSH would be useful, although eventually a cohort from another

transplantation centre would be needed. This is to ensure that the biomarkers identified are applicable across the HSCT paediatric population.

It was surprising, yet promising, that we were able to find biomarkers in a heterogeneous cohort. However, it may be useful to stratify clinical outcomes such as mortality by cause and viraemia by viral load in order to disentangle biomarkers and their usefulness in specific populations. Similarly, it may be useful to investigate viraemia by viral families such as *adenoviridae* and *caliciviridae*. It is likely that certain biomarkers may be more applicable to certain sub-groups of clinical outcomes such as mortality by infectious causes *versus* mortality by disease relapse.

5.4.1 Limitations

This work has several limitations. The pre-engraftment sampling timepoints were adapted to each person hence they are very variable. It is therefore unclear whether the findings would hold up in a pre-determined timepoint such as day 16 (the median engraftment timepoint in this cohort). A future study with pre-planned timepoints would be beneficial. As previously mentioned, the baseline timepoint would be more clinically applicable.

Additionally, although baseline samples were collected prior to transplantation, they were not necessarily true baseline samples prior to any medication as patients often started antimicrobial prophylaxis within a day of admission and conditioning within a few days. Although we aimed to consent and collect samples at the earliest moment, this was not always possible. In fact, due to complex underlying diseases, most patients would have received medications, including antimicrobials, prior to admission, meaning that getting a true baseline sample in this population is most likely not possible.

The approach to biomarker discovery taken here did not use more sophisticated machine learning methods such as random forests with cross-validation as it is found to be unreliable for small sample sizes[172,175]. Both this and bootstrapping however, may also be useful as discovery/validation tools in the future.

As 16S rRNA gene sequencing only provides relative taxa numbers, other methods, such as qPCR, will be necessary to ascertain the absolute levels of the

taxa in the population. Using an established test, such as a qPCR, would also be currently more clinically applicable than using 16S rRNA sequencing. Although not included here, it would also be interesting to investigate the autologous cohort in more detail.

5.5 Conclusion

In this chapter, potential associations between the gut microbiota at baseline and pre-engraftment with clinical outcomes were studied with a view of identifying potential microbial biomarkers for overall survival, GvHD and viraemia in the paediatric HSCT.

At baseline *Clostridium_XVIII* was associated with a decreased risk of viraemia and *Klebsiella* was associated with an increased risk of GvHD. Additionally, *Enterobacteriaceae* was associated with an increased the risk of viraemia at the pre-engraftment timepoint. Focusing on viraemia, at baseline the outcome appears to be highly reflective of pre-existing gut health. At baseline, those who carried health-promoting taxa showed better clinical outcome post-transplantation. At engraftment however, the composition was reflective of detrimental gut domination and we provide evidence that *Klebsiella* at baseline may be indicative of risk of GvHD. If these findings are validated in future studies, once a patient is identified at risk for a particular outcome, clinical treatment may be adapted to mitigate the risk. In addition, selective approaches designed to increase or decrease particular taxa as well as to improve the composition of the microbiota (pre/probiotics/custom diet/FMT) may be useful in this population.

The approach taken here is somewhat incomplete, as we did not have the opportunity to validate our findings in another cohort, yet it is encouraging that we were able to find biomarkers in a highly heterogeneous population. Our findings, as well as the feasibility, lack of invasiveness and the ability to complement existing clinical approaches warrant further investigation of microbial biomarkers from stool as a method for adverse HSCT outcome prediction.

Chapter 6- Faecal metabolite profiling in paediatric HSCT

6.1 Introduction

The intestinal microbiota plays an important role within the host including protein, amino acid, xenobiotic and carbohydrate metabolism among others, which can result in both beneficial and deleterious effects on the host intestinal pathophysiology.

SCFAs acetate, butyrate and propionate for example, are produced by bacterial fermentation of carbohydrates/amino acids and are known to have multiple beneficial effects. Butyrate is an energy source for the colonocytes and is known to inhibit LPS-induced epithelial pro-inflammatory cytokine release by inducing colonic Treg cells via epigenetic modification of forkhead box-P3 promoter[176]. Additionally, butyrate can bind GPR109a, a receptor expressed by intestinal dendritic cells and macrophages, which activates anti-inflammatory IL-10 expression[176]. Propionate and acetate on the other hand, can undergo partial oxidation in colonocytes and are taken up to the liver to serve as substrates for gluconeogenesis and lipogenesis.

In contrast, the gut microbiota is also able to synthesise metabolites with detrimental effects on the host epithelium. P-cresol is synthesised from tyrosine, primarily by the *Clostridium* species, taken up by the colonocytes and excreted by the kidneys. Despite this, p-cresol is known to be genotoxic towards epithelial colonocytes[176]. Indole, on the other hand, is synthesised from tryptophan and its administration to colonocytes results in increased expression of genes involved in mucosal barrier function, yet it undergoes hydroxylation to indoxyl sulfate, a uraemic toxin, in the liver. Overall, the gut microbiota mediates a range of both beneficial and detrimental systemic and local effects on the host epithelium.

6.1.1 The effects of HSCT on the host and microbiota-derived metabolites

As pertained to in the introduction, HSCT is a complex procedure which can comprise conditioning (chemotherapy/radiotherapy), immunotherapy, cell infusion and the administration of a variety of medications. The side effects of the procedure include mucositis (inflammation of the mucosal surfaces), GvHD and viral and bacterial infections.

Epithelial damage during HSCT

The host defence system comprises of cellular immunity, the gut microbiota and an intact intestinal barrier. In patients undergoing HSCT, these attributes are often disturbed and a release of endogenous danger-associated molecular patterns (DAMPs) and exogenous pathogen-associated molecular patterns (PAMPs) is observed.

Chemotherapy and radiotherapy have been found to induce mucosal damage, which results in the release of pro-inflammatory cytokines including IL-1, IL-6 and type 1 interferons as well as reactive oxygen species and reactive nitrogen species[177]. Additionally, the damaged epithelium leads to increased translocation of bacterial species and an increased risk of bacteraemia as well as translocation of PAMPs, which can enhance pro-inflammatory activation. During GvHD for example, antigen-presenting cells sense PAMPs, which leads to their activation, release of pro-inflammatory cytokines and augmented presentation of the major histocompatibility complex to T cells[177]. Overall, HSCT damages the gut mucosa, leading to the release of a variety of danger signals, which can result in GvHD and is also likely to disturb the host-microbiota equilibrium[178]. A schematic of some of the changes to the epithelium and resulting GvHD are detailed below (Figure 6.1).

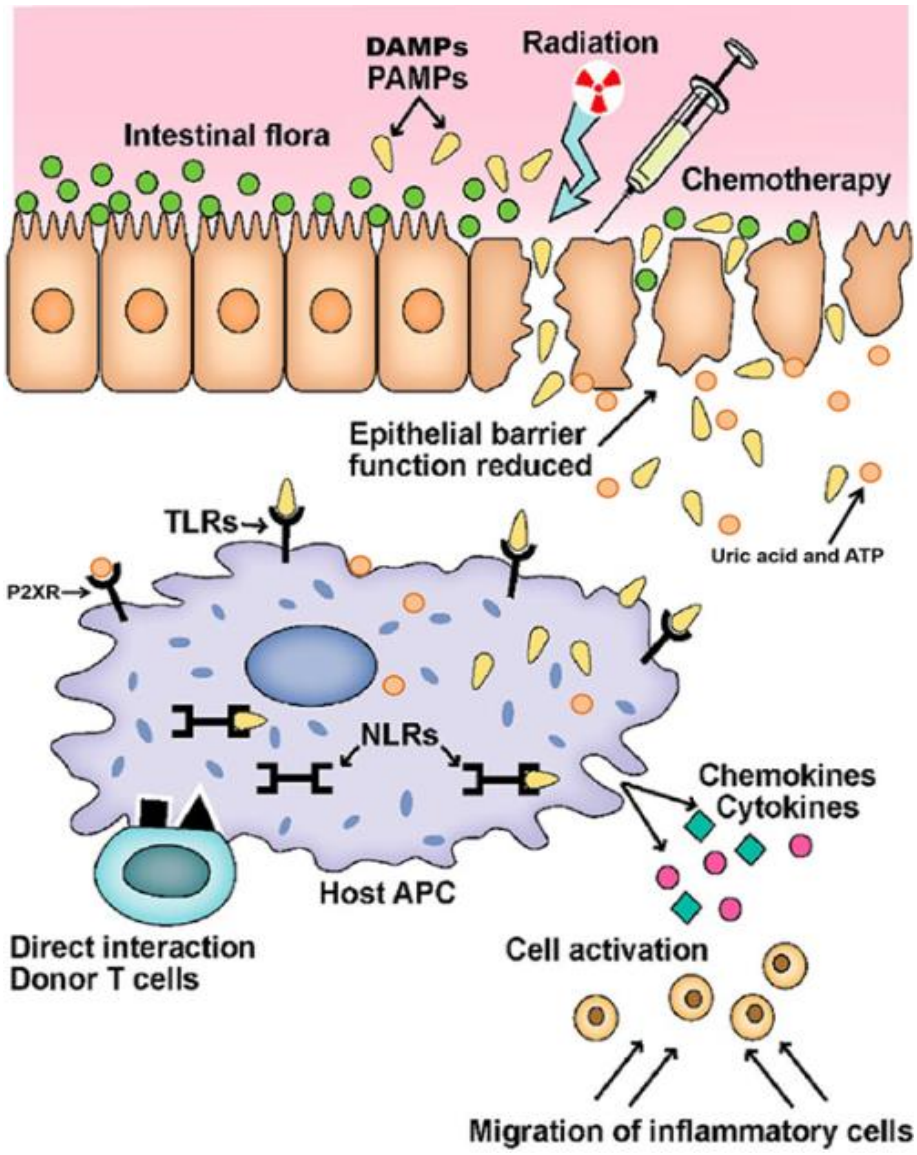


Figure 6.1 Gut damage during HSCT and GvHD initiation. Radiation and chemotherapy lead to damage to the gut mucosa including its integrity. The damaged cells release danger signals including uric acid, which results in pro-inflammatory cytokine release. PAMPs and DAMPs activate receptors (NOD-like receptors (NLR); Toll-like receptors (TLR)). Host antigen-presenting cells (APC) are activated, which leads to donor T-cell activation. Adapted from Ghimire *et al*[178].

Microbiota and metabolites during HSCT

As previously mentioned, (Chapters 4 and 5) HSCT as well as antimicrobial and other medications also affect the gut microbiota, resulting in prolonged reduced diversity and the overgrowth of facultative anaerobes such as *Enterococcaceae* and *Enterobacteriaceae* families. In total, the damage to the overall gut environment, including effects on the gut microbiota as well as the damage to the colonic mucosa is likely to have an impact on both host and microbial metabolism and the host-microbe interface.

Given our knowledge of the importance of SCFA in the maintenance of the gut epithelium, several studies to date have reported changes in SCFA during HSCT. A recent longitudinal study found butyrate and indole levels to be lower in adult HSCT recipients at baseline and at engraftment in comparison to healthy controls and that both positively correlate to diversity at baseline[179]. Most studies, however, have focused on the role of microbiota-derived metabolites in GvHD. Mathewson *et al* found no changes in SCFA levels in mice faecal samples, however they did find a decrease of butyrate in IEC[34]. A decrease in butyrate resulted in decreased histone acetylation, which lead to a decrease in luminal butyrate uptake by the IEC. Interestingly, butyrate uptake was restored with exogenous butyrate administration, which also mediated GvHD in the mouse model[34].

Michonneau *et al* found serum metabolite levels to be markedly different between patients with GvHD and healthy donors, such as a decrease in tryptophan and indole metabolites[180]. Indole compounds are aryl hydrocarbon receptor (AhR) ligands, which regulate the indoleamine 2,3-dioxygenase induction in cells, which suppresses the innate and adaptive immune cells. Additionally, AhR ligands also modulate Th17 responses and promote tolerance through the differentiation and the activation of Treg and Th1 cells. They concluded that dysbiosis as well as HSCT-related alteration of host metabolism induce a change in circulating metabolites that may influence allogeneic immune cell reactivity. Overall, gut microbiota-derived metabolites exert multiple effects on its host and certain metabolites are known to be altered during HSCT as well as during GvHD.

As few longitudinal studies on metabolite patterns post-HSCT exist to date, we chose to investigate several metabolite groups longitudinally including SCFA, amino acid (AA) and glycolysis and tricarboxylic acid (TCA) metabolites, from here on referred to as TCA metabolites. We thought they may be of relevance in this population as the microbiota is known to be involved in SCFA production and AA and TCA metabolites have not been previously investigated in detail in this population. Whilst other microbiota-associated metabolites including bile acids and tryptophan by-products were of interest, we were unable to reliably detect them using an untargeted NMR method.

6.1.2 Aims

The main aim of this chapter was to profile the faecal metabolome of paediatric HSCT patients and to link it to the microbial composition detailed in the previous chapters.

6.2 Methods

Methods for this chapter are detailed in Chapter 2.

6.3 Results

6.3.1 Baseline metabolites of patients and healthy controls

With a better understanding of the changes in faecal microbial composition (Chapters 4 and 5) in patients undergoing HSCT, it was imperative to identify corresponding changes in microbial function. For this purpose, dynamic changes in faecal metabolites were investigated by NMR spectroscopy.

We initially compared samples from unmatched HC to baseline samples of the patients (first sample upon admission prior to transplant). Figure 6.2 details the PCA plot for the comparison. Despite the range between the ages of the HC, the samples clustered closely. Some patient samples clustered close to the HC cohort, the majority however were dissimilar not only to HC but also with one another, indicating a high level of variability among the patient cohort.

Investigating these differences further we found that levels of SCFA butyrate and branched chain fatty acid isobutyrate were the main drivers of the observed separation, as both metabolites are highly enriched in HC, in contrast sugars were enriched in patient samples (Appendix Figure A12).

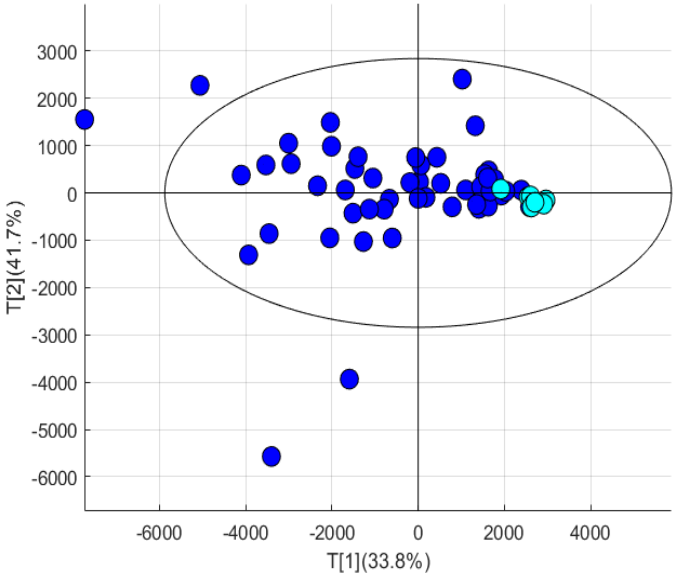


Figure 6.2 PCA plot of unmatched healthy controls (aqua; n=8) and baseline patient samples (dark blue, n=48)

Next, we chose to investigate several metabolite families, these included SCFA, AA and TCA metabolites.

Figures 6.3, 6.4 and 6.5 detail the comparisons in SCFA, AA and TCA metabolites between the HC and patient baseline samples. The 2-carbon SCFA acetate levels were not significantly different between the two groups (Fig 6.3 A); interestingly butyrate and isobutyrate were significantly lower in the patient group when compared to the HC (6.3 B, C; $p<0.001$), in contrast propionate levels were higher in in the patient group (6.3 D; $p<0.05$). A trend for increase in formate levels was seen in the patient group, but this trend did not achieve statistical significance. Most SCFA seem quite variable among patients, that is, some patients may be closer to the healthy metabolite levels than others upon admission.

AA analyses included investigation of alanine, glutamate, glycine, leucine, isoleucine and lactate (Figure 6.4). Some patients showed very high levels of glutamate and leucine, however no statistical differences in alanine, glutamate and leucine levels was recorded between the two groups (Figure 6.4 A/B/D). Glycine and lactate levels achieved significance with higher levels in patients *versus* HC (Figure 6.4 C/F), whilst isoleucine was higher in HC (Figure 6.4 E).

Several metabolites of the TCA cycle were also measured (Figure 6.5). These included D-glucose, citrate, pyruvate, pyruvic acid and succinate. Some patients showed levels of TCA metabolites similar to HC, others in contrast showed very high levels in all metabolites. Statistical significance was only achieved in D-glucose levels which were markedly greater in the patient group compared to HC (Figure 6.5 A).

In addition to the model above (Figure 6.2), we also explored the effects of age, sex, diagnosis and autologous *vs* allogeneic transplantation within the initial baseline samples, however no significant differences were found (Appendix; Table A17). As the autologous sub-cohort was very small, it was not investigated further.

Next, we investigated the effects of HSCT in patients receiving an allogeneic transplant by comparing the pre-transplant baseline samples to: *(a)* sample one

week post-transplantation, *(b)* sample closest to engraftment and *(c)* the last sample obtained. Significant differences between the pre-transplant *versus* the last sample for each patient were observed (Appendix; Table A17). With the available data we were not able to stratify patients by outcome, therefore these differences were not investigated any further.

Finally, we compared samples at baseline between several outcomes including overall survival, GvHD and viraemia, however no significant differences were noted. All models are detailed in the appendix (Appendix; Table A17).

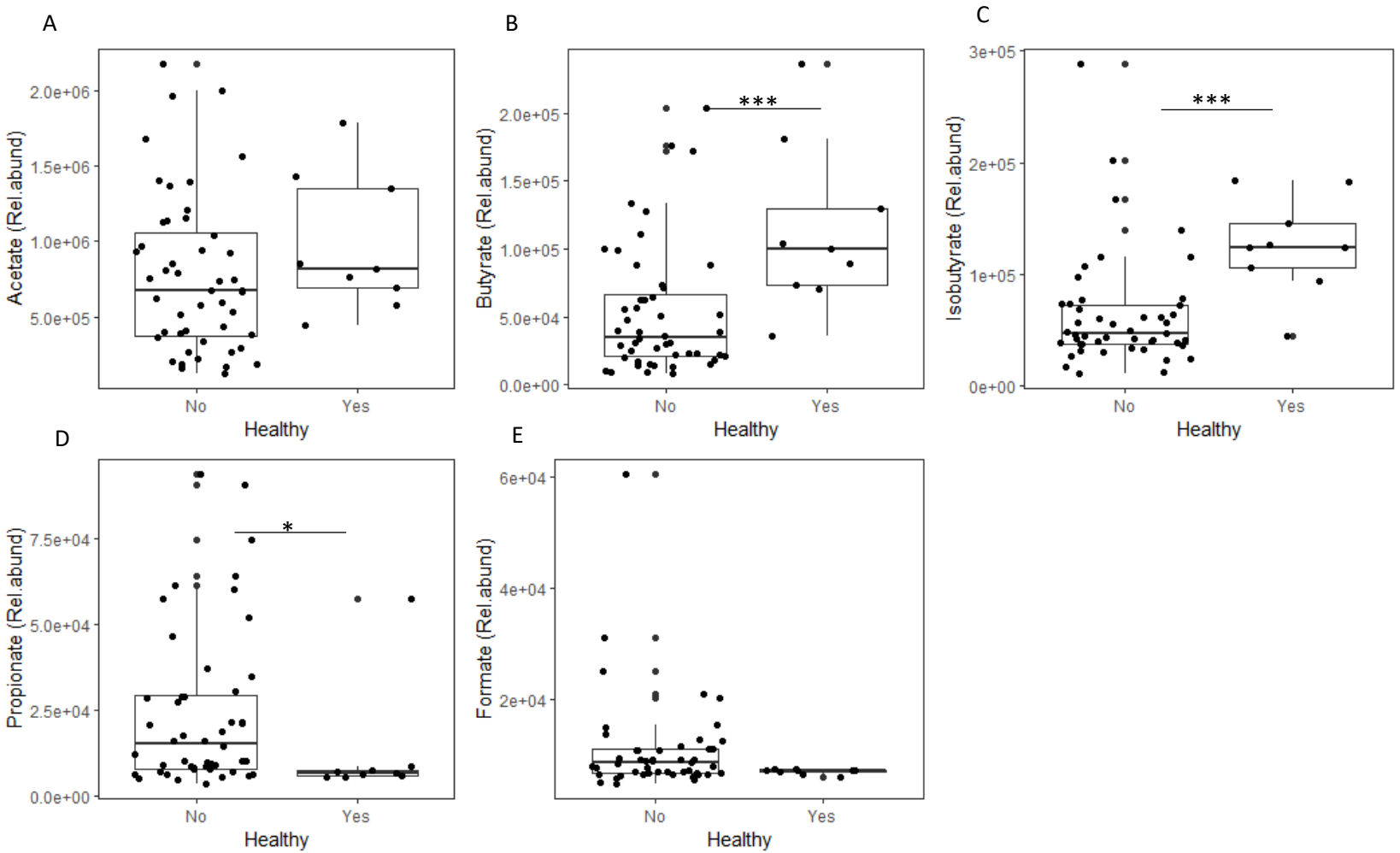


Figure 6.3 SCFA between unmatched healthy controls (n=8) and patient baseline samples (n=48). Rel.abund -relative abundance. Wilcoxon rank sum (***<0.001; *<0.05).

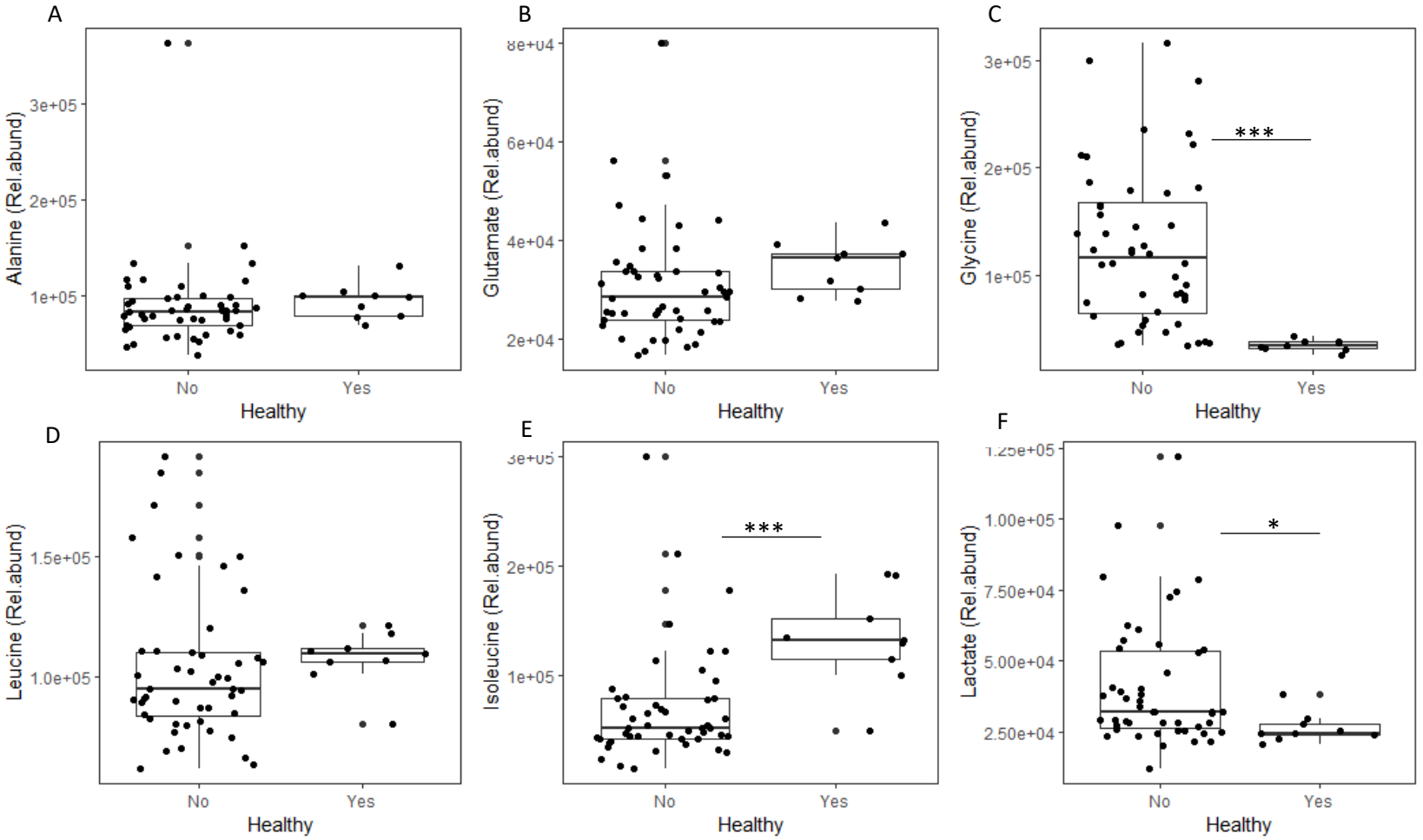


Figure 6.4 Amino acids and lactate between unmatched healthy controls (n=8) and patient baseline samples (n=48). Rel.abund-relative abundance. Wilcoxon rank sum (***<0.001; *<0.05).

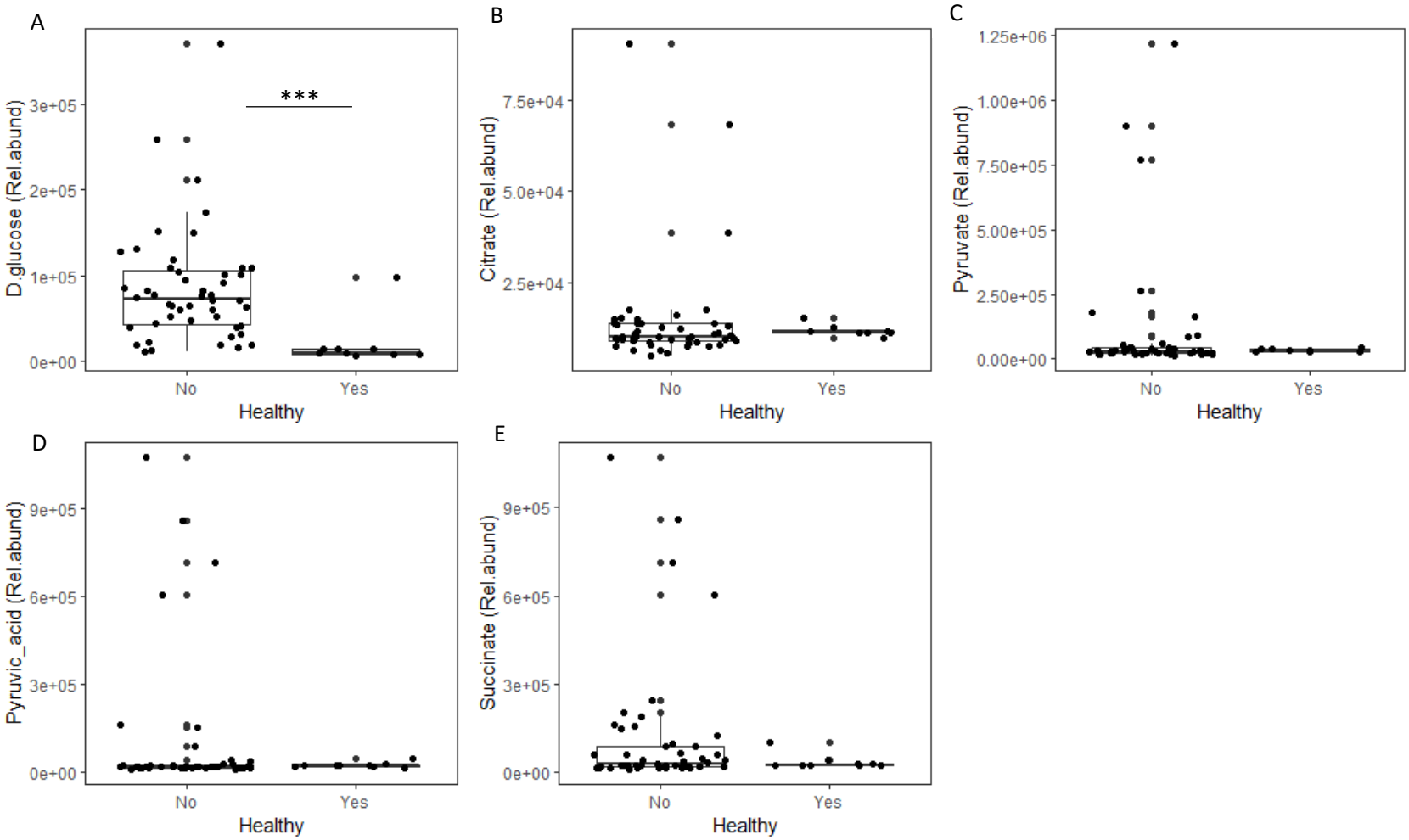


Figure 6.5 TCA and organic acids between unmatched healthy controls (n=8) and patient baseline samples (n=48). Rel.abund-relative abundance. Wilcoxon rank sum (***<0.001).

6.3.2 Longitudinal metabolite patterns in HSCT

We then investigated metabolites pre- and post-transplantation. Figures 6.6, 6.7 and 6.8 detail SCFA, AA and TCA metabolites measured during the first 100 days of inpatient stay (HC levels plotted alongside in red).

Acetate levels (Figure 6.6 A) were more robust as some patients recovered the acetate levels to pre-transplant levels. Both acetate and butyrate generally decreased and continued to be low post-transplantation, this was more evident for butyrate levels (Figure 6.6 B), potentially indicating a continued decrease of butyrate-generating microbiota. Some patients showed marked increase in iso-butyrate, propionate and formate post-transplantation (Figure 6.6 C-E).

AA analyses revealed complex patterns. Glycine levels were higher than the levels observed in HC and show a slight increase with time post-HSCT (Figure 6.7 C). Some patients showed marked increase in isoleucine post-transplantation (Figure 6.7 E). The rest of the amino acids or lactate do not show clear time-dependent patterns and appear variable among individuals (Figure 6.7 A/B/D/F).

In terms of TCA metabolites, glucose appears higher in patient samples than in HC during transplantation (Figure 6.8 A). Pyruvate, pyruvic acid, succinate and citrate show no clear patterns (Figure 6.8 B-E).

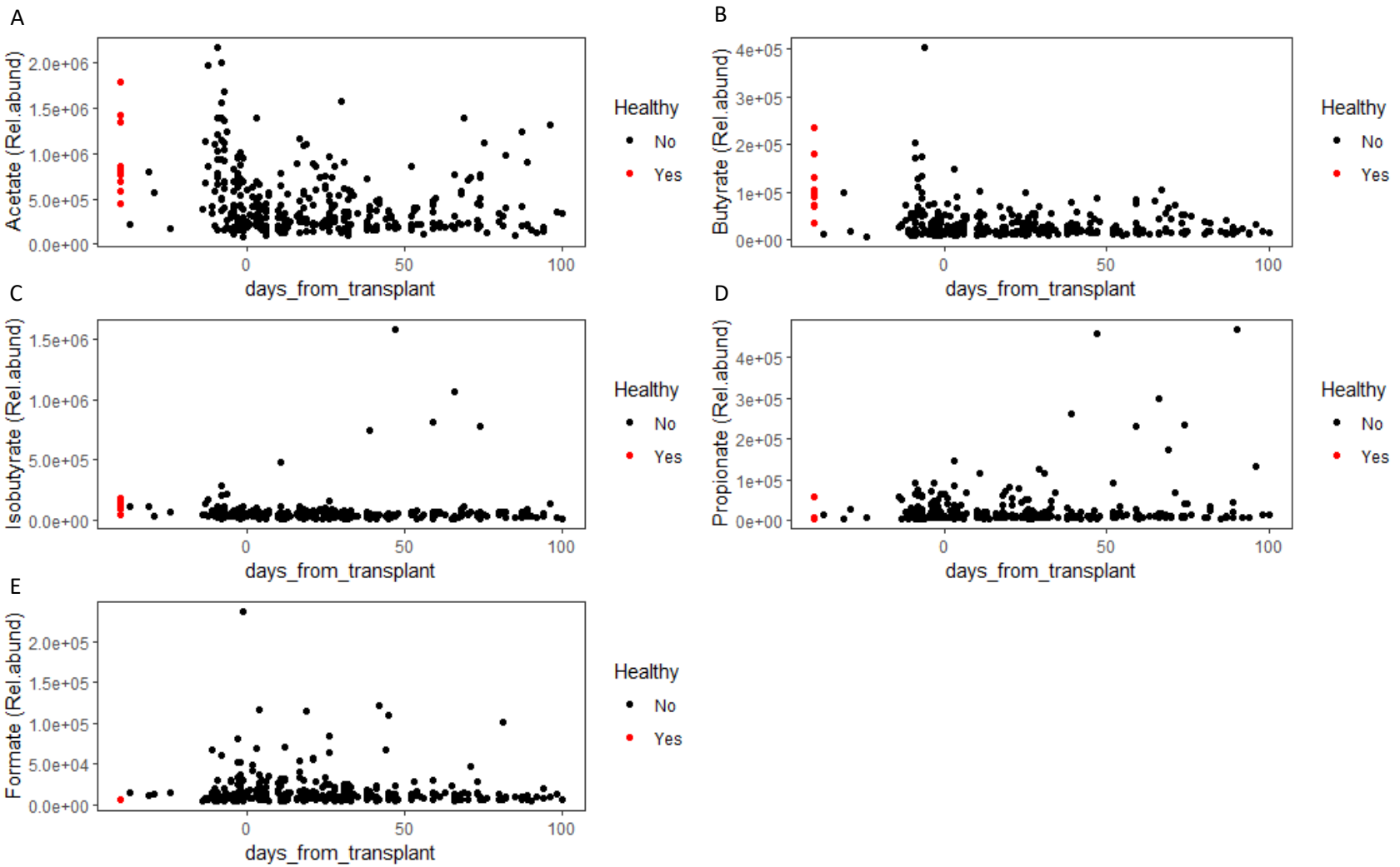


Figure 6.6 SCFA levels recorded during the first 100 days post-transplantation (n=339). Rel.abund-relative abundance. Red points denote healthy controls (n=8).

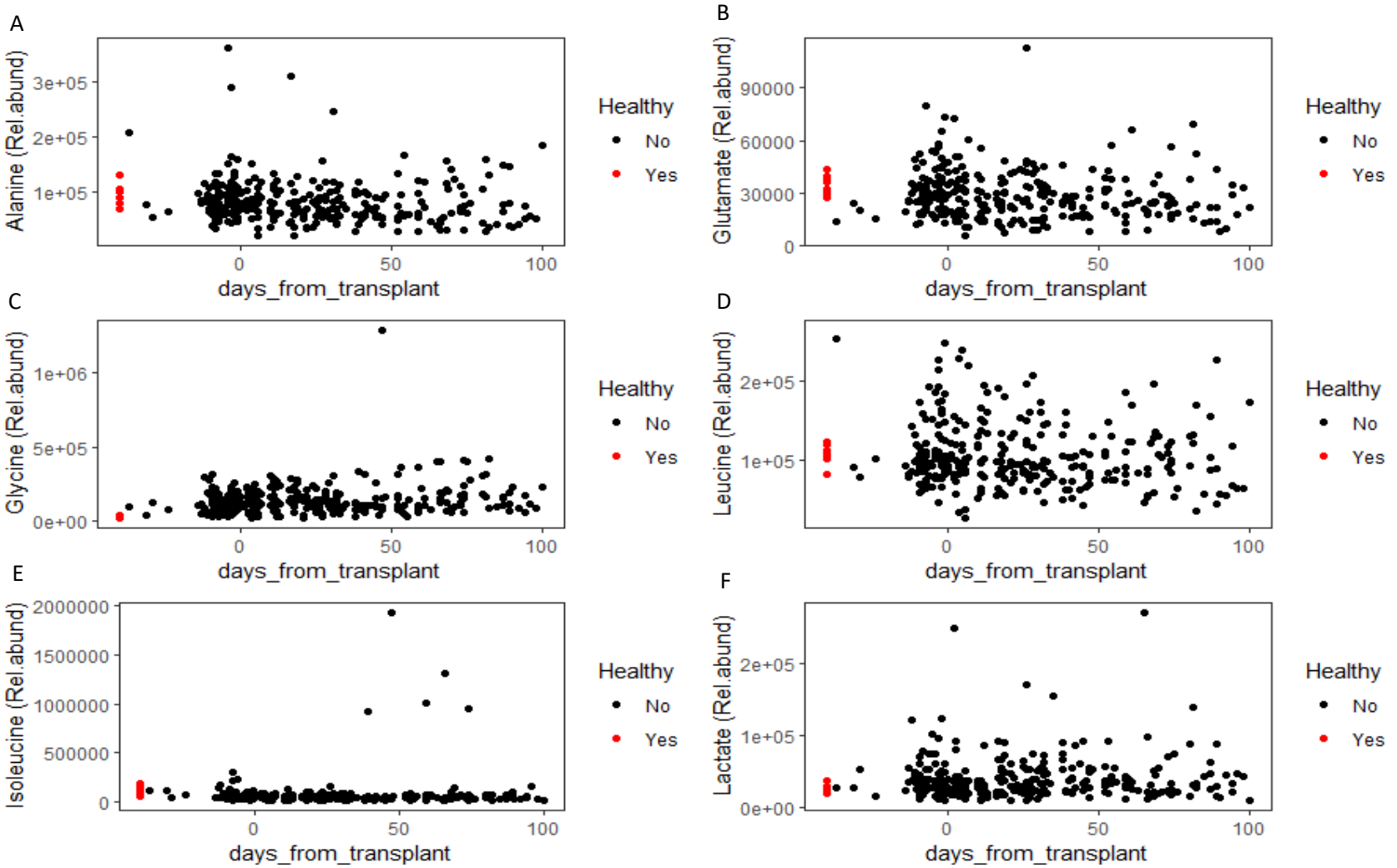


Figure 6.7 Amino acid and lactate levels during the first 100 days post-transplantation (n=339). Rel.abund-relative abundance. Red points denote healthy controls (n=8).

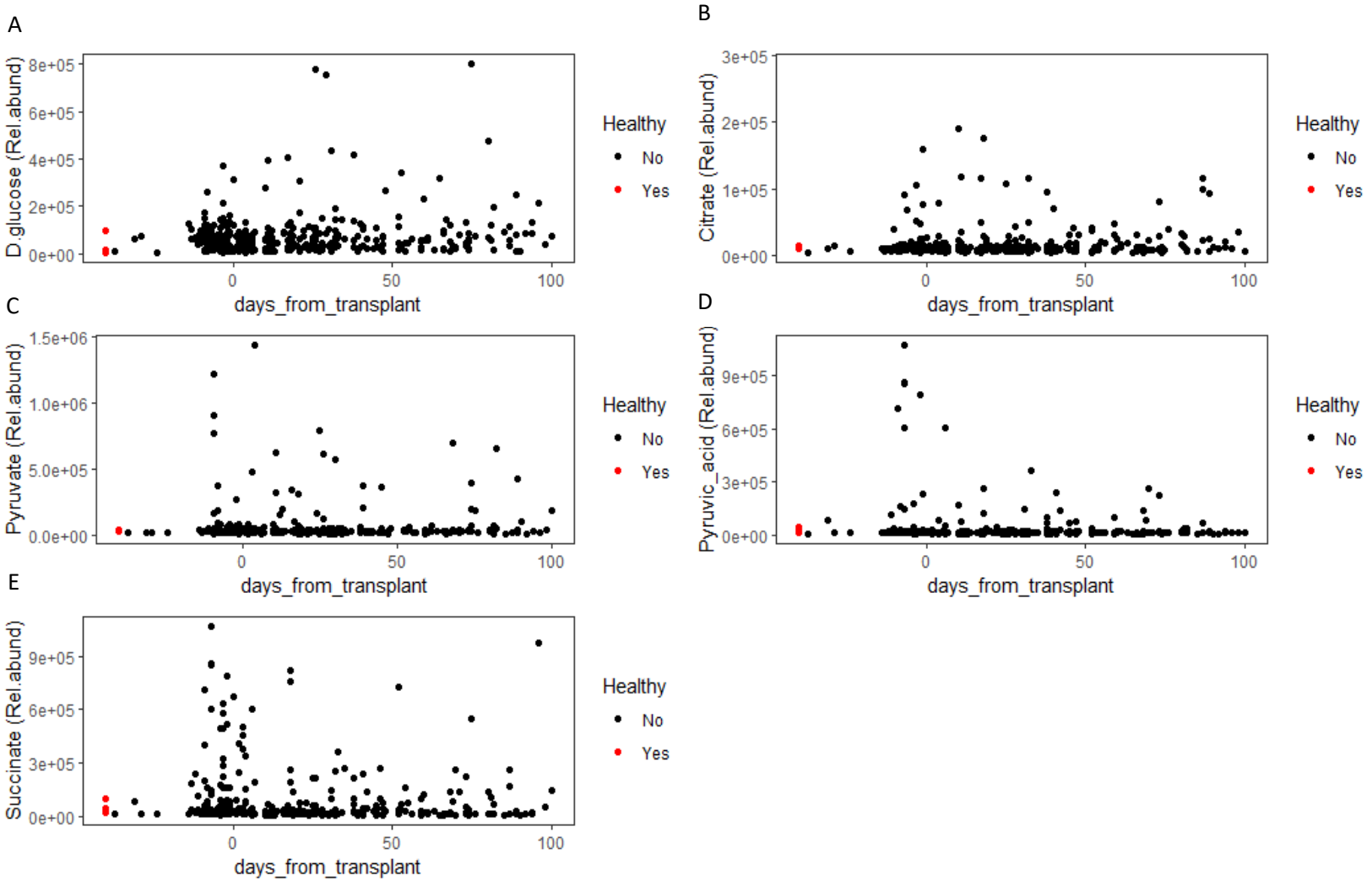


Figure 6.8 TCA and organic acid levels during the first 100 days post-transplantation (n=339). Rel.abund-relative abundance. Red points denote healthy controls (n=8).

6.3.3 Metabolite patterns during the first five weeks

Once the overall metabolite patterns had been identified, we next focused on the first five weeks (-7 days to 28 days relative to transplantation) of treatment, starting at admission, since most patients engraft and remain in hospital during this period. Figure 6.9 details the PCA plot for patients with two or more samples within the time period. Post transplantation (weeks 1-4) samples are similar to each other, however there is a separation between the pre-transplant samples (week -1) and post-transplant samples (weeks 1-4).

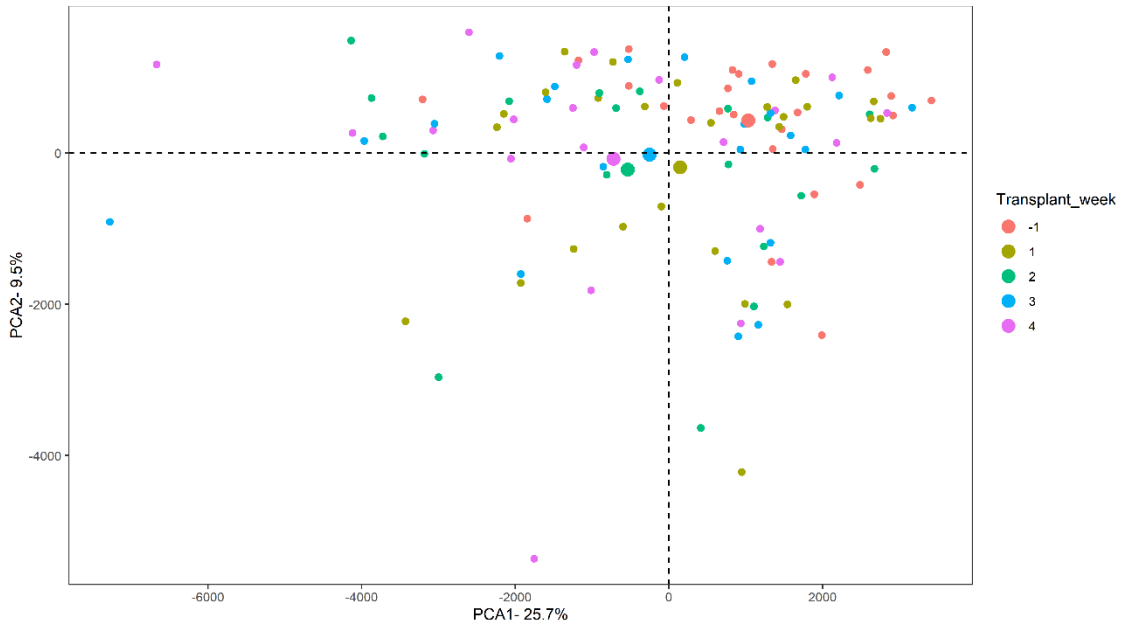


Figure 6.9 PCA plot of patient samples pre-transplant and over the first four weeks post-transplant (n=114; range 19-26). Week -1 denotes days -7 to -1 relative to transplantation. Larger dots denote the centroid for each week.

We then profiled all metabolites pre-transplant and over the first four weeks post-transplant (Figure 6.10). As previously observed, both acetate and butyrate decrease over the course of transplantation. Inversely, both glycine and lactate increase throughout.

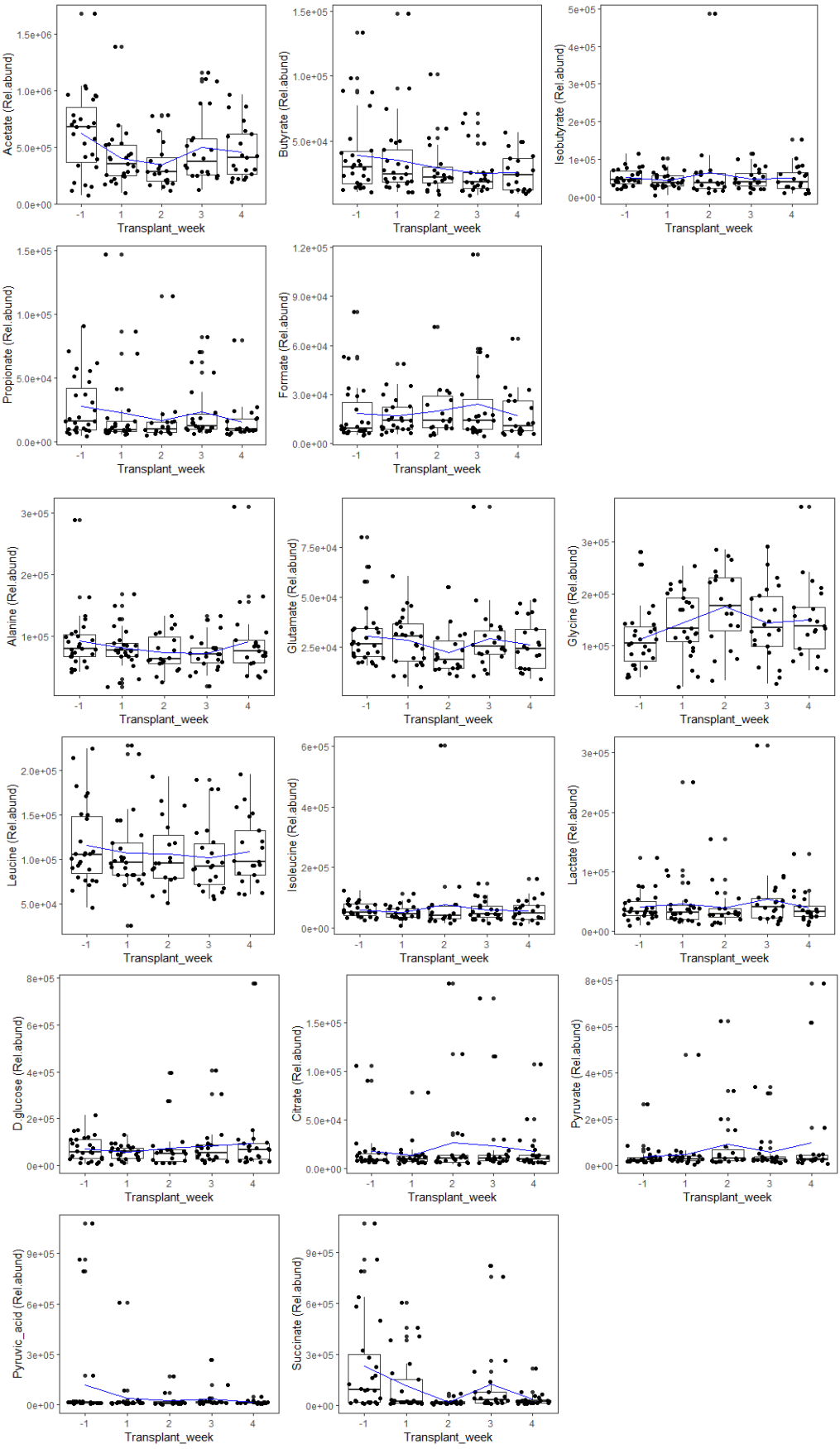


Figure 6.10 Metabolite profiling during the first five weeks of transplantation. Week -1 denotes days -7 to -1 relative to transplantation (sample range 19-26/week). The blue line indicates the mean for each week. Rel.abund-relative abundance.

6.3.4 Metabolites and microbial diversity at baseline

Having investigated metabolite trends in response to HSCT, we then wanted to link this to the 16S rRNA data discussed in the previous chapters. We were initially interested in whether any of the metabolites would correlate to alpha diversity at baseline. Butyrate positively correlated to diversity ($R=0.42$, $p<0.01$), and glutamate and glycine both showed negative correlations to diversity ($R=-0.29/-0.44$, $p=0.042/0.002$). Non-significant metabolites are detailed in the appendix (Figure A13).

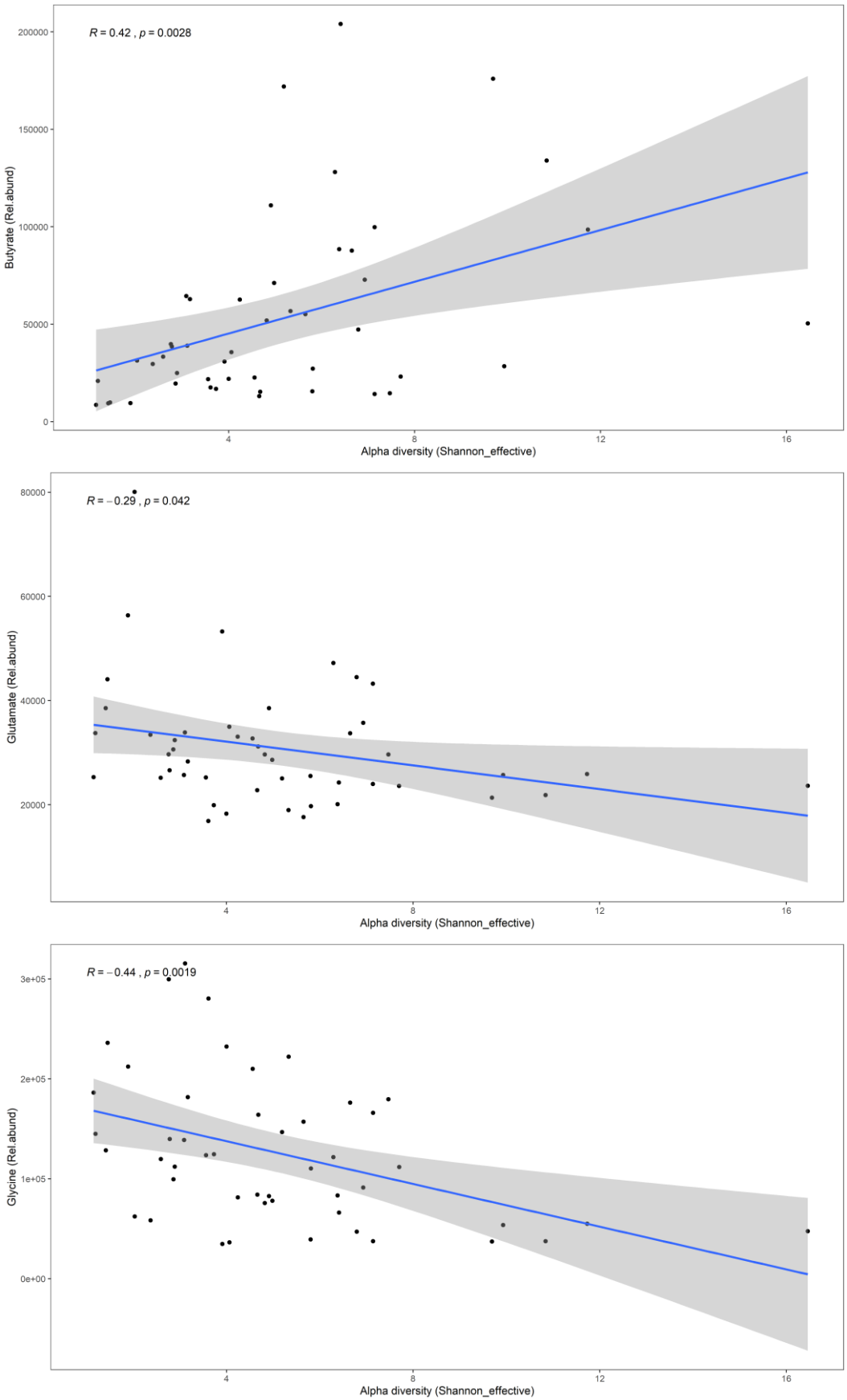


Figure 6.11 Metabolites correlated to alpha diversity at baseline (n=42). Pearson correlation coefficient the corresponding p-values are also displayed. Rel.abund- relative abundance.

6.3.5 Metabolites and microbial CSTs at baseline

Following this, we were also interested in the metabolic composition of the CSTs (described in Chapter 4). The taxonomic composition of CST1 broadly indicated that it showed greater resemblance to the HC in comparison to CST2 and CST3. We investigated this in baseline samples for each patient.

Despite the small sample number, we observed differences in metabolite profiles between the three CSTs in comparison to HC. Acetate levels in CST1, for example, were comparable to those in HC, whilst CST2 and CST3 were lower in comparison (Figure 6.12 A). In contrast, butyrate was reduced in all CSTs compared to HC, each type showing a progressive decrease (Figure 6.12 B). Isobutyrate levels showed a similar pattern (Figure 6.12 C). There was an increase in propionate and formate across the CSTs (Figure 6.12 D/E), this was evident between CST1 and CST3.

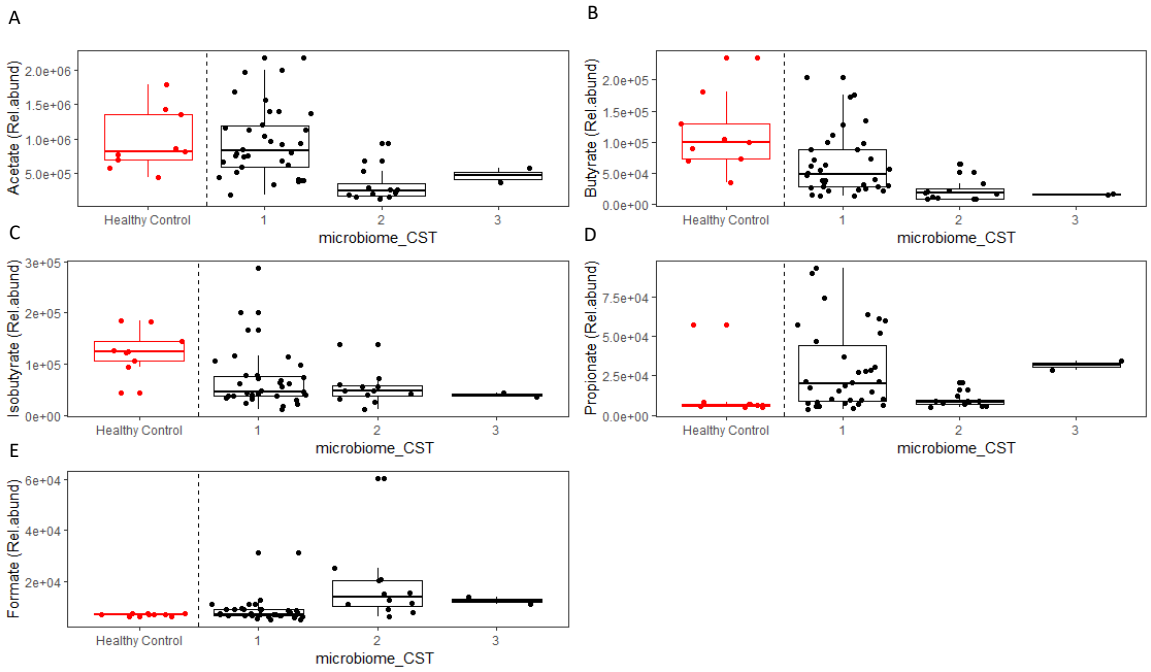


Figure 6.12 SCFA in baseline samples by their respective CST. Rel.abund-relative abundance.

We then compared trajectories of all SCFA for the first 100 days during HSCT based on the baseline CST for each patient. Acetate was the only SCFA that showed significant trajectories prior to false discovery correction (p=0.04) (Figure 6.13). Patients who started out in CST1 revealed a recovered acetate trajectory in comparison to those that started out in CST2 or CST3. Once corrected using

Benjamini-Hochberg however, the trajectories became non-significant (p=0.5). Trajectories for the other metabolites were not significant pre- or post-correction.

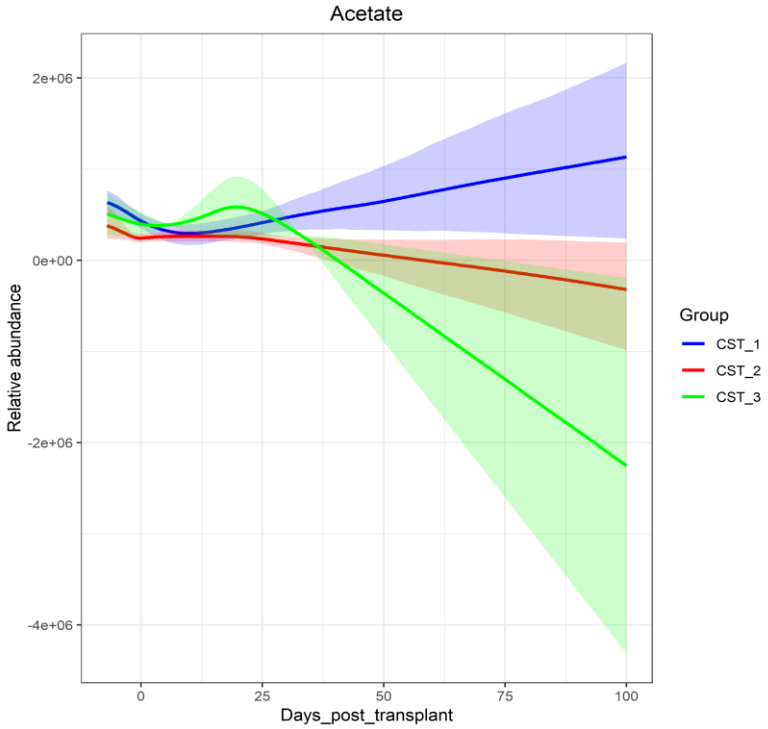


Figure 6.13 Trajectories of acetate throughout the first 100 days post-transplantation by the baseline CST for each patient. A smoothing spline is fitted with 95% confidence intervals coloured for each line.

Similar trends were seen with the AA profiles, whereby glycine was higher in all CSTs than in HC (Figure 6.14). Lactate levels in CST1 were similar to those of HC, whilst further increases were recorded in both CST2 and CST3. Isoleucine levels were broadly similar in all CSTs.

Most TCA metabolites were comparable between the three CSTs, with some variability, predominately within CST1 (Figure 6.15). Glucose was similar across the CSTs, and higher than HC. Graphs of metabolites in all samples stratified by their CST are detailed in the Appendix (Figures A14-A16).

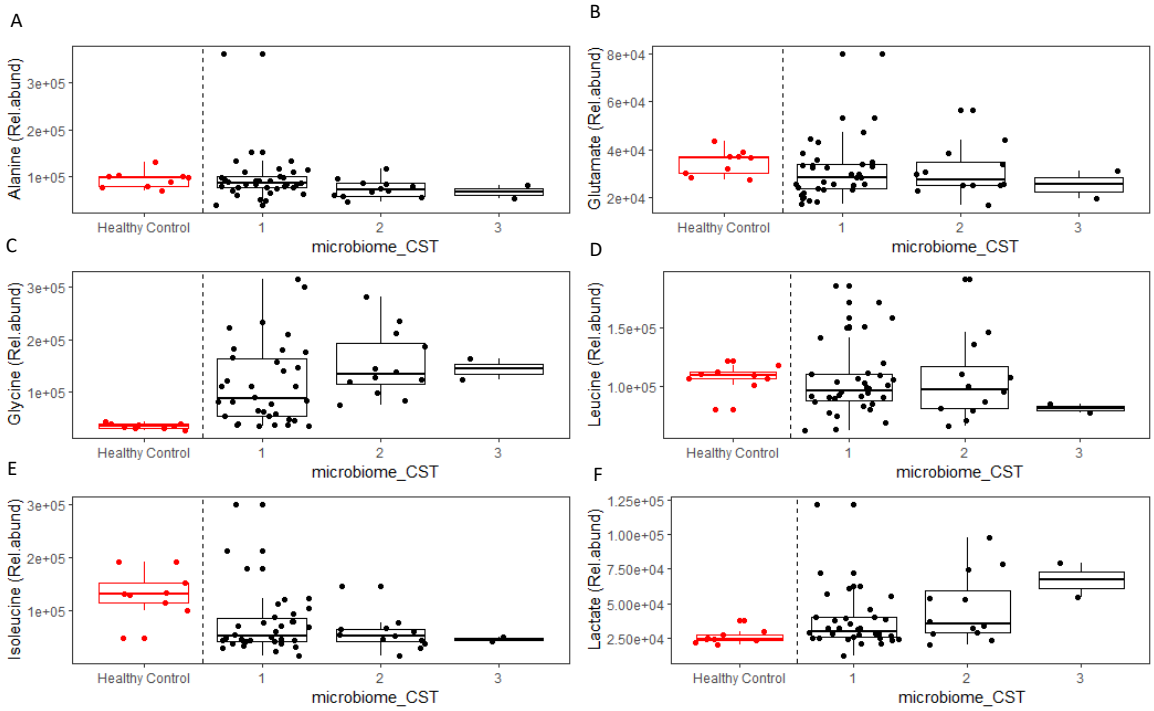


Figure 6.14 Amino acids and lactate in baseline samples by their respective CST. Rel.abund-relative abundance.

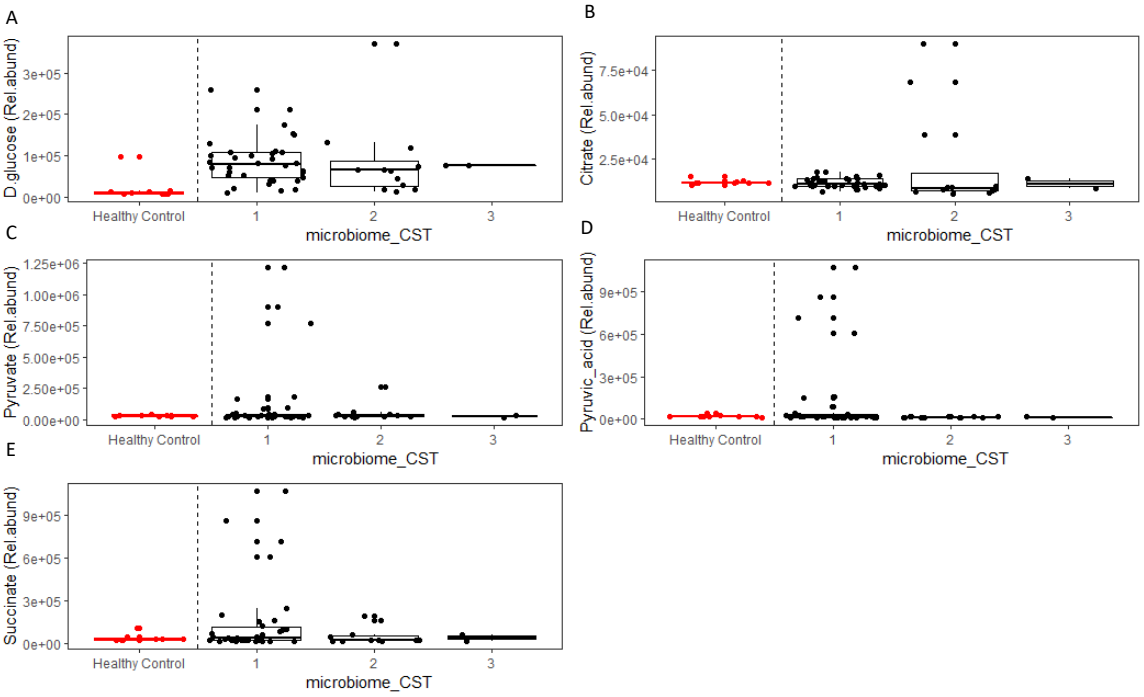


Figure 6.15 TCA and organic acids in baseline samples by their respective CST. Rel.abund-relative abundance.

6.3.5 Metabolites association to clinical outcomes at baseline

Finally, we were interested whether any of the metabolites at baseline, alone or in combination, would be useful predictors of clinical outcomes. To match predictive analysis in Chapter 5, only samples from allogeneic-HSCT recipients were included here. No associations were found for GvHD or survival. The final regression model for viraemia after backwards stepwise elimination of variables using AIC as a criterion for model selection is detailed in Table 6.1. The model shows that butyrate is associated with a decreased risk of viraemia (OR-0.99, 95 CI 0.85- 1.00; probability-0.49).

Table 6.1 Logistic regression model to predict viraemia at baseline

Viraemia	Estimate (Standard error)	Pr(>|z|)
(Intercept)	0.16 (1.72)	0.93
Butyrate	-3.52E-05 (1.46E-05)	0.02
Glutamate	1.02E-04 (6.47E-05)	0.11
Pyruvate	5.34E-06 (3.79E-06)	0.16

Figure 6.16 details the levels of the metabolites included in the final model for viraemia. At baseline butyrate does appear lower in patients who will go on to develop viraemia, although the levels can be variable. Butyrate levels do not increase the predictive power of the model created to predict viraemia at baseline in Chapter 5.

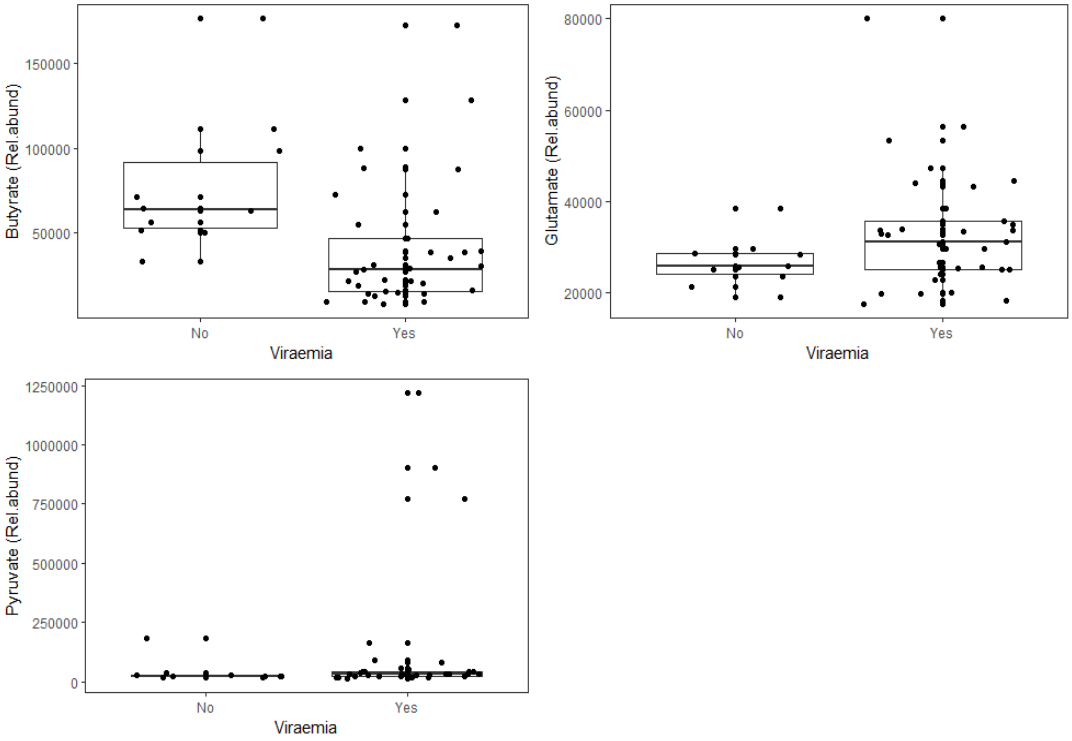


Figure 6.16 Relative abundance levels of metabolites at baseline between patients who did not (No) or who did (Yes) go on to develop viraemia (n=42). Rel.abund- relative abundance.

6.4 Discussion

The overall aim of this chapter was to profile the longitudinal faecal metabolome of paediatric HSCT patients and to broadly relate it to the microbial composition detailed in previous chapters.

We investigated the levels of select groups of metabolites, such as SCFA, AA and TCA, at baseline, during and post-transplantation. We also obtained data from unmatched HC at a single time point.

6.4.1 SCFA, AA and TCA patterns

As expected from patients with varying underlying disease and variation in the treatment protocols such as conditioning and drug administration, heterogeneity in metabolite levels was evident. Despite this heterogeneity certain trends were observed.

Baseline

At baseline, the levels of certain SCFAs, such as butyrate and isobutyrate, which can be utilised as fuel sources by IEC when food is scarce, as well as the AA isoleucine were decreased in comparison to healthy controls. On the other hand, the SCFA propionate and AAs lactate and glycine, as well as the TCA metabolite glucose were higher in patients than in the healthy controls. Collectively, this may be indicative of metabolic dysregulation in some patients even prior to the HSCT procedure. The variation in baseline microbiota taxonomic profiles seen in most patients supports this hypothesis.

Whilst other trends, such as a decrease in butyrate have been previously observed, it is unclear why propionate is higher at baseline in patients in comparison to HC, as it is known to promote intestinal epithelium homeostasis[181]. Despite this, the trend is likely to be driven by a few individuals and we did observe a decrease in propionate with time, in agreement with published work, which is likely a result of depletion in propionate-producing taxa such as *Bacteroides*. It is also unclear why isoleucine, an essential branched amino acid, a precursor to branched-chain fatty acids, is lower at baseline in patients; however it may be indicative of amino acid metabolism dysregulation or a decreased dietary intake.

During HSCT

Throughout transplantation, certain metabolites show time-dependent trends, such as the continuing decrease of both butyrate and acetate. This might indicate a continuous loss of anaerobes able to produce specific SCFA, particularly the *Lachnospiraceae* and the *Ruminococcaceae* families, which we observed previously (Chapter 4)[171]. Others, such as glucose, glycine and lactate were high form the onset and remained elevated throughout transplantation, with somewhat more complex trends. Additionally, glucose increased continuously (Figure 6.10), however its origin is unclear. Elevated levels may be indicative of malabsorption at the intestinal epithelium or it may be host derived as a result of treatment.

The decrease in butyrate, amongst other metabolites, is likely to have a detrimental effect on the reconstituting immune system. Butyrate is able to regulate the differentiation and expansion of the colonic Treg population, and thus it is likely that a microbiota and corresponding metabolites are necessary for the establishment of immunological tolerance[182]. This is particularly important in this population as the immune system begins to engraft approximately two weeks post-HSCT and thus changes in the microbiota and the associated metabolites in this period may affect the rate of engraftment and in turn affect clinical outcomes[183].

Due to both the 16S rRNA sequencing method being unable to identify taxa to the species level and the broad metabolomics approach taken here we cannot identify specific host-microbial interactions and mechanisms at play. It is most likely however, that the patterns observed are a result of changes in both the host and the gut microbiota metabolism. Potential scenarios to explain some of the observed metabolite changes are proposed below.

Metabolism of the colonic epithelium

In a homeostatic state the colonic mucosa, particularly the colonic lining, which is predominantly made up of colonocytes, is the first host cell to interact with the resident microbiota.

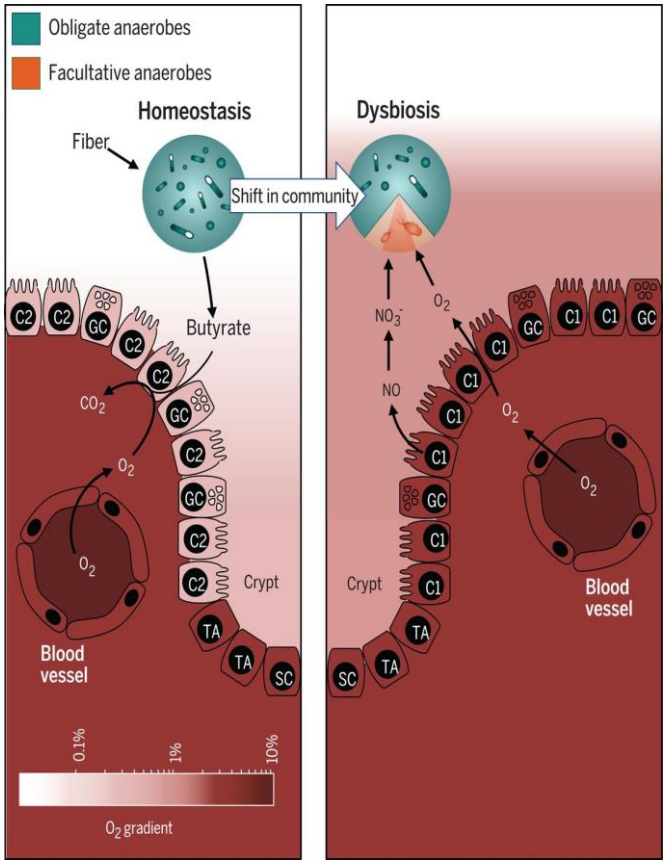


Figure 6.17 Epithelial metabolism shapes the colonic microbiota. During homeostasis (left), obligate anaerobic bacteria convert fibre into fermentation products, such as butyrate, which maintains the epithelium in a state leading to high oxygen consumption and maintains epithelial hypoxia (<1% oxygen) to limit oxygen diffusion to the lumen. During dysbiosis (right), colonocyte metabolic reorientation towards low oxygen consumption leads to oxygen diffusion to the lumen, which causes a shift in the microbiota from obligate to facultative anaerobes. TA- undifferentiated transit-amplifying cell; C2/C1- terminally differentiated C2-like/C1-like colonocyte; GC- goblet cell, NO- nitric oxide. Adapted from Litvak *et al*[184].

Colonocytes are well known to utilise fermentation by-products such as butyrate, leading to ATP production via β-oxidation[184] (Figure 6.17). This in turn reduces luminal oxygen availability and provides favourable conditions for obligate anaerobes[185]. The lack of oxygen leads to the stabilisation of hypoxia inducible factor, a transcription factor promoting maintenance of the epithelial barrier function. Administration of certain antibiotics may aggravate microbial ecology, as they help deplete obligate anaerobes resulting in the loss of butyrate and other SCFA. Taken together, these cellular events may trigger a shift in the colonocyte energy metabolism towards anaerobic glycolysis, with the preferential use of glucose and release of lactate, as observed in this cohort[184]. The subsequent increase in oxygenation and potentially an increase in pH, may favour the growth of facultative anaerobes, which are better adapted to utilise luminal oxygen[184].

Byndloss *et al* have found that a decrease in butyrate-producing organisms downregulates signalling through the butyrate sensor peroxisome proliferator activated receptor γ (PPAR-γ) and Treg cells, which lead to an increase in luminal nitrate and oxygen, resulting in *E. coli* and *Salmonella* outgrowth in a murine model[186]. The authors propose that colonocytes exist in two energy states, C2-

the homeostatic energy state driven towards β-oxidation and C1- the inflamed state with a trend towards anaerobic glycolysis, and they liken this to macrophage polarisation[186] (Figure 6.18). They found that proinflammatory signals led to a change in colonocyte polarisation to the C1 state, favouring anaerobic glycolysis, which leads to oxygenation of the gut lumen.

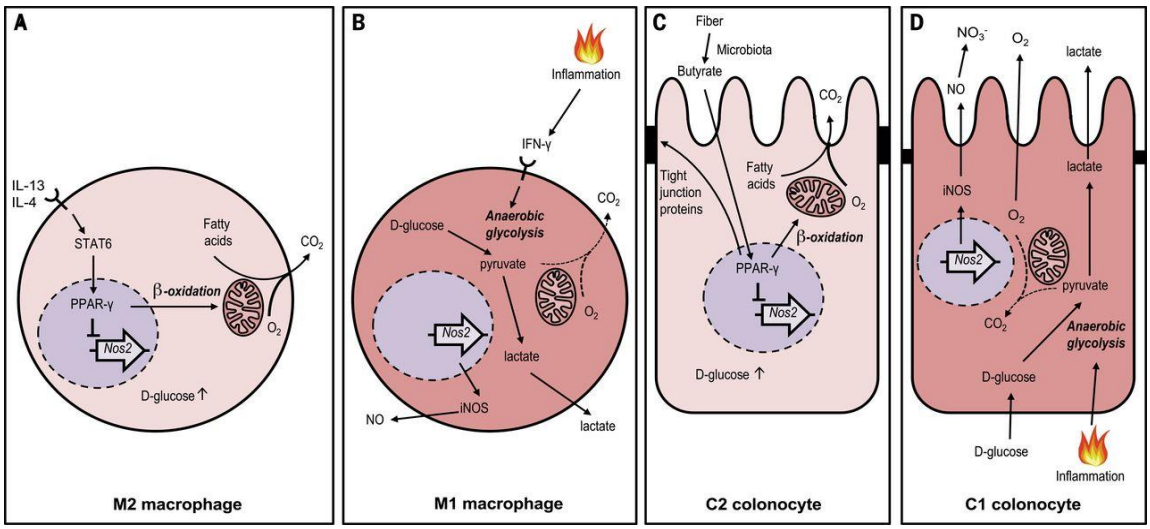


Figure 6.18 Extending the M1/M2 paradigm to colonocytes. **A** IL-4 and IL-13 stimulate polarization into alternatively activated M2 macrophages. **B** Proinflammatory signals, such as IFN-γ, stimulate polarization into classically activated M1 macrophages by shifting the host cell metabolism toward anaerobic glycolysis. **C** Microbiota converts fibre into fermentation products, which stimulate a metabolic polarization into C2 colonocytes by inducing a PPAR-γ–dependent activation of mitochondrial β-oxidation, lowering epithelial oxygenation. **D** Proinflammatory signals stimulate a metabolic polarization into C1 colonocytes by shifting the host cell metabolism toward anaerobic glycolysis, which increases epithelial oxygenation. This lead in an increase in oxygen in the lumen. Lactate produced during anaerobic glycolysis is released into the gut lumen, whereas nitric oxide (NO) produced by iNOS is converted to nitrate (NO_3^-). Adapted from Litvak *et al*[184].

This is a plausible situation for the HSCT patients, as they receive a plethora of antimicrobials and exhibit gut domination with *Proteobacteria* and *Enterococcaceae*, as well as the corresponding decrease in butyrate and increases in lactate and the relative lack of *Ruminococcaceae* and *Lachnospiraceae* families.

Butyrate does however have other receptors; therefore, this theory is unlikely to be the whole picture, however it fits well with the metabolite patterns we have observed. Finally, the authors primarily focus on *Enterobacteriaceae* expansion, however we observed expansion of taxa from other families, therefore other mechanisms could also be at play.

Epithelial inflammation and integrity

A large proportion of patients in the cohort suffer from varying degrees of mucositis (inflammation in the gut epithelium as a result of conditioning) and receive repeated antibiotic administration during hospitalisation. Mathewson *et al* found that butyrate in the intestinal tissue was significantly reduced seven days post-HSCT, which was associated with impaired IEC junction integrity. Importantly, administration of butyrate restores epithelial integrity[34]. Similarly, Holota *et al* found that repeated administration of ceftriaxone in a murine model resulted in decreased immunoreactivity of the free fatty acid receptor 2 and 3, which in turn caused a significant shift in colonic mucosal homeostasis with a disturbance of oxidation balance, activation of HIF1α, an increase in epithelial permeability, bacterial translocation and remodelling of the colonic tissue[187]. Finally, de Pietri *et al* suggest that overgrowth of *Enterococcus* may exacerbate intestinal injury and inflammation during chemotherapy[188].

Malabsorption, also observed in conditions such as Crohn's disease, whereby certain metabolites such as bile acids are not absorbed as a result of inflammation, could also play a part in metabolic dysregulation[189]. It is likely that both repeated administration of antimicrobials as well as the associated conditioning promotes the loss of epithelial integrity, which could additionally disturb the host-microbe metabolic inter-dependence and lead to the metabolic changes observed in the study.

6.4.2 Metabolites and the microbiota

Metabolites and microbial diversity

Having explored metabolite patterns, we were particularly interested in linking the findings from the 16S rRNA sequencing (Chapters 4 and 5) to metabolite findings. We found that both glycine and glutamate were negatively associated with alpha diversity.

Colonocytes in germ free mice (which lack any bacterial metabolic by-products), adapt to their suboptimal energy state by upregulating autophagy in order to utilise amino acids for energy and to prevent apoptosis[190]. We quantified a range of AA and found increased levels of glycine both prior to and during HSCT, which

may reflect a degree of dysregulation in AA metabolism. Glycine is a dietary and an endogenous amino acid. It is converted to acetate and glutathione, among other metabolites, by the gut microbiota and is also taken up by the liver for catabolism. An increase in glycine may highlight a higher pH and a low carbohydrate environment as well as a lack of microbial species that utilise this amino acid and this increase may in turn affect the host metabolism[191]. Mardinoglu *et al* found higher levels of glycine in germ-free compared to conventionally raised mice[192]. The gut microbiota of conventionally raised mice modulated glycine availability by downregulating genes involved in glycine metabolism and synthesis in the host[192]. The extent to which different microbial taxa contribute to host glycine metabolism has not yet been elucidated.

Glutamate is another non-essential dietary/endogenous amino acid, which is a pre-cursor for a variety of important metabolites including acetate, butyrate and 4-aminobutanoic acid, an important neurotransmitter, and is a fuel source for enterocytes. It is unclear why patients had higher levels of glutamate at baseline in comparison to HC and why glutamate was inversely correlated with gut microbiota diversity. It may be that as a result of the epithelial damage and drug administration the intestinal epithelium is either unable to take up glutamate or there is a lack of taxa involved in glutamate processing. The negative correlation with diversity may therefore be an indication of overall dysregulation and damage in the gut.

Additionally, butyrate at baseline was positively correlated to microbial diversity and a reduced risk of viraemia. The link to diversity has been observed before and is not surprising due to evidence linking butyrate to IEC integrity and health[179,183]. It is likely that in agreement with the findings from the previous chapter, increased butyrate levels at baseline highlight gut and epithelial health and can be considered a ‘proxy’ for a more healthy-like gut microbiota composition. Similarly, patients with higher butyrate levels at baseline may be more likely to have a greater degree of host-microbial fitness and may therefore be less likely to suffer from new viral infections and/or viral reactivation.

Metabolites and CSTs

Finally, we aimed to understand the metabolic make-up of the CSTs defined in the preceding chapters. The results of samples at baseline broadly supported our previous observations that CST1 is more similar to HC than CST2 and CST3 (based on taxonomies). With certain metabolites, such as acetate and lactate, CST1 levels were found to be comparable to HC. In contrast butyrate levels were reduced across all CSTs with a progressive decline between CST1 and CST3. It is therefore evident that CST1 is more healthy-like than the other CSTs from both the taxonomic and metabolite perspectives, albeit not necessarily on par with the HC. Some metabolites, such as the SCFA show greater association to the microbiota composition than others such as AA and the TCA metabolites. As little variation between the CSTs was recorded for these, one may speculate that host metabolism may play a greater role in modulating their levels compared to SCFA. It would be interesting to investigate whether and to which extent starting in a particular CST may impact patient metabolite profiles. With acetate for example, the trajectories appear to be better for those starting in CST1 in comparison to CST2 and CST3, however this warrants further investigation.

6.4.3 Limitations

This work has several limitations. Firstly, as this was a pilot study, we used an untargeted approach. As a result, we do not know absolute metabolite levels, although a targeted approach could be undertaken in the future. Additionally, faecal metabolites reflect both host and microbial metabolism, therefore the origins of certain metabolites such as lactate are not necessarily clear, as it is both a fermentation product of lactic-acid bacteria and a host metabolite, making hypotheses more difficult.

As seen for microbial composition, the metabolome is known to be fairly stable in adults, in contrast paediatric metabolome is likely to show greater variation, especially in those undergoing HSCT. This makes finding and appropriate intervention for this paediatric cohort more difficult. Additionally, certain metabolites, such as acetate, are age dependent. The current study was preliminary and the impact of co-variates such as age were not studied in detail.

Stratifying into age groups would be useful in the future. Similarly, it would be worth stratifying by other co-variates including the underlying diagnoses.

Additionally, whilst the faecal microbiome may not precisely reflect the microbiome-host interface at the epithelial layer, using faecal samples has a number of advantages such as being non-invasive, sampling the gastrointestinal tract, providing the possibility of repeat measurements and largely reflecting gut microbiota composition[193]. Despite this, the close-knit relationships between certain bacteria and the host mean that faecal metabolites may not be descriptive of these relationships and metabolomics of the intestinal tissue may also be of interest.

Finally, the impact of nutrition such as TPN and EN while undergoing treatment was not investigated here. Patients on TPN receive a variety of amino acids, proteins, lipids and glucose, which bypass the gut, which may in turn have an indirect impact on the metabolite availability in the gut and thus availability for microbial metabolism. This could, indirectly, have an effect on the observations reported here. Additionally, only one sample was obtained from each HC, meaning further age-matched comparative studies are needed.

6.5 Conclusion

In conclusion, untargeted NMR profiling revealed the complexity of metabolite patterns in children undergoing HSCT. In concord with the 16S rRNA data, the baseline samples of most patients were found to be dissimilar to the HC, most notably with SCFA and glucose.

Host-microbiota metabolic interactions, specifically relating to colonocyte energy and AA metabolism are disturbed, likely as a result of the underlying disease and ongoing/previous treatment. The transplantation procedure and repeated antimicrobial administrations are likely to exacerbate the initial imbalance.

As metabolites of microbial origin are integral to the host's wellbeing and are clearly dysregulated throughout HSCT, further exploration of host-microbiota interactions as well as their use as clinical biomarkers and in clinical interventions in this population is warranted.

Chapter 7- Conclusions

Little work to date has investigated the paediatric gut microbiota during HSCT, however it has been extensively explored and linked to clinical outcomes post-HSCT in adult cohorts. This work aimed to explore the paediatric gut microbiota and metabolome during haematopoietic stem cell transplantation in several ways. Initially, the dynamics of both the longitudinal microbiota and metabolome in the cohort were explored, followed by biomarker discovery in both datasets.

7.1 Summary of findings

In Chapter 4 I focused on exploring longitudinal dynamics of the gut microbiota in this cohort. There was a decrease in alpha diversity throughout transplantation, particularly around the time of conditioning, which is in accord with findings in adult cohorts. In contrast to others' findings however, most patients did not recover their initial diversity. The samples grouped into three distinct, yet overlapping clusters, with CST1 being more healthy-like. This revealed interesting dynamic patterns. Some patients start out in more healthy-like CST1 and eventually progress into CST2 or CST3. Others are already in CST2 or CST3 to begin with. CST2 and CST3, which had high abundances of taxa such as *Enterococcus* and *Staphylococcus*, had high self-transition probabilities, making them more stable than CST1. CST3 was in turn linked to a higher probability of viraemia.

In Chapter 5 I explored whether the gut microbiota may be useful as a predictive biomarker of clinical outcomes at both baseline and pre-engraftment timepoints. Initial baseline samples were already dissimilar to samples from healthy controls at baseline in both composition and diversity. Reduced abundance of *Clostridium_XVIII* at baseline and increased abundance of *Enterobacteriaceae* at pre-engraftment increased the risk of viraemia respectively, whereas the presence of higher levels of *Klebsiella* at baseline increased the risk of subsequent GvHD. The findings are likely a reflection of gut microbiota health at baseline and of gut domination at engraftment. It was encouraging to find biomarkers in such a heterogeneous cohort of patients.

Finally, in Chapter 6 I investigated the faecal metabolome of the cohort by way of NMR profiling. In agreement with the 16S rRNA sequencing data, patient baseline samples were dissimilar to the healthy controls. Specifically, they had lower levels

of SCFA such as butyrate and isobutyrate and had higher levels of propionate, glucose, lactate and glycine. These changes are suggestive of the loss of SCFA producing anaerobes, which we observe from the sequencing data, as well as changes to epithelial metabolism and epithelial damage.

Butyrate was, unsurprisingly, positively correlated to alpha diversity and decreased the risk of viraemia at baseline. I was also interested in the metabolic makeup of the three clusters identified in the dynamics chapter. As expected, CST1 was more healthy-like from metabolite perspective as it was on par with or closer to the healthy control levels of certain metabolites such as acetate and butyrate in comparison to CST2 and CST3.

7.2 Discussion

Overall, the findings indicate that HSCT impacts the gut microbiota in a similar way in both children and adults. Figure 7.1 summarises previously observed effects of allo-HSCT on the gut microbiota[194].

In agreement with published work in both adults and children, we observed decreased alpha diversity during HSCT and an expansion of facultative anaerobes including *Enterococcaceae* and *Streptococcaceae* families. Although previously published work indicated *Lactobacillus* dominance, this was not observed in the current study[96]. We observed expansion of *Staphylococcaceae*, which has not been previously observed in a paediatric cohort. Differences between the cohorts suggests that these patterns may be specific to the transplant centre, potentially as a result of drugs and/or dietary regimes[96]. Additionally, we observed decreases in certain beneficial short-chain and branched-chain fatty acids such as butyrate and isobutyrate and increases in lactate and glucose during HSCT, which are partly novel findings.

Together, these findings are indicative of a disruption in the gut microbiota and a dysregulation of the host/microbiota metabolism. More specifically, as conditioning and drug administration impact the gut epithelium, we hypothesize that patients' epithelial colonocytes may be at a suboptimal energy state and can switch to an inflamed state, which increases oxygen release into the lumen and drives overgrowth of facultative anaerobes.

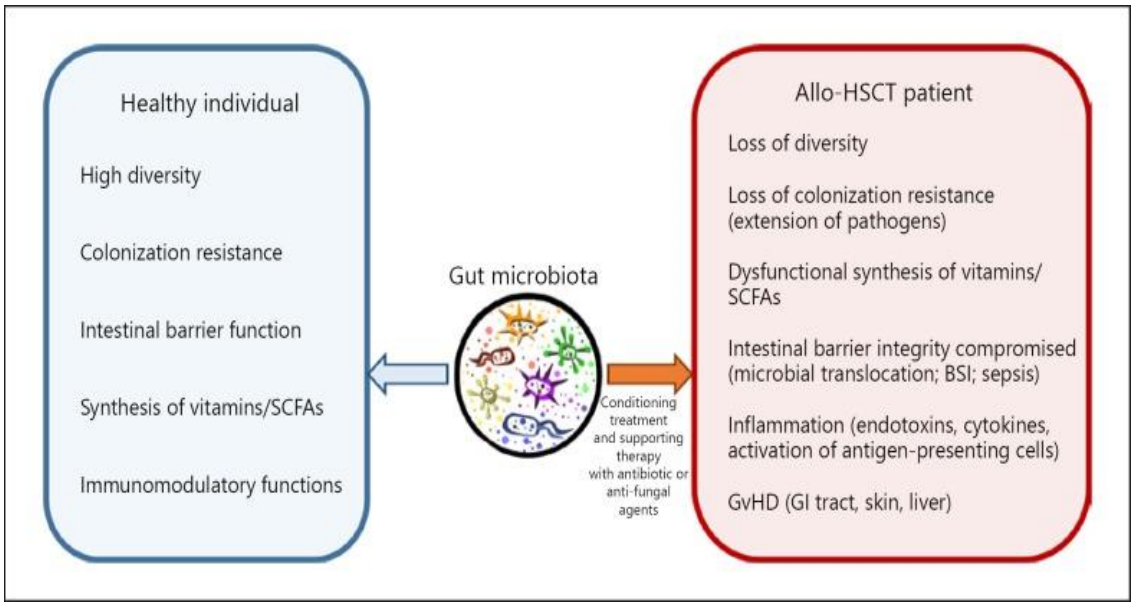


Figure 7.1 Gut microbiota in a healthy individual *versus* allo-HSCT patients. Conditioning and drugs disrupt the microbiota and its functions, which results in a range of effects including loss of diversity, a compromised intestinal barrier and inflammation. Adapted from Noor *et al*[193].

Interestingly, we observed that domination by a single taxon was already common in baseline samples, which are the earliest samples collected for each patient upon admission, which again, are novel findings. Similarly, metabolite levels at baseline, specifically those of butyrate, isobutyrate, lactate and glucose were already different to levels observed in HC, which raises an interesting premise. It is tempting to speculate that the patients' epithelial colonocyte metabolic state is already altered prior to admission. The exact role of the underlying clinical condition and subsequent drug use on colonocyte homeostasis during transplantation is not known. It is also unclear whether the underlying clinical condition may contribute to our observations or whether colonocyte metabolic state switching is largely at play. It may be that the underlying clinical condition and the associated treatment contributes towards the phenomenon prior to admission and the HSCT exacerbates this, leading to further pathogen expansion. Condition-specific metabolite signatures would be of interest in the future. It would also be interesting to investigate whether there is a gradient in colonic epithelial metabolic dysregulation in these patients and whether this is reversible with SCFA administration.

Another interesting finding is the stratification of samples into three CSTs. Although an artificial stratification, it revealed instinctive patterns within the data. CST1 was generally followed by a transition into CST2 or CST3. Additionally both CST2 and CST3, predominated by dominant taxa such as *Enterococcaceae* and *Staphylococcaceae*, were stable in comparison to CST1, meaning that once the gut is dominated by a specific taxon it is more likely to remain in that CST than to return to a more diverse CST1. This is an important finding, which may help with deciding the best time for interventions in this population.

The impact of this work is two-fold. On the one hand, the patterns observed from the longitudinal data of the gut microbiota and metabolome gives us an insight into the basic dynamics of the microbiota and the status of the host and microbiota metabolism during HSCT. Biomarker discovery on the other hand, provides insights into the usefulness of the gut microbiota and metabolome, which may allow us to stratify patients at risk of specific outcomes. The findings collectively point to the gut microbiota as an important player in HSCT outcomes.

7.3 Interventions and future questions

Given the potential associations between intestinal microbiota and clinical outcomes following HSCT, the intestinal microbiota poses as an attractive modulation target. Modulating the gut microbiota may improve recovery and reduce risk of adverse clinical outcomes such as bacteraemia, viraemia and GvHD.

More broadly, being able to identify patients who *(a)* will develop domination and therefore greater propensity to certain adverse transplant-related outcomes and *(b)* undergo delayed reconstitution, will be of utmost importance as appropriate identification could result in improved clinical management in the future.

FMT has recently been undertaken in HSCT adult recipients, whereby autologous FMT was administered post-transplantation, upon observing a decrease in the *Bacteroidetes* phylum[195]. The most recent report shows promising results, where autologous FMT led to improved diversity and re-establishment of the intestinal microbiota composition observed prior to transplantation. Despite success in adults, many challenges remain. It is unclear how useful autologous FMT would

be in the paediatric HSCT population. FMT may be of less use in children under the age of two, since their microbiome is still developing. More importantly, most patients present with a dysregulated microbiota at admission, highlighting major differences compared to healthy controls, therefore returning the composition to its initial composition may not be favourable for these patients. An FMT trial for paediatric patients with refractory GvHD is currently underway in Shanghai (ClinicalTrials.gov Identifier: NCT03148743).

Another plausible approach may be the administration of pre- and probiotics, for which there is limited evidence in this population. A recent trial supplemented adult HSCT patients with *Lactobacillus rhamnosus* starting at day -7 to day 14 relative to transplantation with the intent to modulate incidence of GvHD[196], the trial was halted as there was no evidence of benefit. A similar trial with *Lactobacillus plantarum* and another administering inulin starting at day -7 relative to transplantation are currently underway in paediatric HSCT patients (ClinicalTrials.gov Identifiers: NCT03057054/NCT04111471). Although it is feasible to think that prebiotics and probiotics may be useful in improving the microbiome post-transplantation, recent studies highlight individualised responses to probiotics and their ability to slow microbiome recovery post antibiotic administration, which makes finding a useful probiotic in an already heterogeneous population challenging[197,198]. Faecal filtrates to administer beneficial bacterial metabolite by-products may also be useful.

Additionally, antibiotic interventions, such as modulating administration to prevent *Enterococcus* or *Proteobacteria* domination may be helpful. Current trials in include comparing cefepime *versus* piperacillin-tazobactam use during febrile neutropenia in adults and assessing total gut decontamination *versus* selective gut decontamination in children (ClinicalTrials.gov Identifiers: NCT03078010/NCT02641236).

Investigating microbial metabolites and their effects on the host's metabolism and immune system offers another approach for potential interventions in this cohort.

The damage to the gut as a result of conditioning and antimicrobial administration mean that approaches for tissue regeneration and protection of the intestinal epithelium are of interest. The finding that the gut microbiota is able to maintain

epithelial cell homeostasis via the PPAR-γ receptor makes this a potentially attractive target in order to shift colonocyte metabolism towards β-oxidation[184]. Topical administration of the PPAR-γ receptor agonist, 5-aminosalicylic acid (5-ASA), is common in inflammatory bowel diseases and administration of 5-ASA to a murine model reduced crypt hyperplasia[199,200]. Whether this could be useful in the paediatric HSCT population remains to be seen. Given the complexity and heterogeneity of the HSCT paediatric population, this may be a more useful approach than prebiotics or FMT and may be a way to restore microbial-host homeostasis, at least for a defined period of time.

Dietary or intravenous supplementation with microbial metabolites, which may bypass some of the complexities associated with giving pre- and probiotics, is another potential approach. Butyrate-supplemented TPN for example, resulted in increased production of tight junction proteins and antimicrobial peptides in the ileal mucosa in a murine model, which means that it alleviated TPN-induced intestinal damage and thus could be useful in our population[201]. Butyrate also led to a decrease in Treg cells, which are thought to play a part in limiting inflammatory responses in the intestine and thus reflecting lower inflammation of the intestinal epithelium[201]. Similarly, butyrate administration in a murine *C. rodentium* model reduced histological scores of intestinal inflammation as well as increasing levels of IL-10 and TGF-β, which are known to be anti-inflammatory, and Muc2, a protein that is an integral part of the mucous layer and prevents bacterial access to the epithelium[202]. In contrast, administration of butyrate via enema, although safe in healthy humans, did not alter colonic expression of Muc2 and TFF3, although it is reported to indirectly interact with both receptors[203].

This highlights that administration of microbial by-products may be more complicated than previously thought. Golob *et al* have recently found that despite the perceived benefits of butyrate, the presence of butyrogenic bacteria at the onset of GvHD is associated with the development of steroid-refractory GvHD as butyrate is able to inhibit colonic stem cells[204]. Similarly, in a healthy gut, butyrate is metabolised by differentiated stem cells, thus protecting stem cells from undue butyrate exposure[205]. During HSCT, where colonocytes may be damaged and lower in numbers administering butyrate may worsen epithelial repair and the observed loss of butyrate producers may actually be beneficial for the diseased

host, whose gut may not be in the best condition to tolerate luminal butyrate. Despite this, metabolite administration as a potential therapeutic intervention warrants further exploration.

It would be of interest to explore microbial metabolites, such as tryptophan and indole derivatives as well as primary and secondary bile acids further as the gut microbiota is known to be involved in their metabolism. Additionally, it would be of interest to investigate associations between specific metabolite levels and taxa found in 16S rRNA analysis. This may help with elucidating metabolic patterns throughout HSCT. We did not observe identifiable bile acids here, therefore more targeted approaches, such as liquid chromatography-mass spectrometry, may be necessary.

In addition to this, investigating the effect of current dietary supplementation on the gut microbiota and relevant metabolites, such as TPN and EN would be of interest. A recent study compared gut microbiota composition between patients predominately on either TPN or EN. Patients predominately on EN had greater levels of *R. bromii* and *Faecalibacterium,* the latter being a primary carbohydrate degrader, whereas those predominately on TPN had higher levels of *Enterococcus* and *Klebsiella*[67,206]. The effect of nutrition throughout HSCT on the gut microbiota and metabolome in paediatric patients has not been investigated to date. It would additionally be useful to assess the levels of resistant starch degraders at baseline as well as throughout transplantation in this cohort, as the resultant SCFA are dependent on their presence. This may inform us on whether probiotics would be useful in this population. Finally, as previously mentioned the effects of microbial domination on immune reconstitution and on viral reactivation in this patient population would be of interest.

In this work I have explored the longitudinal dynamics of the gut microbiota of paediatric HSCT recipients. Whilst similar to adults in many respects, I also found certain unique characteristic of the cohort such as domination by specific taxa. Additionally, we found longitudinal metabolite patterns, some of which have not been previously reported on in this cohort to date. Validation and further exploration of this data is warranted, before moving into in vitro/murine models to

recapitulate and confirm some of these findings and considering interventional approaches.

References

1 Passweg, J. R. *et al.* Hematopoietic SCT in Europe: data and trends in 2012 with special consideration of pediatric transplantation. *Bone Marrow Transplant* **49**, 744-750, doi:10.1038/bmt.2014.55 (2014).

2 Barker, J. N. & Wagner, J. E. Umbilical-cord blood transplantation for the treatment of cancer. *Nat Rev Cancer* **3**, 526-532, doi:10.1038/nrc1125 (2003).

3 Choo, S. Y. The HLA system: genetics, immunology, clinical testing, and clinical implications. *Yonsei Med J* **48**, 11-23, doi:10.3349/ymj.2007.48.1.11 (2007).

4 Tiercy, J. M. How to select the best available related or unrelated donor of hematopoietic stem cells? *Haematologica* **101**, 680-687, doi:10.3324/haematol.2015.141119 (2016).

5 Elfeky, R. *et al.* New graft manipulation strategies improve the outcome of mismatched stem cell transplantation in children with primary immunodeficiencies. *J Allergy Clin Immunol* **144**, 280-293, doi:10.1016/j.jaci.2019.01.030 (2019).

6 Miano, M. *et al.* Haematopoietic stem cell transplantation trends in children over the last three decades: a survey by the paediatric diseases working party of the European Group for Blood and Marrow Transplantation. *Bone Marrow Transplant* **39**, 89-99, doi:10.1038/sj.bmt.1705550 (2007).

7 *UK & ROI Transplant Table Indications*, <http://bsbmtct.org/activity/2018/> (2018).

8 Gkazi, A. S. *et al.* Clinical T Cell Receptor Repertoire Deep Sequencing and Analysis: An Application to Monitor Immune Reconstitution Following Cord Blood Transplantation. *Front Immunol* **9**, 2547, doi:10.3389/fimmu.2018.02547 (2018).

9 Gluckman, E. & Rocha, V. Cord blood transplantation: state of the art. *Haematologica* **94**, 451-454, doi:10.3324/haematol.2009.005694 (2009).

10 Keating, A. K. *et al.* The influence of stem cell source on transplant outcomes for pediatric patients with acute myeloid leukemia. *Blood Adv* **3**, 1118-1128, doi:10.1182/bloodadvances.2018025908 (2019).

11 Arndt, C., Beck, J. F. & Gruhn, B. A pediatric prognostic score for patients undergoing allogeneic hematopoietic stem cell transplantation. *Eur J Haematol* **93**, 509-515, doi:10.1111/ejh.12390 (2014).

12 Henig, I. & Zuckerman, T. Hematopoietic stem cell transplantation-50 years of evolution and future perspectives. *Rambam Maimonides Med J* **5**, e0028, doi:10.5041/RMMJ.10162 (2014).

13 Ringdén, O. *et al.* A prospective randomized toxicity study to compare reduced-intensity and myeloablative conditioning in patients with myeloid leukaemia undergoing allogeneic haematopoietic stem cell transplantation. *J Intern Med* **274**, 153-162, doi:10.1111/joim.12056 (2013).

14 Lucchini, G. *et al.* Impact of Conditioning Regimen on Outcomes for Children with Acute Myeloid Leukemia Undergoing Transplantation in First Complete Remission. An Analysis on Behalf of the Pediatric Disease Working Party of the European Group for Blood and Marrow Transplantation. *Biol Blood Marrow Transplant* **23**, 467-474, doi:10.1016/j.bbmt.2016.11.022 (2017).

15 Khandelwal, P. *et al.* Hematopoietic Stem Cell Transplantation Activity in Pediatric Cancer between 2008 and 2014 in the United States: A Center for International Blood and Marrow Transplant Research Report. *Biol Blood Marrow Transplant* **23**, 1342-1349, doi:10.1016/j.bbmt.2017.04.018 (2017).

16 Stern, L. *et al.* Mass Cytometry for the Assessment of Immune Reconstitution After Hematopoietic Stem Cell Transplantation. *Front Immunol* **9**, 1672, doi:10.3389/fimmu.2018.01672 (2018).

17 Hatzimichael E, T. M. Hematopoietic stem cell transplantation. *Stem Cells* **3**, 105-117 (2010).

18 Gonçalves, T. L., Benvegnú, D. M. & Bonfanti, G. Specific factors influence the success of autologous and allogeneic hematopoietic stem cell transplantation. *Oxid Med Cell Longev* **2**, 82-87, doi:10.4161/oxim.2.2.8355 (2009).

19 Elfeky, R., Lazareva, A., Qasim, W. & Veys, P. Immune reconstitution following hematopoietic stem cell transplantation using different stem cell sources. *Expert Rev Clin Immunol* **15**, 735-751, doi:10.1080/1744666X.2019.1612746 (2019).

20 Zaucha-Prazmo, A. *et al.* Transplant-related mortality and survival in children with malignancies treated with allogeneic hematopoietic stem cell transplantation. A multicenter analysis. *Pediatr Transplant* **22**, e13158, doi:10.1111/petr.13158 (2018).

21 Mohty, B. & Mohty, M. Long-term complications and side effects after allogeneic hematopoietic stem cell transplantation: an update. *Blood Cancer J* **1**, e16, doi:10.1038/bcj.2011.14 (2011).

22 Perez, P. *et al.* Bacteremia in pediatric patients with hematopoietic stem cell transplantation. *Hematol Transfus Cell Ther*, doi:10.1016/j.htct.2019.05.006 (2019).

23 Poutsiaka, D. D. *et al.* Blood stream infection after hematopoietic stem cell transplantation is associated with increased mortality. *Bone Marrow Transplant* **40**, 63-70, doi:10.1038/sj.bmt.1705690 (2007).

24 Styczynski, J., Gil, L. & Party, E. P. D. W. Prevention of infectious complications in pediatric HSCT. *Bone Marrow Transplant* **42 Suppl 2**, S77-81, doi:10.1038/bmt.2008.289 (2008).
25 Gennery AR, M. P. Infection following haematopoietic stem cell transplantation. (2014).
26 Robles, J. D., Cheuk, D. K., Ha, S. Y., Chiang, A. K. & Chan, G. C. Norovirus infection in pediatric hematopoietic stem cell transplantation recipients: incidence, risk factors, and outcome. *Biol Blood Marrow Transplant* **18**, 1883-1889, doi:10.1016/j.bbmt.2012.07.005 (2012).
27 Meisel, R. *et al.* Comparable long-term survival after bone marrow versus peripheral blood progenitor cell transplantation from matched unrelated donors in children with hematologic malignancies. *Biol Blood Marrow Transplant* **13**, 1338-1345, doi:10.1016/j.bbmt.2007.07.009 (2007).
28 Wagner, J. E., Kernan, N. A., Steinbuch, M., Broxmeyer, H. E. & Gluckman, E. Allogeneic sibling umbilical-cord-blood transplantation in children with malignant and non-malignant disease. *Lancet* **346**, 214-219, doi:10.1016/s0140-6736(95)91268-1 (1995).
29 Baird, K., Cooke, K. & Schultz, K. R. Chronic graft-versus-host disease (GVHD) in children. *Pediatr Clin North Am* **57**, 297-322, doi:10.1016/j.pcl.2009.11.003 (2010).
30 Kato, M. *et al.* Impact of graft-versus-host disease on relapse and survival after allogeneic stem cell transplantation for pediatric leukemia. *Bone Marrow Transplant* **54**, 68-75, doi:10.1038/s41409-018-0221-6 (2019).
31 Blazar, B. R., Murphy, W. J. & Abedi, M. Advances in graft-versus-host disease biology and therapy. *Nat Rev Immunol* **12**, 443-458, doi:10.1038/nri3212 (2012).
32 Holler, E. *et al.* Metagenomic Analysis of the Stool Microbiome in Patients Receiving Allogeneic Stem Cell Transplantation: Loss of Diversity Is Associated with Use of Systemic Antibiotics and More Pronounced in Gastrointestinal Graft-versus-Host Disease. *Biology of Blood and Marrow Transplantation* **20**, 640-645, doi:10.1016/j.bbmt.2014.01.030 (2014).
33 Simms-Waldrip, T. R. *et al.* Antibiotic-Induced Depletion of Anti-inflammatory Clostridia Is Associated with the Development of Graft-versus-Host Disease in Pediatric Stem Cell Transplantation Patients. *Biol Blood Marrow Transplant* **23**, 820-829, doi:10.1016/j.bbmt.2017.02.004 (2017).
34 Mathewson, N. D. *et al.* Gut microbiome-derived metabolites modulate intestinal epithelial cell damage and mitigate graft-versus-host disease (vol 17, pg 505, 2016). *Nature Immunology* **17**, 1235-1235, doi:DOI 10.1038/ni1016-1235b (2016).
35 Swimm, A. *et al.* Indoles derived from intestinal microbiota act via type I interferon signaling to limit graft-versus-host disease. *Blood* **132**, 2506-2519, doi:10.1182/blood-2018-03-838193 (2018).
36 Stripling, J. & Rodriguez, M. Current Evidence in Delivery and Therapeutic Uses of Fecal Microbiota Transplantation in Human Diseases-Clostridium difficile Disease and Beyond. *American Journal of the Medical Sciences* **356**, 424-432, doi:DOI 10.1016/j.amjms.2018.08.010 (2018).
37 Turnbaugh, P. J. *et al.* An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* **444**, 1027-1031, doi:10.1038/nature05414 (2006).
38 Huttenhower, C. *et al.* Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207-214, doi:10.1038/nature11234 (2012).
39 Segata, N. *et al.* Composition of the adult digestive tract bacterial microbiome based on seven mouth surfaces, tonsils, throat and stool samples. *Genome Biology* **13**, doi:ARTN R42 10.1186/gb-2012-13-6-r42 (2012).
40 Aagaard, K. *et al.* The Placenta Harbors a Unique Microbiome. *Science Translational Medicine* **6**, doi:ARTN 237ra65 10.1126/scitranslmed.3008599 (2014).
41 de Goffau M.C, L. S., Sovio U, Gaccioli F, Cook E, Peacock S.J, Parkhill J, Charnock-Jones S, Smith G.C.S. Human placenta has no microbiome but can contain potential pathogens. *Nature* (2019).
42 Stinson, L. F., Boyce, M. C., Payne, M. S. & Keelan, J. A. The Not-so-Sterile Womb: Evidence That the Human Fetus Is Exposed to Bacteria Prior to Birth. *Front Microbiol* **10**, 1124, doi:10.3389/fmicb.2019.01124 (2019).
43 Perez-Munoz, M. E., Arrieta, M. C., Ramer-Tait, A. E. & Walter, J. A critical assessment of the "sterile womb" and "in utero colonization" hypotheses: implications for research on the pioneer infant microbiome. *Microbiome* **5**, doi:ARTN 48 10.1186/s40168-017-0268-4 (2017).
44 Stewart, C. J. *et al.* Temporal development of the gut microbiome in early childhood from the TEDDY study. *Nature* **562**, 583-+, doi:10.1038/s41586-018-0617-x (2018).
45 Derrien, M., Alvarez, A. S. & de Vos, W. M. The Gut Microbiota in the First Decade of Life. *Trends Microbiol* **27**, 997-1010, doi:10.1016/j.tim.2019.08.001 (2019).
46 Hansen, C. H. F. *et al.* Patterns of Early Gut Colonization Shape Future Immune Responses of the Host. *Plos One* **7**, doi:ARTN e34043 10.1371/journal.pone.0034043 (2012).
47 Sommer, F. & Backhed, F. The gut microbiota - masters of host development and physiology. *Nature Reviews Microbiology* **11**, 227-238, doi:10.1038/nrmicro2974 (2013).

48 Ihekweazu, F. D. & Versalovic, J. Development of the Pediatric Gut Microbiome: Impact on Health and Disease. *American Journal of the Medical Sciences* **356**, 413-423, doi:DOI 10.1016/j.amjms.2018.08.005 (2018).
49 Fujimura, K. E. *et al.* Neonatal gut microbiota associates with childhood multisensitized atopy and T cell differentiation. *Nature Medicine* **22**, 1187-1191, doi:10.1038/nm.4176 (2016).
50 Mauras, A. *et al.* Gut microbiota from infant with cow's milk allergy promotes clinical and immune features of atopy in a murine model. *Allergy* **74**, 1790-1793, doi:10.1111/all.13787 (2019).
51 Turta, O. & Rautava, S. Antibiotics, obesity and the link to microbes - what are we doing to our children? *Bmc Medicine* **14**, doi:ARTN 57 10.1186/s12916-016-0605-7 (2016).
52 Belkaid, Y. & Hand, T. W. Role of the Microbiota in Immunity and Inflammation. *Cell* **157**, 121-141, doi:10.1016/j.cell.2014.03.011 (2014).
53 Olszak, T. *et al.* Microbial Exposure During Early Life Has Persistent Effects on Natural Killer T Cell Function. *Science* **336**, 489-493, doi:10.1126/science.1219328 (2012).
54 Russell, S. L. *et al.* Early life antibiotic-driven changes in microbiota enhance susceptibility to allergic asthma. *Embo Reports* **13**, 440-447, doi:10.1038/embor.2012.32 (2012).
55 Ni, J. *et al.* Early antibiotic exposure and development of asthma and allergic rhinitis in childhood. *Bmc Pediatrics* **19**, doi:10.1186/s12887-019-1594-4 (2019).
56 Donaldson, G. P., Lee, S. M. & Mazmanian, S. K. Gut biogeography of the bacterial microbiota. *Nature Reviews Microbiology* **14**, 20-32, doi:10.1038/nrmicro3552 (2016).
57 Atarashi, K. *et al.* Induction of Colonic Regulatory T Cells by Indigenous Clostridium Species. *Science* **331**, 337-341, doi:10.1126/science.1198469 (2011).
58 Peterson, D. A., McNulty, N. P., Guruge, J. L. & Gordon, J. I. IgA response to symbiotic bacteria as a mediator of gut homeostasis. *Cell Host & Microbe* **2**, 328-339, doi:10.1016/j.chom.2007.09.013 (2007).
59 Lloyd-Price, J., Abu-Ali, G. & Huttenhower, C. The healthy human microbiome. *Genome Med* **8**, 51, doi:10.1186/s13073-016-0307-y (2016).
60 Tap, J. *et al.* Towards the human intestinal microbiota phylogenetic core. *Environmental Microbiology* **11**, 2574-2584, doi:10.1111/j.1462-2920.2009.01982.x (2009).
61 David, L. A. *et al.* Diet rapidly and reproducibly alters the human gut microbiome. *Nature* **505**, 559-+, doi:10.1038/nature12820 (2014).
62 David, L. A. *et al.* Host lifestyle affects human microbiota on daily timescales. *Genome Biology* **15**, doi:ARTN R89 10.1186/gb-2014-15-7-r89 (2014).
63 Martinson, J. N. V. *et al.* Rethinking gut microbiome residency and the Enterobacteriaceae in healthy human adults. *Isme Journal* **13**, 2306-2318, doi:10.1038/s41396-019-0435-7 (2019).
64 Schluter, J. & Foster, K. R. The Evolution of Mutualism in Gut Microbiota Via Host Epithelial Selection. *Plos Biology* **10**, doi:10.1371/journal.pbio.1001424 (2012).
65 Beckerson, J. *et al.* Impact of route and adequacy of nutritional intake on outcomes of allogeneic haematopoietic cell transplantation for haematologic malignancies. *Clin Nutr* **38**, 738-744, doi:10.1016/j.clnu.2018.03.008 (2019).
66 Guièze, R. *et al.* Enteral versus parenteral nutritional support in allogeneic haematopoietic stem-cell transplantation. *Clin Nutr* **33**, 533-538, doi:10.1016/j.clnu.2013.07.012 (2014).
67 Andersen, S. *et al.* Pilot study investigating the effect of enteral and parenteral nutrition on the gastrointestinal microbiome post-allogeneic transplantation. *British Journal of Haematology*, doi:10.1111/bjh.16218 (2019).
68 Gafter-Gvili, A. *et al.* Antibiotic prophylaxis for bacterial infections in afebrile neutropenic patients following chemotherapy. *Cochrane Database Syst Rev* **1**, CD004386, doi:10.1002/14651858.CD004386.pub3 (2012).
69 Taur, Y. *et al.* The effects of intestinal tract bacterial diversity on mortality following allogeneic hematopoietic stem cell transplantation. *Blood* **124**, 1174-1182, doi:10.1182/blood-2014-02-554725 (2014).
70 Shono, Y. *et al.* Increased GVHD-related mortality with broad-spectrum antibiotic use after allogeneic hematopoietic stem cell transplantation in human patients and mice. *Science Translational Medicine* **8**, doi:10.1126/scitranslmed.aaf2311 (2016).
71 D'Amico, F. *et al.* Gut resistome plasticity in pediatric patients undergoing hematopoietic stem cell transplantation. *Scientific Reports* **9**, doi:10.1038/s41598-019-42222-w (2019).
72 Routy, B. *et al.* Gut microbiome influences efficacy of PD-1-based immunotherapy against epithelial tumors. *Science* **359**, 91-+, doi:10.1126/science.aan3706 (2018).
73 Jakobsson, H. E. *et al.* Short-Term Antibiotic Treatment Has Differing Long-Term Impacts on the Human Throat and Gut Microbiome. *Plos One* **5**, doi:10.1371/journal.pone.0009836 (2010).
74 Dethlefsen, L. & Relman, D. A. Incomplete recovery and individualized responses of the human distal gut microbiota to repeated antibiotic perturbation. *Proceedings of the*

National Academy of Sciences of the United States of America **108**, 4554-4561, doi:10.1073/pnas.1000087107 (2011).

75 Palleja, A. *et al.* Recovery of gut microbiota of healthy adults following antibiotic exposure. *Nat Microbiol* **3**, 1255-1265, doi:10.1038/s41564-018-0257-9 (2018).

76 van Bekkum, D. W. & Knaan, S. Role of bacterial microflora in development of intestinal lesions from graft-versus-host reaction. *J Natl Cancer Inst* **58**, 787-790, doi:10.1093/jnci/58.3.787 (1977).

77 Heit, H., Heit, W., Kohne, E., Fliedner, T. M. & Hughes, P. Allogeneic bone marrow transplantation in conventional mice: I. Effect of antibiotic therapy on long term survival of allogeneic chimeras. *Blut* **35**, 143-153, doi:10.1007/bf00996294 (1977).

78 Routy, B. *et al.* The influence of gut-decontamination prophylactic antibiotics on acute graft-versus-host disease and survival following allogeneic hematopoietic stem cell transplantation. *Oncoimmunology* **6**, e1258506, doi:10.1080/2162402X.2016.1258506 (2017).

79 Taur, Y. *et al.* Intestinal domination and the risk of bacteremia in patients undergoing allogeneic hematopoietic stem cell transplantation. *Clin Infect Dis* **55**, 905-914, doi:10.1093/cid/cis580 (2012).

80 Ubeda, C. *et al.* Vancomycin-resistant Enterococcus domination of intestinal microbiota is enabled by antibiotic treatment in mice and precedes bloodstream invasion in humans. *J Clin Invest* **120**, 4332-4341, doi:10.1172/JCI43918 (2010).

81 Taur, Y., Jenq, R. R., Ubeda, C., van den Brink, M. & Pamer, E. G. Role of intestinal microbiota in transplantation outcomes. *Best Practice & Research Clinical Haematology* **28**, 155-161, doi:10.1016/j.beha.2015.10.013 (2015).

82 Biagi, E. *et al.* Gut microbiota trajectory in pediatric patients undergoing hematopoietic SCT. *Bone Marrow Transplant* **50**, 992-998, doi:10.1038/bmt.2015.16 (2015).

83 Baxter, N. T. *et al.* Dynamics of Human Gut Microbiota and Short-Chain Fatty Acids in Response to Dietary Interventions with Three Fermentable Fibers. *Mbio* **10**, doi:10.1128/mBio.02566-18 (2019).

84 Lahteenmaki, K. *et al.* Haematopoietic stem cell transplantation induces severe dysbiosis in intestinal microbiota of paediatric ALL patients. *Bone Marrow Transplantation* **52**, 1479-1482, doi:10.1038/bmt.2017.168 (2017).

85 Peled, J. U. *et al.* Intestinal Microbiota and Relapse After Hematopoietic-Cell Transplantation. *J Clin Oncol* **35**, 1650-1659, doi:10.1200/JCO.2016.70.3348 (2017).

86 Hakim, H. *et al.* Gut Microbiome Composition Predicts Infection Risk during Chemotherapy in Children with Acute Lymphoblastic Leukemia. *Clin Infect Dis*, doi:10.1093/cid/ciy153 (2018).

87 Kelly, M. S. *et al.* Gut Colonization Preceding Mucosal Barrier Injury Bloodstream Infection in Pediatric Hematopoietic Stem Cell Transplantation Recipients. *Biol Blood Marrow Transplant*, doi:10.1016/j.bbmt.2019.07.019 (2019).

88 Jenq, R. R. *et al.* Intestinal Blautia Is Associated with Reduced Death from Graft-versus-Host Disease. *Biology of Blood and Marrow Transplantation* **21**, 1373-1383, doi:10.1016/j.bbmt.2015.04.016 (2015).

89 Jenq, R. R. *et al.* Regulation of intestinal inflammation by microbiota following allogeneic bone marrow transplantation. *J Exp Med* **209**, 903-911, doi:10.1084/jem.20112408 (2012).

90 Varelias, A. *et al.* Acute graft-versus-host disease is regulated by an IL-17-sensitive microbiome. *Blood* **129**, 2172-2185, doi:10.1182/blood-2016-08-732628 (2017).

91 Weber, D. *et al.* Low urinary indoxyl sulfate levels early after transplantation reflect a disrupted microbiome and are associated with poor outcome. *Blood* **126**, 1723-1728, doi:10.1182/blood-2015-04-638858 (2015).

92 Rayes, A. *et al.* A Genetic Modifier of the Gut Microbiome Influences the Risk of Graft-versus-Host Disease and Bacteremia After Hematopoietic Stem Cell Transplantation. *Biology of Blood and Marrow Transplantation* **22**, 418-422, doi:10.1016/j.bbmt.2015.11.017 (2016).

93 Mancini, N. *et al.* Enteric Microbiome Markers as Early Predictors of Clinical Outcome in Allogeneic Hematopoietic Stem Cell Transplant: Results of a Prospective Study in Adult Patients. *Open Forum Infect Dis* **4**, ofx215, doi:10.1093/ofid/ofx215 (2017).

94 Montassier, E. *et al.* Pretreatment gut microbiome predicts chemotherapy-related bloodstream infection. *Genome Med* **8**, 49, doi:10.1186/s13073-016-0301-4 (2016).

95 Golob, J. L. *et al.* Stool Microbiota at Neutrophil Recovery Is Predictive for Severe Acute Graft vs Host Disease After Hematopoietic Cell Transplantation. *Clinical Infectious Diseases* **65**, 1984-1991, doi:10.1093/cid/cix699 (2017).

96 Ingham, A. C. *et al.* Specific gut microbiome members are associated with distinct immune markers in pediatric allogeneic hematopoietic stem cell transplantation. *Microbiome* **7**, doi:ARTN 131 10.1186/s40168-019-0745-z (2019).

97 Glucksberg, H. *et al.* Clinical manifestations of graft-versus-host disease in human recipients of marrow from HL-A-matched sibling donors. *Transplantation* **18**, 295-304, doi:10.1097/00007890-197410000-00001 (1974).

98 *Bloodstream infection event (central line-associated bloodstream infection and non-central line-associated bloodstream infection)*, <https://www.cdc.gov/nhsn/pdfs/pscmanual/4psc_clabscurrent.pdf> (2018).

99 Kozich, J. J., Westcott, S. L., Baxter, N. T., Highlander, S. K. & Schloss, P. D. Development of a Dual-Index Sequencing Strategy and Curation Pipeline for Analyzing Amplicon Sequence Data on the MiSeq Illumina Sequencing Platform. *Applied and Environmental Microbiology* **79**, 5112-5120, doi:10.1128/Aem.01043-13 (2013).
100 Schloss, P. D. *et al.* Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* **75**, 7537-7541, doi:10.1128/AEM.01541-09 (2009).
101 Lagkouvardos, I., Fischer, S., Kumar, N. & Clavel, T. Rhea: a transparent and modular R pipeline for microbial profiling based on 16S rRNA gene amplicons. *PeerJ* **5**, e2836, doi:10.7717/peerj.2836 (2017).
102 van der Maaten, L. & Hinton, G. Visualizing Data using t-SNE. *Journal of Machine Learning Research* **9**, 2579-2605 (2008).
103 Therneau, T. A Package for Survival Analysis in S. version 2.38. (2015).
104 Segata, N. *et al.* Metagenomic biomarker discovery and explanation. *Genome Biology* **12**, doi:ARTN R60 10.1186/gb-2011-12-6-r60 (2011).
105 Veselkov, K. A. *et al.* Recursive segment-wise peak alignment of biological (1)h NMR spectra for improved metabolic biomarker recovery. *Anal Chem* **81**, 56-66, doi:10.1021/ac8011544 (2009).
106 Dieterle, F., Ross, A., Schlotterbeck, G. & Senn, H. Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1H NMR metabonomics. *Anal Chem* **78**, 4281-4290, doi:10.1021/ac051632c (2006).
107 Cloarec, O. *et al.* Statistical total correlation spectroscopy: an exploratory approach for latent biomarker identification from metabolic 1H NMR data sets. *Anal Chem* **77**, 1282-1289, doi:10.1021/ac048630x (2005).
108 Clooney, A. G. *et al.* Comparing Apples and Oranges?: Next Generation Sequencing and Its Impact on Microbiome Analysis. *PLoS One* **11**, e0148028, doi:10.1371/journal.pone.0148028 (2016).
109 Yuan, S., Cohen, D. B., Ravel, J., Abdo, Z. & Forney, L. J. Evaluation of methods for the extraction and purification of DNA from the human microbiome. *PLoS One* **7**, e33865, doi:10.1371/journal.pone.0033865 (2012).
110 Wu, G. D. *et al.* Sampling and pyrosequencing methods for characterizing bacterial communities in the human gut using 16S sequence tags. *BMC Microbiol* **10**, 206, doi:10.1186/1471-2180-10-206 (2010).
111 Maukonen, J., Simoes, C. & Saarela, M. The currently used commercial DNA-extraction methods give different results of clostridial and actinobacterial populations derived from human fecal samples. *FEMS Microbiol Ecol* **79**, 697-708, doi:10.1111/j.1574-6941.2011.01257.x (2012).
112 Janabi, A. H., Kerkhof, L. J., McGuinness, L. R., Biddle, A. S. & McKeever, K. H. Comparison of a modified phenol/chloroform and commercial-kit methods for extracting DNA from horse fecal material. *J Microbiol Methods* **129**, 14-19, doi:10.1016/j.mimet.2016.07.019 (2016).
113 Mackenzie, B. W., Waite, D. W. & Taylor, M. W. Evaluating variation in human gut microbiota profiles due to DNA extraction method and inter-subject differences. *Frontiers in Microbiology* **6**, doi:10.3389/fmicb.2015.00130 (2015).
114 Wesolowska-Andersen, A. *et al.* Choice of bacterial DNA extraction method from fecal material influences community structure as evaluated by metagenomic analysis. *Microbiome* **2**, doi:10.1186/2049-2618-2-19 (2014).
115 de Boer, R. *et al.* Improved detection of microbial DNA after bead-beating before DNA isolation. *Journal of Microbiological Methods* **80**, 209-211, doi:10.1016/j.mimet.2009.11.009 (2010).
116 Santiago, A. *et al.* Processing faecal samples: a step forward for standards in microbial community analysis. *Bmc Microbiology* **14**, doi:Artn 112 10.1186/1471-2180-14-112 (2014).
117 Kim, M., Morrison, M. & Yu, Z. T. Evaluation of different partial 16S rRNA gene sequence regions for phylogenetic analysis of microbiomes. *Journal of Microbiological Methods* **84**, 81-87, doi:10.1016/j.mimet.2010.10.020 (2011).
118 Wang, Q., Garrity, G. M., Tiedje, J. M. & Cole, J. R. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol* **73**, 5261-5267, doi:10.1128/AEM.00062-07 (2007).
119 Chakravorty, S., Helb, D., Burday, M., Connell, N. & Alland, D. A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. *J Microbiol Methods* **69**, 330-339, doi:10.1016/j.mimet.2007.02.005 (2007).
120 Claesson, M. J. *et al.* Comparison of two next-generation sequencing technologies for resolving highly complex microbiota composition using tandem variable 16S rRNA gene regions. *Nucleic Acids Research* **38**, doi:10.1093/nar/gkq873 (2010).
121 Bukin, Y. S. *et al.* The effect of 16S rRNA region choice on bacterial community metabarcoding results. *Sci Data* **6**, 190007, doi:10.1038/sdata.2019.7 (2019).
122 Debelius, J. *et al.* Tiny microbes, enormous impacts: what matters in gut microbiome studies? *Genome Biology* **17**, doi:10.1186/s13059-016-1086-x (2016).
123 Lozupone, C. A. *et al.* Meta-analyses of studies of the human microbiota. *Genome Res* **23**, 1704-1714, doi:10.1101/gr.151803.112 (2013).

124 Fuks, G. *et al.* Combining 16S rRNA gene variable regions enables high-resolution microbial community profiling. *Microbiome* **6**, doi:10.1186/s40168-017-0396-x (2018).
125 Sinclair, L., Osman, O. A., Bertilsson, S. & Eiler, A. Microbial Community Composition and Diversity via 16S rRNA Gene Amplicons: Evaluating the Illumina Platform. *Plos One* **10**, doi:10.1371/journal.pone.0116955 (2015).
126 Schloss, P. D., Gevers, D. & Westcott, S. L. Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. *PLoS One* **6**, e27310, doi:10.1371/journal.pone.0027310 (2011).
127 Glassing, A., Dowd, S. E., Galandiuk, S., Davis, B. & Chiodini, R. J. Inherent bacterial DNA contamination of extraction and sequencing reagents may affect interpretation of microbiota in low bacterial biomass samples. *Gut Pathogens* **8**, doi:10.1186/s13099-016-0103-7 (2016).
128 Schloss, P. D. The Effects of Alignment Quality, Distance Calculation Method, Sequence Filtering, and Region on the Analysis of 16S rRNA Gene-Based Studies. *Plos Computational Biology* **6**, doi:10.1371/journal.pcbi.1000844 (2010).
129 Weyrich L.S *et al.* Laboratory contamination over time during low-biomass sample analysis. (2018).
130 Salter, S. J. *et al.* Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *Bmc Biology* **12**, doi:ARTN 87 10.1186/s12915-014-0087-z (2014).
131 Stinson, L. F., Keelan, J. A. & Payne, M. S. Identification and removal of contaminating microbial DNA from PCR reagents: impact on low-biomass microbiome analyses. *Lett Appl Microbiol* **68**, 2-8, doi:10.1111/lam.13091 (2019).
132 Doyle, R. M. *et al.* Bacterial communities found in placental tissues are associated with severe chorioamnionitis and adverse birth outcomes. *Plos One* **12**, doi:10.1371/journal.pone.0180167 (2017).
133 Joss, T. V. *et al.* Bacterial Communities Vary between Sinuses in Chronic Rhinosinusitis Patients. *Front Microbiol* **6**, 1532, doi:10.3389/fmicb.2015.01532 (2015).
134 Albertsen, M., Karst, S. M., Ziegler, A. S., Kirkegaard, R. H. & Nielsen, P. H. Back to Basics--The Influence of DNA Extraction and Primer Choice on Phylogenetic Analysis of Activated Sludge Communities. *PLoS One* **10**, e0132783, doi:10.1371/journal.pone.0132783 (2015).
135 Minich J.J, S. J. G., Amir A, Humphrey G, Gilbert J, Knight R. Quantifying and understanding well-to-well contamination in microbiome research. *BioRxiv* (2019).
136 Walker, A. W. A Lot on Your Plate? Well-to-Well Contamination as an Additional Confounder in Microbiome Sequence Analyses. *mSystems* **4** (2019).
137 Stammler, F. *et al.* Adjusting microbiome profiles for differences in microbial load by spike-in bacteria. *Microbiome* **4**, doi:10.1186/s40168-016-0175-0 (2016).
138 Marcel Martínez-Porchas, E. V.-C., Francisco Vargas-Albores. Significant loss of sensitivity and specificity in the taxonomic classification occurs when short 16S rRNA gene sequences are used. *Heliyon* **2** (2016).
139 Hong, S., Bunge, J., Leslin, C., Jeon, S. & Epstein, S. S. Polymerase chain reaction primers miss half of rRNA microbial diversity. *ISME J* **3**, 1365-1373, doi:10.1038/ismej.2009.89 (2009).
140 Wu, J. Y. *et al.* Effects of polymerase, template dilution and cycle number on PCR based 16 S rRNA diversity analysis using the deep sequencing method. *Bmc Microbiology* **10**, doi:10.1186/1471-2180-10-255 (2010).
141 Aird, D. *et al.* Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol* **12**, R18, doi:10.1186/gb-2011-12-2-r18 (2011).
142 Ahn, J. H., Kim, B. Y., Song, J. & Weon, H. Y. Effects of PCR cycle number and DNA polymerase type on the 16S rRNA gene pyrosequencing analysis of bacterial communities. *J Microbiol* **50**, 1071-1074, doi:10.1007/s12275-012-2642-z (2012).
143 Gohl, D. M. *et al.* Systematic improvement of amplicon marker gene methods for increased accuracy in microbiome studies. *Nat Biotechnol* **34**, 942-949, doi:10.1038/nbt.3601 (2016).
144 Suzuki, M. T. & Giovannoni, S. J. Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. *Appl Environ Microbiol* **62**, 625-630 (1996).
145 Kurata, S. *et al.* Reevaluation and reduction of a PCR bias caused by reannealing of templates. *Applied and Environmental Microbiology* **70**, 7545-7549, doi:10.1128/Aem.70.12.7545-7549.2004 (2004).
146 Kennedy, N. A. *et al.* The Impact of Different DNA Extraction Kits and Laboratories upon the Assessment of Human Gut Microbiota Composition by 16S rRNA Gene Sequencing. *Plos One* **9**, doi:10.1371/journal.pone.0088982 (2014).
147 Dore, J., Ehrlich, S.D., Levenez, F., Pelletier, E., Alberti, A., Bertrand, L., Bork, P., Costea, P.I., Sunagawa, S., Guarner, F., Manichanh, C., Santiago, A., Zhao, L., Shen, J., Zhang, C., Versalovic, J., Luna, R.A., Petrosino, J., Yang, H., Li, S., Wang, J., Allen-Vercoe, E., Gloor, G., Singh, B. and IHMS Consortium (2015). IHMS_SOP 06 V1: Standard operating procedure for fecal samples DNA extraction, Protocol Q. International Human Microbiome Standards. (2015).
148 Pollock, J., Glendinning, L., Wisedchanwet, T. & Watson, M. The Madness of Microbiome: Attempting To Find Consensus "Best Practice" for 16S Microbiome

Studies. *Applied and Environmental Microbiology* **84**, doi:UNSP e02627-17 10.1128/AEM.02627-17 (2018).
149 Bekker, V. *et al.* Microbiome dynamics during stem cell transplantation in children using total gut decontamination as graft-versus-host prophylaxis. *Bone Marrow Transplantation* **52**, S370-S371 (2019).
150 Schluter, J. *et al.* The gut microbiota influences how circulating immune cells in humans change from one day to the next. *BioRxiv*, doi:10.1101/618256 (2019).
151 Khan,N. *et al.* Loss of Microbiota Diversity after Autologous Stem Cell Transplant Is Comparable to Injury in Allogeneic Stem Cell Transplant. **132**, 608 (2018).
152 Kriss, M., Hazleton, K. Z., Nusbacher, N. M., Martin, C. G. & Lozupone, C. A. Low diversity gut microbiota dysbiosis: drivers, functional implications and recovery. *Current Opinion in Microbiology* **44**, 34-40, doi:10.1016/j.mib.2018.07.003 (2018).
153 Biagi, E. *et al.* Early gut microbiota signature of aGvHD in children given allogeneic hematopoietic cell transplantation for hematological disorders. *Bmc Medical Genomics* **12**, doi:ARTN 49 10.1186/s12920-019-0494-7 (2019).
154 Shono, Y. & van den Brink, M. R. M. Gut microbiota injury in allogeneic haematopoietic stem cell transplantation. *Nature Reviews Cancer* **18**, 283-295, doi:10.1038/nrc.2018.10 (2018).
155 Josefsdottir, K. S., Baldridge, M. T., Kadmon, C. S. & King, K. Y. Antibiotics impair murine hematopoiesis by depleting the intestinal microbiota. *Blood* **129**, 729-739, doi:10.1182/blood-2016-03-708594 (2017).
156 Khosravi, A. *et al.* Gut Microbiota Promote Hematopoiesis to Control Bacterial Infection. *Cell Host & Microbe* **15**, 374-381, doi:10.1016/j.chom.2014.02.006 (2014).
157 Staffas, A. *et al.* Nutritional Support from the Intestinal Microbiota Improves Hematopoietic Reconstitution after Bone Marrow Transplantation in Mice. *Cell Host Microbe* **23**, 447-457 e444, doi:10.1016/j.chom.2018.03.002 (2018).
158 Morjaria, S. *et al.* Antibiotic-Induced Shifts in Fecal Microbiota Density and Composition during Hematopoietic Stem Cell Transplantation. *Infection and Immunity* **87**, doi:ARTN e00206-19 10.1128/IAI.00206-19 (2019).
159 Maier, L. *et al.* Extensive impact of non-antibiotic drugs on human gut bacteria. *Nature* **555**, 623-628, doi:10.1038/nature25979 (2018).
160 Lavelle, A. *et al.* Baseline microbiota composition modulates antibiotic-mediated effects on the gut microbiota and host. *Microbiome* **7**, doi:ARTN 111 10.1186/s40168-019-0725-3 (2019).
161 Raymond, F. *et al.* The initial state of the human gut microbiome determines its reshaping by antibiotics. *ISME J* **10**, 707-720, doi:10.1038/ismej.2015.148 (2016).
162 Shaw L, B. C. P., Walker A.S, Klein N, Balloux F. A perturbation model of the gut microbiome's response to antibiotics. *BioRxiv* (2017).
163 Naum, M., Brown, E. W. & Mason-Gamer, R. J. Is 16S rDNA a reliable phylogenetic marker to characterize relationships below the family level in the Enterobacteriaceae? *Journal of Molecular Evolution* **66**, 630-642, doi:10.1007/s00239-008-9115-3 (2008).
164 Winkler, E. S. & Thackray, L. B. A long-distance relationship: the commensal gut microbiota and systemic viruses. *Current Opinion in Virology* **37**, 44-51, doi:10.1016/j.coviro.2019.05.009 (2019).
165 Tanaka, K., Sawamura, S., Satoh, T., Kobayashi, K. & Noda, S. Role of the indigenous microbiota in maintaining the virus-specific CD8 memory T cells in the lung of mice infected with murine cytomegalovirus. *J Immunol* **178**, 5209-5216, doi:10.4049/jimmunol.178.8.5209 (2007).
166 Holmes, I. H., K. Quince, K. Dirichlet Multinomial Mixtures: Generative Models for Microbial Metagenomics. *PLoS One* **7** (2012).
167 Nearing, J. T. *et al.* Infectious Complications Are Associated With Alterations in the Gut Microbiome in Pediatric Patients With Acute Lymphoblastic Leukemia. *Frontiers in Cellular and Infection Microbiology* **9**, doi:ARTN 28 10.3389/fcimb.2019.00028 (2019).
168 Han. L, Z. H., Chen. S, Zhou. L, Li. Y, Zhao. K, Huang. F, Fan. Z, Xuan. L, Zhang. X, Dai. M, Lin. Q, Jiang. Z, Peng. J, Jin. H, Liu. K. Intestinal Microbiota Can Predict Acute Graft-versus-Host Disease Following Allogeneic Hematopoietic Stem Cell Transplantation. *Biol Blood Marrow Transplant* **25** (2019).
169 Liu, C. *et al.* Associations between acute gastrointestinal GvHD and the baseline gut microbiota of allogeneic hematopoietic stem cell transplant recipients and donors. *Bone Marrow Transplantation* **52**, 1643-1650, doi:10.1038/bmt.2017.200 (2017).
170 Andermann, T. M. *et al.* The Microbiome and Hematopoietic Cell Transplantation: Past, Present, and Future. *Biol Blood Marrow Transplant*, doi:10.1016/j.bbmt.2018.02.009 (2018).
171 den Besten, G. *et al.* The role of short-chain fatty acids in the interplay between diet, gut microbiota, and host energy metabolism. *Journal of Lipid Research* **54**, 2325-2340, doi:10.1194/jlr.R036012 (2013).
172 Zackular, J. P., Rogers, M. A. M., Ruffin, M. T. & Schloss, P. D. The Human Gut Microbiome as a Screening Tool for Colorectal Cancer. *Cancer Prevention Research* **7**, 1112-1121, doi:10.1158/1940-6207.Capr-14-0129 (2014).
173 Subramanian, S. *et al.* Persistent gut microbiota immaturity in malnourished Bangladeshi children. *Nature* **510**, 417-+, doi:10.1038/nature13421 (2014).

174 Burdet, C. *et al.* Impact of Antibiotic Gut Exposure on the Temporal Changes in Microbiome Diversity. *Antimicrobial Agents and Chemotherapy* **63**, doi:ARTN e00820-19 10.1128/AAC.00820-19 (2019).
175 Isaksson, A., Wallman, M., Goransson, H. & Gustafsson, M. G. Cross-validation and bootstrapping are unreliable in small sample classification. *Pattern Recognition Letters* **29**, 1960-1965, doi:10.1016/j.patrec.2008.06.018 (2008).
176 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* **57**, 213-232, doi:10.1016/j.tifs.2016.08.011 (2016).
177 Ramadan, A. & Paczesny, S. Various forms of tissue damage and danger signals following hematopoietic stem-cell transplantation. *Frontiers in Immunology* **6**, doi:10.3389/fimmu.2015.00014 (2015).
178 Ghimire, S. *et al.* Pathophysiology of GvHD and Other HSCT-Related Major Complications. *Front Immunol* **8**, 79, doi:10.3389/fimmu.2017.00079 (2017).
179 Galloway-Pena, J. R. *et al.* Fecal Microbiome, Metabolites, and Stem Cell Transplant Outcomes: A Single-Center Pilot Study. *Open Forum Infect Dis* **6**, ofz173, doi:10.1093/ofid/ofz173 (2019).
180 Michonneau D *et al.* Metabolomics Profiling after Allogeneic Hematopoietic Stem Cell Transplantation Unravels a Specific Signature in Human Acute GVHD. *Blood* **132** (2018).
181 Romick-Rosendale, L. E. *et al.* Antibiotic Exposure and Reduced Short Chain Fatty Acid Production after Hematopoietic Stem Cell Transplant. *Biology of Blood and Marrow Transplantation* **24**, 2418-2424, doi:10.1016/j.bbmt.2018.07.030 (2018).
182 Furusawa, Y. *et al.* Commensal microbe-derived butyrate induces the differentiation of colonic regulatory T cells. *Nature* **504**, 446-+, doi:10.1038/nature12721 (2013).
183 Maslowski, K. M. Metabolism at the centre of the host-microbe relationship. *Clinical and Experimental Immunology* **197**, 193-204, doi:10.1111/cei.13329 (2019).
184 Litvak, Y., Byndloss, M. X. & Baumler, A. J. Colonocyte metabolism shapes the gut microbiota. *Science* **362**, 1017, doi:ARTN eaat9076 10.1126/science.aat9076 (2018).
185 Yoon, M. Y. & Yoon, S. S. Disruption of the Gut Ecosystem by Antibiotics. *Yonsei Medical Journal* **59**, 4-12, doi:10.3349/ymj.2018.59.1.4 (2018).
186 Byndloss, M. X. *et al.* Microbiota-activated PPAR-gamma signaling inhibits dysbiotic Enterobacteriaceae expansion. *Science* **357**, 570-+, doi:10.1126/science.aam9949 (2017).
187 Holota, Y. *et al.* The long-term consequences of antibiotic therapy: Role of colonic short-chain fatty acids (SCFA) system and intestinal barrier integrity. *Plos One* **14**, doi:ARTN e0220642 10.1371/journal.pone.0220642 (2019).
188 De Pietri, S. *et al.* Gastrointestinal toxicity during induction treatment for childhood acute lymphoblastic leukemia: The impact of the gut microbiota. *Int J Cancer*, doi:10.1002/ijc.32942 (2020).
189 Marchesi, J. R. *et al.* Rapid and noninvasive metabonomic characterization of inflammatory bowel disease. *Journal of Proteome Research* **6**, 546-551, doi:10.1021/pr060470d (2007).
190 Donohoe, D. R. *et al.* The Microbiome and Butyrate Regulate Energy Metabolism and Autophagy in the Mammalian Colon. *Cell Metabolism* **13**, 517-526, doi:10.1016/j.cmet.2011.02.018 (2011).
191 Alves, A., Bassot, A., Bulteau, A. L., Pirola, L. & Morio, B. Glycine Metabolism and Its Alterations in Obesity and Metabolic Diseases. *Nutrients* **11**, doi:ARTN 1356 10.3390/nu11061356 (2019).
192 Mardinoglu, A. *et al.* The gut microbiota modulates host amino acid and glutathione metabolism in mice. *Molecular Systems Biology* **11**, doi:ARTN 834 DOI 10.15252/msb.20156487 (2015).
193 Zierer, J. *et al.* The fecal metabolome as a functional readout of the gut microbiome. *Nature Genetics* **50**, 790-+, doi:10.1038/s41588-018-0135-7 (2018).
194 Noor, F., Kaysen, A., Wilmes, P. & Schneider, J. G. The Gut Microbiota and Hematopoietic Stem Cell Transplantation: Challenges and Potentials. *J Innate Immun*, 1-11, doi:10.1159/000492943 (2018).
195 Taur, Y. *et al.* Reconstitution of the gut microbiota of antibiotic-treated patients by autologous fecal microbiota transplant. *Sci Transl Med* **10**, doi:10.1126/scitranslmed.aap9489 (2018).
196 Gorshein, E. *et al.* Lactobacillus rhamnosus GG probiotic enteric regimen does not appreciably alter the gut microbiome or provide protection against GVHD after allogeneic hematopoietic stem cell transplantation. *ClinTransplant* **31**, doi:ARTN e12947 10.1111/ctr.12947 (2017).
197 Zmora, N. *et al.* Personalized Gut Mucosal Colonization Resistance to Empiric Probiotics Is Associated with Unique Host and Microbiome Features. *Cell* **174**, 1388-+, doi:10.1016/j.cell.2018.08.041 (2018).
198 Suez, J. *et al.* Post-Antibiotic Gut Mucosal Microbiome Reconstitution Is Impaired by Probiotics and Improved by Autologous FMT. *Cell* **174**, 1406-+, doi:10.1016/j.cell.2018.08.047 (2018).

199 Rousseaux, C. *et al.* Intestinal antiinflammatory effect of 5-aminosalicylic acid is dependent on peroxisome proliferator-activated receptor-gamma. *Journal of Experimental Medicine* **201**, 1205-1215, doi:10.1084/jem.20041948 (2005).

200 Rousseaux, C. *et al.* The 5-aminosalicylic acid antineoplastic effect in the intestine is mediated by PPAR gamma. *Carcinogenesis* **34**, 2580-2586, doi:10.1093/carcin/bgt245 (2013).

201 Jirsova, Z. *et al.* The Effect of Butyrate-Supplemented Parenteral Nutrition on Intestinal Defence Mechanisms and the Parenteral Nutrition-Induced Shift in the Gut Microbiota in the Rat Model. *Biomed Research International*, doi:Artn 7084734 10.1155/2019/7084734 (2019).

202 Jiminez, J. A., Uwiera, T. C., Abbott, D. W., Uwiera, R. R. E. & Inglis, G. D. Butyrate Supplementation at High Concentrations Alters Enteric Bacterial Communities and Reduces Intestinal Inflammation in Mice Infected with Citrobacter rodentium. *Msphere* **2**, doi:ARTN e00243-17 10.1128/mSphere.00243-17 (2017).

203 Hamera, H. *et al.* Butyrate enemas do not affect human colonic MUC2 and TFF3 expression. *EUR J GASTROEN HEPAT* **22**, 1134-1140 (2010).

204 Golob, J. L. *et al.* Butyrogenic bacteria after acute graft-versus-host disease (GVHD) are associated with the development of steroid-refractory GVHD. *Blood Advances* **3**, 2866-2869, doi:10.1182/bloodadvances.2019000362 (2019).

205 Kaiko, G. E. *et al.* The Colonic Crypt Protects Stem Cells from Microbiota-Derived Metabolites. *Cell* **165**, 1708-1720, doi:10.1016/j.cell.2016.05.018 (2016).

206 Ze, X. L., Duncan, S. H., Louis, P. & Flint, H. J. Ruminococcus bromii is a keystone species for the degradation of resistant starch in the human colon. *Isme Journal* **6**, 1535-1543, doi:10.1038/ismej.2012.4 (2012).

Appendix

Table A1 Primers used in this study

341F (V3 F)	CCTACGGGNGGCWGCAG
785R (V4 R)	GGACTACHVGGGTWTCTAAT
V3V4 index	ATT AGA WAC CCB DGT AGT CCG GCT GAC TGA CT
V3V4 read 1	TAT GGT AAT TGG CCT ACG GGN GGC WGC AG
V3V4 read 2	AGT CAG TCA GCC GGA CTA CHV GGG TWT CTA AT
SA501	AATGATACGGCGACCACCGAGATCTACACATCGTACGTATGGTAATTGGCCTACGGGNGGCWGCAG
SA502	AATGATACGGCGACCACCGAGATCTACACACTATCTGTATGGTAATTGGCCTACGGGNGGCWGCAG
SA503	AATGATACGGCGACCACCGAGATCTACACTAGCGAGTTATGGTAATTGGCCTACGGGNGGCWGCAG
SA504	AATGATACGGCGACCACCGAGATCTACACCTGCGTGTTATGGTAATTGGCCTACGGGNGGCWGCAG
SA505	AATGATACGGCGACCACCGAGATCTACACTCATCGAGTATGGTAATTGGCCTACGGGNGGCWGCAG
SA506	AATGATACGGCGACCACCGAGATCTACACCGTGAGTGTATGGTAATTGGCCTACGGGNGGCWGCAG
SA507	AATGATACGGCGACCACCGAGATCTACACGGATATCTTATGGTAATTGGCCTACGGGNGGCWGCAG
SA508	AATGATACGGCGACCACCGAGATCTACACGACACCGTTATGGTAATTGGCCTACGGGNGGCWGCAG
SA701	CAAGCAGAAGACGGCATACGAGATAACTCTCGAGTCAGTCAGCCGGACTACHVGGGTWTCTAAT
SA702	CAAGCAGAAGACGGCATACGAGATACTATGTCAGTCAGTCAGCCGGACTACHVGGGTWTCTAAT
SA703	CAAGCAGAAGACGGCATACGAGATAGTAGCGTAGTCAGTCAGCCGGACTACHVGGGTWTCTAAT
SA704	CAAGCAGAAGACGGCATACGAGATCAGTGAGTAGTCAGTCAGCCGGACTACHVGGGTWTCTAAT

SA705	CAAGCAGAAGACGGCATACGAGATCGTACTCAAGTCAGTCAGCCGGACTACHVGGGTWTCTAAT
SA706	CAAGCAGAAGACGGCATACGAGATCTACGCAGAGTCAGTCAGCCGGACTACHVGGGTWTCTAAT
SA707	CAAGCAGAAGACGGCATACGAGATGGAGACTAAGTCAGTCAGCCGGACTACHVGGGTWTCTAAT
SA708	CAAGCAGAAGACGGCATACGAGATGTCGCTCGAGTCAGTCAGCCGGACTACHVGGGTWTCTAAT
SA709	CAAGCAGAAGACGGCATACGAGATGTCGTAGTAGTCAGTCAGCCGGACTACHVGGGTWTCTAAT
SA710	CAAGCAGAAGACGGCATACGAGATTAGCAGACAGTCAGTCAGCCGGACTACHVGGGTWTCTAAT
SA711	CAAGCAGAAGACGGCATACGAGATTCATAGACAGTCAGTCAGCCGGACTACHVGGGTWTCTAAT
SA712	CAAGCAGAAGACGGCATACGAGATTCGCTATAAGTCAGTCAGCCGGACTACHVGGGTWTCTAAT

Table A2 Relative abundance (16S rRNA copy number adjusted) of the 8-strain bacterial mock community used in this chapter (D6305, Zymo, USA)

	Relative abundance
Pseudomonas aeruginosa	4.6
Escherichia coli	10
Salmonella enterica	11.3
Lactobacillus fermentum	18.8
Staphylococcus aureus	13.3
Enterococcus faecalis	10.4
Listeria monocytogenes	15.9
Bacillus subtilis	15.7

Table A3 Primers for the V3-4 and the V5-7 16S rRNA region used in this chapter

V3-4 Region primers	
341F (V3 F)	CCTACGGGNGGCWGCAG
785R (V4 R)	GGACTACHVGGGTWTCTAAT
Index_V34_read (Index read)	ATTAGAWACCCBDGTAGTCCGGCTGACTGACT
Read_V34_1 (Read 1)	TATGGTAATTGGCCTACGGGNGGCWGCAG
Read_V34-_2 (Read 2)	AGTCAGTCAGCCGGACTACHVGGGTWTCTAAT
16S_V3F_2	AATGATACGGCGACCACCGAGATCTACACAACGCTAATATGGTAATTGGCCTACGGGNGGCWGCAG
16S_V3F_3	AATGATACGGCGACCACCGAGATCTACACAAGACTACTATGGTAATTGGCCTACGGGNGGCWGCAG
16S_V3F_4	AATGATACGGCGACCACCGAGATCTACACAATCGATATATGGTAATTGGCCTACGGGNGGCWGCAG
16S_V3F_5	AATGATACGGCGACCACCGAGATCTACACACCAATTGTATGGTAATTGGCCTACGGGNGGCWGCAG
16S_V3F_6	AATGATACGGCGACCACCGAGATCTACACACTGAAGTTATGGTAATTGGCCTACGGGNGGCWGCAG
16S_V3F_8	AATGATACGGCGACCACCGAGATCTACACCAACCTTATATGGTAATTGGCCTACGGGNGGCWGCAG
16S_V3F_10	AATGATACGGCGACCACCGAGATCTACACCCTCTGATTATGGTAATTGGCCTACGGGNGGCWGCAG
16S_V3F_14	AATGATACGGCGACCACCGAGATCTACACGAACGGAGTATGGTAATTGGCCTACGGGNGGCWGCAG
16S_V3F_18	AATGATACGGCGACCACCGAGATCTACACGGATGCCATATGGTAATTGGCCTACGGGNGGCWGCAG
16S_V3F_20	AATGATACGGCGACCACCGAGATCTACACGTTGGCCGTATGGTAATTGGCCTACGGGNGGCWGCAG
16S_V3F_22	AATGATACGGCGACCACCGAGATCTACACTGACTGCTTATGGTAATTGGCCTACGGGNGGCWGCAG
16S_V4R_1	CAAGCAGAAGACGGCATACGAGATAACCAGTCAGTCAGTCAGCCGGACTACHVGGGTWTCTAAT
16S_V4R_7	CAAGCAGAAGACGGCATACGAGATATTGCCGCAGTCAGTCAGCCGGACTACHVGGGTWTCTAAT
16S_V4R_9	CAAGCAGAAGACGGCATACGAGATCCTAATAAAGTCAGTCAGCCGGACTACHVGGGTWTCTAAT

16S_V4R_15	CAAGCAGAAGACGGCATACGAGATGCCTACGCAGTCAGTCAGCCGGACTACHVGGGTWTCTAAT
16S_V4R_16	CAAGCAGAAGACGGCATACGAGATGCGTTACCAGTCAGTCAGCCGGACTACHVGGGTWTCTAAT
16S_V4R_17	CAAGCAGAAGACGGCATACGAGATGGAGGCTGAGTCAGTCAGCCGGACTACHVGGGTWTCTAAT
16S_V4R_23	CAAGCAGAAGACGGCATACGAGATTGGCGATTAGTCAGTCAGCCGGACTACHVGGGTWTCTAAT
16S_V4R_25	CAAGCAGAAGACGGCATACGAGATTTGGCTATAGTCAGTCAGCCGGACTACHVGGGTWTCTAAT
V5-7 Region primers	
785F (V5F)	GGATTAGATACCCBRGTAGTC
1175R (V5R)	ACGTCRTCCCCDCCTTCCTC
Index_V57_read (Index read)	ACGTACGTACGTGGATTAGATACCCBRGTAGTC
Read_V57_1 (Read 1)	AGTCAGTCAGCCACGTCRTCCCCDCCTTCCTC
Read_V57_2 (Read 2)	GAGGAAGGHGGGGAYGACGTGGCTGACTGACT
FWD01	AATGATACGGCGACCACCGAGATCTACACTAGATCGACGTACGTACGTGGATTAGATACCCBRGTAGTC
FWD02	AATGATACGGCGACCACCGAGATCTACACCTCTCTATACGTACGTACGTGGATTAGATACCCBRGTAGTC
FWD03	AATGATACGGCGACCACCGAGATCTACACTATCCTCTACGTACGTACGTGGATTAGATACCCBRGTAGTC
FWD04	AATGATACGGCGACCACCGAGATCTACACAGAGTAGAACGTACGTACGTGGATTAGATACCCBRGTAGTC
FWD05	AATGATACGGCGACCACCGAGATCTACACGTAAGGAGACGTACGTACGTGGATTAGATACCCBRGTAGTC
FWD06	AATGATACGGCGACCACCGAGATCTACACACTGCATAACGTACGTACGTGGATTAGATACCCBRGTAGTC
FWD07	AATGATACGGCGACCACCGAGATCTACACAAGGAGTAACGTACGTACGTGGATTAGATACCCBRGTAGTC
FWD08	AATGATACGGCGACCACCGAGATCTACACCTAAGCCTACGTACGTACGTGGATTAGATACCCBRGTAGTC
REV01	CAAGCAGAAGACGGCATACGAGATTCGCCTTAAGTCAGTCAGCCACGTCRTCCCCDCCTTCCTC
REV02	CAAGCAGAAGACGGCATACGAGATCTAGTACGAGTCAGTCAGCCACGTCRTCCCCDCCTTCCTC
REV03	CAAGCAGAAGACGGCATACGAGATTTCTGCCTAGTCAGTCAGCCACGTCRTCCCCDCCTTCCTC

REV04	CAAGCAGAAGACGGCATACGAGATGCTCAGGAAGTCAGTCAGCCACGTCRTCCCCDCCTTCCTC
REV05	CAAGCAGAAGACGGCATACGAGATAGGAGTCCAGTCAGTCAGCCACGTCRTCCCCDCCTTCCTC
REV06	CAAGCAGAAGACGGCATACGAGATCATGCCTAAGTCAGTCAGCCACGTCRTCCCCDCCTTCCTC
REV07	CAAGCAGAAGACGGCATACGAGATGTAGAGAGAGTCAGTCAGCCACGTCRTCCCCDCCTTCCTC
REV08	CAAGCAGAAGACGGCATACGAGATCCTCTCTGAGTCAGTCAGCCACGTCRTCCCCDCCTTCCTC
REV09	CAAGCAGAAGACGGCATACGAGATAGCGTAGCAGTCAGTCAGCCACGTCRTCCCCDCCTTCCTC
REV10	CAAGCAGAAGACGGCATACGAGATCAGCCTCGAGTCAGTCAGCCACGTCRTCCCCDCCTTCCTC
REV11	CAAGCAGAAGACGGCATACGAGATTGCCTCTTAGTCAGTCAGCCACGTCRTCCCCDCCTTCCTC

Table A4 Relative abundance (16S rRNA copy number adjusted) of the 20-strain bacterial mock community used in this chapter (HM-783D, low concentration, staggered, BEI Resources, USA)

	Relative abundance
Acinetobacter baumannii	0.24
Actinomyces odontolyticus	0.02
Bacillus cereus	1.79
Bacteroides vulgatus	0.03
Clostridium beijerinckii	1.94
Deinococcus radiodurans	0.02
Enterococcus faecalis	0.01
Escherichia coli	23.16
Helicobacter pylori	0.11
Lactobacillus gasseri	0.04
Listeria monocytogenes	0.11
Neisseria meningitidis	0.10
Propionibacterium acnes	0.17
Pseudomonas aeruginosa	7.35
Rhodobacter sphaeroides	47.28
Staphylococcus aureus	1.26
Staphylococcus epidermidis	9.60
Streptococcus agalactiae	0.51
Streptococcus mutans	6.26
Streptococcus pneumoniae	0.01

Table A5 Indicator species analysis comparing the DNA extraction methods within the V3-4 16S region. Indicator values were calculated using the abundance distribution of OTUs at the 97% identity cut-off. P values are Holm-adjusted.

	Indicator of	Adjusted p value
Enteric_Bacteria_cluster	MPBio	0.04
Incertae_Sedis	MPBio	0.04
Enterococcus	MPBio	0.04
Incertae_Sedis.3	MPBio	0.04
Peptoniphilus	MPBio	0.04
Anaerococcus	MPBio	0.04
Gemella	MoBio	0.04
Ralstonia	MoBio	0.04
Bacteroides	Phenol	0.04
Faecalibacterium	Phenol	0.04
Blautia	Phenol	0.04
Subdoligranulum	Phenol	0.04
Phascolarctobacterium	Phenol	0.04
Incertae_Sedis.2	Phenol	0.04
Parabacteroides	Phenol	0.04
Alistipes	Phenol	0.04
Clostridium	Phenol	0.04
Parasutterella	Phenol	0.04
Odoribacter	Phenol	0.04
Coriobacterineae	Phenol	0.04
Bifidobacteriaceae	Phenol	0.04
Bilophila	Phenol	0.04
Clostridiaceae_unclassified	Phenol	0.04
Coprococcus	Phenol	0.04
Lachnospira	Phenol	0.04
Oscillibacter	Phenol	0.04
Bacteroidetes_unclassified	Phenol	0.04
Veillonella	Phenol	0.04
Haemophilus	Phenol	0.04
Bacteria_unclassified	Phenol	0.04
Streptococcus	Phenol	0.04
Prevotella	Phenol	0.04
Porphyromonadaceae_unclassified	Phenol	0.04
Bacteroidales_unclassified	Phenol	0.04
Micrococcineae	Phenol	0.04
Peptostreptococcaceae_unclassified	Phenol	0.04
Eubacterium	Qiagen	0.04

Table A6 Indicator species analysis comparing the DNA extraction methods within the V5-7 16S region. Indicator values were calculated using the abundance distribution of OTUs at the 97% identity cut-off. P values are Holm-adjusted.

	Indicator of	adjusted p value
Enteric_Bacteria_cluster	MPBio	0.04
Enterococcus	MPBio	0.04
Incertae_Sedis.4	MPBio	0.04
Bacteroides	MoBio	0.04
Blautia	MoBio	0.04
Parabacteroides	MoBio	0.04
Staphylococcus	MoBio	0.04
Parasutterella	MoBio	0.04
Lachnospira	MoBio	0.04
Collinsella	MoBio	0.04
Gemella	MoBio	0.04
Finegoldia	MoBio	0.04
Incertae_Sedis.2	Phenol	0.04
Subdoligranulum	Phenol	0.04
Incertae_Sedis.3	Phenol	0.04
Clostridium	Phenol	0.04
Oscillibacter	Phenol	0.04
Bacteroidetes_unclassified	Phenol	0.04
Roseburia	Phenol	0.04
Clostridiales_unclassified	Phenol	0.04
Turicibacter	Phenol	0.04
Bifidobacteriaceae	Phenol	0.04
Family_XIII_Incertae_Sedis_unclassified	Phenol	0.04
Eggerthella	Phenol	0.04
Coprococcus	Phenol	0.04
Veillonella	Phenol	0.04
Actinomyces	Phenol	0.04
Coriobacteriaceae_unclassified	Phenol	0.04
Eubacterium	Phenol	0.04
Rothia	Phenol	0.04
Bacteroidales_unclassified	Phenol	0.04
Prevotella	Phenol	0.04
Candidate_division_TM7_unclassified	Phenol	0.04
Clostridia_unclassified	Phenol	0.04
Corynebacterium	Phenol	0.04
Eubacterium.2	Phenol	0.04
Ralstonia	Phenol	0.04
Clostridiaceae_unclassified	Phenol	0.04
Rhodobacter	Qiagen	0.04
Pseudomonas	Qiagen	0.04
Helicobacter	Qiagen	0.04

Table A7 Composition of the 20-species mock community for the V3-4 and the V5-7 16 rRNA regions amplified using a single (1S) or double (2S) PCR steps.

	V34-1S	**V34-2S**	**V57-1S**	**V57-2S**
Overall number of OTUs	22	25	103	53
Acinetobacter baumannii	1	1	1	1
Actinomyces odontolyticus	0	0	1	1
Bacillus cereus	1	1	1	1
Bacteroides vulgatus	1	2	16	7
Clostridium beijerinckii	1	1	1	1
Deinococcus radiodurans	1	1	0	1
Enterococcus faecalis	1	1	0	1
Escherichia coli	2	4	5	3
Helicobacter pylori	1	1	1	0
Lactobacillus gasseri	1	1	1	1
Listeria monocytogenes	1	1	1	2
Neisseria meningitidis	1	1	1	1
Propionibacterium acnes	0	0	1	0
Pseudomonas aeruginosa	1	1	1	1
Rhodobacter sphaeroides	1	1	1	1
Staphylococcus aureus	2	2	4	3
Staphylococcus epidermidis	3	2	3	3
Streptococcus agalactiae	1	1	0	1
Streptococcus mutans	1	1	1	1
Streptococcus pneumoniae	0	1	1	1
Other*	0	0	4	6
Other (not classified enough)	1	1	58	16

*Other category contains bacterial taxa not found in the original Mock Community.

Table A8 Composition of all mock communities sequenced prior to- and as a result of varying levels of filtering.

Filter criteria	Mean abundance	Correct genera	Incorrect genera	Unassigned genera
Mock 1				
No filtering	-	10	69	676
Most abundant erroneous OTU	8.10E-05	9	0	3
2nd most abundant erroneous OTU	1.80E-05	10	0	36
Mock 2				
No filtering	-	10	8	636
Most abundant erroneous OTU	3.68E-05	7	0	17
2nd most abundant erroneous OTU	2.76E-05	8	1	23
3rd most abundant erroneous OTU	9.19E-06	8	2	87
Mock 3				
No filtering	-	10	41	529
Most abundant erroneous OTU	2.49E-04	8	0	0
2nd most abundant erroneous OTU	1.19E-04	8	1	0
3rd most abundant erroneous OTU	7.11E-05	8	2	1
4th most abundant erroneous OTU	5.93E-05	8	3	2
Mock 4				

No filtering	-	10	84	851
Most abundant erroneous OTU	1.13E-04	8	0	0
2nd most abundant erroneous OTU	8.11E-05	8	1	1
3rd most abundant erroneous OTU	6.48E-05	8	2	2
Mock 5				
No filtering	-	10	77	1209
Most abundant erroneous OTU	7.05E-06	8	0	135
Mock 6				
No filtering	-	10	155	2426
Most abundant erroneous OTU	1.30E-04	8	0	3
2nd most abundant erroneous OTU	2.17E-05	8	1	19
3rd most abundant erroneous OTU	8.69E-06	8	2	97
Mock 7				
No filtering	-	10	53	513
Most abundant erroneous OTU	9.50E-04	8	0	0
2nd most abundant erroneous OTU	4.12E-04	8	1	1
3rd most abundant erroneous OTU	1.11E-04	8	2	3

Table A9 Common genera found to be contaminants in Negative extraction controls (NECs)

Present in all NEC	Present in 6/7 NEC	Present in 5/7 NEC	Present in 4/7 NEC
Enterococcus	Porphyromonadaceae_unclassified	Rothia	Prevotella
Bacillus	Actinomyces	Clostridium_XI	Ruminococcaceae_unclassified
Escherichia_Shigella	Bacteroidetes_unclassified	Gammaproteobacteria_unclassified	Pelomonas
Bacteria_unclassified	Lachnospiraceae_unclassified	Bradyrhizobium	Gp4
Bifidobacterium	Veillonella	Clostridiales_unclassified	Bradyrhizobiaceae_unclassified
Staphylococcus	Parabacteroides	Enterococcaceae_unclassified	Methylobacterium
Streptococcus	Lactobacillus		Lactobacillaceae_unclassified
Bacteroides			TM7_genus_incertae_sedis
Clostridium_XlVa			Betaproteobacteria_unclassified
Pseudomonas			Planctomyces
Enterobacteriaceae_unclassified			Roseburia
Lactobacillales_unclassified			

Table A10 Details of contaminants found within each sequencing plate as well as taxa removed from each dataset

Sequencing plate	Number of OTUs removed	Taxa removed entirely
1	136	Bradyrhizobiaceae_unclassified[127,129,130], Rhizobiales_unclassified[127,129,130], Novosphingobium[129,130], Thermicanus[129], Defluviicoccus, Curvibacter[127,130]
2	260	Bradyrhizobium[127,129,130], Methylophilus[130], Limnohabitans[129], Thermicanus, Bosea[130], Rubellimicrobium[127], Halomonas[129]
3	63	-
4	65	Bradyrhizobium
5	166	Thermicanus, Novosphingobium, Comamonas[127,129,130]
6	326	Pelomonas[127,130], Bradyrhizobium, Curvibacter
7	43	-

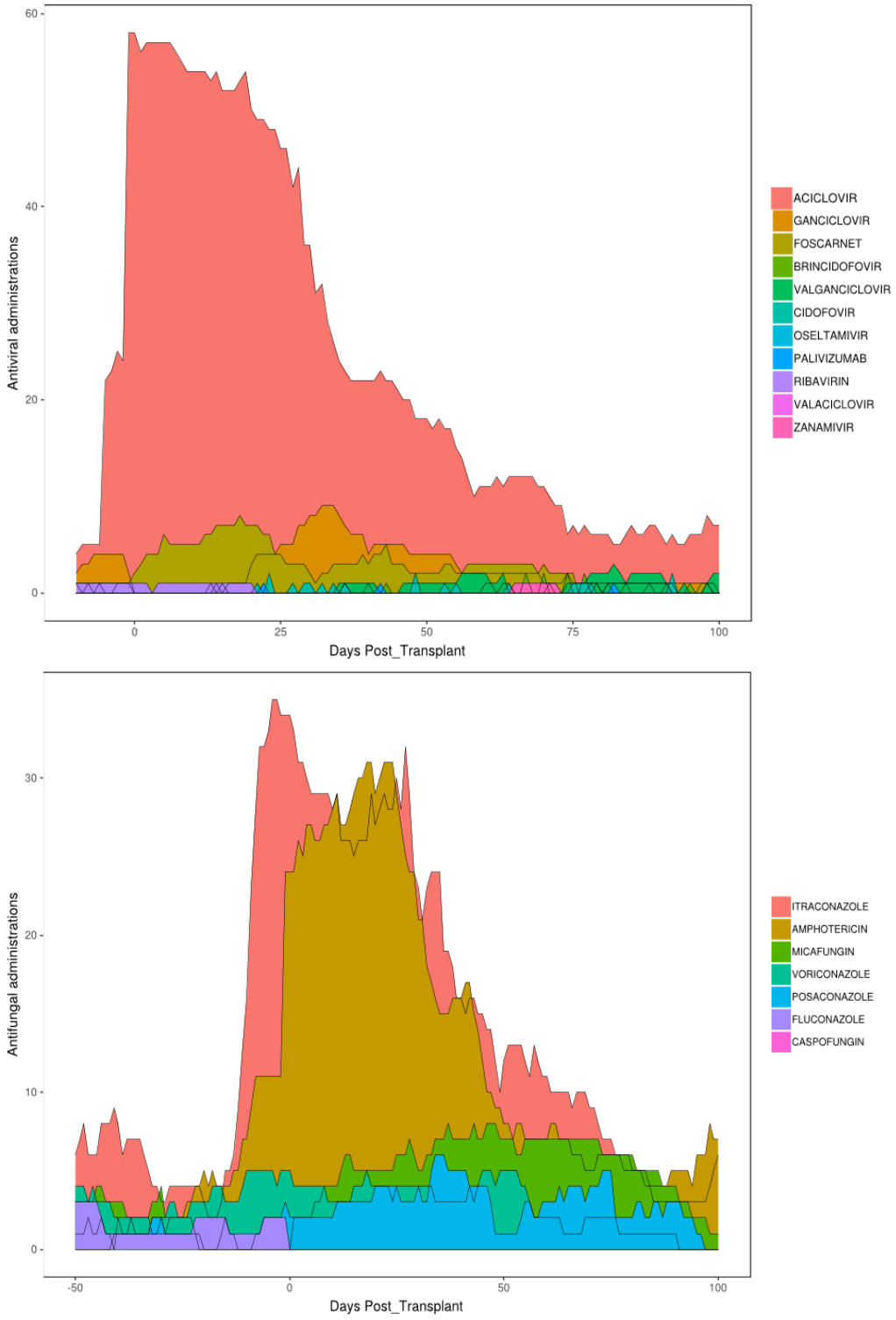
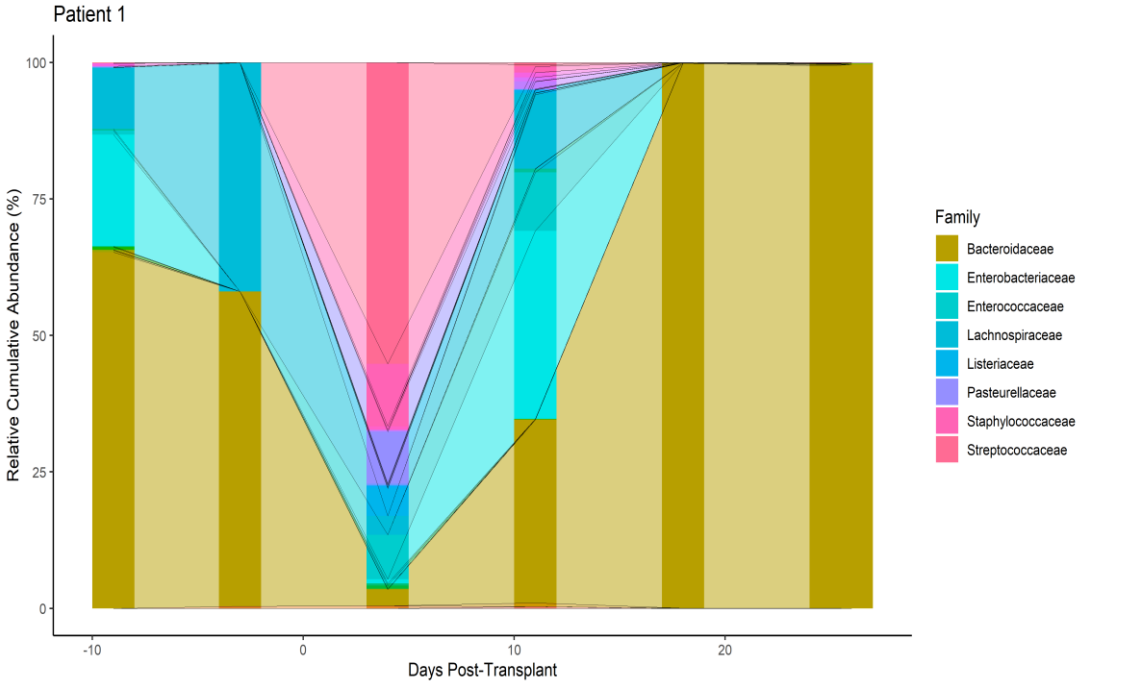
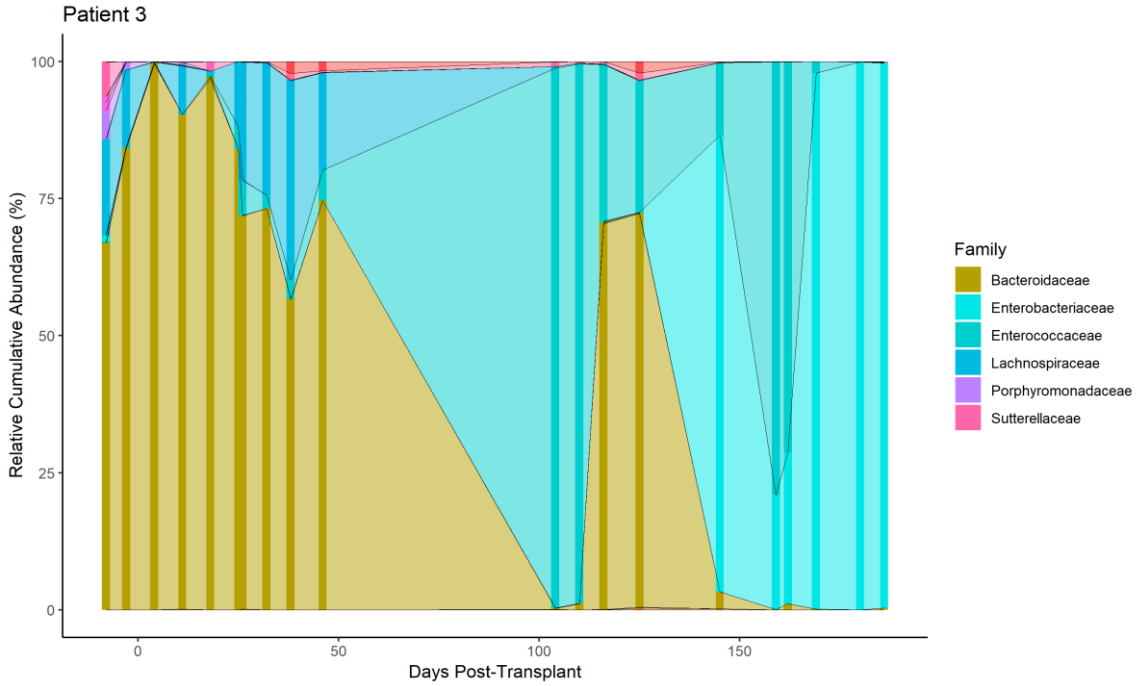
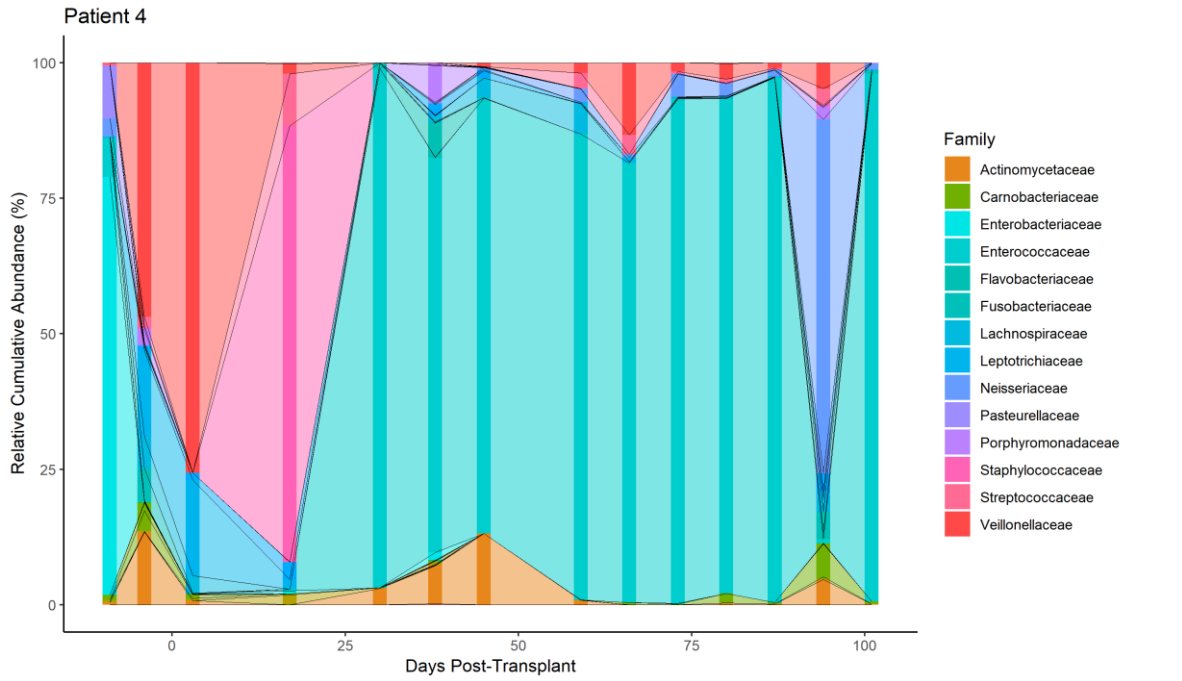
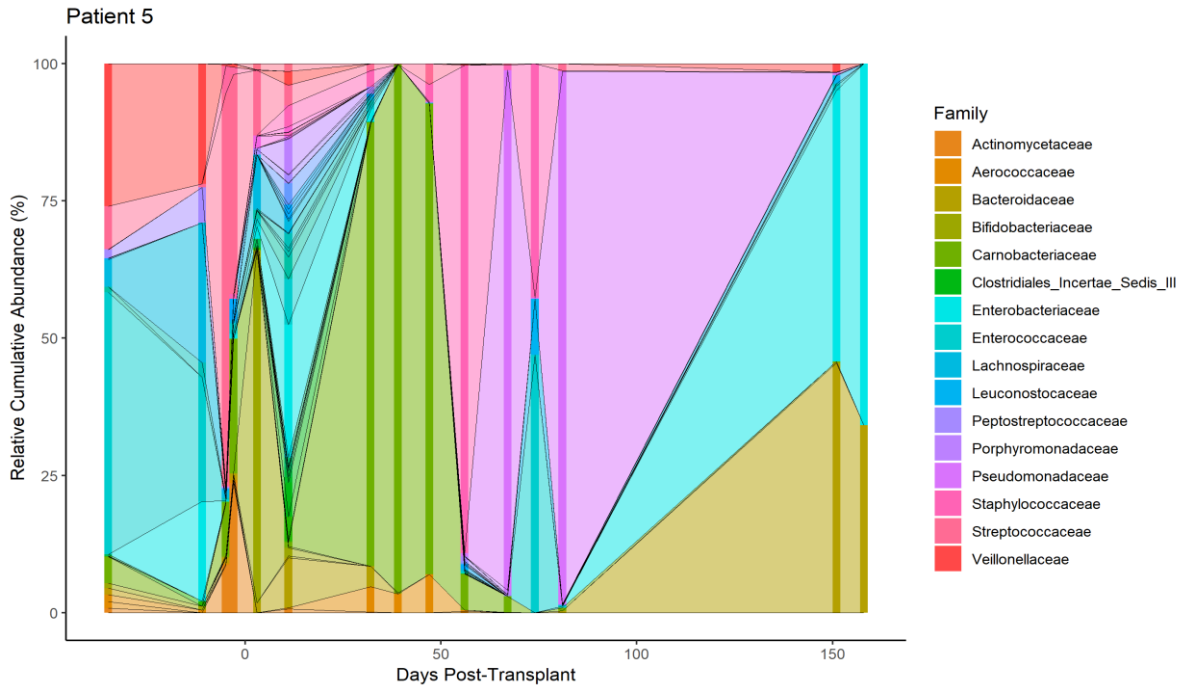


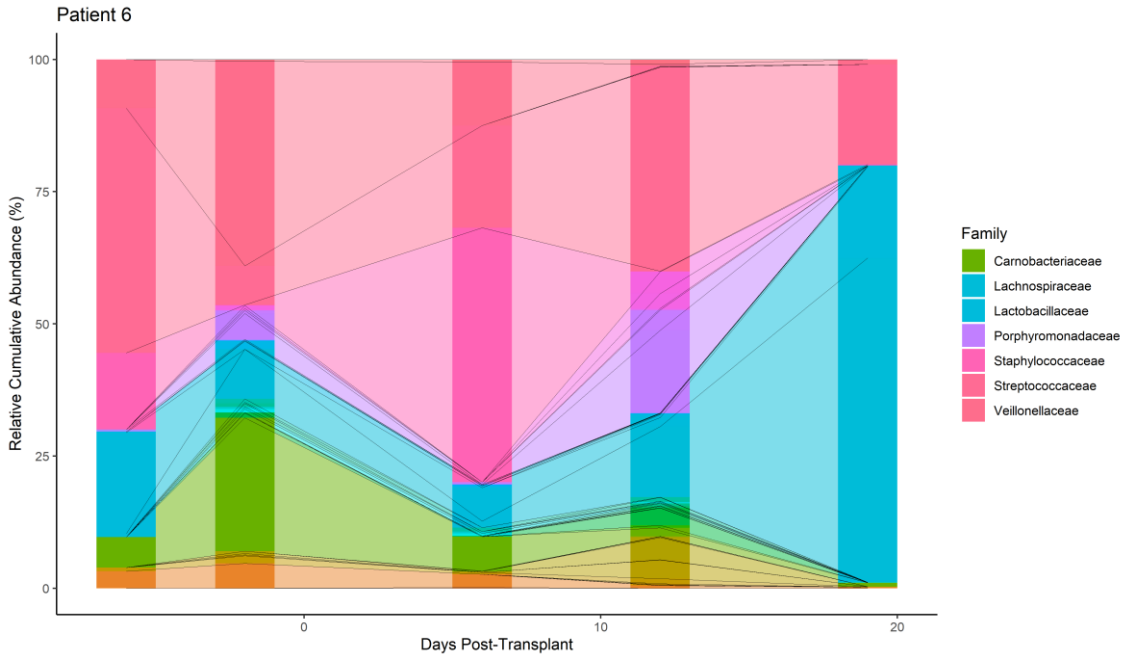
Figure A1 Medication trends throughout HSCT; A) Antiviral and B) Antifungal administrations throughout the first 100 days post-HSCT

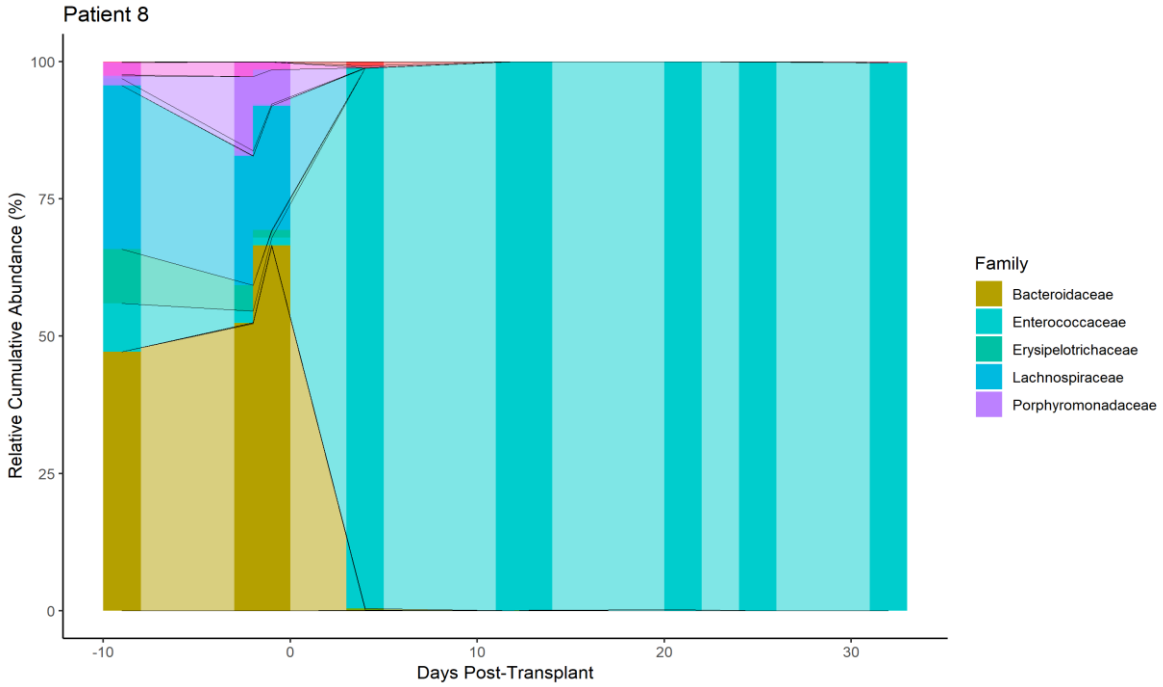

Patient 1
Relative Cumulative Abundance (%)
100
75
50
25
0
-10
0
10
20
Days Post-Transplant
Family
Bacteroidaceae
Enterobacteriaceae
Enterococcaceae
Lachnospiraceae
Listeriaceae
Pasteurellaceae
Staphylococcaceae
Streptococcaceae

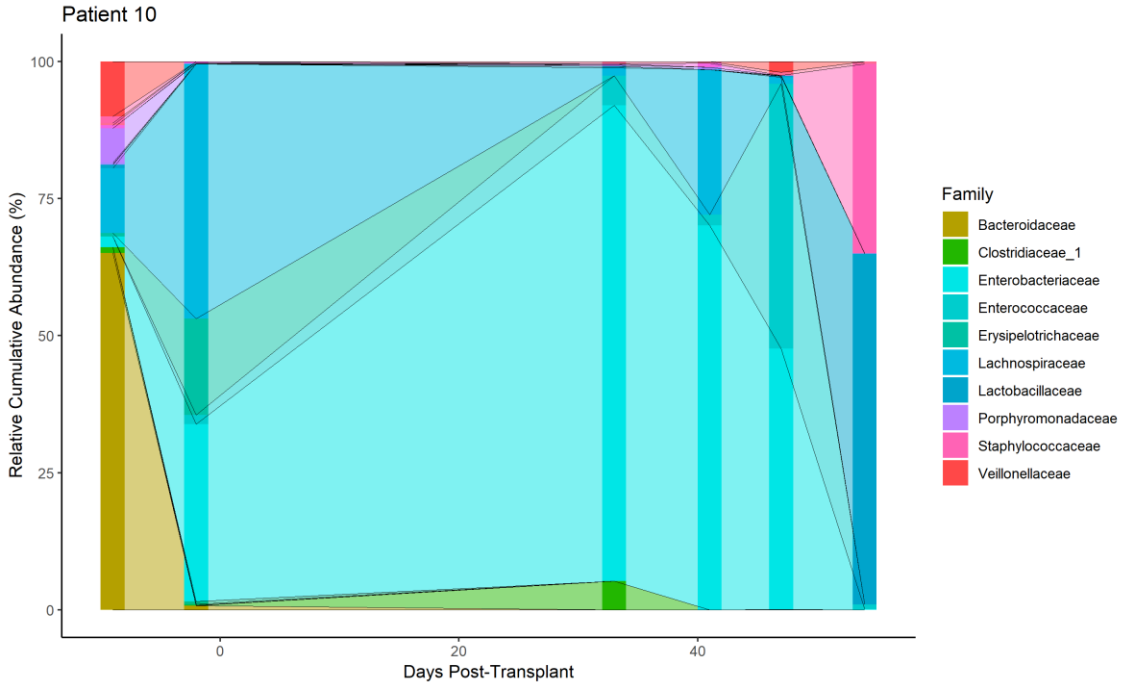

Patient 3
Relative Cumulative Abundance (%)
100
75
50
25
0
0
50
100
150
Days Post-Transplant
Family
Bacteroidaceae
Enterobacteriaceae
Enterococcaceae
Lachnospiraceae
Porphyromonadaceae
Sutterellaceae

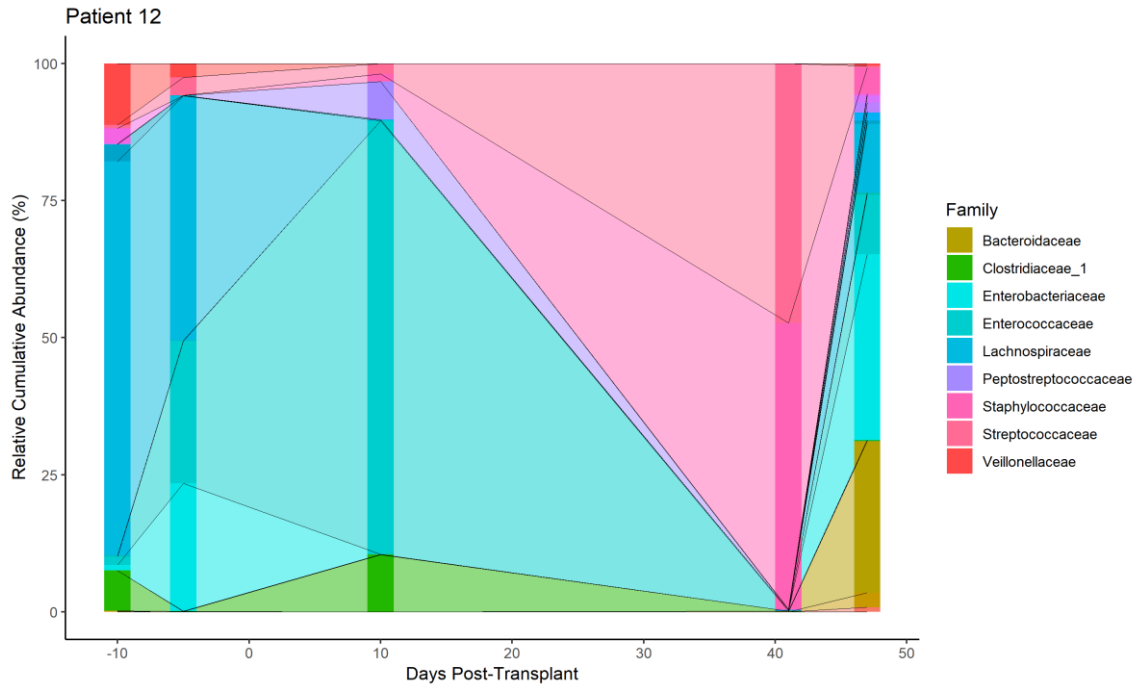

Patient 4
Relative Cumulative Abundance (%)
100
75
50
25
0
0
25
50
75
100
Days Post-Transplant
Family
Actinomycetaceae
Carnobacteriaceae
Enterobacteriaceae
Enterococcaceae
Flavobacteriaceae
Fusobacteriaceae
Lachnospiraceae
Leptotrichiaceae
Neisseriaceae
Pasteurellaceae
Porphyromonadaceae
Staphylococcaceae
Streptococcaceae
Veillonellaceae

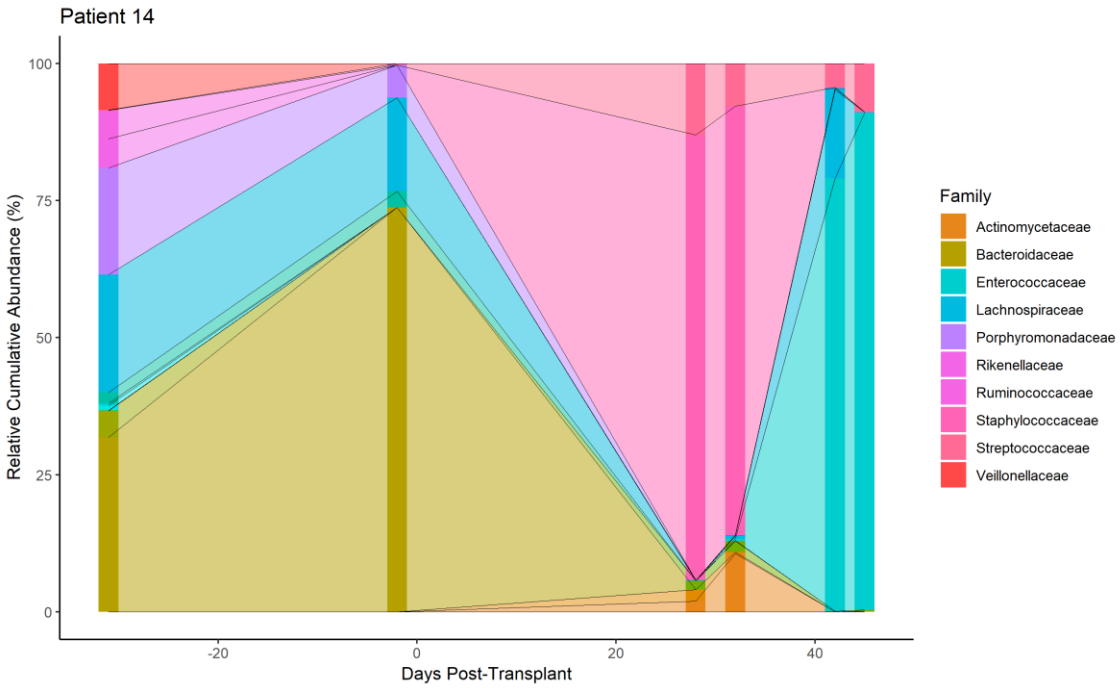

Patient 5
Relative Cumulative Abundance (%)
100
75
50
25
0
0
50
100
150
Days Post-Transplant
Family
Actinomycetaceae
Aerococcaceae
Bacteroidaceae
Bifidobacteriaceae
Carnobacteriaceae
Clostridiales_Incertae_Sedis_III
Enterobacteriaceae
Enterococcaceae
Lachnospiraceae
Leuconostocaceae
Peptostreptococcaceae
Porphyromonadaceae
Pseudomonadaceae
Staphylococcaceae
Streptococcaceae
Veillonellaceae

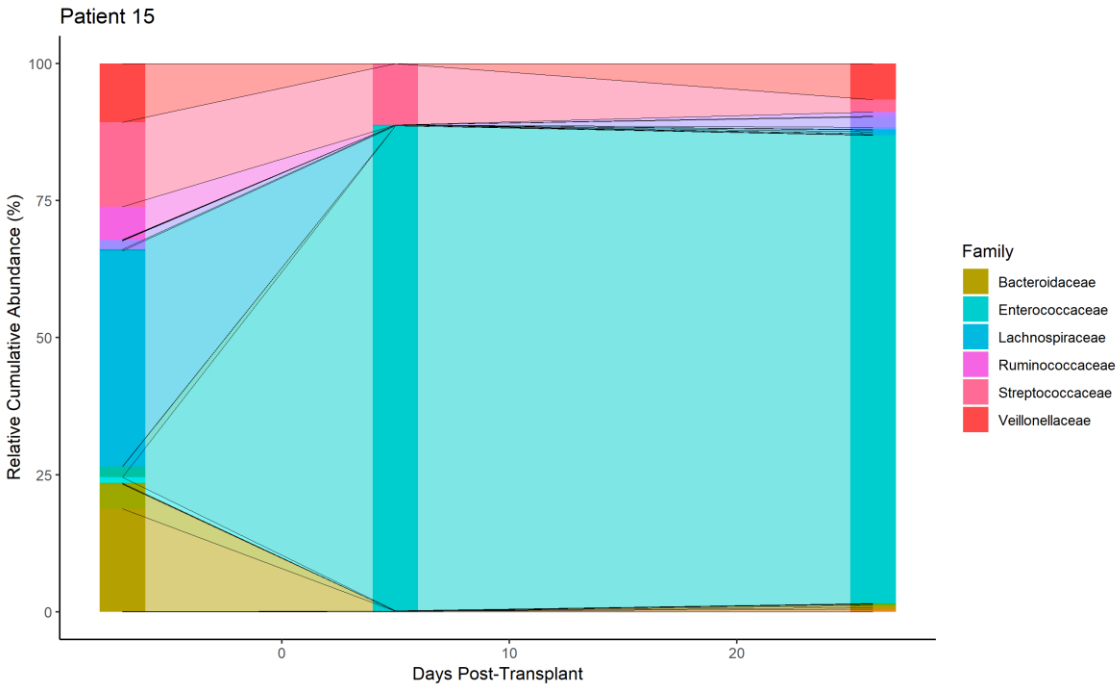

Patient 6
Relative Cumulative Abundance (%)
100
75
50
25
0
0
10
20
Days Post-Transplant
Family
Carnobacteriaceae
Lachnospiraceae
Lactobacillaceae
Porphyromonadaceae
Staphylococcaceae
Streptococcaceae
Veillonellaceae

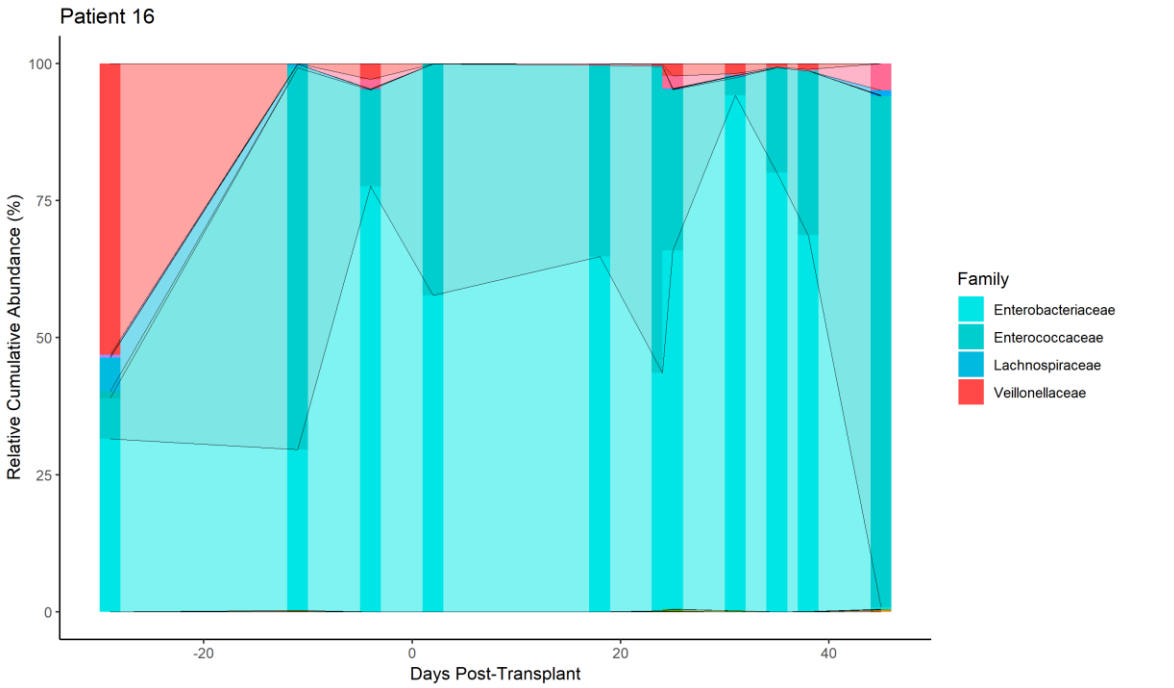

Patient 8
Relative Cumulative Abundance (%)
100
75
50
25
0
-10
0
10
20
30
Days Post-Transplant
Family
Bacteroidaceae
Enterococcaceae
Erysipelotrichaceae
Lachnospiraceae
Porphyromonadaceae

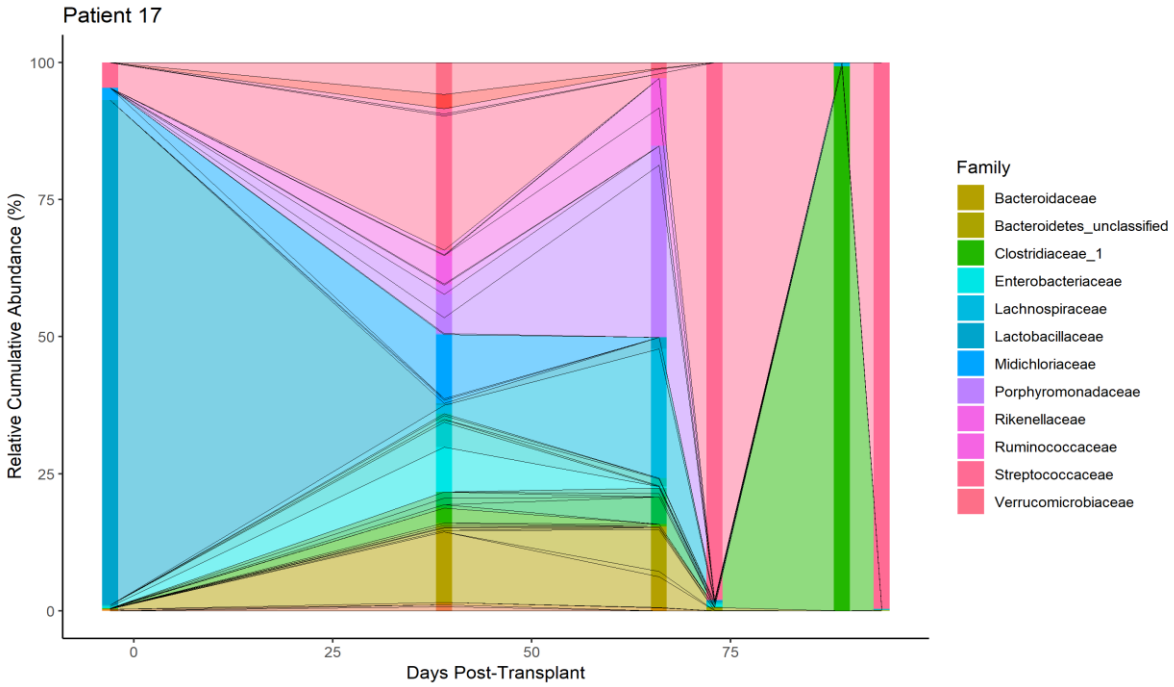

Patient 10
Relative Cumulative Abundance (%)
100
75
50
25
0
0
20
40
Days Post-Transplant
Family
Bacteroidaceae
Clostridiaceae_1
Enterobacteriaceae
Enterococcaceae
Erysipelotrichaceae
Lachnospiraceae
Lactobacillaceae
Porphyromonadaceae
Staphylococcaceae
Veillonellaceae

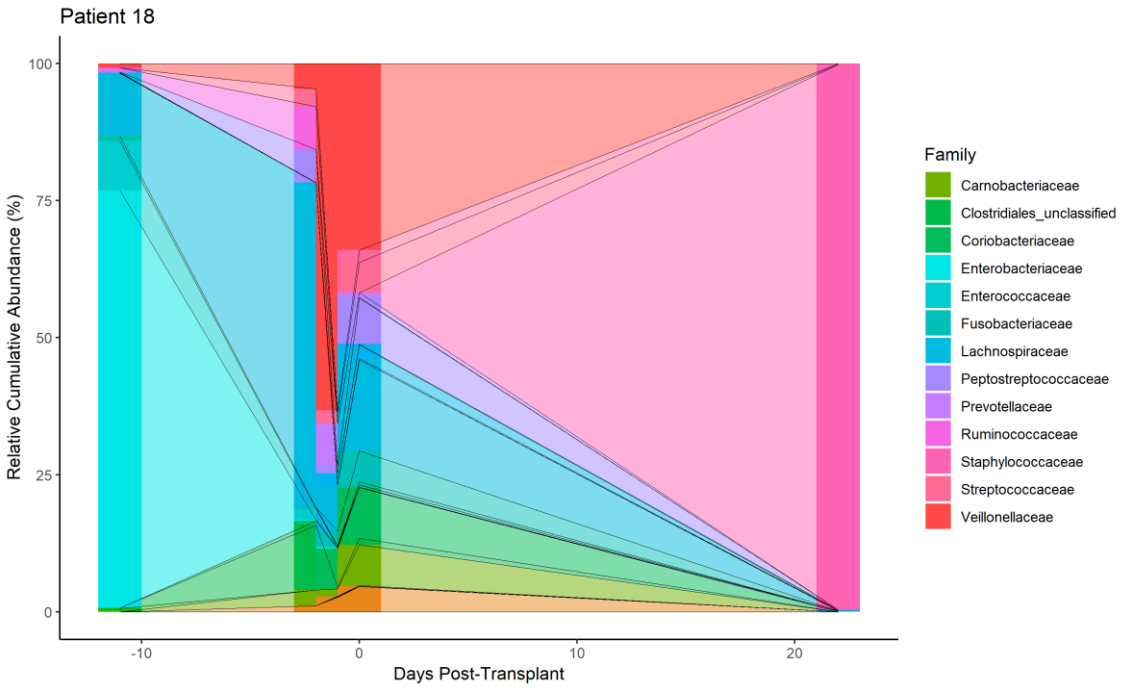

Patient 12
Relative Cumulative Abundance (%)
100
75
50
25
0
-10
0
10
20
30
40
50
Days Post-Transplant
Family
Bacteroidaceae
Clostridiaceae_1
Enterobacteriaceae
Enterococcaceae
Lachnospiraceae
Peptostreptococcaceae
Staphylococcaceae
Streptococcaceae
Veillonellaceae

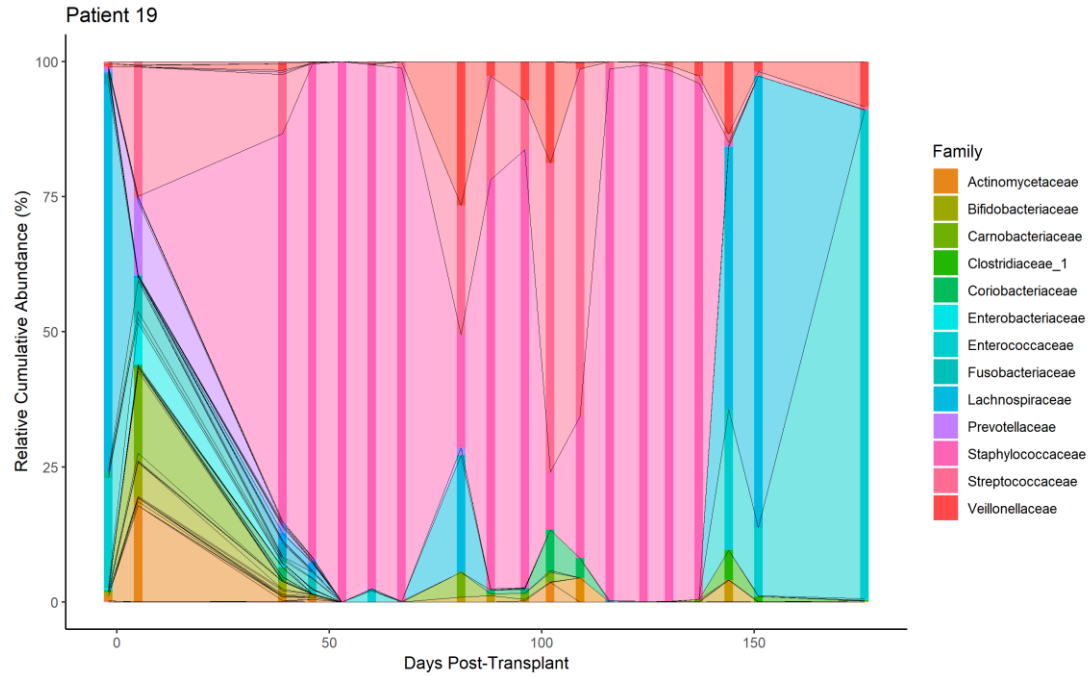

Patient 14
Relative Cumulative Abundance (%)
100
75
50
25
0
-20
0
20
40
Days Post-Transplant
Family
Actinomycetaceae
Bacteroidaceae
Enterococcaceae
Lachnospiraceae
Porphyromonadaceae
Rikenellaceae
Ruminococcaceae
Staphylococcaceae
Streptococcaceae
Veillonellaceae

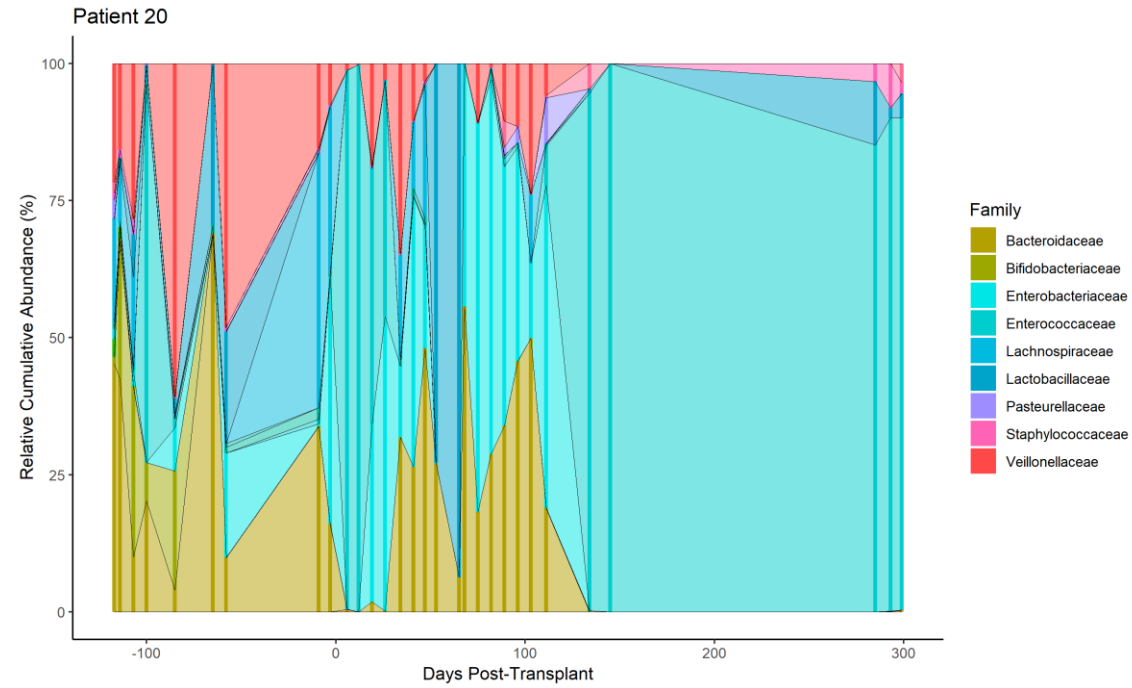

Patient 15
Relative Cumulative Abundance (%)
100
75
50
25
0
0
10
20
Days Post-Transplant
Family
Bacteroidaceae
Enterococcaceae
Lachnospiraceae
Ruminococcaceae
Streptococcaceae
Veillonellaceae

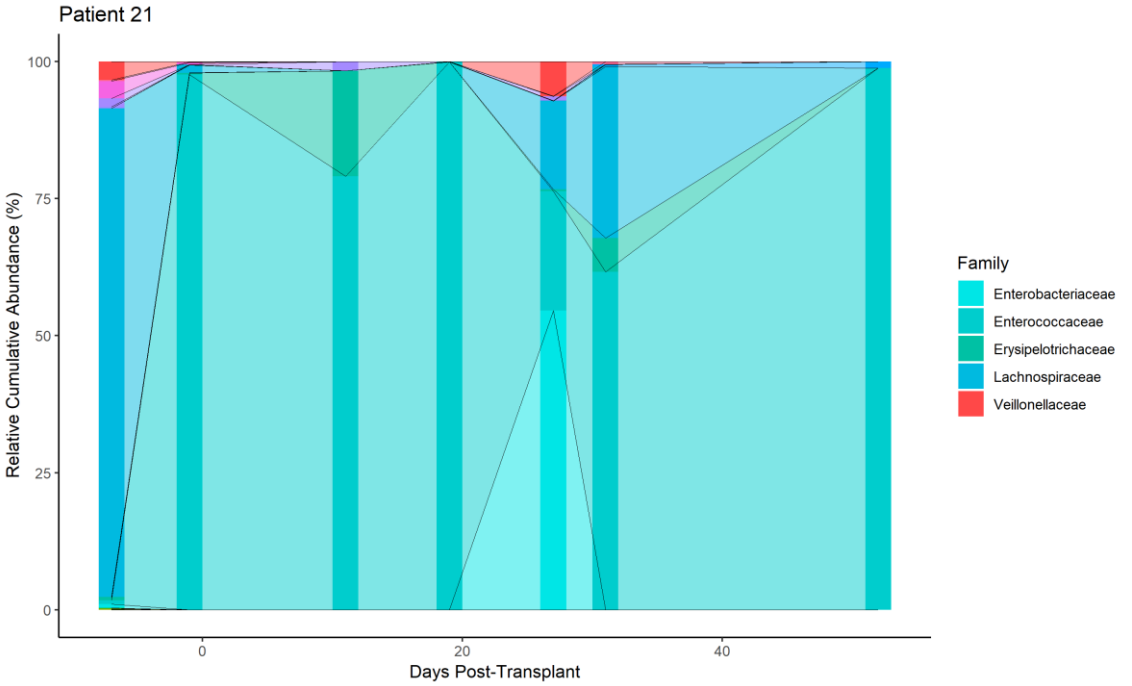

Patient 16
Relative Cumulative Abundance (%)
100
75
50
25
0
-20
0
20
40
Days Post-Transplant
Family
Enterobacteriaceae
Enterococcaceae
Lachnospiraceae
Veillonellaceae

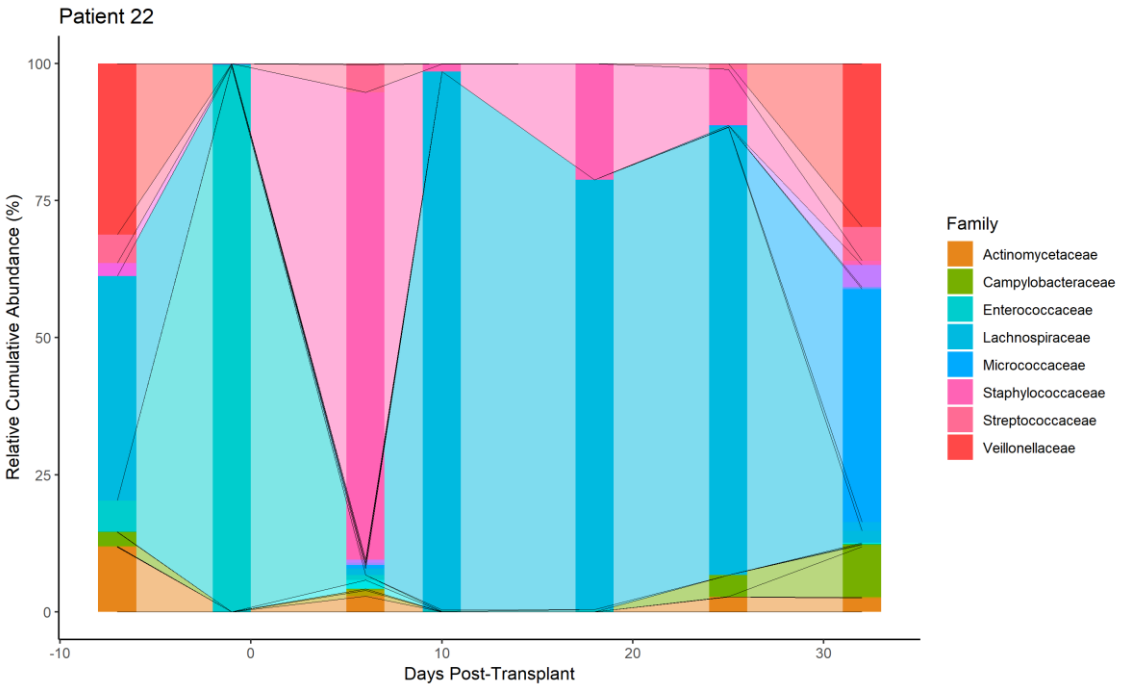

Patient 17
Relative Cumulative Abundance (%)
100
75
50
25
0
0
25
50
75
Days Post-Transplant
Family
Bacteroidaceae
Bacteroidetes_unclassified
Clostridiaceae_1
Enterobacteriaceae
Lachnospiraceae
Lactobacillaceae
Midichloriaceae
Porphyromonadaceae
Rikenellaceae
Ruminococcaceae
Streptococcaceae
Verrucomicrobiaceae

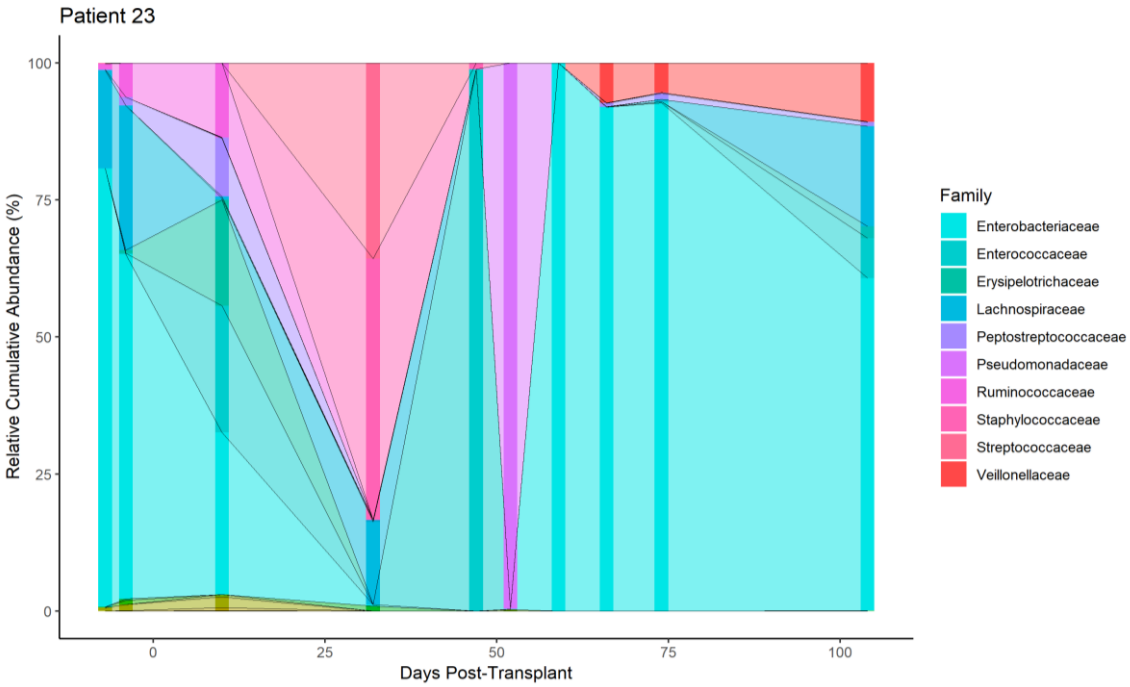

Patient 18
Relative Cumulative Abundance (%)
100
75
50
25
0
-10
0
10
20
Days Post-Transplant
Family
Carnobacteriaceae
Clostridiales_unclassified
Coriobacteriaceae
Enterobacteriaceae
Enterococcaceae
Fusobacteriaceae
Lachnospiraceae
Peptostreptococcaceae
Prevotellaceae
Ruminococcaceae
Staphylococcaceae
Streptococcaceae
Veillonellaceae

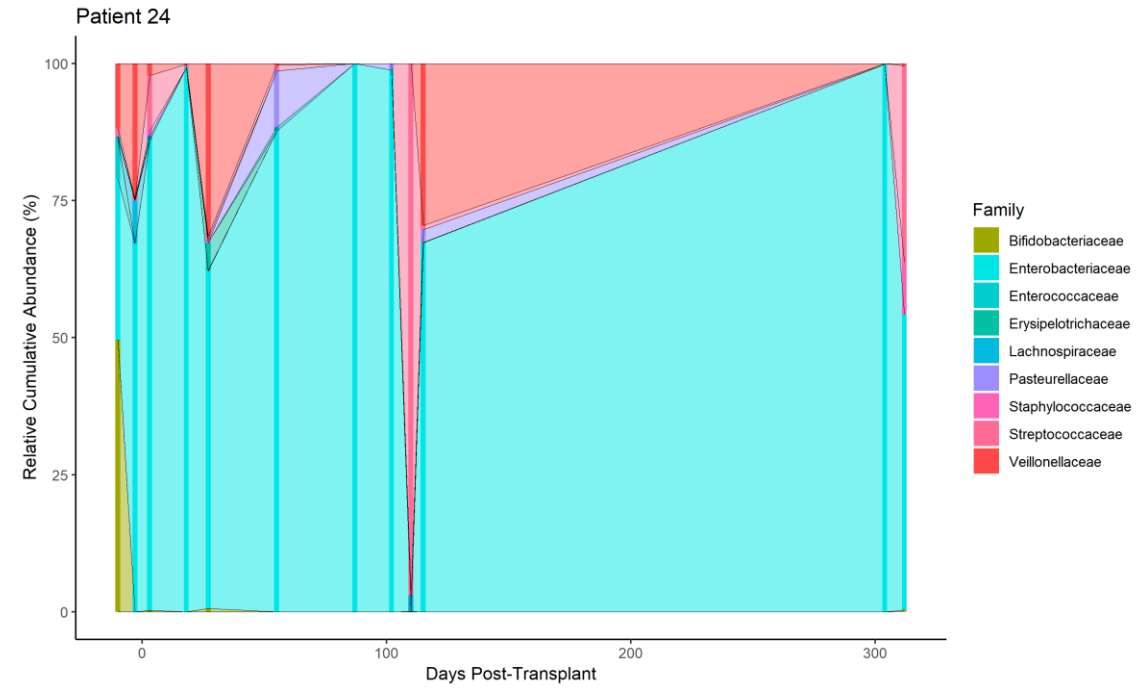

Patient 19
Relative Cumulative Abundance (%)
100
75
50
25
0
0
50
100
150
Days Post-Transplant
Family
Actinomycetaceae
Bifidobacteriaceae
Carnobacteriaceae
Clostridiaceae_1
Coriobacteriaceae
Enterobacteriaceae
Enterococcaceae
Fusobacteriaceae
Lachnospiraceae
Prevotellaceae
Staphylococcaceae
Streptococcaceae
Veillonellaceae

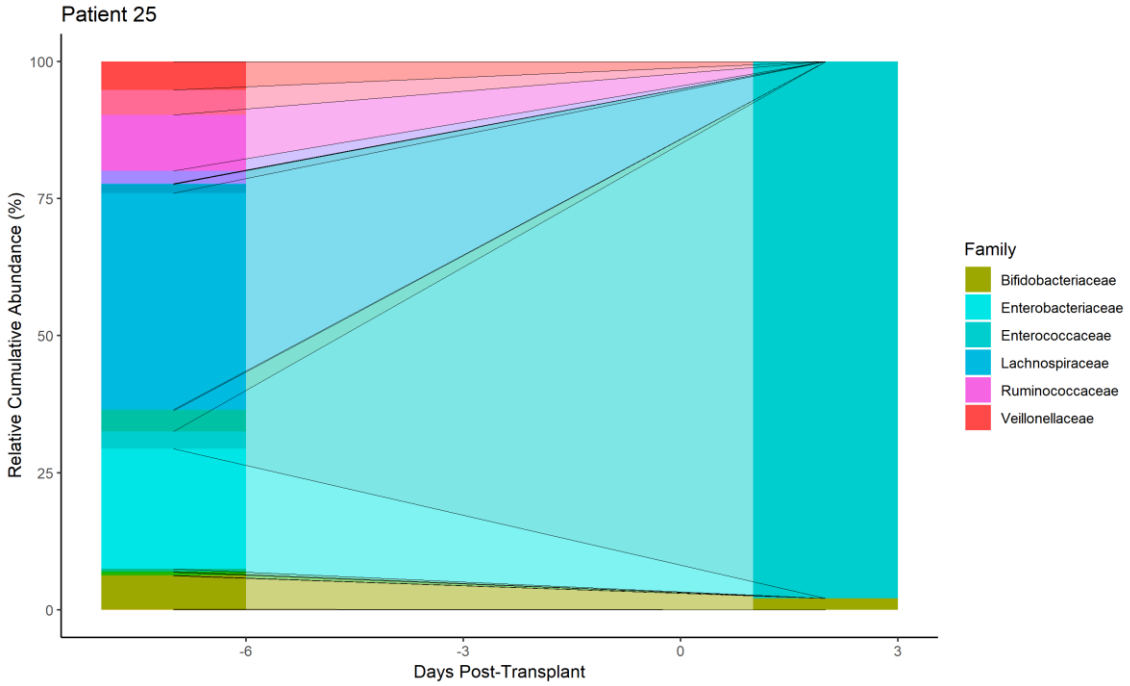

Patient 20
Relative Cumulative Abundance (%)
100
75
50
25
0
-100
0
100
200
300
Days Post-Transplant
Family
Bacteroidaceae
Bifidobacteriaceae
Enterobacteriaceae
Enterococcaceae
Lachnospiraceae
Lactobacillaceae
Pasteurellaceae
Staphylococcaceae
Veillonellaceae

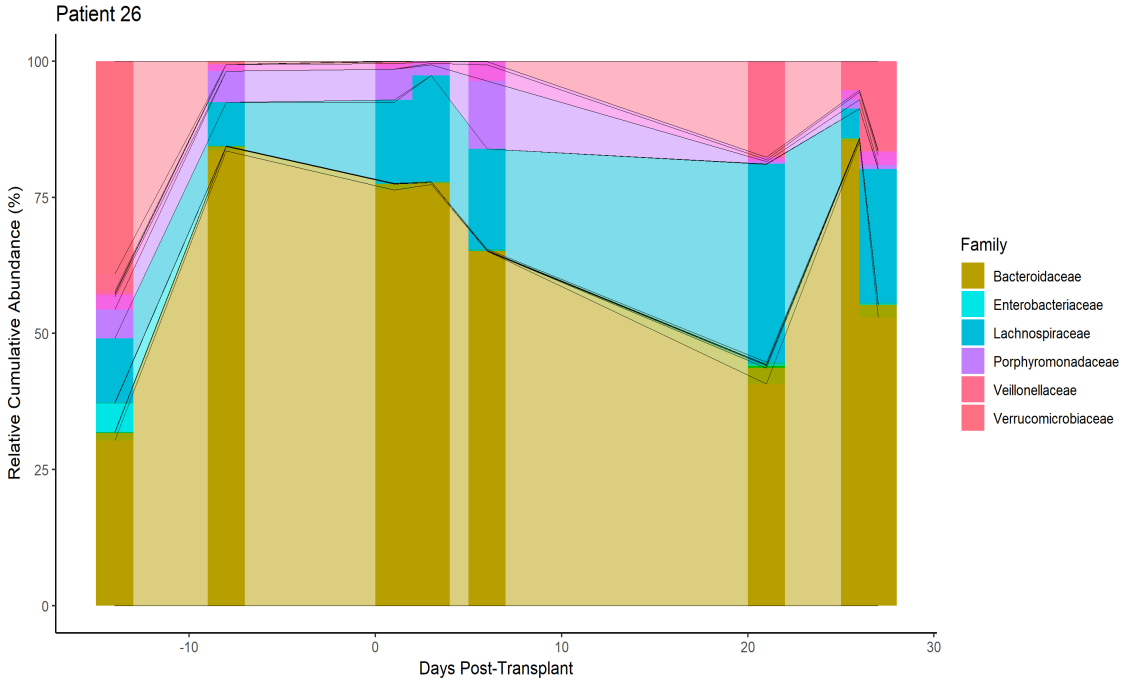

Patient 21
Relative Cumulative Abundance (%)
100
75
50
25
0
0
20
40
Days Post-Transplant
Family
Enterobacteriaceae
Enterococcaceae
Erysipelotrichaceae
Lachnospiraceae
Veillonellaceae

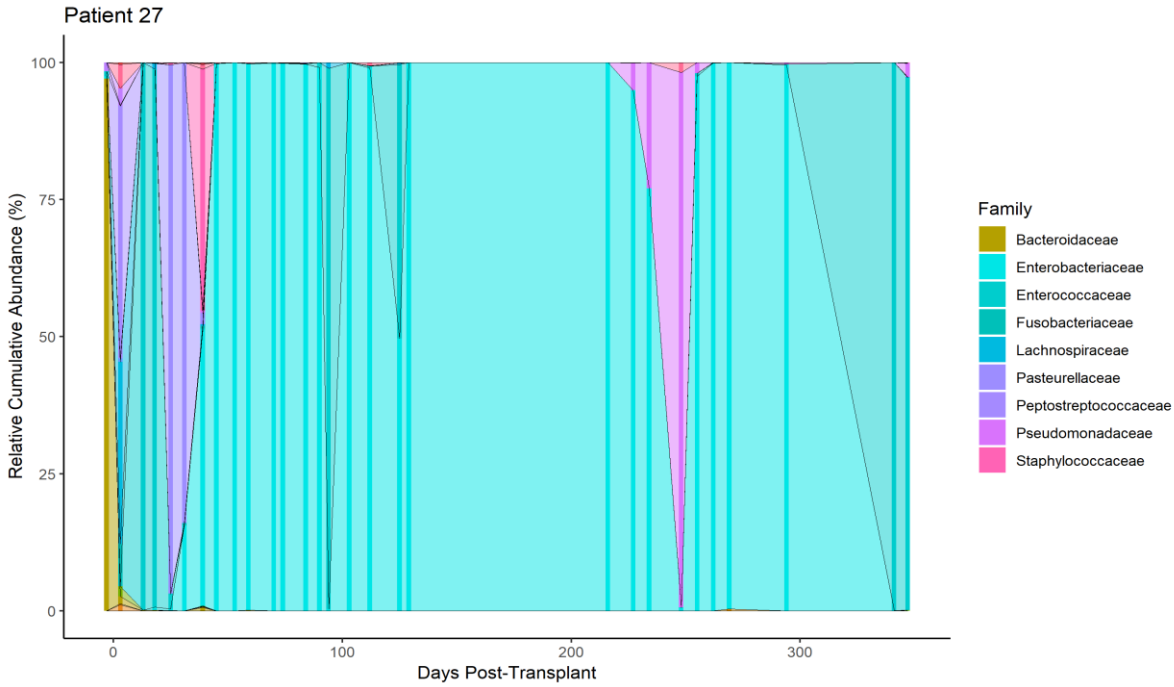

Patient 22
Relative Cumulative Abundance (%)
100
75
50
25
0
-10
0
10
20
30
Days Post-Transplant
Family
Actinomycetaceae
Campylobacteraceae
Enterococcaceae
Lachnospiraceae
Micrococcaceae
Staphylococcaceae
Streptococcaceae
Veillonellaceae

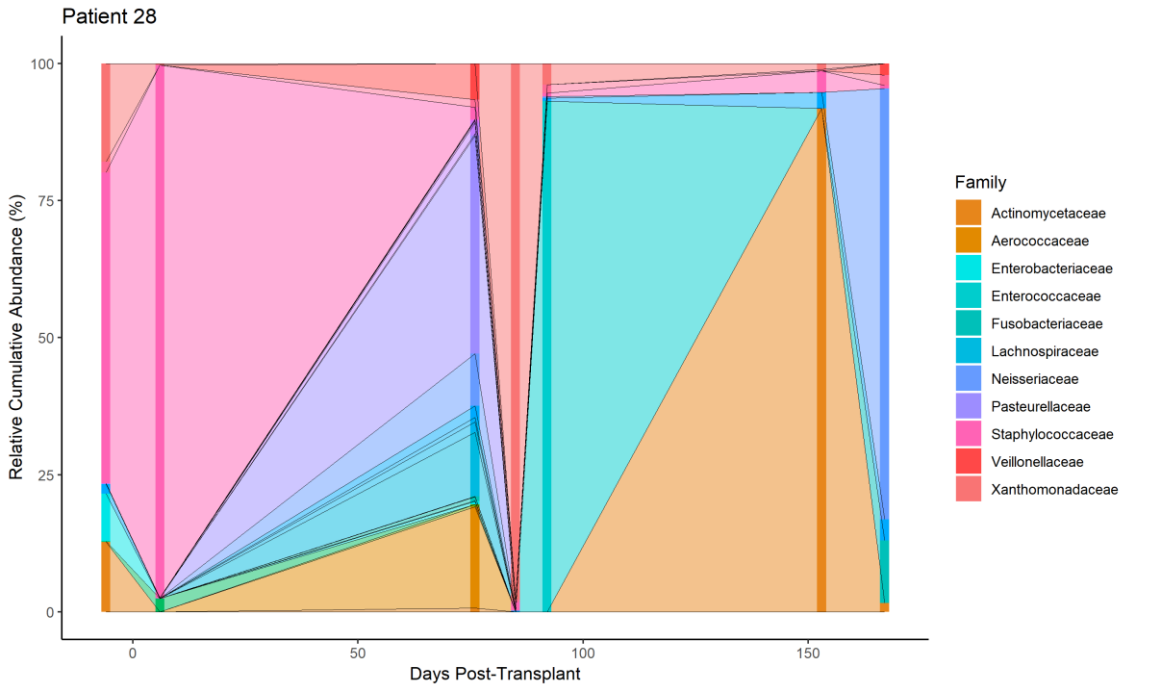

Patient 23
Relative Cumulative Abundance (%)
100
75
50
25
0
0
25
50
75
100
Days Post-Transplant
Family
Enterobacteriaceae
Enterococcaceae
Erysipelotrichaceae
Lachnospiraceae
Peptostreptococcaceae
Pseudomonadaceae
Ruminococcaceae
Staphylococcaceae
Streptococcaceae
Veillonellaceae

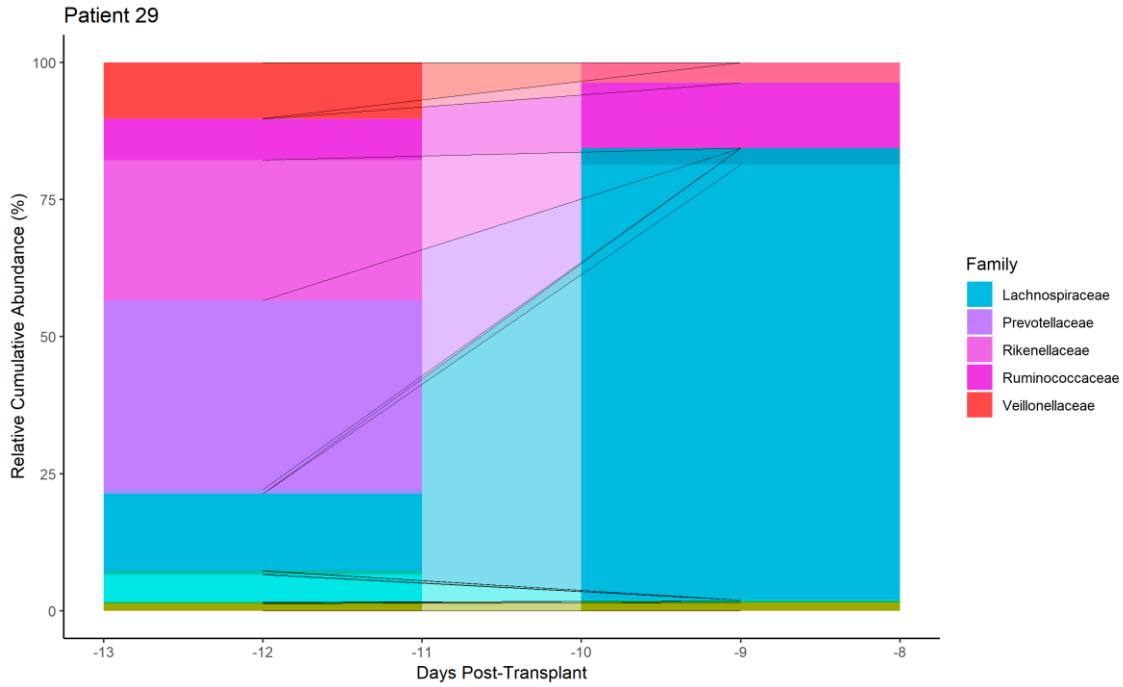

Patient 24
Relative Cumulative Abundance (%)
100
75
50
25
0
0
100
200
300
Days Post-Transplant
Family
Bifidobacteriaceae
Enterobacteriaceae
Enterococcaceae
Erysipelotrichaceae
Lachnospiraceae
Pasteurellaceae
Staphylococcaceae
Streptococcaceae
Veillonellaceae

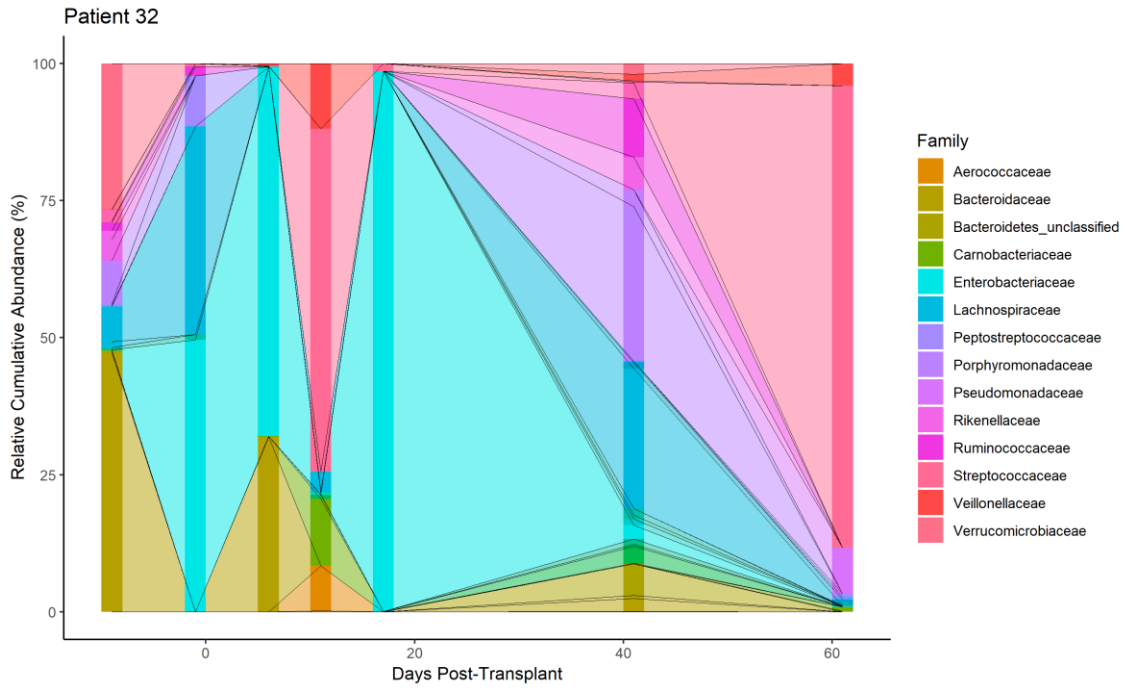

Patient 25
Relative Cumulative Abundance (%)
100
75
50
25
0
-6
-3
0
3
Days Post-Transplant
Family
Bifidobacteriaceae
Enterobacteriaceae
Enterococcaceae
Lachnospiraceae
Ruminococcaceae
Veillonellaceae

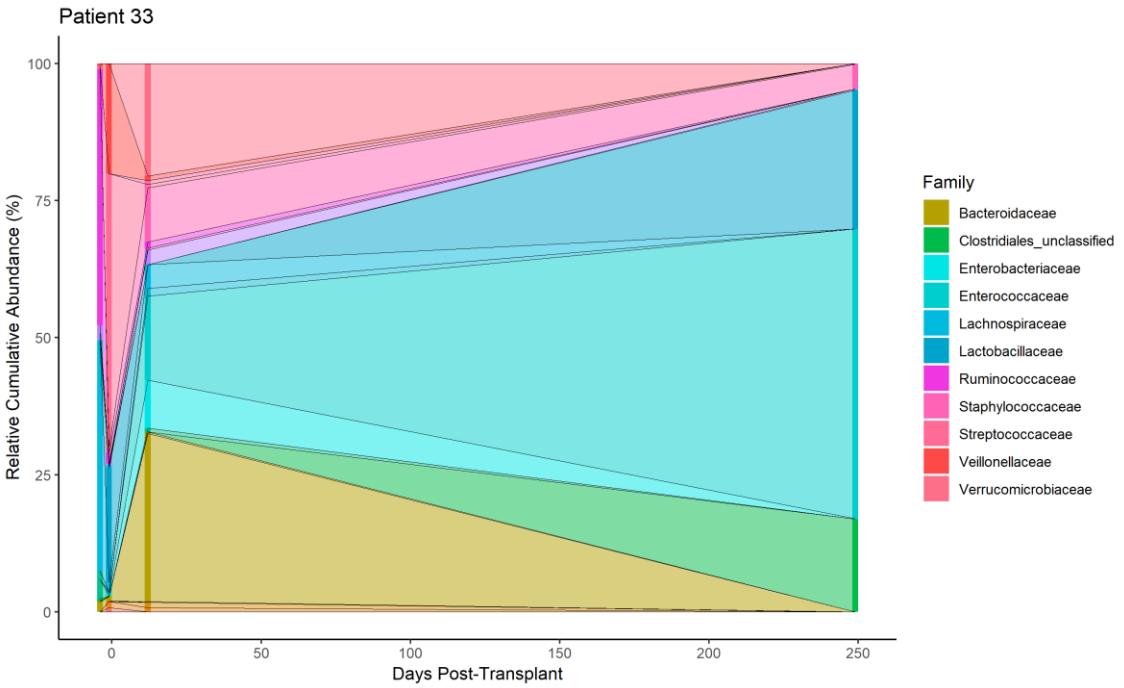

Patient 26
Relative Cumulative Abundance (%)
100
75
50
25
0
-10
0
10
20
30
Days Post-Transplant
Family
Bacteroidaceae
Enterobacteriaceae
Lachnospiraceae
Porphyromonadaceae
Veillonellaceae
Verrucomicrobiaceae

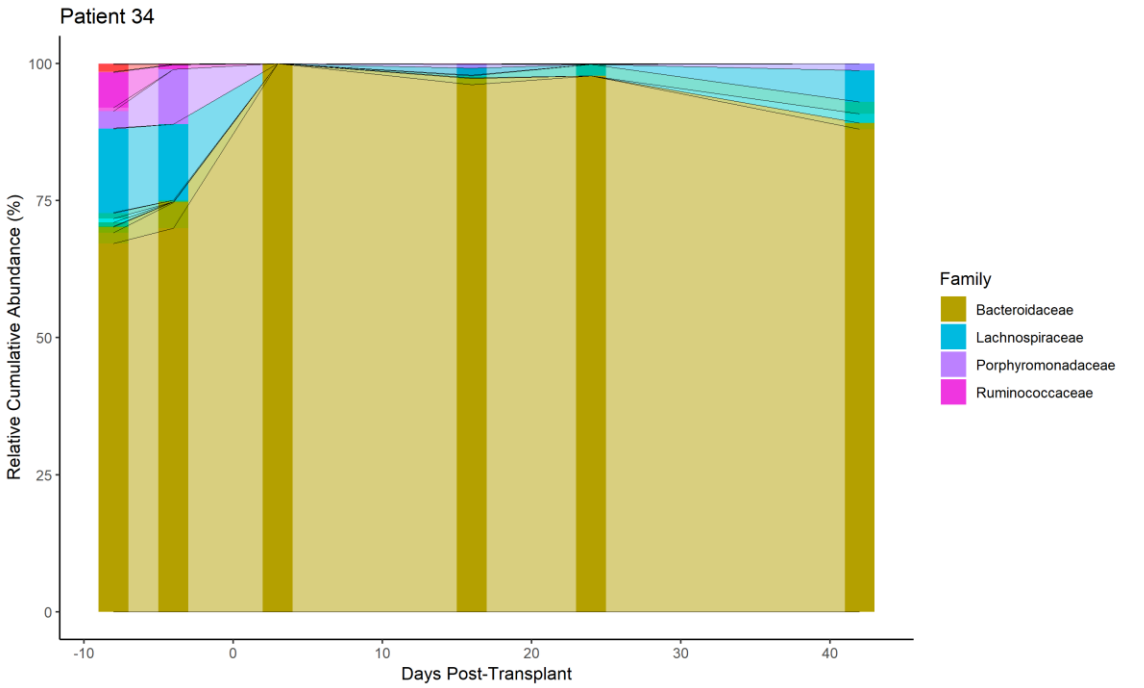

Patient 27
Relative Cumulative Abundance (%)
100
75
50
25
0
0
100
200
300
Days Post-Transplant
Family
Bacteroidaceae
Enterobacteriaceae
Enterococcaceae
Fusobacteriaceae
Lachnospiraceae
Pasteurellaceae
Peptostreptococcaceae
Pseudomonadaceae
Staphylococcaceae

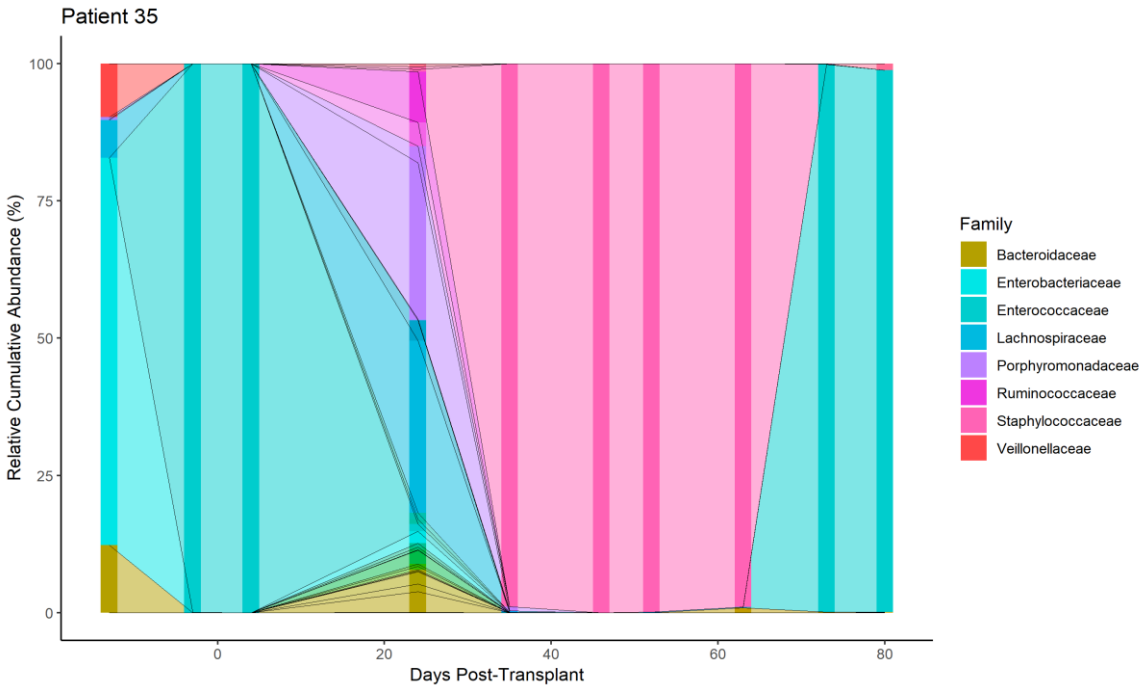

Patient 28
Relative Cumulative Abundance (%)
100
75
50
25
0
0
50
100
150
Days Post-Transplant
Family
Actinomycetaceae
Aerococcaceae
Enterobacteriaceae
Enterococcaceae
Fusobacteriaceae
Lachnospiraceae
Neisseriaceae
Pasteurellaceae
Staphylococcaceae
Veillonellaceae
Xanthomonadaceae

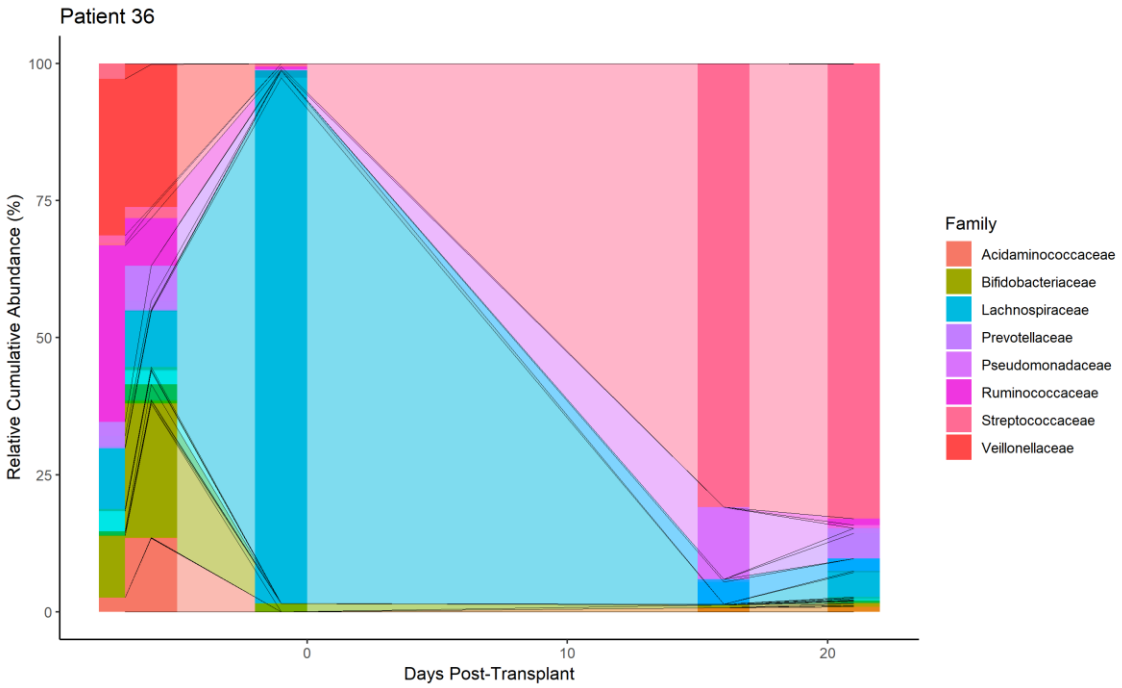

Patient 29
Relative Cumulative Abundance (%)
100
75
50
25
0
-13
-12
-11
-10
-9
-8
Days Post-Transplant
Family
Lachnospiraceae
Prevotellaceae
Rikenellaceae
Ruminococcaceae
Veillonellaceae

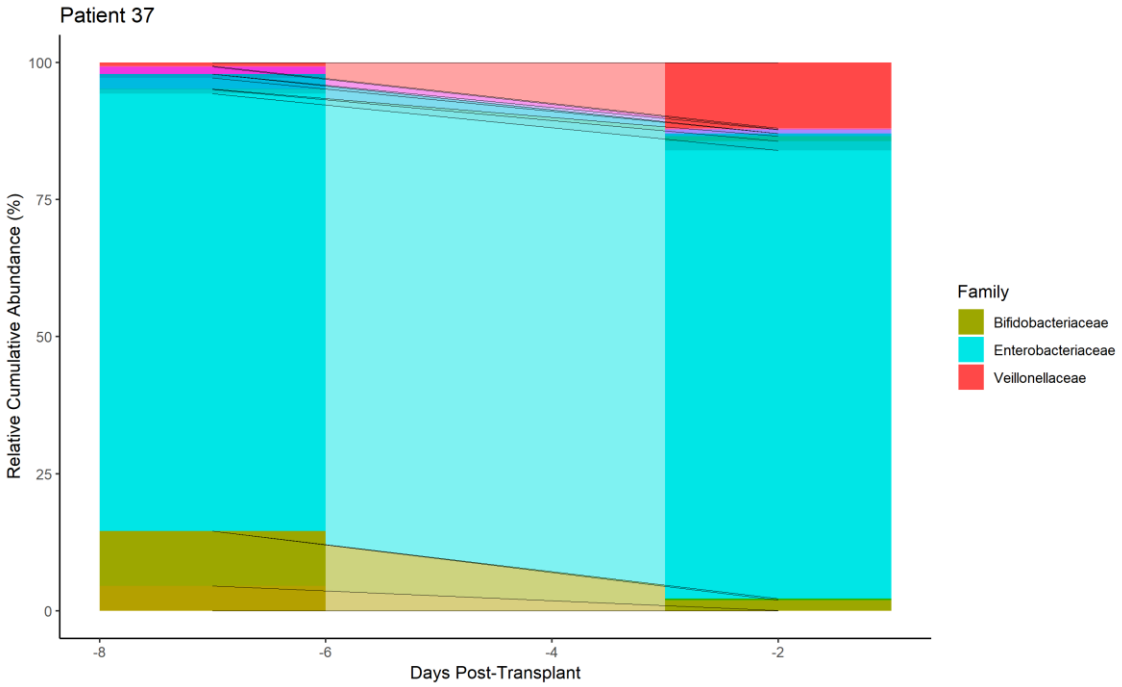

Patient 32
Relative Cumulative Abundance (%)
100
75
50
25
0
0
20
40
60
Days Post-Transplant
Family
Aerococcaceae
Bacteroidaceae
Bacteroidetes_unclassified
Carnobacteriaceae
Enterobacteriaceae
Lachnospiraceae
Peptostreptococcaceae
Porphyromonadaceae
Pseudomonadaceae
Rikenellaceae
Ruminococcaceae
Streptococcaceae
Veillonellaceae
Verrucomicrobiaceae

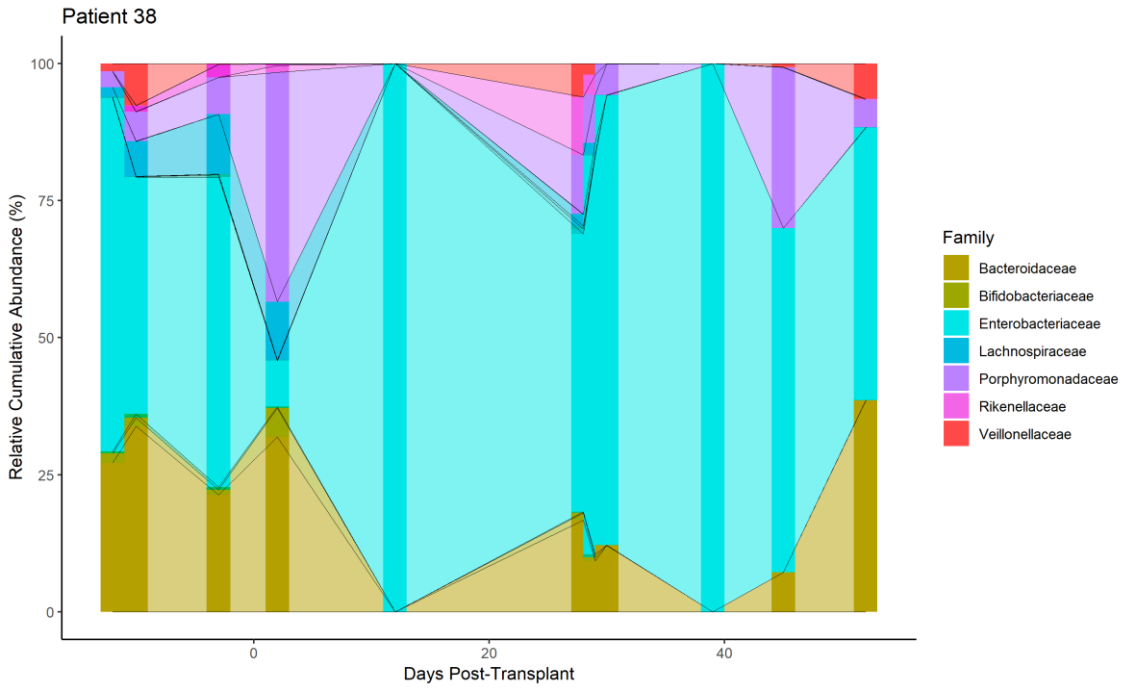

Patient 33
Relative Cumulative Abundance (%)
100
75
50
25
0
0
50
100
150
200
250
Days Post-Transplant
Family
Bacteroidaceae
Clostridiales_unclassified
Enterobacteriaceae
Enterococcaceae
Lachnospiraceae
Lactobacillaceae
Ruminococcaceae
Staphylococcaceae
Streptococcaceae
Veillonellaceae
Verrucomicrobiaceae

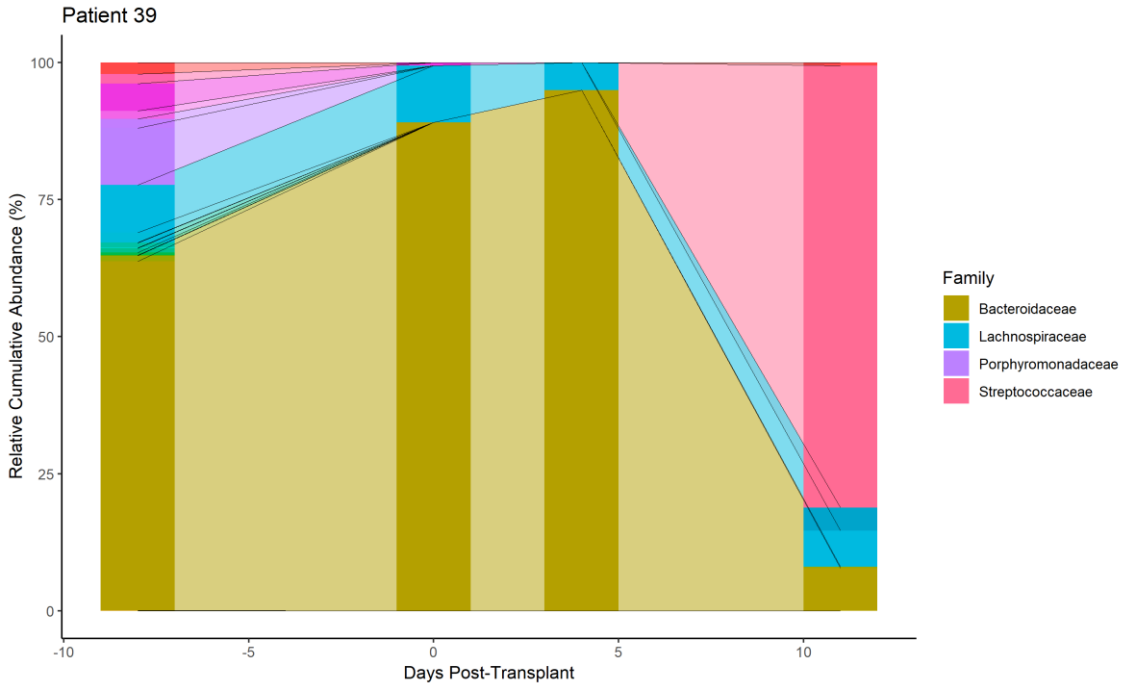

Patient 34
Relative Cumulative Abundance (%)
100
75
50
25
0
-10
0
10
20
30
40
Days Post-Transplant
Family
Bacteroidaceae
Lachnospiraceae
Porphyromonadaceae
Ruminococcaceae

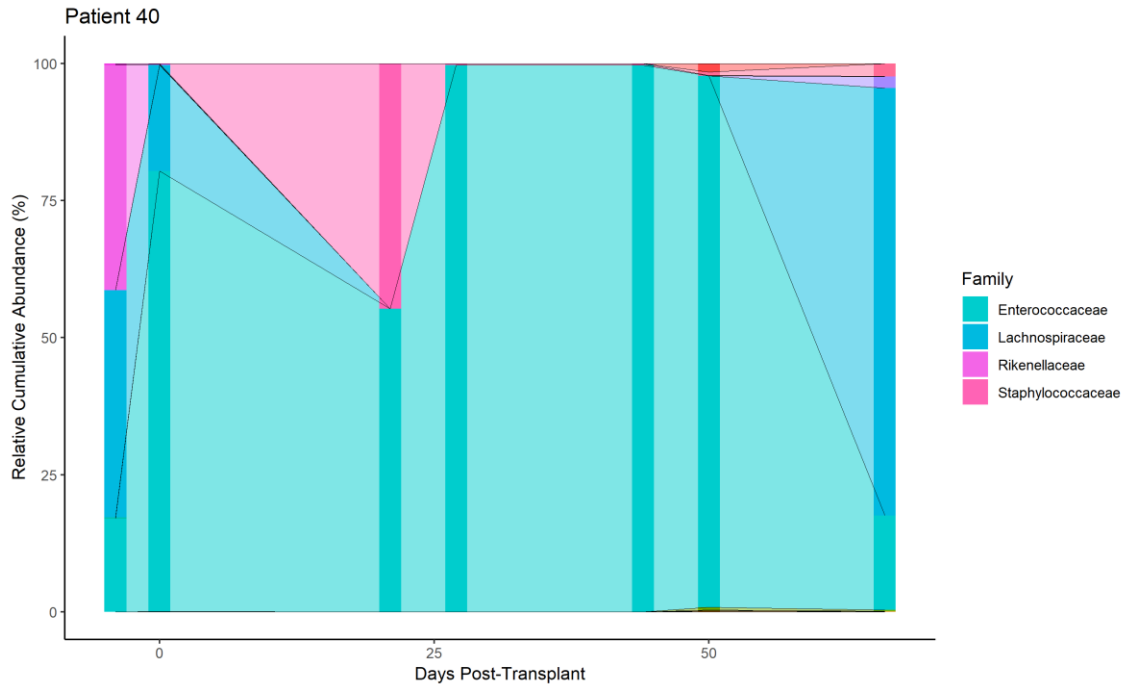

Patient 35
Relative Cumulative Abundance (%)
100
75
50
25
0
0
20
40
60
80
Days Post-Transplant
Family
Bacteroidaceae
Enterobacteriaceae
Enterococcaceae
Lachnospiraceae
Porphyromonadaceae
Ruminococcaceae
Staphylococcaceae
Veillonellaceae

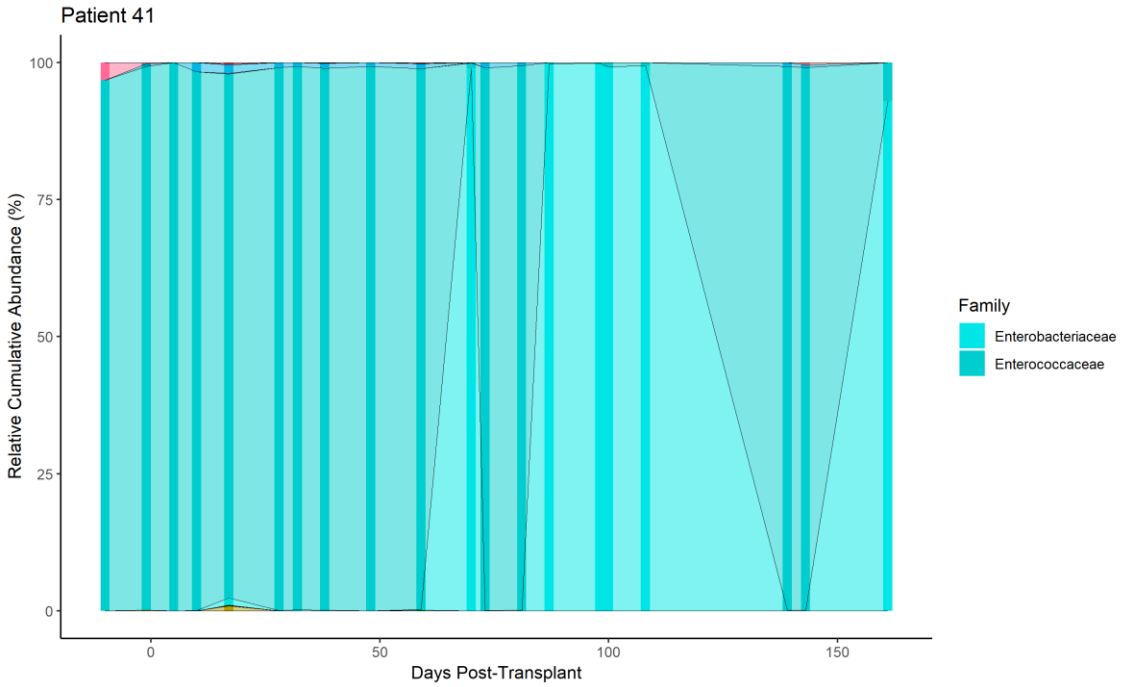

Patient 36
Relative Cumulative Abundance (%)
100
75
50
25
0
0
10
20
Days Post-Transplant
Family
Acidaminococcaceae
Bifidobacteriaceae
Lachnospiraceae
Prevotellaceae
Pseudomonadaceae
Ruminococcaceae
Streptococcaceae
Veillonellaceae

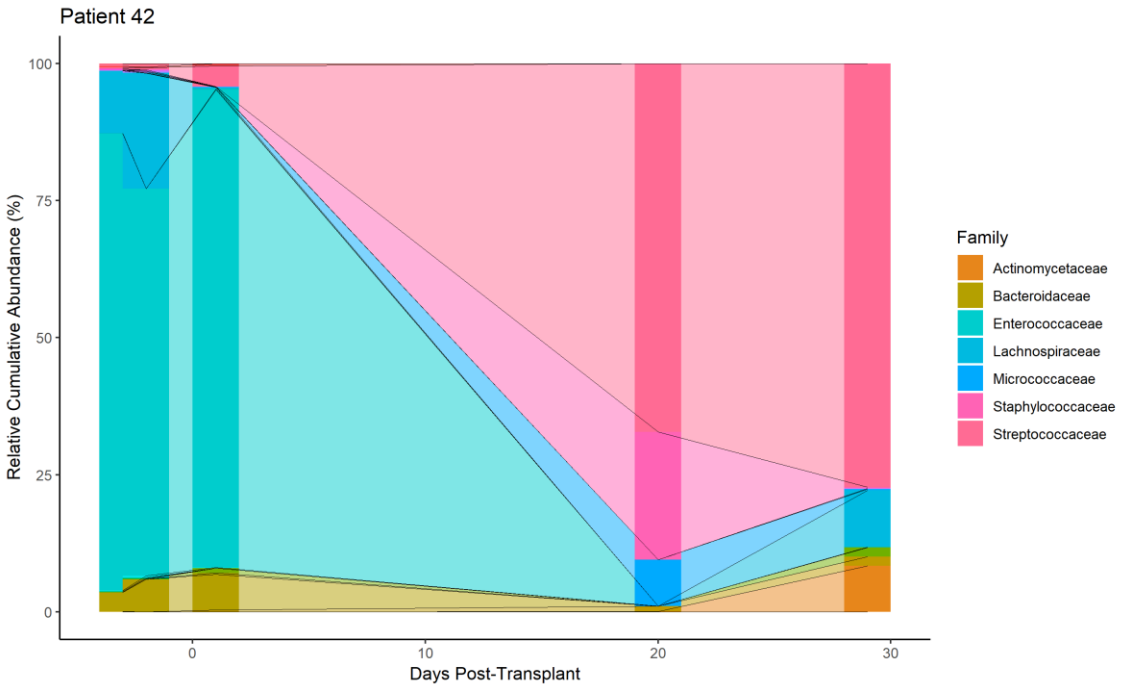

Patient 37
Relative Cumulative Abundance (%)
100
75
50
25
0
-8
-6
-4
-2
Days Post-Transplant
Family
Bifidobacteriaceae
Enterobacteriaceae
Veillonellaceae

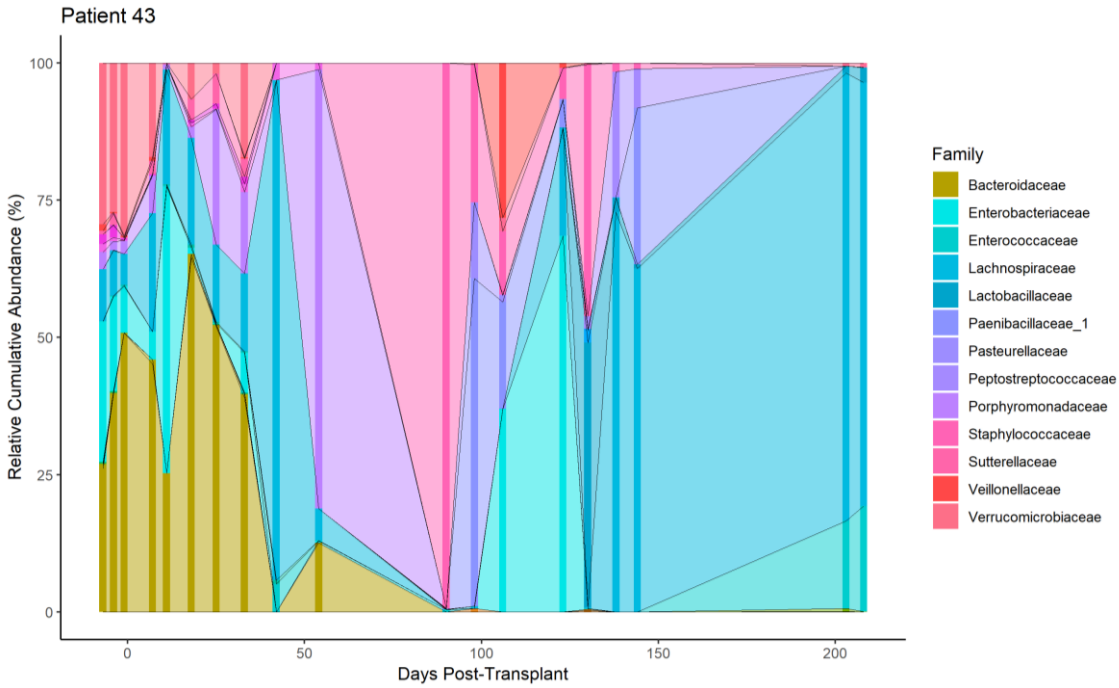

Patient 38
Relative Cumulative Abundance (%)
100
75
50
25
0
0
20
40
Days Post-Transplant
Family
Bacteroidaceae
Bifidobacteriaceae
Enterobacteriaceae
Lachnospiraceae
Porphyromonadaceae
Rikenellaceae
Veillonellaceae

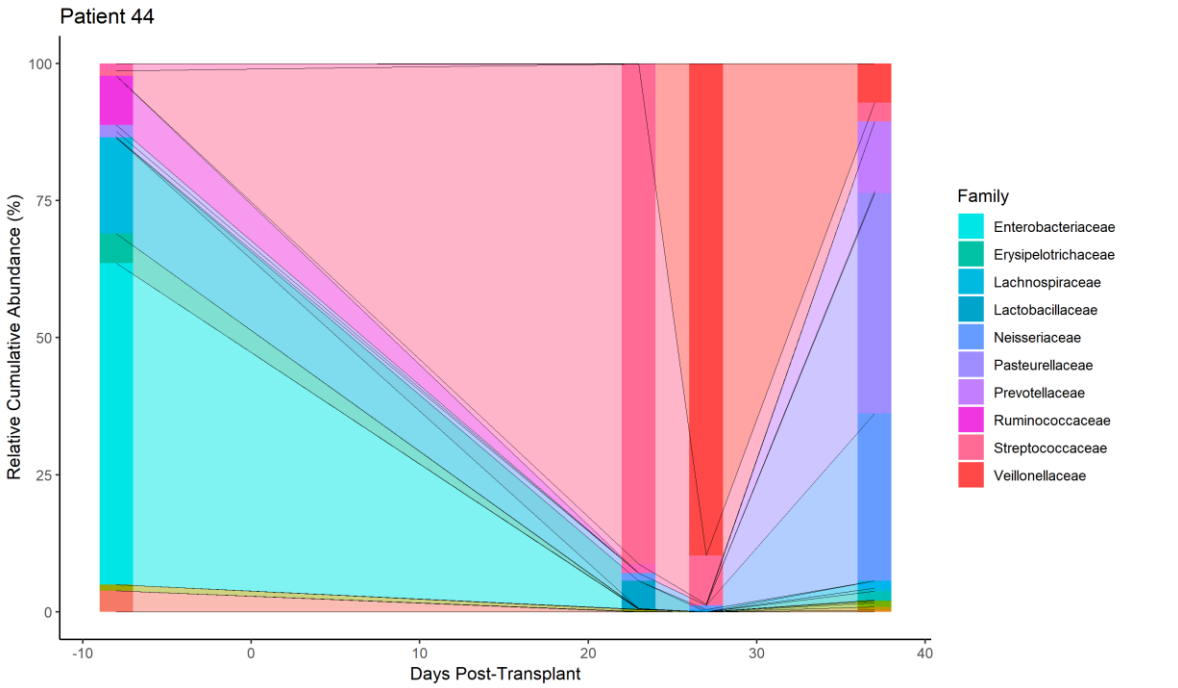

Patient 39
Relative Cumulative Abundance (%)
100
75
50
25
0
-10
-5
0
5
10
Days Post-Transplant
Family
Bacteroidaceae
Lachnospiraceae
Porphyromonadaceae
Streptococcaceae

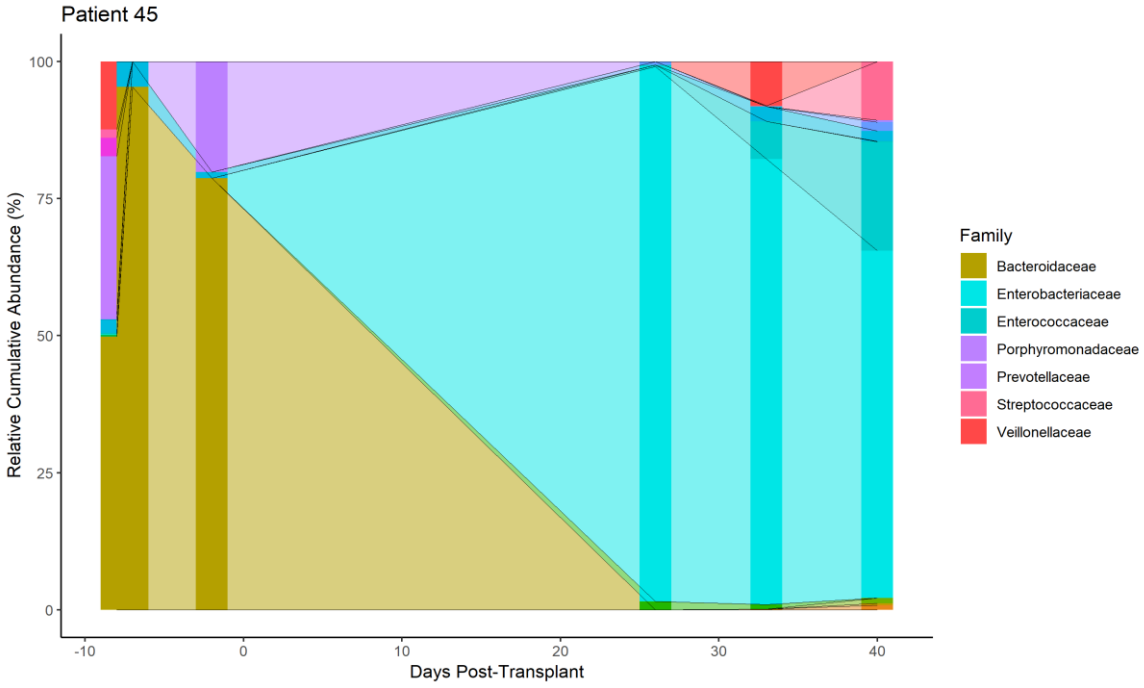

Patient 40
Relative Cumulative Abundance (%)
100
75
50
25
0
0
25
50
Days Post-Transplant
Family
Enterococcaceae
Lachnospiraceae
Rikenellaceae
Staphylococcaceae

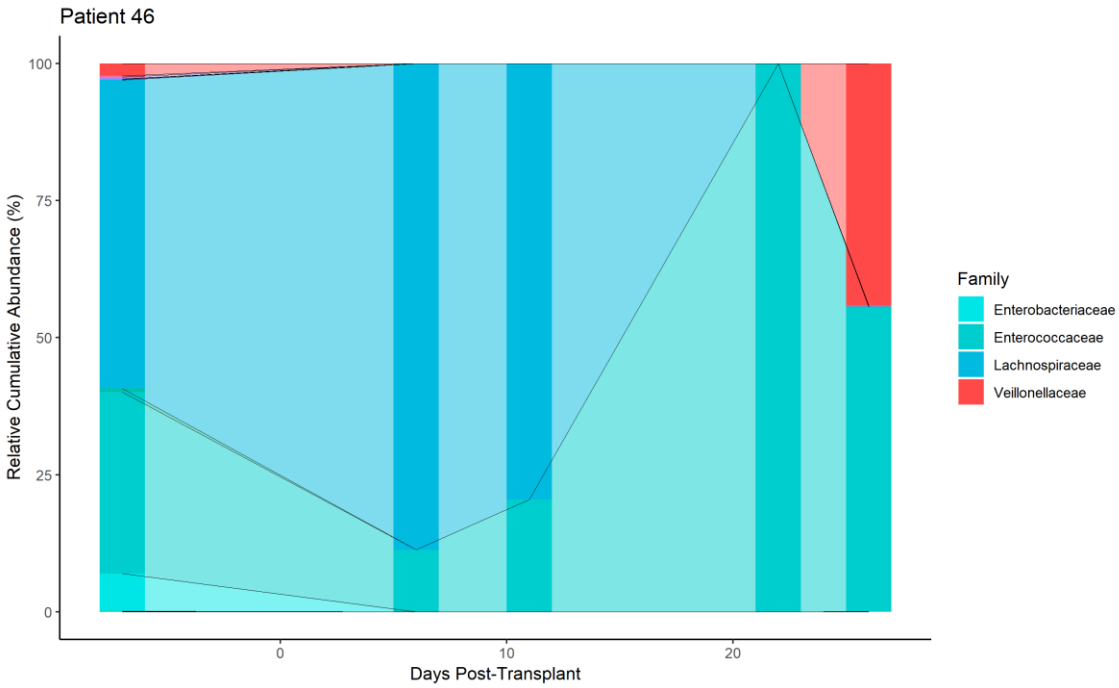

Patient 41
Relative Cumulative Abundance (%)
100
75
50
25
0
0
50
100
150
Days Post-Transplant
Family
Enterobacteriaceae
Enterococcaceae

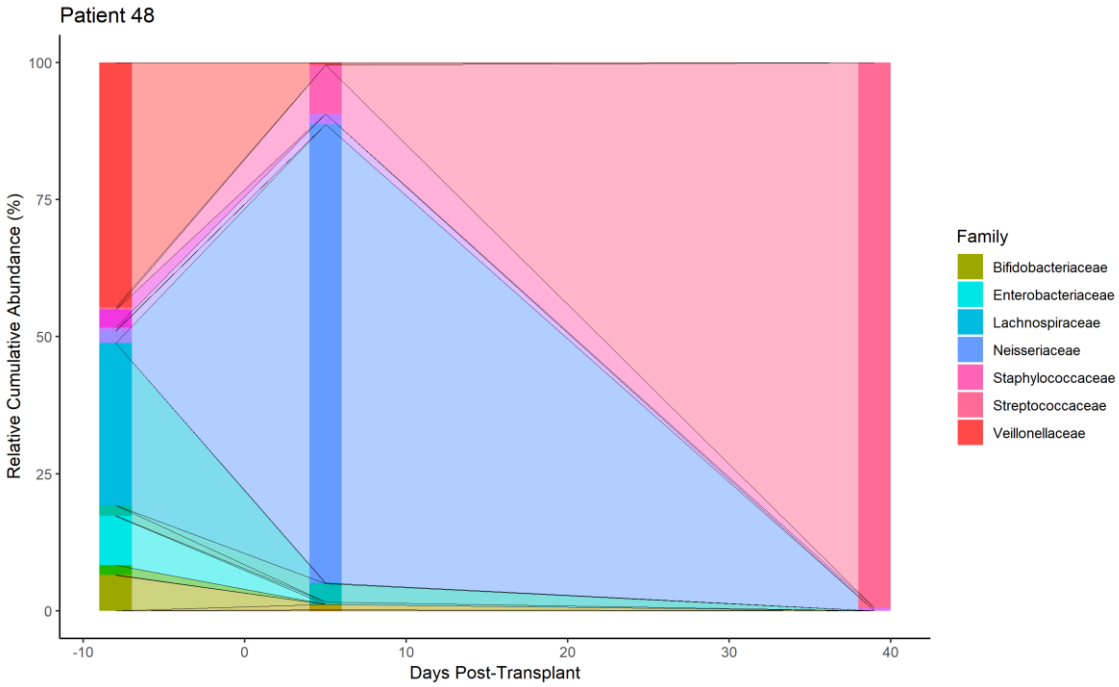

Patient 42
Relative Cumulative Abundance (%)
100
75
50
25
0
0
10
20
30
Days Post-Transplant
Family
Actinomycetaceae
Bacteroidaceae
Enterococcaceae
Lachnospiraceae
Micrococcaceae
Staphylococcaceae
Streptococcaceae

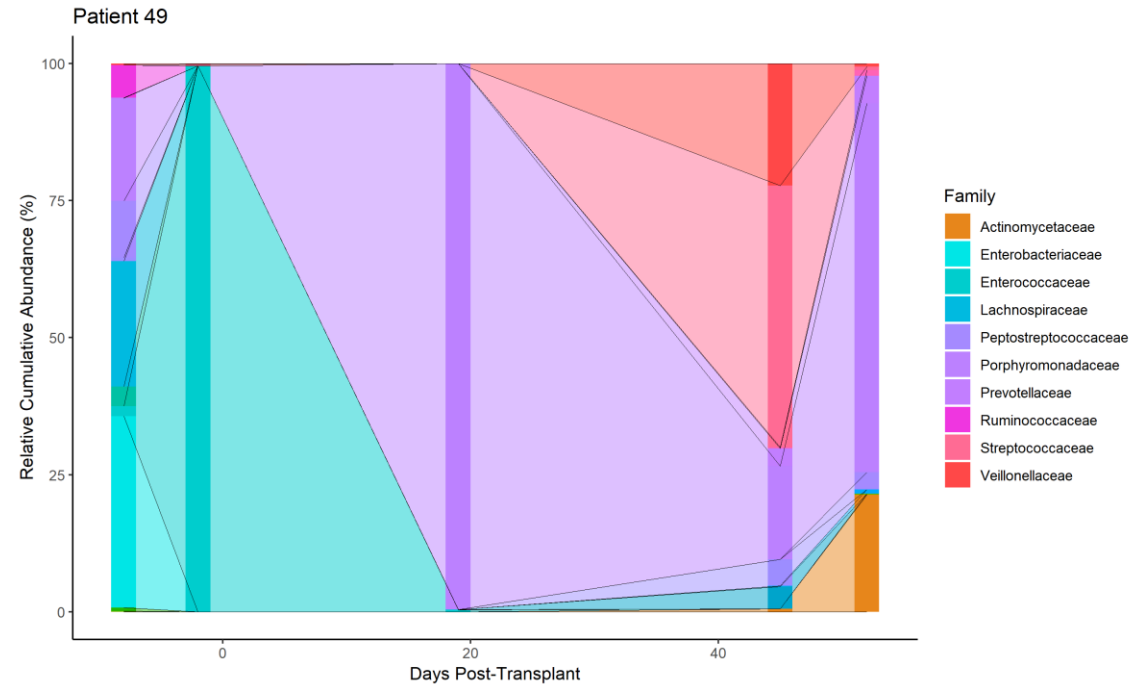

Patient 43
Relative Cumulative Abundance (%)
100
75
50
25
0
0
50
100
150
200
Days Post-Transplant
Family
Bacteroidaceae
Enterobacteriaceae
Enterococcaceae
Lachnospiraceae
Lactobacillaceae
Paenibacillaceae_1
Pasteurellaceae
Peptostreptococcaceae
Porphyromonadaceae
Staphylococcaceae
Sutterellaceae
Veillonellaceae
Verrucomicrobiaceae

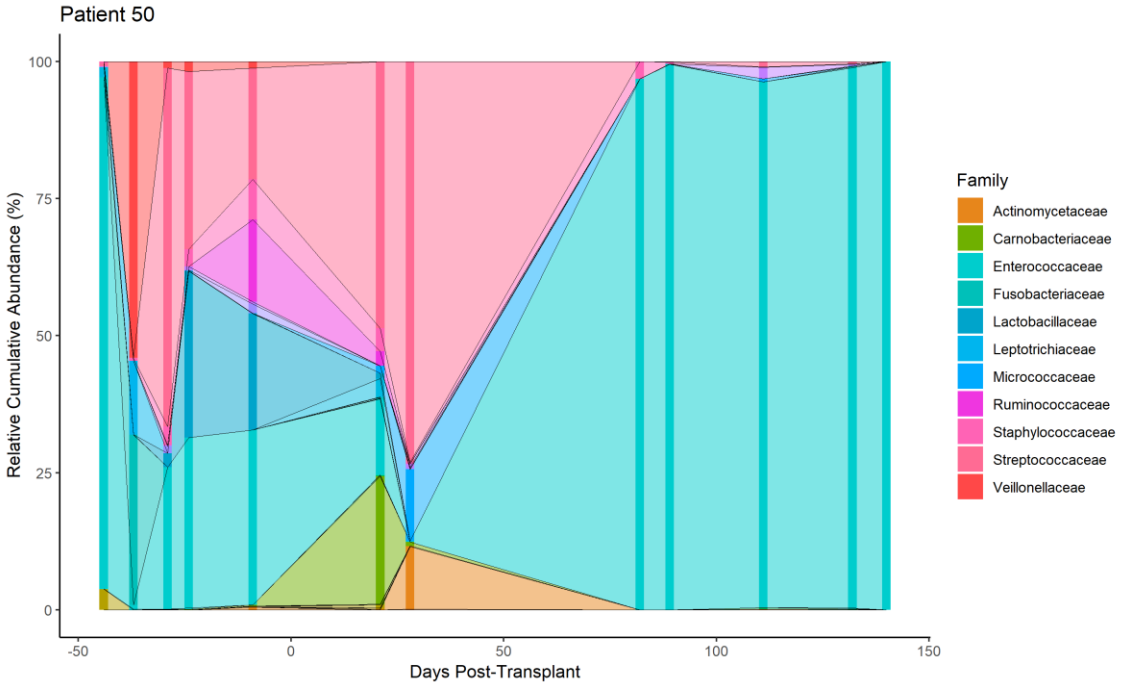

Patient 44
Relative Cumulative Abundance (%)
100
75
50
25
0
-10
0
10
20
30
40
Days Post-Transplant
Family
Enterobacteriaceae
Erysipelotrichaceae
Lachnospiraceae
Lactobacillaceae
Neisseriaceae
Pasteurellaceae
Prevotellaceae
Ruminococcaceae
Streptococcaceae
Veillonellaceae

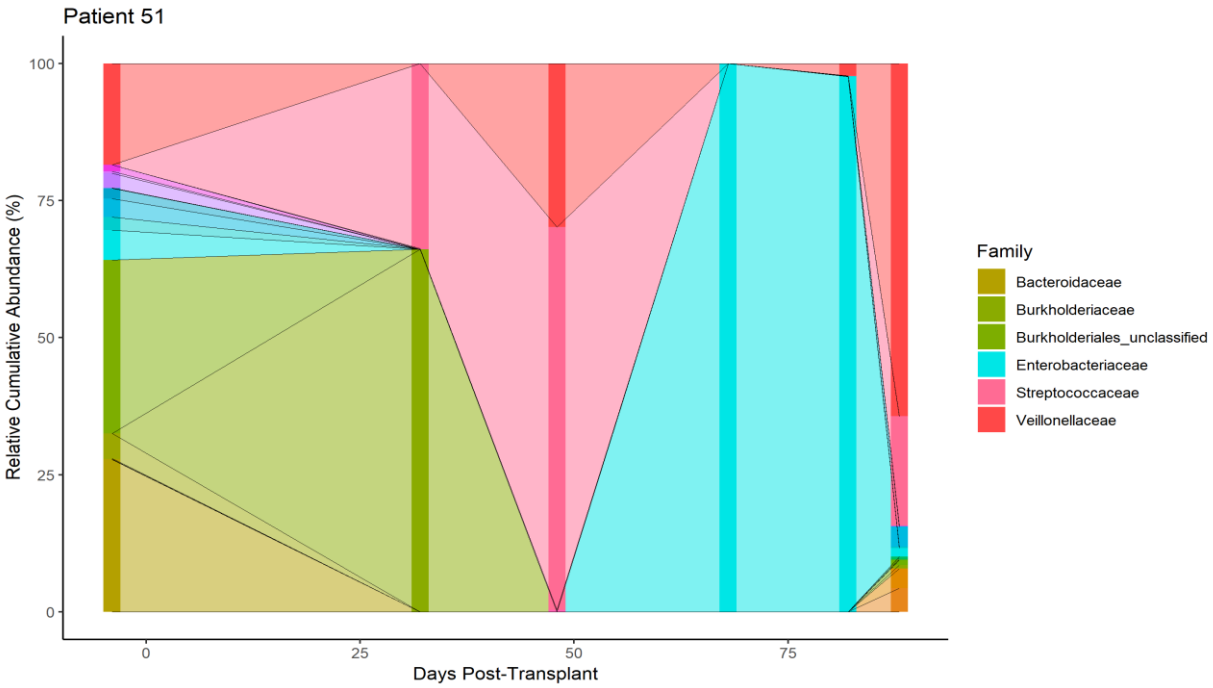

Patient 45
Relative Cumulative Abundance (%)
100
75
50
25
0
-10
0
10
20
30
40
Days Post-Transplant
Family
Bacteroidaceae
Enterobacteriaceae
Enterococcaceae
Porphyromonadaceae
Prevotellaceae
Streptococcaceae
Veillonellaceae

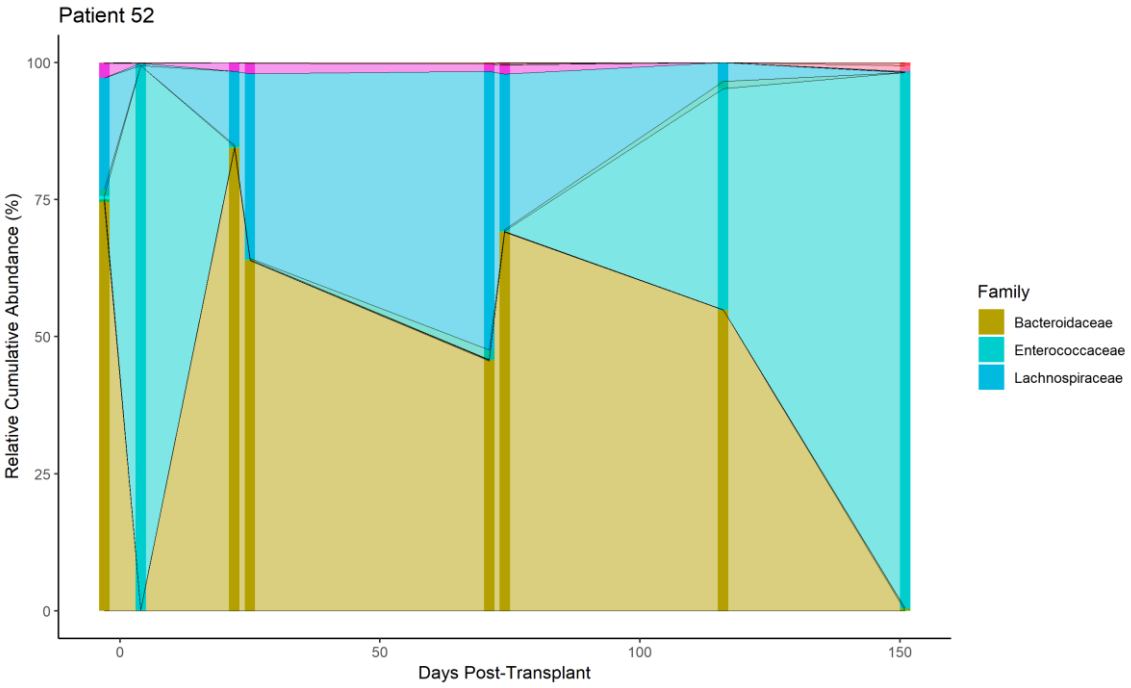

Patient 46
Relative Cumulative Abundance (%)
100
75
50
25
0
0
10
20
Days Post-Transplant
Family
Enterobacteriaceae
Enterococcaceae
Lachnospiraceae
Veillonellaceae

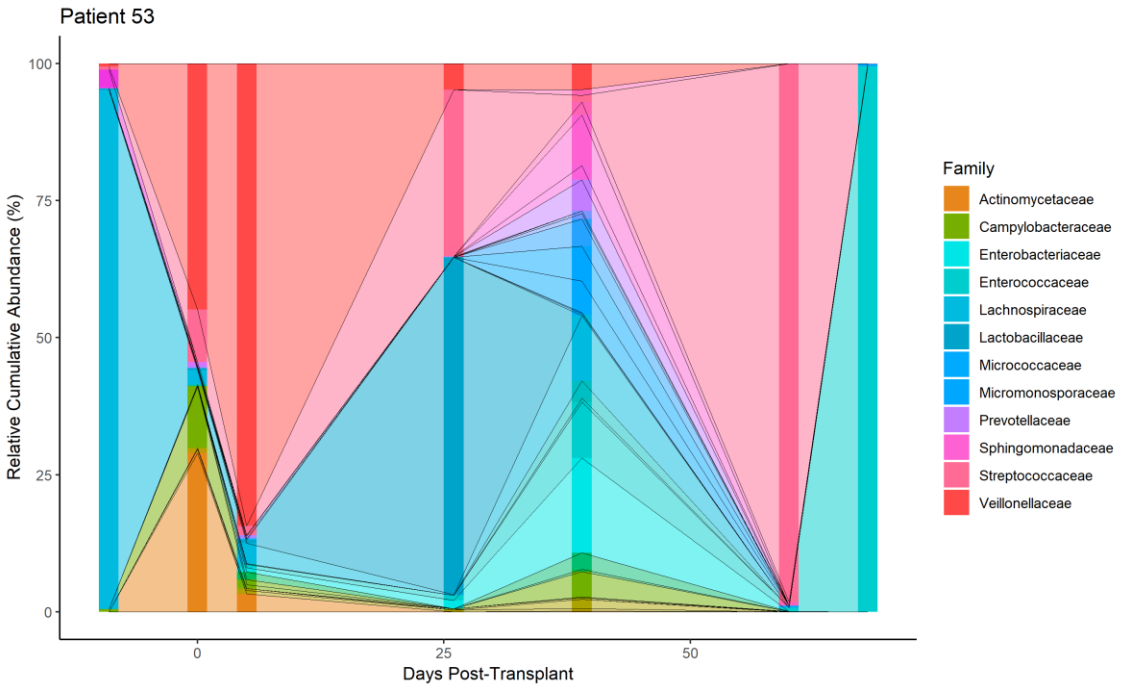

Patient 48
Relative Cumulative Abundance (%)
100
75
50
25
0
-10
0
10
20
30
40
Days Post-Transplant
Family
Bifidobacteriaceae
Enterobacteriaceae
Lachnospiraceae
Neisseriaceae
Staphylococcaceae
Streptococcaceae
Veillonellaceae

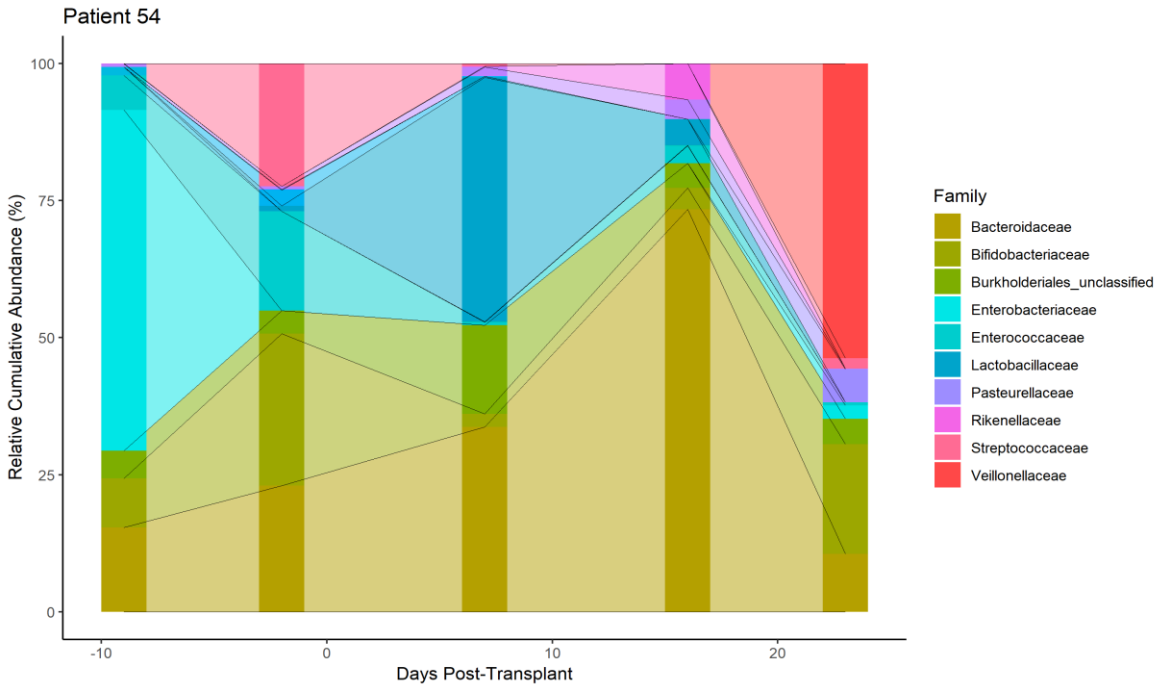

Patient 49
Relative Cumulative Abundance (%)
100
75
50
25
0
0
20
40
Days Post-Transplant
Family
Actinomycetaceae
Enterobacteriaceae
Enterococcaceae
Lachnospiraceae
Peptostreptococcaceae
Porphyromonadaceae
Prevotellaceae
Ruminococcaceae
Streptococcaceae
Veillonellaceae

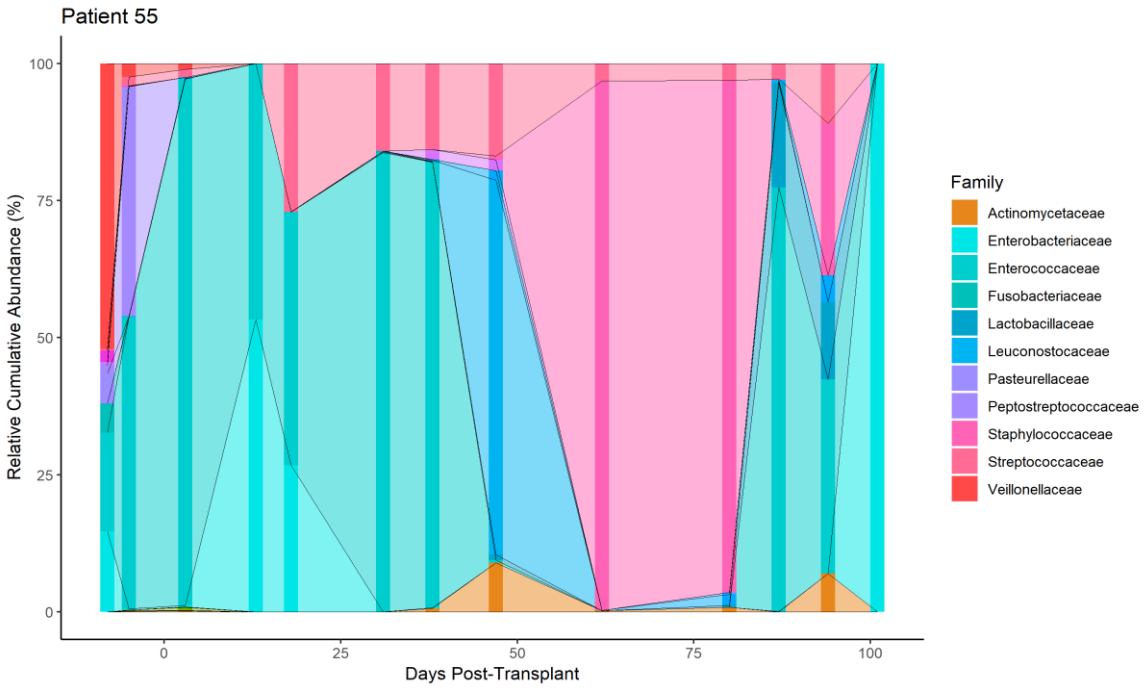

Patient 50
Relative Cumulative Abundance (%)
100
75
50
25
0
-50
0
50
100
150
Days Post-Transplant
Family
Actinomycetaceae
Carnobacteriaceae
Enterococcaceae
Fusobacteriaceae
Lactobacillaceae
Leptotrichiaceae
Micrococcaceae
Ruminococcaceae
Staphylococcaceae
Streptococcaceae
Veillonellaceae

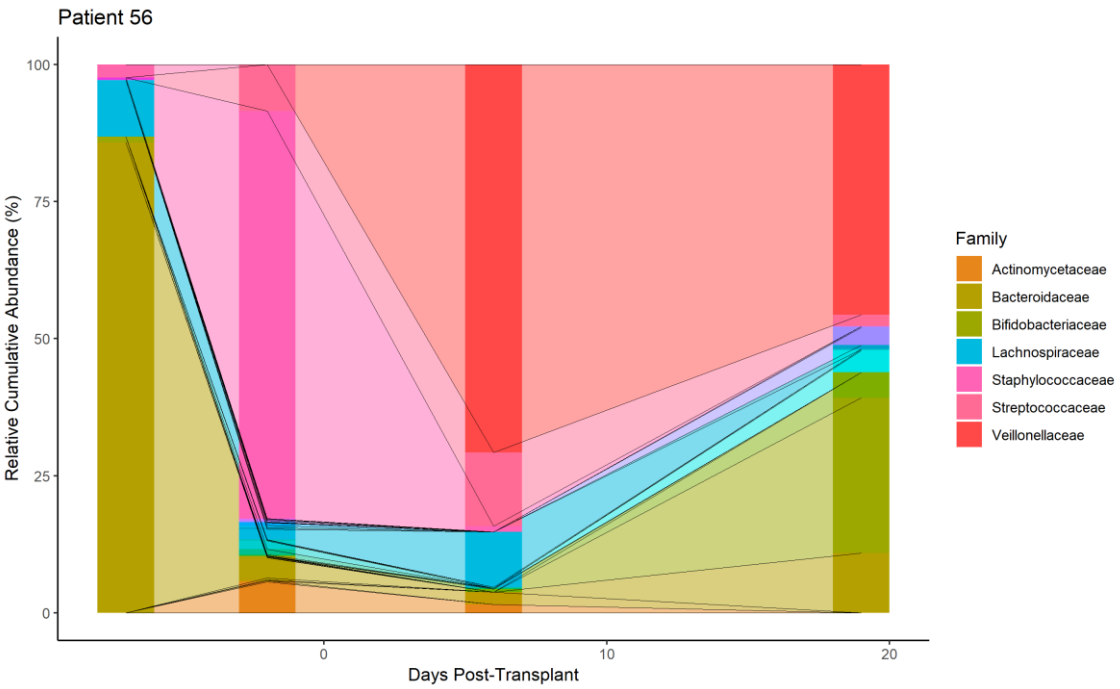

Patient 51
Relative Cumulative Abundance (%)
100
75
50
25
0
0
25
50
75
Days Post-Transplant
Family
Bacteroidaceae
Burkholderiaceae
Burkholderiales_unclassified
Enterobacteriaceae
Streptococcaceae
Veillonellaceae

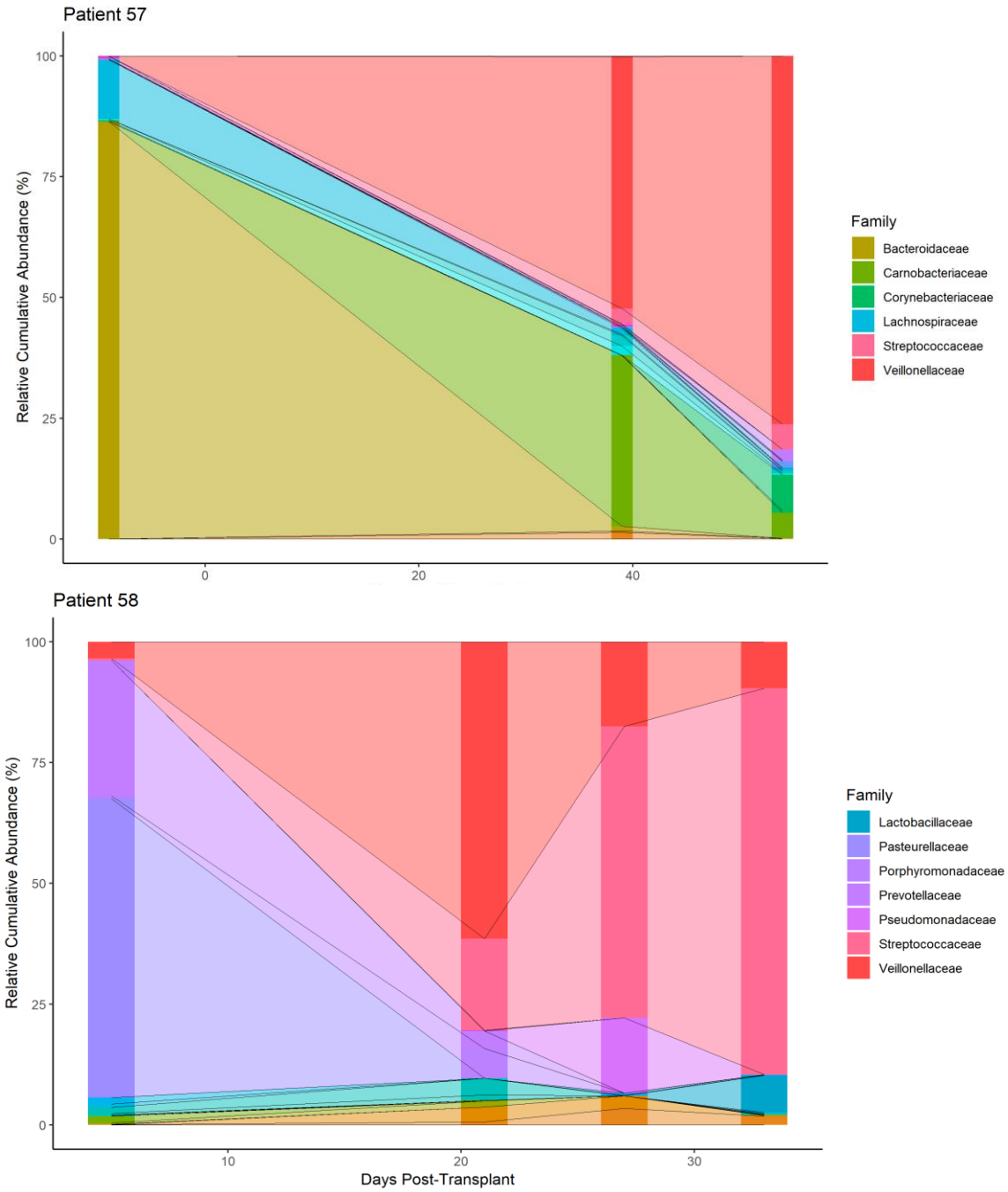

Patient 52
Relative Cumulative Abundance (%)
100
75
50
25
0
0
50
100
150
Days Post-Transplant
Family
Bacteroidaceae
Enterococcaceae
Lachnospiraceae

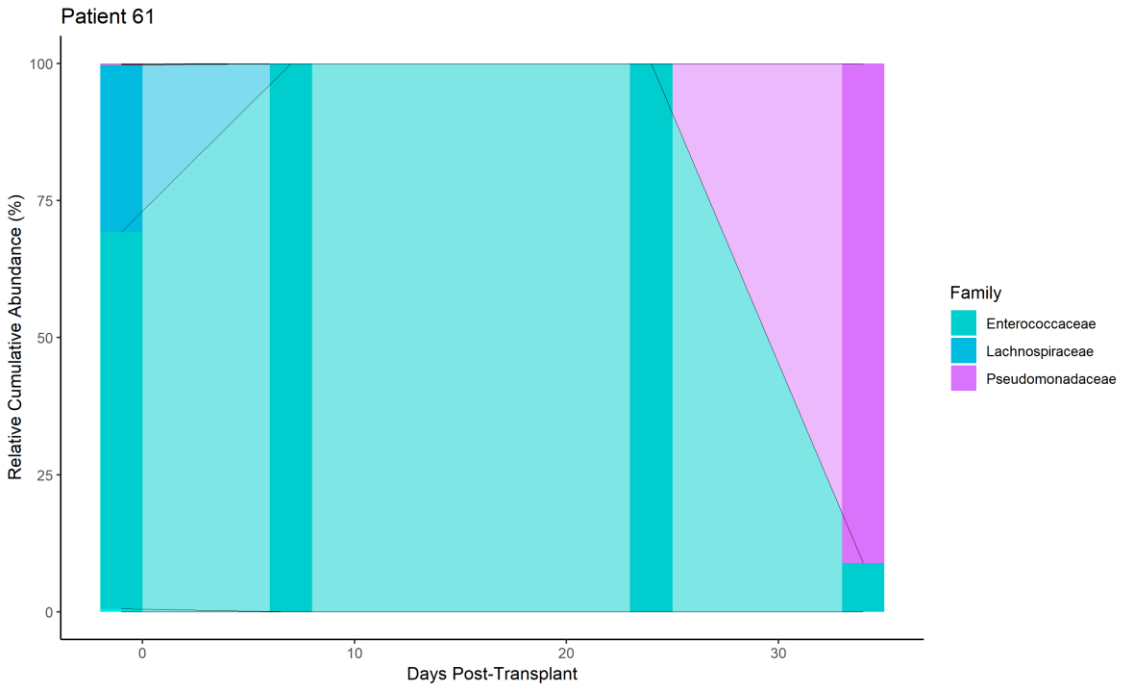

Patient 53
Relative Cumulative Abundance (%)
100
75
50
25
0
0
25
50
Days Post-Transplant
Family
Actinomycetaceae
Campylobacteraceae
Enterobacteriaceae
Enterococcaceae
Lachnospiraceae
Lactobacillaceae
Micrococcaceae
Micromonosporaceae
Prevotellaceae
Sphingomonadaceae
Streptococcaceae
Veillonellaceae

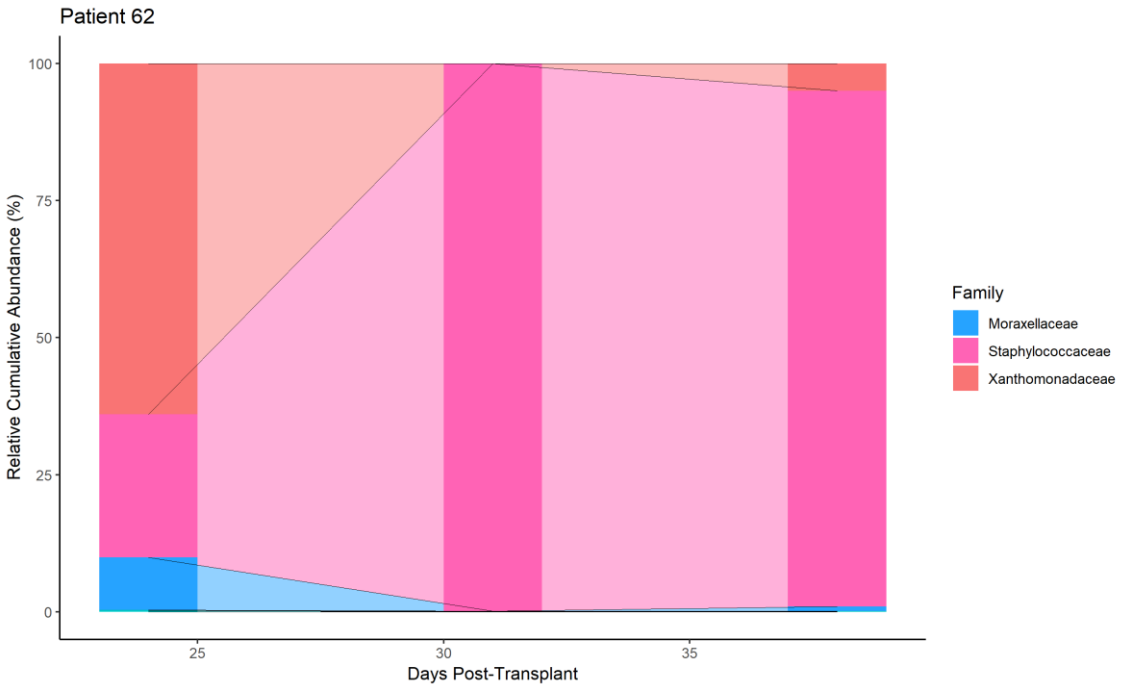

Patient 54
Relative Cumulative Abundance (%)
100
75
50
25
0
-10
0
10
20
Days Post-Transplant
Family
Bacteroidaceae
Bifidobacteriaceae
Burkholderiales_unclassified
Enterobacteriaceae
Enterococcaceae
Lactobacillaceae
Pasteurellaceae
Rikenellaceae
Streptococcaceae
Veillonellaceae

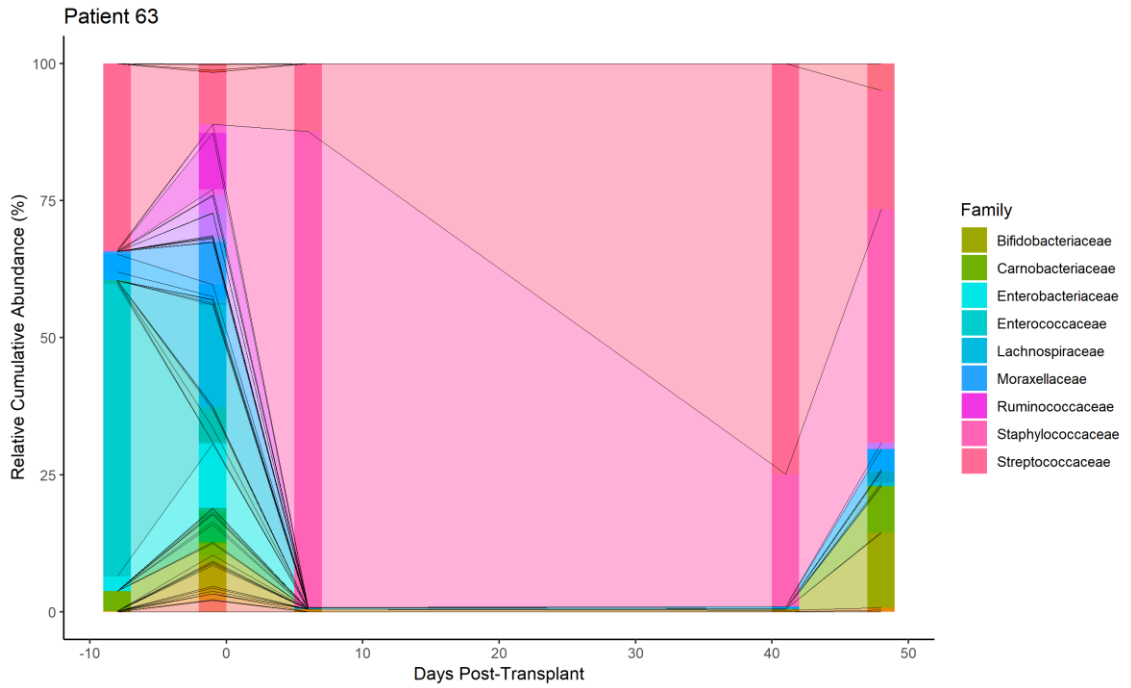

Patient 55
Relative Cumulative Abundance (%)
100
75
50
25
0
0
25
50
75
100
Days Post-Transplant
Family
Actinomycetaceae
Enterobacteriaceae
Enterococcaceae
Fusobacteriaceae
Lactobacillaceae
Leuconostocaceae
Pasteurellaceae
Peptostreptococcaceae
Staphylococcaceae
Streptococcaceae
Veillonellaceae

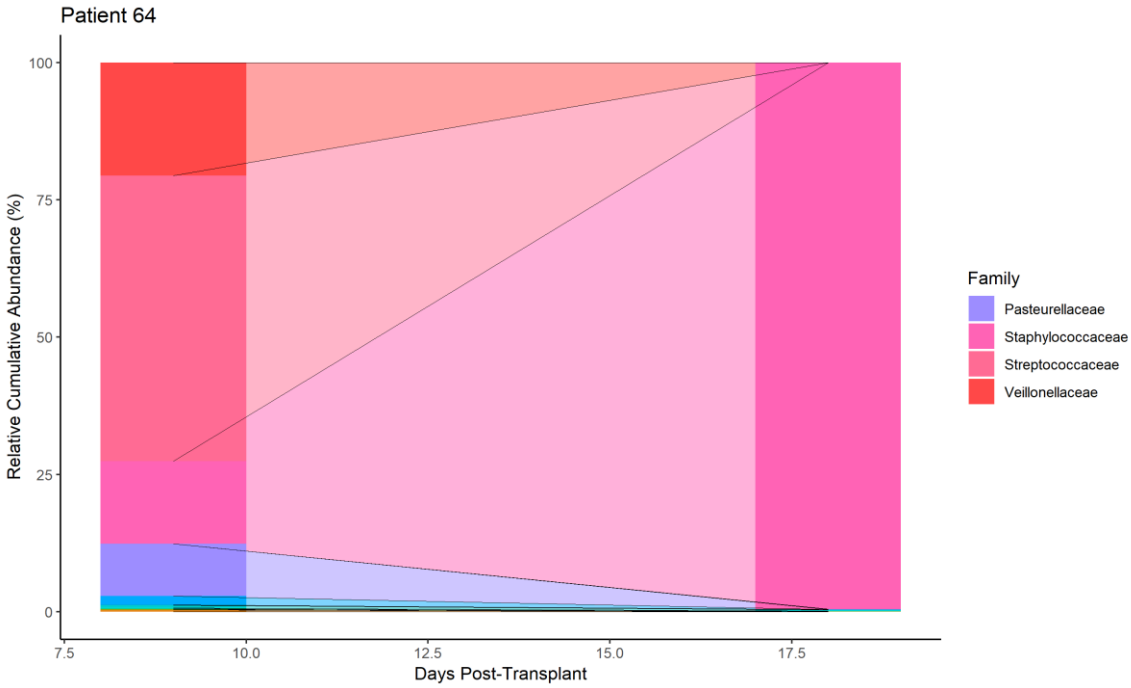

Patient 56
Relative Cumulative Abundance (%)
100
75
50
25
0
0
10
20
Days Post-Transplant
Family
Actinomycetaceae
Bacteroidaceae
Bifidobacteriaceae
Lachnospiraceae
Staphylococcaceae
Streptococcaceae
Veillonellaceae


Patient 57
Relative Cumulative Abundance (%)
100
75
50
25
0
0
20
40
Family
Bacteroidaceae
Carnobacteriaceae
Corynebacteriaceae
Lachnospiraceae
Streptococcaceae
Veillonellaceae
Patient 58
Relative Cumulative Abundance (%)
100
75
50
25
0
10
20
30
Days Post-Transplant
Family
Lactobacillaceae
Pasteurellaceae
Porphyromonadaceae
Prevotellaceae
Pseudomonadaceae
Streptococcaceae
Veillonellaceae


Patient 61
Relative Cumulative Abundance (%)
100
75
50
25
0
0
10
20
30
Days Post-Transplant
Family
Enterococcaceae
Lachnospiraceae
Pseudomonadaceae


Patient 62
Relative Cumulative Abundance (%)
100
75
50
25
0
25
30
35
Days Post-Transplant
Family
Moraxellaceae
Staphylococcaceae
Xanthomonadaceae


Patient 63
Relative Cumulative Abundance (%)
100
75
50
25
0
-10
0
10
20
30
40
50
Days Post-Transplant
Family
Bifidobacteriaceae
Carnobacteriaceae
Enterobacteriaceae
Enterococcaceae
Lachnospiraceae
Moraxellaceae
Ruminococcaceae
Staphylococcaceae
Streptococcaceae


Patient 64
Relative Cumulative Abundance (%)
100
75
50
25
0
7.5
10.0
12.5
15.0
17.5
Days Post-Transplant
Family
Pasteurellaceae
Staphylococcaceae
Streptococcaceae
Veillonellaceae

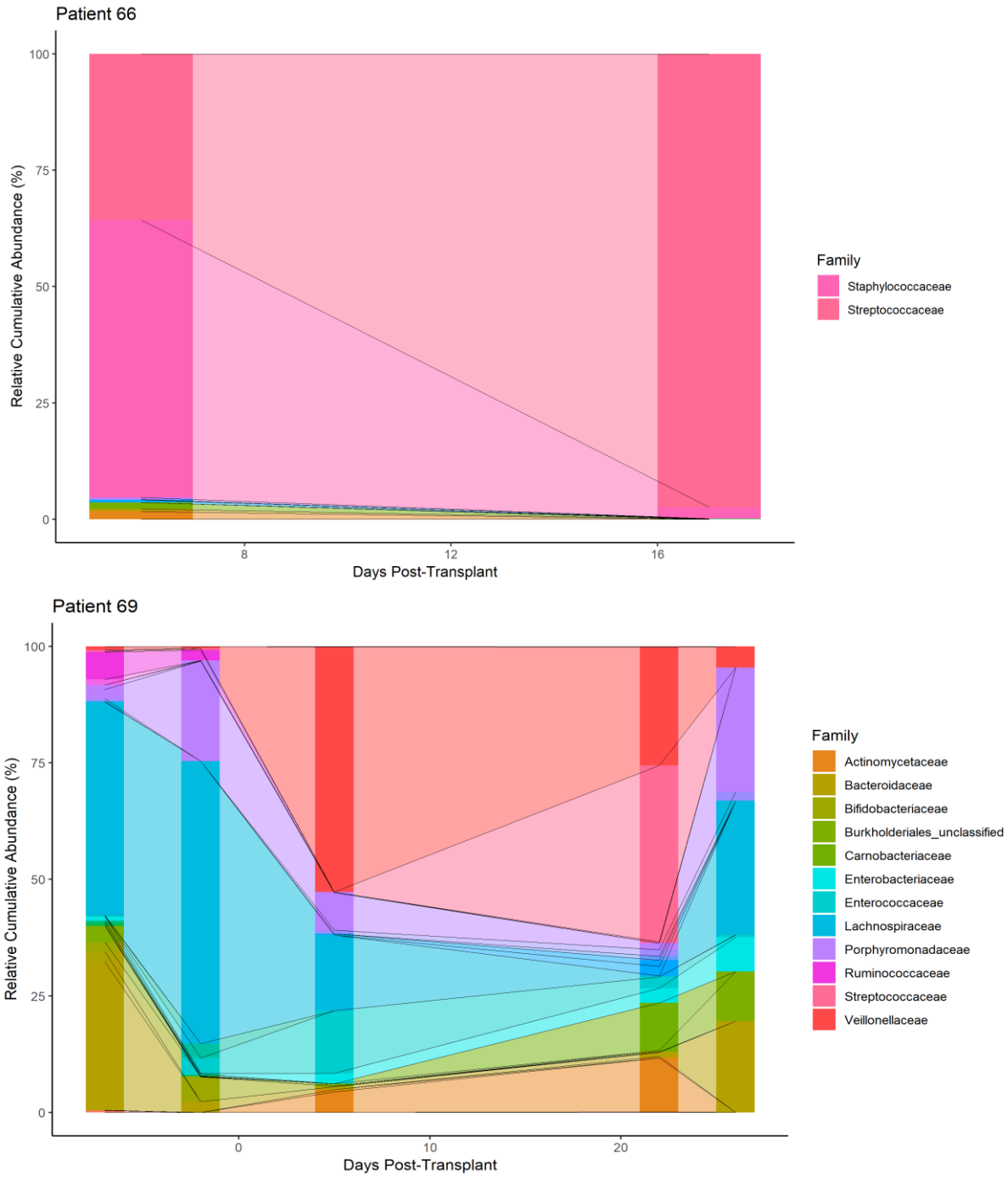
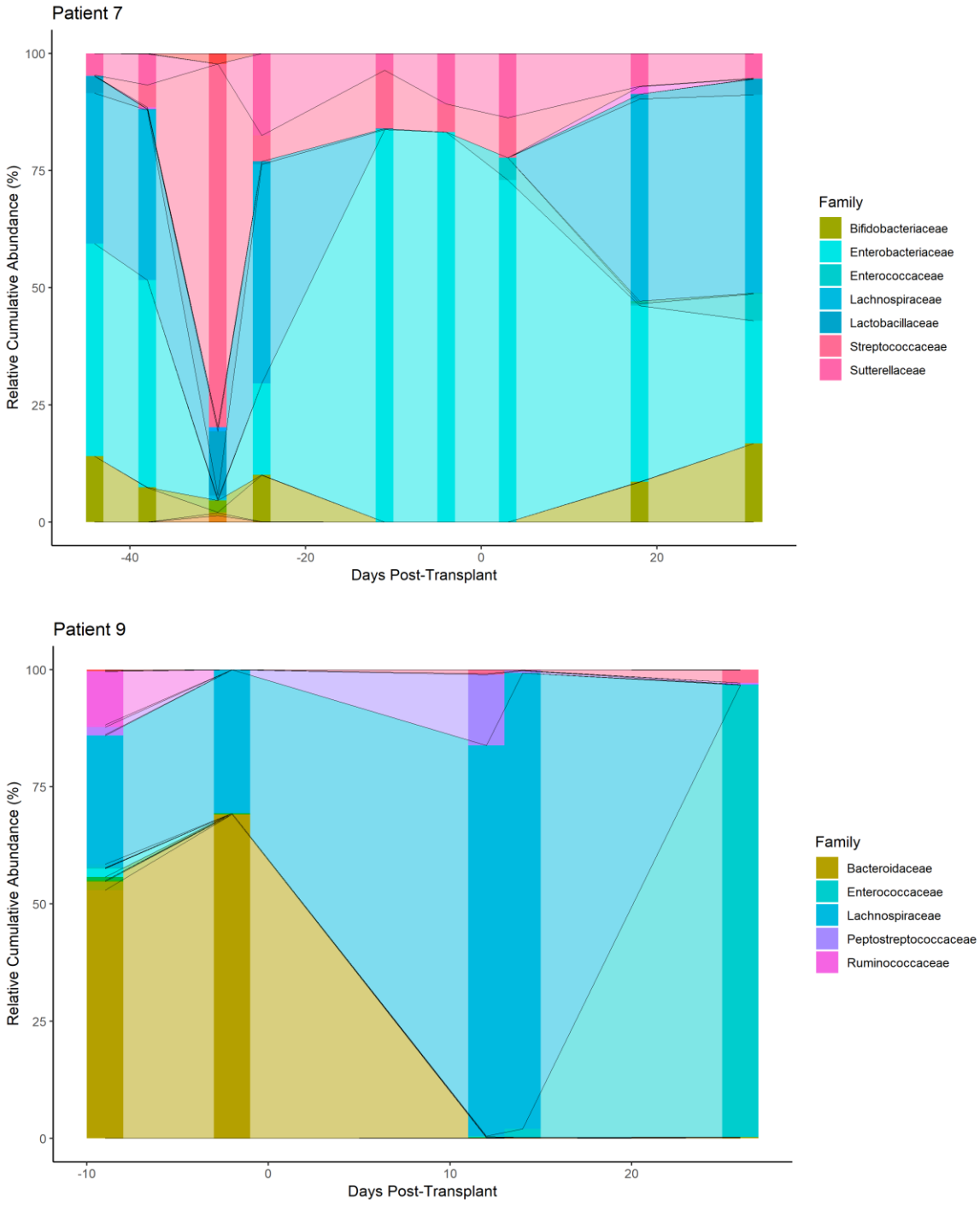
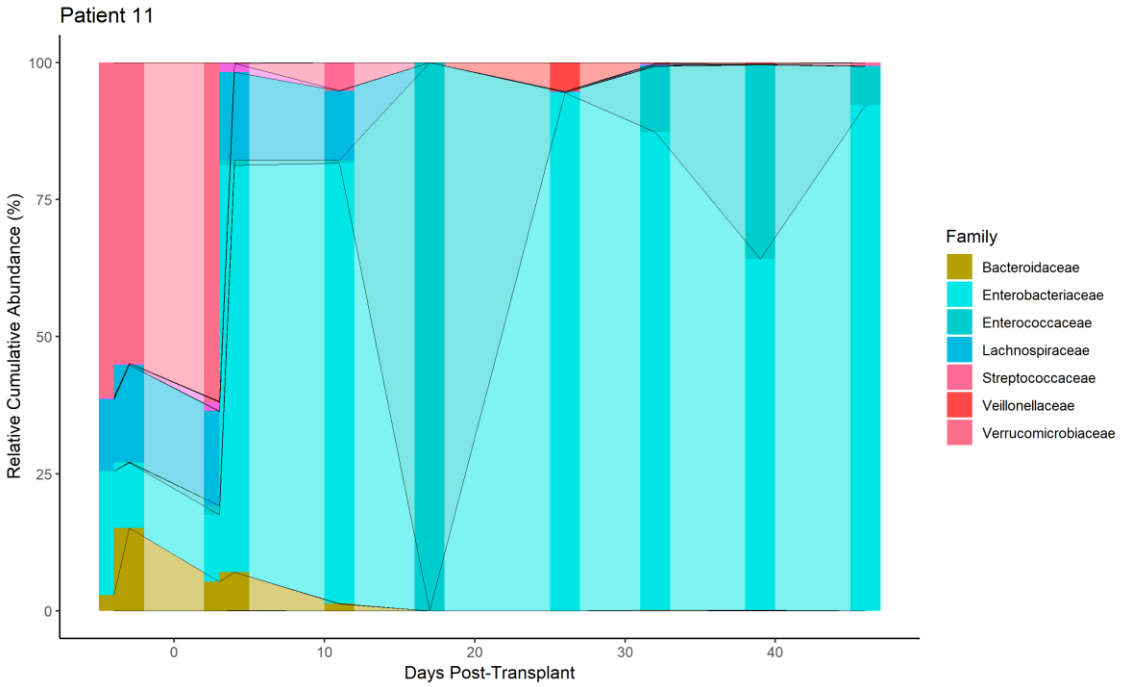
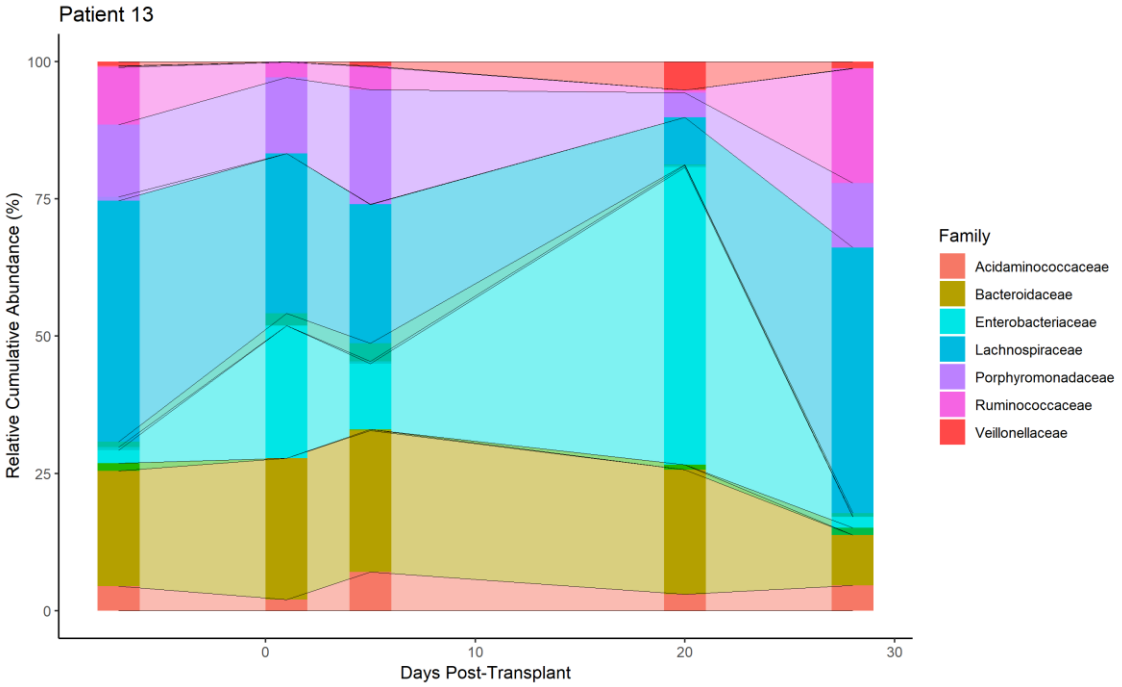
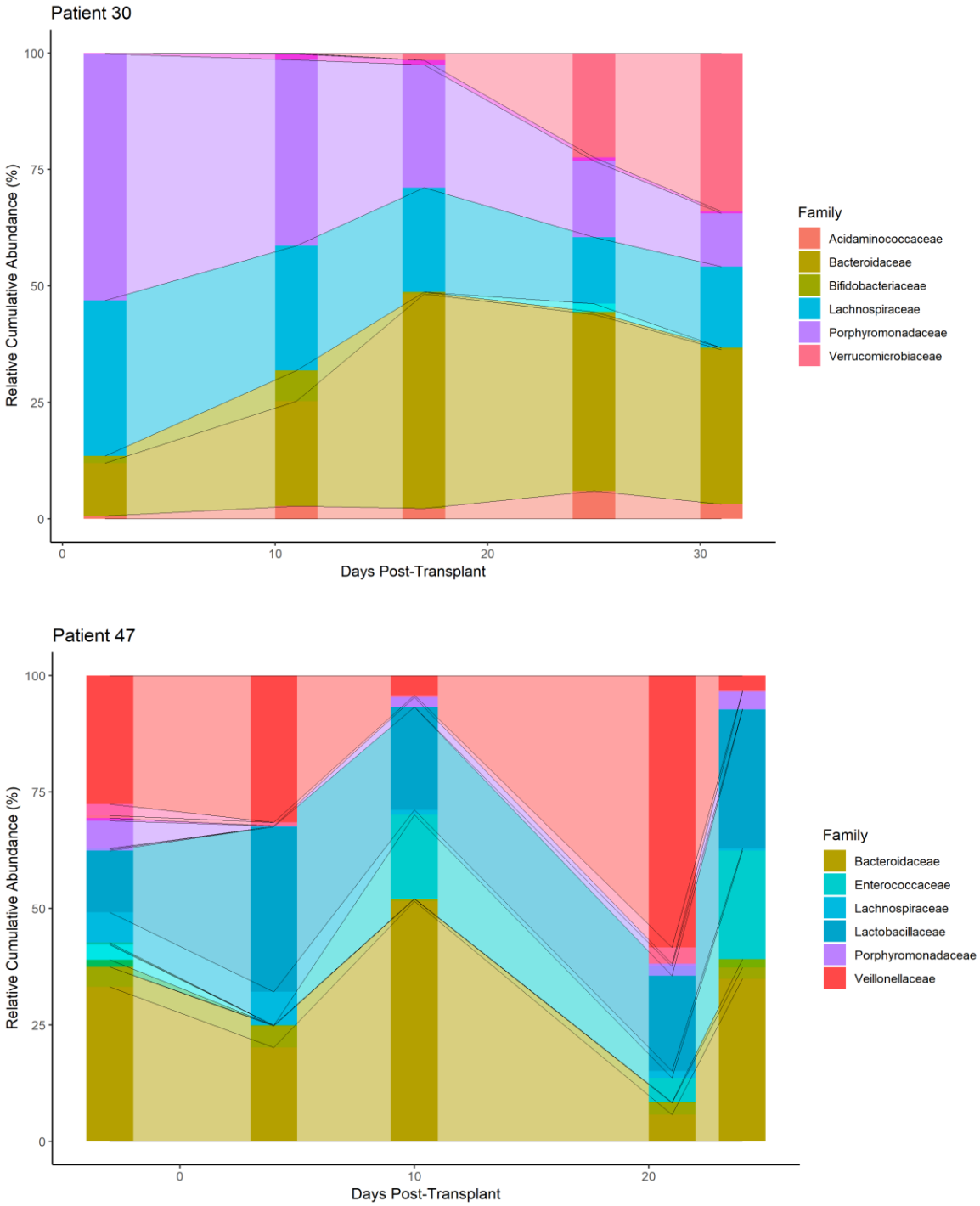


Figure A2 Taxonomic plots at family level for allogeneic HSCT recipients in the study cohort. Only families with an overall abundance of >10% are labelled for clarity. Darker vertical bars reflect sample collection points, whereas abundances between these are inferred.


Patient 7
Relative Cumulative Abundance (%)
100
75
50
25
0
-40
-20
0
20
Days Post-Transplant
Family
Bifidobacteriaceae
Enterobacteriaceae
Enterococcaceae
Lachnospiraceae
Lactobacillaceae
Streptococcaceae
Sutterellaceae
Patient 9
Relative Cumulative Abundance (%)
100
75
50
25
0
-10
0
10
20
Days Post-Transplant
Family
Bacteroidaceae
Enterococcaceae
Lachnospiraceae
Peptostreptococcaceae
Ruminococcaceae


Patient 11
Relative Cumulative Abundance (%)
100
75
50
25
0
0
10
20
30
40
Days Post-Transplant
Family
Bacteroidaceae
Enterobacteriaceae
Enterococcaceae
Lachnospiraceae
Streptococcaceae
Veillonellaceae
Verrucomicrobiaceae


Patient 13
Relative Cumulative Abundance (%)
100
75
50
25
0
0
10
20
30
Days Post-Transplant
Family
Acidaminococcaceae
Bacteroidaceae
Enterobacteriaceae
Lachnospiraceae
Porphyromonadaceae
Ruminococcaceae
Veillonellaceae


Patient 30
Relative Cumulative Abundance (%)
100
75
50
25
0
0
10
20
30
Days Post-Transplant
Family
Acidaminococcaceae
Bacteroidaceae
Bifidobacteriaceae
Lachnospiraceae
Porphyromonadaceae
Verrucomicrobiaceae
Patient 47
Relative Cumulative Abundance (%)
100
75
50
25
0
0
10
20
Days Post-Transplant
Family
Bacteroidaceae
Enterococcaceae
Lachnospiraceae
Lactobacillaceae
Porphyromonadaceae
Veillonellaceae

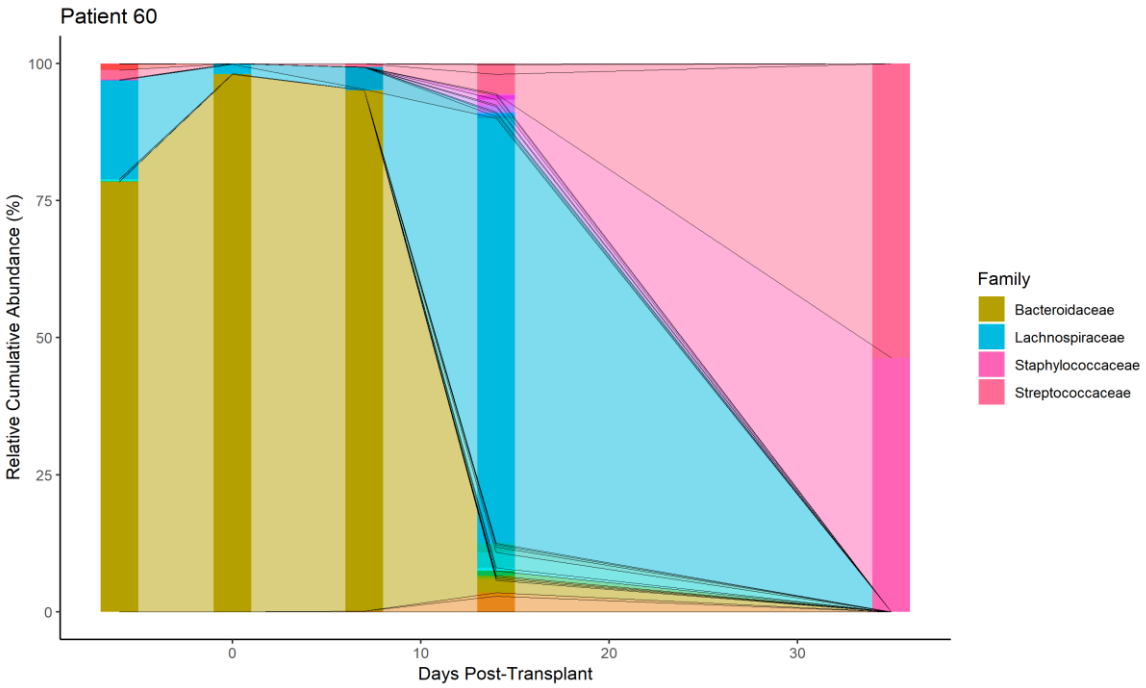
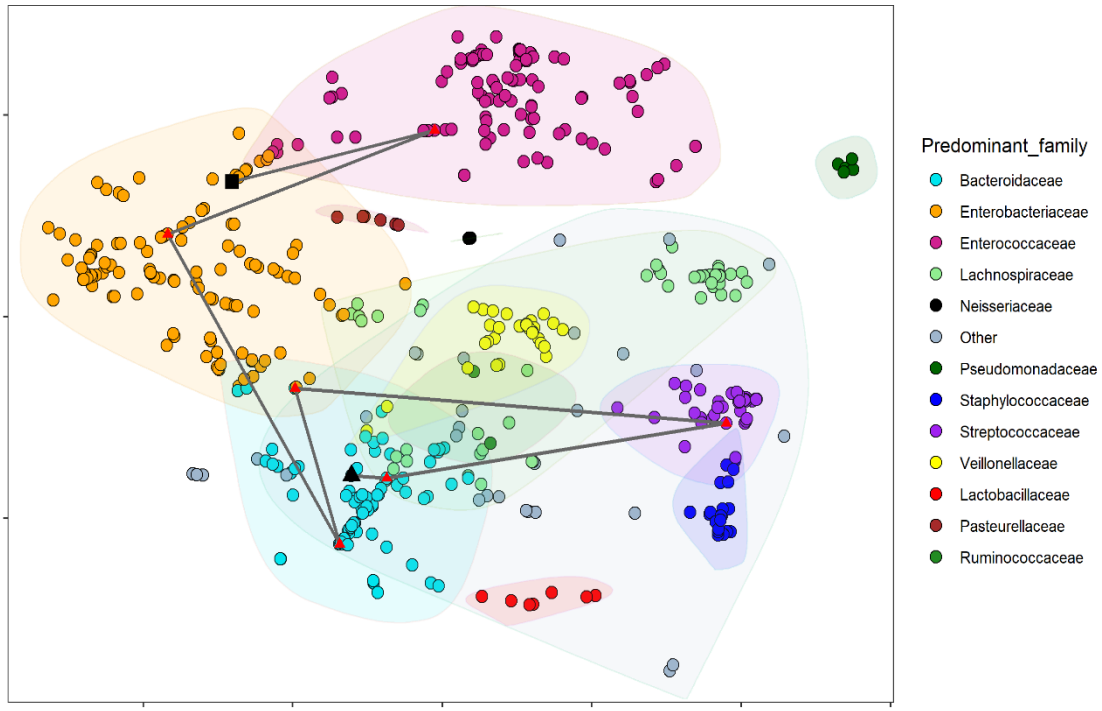
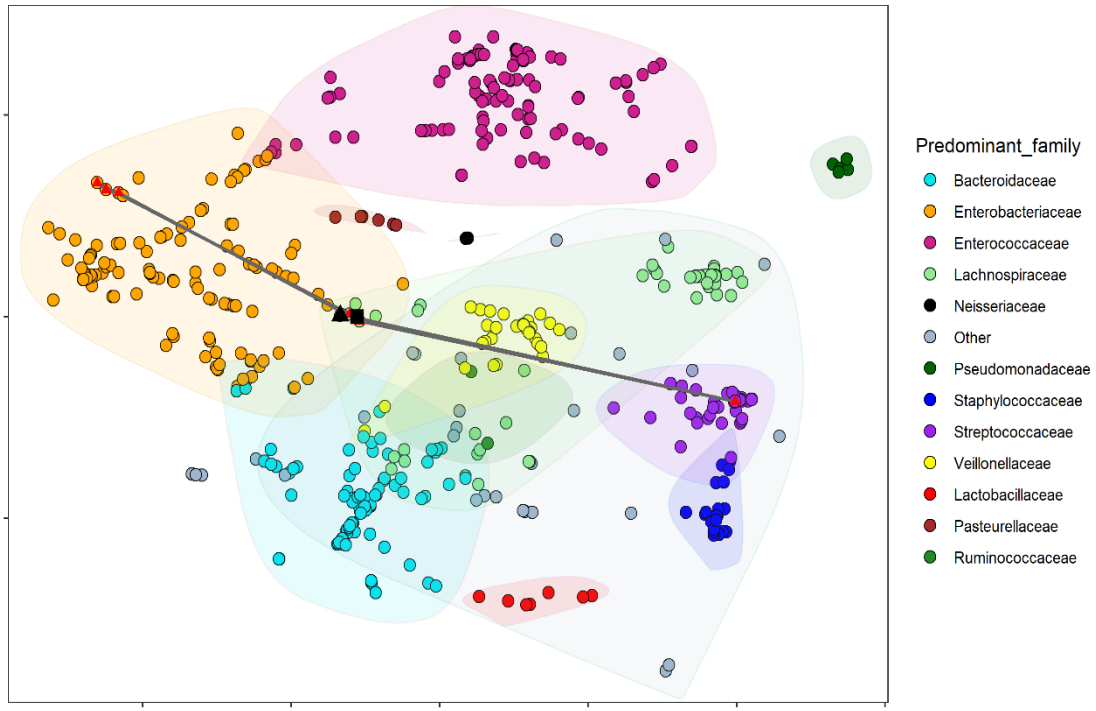


Figure A3 Taxonomic plots at family level for autologous HSCT recipients in the study cohort. Only families with an overall abundance of >10% are labelled for clarity. Darker vertical bars reflect sample collection points, whereas abundances between these are inferred.

Patient 1



Patient 7



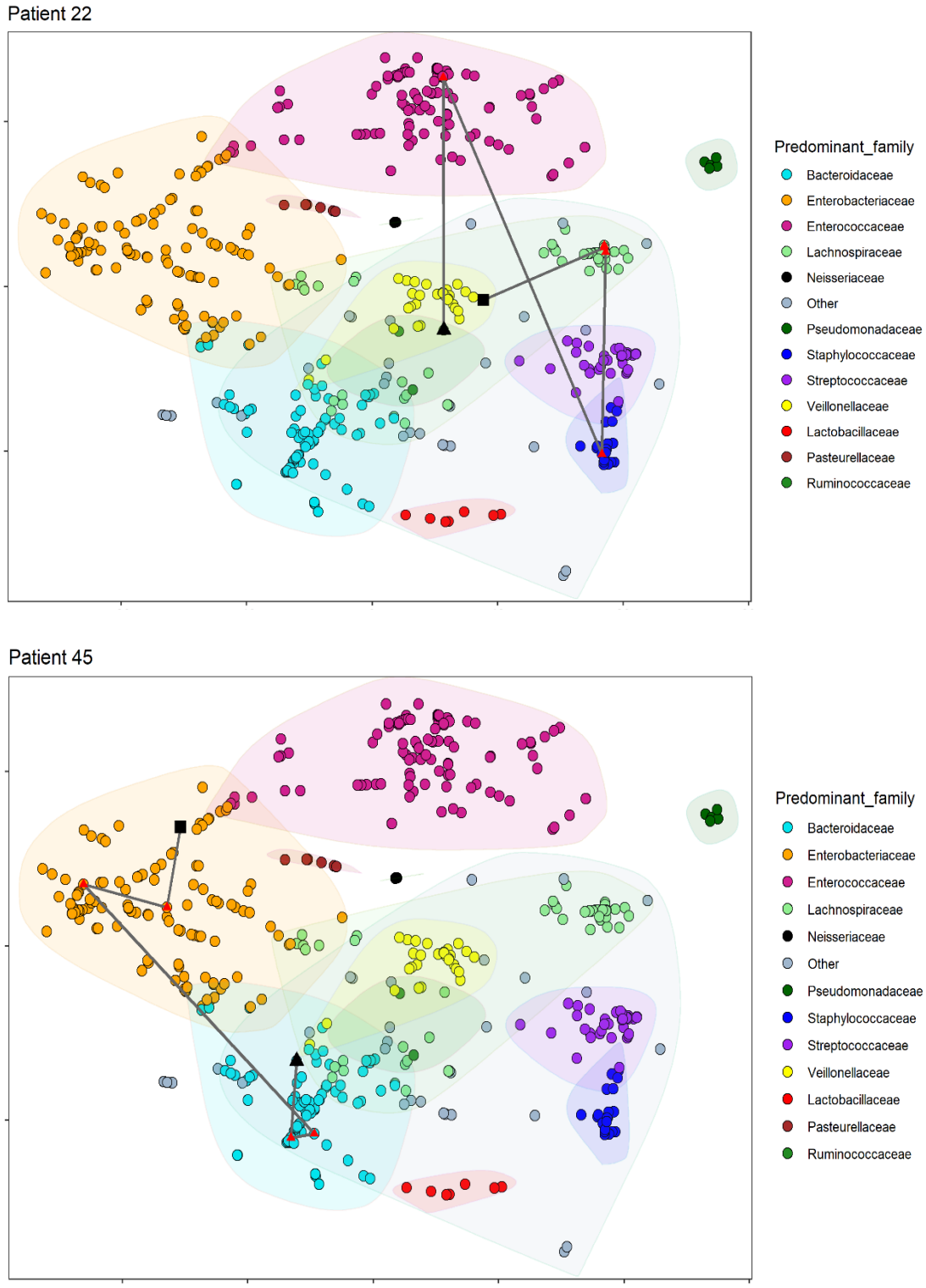


Figure A4 A t-SNE plot of all samples collected in the study with select individual trajectories. Each sample is coloured by its predominant taxa (>30%) A black triangle signifies the first collected sample and the black square signifying the last collected sample. Red triangles indicate all samples collected for this individual. Other is composed of several infrequently dominant taxa.

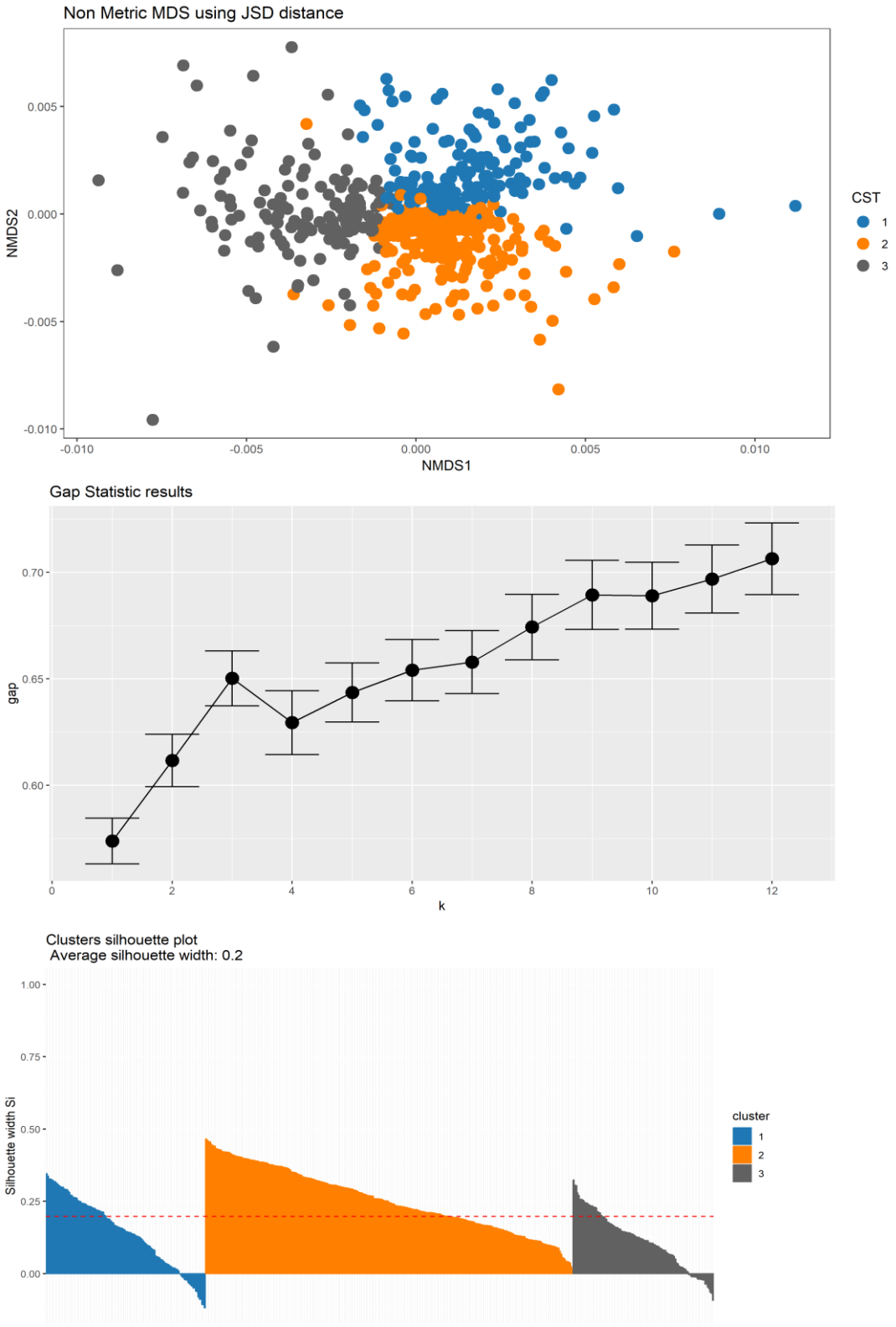


Figure A5 CST evaluation **A)** NMDS ordination of all samples using Jensen-Shannon divergence of the variance stabilized microbial count data **B)** Gap statistic evaluation of the variance stabilized microbial count data **C)** Cluster validation by silhouette assessment. The average width was 0.2, with the width of 0.14 for CST1, 0.25 for CST2 and 0.11 for CST3.

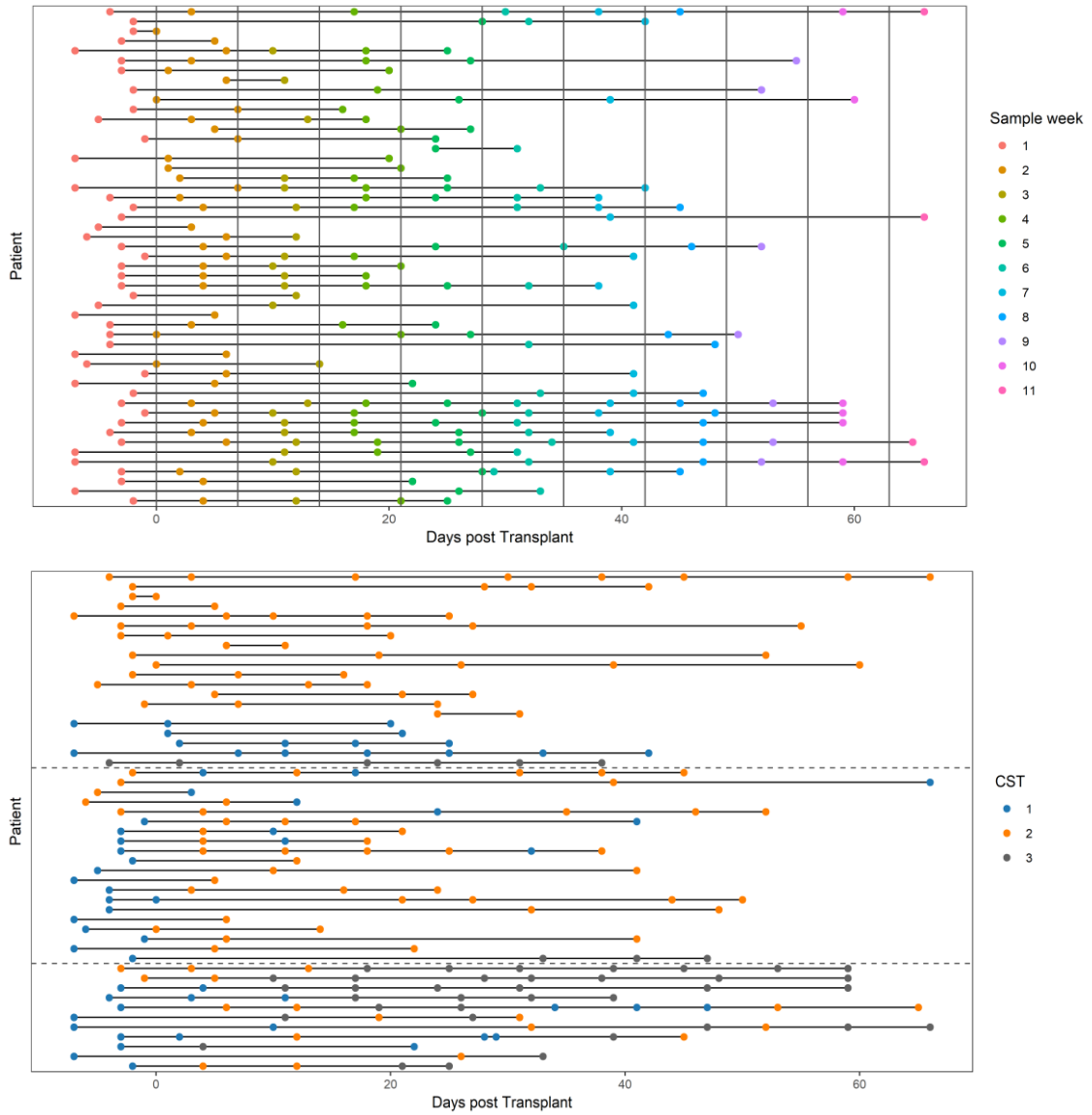


Figure A6 Timelines of the samples utilised in the time-dependent transition model. Samples are coloured by **A)** the sample week they fall into and **B)** the CST they have been classified into. Horizontal dashed lines detail CST patterns.

Table A12 Univariate and multivariable Cox models with GvHD as the dependent variable

Variable	HR	95% CI	P value
Sex_Female:Yes	0.9	0.47-1.72	0.75
Age_under_2:Yes	1.3	0.69-2.44	0.42
Diagnosis:Malignant	0.7	0.37-1.31	0.26
More than 1 transplant:Yes	1.27	0.55-2.95	0.58
Conditioning:Myeloablative	1.21	0.66-2.21	0.53
Serotherapy:Yes	0.82	0.28-2.42	0.71
Graft_manipulation:Yes	0.69	0.16-3.09	0.63
Shannon_effective	0.02	0.91-1.14	0.73
Microbiome_CST:2	0.9	0.28-2.84	0.85
Microbiome_CST:3	0.58	0.14-2.44	0.46

Abbreviations: CI, confidence interval; HR, hazard ratio

Table A13 Univariate Cox models with viraemia as the dependent variable

Univariate				Multivariable		
Variable	HR	95% CI	P value	HR	95% CI	P value
Sex_Female:Yes	0.95	0.58-1.53	0.82	-	-	-
Age_under_2:Yes	0.74	0.35-1.56	0.42	-	-	-
Diagnosis:Malignant	1.08	0.66-1.76	0.77	-	-	-
More than 1 transplant:Yes	1.59	1.01-2.52	0.05	1.44	0.93-2.23	0.10
Conditioning:Myeloablative	1.13	0.67-1.91	0.65	-	-	-
Shannon_effective	0.96	0.90-1.03	0.27	-	-	-
Stability:2	1.01	0.56-1.82	0.98	-	-	-
Stability:3	1.27	0.72-2.27	0.41	-	-	-
Microbiome_CST:2	1.32	0.73-2.36	0.36	1.28	0.71-2.32	0.41
Microbiome_CST:3	2.19	1.27-3.76	0.005	2.07	0.29-2.52	0.01

Abbreviations: CI, confidence interval; HR, hazard ratio; -, not significant.

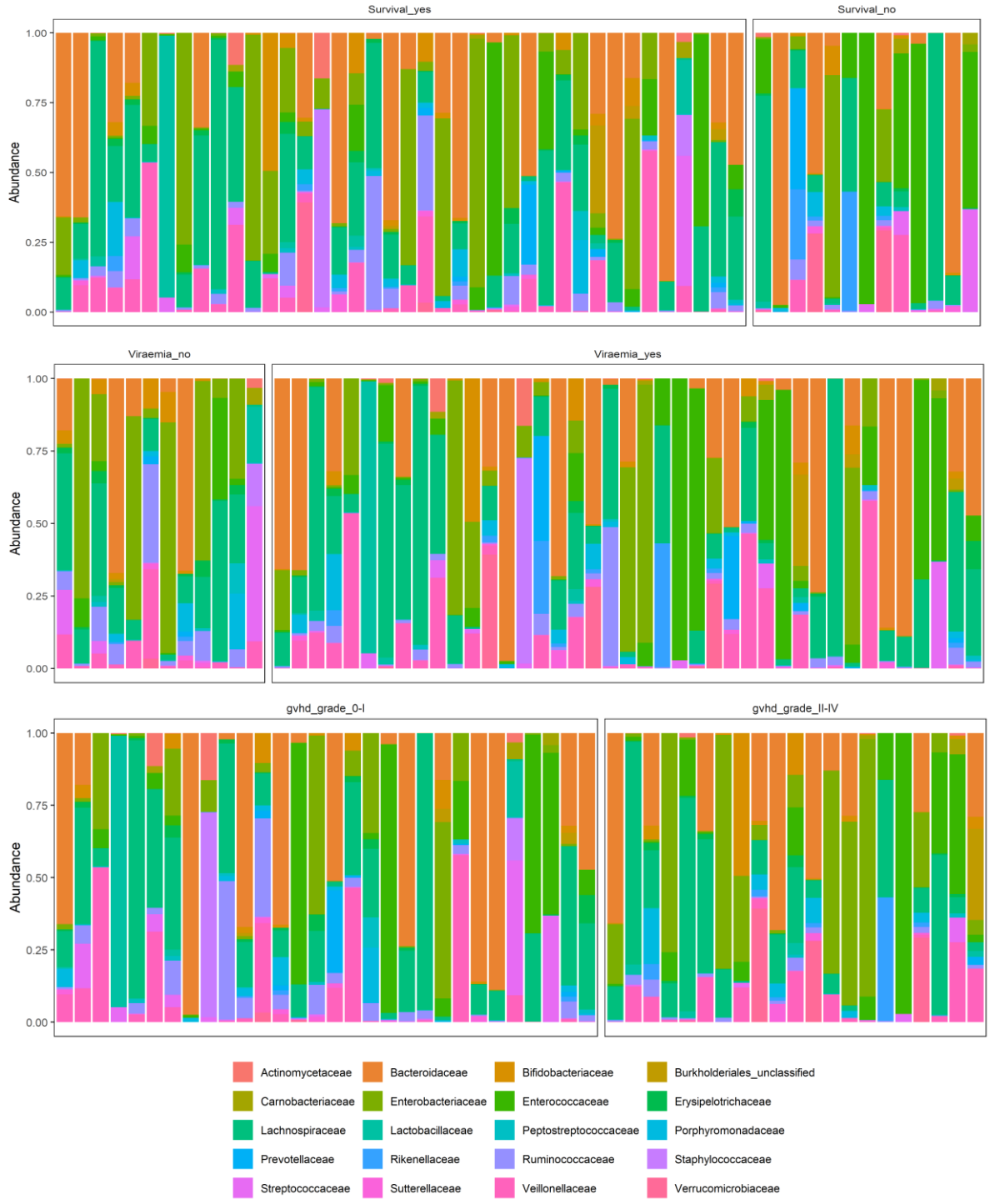


Figure A7 Family level taxonomic plots comparing baseline samples of patients with varying clinical outcomes

Table A14 Optimal cut-offs for significant taxa for the baseline sub-cohort

Taxa	AUC	Cut-off	Sensitivity	Specificity	P- value
Outcome: Survival					
Veillonella	0.72	0.0005	0.62	0.8	0.02
Veillonellaceae	0.72	0.01	0.87	0.58	0.02
Enterobacteriaceae	0.72	0.002	0.62	0.83	0.02
Peptostreptococcaceae	0.69	0.001	0.92	0.53	0.04
Shannon_effective	0.65	4.56	0.69	0.55	0.11
Outcome: Viraemia					
Clostridium_XVIII	0.77	0.0024	0.88	0.73	0.01
Faecalibacterium	0.71	0.0409	0.98	0.46	0.03
Dysgonomonas	0.68	0	1	0.36	0.07
Holdemania	0.62	0	0.27	0.62	0.22
Neisseria	0.64	0.0001	0.95	0.36	0.15
Robinsoniella	0.63	0	0.98	0.27	0.21
Turicibacter	0.63	0	0.9	0.36	0.19
Shannon_effective	0.67	4.74	0.65	0.73	0.09
Outcome: GvHD					
Klebsiella	0.67	0.000034	0.65	0.74	0.04
Shannon_effective	0.50	3.13	0.45	0.71	0.65

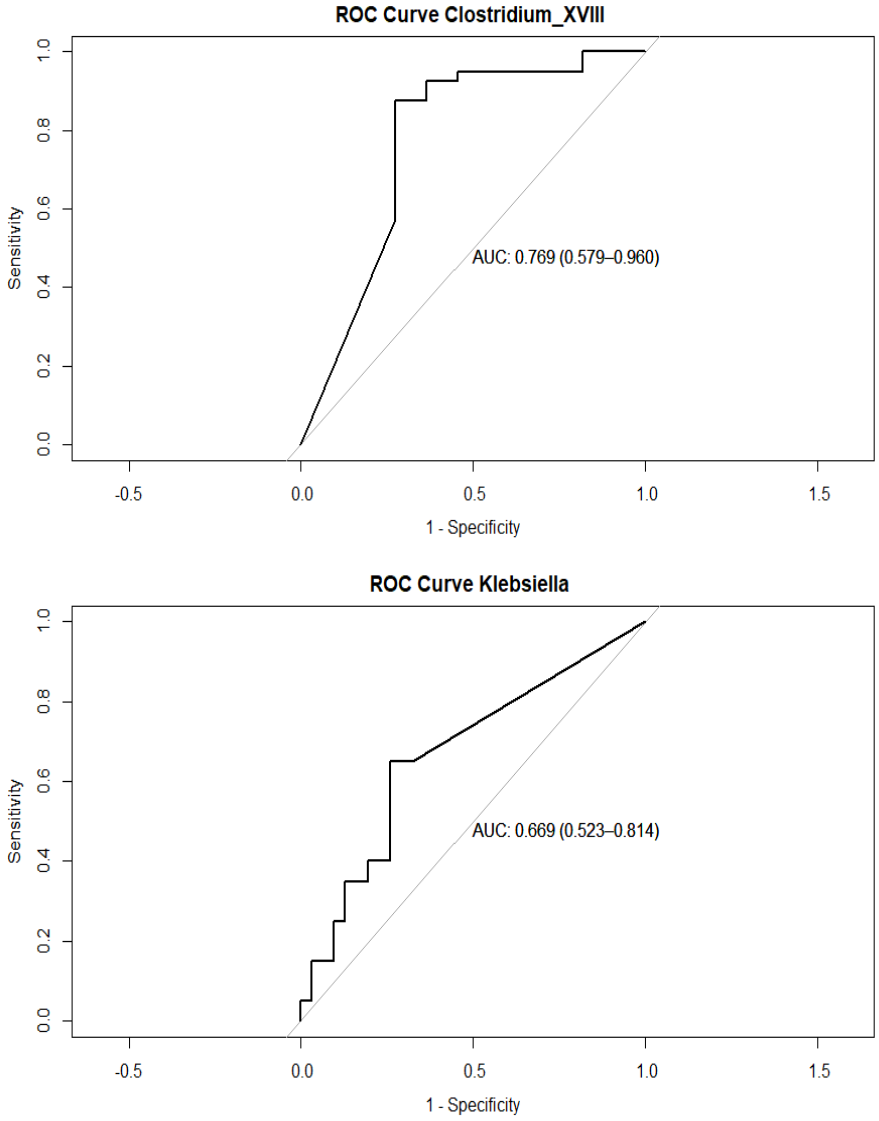


Figure A8 *Clostridium*_XVIII (viraemia) and *Klebsiella* (GvHD) ROC curves at baseline (95% CI)

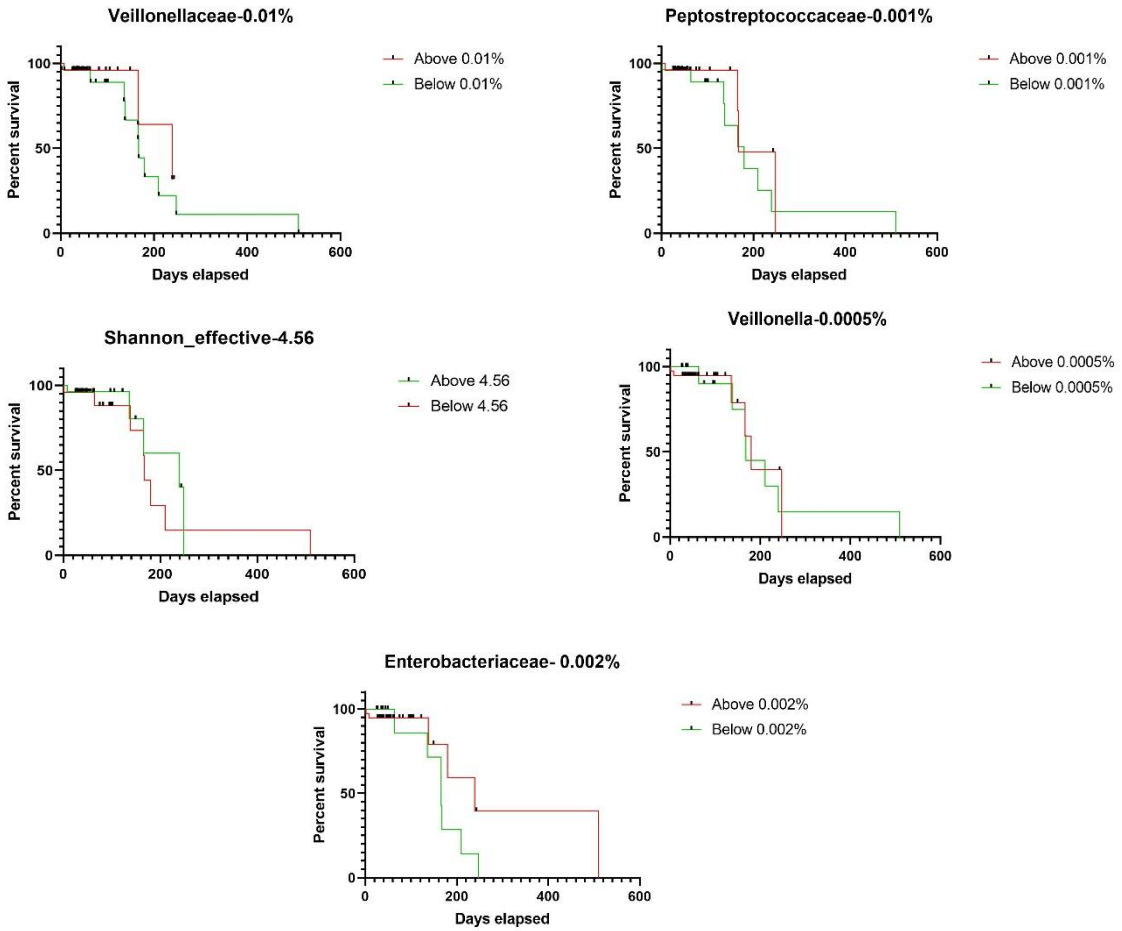


Figure A9 KM survival curves for differentially abundant taxa at baseline for overall survival

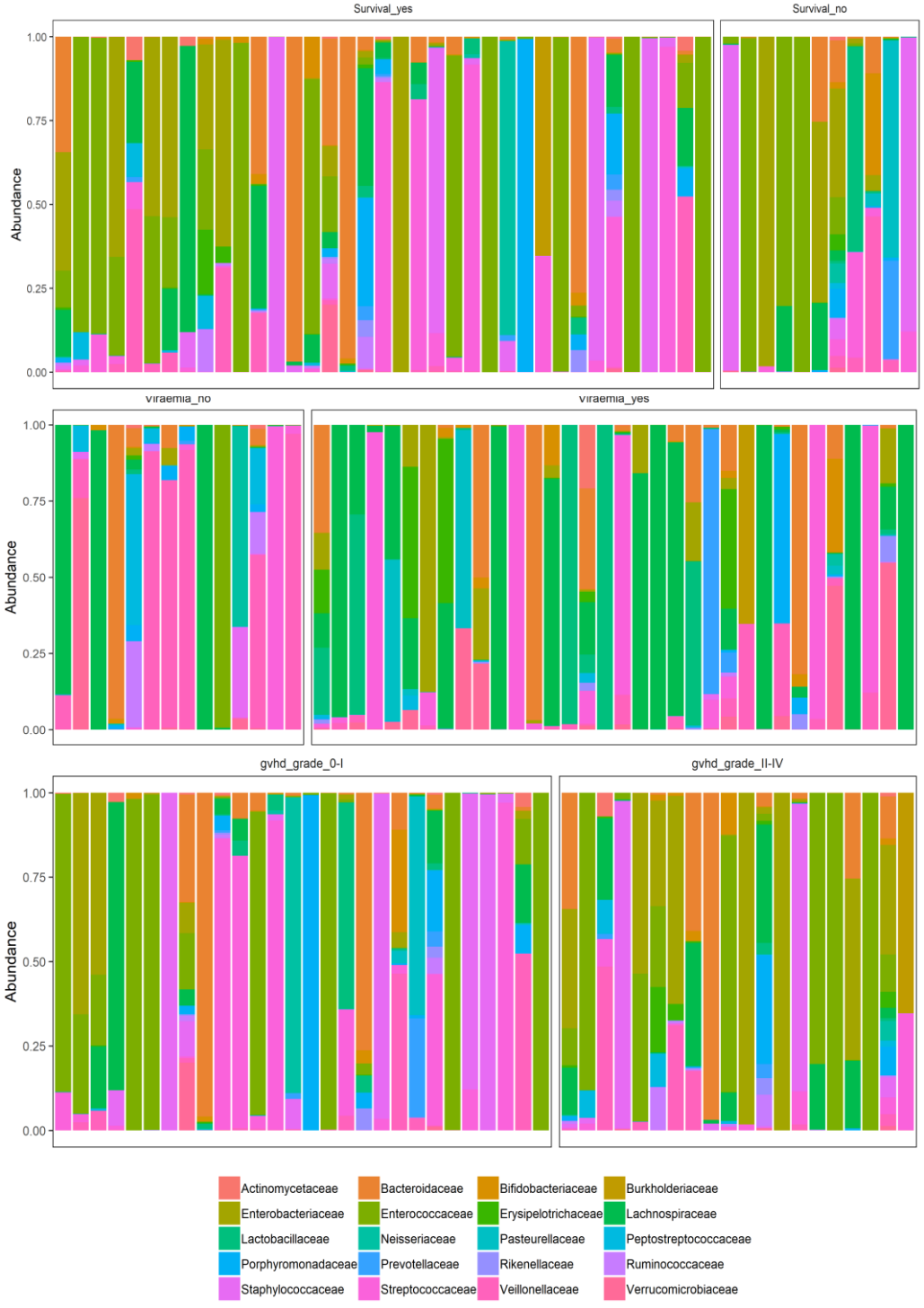


Figure A10 Genus level taxonomic plots comparing pre-engraftment samples of patients with varying clinical outcomes

Table A15 Differentially abundant taxa for GvHD and viraemia in pre-engraftment samples

Feature	Enriched in	LDA effect size	P-value
Outcome: Viraemia			
Mucispirillum	No	2.96	0.01
Odoribacter	No	3.14	0.01
Allobaculum	No	2.79	0.01
Prevotellaceae_unclassified	No	3.44	0.01
Bacteroidales_unclassified	No	2.87	0.01
Akkermansia	Yes	3.34	0.01
Clostridiales_unclassified	No	3.38	0.02
Barnesiella	No	3.69	0.02
Oribacterium	No	3.61	0.03
Bacteroidetes_unclassified	No	3.48	0.03
Oscillibacter	No	2.85	0.03
Firmicutes_unclassified	No	2.80	0.03
Ruminococcus	No	2.59	0.04
Clostridia_unclassified	No	2.91	0.04
Parasutterella	No	2.65	0.04
Deferribacteraceae	No	4.79	0.01
Clostridiales_unclassified	No	3.68	0.02
Bacteroidetes_unclassified	No	3.90	0.03
Firmicutes_unclassified	No	3.64	0.03
Clostridia_unclassified	No	4.24	0.04
Flavobacteriaceae	No	3.00	0.02
Bacteroidales_unclassified	No	3.69	0.01
Enterobacteriaceae	Yes	4.92	0.01
Outcome: GvHD			
Clostridium_XI	Yes	3.79	0.03
Flavonifractor	Yes	3.66	0.04
Abiotrophia	No	3.60	0.04
Peptostreptococcaceae	Yes	3.88	0.01
Aerococcaceae	No	3.80	0.04

Table A16 Optimal cut-offs for significant taxa for pre-engraftment sub-cohort

Taxa	AUC	Cut-off	Sensitivity	Specificity	P value
Outcome: Viraemia					
Enterobacteriaceae	0.74	0.0015	0.62	0.86	0.01
Flavobacteriaceae	0.64	0	0.91	0.36	0.14
Firmicutes_unclassified	0.61	0	0.94	0.29	0.22
Clostridiales_unclassified	0.62	0.0001	1	0.29	0.19
Bacteroidales_unclassified	0.63	0	1	0.29	0.15
Bacteroidetes_unclassified	0.60	0	1	0.21	0.30
Clostridia_unclassified	0.59	0	0.97	0.21	0.32
Defferibacteraceae	0.61	0	1	0.21	0.25
Akkermansia	0.64	0	0.94	0.36	0.12
Ruminococcus	0.59	0	0.97	0.21	0.43
Parasuterella	0.59	0	0.97	0.21	0.32
Allobaculum	0.61	0	1	0.21	0.25
Oscillibacter	0.60	7.00E-04	1	0.21	0.30
Mucispirillum	0.61	0	1	0.21	0.25
Odoribacter	0.61	0	1	0.21	0.25
Prevotellaceae_unclassified	0.66	0	0.97	0.36	0.09
Oribacterium	0.60	2.00E-04	1	0.21	0.30
Barnesiella	0.64	1.00E-04	0.94	0.36	0.14
Shannon_effective	0.55	1.57	0.68	0.57	0.57
Outcome: GvHD					
Clostridium_XI	0.62	0.00025	0.30	0.96	0.16
Flavonifractor	0.64	0.00012	0.40	0.89	0.09
Abiotropia	0.60	0.000038	0.25	0.96	0.23
Aerococcaceae	0.60	0.0136	0.55	0.57	0.23
Peptostreptococcaceae	0.65	0.00025	0.35	0.96	0.08
Shannon_effective	0.60	1.89	0.70	0.54	0.23

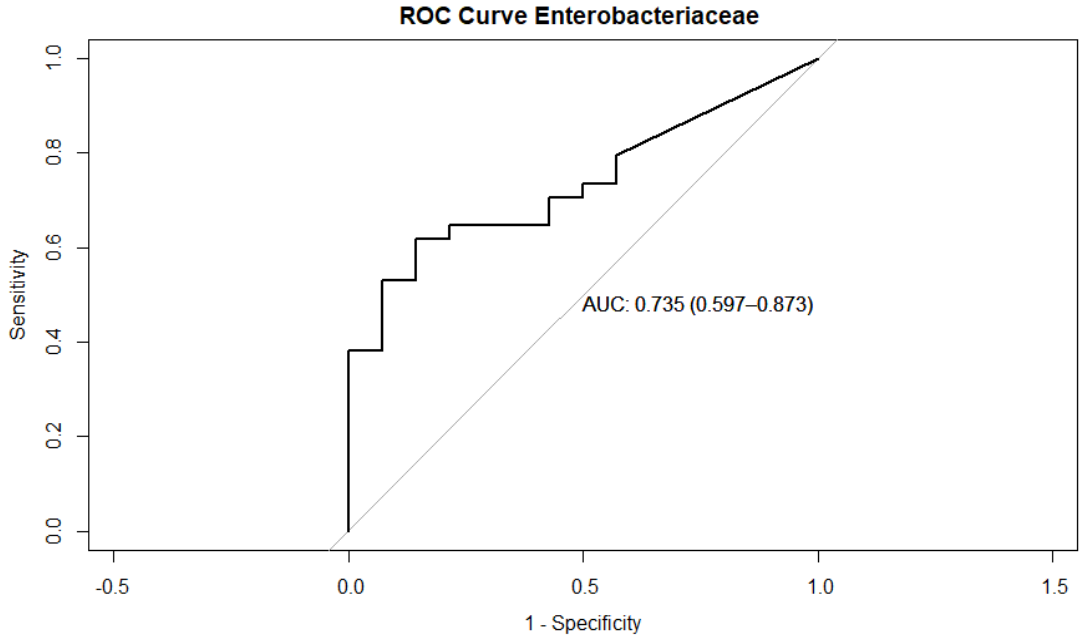
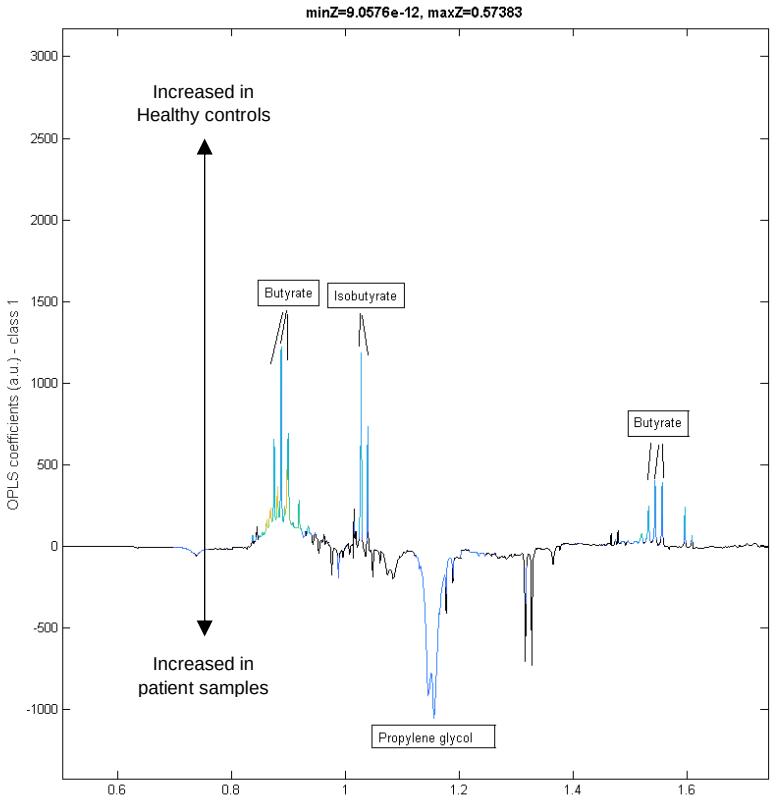
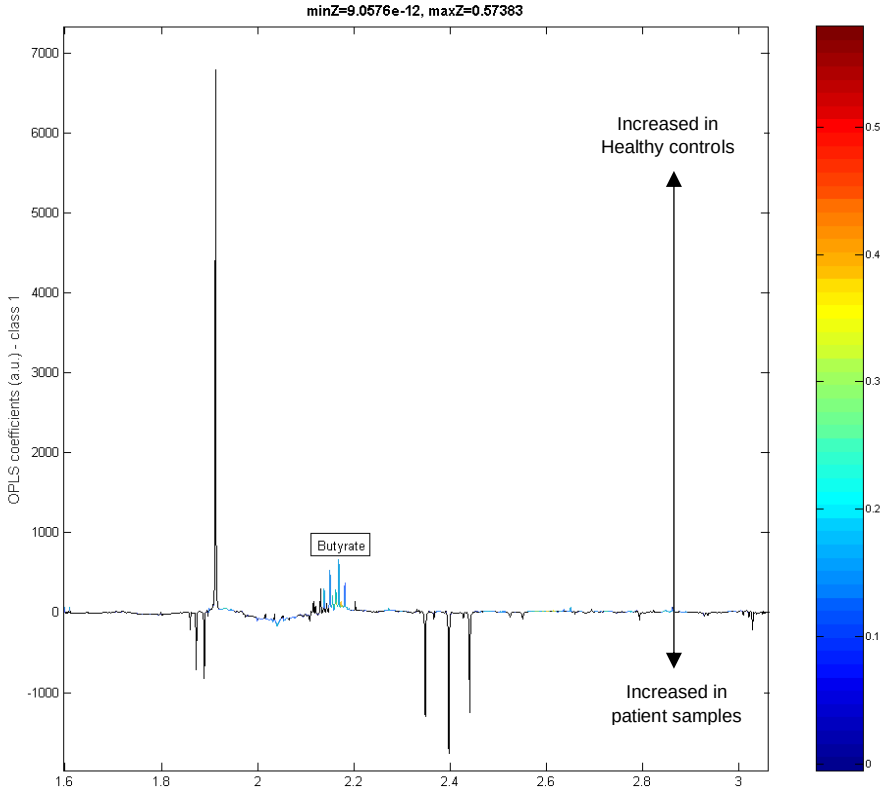


Figure A11 *Enterobacteriaceae* (viraemia) ROC curve at pre-engraftment (95% CI)


minZ=9.0576e-12, maxZ=0.57383
OPLS coefficients (a.u.) - class 1
Increased in
Healthy controls
Increased in
patient samples
Butyrate
Isobutyrate
Butyrate
Propylene glycol
3000
2500
2000
1500
1000
500
0
-500
-1000
0.6
0.8
1
1.2
1.4
1.6


minZ=9.0576e-12, maxZ=0.57383
OPLS coefficients (a.u.) - class 1
Increased in
Healthy controls
Increased in
patient samples
Butyrate
7000
6000
5000
4000
3000
2000
1000
0
-1000
1.6
1.8
2
2.2
2.4
2.6
2.8
3
0.5
0.4
0.3
0.2
0.1
0

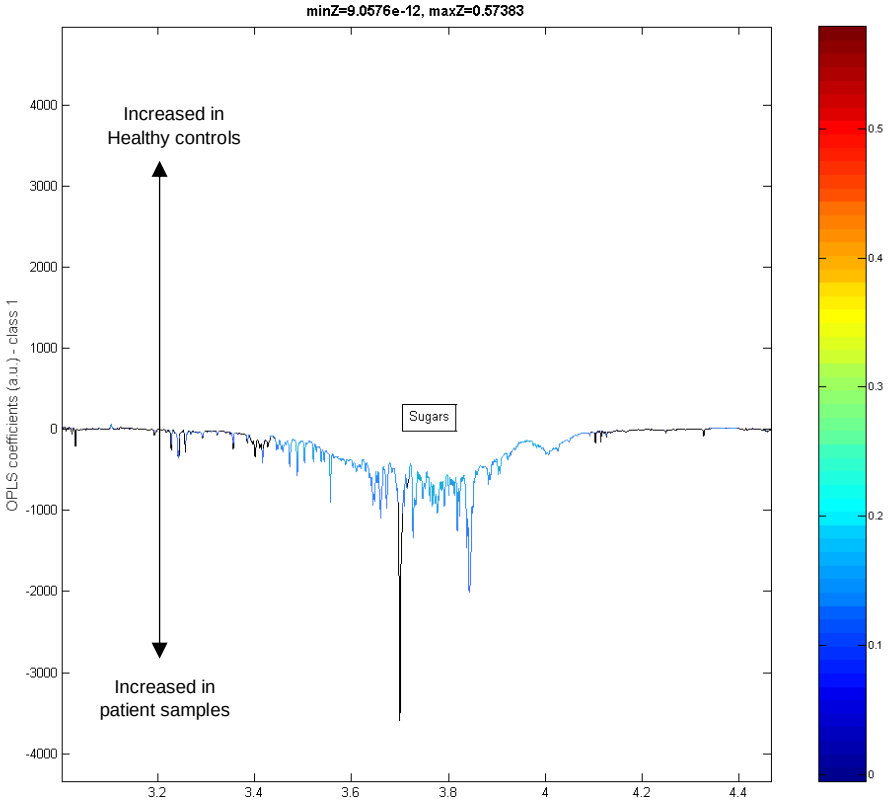
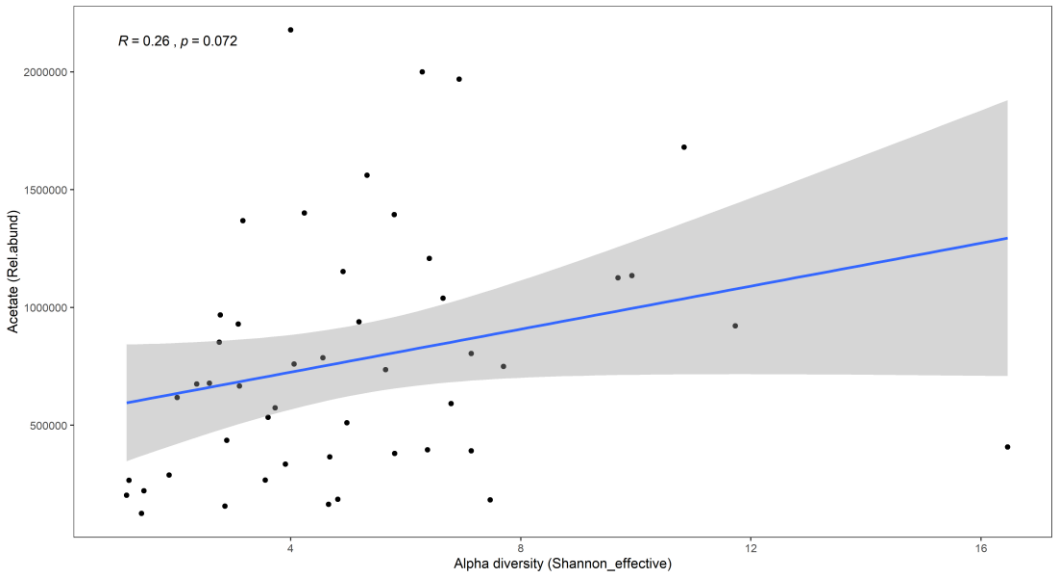
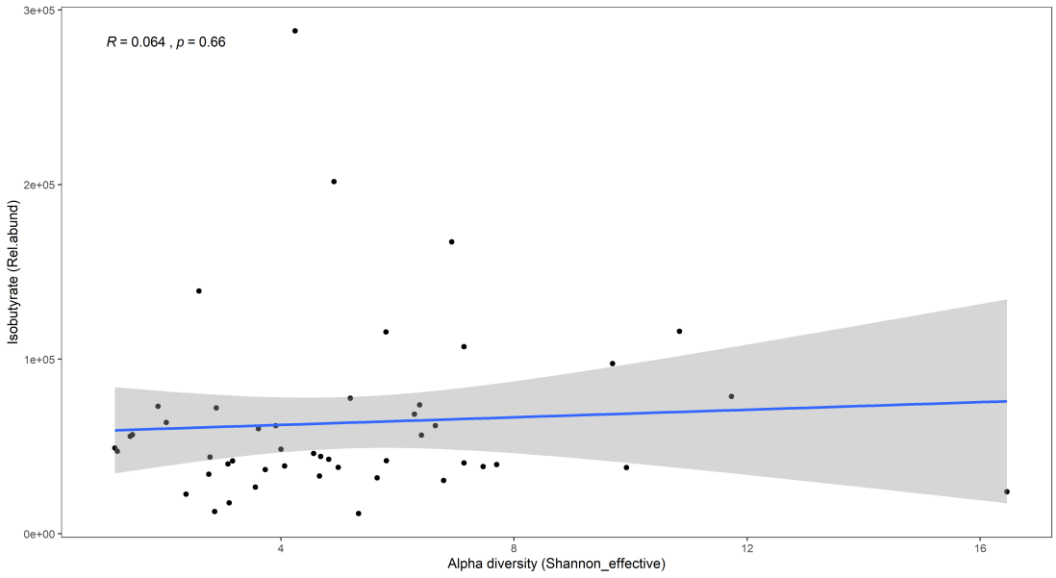


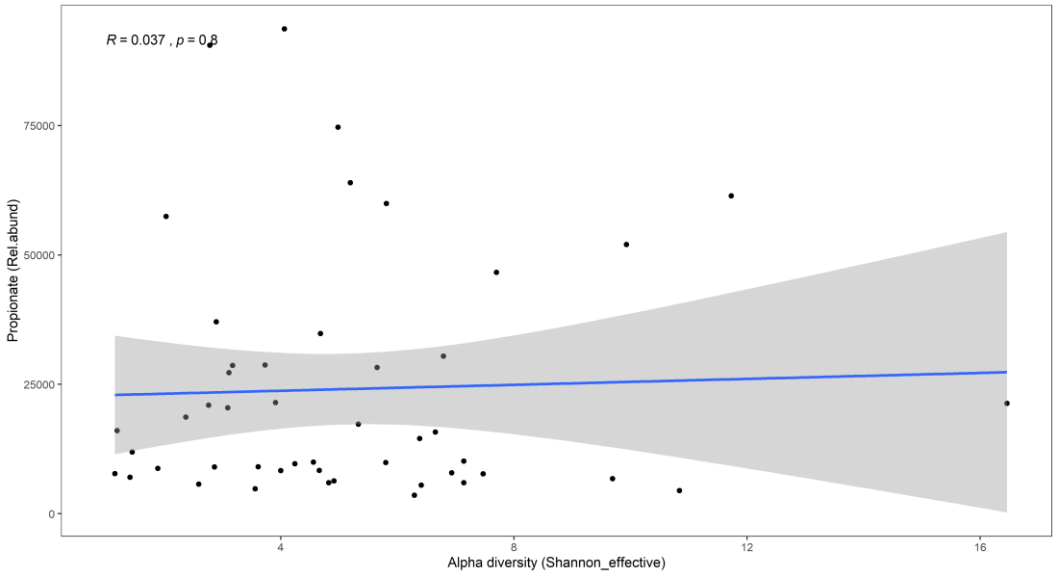
Figure A12 Spectra comparing healthy controls to baseline patient samples. Coloured peaks highlight those enriched in either of the groups.

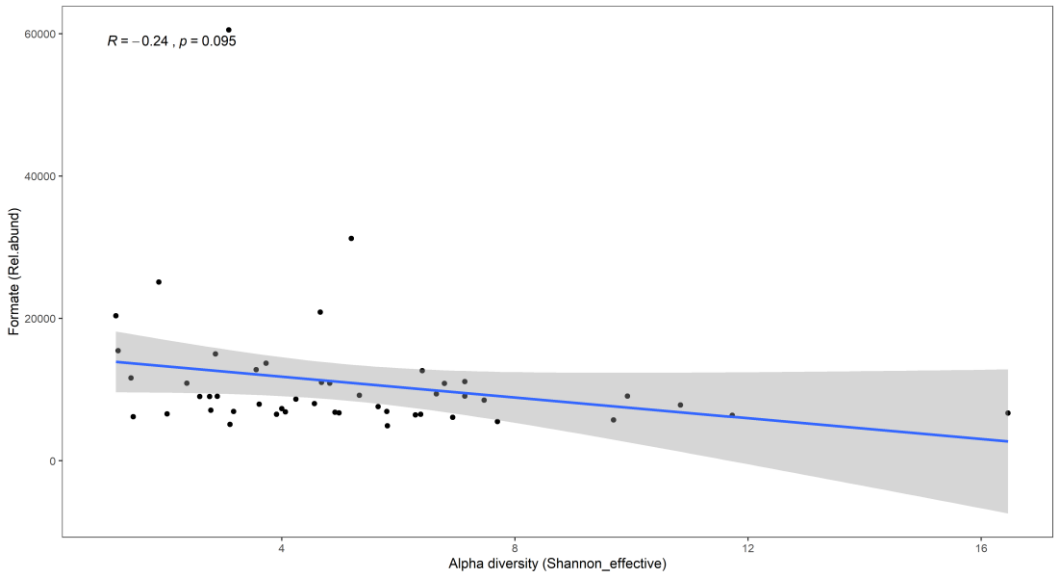
Table A17 Models run in this cohort

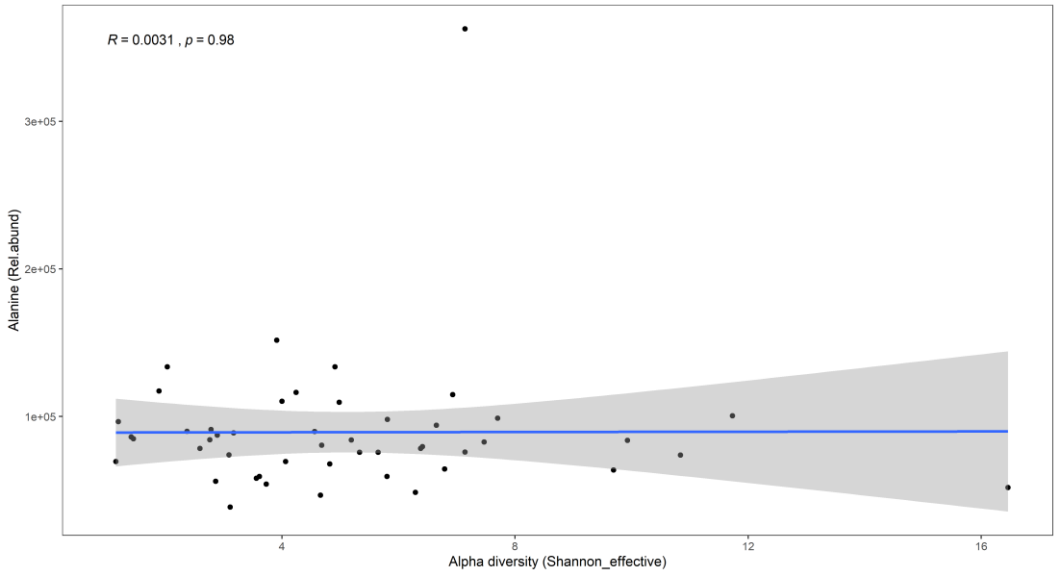
	Timepoint	Explains	N	Breakdown	X2	Q2	p-value
Clinical models							
Healthy	Baseline	Healthy controls vs baseline samples (with autologous)	62	9 vs 53	0.8193	0.0226	0.01
Age	Baseline	Age under 2 vs age over 2 at baseline-HSCT only	47	-	-	-	0.18
Age	Baseline	Age at baseline- continuous-HSCT only	47	-	-	-	0.92
Sex	Baseline	Female vs male at baseline-HSCT only	47	-	-	-	0.26
Diagnosis	Baseline	Malignant vs non-malignant diagnosis at baseline-HSCT only	47	-	-	-	0.71
Allogeneic	Baseline	Allogeneic vs autologous at baseline	47	-	-	-	0.92
Treatment effects							
Transplant effect	Pre-engraftment	Transplant effect (1 timepoint pre vs 1 timepoint post)-HSCT only	68	-	-	-	0.07
Transplant effect	Pre-engraftment	Transplant effect (1 timepoint pre vs 1 timepoint post)-HSCT only	68	-	-	-	0.36
Transplant effect	Pre-engraftment	Transplant effect (1 timepoint pre vs last individual timepoint post-last)-HSCT only	68	34 vs 34	0.7535	0.0209	0.02
Biomarker models							
Viraemia	Baseline	Viraemia vs non-viraemia in pre-samples- HSCT only	47	-	-	-	0.74
GvHD	Baseline	GvHD vs non-GvHD in pre-samples- HSCT only	45	-	-	-	0.36
Overall survival	Baseline	Survival vs non-survival in pre-samples- HSCT only	47	-	-	-	0.18
Viraemia	Pre-engraftment	Viraemia vs non-bacteraemia in pre-engraftment samples- HSCT only	38	-	-	-	0.94
GvHD	Pre-engraftment	GvHD vs non-GvHD in pre-engraftment samples- HSCT only	37	-	-	-	0.74
Overall survival	Pre-engraftment	Survival vs non-survival in pre-engraftment samples- HSCT only	38	-	-	-	0.77

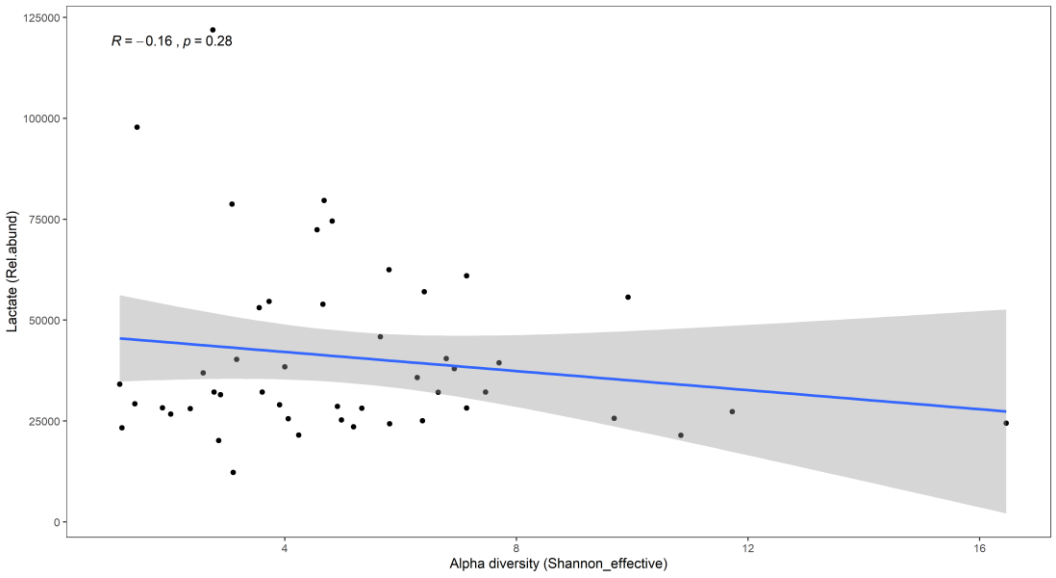

R = 0.26 , p = 0.072
Acetate (Rel.abund)
2000000
1500000
1000000
500000
4
8
12
16
Alpha diversity (Shannon_effective)

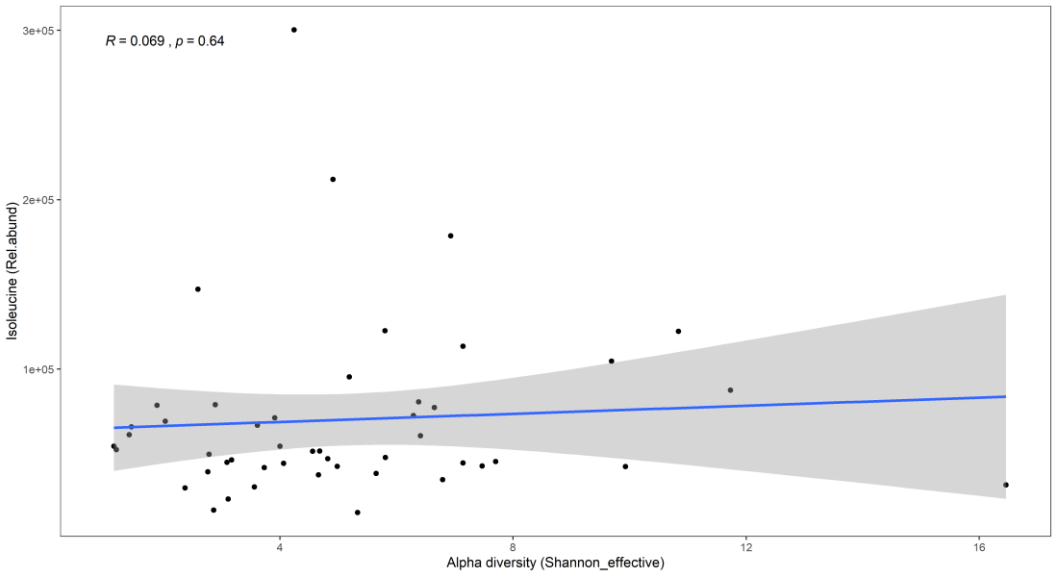

R = 0.064 , p = 0.66
Isobutyrate (Rel.abund)
3e+05
2e+05
1e+05
0e+00
4
8
12
16
Alpha diversity (Shannon_effective)

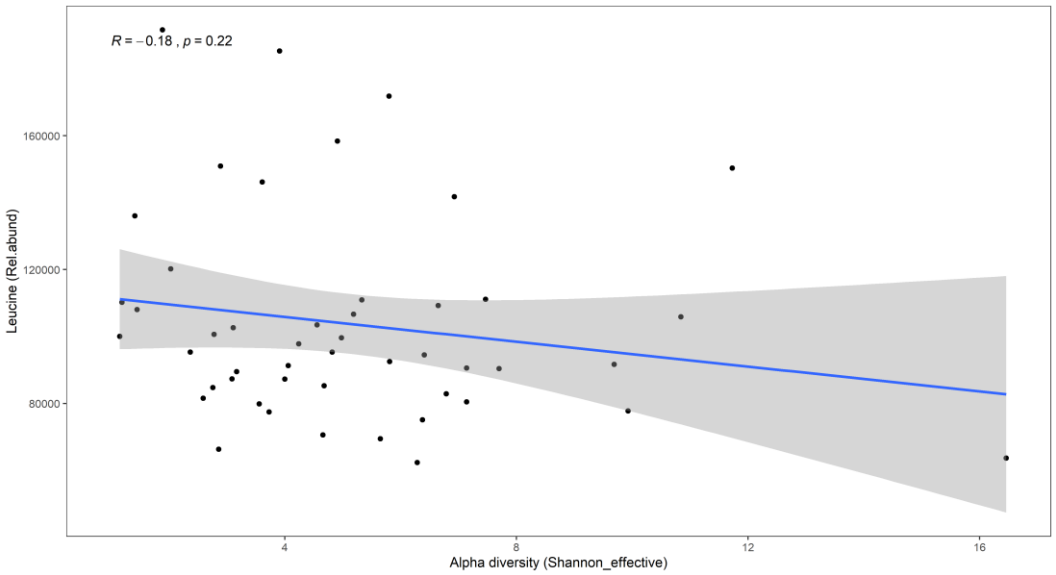

R = 0.037 , p = 0.8
Propionate (Rel.abund)
75000
50000
25000
0
4
8
12
16
Alpha diversity (Shannon_effective)

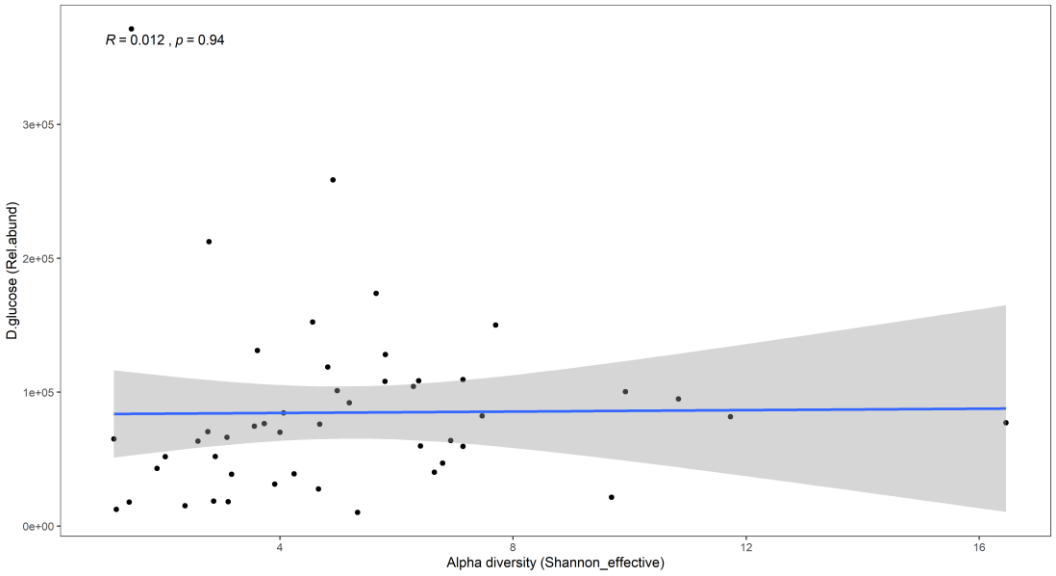

R = −0.24 , p = 0.095
Formate (Rel.abund)
60000
40000
20000
0
4
8
12
16
Alpha diversity (Shannon_effective)

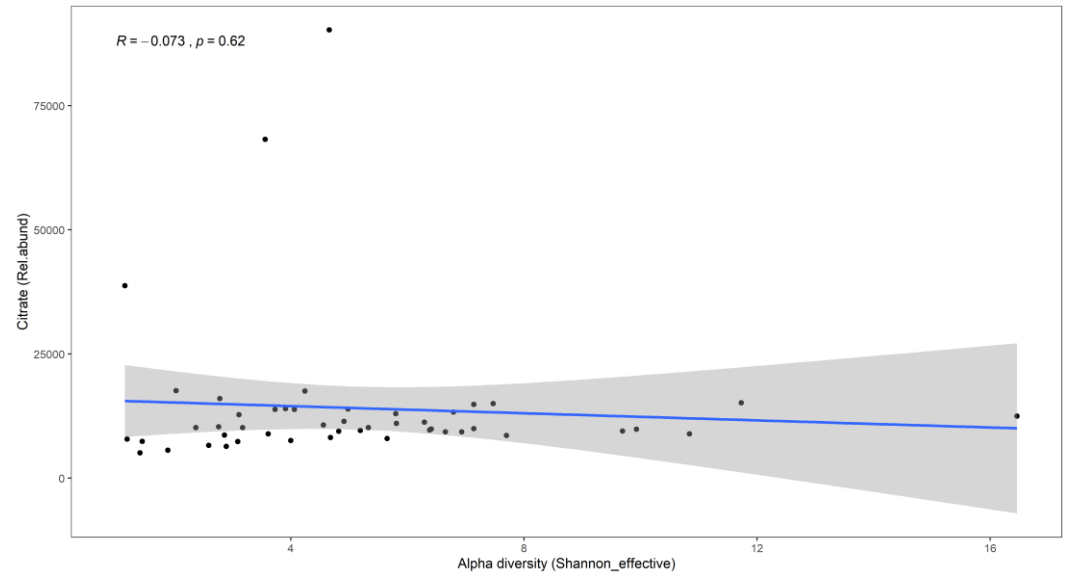

R = 0.0031 , p = 0.98
Alanine (Rel.abund)
3e+05
2e+05
1e+05
4
8
12
16
Alpha diversity (Shannon_effective)

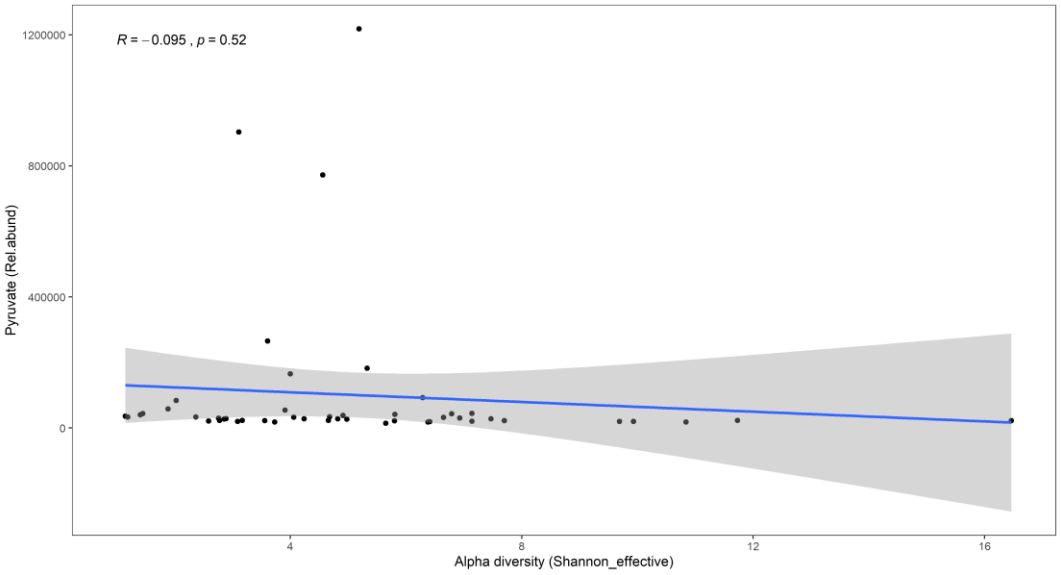

R = −0.16 , p = 0.28
Lactate (Rel.abund)
125000
100000
75000
50000
25000
0
4
8
12
16
Alpha diversity (Shannon_effective)

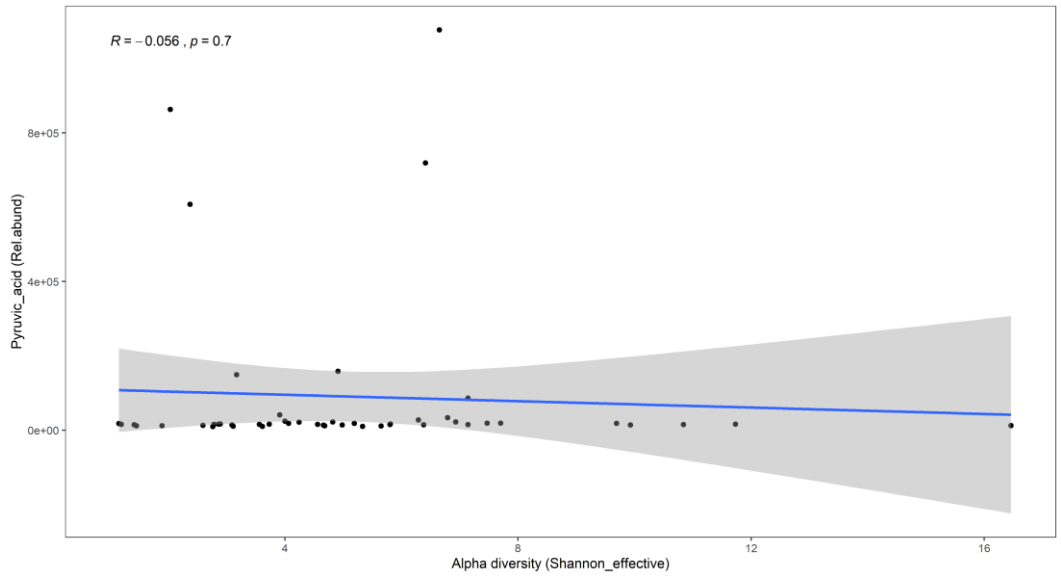

R = 0.069 , p = 0.64
Isoleucine (Rel.abund)
3e+05
2e+05
1e+05
4
8
12
16
Alpha diversity (Shannon_effective)


R = −0.18 , p = 0.22
Leucine (Rel.abund)
160000
120000
80000
4
8
12
16
Alpha diversity (Shannon_effective)


R = 0.012 , p = 0.94
D.glucose (Rel.abund)
3e+05
2e+05
1e+05
0e+00
4
8
12
16
Alpha diversity (Shannon_effective)


R = −0.073 , p = 0.62
Citrate (Rel.abund)
75000
50000
25000
0
4
8
12
16
Alpha diversity (Shannon_effective)


R = −0.095 , p = 0.52
Pyruvate (Rel.abund)
1200000
800000
400000
0
4
8
12
16
Alpha diversity (Shannon_effective)


R = −0.056 , p = 0.7
Pyruvic_acid (Rel.abund)
8e+05
4e+05
0e+00
4
8
12
16
Alpha diversity (Shannon_effective)

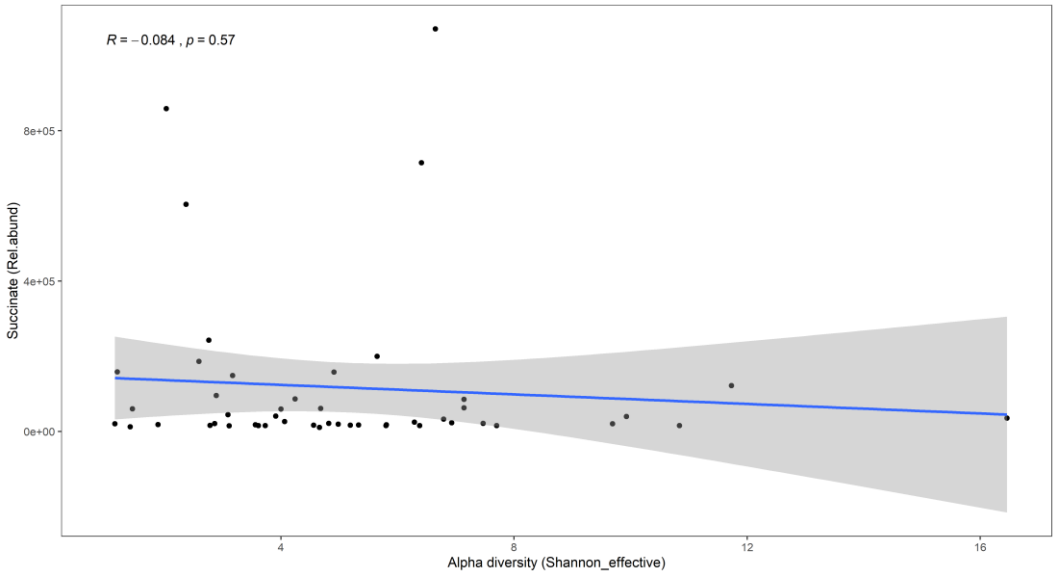


Figure A13 Metabolite correlation to alpha diversity at baseline. Pearson correlation coefficient the corresponding p-values are displayed.

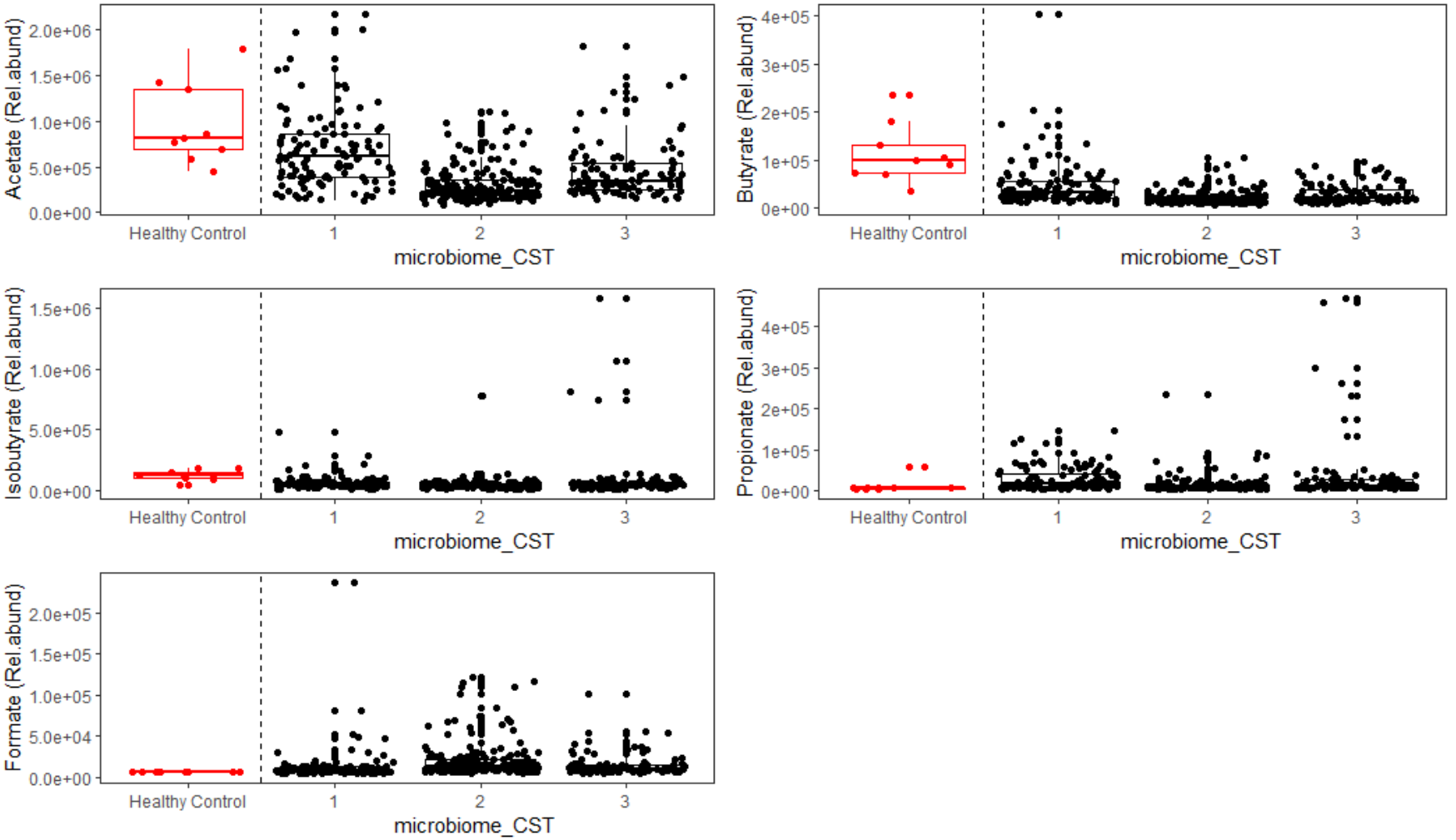


Figure A14 SCFA metabolites in all samples by their respective CST

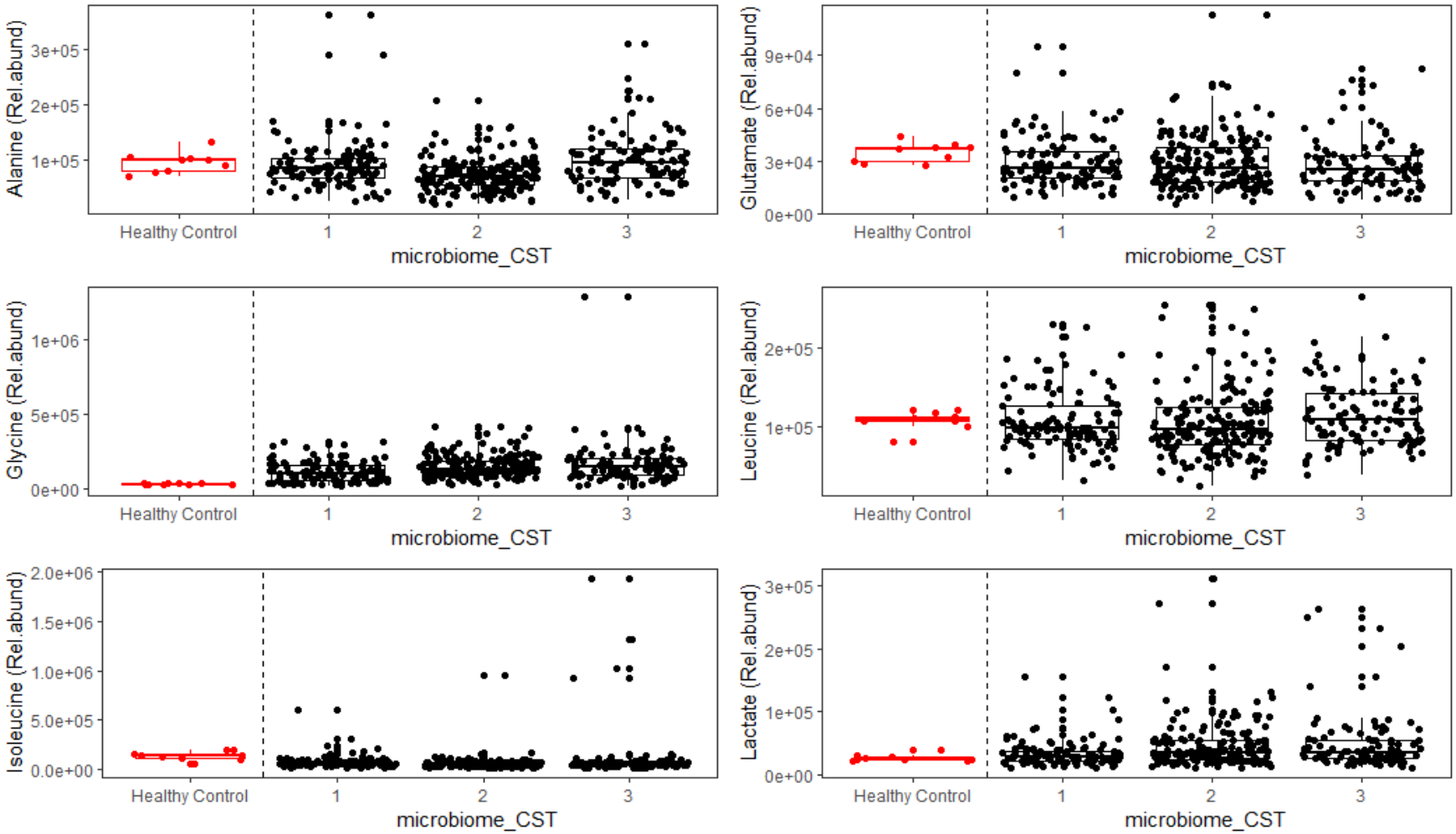


Figure A15 Amino acid metabolites and lactate in all samples by their respective CST

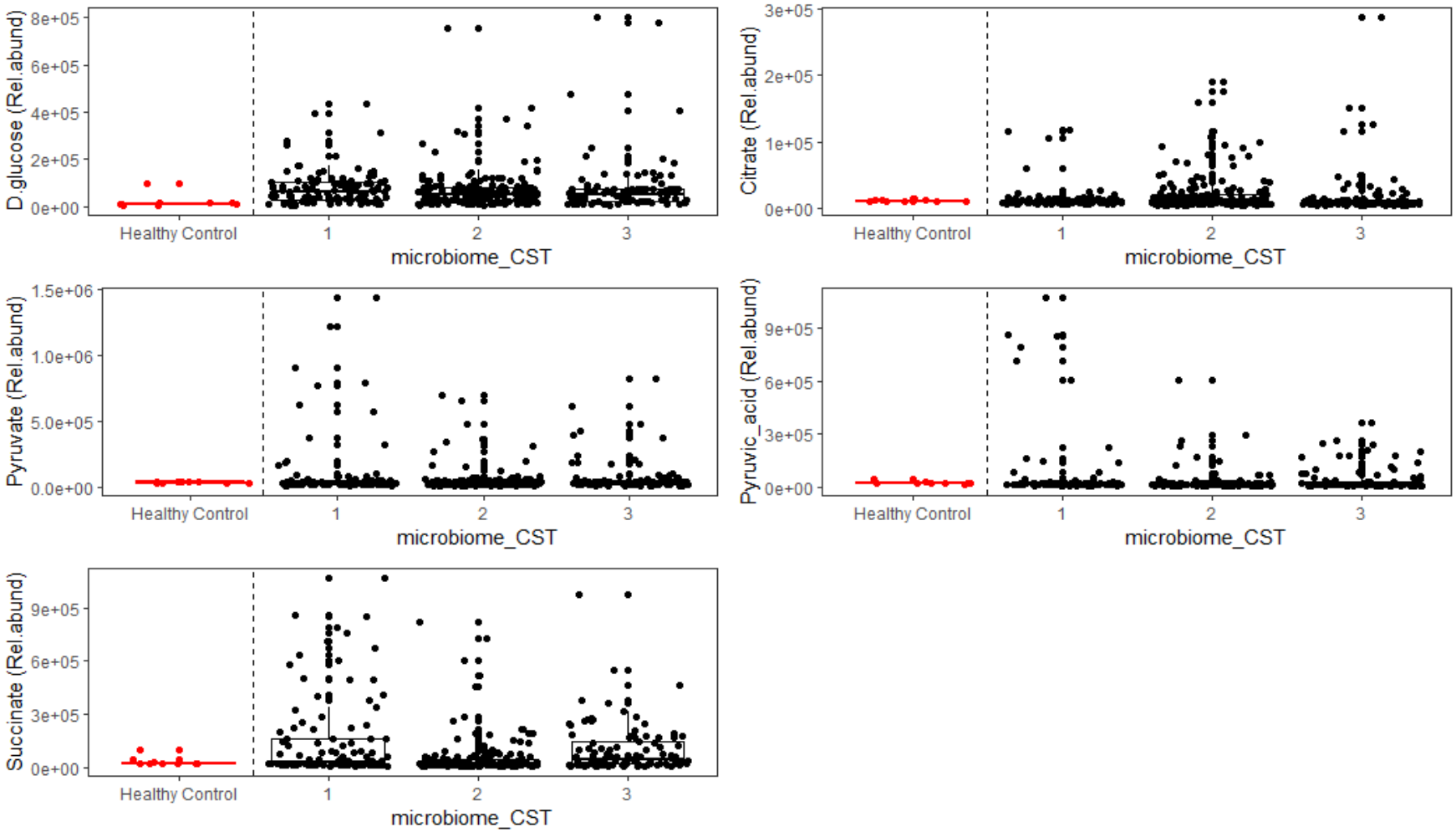


Figure A16 TCA metabolites in all samples by their respective CST