# A Large Scale Metagenomic Analysis of the Faecal Microbiota in Preterm Infants Developing Necrotising Enterocolitis

Thesis submitted in accordance with the requirements of the University of Liverpool for the degree of Doctor of Philosophy by

Nicholas Alexander Franklin Ellaby

January 2018

# Abstract

Necrotising enterocolitis (NEC) is an inflammatory intestinal disorder affecting premature infants. Despite the worldwide improvement of health care practices and facilities raising the survival rates of neonatal premature infants, there has not been any improvement in treatment options or mortality rates for NEC. There has been an extensive volume of research into NEC, though to date there has not been any evidence to directly associate a causal agent to this devastating disease, nor have there been any conclusive observations of NEC prior to birth. The only key prognostic signal for NEC is that onset and severity of the disease are significantly associated with the prematurity of the neonatal infant.

During the process of birth, the infant transitions from the near sterile conditions of the womb to the outside environment teeming with bacteria. Only then does the infant develop a symbiotic relationship as a host to beneficial bacteria. Upon this transition the community of gut microbes develops and aids in a range of host functions, namely digestion and absorption of nutrients, as well as the immune response. It is also at this time that NEC can begin to develop in the lower gastrointestinal tract. This coinciding of factors implicates the colonisation of the gut with bacteria in the development of NEC, however studies to date have failed to provide consistent reports of a causative pathogen or a characteristic gut microbiome structure associated with NEC.

The purpose of this study was to characterise the structure of the bacterial community in the gastrointestinal tract of infants with and without NEC in order to identify a community distinctive to infants with NEC. This was addressed using non-invasive faecal sampling of premature infants in a large, prospectively enrolled cohort from across England, sampled over a twenty-day window spanning ten days prior to and ten days following the onset of NEC.

Using established V4 16S rRNA protocols, bacterial taxa within environmental samples could be characterised without relying on classic culture dependent methods. With this amplification technique it was possible to utilise small amounts of bacterial DNA isolated from infant faecal samples while at the same time mitigating the selection bias associated with culture-based techniques.

Short read sequencing (illumina MiSeq) was performed on a total of 656 faecal samples from 132 infants spanning eight neonatal intensive care units across England. All infants had gestational durations of less than 35 weeks. 44 infants had NEC (225 samples) and 88 were assigned as controls (431 samples) according to key risk factors defined by medical practitioners. Taxonomic abundances assigned with the QIIME informatics pipeline were normalised using non-reductive negative binomial normalisation. Local contributions to beta-diversity (LCBD) scores were used to quantify taxonomic changes in the community structure through subset regression. Non-metric multidimensional scaling (NMDS) was used to establish risk factors that best described NEC and control samples. The Random Forest machine learning algorithm was used to establish taxa that best discriminated between NEC and control infants, as well as to identify any conserved pathogens.

Subset regression identified feeding regime, mode of delivery and age at sampling as significant discriminating factors for the NEC status of infants based on sample LCBD values. However, NMDS plots of sample LCBD values showed no clear clustering of samples according to NEC status. Canonical correlation analysis (CCA) indicated that this variability was due to inter-individual differences. Of the risk factors that could be accounted for, feeding regime was the most effective in differentiating community structures of NEC and control infant samples. There was also evidence that initial communities were influenced by delivery method. Three subgroups of infants based on these influential risk factors and with sufficient sampling depth were established and analysed separately in addition to collective, population-scale analysis.

Random Forest analysis demonstrated that reduced abundance of the genus *Bifidobacterium* was significantly associated with NEC across all sub-groups of infants. Additionally, this method of analysis indicated no clear pathogenic taxa that consistently spanned the population. For infants with NEC that were delivered by caesarean section and fed both formula and breast milk, there was increased abundance of the genus *Dialister* when sampled shortly after birth. Infants that did not develop NEC, who were delivered vaginally and fed both formula and breast milk were seen to have a significantly greater abundance of the genus *Veillonella* over time, relative to NEC subjects. There were no additional taxonomic differences that could be ascertained in the sub-group of infants who were vaginally delivered and fed breast milk exclusively.

Overall, the unique nature of the microbiome and the high degree of inter-individual variation within the community made direct comparisons between NEC and non-NEC subjects difficult. However, by accounting for factors that were significantly associated with NEC status it was possible to observe consistent association of increased bifidobacterial abundance in infants that did not develop NEC. This highlights the importance of large scale studies and case-control assignment when analysing complex community structures such as that of the human gastrointestinal tract, as well as the powerful, deep analytical analysis provided by machine learning algorithms. Further work should look to establish the impact and role of Bifidobacteria in the human gut community to inform early interventions in healthcare settings for infants at risk of NEC, focussing specifically on encouraging the development of a community structure reflective of that observed in premature infants that do not develop NEC.

# Acknowledgments

To my parents John and Christina Ellaby who never doubted my ability and raised me to strive for the highest level with the greatest work ethic, I would be nowhere without you both. To my supervisors who stuck by me through thick and thin, Chris Probert, Neil Hall, Alistair Darby and Andy Ewer. They never doubted my ability and worked tirelessly over five years to help produce this work. To Hannah Chillingworth and Fiona Probert who had to endure proof reading and know my biggest secret, English is not my first language. To Richard, Shirley and Peter Evans who always supported me and provided a roof over my head when I most needed one, I cannot thank you enough for everything you have done for me. Finally, all the family, friends and colleagues who are too numerous to mention but who all were incredibly supportive in my work.

Word Count: 75,635

# Glossary of Terms

| | |
|---|---|
| AMY1 | Amylase Gene |
| ANOVA | Analysis of Variance |
| BHH | Birmingham Heartlands Hospital |
| BWH | Birmingham Women's Hospital |
| *C. diff* | *Clostridium difficile* |
| CHIP | Carboxyl Terminus of Hsp70-Interacting Protein |
| DNA | Deoxyribonucleic Acid |
| dNTPs | Deoxynucleotides |
| EGF | Epidermal Growth Factor |
| GI | Gastrointestinal |
| HGT | Horizontal Gene Transfer |
| HMO | Human Milk Oligosaccharides |
| Hsp | Heat Shock Protein |
| IAP | Intestinal Alkaline Phosphatase |
| IBS | Irritable Bowel Syndrome |
| IEL | Intraepithelial Lymphocyte |
| IL-8 | Interleukin-8 |
| IκB | I-kappa-B |
| LCBD | Local contributions to beta-diversity |
| LPS | Lipopolysaccharides |
| LWH | Liverpool Women's Hospital |
| MetaHIT | Metagenomics Project of the Human Intestinal Tract |
| NEC | Necrotising Enterocolitis |
| NFκB | Nuclear Factor Kappa-Light-Chain-Enhancer of Activated B Cells |
| NHS | National Health Service |
| NICU | Neonatal Intensive Care Unit |
| NMDS | Non-metric multidimensional scaling |
| NOD2 | Nucleotide-Binding-Oligomerization Domain-2 |
| OOB | Out of Bag |
| OTUs | Operational Taxonomic Units |

| | |
|---|---|
| PAF | Platelet-Activating Factor |
| PCR | Polymerase Chain Reaction |
| PERMANOVA | permutational multivariate analysis of variance |
| PPROM | Preterm Premature Rupture of the Membrane |
| RNA | Ribonucleic Acid |
| rRNA | Ribosomal Ribonucleic Acid |
| RSH | Royal Shrewsbury Hospital |
| RWH | Royal Wolverhampton Hospital |
| SCFA | Short Chain Fatty Acids |
| SMAC | Second Mitochondria-Derived Activator of Caspases |
| STH | Sheffield Teaching Hospital |
| TJ | Tight Junction |
| TLR | Toll-Like Receptors |
| TMAO | Trimethylamine-*N*-Oxide |
| UCHW | University Hospital of Coventry and Warwickshire |
| UHL | University Hospital of Leicester |
| UHNS | University Hospital of North Staffordshire |
| UK | United Kingdom |
| USA | United States |
| VLBW | Very Low Birthweight |

# Contents

# 1 Introduction

Necrotising enterocolitis (NEC) is the most common life threatening gastrointestinal (GI) disorder found to affect premature infants, and remains a major cause of morbidity and mortality in survivors[1,2].

NEC is an acute inflammatory disease that occurs in the intestinal tract of newborn infants, predominantly infants born prematurely. NEC is characterised by haemorrhagic necrosis of the intestinal tissue which may lead to perforation and destruction of the gut. Symptoms include abdominal distension, bilious vomiting, and bloody stools, which if not alleviated can lead to septic shock, disseminated vascular coagulation, peritonitis and intestinal perforation.

The function of the gut has been associated with the presence and structure of a microbial community referred to as the gut microbiome[3]. The community structure fluctuates through childhood[4] before stabilising in early adulthood. These changes have been shown to be influenced by different environmental factors[5,6].

There have been numerous small and large scale experiments designed to understand if there is an association between NEC and altered gut microbiota development[7,8], beneficial bacterial taxa[9] or if there are strains of bacteria associated with increased risk of NEC[10]. In general there has been an increased interest into research on the gut microbiota and gastrointestinal diseases,[11] such as irritable bowel syndrome[12] and Crohn's disease[13].

Despite many years of research into the pathogenesis and treatment of NEC, relatively little progress has been made towards improving the outcomes of affected infants. In light of recent advances in understanding the association between the gut microbiome and host health, there has been reason to further investigate potential associations with NEC, especially within large scale, population cohorts.

The dramatic decline in the cost of DNA sequencing and the development of techniques which allow for the quantification of diversity within bacterial communities, without the need for invasive or time intensive techniques, have enabled large, population level cohort studies into diseases associated with metabolic, intestinal disorders within fragile hosts.

## 1.1 Necrotising Enterocolitis

### 1.1.1 History of Necrotising Enterocolitis

The first recorded instance of NEC was considered to be that of Caroline Jossey, who was 9 days old. This case was observed by Charles Billard in 1826 but the book in which it was described was not published until 1928, wherein he termed the disease 'gangrenous enterocolitis'[14]. Following a post mortem, he described the same features used today in the diagnosis of NEC; a swollen abdomen, bloody stools, and the terminal ileum being intensely red and swollen with a friable and bloody mucosa. Inspection of the membrane found that the surface was furrowed by numerous wrinkles with black lines within. Additionally, a large number of spots associated with bleeding underneath the surface tissue were observed, these spots being seen in different regions of the colon.

The first reporting and labelling of an NEC case was by Genersich in 1891. He described the classical NEC symptoms including septic shock, gastric retention, bile stained vomitus, abdominal distention with ileus, and bloody stools[15]. The pathology was first defined and articulated only in 1952 by Schmidt and Quaiser, who named it as *enterocolitis ulcerosa necroticans*[16,17] - 126 years after its original description[14]. After the early 20th century there appears to have been an increase in the number of NEC cases around the world, in parallel with the development of neonatal intensive care units (NICU). It is likely that this was linked to the increased number of premature infants surviving birth due to advances in protocols associated with infant care[18,19,20,21].

NEC was brought to public awareness by Dr Touloukian in the 1960s, who suggested that many of the poorly defined reports relating to functional ileus, intra-abdominal abscesses, spontaneous perforations of the ileum, appendicitis, and colitis with perforation in newborn infants appear to represent the same entity. The purpose of his report was to suggest the surgical management of such diseases and how they have evolved between 1955 and 1966[22]. In his review, Dr Touloukian identified the first instance of infant survival after NEC related surgery (attributed to Agerty *et al* in 1943[23]). 25 cases were described in this report (18 female, 7 male) between 1955 and 1966 in which the incidence of prematurity (22-39 weeks gestation) was high and birthweight low, (only 7 patients weighed over 2500 grams at birth). His primary findings included:

> "*(1) An extremely high incidence of fetal or neonatal hypoxia, (2) a latent period prior to onset of clinical signs, (3) a variability in the length of necrotic intestine and (4) mucosal necrosis as the constant pathologic finding.*"

It was postulated that a decreased blood flow to the intestine of a hypoxic infant could induce ischaemia, leading to a diminished mucin production and degenerative alteration of the intestinal tissue. Touloukian *et al* suggested that this may be linked to the infant microbiome. It was stated that "gas forming organisms" became more abundant in the lower intestinal tract after 24 hours of life and invaded the mucosa, thereby suggesting that these were responsible for forming the gas pockets located in the submucosa and subserosa. The large number of patients could help define the radiological and clinical presentations of the disease. Following this publication, the interest and research into NEC has risen over the years (Figure 1).

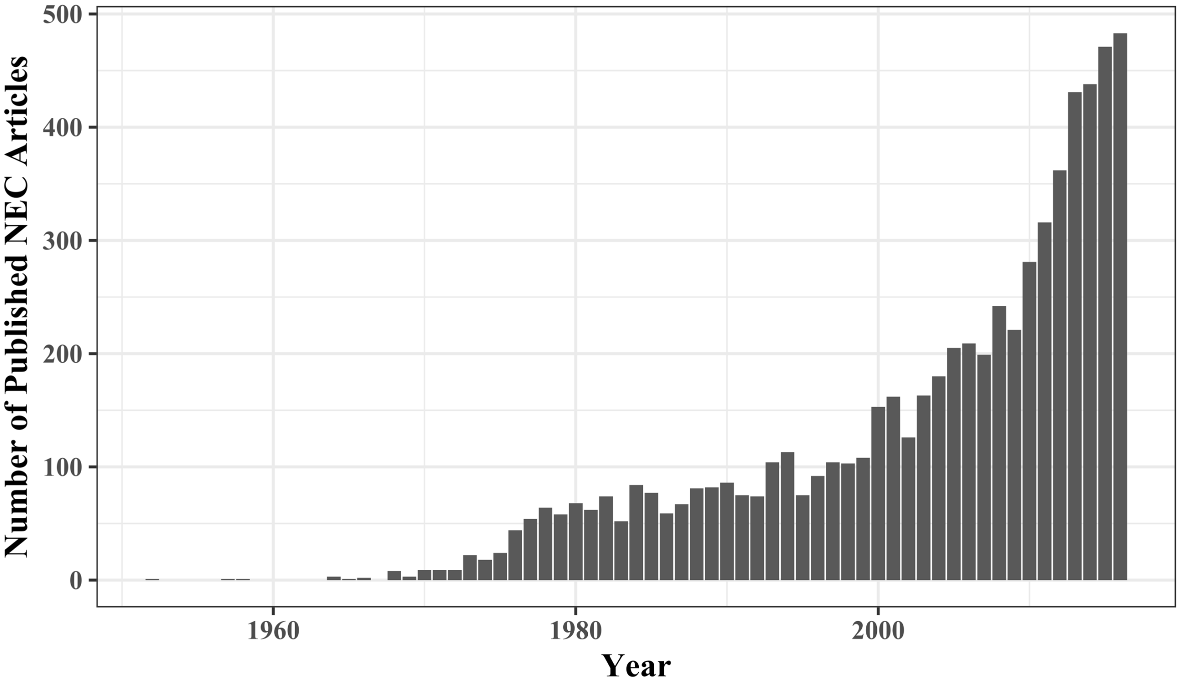


Figure 1 Number of articles relating to NEC published each year 1952-2016

By 1975, detailed analysis of survival rates for NEC patients treated with gastric decompression, antibiotics, intensive support therapy, and intravenous hyperalimentation suggested a significant drop in platelet counts could be used to identify gangrenous bowels. This was considered a predictor of NEC onset and infants should be considered for surgical intervention when this was observed[24].

In 1978 M. J. Bell devised a grading system on which to base therapeutic decisions. He defined NEC as two polar forms with a spectrum existing between[25] the fulminant form, which progresses to intestinal necrosis in 12-24 hours, and the more slowly evolving and benign form. The latter was first observed by Richmond, in 1975, who focused on seven infants with comparatively mild clinical signs and clear radiographic findings of NEC but without pneumoperitoneum[26].

In 1982 R. Wilson *et al* indicated that the age at onset of NEC was primarily determined by the maturity of the gastrointestinal tract. It was suggested that as birthweight increased, the risk of NEC declined. These observations correlated with the gestational age for each birth weight group, with a sharply declining risk for infants between 35 and 36 weeks' gestation being observed[27].

By 1985, research on pathogenic bacteria and the occurrence of NEC had begun. Initial studies focused on species of *Clostridium*[28,29] and *Escherichia coli*[30]. One of the first descriptions indicating an association with the gut microbiota was suggested by J. Blakey *et al,* who described a decrease in the presence of *Bacteroides* and *Lactobacilli* species in infants with NEC compared to controls[31].

In 1986, Walsh was one of the first authors to suggest multiple factors associated with the development of NEC and further expanded on the staging criteria proposed by Bell in 1978[32]. This was followed through the 90's with further reviews of multifactorial theories in the prevention of NEC[33,34]. A. Kosloke suggested that ischemic episodes, bacterial interference, and mucosal immaturity had to be treated in order to prevent the onset of NEC[35]. This was built upon with studies into the influence of feeding regimes and the onset NEC,[36] as well as intestinal maturation[37,38].

In 1991 the PCR amplification of the 16S subunit of ribosomal ribonucleic acid (rRNA) was identified as a method that could be used in the characterisation of bacterial strain phylogeny[39].

The advent of high throughput sequencing technology[40] along with multiplexing of samples[41] has enabled researchers to explore the community structure of the gut microbiota in detail for relatively little cost and time when compared to culture based techniques.

This led to greater understanding of environmental factors associated with the development of the gut microbiota in early infancy,[42,43,44] as well as the exploration of potentially beneficial bacterial taxa[9] and strains associated with increased risk of NEC[10]. However, these studies were predominantly small scale and many were contradictory[45]. Only recently have studies aimed to elucidate the microbiota's influence in the development of NEC in large scale cohort analysis[46].

### 1.1.2 Definition of Necrotising Enterocolitis

From a clinical perspective it is clear that NEC symptoms vary among infants, therefore it is important to define a clinical spectrum from which medical staff can gauge the extent of the disease, and the level of treatment required to mitigate the symptoms. The disease ranges from rapid onset and extremely dangerous,[25] to slow and benign[26]. Systemic, intestinal, and radiologic signs are all important in the diagnosis, but these can variable.

Prior to the development of a standardised method by M. J. Bell in 1978 clinicians would primarily rely on stool observation. If the stool was bloody it could be indicative of NEC when presented alongside abdominal distention in the form of air pockets around the intestines and pneumatosis. However, these symptoms could be associated with various other diseases such as focal intestinal perforation[47,48], septicaemia with ileus[23], and neonatal pseudomembranous colitis[49] .

Table 1 defines how Bell described three key stages of NEC[25]. Infants considered to be in Stage I ('Suspected' NEC) were seen to have minor symptoms relating to poor feeding, distention, lethargy, and a high temperature. Stage II ('Definitive' NEC) built upon the same symptoms as Stage I but with the addition of gastrointestinal bleeding, marked abdominal distention, rigid bowel loops, pneumatosis intestinalis, and portal vein gas. These would have been observable clinically and by radiographic evaluation. Other disorders such as malrotation and Hirschsprung's disease must, however, be excluded. Stage III is classed as 'Advanced' NEC, and patients display bowel necrosis, peritonitis, perforation, and septic shock. It has also been noted that a small group of infants with NEC can have less severe symptoms but radiographic evaluation shows pneumoperitoneum.

Table 1 Bell's NEC grading system based on historical, clinical and radiographic data[25].

| | |
|---|---|
| **Stage I (Suspected)** | |
| a | Any one or more historical factors producing perinatal stress. |
| b | Systemic manifestations - temperature instability, lethargy, apnoea, bradycardia. |
| c | Gastrointestinal manifestations - poor feeding, increasing pregavage residuals, emesis (may be bilious or test positive for occult blood) mild abdominal distension, occult blood may be present in stool (no fissure) |
| d | Abdominal radiographs show distension with mild ileus. |
| **Stage II (Definite)** | |
| a | Any one or more historical factors. |
| b | Above signs and symptoms plus persistent occult or gross gastrointestinal bleeding; marked abdominal distension. |
| c | Abdominal radiographs show significant intestinal distension with ileus; small bowel separation (edema in bowel wall or peritoneal fluid), unchanging or persistent "rigid" bowel loops, pneumatosis intestinalis, portal vein gas. |
| **Stage III (Advanced)** | |
| a | Any one or more historical factors. |
| b | Above signs and symptoms plus deterioration of vital signs, evidence of septic shock or marked gastrointestinal haemorrhage. |
| c | Abdominal radiographs may show pneumoperitoneum in addition to others listed in II c. |

Further refinement of the diagnosis of NEC was proposed by Walsh *et al* in 1986. They suggested that not all infants who develop NEC have observable risk signs associated with

Bell's Grading. For example, two studies identified infants with no obvious signs of NEC were seen to have symptoms following a more thorough investigation; five infants in a sample of fifty in the paper published by O'Neil *et al*[50] and six infants in forty-four in another report by Yu and Tudehope[32]. Walsh *et al* suggested that there are two distinct epidemiologic classifications of NEC, endemic and epidemic[32].

Walsh suggested that there was an endemic (sporadic) incidence of NEC in all nurseries studied, ranging between 0 and 2 cases per month. In addition to these there are epidemics of NEC, in which a large number of cases are clustered in space and time[51,52]. It was reported that during epidemics, patients tended to have higher birthweights and Apgar scores as well as later onset of the symptoms than those occurring in endemic cases of NEC[53].

Based on these findings, Walsh proposed modifying Bell's staging criteria to include systemic, intestinal, and radiographic signs, and he suggested that treatment should be based on the added stages and the severity of the illness (Table 2). Progression from Stage I to Stage II usually occurs within 24-48 hours. The progression from Stage II or IIIA to Stage IIIB could be delayed for 5-7 days.

Table 2 Walsh's Suggested Modified Bell's staging Criteria for NEC [32]

| Bell's Stage | Sub-stage Proposed | Description | Systemic Signs | Intestinal Signs | Radiological Signs | Treatment Options |
|---|---|---|---|---|---|---|
| I | A | Suspected NEC | Temperature instability,<br>apnoea,<br>Bradycardia,<br>lethargy. | Elevated pre-gavage residuals,<br>mild abdominal distention<br>emesis,<br>guaiac-positive stool. | Normal or intestinal dilation,<br>mild ileus. | Nil-by-mouth and antibiotics for three days pending culture. |
| | B | Suspected NEC | Same as previous sub-stage. | Bright red blood from rectum. | Same as previous sub-stage. | Same as previous sub-stage. |
| II | A | Definitive NEC | Same as I-A. | Same as previous stage, plus;<br>absent bowel sounds,<br>+/- abdominal tenderness. | Intestinal dilation,<br>ileus pneumatosis intestinalis. | Nil-by-mouth and antibiotics for seven to ten days, if examination is normal in 24-48 hours |
| | B | Definitive NEC | Same as I-A, plus;<br>mild metabolic acidosis,<br>mild thrombocytopenia. | Same as previous sub-stage, plus;<br>abdominal tenderness,<br>+/- abdominal cellulitis or right lower quadrant mass. | Same as previous sub-stage, plus;<br>portal vein gas,<br>+/- ascites. | Nil-by-mouth and antibiotics for 14 days,<br>$NaHCO_3$ for acidosis |
| III | A | Advanced NEC | Same as previous sub-stage, plus;<br>hypotension,<br>severe apnoea,<br>combined respiratory and metabolic acidosis,<br>disseminated intravascular coagulation. | Same as previous sub-stage, plus;<br>signs of generalised peritonitis,<br>marked tenderness,<br>distention of abdomen. | Same as previous sub-stage, plus;<br>definite ascites. | Same as previous subs-stage, plus;<br>200+ ml/kg fluids,<br>inotropic agents,<br>ventilation therapy,<br>paracentesis. |
| | B | Advanced NEC | Same as previous sub-stage. | Same as previous sub-stage. | Same as previous sub-stage, plus;<br>pneumoperitoneum. | Same as previous sub-stage, plus;<br>surgical intervention. |

## 1.2 Affected individuals & Rate of Incidence

The primary group affected by NEC are premature infants, with 90% of cases occurring in infants born at less than 34 weeks gestation[54,55,56,57]. Additionally, it has a much greater prevalence in infants at very low birth weight (VLBW); 14% of those born at less than 1000g are observed to have NEC. This rate of occurrence decreases significantly in association with increasing gestation and birthweight[58]. NEC predominantly affects infants in NICU units with evidence of individual, sporadic, and nosocomial NEC outbreaks[59]

The average incidence rate of NEC has been shown to be approximately 1 in 1000 live births, however there is considerable variation observed both among and within institutes[56]. Incidence rates vary globally; within the UK the overall rate of occurrence in 2010 for infants in NICUs was approximately 2%[58]. For comparison, the rates in other high income countries were as follows: Australia and New Zealand 1% to 3.5%, Canada 2.5% to 8.7%, Finland 22%, Germany 2.9%, Italy 1% to 13.1%, Japan 0.1% to 5.7%, Korea 6.8%, Poland 8.7%, Spain 2.8% to 10.9%, Sweden 0.1% to 4.6%, Switzerland 0.7% to 11.1% and the USA 2.1% to 9.8%[60]. Data obtained from the UK National Neonatal Research Database[61] between December 2011 and September 2014 described a 23% incidence rate for NEC in all 163 neonatal units across England, from a total of 3,866 infants admitted[62].

When compared to infants born full term, VLBW infants at risk of NEC have abnormal faecal colonisation[63]. They demonstrate a relatively small population of normal enteric bacterial species in addition to a delayed colonisation pattern that would be observed in healthy preterm or term infants[64,65]. Relative to term infants it has been shown that premature infant microbiotas have a limited, non-stable presence of *Bifidobacterium*, a high prevalence of *Staphylococcus*, *Enterobacteriaceae*, *Enterococcaceae* as well as other lactic acid bacteria such as *Lactobacillus* in a low diversity microbiota[66,66,67].

## 1.3 Pathology & Pathogenesis of Necrotising Enterocolitis

Despite the extensive depth of research since 1965, the key frustration of NEC is the lack of understanding of its aetiology. Even within the great wealth of carefully planned and designed experiments utilising the latest models and technology, there is still no clear understanding of NEC development.

### 1.3.1 Pathology

The current hypothesis for the pathology of NEC is, primarily, the coinciding of two or three pathogenic characteristics; coagulation (ischaemic) necrosis; excess protein substrate in the intestinal lumen; and bacterial overgrowth within the intestine[33]. It is characterised by bowel wall necrosis of various lengths and depth, with bowel perforation affecting one in three infants[68]. NEC is most commonly referred to as a multifactorial disease, with a complex interaction of factors leading to mucosal injury[34]. Critically, bacterial colonisation is believed to be necessary for its occurrence[35,69].

#### *1.3.1.1 Prematurity with impaired host defence - Epithelial Barrier*

The most consistently associated risk factor of NEC is prematurity. This inherently has multiple implications in the host's physiology and responses[70,71]. In the context of NEC, the primary region of interest is the intestinal mucosa. This is where necrosis occurs initially in neonates and it appears to be in a constant equilibrium of injury and repair. This may depend on a variety of conditions associated with prematurity including hypoxia[72], perinatal infection (primarily indicated by the presence of interleukin-6)[73,74,75], and starvation[76] (a common approach to mitigate respiratory distress syndrome, immaturity of gastrointestinal function, and systemic hypoxia[77]). Additionally, microcirculatory dysfunction also contributes to epithelial damage[78].

Under normal physiological conditions, repair of the epithelium begins immediately after injury when mature enterocytes migrate into the affected area[79]. This is followed by proliferation of new enterocytes within the crypts of Lieberkühn to complete the repair process[80]. Recent literature has suggested that infants with NEC have both enterocyte migration and proliferation inhibition, resulting in the host being susceptible to further injury. The potential loss of the epithelial barrier could facilitate the translocation of microbial pathogens from the intestinal lumen into the mucosa[81]. The loss of the epithelial barrier allows for the translocation of bacteria from the intestinal lumen into the mucosa layer.

#### *1.3.1.2 Innate Immunity*

Innate immunity relates to the host's nonspecific defence mechanisms that become activated immediately or within hours of a pathogen appearance in the body. These mechanisms include physical barriers such as skin and gastric acid, as well as the immune system that attacks foreign cells in the body. Inflammation is one of the first responses of the immune system to infection. This process is stimulated by chemical factors released by injured cells, with the purpose of establishing a physical barrier against the spread of infection, in addition to promoting the healing of the damaged tissue.

Key components of the innate immune system are located on the epithelial surface; these play a major role in tissue repair and the recognition of microorganisms. In particular toll-like receptors (TLRs) are known to be some of the primary cell-surface sensors of bacterial components and key to the induction of the proinflammatory responses[82].

TLRs are evolutionarily conserved pattern recognition receptors which recognise highly conserved structural motifs known as pathogen-associated microbial patterns[83]. Pathogen-associated motifs include components such as mannans on the yeast cell wall, formylated peptides, and various bacterial cell wall components such as lipopeptides, peptidoglycans, teichoic acids, and lipopolysaccharides (LPS).

TLR4 appears to have a crucial role in NEC development[84,85,86]. Cynthia Leapheart *et al* demonstrated a clear association in human and mouse models of NEC and increased expression of TLR4 within the intestinal mucosa. Physiological stressors, such as LPS and hypoxia, were associated with NEC development by sensitizing the murine intestinal epithelium to LPS through upregulation of TLR4. It was shown that TLR4-mutant mice were protected from the development of NEC when compared with wild-type mice.

Ward Richardson *et al* showed TLR4 activation causes apoptosis of the small intestine or colon in newborn mice but not in adults. Its activation was influenced by nucleotide-binding-oligomerization domain-2 (NOD2)[85]. NOD2 activation inhibited TLR4 in enterocytes, but not macrophages, and reversed the effects of TLR4 on intestinal mucosal injury and repair. The protection by NOD2 utilised a novel pathway linking NOD2 with second mitochondria-derived activator of caspases (SMAC)-diablo, wherein NOD2 reduced SMAC-diablo expression, attenuated the extent of enterocyte apoptosis, and reduced the severity of NEC. This was further supported by the work of Sodhi *et al* who were also able to show that TLR4 activation significantly impaired enterocyte proliferation in the ileum, but not the colon, of newborn mice; this was not observed in adult mice[86].

Developing foetuses express elevated levels of TLR4 until the end of gestation; this could be linked to the importance of TLR4 in the proliferation and differentiation of the intestinal epithelium tissue during the embryogenic period. When observed in TLR4 knockout mice there was an increase in the frequency of goblet-like cells and these appeared to be more apparent along the duodenum-jejunum ileum axis[87].

These cells are modified, simple, columnar epithelial cells whose function is to secrete gel-forming mucus, thereby protecting the mucous membranes. In this study, it was shown that both the mice with TLR4 knockout isolated within the intestinal tissue, and those with complete TLR4 knockout were observed to have significantly reduced levels of bile acids in the ileal lumen and stool relative to wild-type strains.

The evidence that bile acid concentration in the newborn gut has been linked to the regulation of goblet cells[88] raises the possibility that bile acids could play a role in the increased level of goblet cells observed within TLR4-deficient mice. However, when treated with four antibiotics neither knockout mouse strain showed a difference in the level of goblet cells. This would suggest that bile acid concentrations have a greater influence on the levels of goblet cells than exposure to bacteria.

Expression of TLR4 in preterm babies is very high relative to full term infants. When combined with the introduction of environmental bacteria, relative to the developing foetuses which remain in a quasi-sterile environment, the TLR4 signal could be over activated. This, in turn, decreases the host's ability to repair the epithelium after injury. This results in gut barrier failure, bacterial translocation into the mucosa, intestinal inflammation and activation of systemic inflammatory responses[84].

However, there appear to be additional factors that result in most premature infants not developing NEC, despite the observation that TLR4 activation is significantly increased in preterm infants. This would suggest that mechanisms are involved that limit the consequences of TLR4 activation within the newborn intestinal epithelium. These mechanisms are likely to be influenced or based on intra- and extracellular factors and are probably affected by the microbiota composition.

#### *1.3.1.3 Intra- /Extracellular Mechanisms*

##### 1.3.1.3.1 Hsp70

Heat shock proteins (Hsp) are a family of intracellular proteins activated by a variety of stressors and which contribute to the delivery of proteins into a degradation pathway involving the ubiquitin-proteasome. Hsp70 is the predominant member of this family and chaperones molecules, specifically, the carboxyl terminus of Hsp70-interacting protein (CHIP)[89].

Hsp70 has been shown to have a protective role in CHIP-mediated ubiquitination and associated degradation of TLR4, with intracellular Hsp70 induction in enterocytes being linked to dramatically reduced TLR4 signalling. This was assessed by LPS-induced nuclear factor kappa-light-chain-enhancer of activated B cells (NFκB) translocation, cytokine expression, and apoptosis.

Expression of Hsp70 in the intestinal epithelium was significantly decreased in murine and human NEC compared with healthy controls, suggesting protective effects were imparted by Hsp70 on TLR4 and could be critical in the prevention of NEC[90].

### 1.3.1.3.2 Amniotic Fluid/EGF

The epidermal growth factor (EGF) has been observed to influence the TLR4 signal. EGF is an extracellular factor that is extremely rich within the amniotic fluid. It inhibits the TLR4 signalling by the peroxisome proliferator-activated receptor gamma and NFκB pathway.

During gestation, the foetus takes in high quantities of amniotic fluid which limits the amplification of TLR4 signalling in the intestinal mucosa, this is also apparent in cultured enterocytes exposed to bacterial products. These associations were further confirmed with amniotic fluid-mediated TLR4 inhibition, which reduced the severity of NEC in mice through EGF receptor activation[91].

It was observed that NEC development in both mice and humans was associated with reduced EGF receptor expression, which was subsequently restored by the administration of amniotic fluid-mediated EGF signalling.

### 1.3.1.3.3 PAF

TLR4 signalling has also been shown to upregulate platelet-activating factor (PAF) expression. PAF has been demonstrated to increase the risk of injury in experimental models of NEC[92]. It is an endogenous phospholipid mediator that is synthesised and secreted by many cell types with relevance in a variety of pathophysiological processes. It has been shown that LPS and PAF act synergistically to induce bowel necrosis, in addition to other systemic changes such as hypotension, haemoconcentration, and leukopenia[92].

#### *1.3.1.4 Gut microbiota*

NEC is associated with changes in the abundance of bacteria in many studies, but no single species has been consistently identified as a causative pathogen. Research has suggested that NEC may be a consequence of prematurity, enteral feeding and bacterial colonisation, where feeding results in inappropriate colonisation and an exacerbated inflammatory response. As such, it has been hypothesised that altered colonisation patterns of the premature intestine could cause NEC[93].

A community established from naturally occurring bacterial, known to be members of the gut microbiome and found in stable proportions (equilibrated), have been shown to have an essential role in nutrient digestion and metabolism[94], vitamin synthesis[95], immune tolerance,[96] and maturation of the intestinal mucosa[97]. The composition of this community varies widely along the different regions of the gastrointestinal tract according to factors such as transit time, pH, nutrient availability, oxygen tension, host secretions, mucosal surfaces, and interactions with the immune system[98].

With the introduction of improved sequencing technologies and databases it has been possible to perform direct sequencing and interpretation of bacterial deoxyribonucleic acid (DNA) from stools. However, even with the utilisation of these methods there have been inconsistent reports of the level of association with the gut microbiota and NEC.

Some researchers have observed little or no correlation between bacteria and NEC. Raveh-Sadka *et al* identified potentially pathogenic bacteria of the same species colonising many infants within their cohort. Genome-resolved analysis revealed that strains colonising each infant were typically distinct and that no strain was common to all infants who developed NEC[99].

Erika Claud *et al* demonstrated a temporal pattern in the faecal samples of control infants which converged towards that of a healthy full-term breast-fed infant. In contrast, the microbiota development in infants with NEC diverged from that seen in controls three weeks prior to diagnosis[100].

The majority of differentially abundant genes in the NEC patient were associated with members of the family Enterobacteriaceae, but there was no significant differences associated with specific taxa[100]. In addition, Erik Normann *et al* identified a high relative abundance of Bacillales and Enterobacteriaceae in the early time points of patients with NEC but in contrast, healthy controls were observed to have a greater dominance of *Enterococcus* relative to NEC subjects. However, neither of these differences was seen to be statistically significant[101].

Some researchers have identified associated disparate organisms such as Gram-negative Bacilli. Morrow *et al* identified infants with NEC showing community dysbiosis preceding its onset in addition to lacking propionibacteria when compared with their controls. Earlier communities in the diseased infants appeared to become dominated by *Firmicutes*, followed by Proteobacteria dominance, specifically Enterobacteriaceae. Those NEC diagnoses preceded by *Firmicutes* dominance were observed to have an earlier onset relative to those preceded by Proteobacteria dominance[102].

Proteobacteria were also observed to contribute a significantly higher proportion of the microbiome in infants with NEC and additionally Actinobacteria at one week prior to diagnosis. These were offset by lower numbers of *Bifidobacteria* and Bacteroidetes. Additionally, within this same study a distinct, novel signature most closely resembling *Klebsiella pnuemoniae* was strongly associated with NEC development later in life[103].

Further support for blooms of Proteobacteria within NEC subjects was observed in the work by Mai Volker *et al*, observed to be offset by *Firmicutes* in controls. This group of Proteobacteria was not defined in any database but most closely resembled the Enterobacteriaceae family[104].

Clostridia have also been observed to be in greater abundance in early onset NEC subjects compared with their controls, specifically *Clostridium sensu stricto*. In late onset NEC, *Escherichia/Shigella* in addition to other Gammaproteobacteria were observed to increase in relative abundance prior to the onset of NEC[105].

Large scale 16S rRNA analysis of faecal samples by Kathleen Sim *et al* also demonstrated that Clostridia were seen to be overabundant in pre-diagnosis samples from infants with established NEC. Culture analysis confirmed the presence of *Clostridium perfringens* type A; those NEC infants that did not have *C. perfingens* prior to diagnosis were seen to have an overabundance of *Klebsiella*[106].

There is considerable inconsistency between reports on how the gut microbiome of infants with NEC differs from those who do not develop the disease and this has largely been attributed to small scale studies limited by sampling challenges. These challenges are, in part, due to the unpredictable occurrence of NEC throughout the first two months of life[107].

One of the symptoms of NEC is reduced defecation at the onset of the disease. This is partly attributable to the disease itself but also, in many cases, clinicians treat NEC using nil-by-mouth feeding regimes, further reducing the likelihood that infants will pass stools. This leads to variable frequency of stool production within and between infants of this age and as such timed sample sets are often non-uniform.

Another aspect that further complicates the trends and patterns that are observed within NEC-control cohorts are the sudden population shifts that are inherent to the developing gut community which gradually converge into a community more closely shared between individuals[108].

Recent large-scale analysis has described significant differences emerging after one month of age in NEC infant gut microbiomes. Specifically, the time-by-necrotising-enterocolitis interaction emerged only after the first month of age. Mixed model analysis described a positive association with Gammaproteobacteria and a negative association with strictly anaerobic bacteria, in particular, Negativicutes. A more extensive dataset found the Clostridia-Negativicutes class were negatively associated with NEC subjects. These associations were observed to be strongest for infants born at less than 27 weeks' gestation[46].

As described previously, activation of TLR4 on the intestinal epithelial lining by Gram-negative bacteria leads to a number of deleterious effects, including increased enterocyte apoptosis, impaired mucosal healing and enhanced proinflammatory cytokine release[109].

Infants with NEC show abnormal and inappropriate colonisation manifested by low microbiota diversity[110], increased Gram-negative pathogen colonisation, and high levels of pathogens in the peritoneal cavities[111]. The influence of bacteria in the development of NEC is somewhat supported by evidence that antibiotic administration in newborn mice[87] and NEC infants[112,113] decreased bacterial load and protect against the onset of NEC.

#### *1.3.1.5 Intestinal Tissue Integrity – Immaturity*

The neonatal intestinal barrier is immature in preterm infants and has been shown to mature post-natally[114,115,116]. Multiple factors can induce postnatal intestinal maturation of this barrier including epidermal growth factor[117], endogenous glucocorticoids[118], diet,[119,120,121] and commensal bacterial[114,122]. Commensal bacteria have been shown to induce expression of tight junction (TJ) proteins that can tighten the intestinal barrier[114,123].

As such, neonates with delayed or dysbiotic colonisation of the gut may be at a greater risk of intestinal inflammation and injury due to an immature or defective intestinal barrier, which could allow the translocation of microbes, their products or toxins from the gut lumen[124]. This could explain why prolonged antibiotic treatment has been observed to have increased risk of late onset sepsis and NEC[125], whilst infants administered probiotics were seen to have reduced incidence of NEC[126].

#### *1.3.1.6 Histology*

From a histological perspective, a large, retrospective clinicopathologic 10 year study performed by Balance *et al*[33] analysed intestinal specimens from 84 infants with NEC; 64 surgical and 19 from autopsies. They found that their observations were very similar to other bowel diseases known to be ischaemic in origin; diseases such as necrosis after mesenteric artery occlusion; embolic infarction; volvulus; and intussusception. They identified that coagulation (ischaemic) necrosis, inflammation, and bacterial overgrowth were all present in the intestine of nearly all patients. In many cases they were accompanied by idiopathic medical complications and variability of symptoms.

TJs are an important component to the structural integrity of the gut barrier. These serve as a molecular fence that partitions the cytosolic membrane into apical and basolateral domains, which act to preserve the cellular polarity. In combination with transcellular transportation, these domains also generate distinct internal environments, function to strengthen the epithelial barrier, and prevent translocation of bacteria into the mucosa tissue.

A decreased expression of TJ has been linked to increased intestinal permeability in NEC and other inflammatory intestinal diseases[127,128]. Altered expression and localisation of claudin and occluding proteins that form these TJs, have been observed in NEC and are associated with increased epithelial permeability[127,129].

As previously discussed, failure of the gut barrier function is known to cause systemic inflammation from translocation of bacterial products such as LPS. In preterm infants multiple stressors associated with prematurity (hypoxia, infections, treatment with indomethacin and glucocorticoids) have been suggested to compromise TJs and therefore epithelial integrity[107].

Intestinal alkaline phosphatase (IAP) has been shown to detoxify LPS by dephosphorylating the lipid. This prevents LPS from binding to the TLR4 complex and replacement or supplementation of IAP has been shown to reduce the level of intestinal inflammation and endotoxemia[130,131,132,133].

### *1.3.1.7 Translocation*

Under normal conditions, microbes and their molecular products are prevented from moving through the epithelial barrier by a complex of tightly regulated cellular and molecular processes. However, as discussed, this complex has been shown to be weakened in premature infants both intrinsically, with immaturity in the tissue differentiation and low levels of EGF, and extrinsically via LPS exposure, TLR4, and PAF over-expression leading to inflammation cascades. Therefore, reduced intestinal epithelial integrity encourages the opportunity for bacteria, pathogenic or commensal, to translocate through the membrane.

Bacterial translocation through the intestinal barrier subsequently leads to systemic inflammatory response, as seen in subjects with NEC. Evidence of the clinical impact of bacterial translocation was observed by O'Boyle *et al* who identified a significant increase in the post-operative sepsis of patients, who were observed to have bacterial translocation[134] upon undergoing a laparotomy procedure.

The lack of identification of bacteria and endotoxins in portal blood and the development of the systemic inflammatory response system has suggested that bacterial translocation occurs via a lymphatic route[135]. This response to extreme stress is characterised by an extremely large release of cytokines, endothelial cell damage, tissue oedema, increased tissue permeability, activation of the coagulation system, platelet aggregation, local tissue hypoxia with shunting, and a hypermetabolic state[136].

Bacterial translocation could indeed be a major component in the development of systemic inflammatory response system, however human studies addressing this question have been hindered by methodological issues due to serial cultures from mesenteric lymph nodes not being possible in humans.

## 1.4 Risk Factors for Necrotising Enterocolitis

### 1.4.1 Gestation/Prematurity

As alluded to in the previous section, NEC has been defined as a multifactorial disease, with many potential mechanisms contributing to its occurrence. However, it is consistently associated with premature birth.

An estimated 15 million infants are born preterm; this equates to 1 in 10 live births, with rates ranging from 5% in some European countries to 18% in some African countries[137]. The number of preterm births is rising; in Canada for example, rates increased from 6.6% to 9.8% for births under 36 weeks' gestation, 1.7% to 2.3% at less than 34 weeks and 1.0 to 1.2% at less than 32 weeks[138].

This is confounded by the increase in survival rates in high income countries such as the UK and USA, which at the beginning of the 20th century had similar infant mortality rates to the African nations of 40 per 1000. By 1998 this was reduced to 15 per 1000 live births in the USA as neonatal care became more widely available[137] and by 2008 in the UK, infant deaths were reduced to 6 per 1000[139]. The most recent survey from the Office for National Statistics in England and Wales showed that this has been reduced to 4 deaths per 1000 by 2016[140].

Evidence from the UK and USA has shown that whilst there has been a significant decrease in the total infant mortality rate, there has also been a significant increase in the rate of NEC cases[141]. These statistics demonstrate that it is vital for the aetiology of NEC to be understood and successful treatments developed. However, if preterm births could be reduced this would also be a significant step in reducing the incidence of NEC.

Factors leading to preterm birth can arise from maternal or foetal causes, which lead to induced labour, spontaneous preterm labour with intact membranes, and preterm premature rupture of the membrane (PPROM)[142]. The causes of preterm births differ by ethnic group with spontaneous preterm birth being associated most commonly with white women whilst PPROM is associated most commonly with black women[143].

### 1.4.2 Birthweight

Shorter gestational durations are unavoidably linked to a lower birthweight. It has been commonly reported that NEC has a greater prevalence in infants of lower birthweights[56,144,27], with the highest risk groups being observed between 750-990g. Infants born below 750g are less likely to survive and, therefore, are less likely to be diagnosed with NEC prior to other complications. Evidence suggests that risk factors vary with birthweight and gestational age, with more immature infants and lower birthweights developing NEC onset at significantly later ages. This is potentially due to the delayed development of the intestine and resulting microbiome, without which NEC cannot develop.

### 1.4.3 Microbiota

As discussed previously, it is thought that an inappropriate inflammatory response is likely to be a major factor in the occurrence of NEC and therefore factors that contribute to this pathology are probable risk factors and should be considered in comparative analysis. Whilst immaturity of the gut and immune system are likely to be related to the decreased gestational duration, interaction with the gut microbiota is a key factor in many preterm morbidities such as NEC[145]. In the normal microbiota the relationship with the host forms a mutually beneficial relationship with the community, supporting important functions in obtaining nutrition[146], angiogenesis[147], and mucosal immunity[38,148].

Recent studies using molecular methods to profile the faecal microbiota of infants with NEC and infants who develop normally have suggested that the disease is associated with a different microbial community structure[149,110,104,102,105]. The preterm infant is exceptionally undeveloped relative to term infants and can be considered foetal in physiology in some regards. When considering the intestine, this includes limited contact with bacteria and food substrates, because these immaturities increase the likelihood of an exaggerated immune response to both commensal and pathogenic bacteria.

This interaction between inappropriate colonisation of premature infant microbiota was hypothesised in 2001, and it was suggested that intestinal injury is a consequence of prematurity, enteral feeding, and bacterial colonisation[93]. In addition to the interaction with TL4 and Gram-negative bacteria, immature intestinal epithelial cells in rodents were observed to have significantly lower levels of I-kappa-B (IκB) genes which have been confirmed as key regulators of the NFκB-dependent inflammatory pathways. IκBα expression has been shown to inhibit the interleukin-8 (IL-8) response to bacteria in the immature enterocyte cell line and therefore could suggest that the microbiota and the immature immune response have implications for the pathogenesis of NEC[150].

The preterm infant colonisation pattern differs from that of a healthy term infant because of factors associated with neonatal care. Significant contributors to the establishment of the microbiota include feeds, birth methods, and antibiotic administration. These should be carefully considered when making comparisons of the community structure of the gut microbiome.

### 1.4.4 Mode of Delivery

Delivery method has been shown to alter multiple factors associated with infants' health and wellbeing. It has been shown that infants delivered vaginally have a greater likelihood of being born earlier, smaller, and with a higher prevalence of sepsis[151]. Early colonisation patterns interact with the intestinal mucosa to form an immune response towards homeostasis or dysregulation[152,153,154] (Section 1.3.1.2).

Neonates born by caesarean delivery have been shown to have different bacterial flora from that of those born vaginally. *Klebsiella*, *Enterobacter*, and *Clostridium* were seen to have increased abundances in those infants delivered by caesarean section[155,156,157,158]. This implies that the delivery method, which shapes the initial gut microbiota, may contribute to the onset of NEC due to an altered response and development pattern in infants delivered by different methods.

The microbiota is clearly influenced by the mode of delivery but recent evidence suggests that the mode of delivery is not significantly associated with necrotising enterocolitis[151].

However, infants were seen to be significantly different based on mode of delivery, particularly in relation to size and prevalence of sepsis[151]. It is important to remember that infants born by different methods are likely to have different colonisation patterns and that these should be considered carefully before conclusions are made.

### 1.4.5 Feeding Ability & Regimes

One of the key challenges faced by newborn infants, term or otherwise, is the transition from parenteral to enteral nutrition. The foetal gut is not inactive during pregnancy as it swallows significant volumes of amniotic fluid which contains proteins that can be digested and absorbed[159]. These peptides, together with the appropriate cell lines, have been identified in the human gut from 6 to 16 weeks after conception[160]. As gestation continues there are changes in the spectra of the molecular forms, with an alteration in the peptides present in different regions of the gut.

This could imply that cells secreting these regulatory peptides are inducing growth and functional development of the foetal intestine. This is particularly apparent in the second trimester where the amniotic fluid has been shown to contain substantial concentrations of hormones and peptides. Amniotic fluid may therefore provide a vital role in the development of the gut in preparation for postnatal feeding.

The foetal gut appears to be prepared for the intake of enteral feeds by the third trimester[159], however after oral feeding is initiated there is an associated change of the gut both physiologically and morphologically.

Evidence for the immaturity of the gut in preterm infants has been seen in the lower capacity for fat absorption (this can be compensated for by using medium-chain triglycerides) and a decreased absorptive capacity for carbohydrates[161]. Therefore, there is likely to be an association between issues in feeding and nutrient uptake in preterm infants, the type of feeds administered, and the development of a normal gut.

This was aptly shown in a study by Berseth where early feeding of preterm infants enhanced the maturation of the small intestinal motor activity and peptide responses to food. Additionally, a delay in enteral feeding was shown to prevent normal maturation of the gut[37].

Many preterm infants who develop NEC are fed enteral milk feeds. When there is no introduction of maternal breast milk and infants are fed artificial formula exclusively, there is evidence of an increased risk of NEC[36]. This is likely to be due to breast milk containing many additional non-nutrient factors such as immunoglobulins, which are involved in intestinal adaptation, maturation, and improved enteral feed tolerance, and which provide a protective influence against infective and inflammatory agents[162].

This understanding of the protective effects of maternal breastmilk has been validated in multiple studies[163,164]. In particular, Lucas and Cole showed evidence that infants who are exclusively fed human milk reduced the incidence of NEC prevalence observed from 3.4% in the control group down to 1% (Paired T-test; p-value = $9x10^{-3}$)[165].

Other differences have been observed with feeds and the development of NEC including timing of introduction of feeds and the volume of daily increments. Those infants who have had feeds introduced earlier and feeding volumes advanced more quickly have been observed to have higher incidences of NEC[166].

### 1.4.6 Infection/Antibiotics

Infections of premature infants, whether isolated to the intestinal tract or systemic, have been shown to spread NEC. It is unclear whether enteric infection is an instigator in the development of NEC. Although there is an abundance of bacteria in the premature intestine in early life[167], which includes the evidence of Gram-negative sepsis, a positive blood culture from infants with NEC is uncommon[168].

A study by Bizzarro *et* al found that of 410 infants with NEC, 158 (39%) were diagnosed with at least one blood stream infection (BSI). Of these, 69 (43.7%) had infection prior to NEC onset (NEC-associated)[169]. Two thirds of those with NEC-associated infections were observed to have Gram-negative bacilli. This would suggest that infection as the precipitating factor in NEC is unlikely, although bacteria are clearly associated in the pathogenesis of NEC, as can be seen in Table 3 which describes all incidences of NEC with significant bacterial, viral or fungal associations.

Although intestinal pathogens may contribute to NEC or NEC-like symptoms in animal models, and in some cases clinical outbreaks, there appears to be no evidence of their presence in the majority of NEC cases[170].

Epidemiological studies have shown a correlation between the duration of antibiotic courses and the occurrence of NEC. Prolonged antibiotic regimes immediately after birth may be associated with an increased risk of developing NEC[171,172]. Though, these studies do not necessarily demonstrate a direct cause and effect between antibiotics and NEC, they do suggest an impact on the microbiome composition. When making comparisons antibiotic usage should be considered carefully.

Table 3 Species and references of bacterial infection association with NEC diagnosis. Adapted from Infectious Causes of Necrotising Enterocolitis[173]

| Bacterial | Viral | Fungal |
|---|---|---|
| *Clostridium spp.* | Astrovirus[174,175,176] | *Candida spp*[33,177,178,179,] |
| *Butyricum*[180,181,182,29] | Cytomegalovirus[183,184,185] | |
| *Difficile*[186,28,187] | Coronavirus[188] | |
| *Perfringens*[189,190,191,31] | Coxsackievirus B2[192,193] | |
| *Cronobacter (Enterobacter)* | Echovirus[194] | |
| *Sakazakii*[195,196,197,198] | Human Immunodeficiency virus (maternal exposure)[199,200,201] | |
| *Enterococcus* (VRE)[202] | Norovirus[203,204,205,204,206] | |
| *Escherichia coli*[169,207,208,30,209] | Rotavirus[210,211,211,212] | |
| *Klebsiella spp.*[213,214,59,215,207] | Torovirus[216,217] | |
| *Pseudomonas aeruginosa*[218,219,220] | | |
| *Salmonella*[221,222] | | |
| *Staphylococcus aureus* (MRSA)[223] | | |
| *Staphylococcus epidermidis*[224,225] | | |
| *Ureaplasma urealyticum*[226,227] | | |

### 1.4.7 Other Factors

Additionally there are other potential risk factors that could be considered niche groups, for example, prenatal course including maternal drug use (specifically cocaine)[228], hypoalbuminemia[229], hypoxic ischaemic encephalopathy, and respiratory distress syndrome[230].

## 1.5 The Diagnosis of Necrotising Enterocolitis

NEC diagnosis is based on the presence of abdominal distention and infrequently in rectal bleeding (haem-positive or excessively bloody stools). Assessment of the infant is then performed by abdominal imaging and sepsis evaluation. Although individually the results of these tests are non-specific findings, such as pneumatosis intestinalis, they are likely to be indicative of NEC and should warrant further investigation[231].

Radiographic analysis of the abdomen can confirm NEC diagnosis and follow the disease progression. NEC radiographic features can be abdominal gas patterns in the dilated loops of the bowel. Bubbles of gas in the small bowel wall are representative of pneumatosis intestinalis – this is primarily associated with Bell's grade II or III.

Following bowel perforation, pneumoperitoneum is typically present; this is indicative of Bell's grade IIIB and associated with a substantial amount of intraperitoneal air. Bowels that remain in a fixed position, i.e. sentinel loops, are also suggestive of necrotic bowel and/or perforation when there is no evidence of pneumatosis intestinalis[232].

Portal venous gas was thought to be a predictor of poor prognosis and indicative of the need for surgical intervention. However, this no longer appears to be the rule as demonstrated by a prospective study of 194 infants with confirmed NEC. In this study, the survival rate of infants with portal venous gas treated medically was greater than those who that underwent surgery[233]. The decision to operate should, therefore be based on the severity of the NEC and not solely on the presence of portal venous gas.

Abdominal ultrasonography is becoming increasingly popular in the diagnosis of NEC[234,235,236,237]. Ultrasound can identify intermittent gas bubbles in liver parenchyma and the portal venous system which is not possible using radiography. More severe NEC diagnostic signs include free gas, focal fluid collections, and increased bowel wall thickness and echogenicity[235,237,238].

## 1.6 Management & Treatment of Necrotising Enterocolitis

### 1.6.1 Immediate Supportive Management of Necrotising Enterocolitis

The initial medical management of NEC includes stabilisation of the infant and resting the intestinal tract, this means the immediate discontinuation of oral feeds for 10 days (nil by mouth) and aspiration of gastric contents, decompression of the stomach by nasogastric tube, fluid resuscitation, intravenous antibiotics for 7-14 days, and correction of metabolic abnormalities[239]. Broad spectrum antibiotics used are usually dependent on the local policy which is informed by the regional microbial flora. Anaerobic cover is used if there is evidence of perforation and/or systemic signs that would indicate severe disease[240,241].

In a survey on the management of NEC it was seen that two thirds of surgeons (52, 67%) kept patients on antibiotics for more than 7 days, whereas the remaining third (23, 34%) kept infants on antibiotics up to 7 days. Usually a combination of two, three or more antibiotics were administered[242]. Confirmation of NEC is then established through use of radiographs and histopathology (See Section 1.5).

### 1.6.2 Surgical Treatment of Necrotising Enterocolitis

Perforated necrotising enterocolitis is a major cause of death and morbidity in infants with NEC, between 30% and 50% mortality for those infants[243]. An appropriate means of mitigating this damage is vital in improving the survival rates of those infants with perforated, necrotic intestinal tracts.

There are two options regarding the surgical treatment of NEC; peritoneal drainage and laparotomy. The standard approach for NEC infants with either a perforated or necrotic intestine is the surgical resection of viable intestinal tissue via laparotomy. However, in critically ill infants this carries substantial risks. Peritoneal drainage is considered the less risky in providing some form of reduction in the symptoms[244].

It has been demonstrated that there is no significant difference between the peritoneal drainage and laparotomy and the mortality of infants at 90 days after operation. There were no statistical differences between rates of dependence on parenteral nutrition 90 days post-surgery for those infants that survived[245].

## 1.7 Prevention of Necrotising Enterocolitis

### 1.7.1 Feeding Methods

Multiple analyses have shown that different feeding regimes have a significant impact on the incidence of NEC, these include increased use of human milk instead of formula milk, early introduction of feeds, and minimal enteral nutrition for infants with feed intolerance[246].

There is little consensus on the exact feeding regimes to use in reducing or treating the presence of NEC; most USA NICUs start parenteral nutrition on day one and the first enteral feed as soon as possible after birth, either as human milk or formula[247]. This is partly due to the higher nutrient requirement of preterm infants compared with term infants[248]. However, many of these decisions are not evidence-based but are derived from personal experience or unit culture.

There is a wealth of data that supports the health benefits of human milk, with evidence suggesting that it decreases the incidence rates of sudden infant death syndrome, childhood infectious diseases, allergic diseases, food intolerance, inflammatory bowel disease, obesity and more[249].

As described earlier, breast milk is abundant in EGF and amniotic fluid, these are both essential for intestinal development as demonstrated in rodent models [250,251,252]. Studies using exclusive human milk regimes have shown a significant decrease in the date of onset and the postmenstrual age of NEC in addition to a significant reduction in NEC incidence[165]. Meta-analysis of data from randomised control experiments also indicate that formula milk fed infants had higher rates of feed intolerance and NEC compared with those who were fed donor breast milk[36].

Other feeding treatments are also utilised. Nil by mouth is used by clinical staff to medically manage NEC patients in up to 50% of infants[242]. Additionally, increased incidence of NEC is associated with a more rapid rate of feeding in very low-birth-weight infants, and a slower feeding rate schedule failed to show an increase in NEC incidence[253].

### 1.7.2 Probiotic Management of the Gastrointestinal Tract

Some centres try to influence the composition of the intestinal microbiota through probiotics. These are bacteria that are introduced into the body because they have beneficial qualities. These have been observed to reduce the incidence and mortality of NEC, however, determining the safest and most effective cocktail of microorganisms, especially without a fundamental understanding of the contribution of bacteria to NEC, is still difficult[45].

Probiotics have been shown to increase the quality of the intestinal mucus, improve motility of the gut, and control the production of inflammatory cytokines[254,255]. Probiotics have been observed to compete with pathogenic bacteria, limiting their potential overgrowth within the intestinal tract. Meta-analysis studies have demonstrated that probiotics can reduce the risk of NEC by ~65% and that treating 25 infants could possibly prevent 1 case of NEC[45]. Most of these studies include a combination of *Bifidobacterium* with other taxa while those studies that contained no *Bifidobacteria* were not observed to benefit  the host[255].

### 1.7.3 Antibiotic Management of Secondary Sepsis

Infectious complications of pregnancy, for example chorioamnionitis, increase the risk of NEC either by direct colonisation or by anatomical and immunological changes to the immature inflammatory cascade of the developing intestine[226,256,257,258]. However, independently proving the association between chorioamnionitis and NEC is a challenge as chorioamnionitis associated with prematurity.

Medical management has been suggested to include broad-spectrum antibiotics based on the known sensitivities of prevalent pathogens associated with the NICU specifically. These should typically be ampicillin plus gentamicin to cover for common intestinal bacteria. The addition of a third antibiotic can provide more targeted anaerobic coverage should that be considered necessary. Piperacillin-tazobactam can provide an alternative that acts as a broad-spectrum antibiotic but also includes typical anaerobes found in the intestinal flora[173].

## 1.8 Unmet Needs

The association between NEC and the development of the microbiota in premature infants is based on the combined evidence that the premature infant immune system exhibits inappropriate responses to commensal and pathogenic bacteria[7,68], in addition to the therapeutic effects of both probiotic[259] and antibiotic regimes[76], for infants diagnosed with NEC.

However, the interaction of internal and external factors influencing the microbiota and the intestinal tract are not fully understood and are likely to be vital in understanding and treating infants with NEC. Therefore, understanding how the preterm microbiota develops and how it changes in infants that develop with and without NEC will be crucial in advancing our understanding of this deadly disease.

## 1.9 The Functions of the Gut Microbiota

It was estimated that there could be as many as 100,000 genes within the human genome, but since the completion of the final chromosome in the Human Genome Project in 2005, it has been shown that this was a large overestimation and actually only 20,000 protein-coding genes exist[260]. Research into the human microbiome estimated that there are approximately $3.8x10^{13}$ bacterial cells across the human body, with the majority being found in the gut[261]. Approximately 9 million bacterial genes found the adult human gut[262] and evidence suggests that this microbiome can contribute a large number of genes to metabolic pathways required by the host[263]. This observation has sparked interest into how these bacterial communities found on or in our person, in these many different niches, are complementing the human genome[264].

It has only recently been possible to estimate the true structure of the microbiome communities because of the advent of higher throughput sequencing technologies. Prior to this, traditional methods were unable to cultivate the majority of microbial taxa within a community and sequencing was cost-prohibitive. However, since the emergence of the new sequencing technologies it has been possible to re-assess various diseases[265], health issues[266] and agricultural[267] challenges using the microbiome perspective.

It is widely estimated that between $10^{13}$ to $10^{14}$ micro-organisms reside in the human gastrointestinal (GI) tract, the largest bacterial niche on the human body, which constitutes the majority of all bacteria found on the human host. The size and diversity of the gut microbiome is directly linked to the multitude of carbon and energy sources available in conjunction with substrate availability and colonic transit time[268]. Therefore, understanding how this community develops, grows, and interacts with the host was considered a high priority in understanding many dietary related diseases.

The European Commission and China initiated the Metagenomics project of the Human Intestinal Tract (MetaHIT)[269]. From 154 subjects it was possible able to identify 3.3 million microbial genes. These appear to be shared between American (>70%) and Japanese (>80%), adults, further confirming that the functions of gut micro-organisms are well conserved and serve to facilitate the metabolic processes occurring in the GI tract.

### 1.9.1 Digestion

One of the primary functions of the human gut microbiota appears to be to aid in the absorption of nutrients found in food sources that would otherwise be indigestible for the host. Bacteria within the gut provide access to metabolic pathways that have not been associated with human cell functions. As such, alterations to the community, both in terms of presence/absence and proportional abundances, have been linked to many changes to dietary functions and dietary-related diseases.

The predominant phyla of a health gut microbiota are Firmicutes and Bacteroidetes. To a lesser extent Actinobacteria and Verrucomicrobia are also important members. Whilst this composition is generally true in health individuals it is known to exhibit temporal and spatial differences in the distribution of genera or species[270,271].

The gut microbiota derives nutrients primarily from dietary carbohydrates.  One characteristic function of the adult microbiota is the fermentation complex carbohydrates that escaped proximal digestion into short chain fatty acids (SCFA), which have been observed, along with many other metabolites, circulating through the digestive system. The most abundant SCFAs in the intestine are acetate, propionate, and butyrate, mainly derived from carbohydrates[272]. Butyrate plays a significant part in regulating the epithelial and immune cell growth, in addition to apoptosis[273]. Evidence suggests propionate is removed by the liver and has implications in cholesterol metabolism, and acetate is oxidised in the brain, heart, and peripheral tissues[274]. This process of carbohydrate digestion is predominantly performed by colonic organisms such as *Bacteroides*, *Bifidobacterium*, *Fecalibacterium*, *Enterobacteria* and *Roseburia*[275,276]. The resulting oxalate in the intestine produced because of this carbohydrate fermentation and bacterial metabolism is utilised by organisms such as *Oxalobacter formigenes*, *Lactobacillus* and *Bifidobacterium* species[277].

The gut microbiota has also been shown to have positive impact on lipid metabolism through suppressing the inhibition of lipoprotein lipase activity in adipocytes. In particular *Bacteroides thetaiotoamicron* has been shown improve lipid hydrolysis efficiency by regulating expression of a colipase that is required for lipid digestion[123]. Amino acid transporters are found on bacterial cell walls that aid in the absorption of amino acids from the intestinal lumen into the bacteria where they are converted into small signalling molecules and antimicrobial peptides (bacteroicins). These functions work in tandem with human proteinases and peptidases to increase the overall efficiency of protein metabolism[278,279].

*Bacteroides* have been implicated in the synthesis of conjugated linoleic acid. This is known to be antidiabetic, antiatherogenic, antibesogenic, hypolipidemic and it is known to have immunomodulatory properties[280,281,282]. *Bacteroides intestinali*, *Bacteroides fragilis* and *E. coli* have been shown deconjugate and dehydrate primarily bile acids, converting these into deoxycholic and lithocolic acids (secondary bile acids) in the human colon[282]. A normal gut microbiome has been shown to increase the concentration of high energy metabolism indicators, e.g. pyruvic acid, citric acid, fumaric acid and malic acid[283].

More recently the gut microbiota has been shown to be crucial in the breakdown of polyphenols in the diet. Polyphenolic secondary metabolites are present in a variety of plants and fruits. These usually remain inactive in the diet until they are biotransformed into active compounds after the removal of the sugar moiety, predominantly performed by the gut microbiota. Structural specificity of polyphenol and the richness of the microbiota is associated with the level of biotransformation that occurs in the intestine. The final products of these pathways are absorbed via the portal vein and travel to other tissues and organs, wherein they provide antimicrobial and other metabolic actions, including androgenic and hypolipidemic actions[284]. Primarily this has been shown to be performed by members of *Bacteroides*, *Enterococcus*, *Clostridium*, *Bifidobacterium* and *Lactobacillus* and to a lesser extent other genera[285].

There are clear implications for the microbiome and its influence on host dietary responses. These have been observed in individuals with high starch consumption maintaining a higher copy number of the salivary amylase gene (AMY1) than those with low starch diets[286]. Another prime example of how diet and region can be a major factor in defining the community structure was seen in Japanese populations. A marine member of the Bacteroidetes, *Zobellia galactanivorans,* can process porphyrin derived from marine red algae from the *Porphyra* genus. Homologues of the gene from *Z. galactanivorans* were seen to be present in the microbiomes within the genome of *Bacteroides plebeius.* This gene was only found in the microbiome of Japanese citizens and not observed in the American microbiota. It has been suggested that this gene was obtained in *B. plebeius* through horizontal gene transfer (HGT)[287].

The microbiota has also been linked to health problems associated with diet. For example, differences were observed between female monozygotic and dizygotic twins where one twin was considered lean and the other obese. It was seen that the human gut microbiome was shared among family members, however each host's microbiome was seen to vary in specific bacterial lineages present. Obesity was correlated with phylum-level changes in the microbiota, reduced bacterial diversity, and altered representation of bacterial genes and metabolic pathways[288].

Within term and preterm infants, feed type has been shown to influence the composition of the gut microbiome. Increased levels of *Bifidobacteria* and lactic acid bacteria dominate the microbiota of infants fed breastmilk. In contrast, those infants administered formula feeds were seen to be more diverse and had higher numbers of facultative anaerobic bacteria such as Staphylococci, Streptococci and Enterobacteriaceae[289,290,291,292].

### 1.9.2 Immune Response

The microbiome has been shown to have important roles in priming the immune system as the ability to use macronutrients, i.e. dietary nutrients required in large quantities, has been seen to be essential in the production and maintenance of a protective effector immune response. For example, T-cells are fundamental in recognising antigens presented on potential pathogens, and therefore are integral to the recognition and creation of a downstream lymphocyte response[293]. It has been observed that upon stimulation, T-cells exhibit an increased uptake of glucose and amino acids and, conversely, a deficiency in glucose or amino acid uptake has a negative impact on the T-cell functions[294,295].

SCFAs demonstrate the implications of nutrient processing by the microbiota community and host diet in the formation of an immune response. SCFAs are bacterial end products from macronutrient fermentation, predominantly plant polysaccharides that cannot be digested by humans because of the lack of glycoside and polysaccharide hydrolases[269]. The concentration of intestinal SCFAs in the intestinal lumen can be altered depending on the intake of fibre by the host, which in turn affects the composition of the microbiota[296]. In conjunction with providing an energy source, SCFAs have been seen to have a clear influence on the host's immune response. Butyrate at low levels is known to modify the cytokine production of the T-cells and promote the intestinal epithelial barrier integrity[297].

### 1.9.3 Competitive Inhibition

The various different microbiotas found on the human body have been implicated in competitive inhibition of pathogenic strains. One example of this concerns the notoriously persistent gut pathogen *Clostridium difficile*, which has been shown to be extremely resistant to antibiotic treatments[298]. The infection rates for *C. difficile* have been observed to increase in both frequency and severity, therefore a treatment option which is not based on antibiotic administration is considered ideal.

Faecal microbiota transfer, that is using the faeces of a healthy human host to supplement the infected host microbiome, has achieved therapeutic success rates in excess of 90% for subjects who had recurrent *C. difficile* infections[299]. This was not possible with antibiotic administration which could, potentially, have only encouraged the development of highly resistant strains within the affected individual's microbiome and reduced the competitive inhibition within the bacterial community[300].

## 1.10 Establishment and Development of the Gut Microbiome

The establishment of the gut microbiota is a complex process influenced by microbes with external and internal factors of the host. Factors such as the mode of delivery, the type of diet, and host immune responses have been shown to influence the establishment, development, and stability of the community. It was thought that both term and preterm infants were essentially born sterile, however there have been studies disproving this, with evidence of bacterial translocation in utero,[301] and the presence of microbial life in the preterm infants' meconium[302].

More recent studies have started to refute the long held assumption that linked periodontitis with the risk of spontaneous preterm birth, abortion or miscarriage, and with bacterial colonisation of the placenta. This assumption was based on the hypothesis that circulating endotoxin in the maternal blood reaches the foetus through the placenta and initiates an inflammatory cascade, triggering preterm labour[303].

Stout *et al* demonstrated that 27% of 195 infants were seen to have intracellular bacteria found in the placental basal plate[304]. Within the study, it was shown that this bacterial presence occurred in both preterm and full-term gestations. However, there was a greater proportion of preterm infants with confirmation of placental microbiota.

Two studies that could be compared were those of Di Giulio *et al* and Han *et al* who analysed women with spontaneous preterm labour by both culture-based and DNA-based methods (Supp. Table 1)[305,306]. Within term infants Firmicutes, Tenericutes, Proteobacteria, Bacteroidetes*,* and Fusobacteria phyla have all been identified as commensal non-pathogenic bacteria from the placental tissue, which were considered to be a metabolically rich community[307]. In preterm infants the species colonising the placenta were observed to have a greater association with the oral cavity than with the urogenital tract[306].

An important aspect of the development of the normal microbiome is how the maternal microbiota is passed on to the infant. This is understandably promoted by close proximity of the mother to the child and is primarily initiated upon passing through the birth canal, where there is an abundance of bacteria[308].

Directly after birth it is possible to detect the prevalence of *Lactobacillus*, *Prevotella* or *Sneathia* on the infant's skin, oral mucosa, and nasopharyngeal aspirate, in addition to the first meconium[309]. Whilst the vaginal microbiome is predominantly *Lactobacillus* spp.[310], it has been shown that the *Lactobacillus* do not actively colonise the infant intestinal tract, but rather it is the maternal faecal bacteria, namely the Enterobacteriaceae and *Bifidobacteria*, that initially colonise those infants delivered vaginally[156]. In contrast, infants born by caesarean section harbour bacterial communities similar to the skin microbiota of the mother[309].

In a relatively short time frame the vaginally delivered infant develops its first gut microbiome, comprising predominantly enteric organisms. In comparison, those infants delivered by caesarean section were seen to have reduced richness and diversity within their gut microbiome communities up to four months of age[42].

Throughout the first year of life, the infant gut is colonised by blooms of microbes, most of which are anaerobic or facultative anaerobes; this is due to the lower gastrointestinal tract being highly anoxic. Initial colonisation is thought to be predominantly facultative anaerobes such as *Staphylococcus*, *Streptococcus*, *E. coli* and Enterobacteria. These consume oxygen,[311,312] creating an environment suitable for the obligate anaerobes, primarily consisting of Actinobacteria and Firmicutes[313]. The term infant microbiome reaches maturity at approximately three years of age, at which point it becomes relatively stable and resembles that of the adult microbiome[44,311].

Feed types also influence the composition and development of the microbiome, especially during the first six months[314]. During this time infants are exclusively milk-fed and after six months they are given complementary foods[315]. Increases in the proportion of complementary foods in the diet should be made up to 12 months of age, at which age it is expected that infants be exclusively fed solids. These changes have been reflected in studies which have shown that alterations in the microbiota correspond to key changes in the diet of the host[316].

Breast milk and formula milk-fed infant gut microbiotas have differences in the bacterial taxa identified within their communities and their relative abundances. Breast milk-fed infants were observed to have approximately twice as many intestinal bacterial cells, but with lower diversity relative to those infants fed formula feeds[317].

Breast milk, unlike formula milk, is found to be high in human milk oligosaccharides (HMOs), which are broken down into SCFA which promote the growth of *Bifidobacteria* and *Lactobacillus*. Additionally, breast milk contains maternally generated antibodies, immunoactive compounds, growth factors, and antimicrobial enzymes, all of which have a beneficial impact on the infant's health[318]. The probiotic effect of breast milk results in a colonisation pattern that appears to prepare the innate and adaptive mucosal immune phenotype by communication between molecular patterns on colonising bacteria and pattern recognition receptors such as TLR4[319].

In breast milk-fed infants, the dominant Actinobacteria are *Bifidobacterium* species, predominantly *B. breve*, *B. dentium*, *B. infantis, B. longum* and *B. pseudocatenulatum*[292,292]. The majority of the Firmicutes present are lactic acid bacteria in the form of *Lactobacillus*, *Clostridium,* and *Enterococcus*[313,320].

One study has shown that infants fed breast milk had twice the number of *Bifidobacterium* compared to formula-fed infants[317]. Formula feeding has been shown to correlate with greater abundances of *Atopobium*[321], which are linked to additional factors such as antibiotics administered to the mother and delivery by caesarean section. While there is evidence of breast feed infants having a dominance of *Bifidobacterium* species other conflicting studies have demonstrated that formula-fed infants have a dominance of *Bifidobacterium* spp.

compared with breast milk-fed infants[292,321,322]. These conflicting reports demonstrate how difficult it is to attribute species to environmental factors.

Increased abundance of *Bacteroides* spp. and members of the Enterobacteriaceae were also reported in formula fed infants[292,322]. Despite evidence suggesting the importance of *Bifidobacterium* as an initial coloniser of the gut, it was reported to be absent in significant portions of some formula-fed infants' guts[311]. This could be due to influences such as antibiotic administration, or the mixture of breast and formula feeding regimes within the study; however the variability may also be linked to the variation of different formula milks.

Formulas supplemented with prebiotics may account for high levels of *Bifidobacterium* found in many formula-fed infants[323,324]. A recent report on these prebiotics found that they had very similar levels of *Bifidobacterium* relative to those of infants fed human breast milk, and additionally these groups were reported to have ~20% greater relative abundance in relation to traditional, formula-fed infants[325]. The maternal diet may also have an impact on the bacterial abundances of infants when they are breast fed[326].

Between one to two years of age, the infant gut microbiome alters dramatically for a second time as it develops into the stable adult microbiome[327,320]. It was suggested that, although there were clear differences in the microbiome before and after weaning, characteristics of earlier colonisation events (e.g. delivery method and feeds) were still present within the community[321]. However, prior to the introduction of solid food diets, breast fed infants from different geolocations were seen to cluster together, indicating similar community structures irrelevant of location. However, following weaning, infants were seen to cluster into distinct geographic groups[327].

This significant shift was also observed in a cohort of 330 Danish infants, which reported that between 9 and 18 months the infants' microbiomes changed and that this could be correlated to the introduction of solid foods. Specifically, Bacteroidetes and Firmicute species were seen to increase in abundance, whilst there was a corresponding decline in *Bifidobacteria*, *Lactobacillus*, and Enterobacteriaceae[320].

Butyrate producing taxa were also observed to increase in abundance. These taxa are known to be instrumental in the breakdown of indigestible complex plant polysaccharides and starches[328], which are major constituents of human diets.

The final major change in the gut microbiota community occurs between 18 and 36 months. This change results in a stable microbial profile consisting predominantly of Bacteriodetes and Firmicutes. This is a temporal change that is representative of the continued variation in the solid food diet occurring in the infant's early life[320,321,328].

The proportions of Firmicutes and Bacteroidetes were observed to be strongly influenced by diet. This was aptly demonstrated by De Filippo *et al* where they described the different microbiotas of children with very distinct dietary habits. They demonstrated a clear dominance of *Prevotella* in African children, although these taxa were completely absent from Italian children[327].

This was also detailed in a case study which demonstrated that the introduction of peas, formula, and other solid foods resulted in a co-dominance between Firmicutes and

Bacteroidetes, with the introduction of plant polysaccharides being linked to increases in the abundance of Bacteroidetes[328].

The typical adult microbiome usually consists of six or seven bacterial phyla, dominated by Bacteroidetes and Firmicutes[329], and also includes Proteobacteria, Verrucomicrobiota, Actinobacteria, and Euryarchaeota. A five year study of 37 adults suggested that 60-70% of the bacterial strains present in the gut microbiome remained unchanged over the duration of observation[330]. This study also suggested that Firmicutes and Proteobacteria were much more susceptible to perturbations of the community, whereas Bacteroidetes and Actinobacteria are stable. These findings appear to agree with earlier studies using microarray-based approaches[4].

When health declines with age, so too does the gut microbiome, developing increased instability and a decrease in diversity. However, if health remains intact, the microbiota maintains the stability and structure of a young, healthy adult[331].

The most significant factors associated with age related declines in the microbiota health appear to be physiological changes, dietary choices or malnutrition, location (e.g. community-dwelling, hospital duration or long-term care) and the use of prescription drugs and antibiotics[332,333,334,335].

## 1.11 Diseases associated with Microbiome Dysbiosis

Whilst NEC is the subject of this investigation, there are other disorders associated with the microbiome of infants and adults that exhibit similar effects on the gastrointestinal tract. Many of these appear to be influenced by the recent changes in western diets, which have altered the composition of the microbial community and its associated metabolic functions[336].

These changes in the microbial community are considered to contribute to the increased incidence of epidemics in chronic illness throughout the developing world. This includes obesity, inflammatory bowel disease, cardiovascular disease, and *Clostridium* infections[6].

In subjects with obesity, it has been observed that the gut microbiome has a greater proportion of Firmicutes relative to Bacteroidetes[337] compared to lean controls. Models suggest that the community in obese subjects has an increased capacity to harvest energy from the diet, a trait which appears to be transmissible by faecal transplantation[338].

Cardiovascular disease has been linked to an inability to metabolise phosphatidylcholine into the proatherosclerotic metabolite trimethylamine-*N*-oxide (TMAO). This has been shown to result in an increase of TMAO in plasma levels and is associated with an increased risk of cardiovascular complications in patients at risk[339,340]. Metabolites of phosphatidylcholine were associated with macrophage scavenger receptors linked to atherosclerosis. These studies suggested that a potential method to reduce cardiovascular issues in mice could be through dietary changes that promote a gut microbiome with a lower rate of phosphatidylcholine metabolism.

Irritable bowel syndrome (IBS) has been associated with gut microbiota dysbiosis. Treatments directed at the microbiome are sometimes helpful. These include dietary changes, probiotics and antibiotics[341].

Changes to the diet were observed to decrease the severity of symptoms of IBS when subjects avoided fermentable oligosaccharides, disaccharides, monosaccharides, and polyols[342]. These changes subsequently altered the composition of the gut community by reducing the total bacterial abundance along with the structure of the faecal microbiota. This specialised diet was associated with specific stimulation of the growth of bacterial groups thought to have health benefits[12].

*Clostridium difficile* infection is a well-studied example of how a dysbiotic microbiota can result in a devastating human disease. *C. difficile* is a very resilient opportunistic pathogen, which is known to be resistant to antibiotic administration[343]. The symptoms of the infection appear to be related to changes in the human gut microbiome which can be rectified through the use of human faecal transplantation[344].

Treatment by transplantation of a healthy microbiome into a subject with *C. difficile* infection proved to be effective in treating the infection in 90% of subjects. Analysis after the faecal transplantation demonstrated that subjects which previously had *C. difficile* infection now had a microbiota that closely resembled that of the donor. Changes in the microbiome during *C. difficile* infection appeared to promote a metabolome that supported *C. difficile* germination and growth[345].

These studies demonstrate that the microbiota impact on the host's health and wellbeing. Many of the factors that influence the establishment and development of the microbiota are considered important risk factors in the development of NEC.

Investigations into the microbiota are important in our understanding of the implications it may have on the development of NEC. There are clearly many environmental factors that result in highly idiosyncratic community structures. An individual's microbiome appears to be unique and there is a growing interest in the use of metabolomics within the field of forensics[346].

Therefore, it is important to understand that large scale cohorts will be required to establish not only the important factors that influence and shape the premature infant microbiome, but also what the impacts of these are on the occurrence of NEC. The development of high throughput sequencing technologies has now made it possible to perform these large scale comparisons within a reasonable budget.

## 1.12 Analysis of the microbiome

### 1.12.1 Microbial Culturing

Originally, the only means of establishing which bacteria were present within an environment involved the culturing of organisms from samples. This is a particularly difficult means of profiling the gut community for two reasons, the first being that after initial colonisation of the gut, it is primarily populated by obligate anaerobes which require sophisticated equipment to cultivate. Secondly, the exact culturing requirements for many bacteria species within the community are unknown and would need to be established through substantial trial and error experiments.

It has been shown that culture-dependent techniques are able to successfully identify the dominant microbes present in the instances of acute infection, but there is evidence that these techniques fail to identify pathogens due to a lack of growth or because pathogen growth was misclassified as oral flora[347]. This is also exacerbated in metagenomic analysis of taxonomically rich samples as those species that grow readily on cultures, such as *E. coli,* will be easily detected and cultured. In the past this led to *E. coli* being considered a significant member of the gut microbiome, however only since the introduction of culture-independent techniques was it known that Gammaproteobacteria, of which *E. coli* is a member, represent less than 1% of the community[348]. This reduction in *E. coli* is due to the reduction in oxygen that results from their metabolism favouring their replacement by strict anaerobes[349].

The use of culturing techniques to profile communities is both ineffective and inefficient compared with modern culture-independent methods, but these techniques still represent an important and cost effective role in diagnostics. Primarily, this is when the phylogeny of the bacteria is known along with its particular nutrient and culturing requirements. However, culture based approaches are confirmatory techniques as they assume that the investigator knows either what bacteria they are looking for or growth conditions suitable for the bacteria and as such are not a hypothesis free approach that are likely to introduce biases.

### 1.12.2 High throughput Sequencing

High throughput sequencing became particularly prevalent in the second generation of DNA sequencing technologies, at which point parallel sequencing methods were developed, initially by 454 (later purchased by Roche). In particular, the sequencing by synthesis method designed by Solexa and subsequently acquired by Illumina[40] was the one of the most successful means of sequencing designed in this generation.

The fundamental mechanism was the use of adaptor-annealed DNA molecules and, by binding them to a lawn of complementary oligonucleotides, it was possible to then perform solid phase polymerase chain reaction (PCR) to produce neighbouring clusters of clonal populations from each of the individual DNA strands[350,351] (Figure 2). In this manner the process could be parallelised, greatly increasing the amount of DNA that could be sequenced in one run[352]. This process was called ‘bridge amplification’ due to the replicating DNA strands arching over in order to prime the next round of polymerisation. This method uses fluorescent ‘reversible-terminator’ deoxynucleotides (dNTPs) which prevent the binding of further nucleotides until the fluorophore is cleaved away, meaning that sequencing occurs in a synchronous manner[353]. After each cycle the identity of the nucleotide is established by exciting the fluorophore using lasers and monitored by a charge-coupled device. After this enzymatic cleavage of the blocking fluorophore, further synthesis is performed.

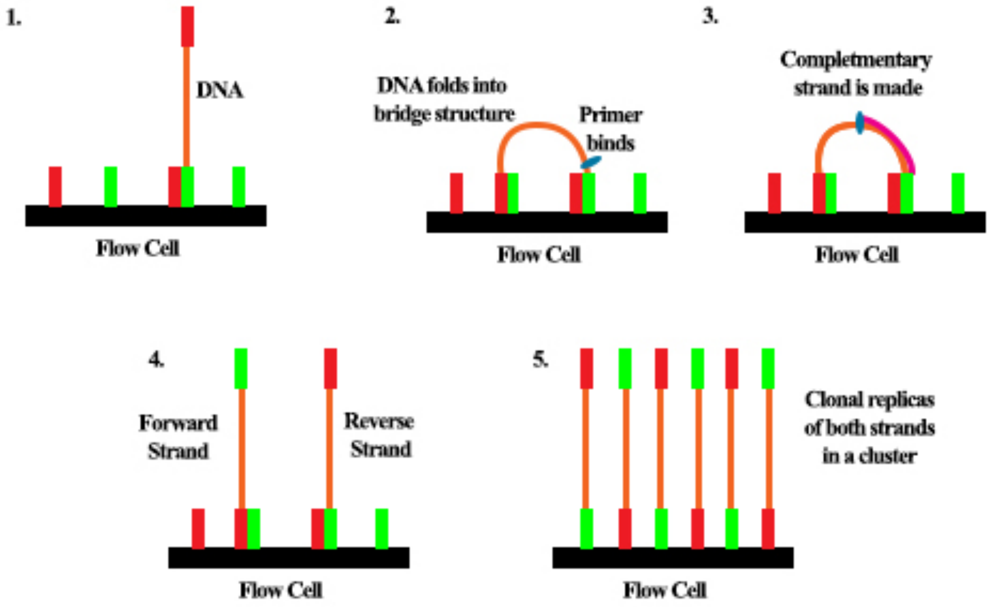

1.
DNA
Flow Cell
2.
DNA folds into bridge structure
Primer binds
Flow Cell
3.
Completmentary strand is made
Flow Cell
4.
Forward Strand
Reverse Strand
Flow Cell
5.
Clonal replicas of both strands in a cluster
Flow Cell

*Figure 2 Diagram of illumina flow cell bridge amplification of DNA sequences used in illumina high throughput sequencing*

Subsequent generations of this technology and the chemistry underpinning it have produced greater read lengths and depths, as well as lower costs both in terms of running the machines and the machines themselves. In particular, the Illumina MiSeq provided a very low point of entry for inexpensive rapid sequencing of samples[354]. These technologies, in conjunction with culture-independent techniques for microbial profiling, have enabled previously unapproachable, large scale, metagenomic analyses.

### 1.12.3 16S rRNA Gene Sequencing

Identifying a universal consensus genomic sequence within bacteria that maintained enough variability to categorise specific operational taxonomic units (OTUs) was also pivotal in the creation of assays that could profile a metagenomic community from a sample. The use of the 16S rRNA gene in this context was first established by Carl Woese in the 1980s, when he demonstrated the use of this gene in the identification of bacterial taxa[355].

Using universal primers on conserved regions of the 16S rRNA gene it is possible to amplify the nine hypervariable regions within this subunit[356] (Figure 3). These sequences demonstrated a high degree of interspecies variability that could be compared with known references within a database. In this manner, it was possible to provide taxonomic identification of sequences amplified from a sample[39].

Analysis of 16S rRNA gene sequences within faecal samples has become the main means of generating non-culture based taxonomic data relating to the composition of the gut microbiota[94,329,357,358,359]. These techniques have enabled scientists to estimate the microbial diversity and dominance within the community and identify specific differences in the GI tract of subjects both in healthy, normal development as well as in diseased states.

Whilst these nine hypervariable regions allow for large scale taxonomic identification of bacteria from environmental samples, they differ in length, position, and taxonomic discrimination[360]. Selection of which region to analyse on the 16S subunit ultimately influences the experimental approach, costs and taxonomic bias,[361,361] and resolution[362].

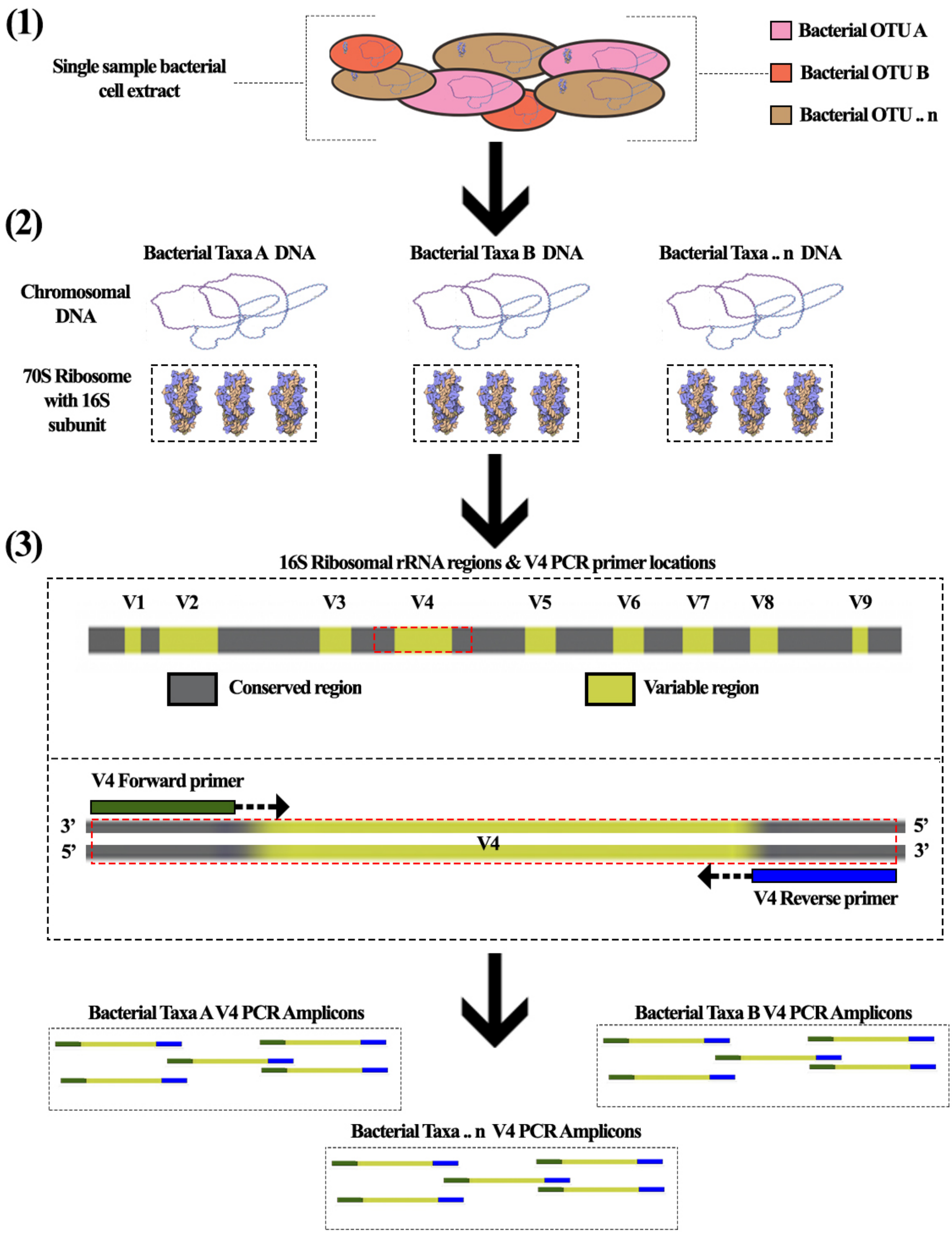


Figure 3 Diagram of the process for taxonomic identification of bacterial taxa (OTUs) from a single sample. (1) Bacterial cell extraction is initially performed on the environmental sample, resulting in a mix of bacterial species. (2) Bacterial cells are lysed and DNA, chromosomal and ribosomal, is purified. (3) PCR Primers designed to anneal to the neighbouring conserved sequences of RNA amplify across the variable region, V4 in this project, producing V4 amplicons that can be used to identify bacterial OTUs when compared against known databases.

### 1.12.4 16S V4 Amplicons: Strengths & Weaknesses

The V4 region in particular has benefits over the other eight hypervariable regions of the 16S subunit. Specifically relating to large scale projects, the V4 amplicons are short enough in length (~395bp) for the Illumina MiSeq machine through the use of paired-end read sequencing (2x250bp), resulting in an overlapped sequence length of 500bp (including adaptors, barcodes, and a known overlap region). It was shown that the topology of trees using the V4 region were seen to have the least geodesic difference to trees constructed using all the sub-regions. Overall these results supported the use of the V4-5-6 regions as the optimal combination[363].This same study also demonstrated that the V4 region tree topology was the most closely related to the tree topology from all sub-regions of the 16S sequence when estimated by their relationship in agglomerative hierarchical clustering. In contrast the V2 and V3 sub-regions were seen to be outliers in this same analysis. V4 rRNA sequence has also been shown to be extremely effective at resolving enteric bacteria such as Firmicutes and Bacteroidetes[364].

By annealing known sequences onto the amplicon it is possible to tag samples with specific DNA barcodes, a method termed multiplexing[41]. This enables the loading of multiple samples onto a single sequence run and dramatically reduces the cost of large scale sequencing projects.

It has been shown that the V4 region of the 16S subunit can provide minimal taxonomic bias when amplified but still fails to annotate many species when compared to a mock community (Supp. Table 2)[362]. However, on a large scale, within a reasonable time frame and cost, this region was shown to ascertain the majority of species likely to be present with the human gut microbiome.

## 1.13 Prospective Cohort Design

One of the most popular means of establishing factors that influence the outcome of diseases or disorders are cohort studies[365]. These involve longitudinal study sampling of a cohort, with a cohort being defined as a group of people sharing a defining characteristic. This characteristic is usually a common event in a selected period, e.g. birth or puberty.

Cohort studies are a form of panel study. These panel studies are a specific design of longitudinal analysis in which the unit of interest is distributed over intervals of time. Within these studies all participants must be at risk of developing the outcome being investigated. This outcome needs to be clearly and unambiguously defined prior to the start of the investigation.

Cases will be defined from the outset of the investigation and usually under highly stringent diagnostic criteria used in the medical diagnosis of the disease. Controls are more complicated in cohort studies, these should closely resemble cases in all important respects, often termed ‘risk factors’, without actually developing the disease. Ideally controls are internal and sourced from subjects from the same time and/or place.

Cohort studies should define the outcomes in advance of the study in a clear, specific and measurable manner. These outcomes should be comparable in both the control and case subjects in every way. Failure to define the objective outcomes leads to uninterpretable results.

The outcome information can be from many sources; in many cases the most information rich data is identified within clinical information. However, the validity of clinical information, especially when it is sourced from multiple sites or collaborators, can be highly variable. When outcomes are from multiple sources it is important that they are graded, e.g. “Definite”, “Probable” or “Suspected”.

It is important to ensure that participants are tracked over the duration of the study, this is vital not just in the sampling period but also after to ensure that subjects do not develop characteristics that that could influence the outcome of the study.

Reporting in cohort studies is often unsatisfactory[365]. An investigator should be able to convince the audience that the exposed and unexposed groups are indeed as similar as possible in all critical factors that influence the outcome. Therefore, reporting the demographic and other prognostic factors for both groups is vital in instilling confidence in the study to establish whether any differences observed are attributable to the dependent variable.

When analysing a dichotomous outcome, i.e. sick or healthy studies, the investigator should provide raw data that is sufficient for the reader to be confident in the conclusion of the study. For incidence rates the value is expressed per unit of time and the relative risks and confidence intervals should be provided. Like any observational experiment, cohort studies should articulate the biases that are present and how these could influence or skew the results of the study.

Prospective cohort design is well suited and has been performed many times in the study of NEC. This is because the study of NEC is appropriate for prospective cohort studies due to its clear definition, albeit with multiple levels of severity, in which infants can be diagnosed by an experienced physician. Also, as NEC occurs almost exclusively in premature infants, who invariably stay in an NICU, it is easy to prospectively enrol infants by approaching the parents prior to birth or complications. This also allows for the establishment of a large control cohort with relative simplicity as those infants found in the same NICU environment are likely to be adequate candidates for control assignment.

The hospital environment in conjunction with NHS  patient records allows for the curation of many data points that could be influential in the development of NEC, for example feeding regime, delivery method or gestational duration. This suits prospective enrolment design by providing a rich, accurate source of data for assignment of controls to NEC cases.

Sample collection was also well suited to this design as the experimental requirements for sequencing the faecal DNA required only a small quantity of faecal matter (less than one gram). This could be easily obtained simply by transferring the contents of the nappy into an appropriate container and storing at -20°C. This meant that large scale sampling could be performed with minimal impact on staff working in the NICU environments and kept in conditions that were relatively cost effective and appropriate for DNA storage until the samples were required.

## 1.14 Hypothesis, Aims and Objectives

The hypothesis proposed was that the microbiome of infants with NEC differs from non-NEC controls around the date of diagnosis. Any differences observed within the microbiome of infants with NEC could be used as potential prognostic biomarkers, whereas differences observed between controls could be indicative of a natural microbiome structure.

The first aim of this project was to identify key risk factors associated with NEC. The first objective was for those that infants that corresponded to worst bells grading III to be selected as cases. Analysis of the control cohort was performed on all known risk factors assigned statistical weighting by medical and statistics experts. The most appropriate controls were then assigned to NEC infants according to those risk factors. This ensured minimal discrepancy between control and NEC infants for those factors most likely to be associated with NEC.

The second objective to this aim was to establish those risk factors that were most influential in the community structure of the gut microbiome. To do this, local contributions to beta-diversity (LCBD) values were used to quantify changes in the community composition and enabled those changes to be interpreted as a function of time. Infants were then grouped according to those risk factors that were influential.

Following control assignment and sample extraction, the second aim was to establish differences between the microbiota of NEC and control counterparts. The first objective of this aim was to analyse the community structure using non-metric multidimensional scaling (NMDS) principal component analysis to observe whether the community structures clustered for case and control infants. The second objective was to analyse taxa that were seen to be significantly different between case and controls over time. The final objective was to utilise machine learning techniques in order to establish the genera that acted as the best predictors of NEC. These techniques were to be applied at the species level to establish any potential pathogens. After subsets were analysed they were to be compared against the whole cohort in order to establish conserved changes across all subsets.

# 2 Methods

## 2.1 Project Initiation

This project was led by Professor Andy Ewer at Birmingham Women's Hospital and the University of Birmingham. This microbiome study built on the study of the neonatal metabolome in neonatal infants developing NEC by Professors Ewer and Probert. Their study suggested new, non-invasive methods that could be utilised in the diagnosis the disease through the detection of volatile organic compounds[366].

## 2.2 Approvals

The study was approved by the West Midlands Research Ethics Committee (11/WM/0078) and sponsored by University of Birmingham. Samples were stored and disposed of in accordance with the Human Tissue Act (2004)[367].

## 2.3 Centres

Eight NICUs took part in this study. They were responsible for acquiring faecal samples from the babies they managed. The NICUs were based in Birmingham Heartlands Hospital (BHH), Birmingham Women's Hospital (BWH), Liverpool Women's Hospital (LWH), Royal Shrewsbury Hospital (RSH), Royal Wolverhampton Hospital (RWH), Sheffield Teaching Hospital (STH), University Hospital of Coventry and Warwickshire (UHCW) and the University Hospital of Leicester (UHL).

## 2.4 Recruitment

All children born before 34 weeks gestation were eligible.

Parents were approached by nursing staff trained in good clinical practice and the aims of the study explained. Parents gave consent, on behalf of their children, for the collection of faecal samples on a daily basis and for the retention of demographic and clinical data. Samples were all subjected to the same SOPs.

Enrolment was initiated in 2010 and closed in September 2014.

## 2.5 Sampling & Storage

### 2.5.1 Sample Collection

Sampling was performed after confirmation of written consent.

Each day stools were removed from the infant's nappy using a plastic spoon; if the stool was in liquid form it was removed using a syringe. The stool was then transferred to a glass vial and the syringe/spoon discarded appropriately. Samples were labelled with a study number, the date-time of collection and then stored in -20°C freezers. If the infants had not passed stool over a 24-hour period, this was noted.

Samples were transported to the University of Liverpool by private courier, on dry ice, in accordance with The Control of Substance Hazardous to Health (2002) as well as the following: Health and Safety at Work Act 1974.; Personal Protective Equipment at work 2002.; The Policy for Effective and Appropriate Hand Hygiene. Birmingham Women's Hospital 2nd March 2010; The Policy for the use of gloves in the clinical area. Birmingham Women's Hospital, December 2008.

### 2.5.2 Sample Storage

In total 13,434 samples were stored at the University of Liverpool in -20°C freezers located in CATII laboratories. Ideally -80°C storage would have limited the effect of degradation of bacterial DNA[368]As the degradation in any one sample should be observed across all samples, comparison within this dataset was considered acceptable. Freezer doors were secured with key or combination locks only accessible to researchers directly involved within the project. Locations of samples were maintained within the MySQL database for future reference. The temperature of each freezer was monitored using an independent thermometer with automated emergency contact capabilities.

### 2.5.3 Sample Disposal

Samples no longer required were anonymised, autoclaved and incinerated according to standards set by the Human Tissue Act of 2004.

## 2.6 Patient Information & Sample Management

### 2.6.1 Clinical and Demographic Data Collection

Demographic data was collected on all infants, including: sex, date of birth, discharge/death date, location of birth and the NICU in which sampling took place. Additional details relating to risk factors for NEC were recorded: birth weight, gestational duration, antibiotic administration, feed type, mode of delivery and severity of NEC were included.

Information that could confound comparisons between disease and controls was also collected, such as medical complications or interventions, surgical procedures and the number of blood transfusions administered. All medical information including backups were anonymised and encrypted on private computers with limited user access, extensive password security and redundancy in the event of theft i.e. remote monitoring, locating and disk formatting.

### 2.6.2 Formatting patient information

All clinical and demographic data were standardised manually whilst dates and times were formatted by automated custom Perl scripts. After formatting all personal and medical information was aggregated into a single text document. MySQL database was created from patient medical data in a relational data (Supp. Fig 1). The parent table, basic_info, contained a unique ID for each patient and some essential information, this was linked to tables specific for antibiotic information, feeding regimes, NEC details, transfer, medical and surgical information and sample details. Formatting this data in a relational database enabled highly specific filtering and fine grain sub-setting of patients based on medical criteria, extent of sampling and experimental data.

### 2.6.3 Defining Critical Risk Factors Associated with Necrotising Enterocolitis & Statistical Weighting

To perform control matching for each NEC subject clinical staff including Prof Ewer, Prof Probert, Sisters Elizabeth Simcox and Rachel Jackson were consulted regarding risk factors associated with NEC from a clinical perspective. Then, each risk factor was assigned a weight with advice from the metabolomic study statistician, Rosemary Greenwood (Table 4). Controls were matched to NEC patients based on the combined score for these risk factors as well as appropriate sampling around the date of NEC diagnosis; this is discussed in detail in Section 2.7.

Table 4 Risk factors and their associated statistical weight assigned by committee with medical staff in conjunction with a statistician.

| Criteria | Calculation | Match Factors Calculation |
|---|---|---|
| **Gestational duration** | Confirmed NEC gestation minus potential control gestation | 10 – i * 10/6 with a minimal factor equal to 0 |
| **Delivery** | Vaginal or Caesarean | 10 if identical and 0 if not |
| **Enteral feeding type** | Mother expressed breastmilk only, formula milk only, mix of both or no feed given | 6 if identical and 0 if not |
| **Hospital** | BHH, BWH, LWH, RSH, RWH, STH, UHCW & UHL | 6 if identical and 0 if not |
| **Birthweight** | Confirmed NEC birth weight minus potential control birth weight | 5 if 0 ≤ i ≤ 20 3 if 21 ≤ i ≤ 50 1 if 51 ≤ i ≤ 75 0 if i > 75 |
| **Intravenous antibiotics (iAB)** | Shared iAB divided by the sum of healthy control's iAB plus sum of confirmed NEC's iAB minus the sum of shared iAB | i * 5 |
| **Gender** | Female or Male | if identical and 0 if not |

### 2.6.4 Filtering Subjects Enrolled in this Study

1,326 subjects were prospectively enrolled. 1,234 had sufficient sampling for consideration as controls or NEC subjects. 70 infants were missing key information regarding risk factors and basic medical information (gestational duration, date of birth, mode of delivery, birthweight, sex and feeds) that were vital in assigning controls and sampling ranges, leaving 1,160 infants within the cohort.

## 2.7 Cohort Selection

In total 88 infants diagnosed with NEC had sufficient medical data to be considered candidates for the study. Of these 44 were considered gold standard NEC cases based on clinical assessment. The changes associated with NEC and the gut microbiome were performed over a 20-day window around the age of NEC diagnosis, +/- 10 days

Each of the NEC subjects were matched to two controls based on weighted risk factors (Supp. Table 3). Confirmed NEC infant BWH120 was diagnosed aged 155 days, only BWH031 was the only suitable control for this subject which was still within the NICU at such a late age. BWH031was also the best candidate for NEC subject BWH258 which was also diagnosed at a relatively late age.

The control infant BHH072 was the only suitable candidate for both BHH095 and BHH107 NEC subjects who were diagnosed very early, aged 2 and 4 days respectively.

BHH061 was also used as a control for the NEC subjects BHH091 and BHH074 because of the late onset of NEC, 50 and 53 days respectively. In total this left 44 confirmed NEC infants with 85 assigned controls.

Across all NEC subjects there was on average 16 potential samples available per day, with a minimum of four samples available on day +4 and a maximum of 28 available on -4. Control subjects were had at least 37 samples (day -9) and at most 65 samples (day +4) (Table 5).

Table 5 The number of samples and subjects with samples for infants with confirmed NEC and their associated controls.

| Age Relative to NEC diagnosis | NEC Samples Available | NEC Subjects | Number of Control Samples | Number of Controls |
|---|---|---|---|---|
| -10 | 18 | 14 | 48 | 48 |
| -9 | 16 | 16 | 37 | 37 |
| -8 | 22 | 18 | 49 | 49 |
| -7 | 18 | 18 | 42 | 41 |
| -6 | 19 | 17 | 49 | 48 |
| -5 | 17 | 16 | 50 | 48 |
| -4 | 28 | 24 | 50 | 44 |
| -3 | 23 | 23 | 61 | 53 |
| -2 | 22 | 21 | 57 | 53 |
| -1 | 19 | 18 | 54 | 52 |
| 0 | 27 | 27 | 58 | 58 |
| 1 | 11 | 11 | 47 | 46 |
| 2 | 11 | 11 | 62 | 54 |
| 3 | 11 | 11 | 53 | 49 |
| 4 | 6 | 6 | 65 | 58 |
| 5 | 12 | 12 | 57 | 54 |
| 6 | 9 | 9 | 54 | 53 |
| 7 | 14 | 11 | 57 | 52 |
| 8 | 14 | 11 | 51 | 49 |
| 9 | 11 | 10 | 59 | 56 |
| 10 | 13 | 10 | 58 | 55 |

Viable samples were then distributed between this project and the affiliated metabolome study by alternate days. This was to aid future comparative analysis of samples using an overlapping time series. Not all subjects had a complete set of samples from the date of diagnosis +/- 10 days, as some infants did not defecate daily.

## 2.8 Laboratory Methods

### 2.8.1 DNA Extractions

The STRATEC PSP Spin Stool DNA Plus kit protocol was used to extract DNA from 982 subject samples (Table 6). 0.2-1.0 grams of faeces was placed into stool stabiliser solution for three days. After three days samples were shaken at 95°C for 5 minutes and the lysate was vortexed with beads. This mechanical and enzymatic lysis was used to yield purer DNA in larger quantities[369].

Table 6 Number of faecal extractions and negative controls performed with the STRATEC PSP Spin Stool DNA Plus Kit.

| Status | Number of Infants | Number of Samples |
|---|---|---|
| Assigned Controls | 78 | 723 |
| Gold Standard NEC | 35 | 228 |
| Suspected - Trial Experiment | 1 | 15 |
| Control - Trial Experiment | 1 | 16 |
| Negative Controls | N/A | 41 |
| **Subject Extractions** | **115** | **982** |
| **Total Extractions** | | **1023** |

### 2.8.2 Manual DI PCR methods

Originally two subjects were assigned for V4 16S rRNA sequencing to test the validity of the protocol and ensure that the methodology was feasible. 31 samples were extracted from subjects UHL022 (16) and BHH004 (15). BHH004 was initially assigned as a confirmed NEC subject however after sequencing this was later revised to a suspected NEC status by clinical staff.

A dual index amplicon library was constructed with a 2-step nested PCR protocol[362] (Supp. Figure 2). The first step selected the V4 region using 25 µL PCR reaction mixture with 1ng DNA isolate, 12.5 µL Q5 High-Fidelity DNA Polymerase (2x), 1 µL NF F515A & 1 µL DR Illumina R706 (3 µM) and 9.5 µL molecular $H_2O$. Primer sequences can be found in Supp. Table 4. PCR amplification was performed on a Mastercycler apparatus (Eppendorf, Hamburg, Germany). The amplification program was 95°C for 120 seconds; 8x cycles of 98°C for 20 seconds, 60°C 15 seconds and 72°C 40 seconds; and 72°C for 60 seconds. Clean-up was performed using Ampure XP magnetic beads (0.8:1)(Roche, Basel, Switzerland) according to the supplier’s instructions and eluted in 15 µL TE.

Second stage PCR was required for indexing and adaptor annealing to the V4 amplicon; (11µL purified DNA, (12.5µL Q5 High-Fidelity DNA Polymerase (2x), 1µL i5 and i7 (TruSeq Dual Index Sequencing Primer, Illumina) (3µM) (Forward and reverse primer sequences can be found in Supp. Table 5). The amplification program was the same as before but for 15x cycles. A second clean-up using Ampure beads was performed as before and the product was eluted in 15µL TE. Amplicon library assessment was performed on the 2100 Bioanalyzer and the High Sensitivity DNA kit (Agilent Technologies, Hewlett-Packard-Straße, Waldbronn, Germany). After normalizing and pooling the libraries, the pools were size selected with Pippen Prep (Sage Science, USA) for the V4 region (432bp).

Sequencing was performed on the MiSEQ Illumina machine using 2x250 paired end reads.

### 2.8.3 Batch DI PCR methods

Samples were normalised to 1ng of DNA in 96-well plates at 0.2ng/µl in 5µl molecular grade $H_20$. First stage PCR master mix (12.5 µL Q5 High-Fidelity DNA Polymerase (2x), 1 µL NF F515A & 1 µL DR Illumina R706 (3 µM) and 5.5 µL molecular $H_2O$) was added with multichannel pipettes. Each PCR plate consisted of 94 DNA extractions as well as positive and negative controls, which were validated V4 products and molecular grade $H_2O$ respectively.

First stage PCR products were purified using Ampure beads on the TECAN Freedom EVO liquid handler. Second round PCR mixture (12.5µL Q5 High-Fidelity DNA Polymerase (2x), 1µL i5 and i7 Dual Index primers (3µM)) was added to a new 96-well PCR plate along with 10.5 µL of Ampure purified first stage PCR product.

Samples were then amplified with the second stage PCR program. PCR products underwent a second round of Ampure clean-up with the TECAN Freedom EVO liquid handler and were eluted, with molecular grade $H_2O$, into a final volume of 15 µL.

After both PCR programs were completed sample concentration was assessed with the QuBit HS dsDNA kit, samples which failed to amplify were re-run through the PCR protocol manually, including appropriate PCR controls. All samples were then diluted to 1ng/µl in 2µl for validation by sequence length visualisation using the Fragment Analyser (Advanced Analytical Technologies, Indiana, USA).

Samples were then grouped into 4-5 sub-pools based on average DNA length, concentration and the proportion of DNA sequences within 300-500bp range in equal quantities. Subsequently, they were size selected within the range of 300-500bp using the Pippin Prep (Sage Science, Massachusetts, USA). Each size selected sub-pool was pooled into a single final volume in equal quantities. The concentration of the 16S rDNA amplicon library was determined (QuBit) and visualised on the Agilent Bioanalyzer before being submitted to the CGR for sequencing.

762 samples were successfully sequenced on Illumina MiSEQ machines using 2x250 paired end reads (Table 7).

Table 7 Number of samples, including PCR positive and negative controls, sequenced on the illumina MiSeq machine

| Status | Subjects Sequenced | Samples Sequenced |
|---|---|---|
| Assigned Control | 75 | 570 |
| Confirmed - Gold Standard | 31 | 175 |
| Control - Trial Experiment | 1 | 7 |
| Suspected - Trial Experiment | 1 | 10 |
| Extraction Negative |  | 24 |
| PCR Negative |  | 13 |
| PCR Positive |  | 13 |
| **Subjects Sequenced** | **108** | **762** |
| **Total Number of Samples Sequenced** |  | **812** |

## 2.9 Sequencing Data Analysis

Pipelines were constructed using bash shell scripts and validated with data from the initial trial experiment. These pipelines included assembling/overlapping reads, quality control, error correction, taxonomic annotation and clustering (Appendix Section 15.1).

All scripts were run from the command line, with Perl v5.10.0, on a Debian 7.0 OS with x86_64-linux-gnu-thread-multi architecture. Programs required for the pipeline include FASTQC v0.11.3[370], PandaSeq v2.5[371], SPAdes v3.3.0[372], blastall v2.2.26[373] and QIIME 1.8.0[374].

### 2.9.1 16S rRNA Sequence Data Processing

The CGR prefixes used in the dual index labelling of 16S rDNA amplicon sequences were removed, sequences were quantified per sample and passed through FASTQC for referencing in quality control. SPAdes Bayes Hammer was then used to perform error correction. Error corrected reads were quantified per sample and passed through FASTQC for reference. PandaSeq performed read overlaps. Overlaps were quantified per sample and passed through FASTQC. FASTQ sequences were then converted into FASTA format using fastq2fasta.

PhiX sequences were removed from FASTA files using blastall. FASTA sequence headers were annotated with sequence run information and formatted appropriately for QIIME. Sequence lengths outside 250-350bp range were removed. All sequences were added into the All_seqs_filtered.fna file.

### 2.9.2 Taxonomic Annotation

The QIIME informatics package was used to process FASTA sequences.

Operational Taxonomic Units (OTUs) were assigned from FASTA sequences using 97% percentage error, 97% sequence identity and a maximum of 1,000 rejects. Taxonomy was assigned with 97% sequence similarity and a confidence level of 0.8 against reference 16S rDNA sequences within the Green Genes database.

Sequences were aligned using the Pynast[375] with pairwise alignment using the UCLUST algorithm with a minimum of 75% ID. Outliers were marked for removal with the gap filter threshold was set to 0.95 and sequence dissimilarity set to 3 standard deviations above the mean.

### 2.9.3 Taxonomic Data Analysis

Taxonomic Analysis was done with the statistics program R using the Phyloseq (v1.16.2)[376], Random Forest (v4.6-12)[377], Vegan (v2.4-4), DESeq2 (v1.14.1)[378], plyr (v1.8.4) and ggplot2 (v2.2.1) libraries. The methods were published by Belen Torondel[379] and described by Umer Zeeshan Ijaz[380].

Local contributions to beta-diversity values calculations are described and added to the metadata table in Appendix Section 15.2. Linear regression analysis was performed using ggplot2 and standard R libraries using the values calculated and placed into the metadata table following completion of this script.

CCA analysis of risk factors is described in the R script in Appendix Section 15.3. Analysis for NMDS plotting, including sub-setting according to metadata variables as well as adding contours and ellipses for continuous data NMDS graphs is described in the R script in Appendix Section 15.4. Log2 fold significant differences between genera and species along with Random Forest analysis are described in the R script in Appendix Section 15.5.

# 3 Cohort Enrolment, Control Selection & Sampling

This project aimed to identify whether the neonatal infant gut microbiota was associated with NEC using a large-scale cohort. It aimed to identify if any changes observed from ten days prior to diagnosis to the date of diagnosis could be used as prognostic signals. Additionally, whether differences after diagnosis could be used to identify a healthy microbiome. Metagenomic research into NEC has predominantly been via small-scale studies, although more recently a few large-scale cohort analyses have been published. However, there has been very little consistency observed between studies with some citing pathogenic strains[209,381] and others indicating that there could be a reduced presence of probiotic taxa[106,382].

It was therefore important that further data from a large-scale perspective could be added to the research community, and that this data be thoroughly analysed alongside associated medical information. By doing so it would be possible to establish consistent trends across multiple locations, age ranges and account for known risk factors defined by the medical staff involved in this project.

Using a large cohort made it possible to observe and report consistent trends within a highly dynamic and unique environment, and be confident in their validity. As described previously, the community of the human infant microbiome undergoes changes until early adulthood, at which point it maintains a relatively stable composition. As the microbiome is highly specific to the host, it is even more important that statistically significant trends are replicated across large cohort populations.

Eight NHS hospitals were used to prospectively enrol 1,326 premature infants and collect their medical data from the UK. Using this cohort, it was possible to describe demographics, identify key risk factors and select a case-control cohort for metagenomic sequencing analysis. As many factors are known to influence the human gut microbiome[383,384], it was important to identify those that were seen to correlate with NEC and establish how effective case-control assignment was with respect to these.

Initial analysis focused on the demographics of the enrolled population of infants in relation to key NEC risk factors described in literature, for example, antimicrobial administration[112], location[59], and birthweight[385]. Factors associated with changing or influencing the infant microbiome were then assessed for an association with NEC. Gold standard NEC subjects were selected from the population based on individual review by the medical team (Section 2.7). Following assignment of controls to NEC subjects, the case-control cohort was statistically analysed regarding risk factors and compared against the enrolled population.

## 3.1 Demographics of Enrolled Population

There were 88 confirmed NEC infants (7.80%) and 902 infants without NEC symptoms (79.96%). An additional 137 infants were suspected of NEC (12.15%) but did not present with the full range of diagnostic criteria to be confirmed NEC cases, and were therefore removed from further analysis and were excluded from control assignment.

### 3.1.1 Association of Gestational Duration with Necrotising Enterocolitis

NEC is primarily associated with premature infants. The gestational durations of case and control infants within the enrolled population were analysed to identify whether there was any significant difference. Infants confirmed with NEC was seen to have significantly shorter gestational durations than infants without NEC (T-test $_{0.95,1}$: t=-10.467, p-value < $2.2x10^{-16}$). On average, infants with NEC were born before 27 weeks' gestation and infants without NEC were born after 30 weeks' gestation (Figure 4). Many existing studies that indicate that NEC is significantly associated with shorter gestational durations[105,386,387], however the significant overlap between the gestations of case and control infants indicated that other factors are likely to influence the onset of NEC.

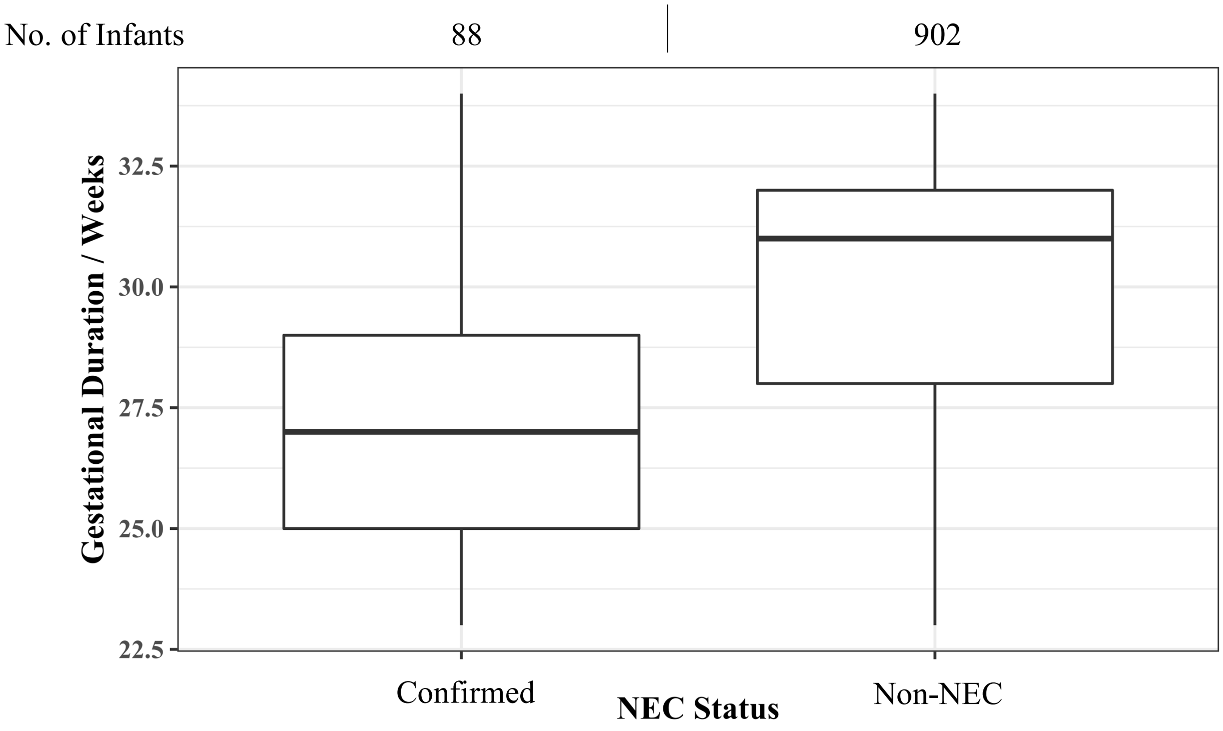


Figure 4 Boxplot describing the significant difference in gestational duration (weeks) for infants with and without NEC.

### 3.1.2 Association of Birthweight and Necrotising Enterocolitis

While prematurity is one of the most associated risks in developing NEC it does not necessarily indicate an infant's physiological immaturity even though the two are closely linked. Birthweight is an important indicator in perinatal mortality and morbidity of infants and NEC incidents rates have been shown to be elevated in extremely low birthweight infants[56]. Infants that are small for their gestational duration are often associated with poor health from the perinatal period up to adulthood[388]. With this population, infants with NEC were seen to have significantly lower birthweights (mean = 955g) when compared to infants without NEC (mean = 1,419g) (T-test $_{0.95,\,1}$: t=-10.743, p-value < $2.2\text{x}10^{-16}$) (Figure 5).

These results support previous evidence that NEC is significantly associated with birthweight[56]. However, as with gestational duration, birthweight is not the only causative factor but could indicate the infant's physiological immaturity when considered with gestation duration.

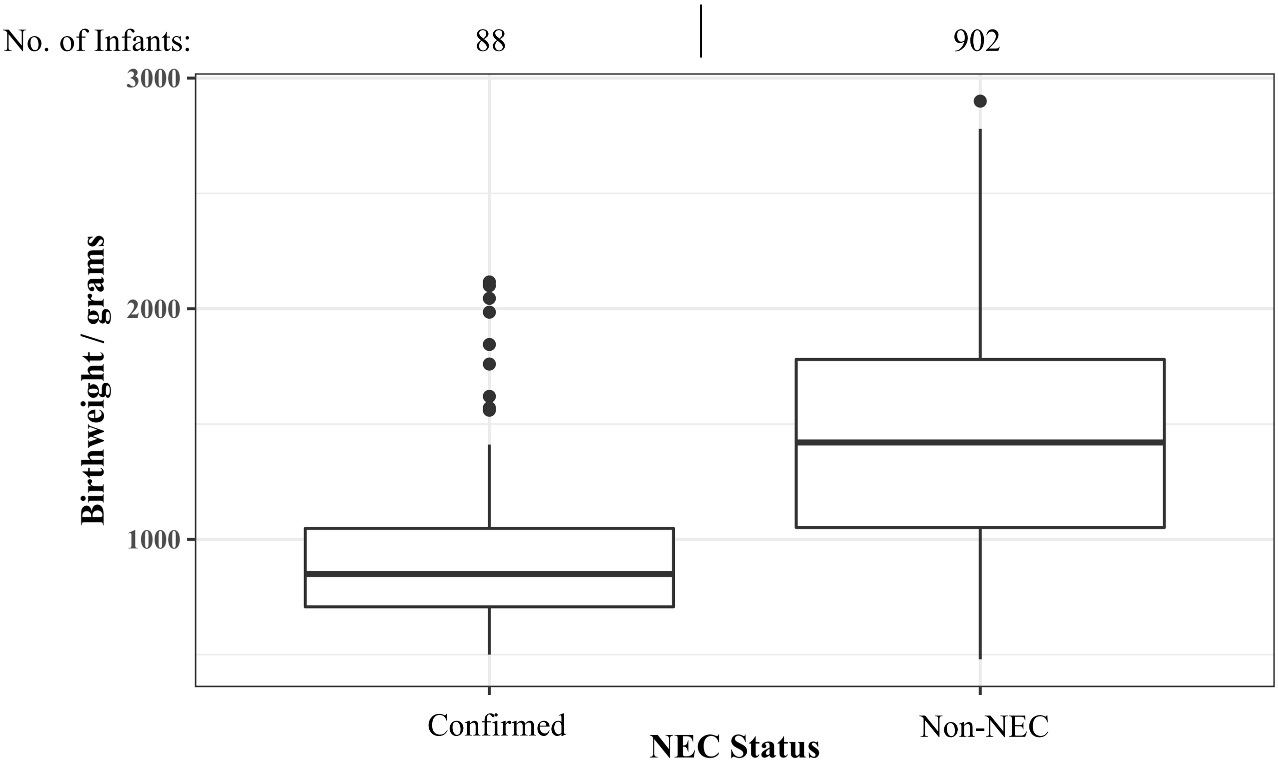


Figure 5 Boxplot describing the significant differences in birthweight relative to NEC status within the sampled population.

### 3.1.3 Association of Delivery Method and Necrotising Enterocolitis

The maturation of the microbiome and the development of dominant community members has been associated with the mode of delivery[155,309,389,390]. Disease states such as asthma[391], allergies[392], type 1 diabetes[393] and obesity[394] have also been observed to occur in infants at significantly different rates depending on the delivery method. Therefore, statistical analyses were used to establish any association between the incidence of NEC and the mode of delivery.

NEC frequency was significantly greater in infants delivered vaginally compared with those delivered by caesarean section. This suggested that the disease status of an infant was not independent of the mode of delivery ($\chi^2_{0.95,\,1} = 6.97$, $p = 8.29\text{x}10^{-3}$, with Yates correction) (Figure 6).

| | Caesarean Delivered | Vaginally Delivered |
|---|---|---|
| NEC Infants | 41 | 47 |
| Non-NEC infants | 556 | 346 |

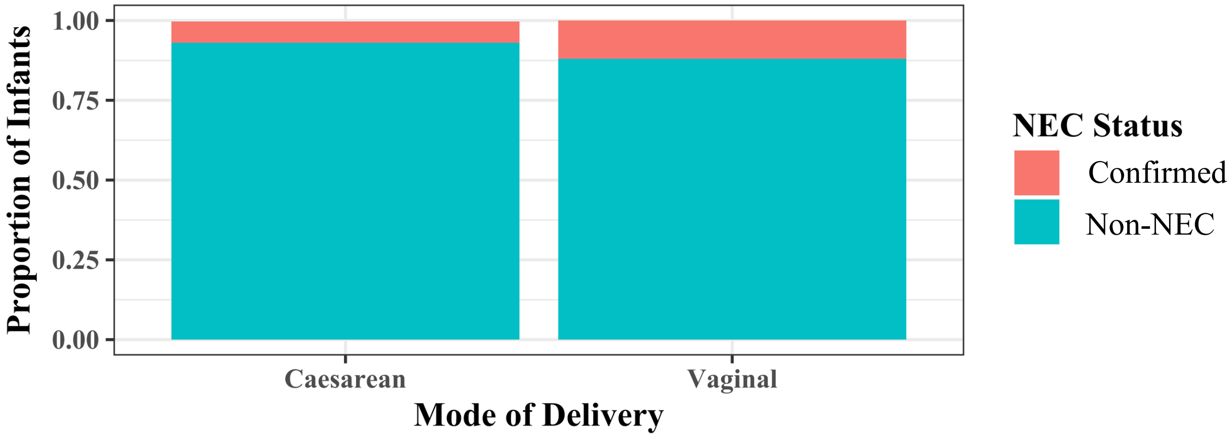


Figure 6 Stacked barplots describing the proportion of infants with and without NEC relative to the mode of delivery.

### 3.1.4 NICU Association with Necrotising Enterocolitis

As many outbreaks of NEC has been reported previously[196,203,381] it was important to establish any biases observed across the recruitment sites that could be indicative of outbreaks or sites with a significantly greater risk of infants developing NEC. The proportion of NEC cases across all locations (Table 8) was observed to be significantly different ($\chi^2_{0.95,\,7} = 22.32$, p-value = $2.24\text{x}10^{-3}$, with Yates correction) but the accuracy was limited by the small number of cases in LWH and RWH. To account for the large differences in subject populations from each unit, the Marascuilo procedural analysis was performed; this method identified no significant difference between any two-way NICU comparisons (Supp. Table 6). Whilst it could be argued that there was an association with the unit and the incidence rate of NEC this is also likely to be influenced by the enrolment rate of infants within the NICU. That is, very proactive units may be enrolling infants at earlier ages whereas some units may enrol only severely ill subjects. Therefore, the staff involved in patient recruitment are likely to be a major contributing factor in the enrolment rates at each unit. As these cannot be accounted for, the implication of an association across units involved is limited.

Table 8 The number of infants and proportion of confirmed NEC cases per NICU

| Infants | BHH | BWH | LWH | RSH | RWH | STH | UHCW | UHL |
|---|---|---|---|---|---|---|---|---|
| Total | 101 | 253 | 44 | 92 | 51 | 150 | 105 | 194 |
| Without NEC | 93 | 214 | 43 | 88 | 49 | 135 | 98 | 182 |
| With NEC | 8 | 39 | 1 | 4 | 2 | 15 | 7 | 12 |
| Proportion with NEC | 0.079 | 0.15 | 0.023 | 0.043 | 0.039 | 0.10 | 0.067 | 0.062 |

### 3.1.5 Association of Enteral Feeding Regimes and Necrotising Enterocolitis

Numerous studies have shown human breast milk to be associated with the incidence rate and severity of NEC[165,395,396] as well as providing long term beneficial impacts that reduce the incidence of other disorders such as respiratory tract and gastrointestinal infections, allergic diseases, coeliac disease, and inflammatory bowel disease[249]. Feeding regimes are also known to influence the normal development of the infant microbiome [42,397] and as such it was considered important that the study population was tested for any significant association with NEC.

Most infants with NEC were exclusively fed human breast milk (47.7%) or a mixture of human breast milk and formula milk (48.9%). The remaining 2.3% were fed formula milk exclusively with exception to a single infant who was not administered any enteral feeds during the study.

Most of the non-NEC infants were fed a mixture of human breast milk and formula milk (61.5%). 29.4% were fed human breast milk only, and 8.87% were exclusively fed formula milk. Two infants (0.2%) were not recorded to have been administered any enteral feeds. There was no significant association between NEC and feeding regimes ($\chi^2_{0.95,\,1} = 2.85$, $p = 0.092$, with Yates correction).

However, there is a limitation in what we can test with regards to the feeding regimes. As there was not a complete record of administration times and dates for all infants it was not possible to know the precise proportions for infants on mixed feeding regimes. That is, it was not possible to establish if an infant was administered breast milk 99% of the time and formula 1% or any variation of this ratio. Nor was it possible to know whether the infant was administered only one type of feed up to and over the duration of sample, to then be switched to another feed thereafter.

### 3.1.6 Association of Antimicrobial Regimes and Necrotising Enterocolitis

Administration of antimicrobial regimes are considered routine for most infants in NICUs however prolonged exposure to antimicrobials has been shown to be associated with NEC[125,172]. As premature infant microbiomes are in the early stages of establishing and culturing a natural, beneficial microbiota it was considered important to understand any association with the onset of NEC and the extent of antimicrobial regimes, especially as antimicrobial compounds have been shown to influence the balance of the gut microbiota[398].

74 infants within the study cohort had no recorded antimicrobials administered (7.47%), one of which was confirmed with NEC. On average, significantly more types of antimicrobials were administered to infants with NEC relative to the rest of the population (Average number

of Antimicrobials ~ NEC status, T-test $_{0.95, 99.3}$: t = 14.86, p-value < 2.00x10$^{-16}$), with NEC infants having on average twice as many (six unique antimicrobials) as infants without NEC (Figure 7).

This association with the number of antimicrobial regimes could be linked to preventative measures often taken in treating and reducing the severity of NEC symptoms[242]. However, as antimicrobials have been shown to affect the microbiota[399] and, importantly, the short-term recovery of the infant microbiota[400] it was considered important that this is factored when comparing case and control cohorts.

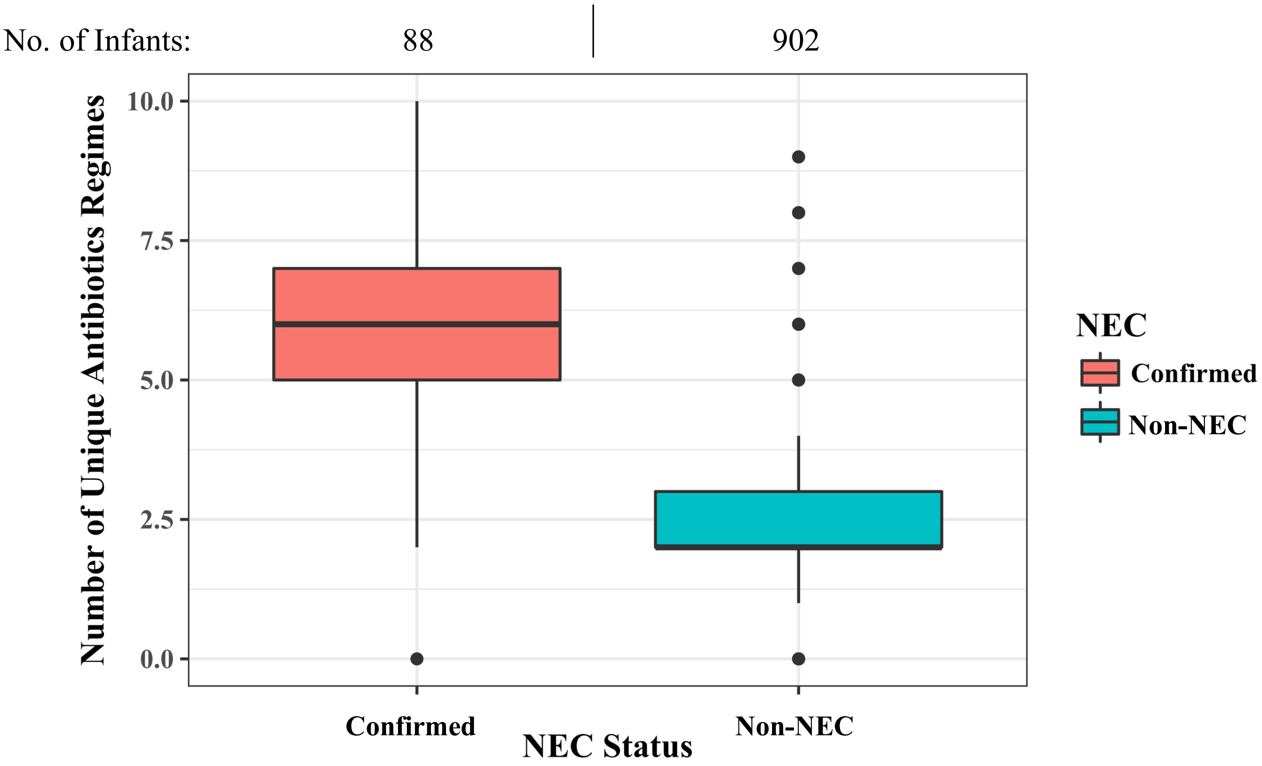


Figure 7 Boxplot describing the distribution of antimicrobial regimes within infants with and without NEC.

### 3.1.7 Age Necrotising Enterocolitis was Diagnosed

The mean age of NEC diagnosis was 21 days, with the maximum age being 79 days and the earliest at one day old. There was no relationship between the age at which an NEC subject was diagnosed with NEC and gestational duration (Linear Regression: $F_{1,86} = 0.84$, p-value = 0.36, $R^2 = 1.80x10^{-3}$), birthweight (Linear Regression: $F_{1,86} = 2.00$, p-value = 0.16, $R^2 = 0.11$) , NICU at which sampling occurred ($ANOVA_{0.95, 7}$: F-value = 0.359, p-value = 0.92), feeding regime ($ANOVA_{0.95, 3}$: F-value = 2.46, p-value = 0.068) or mode of delivery (either simplified $T\text{-test}_{0.95,83.6}$: t = -0.90, p-value = 0.37 or in detail $ANOVA_{0.95, 7}$: F-value = 0.41 p-value = 0.90). This suggests that the onset of NEC was not correlated with any of these factors.

Previous literature suggests NEC on average is diagnosed at 12.5 days of age[27,33,59,110,238,401]. However, these infants were sampled from modern NICUs with specially trained staff, which could be associated with the delayed onset of NEC within this cohort when compared to studies spanning the last 40 years from facilities of varying locations and standards in preterm care.

### 3.1.8 Mortality Rate of Infants within the Cohort

Of the 990 infants with all the key demographic information, 45 died before discharge from the NICU unit (4.55%). The mortality rate in those who had NEC was 26% (23/88 confirmed subjects). This was within the expected range between 15 and 63%[402,403]. 22 infants that died had no NEC diagnosis (40.74%), of those infants one died from a septic ileus, one from meningitis, and one from a congenital heart disease (tetralogy of Fallot). The remaining 19 infants were not observed to have any medical or surgical complications and the reasons for their death are unknown. There were no annotations that indicated other factors, medical or surgical, which contributed to the death of infants with NEC and all were considered to have died from symptoms of the disease.

#### *3.1.8.1 Mortality Rates and Gestational Duration*

Infants with confirmed NEC status that died before discharge were also observed to have significantly lower gestational durations (mean = 184 days) relative to those that were successfully discharged (mean = 193 days) (T-test $_{0.95,\,46.32}$: t = -2.39, p-value = 0.021). No significant difference was observed between the gestational durations of deceased infants with NEC relative to those without NEC (T-test $_{0.95,\,42.01}$: t = 0.27, p-value = 0.021). This shows that there is a negative correlation between the gestational duration and the mortality rates of infants, and that the rate of mortality is irrespective of NEC status.

#### *3.1.8.2 Mortality Rates and Surgical Intervention*

64 infants underwent surgical procedures; of these, 45 were confirmed with NEC, ten were suspected of NEC, and nine did not have any NEC symptoms. One of the nine infants that were not diagnosed with NEC and underwent surgery died before discharge. This infant had undergone both laparotomy and stoma operations.

The mortality rate of infants who were diagnosed with NEC and underwent surgery was 38% (17/45). Ten infants underwent surgical procedures; six had both laparotomy and stoma operations; one had end-to-end anastomosis; two had stoma operations only; and for one infant, there was not any recorded information on the surgery performed. Upon visual inspection seven infants were deemed to have NEC that was too severe for surgical intervention.

#### *3.1.8.3 Mortality Rate & Mode of Delivery*

Infants diagnosed with NEC were observed to have significantly higher mortality rates when delivered vaginally relative to those delivered by caesarean section (T-test $_{0.95,\,3}$: t = 3.53, p-value = 0.04), although this was marginal.

As birthweight and gestational duration had already been associated with mortality within the NEC cohort, the vaginally delivered infants were assessed for associations so see whether these factors were confounding the mortality rate. However, no significant difference was observed for gestational duration (ANOVA $_{0.95,1}$: F-value = 2.31, p-value = 0.14) or birthweight (ANOVA $_{0.95,1}$: F-value = 0.57, p-value = 0.45). This was also true for the mortality rates of NEC infants delivered by caesarean section; gestation (ANOVA $_{0.95,1}$: F-value = 0.86, p-value = 0.36); birthweight (ANOVA $_{0.95,1}$: F-value = 2.53, p-value = 0.12).

## 3.2 Analysis of Case-Control Cohort Selection

44 infants diagnosed with NEC met clinical criteria for control selection; of these, 37 had sufficient sampling. These were assigned with 78 controls based on NEC risk factors (Table 4) and sample availability. Infants were matched based on risk factors to reduce the impact of these factors on further analysis.

Following assignment of controls to NEC subjects risk factors were still observed to be significantly different; gestational duration (T-test $_{0.95,\,72.86}$: t= -3.81, p-value = 2.92x10$^{-3}$), birthweight (T-test $_{0.95,\,78.19}$: t= -3.62, p-value = 5.10x10$^{-3}$), feeding regimes (T-test $_{0.95,\,7}$: t= 2.40, p-value = 0.048), infants per NICU (T-test $_{0.95,\,15}$, t = 5.00, p-value = 1.60x10$^{-4}$), the number of antimicrobials administered (T-test $_{0.95,\,73.83}$, t = 6.6028, p-value = 5.40x10$^{-09}$), and gender (T-test $_{0.95,\,3}$: t= 4.87, p-value = 0.021).

## 3.3 Discussion

### 3.3.1 Demographics of Enrolled Population

Within our population, NEC was significantly associated with gestational duration, birthweight, mode of delivery and antibiotic administration. However, due to a lack of administration dates it was not possible to account for antimicrobials with metagenomic data analysis. Neither breast feeding nor antimicrobials could be considered in cause-effect models as infants with NEC are known to be treated through antibiotic administration[55,7] and encouraged to feed on human breast milk[404,249]. This would artificially inflate the number of NEC infants being fed breast milk and the number of antimicrobials administered to them.

NHS statistics for 2015 – 2016 suggested that the incidence rate of NEC in England and Wales was 509 infants per 100,000. This was much lower than the incidence rate calculated in this study (7,808 per 100,000), however this could have been due to the exclusive use of the NICUs for sampling in this study. These are highly specialised facilities with specialised staff and therefore probably have a greater rate of high-risk infants relative to the hospitals surveyed in the NHS statistics.

One weakness in particular was a lack of information regarding the deaths of 19 infants who did not have NEC. There were no complications annotated which suggests that the medical information for these infants was not as thorough as it should have been. This may have been due to difficulties in obtaining the data or approaching the family to ensure that the data could be contributed to the project. These subjects and their samples were removed from further analysis

### 3.3.2 Case-Control Cohort Selection

In selecting for controls prioritisation was given to gestational duration and birthweight based on the significant association identified within this population. Given two infants with the same gestational duration, there was an increased morbidity associated with decreasing birthweight, which indicated that in addition to gestational duration, birthweight was also a significant indicator in the overall maturity of an infant.

An important priority of this project was the understanding of antimicrobial administration regimes and their association with NEC. It was hypothesised that the actions of antimicrobials would be likely to influence the community structure of the microbiome, therefore high weighting was assigned to direct matches of each antimicrobial NEC and control infants were administered. However, over the course of the project it was not possible to retrieve administration dates for each antimicrobial an infant was assigned. This was a fundamental limitation associated with the scale of the project.

Due to the high association of NEC with gestation, birthweight, mode of delivery, and antibiotic administration, other match factors were seen to be have a greater statistical difference within the NEC-control cohort relative to the overall population. Whilst matching of risk the factors defined by clinical staff reduced the impact of the risk factors in the studied samples there was a limitation in using too many match factors, that is all match factors were still seen to be significantly different, whereas if infants were selected based on a single match factor it would be possible to control for that factor.

### 3.3.3 Conclusion

After researching the literature for factors that were associated with NEC we sought to test the enrolled population of premature infants for those risk factors to establish a case-control cohort according to those factors seen to be significantly associated with NEC. It was also important to control for factors known to be influential in the development of the microbiome. This method aimed to increase the ability to identify differences in the gut microbiome that were associated with NEC and a product of differences in risk factors.

The primary risk factors described in the literature to be associated with NEC are gestational duration, birthweight, feeding regime and antimicrobial administration. These were statistically tested within this cohort population and found to be significantly associated with NEC.

Following case-control specific matching, the significant difference of the case-control cohort for key factors associated with NEC and influencing the gut microbiome composition was seen to be reduced relative to the population. Due to accounting for multiple risk factors NEC subjects were still observed to be significantly different for the risk factors, but this significance was reduced. Limiting the number of match factors could potentially reduce the difference of a given factor, and sub-setting of case-controls would also be able to account for this in some circumstances such as mode of delivery or feeding regime.

It was not possible to account for an important factor that influences the microbiome composition; antimicrobial administration. This was due to a lack of administration dates for all subjects within the case-control cohort. Thus, it was not possible to integrate this factor into future analyses. Further research should focus on controlled sampling based on antimicrobial administration, though performing this in a large cohort would be challenging so these might have to be limited to small scale projects.

# 4 Changes in Community Composition in Relation to Necrotising Enterocolitis and Associated Medical Data

The bacterial community composition has been shown to be associated with the onset of NEC[102,110], rather than an individual taxa i.e. a pathogen. However, these studies were focused on single sites and with small sample numbers per subject, which limited the assumptions that could be made in the absence of further studies that could reinforce these findings. With recent advances in 16S rRNA gene profiling of the microbiome, it has been possible to focus research on the intestinal colonisation of preterm infants at larger scales.

Multiple studies have demonstrated that the preterm infant gut has significantly reduced abundances of beneficial species and diversity, as well as containing a greater proportion of potential pathogens relative to full term infants gut microbiomes[110,104,405,406]. Therefore it was important to establish those changes that occurred in preterm infants at the community level before and after the onset of NEC as well as addressing how those changes differed from infants that did not develop NEC.

Initially the gut microbiota undergoes a dynamic colonisation pattern as the intestinal tract develops to maturity following the introduction of pioneering bacteria after birth[329]. Colonising bacteria are derived from the maternal microbiota (vaginal, faecal, mouth, skin, human milk) as well as the environment[407,408,409] and develop into a community structure shaped by feeding regimes. Infants fed breast milk were seen to have significantly greater diversity within the gut microbiomes compared to infants fed formula milk[64]. This community structure changes substantially from birth to weaning and then from weaning to early adulthood[410].

It was important to establish how key risk factors were associated with differing patterns of colonisation and the impact these had on NEC and control samples comparisons. Once those differences were established appropriate subgroups of samples from NEC and control subjects could be compared. Beta-diversity metrics quantify the diversity of an individual sample relative to the diversity in all the other samples within a population, enabling the association of trends in the community structure for groups of samples and match factors. By controlling for the trends observed within match factor subsets it was then possible to identify taxa that were significantly different in abundances relative to the NEC status of an infant (Chapter 5).

This chapter aimed to establish those match factors that were significantly associated with changes in the development of the gut microbiome and if samples would cluster according to their community composition and the NEC status of the infants within the subset. Additionally, this enabled the detection and visualisation of unique gut communities associated subsets of samples within a population.

## 4.1 DNA Extractions, V4 PCR Amplification, Sequencing Depth and Success

246 successful DNA extractions were performed for 41 of the 45 gold standard NEC infants, and 626 sample extractions were successfully performed on 82 assigned controls. Some NEC subjects' samples were not sufficiently large for the isolation of utilisable quantities of DNA (quantified by QuBit). 228 samples from 39 NEC infants were successfully amplified using dual index V4 PCR protocol alongside 436 samples associated with 68 controls. Some sample DNA did not amplify as well as expected; this was due to the relatively harsh DNA extraction procedures that included vortexing and heating. As some samples did not contain much solid faecal matter, it was likely that the DNA was fragmented to such an extent that amplification was not successful. 225 amplicon preparations for the 39 NEC infants were successfully sequenced alongside 431 amplicon preparations for the 68 controls.

### 4.1.1 Control Group Sampling Depth

26 control groups, which were represented by a single NEC subject and two assigned control subjects, had at least three samples successfully sequenced for every subject in the group. Ensuring at least three samples per subject provided statistical weight to longitudinal analysis. In total these groups corresponded to 157 samples from 25 NEC subjects, and 313 samples from 44 controls. These groups were considered robust enough for thorough profiling across the 21-day time frame around the date of diagnosis.

## 4.2 Summarising Differences in Community Structure: Local Contributions to Beta-Diversity

Local contributions to beta-diversity (LCBD) values summarised the changes in the ratio of regional and local species diversity for samples. These values represented the beta-diversity of a sample (local)  by comparing it's diversity to all the other samples in the population (regional)[411].

Changes in LCBD values indicated whether the local species diversity was altered relative to the regional species diversity. This method normalised changes that occurred across the entire population and allowed the identification of changes missed through the more traditional analysis such as proportional abundances associated with rarefied samples.

Additionally, using LCBD values allowed for the data to be partitioned according to factors of interest and to test the hypotheses that the beta-diversity of samples within different subsets differed between case and control subjects. By ensuring that regression analysis was factored with time it maintained statistical viability when making assumptions across the sample population[412], therefore all statistical analysis of the cohort was analysed as a population relative to time at sampling (Figure 8).

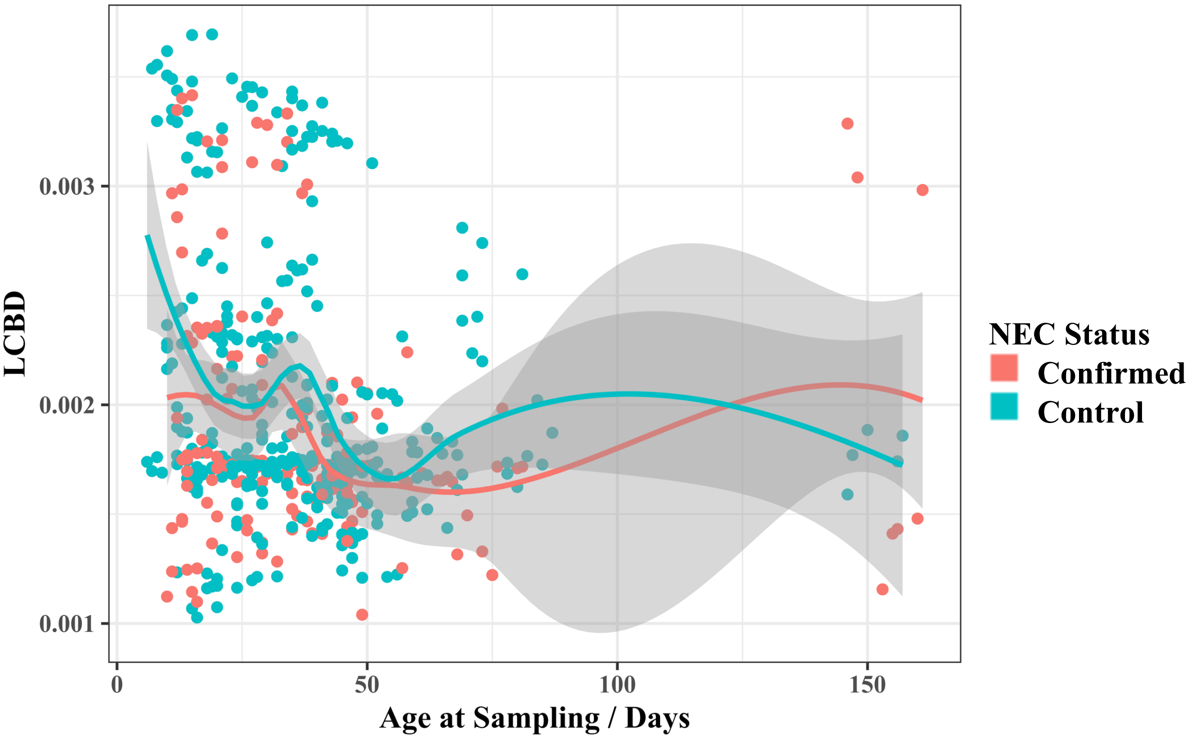


Figure 8 Regression analysis of LCBD values relative to the age at sampling, coloured by NEC status.

Controls and NEC subjects were seen to have a greater difference in the beta-diversity of samples at earlier ages with controls having elevated LCBD values prior to 10 days of age. Between day 20 and 25 control samples were observed to have a decrease in LCBD values until they reached similar values observed for NEC samples. Between 28 and 45 days both NEC and control samples exhibited a similar peak in LCBD values, this peak occurred earlier in NEC samples (~28 days) compared to control samples (~32days). By 50 days of age both groups reached the lowest LCBD values observed. Subject and sampling density after 100 days of age was considered low to be statistically viable and these samples were subsequently removed from further analysis. Over the entire course of sampling, the error margins between NEC and controls (denoted by the grey shading in Figure 8) were never distinct. This indicated that at no point was there a clear distinction in the LCBD values of samples between case and control subjects.

### 4.2.1 Subset Regression of Match Factors with Beta-Diversity

Subset regression analysis was used to identify the factors associated with changes in LCBD. Antibiotics were omitted from this analysis due to the lack of administration dates for most subjects. The regression models aimed to account for most of the variance whilst using the minimum number of factors, thereby creating a minimal model that accounted for most of the variance associated with sample LCBD values. Only models utilising the age at sampling were considered to ensure statistical viability[412].

The best model identified included age, gestational duration and feeding regime (Table 9), these factors accounted for the most variance associated with LCBD values (Linear Regression: LCBD ~ Age + Gestation + Feeding Regime, p = 1.33x10$^{-15}$) (Figure 9). Subset analysis with these factors was used to identify how they were associated with changes in the beta-diversity of samples.

Table 9 Coefficients for the optimum, minimal subset regression model identified between LCBD and match factors (excluding antibiotic administration). Ordered by significance. *** = p-value less than 1x10$^{-3}$.

| | Estimate | Std. Error | t value | Pr(>\|t\|) | |
|---|---|---|---|---|---|
| (Intercept) | 2.70x10$^{-3}$ | 3.34x10$^{-4}$ | 8.079 | 1.67x10$^{-14}$ | *** |
| Formula & Breast Milk | -7.55x10$^{-4}$ | 1.06x10$^{-4}$ | -7.114 | 1.38x10$^{-12}$ | *** |
| Breast Milk only | -6.12x10$^{-4}$ | 1.11x10$^{-4}$ | -5.487 | 4.70x10$^{-10}$ | *** |
| Age at Sampling (days) | -6.27x10$^{-6}$ | 1.75x10$^{-6}$ | -3.581 | 1.07x10$^{-3}$ | *** |
| Gestational Duration (days) | 8.45x10$^{-7}$ | 1.51x10$^{-6}$ | 0.558 | 0.42 | |

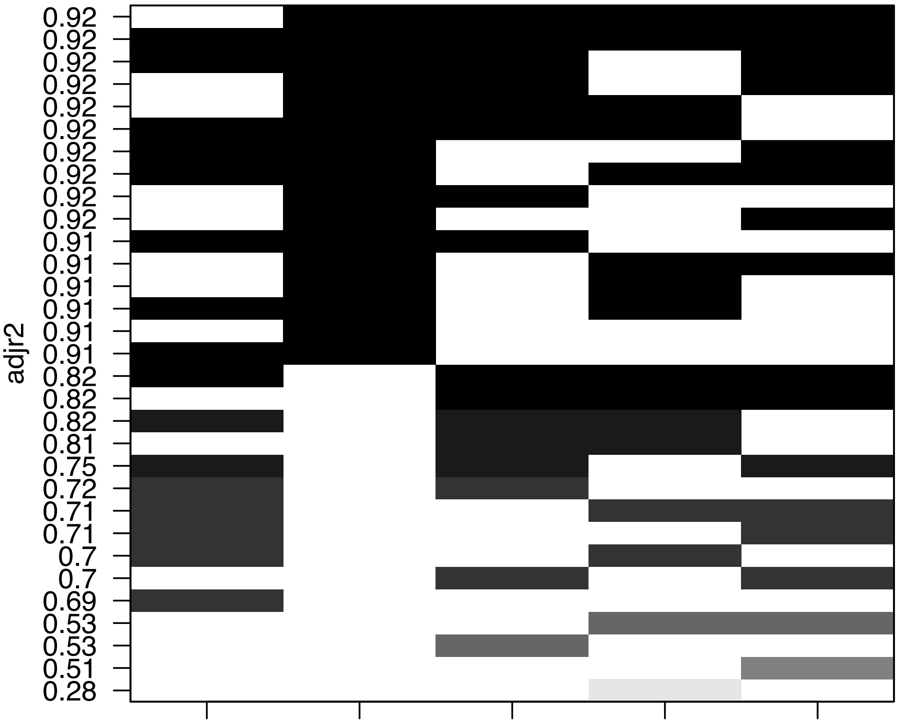


Figure 9 Subset regression summary based on adjusted $R^2$ values. Blocks indicate the inclusion of a factor into a model. Opaqueness represents how important that factor is. Age was to be maintained within the model to ensure statistical viability. $R^2$ value represents the proportion of the data explained in the subset model.

### 4.2.2 Association of Feeding Regime and Community Composition over Time

Feeding regime was the most influential factor in the regression model therefore subsets of samples from each feeding regime were plotted separately to identify how the trends differed over time (Figure 10). Infants exclusively fed formula feeds were considered to lack statistical viability due to the low number of subjects and samples and were not explored in detail.

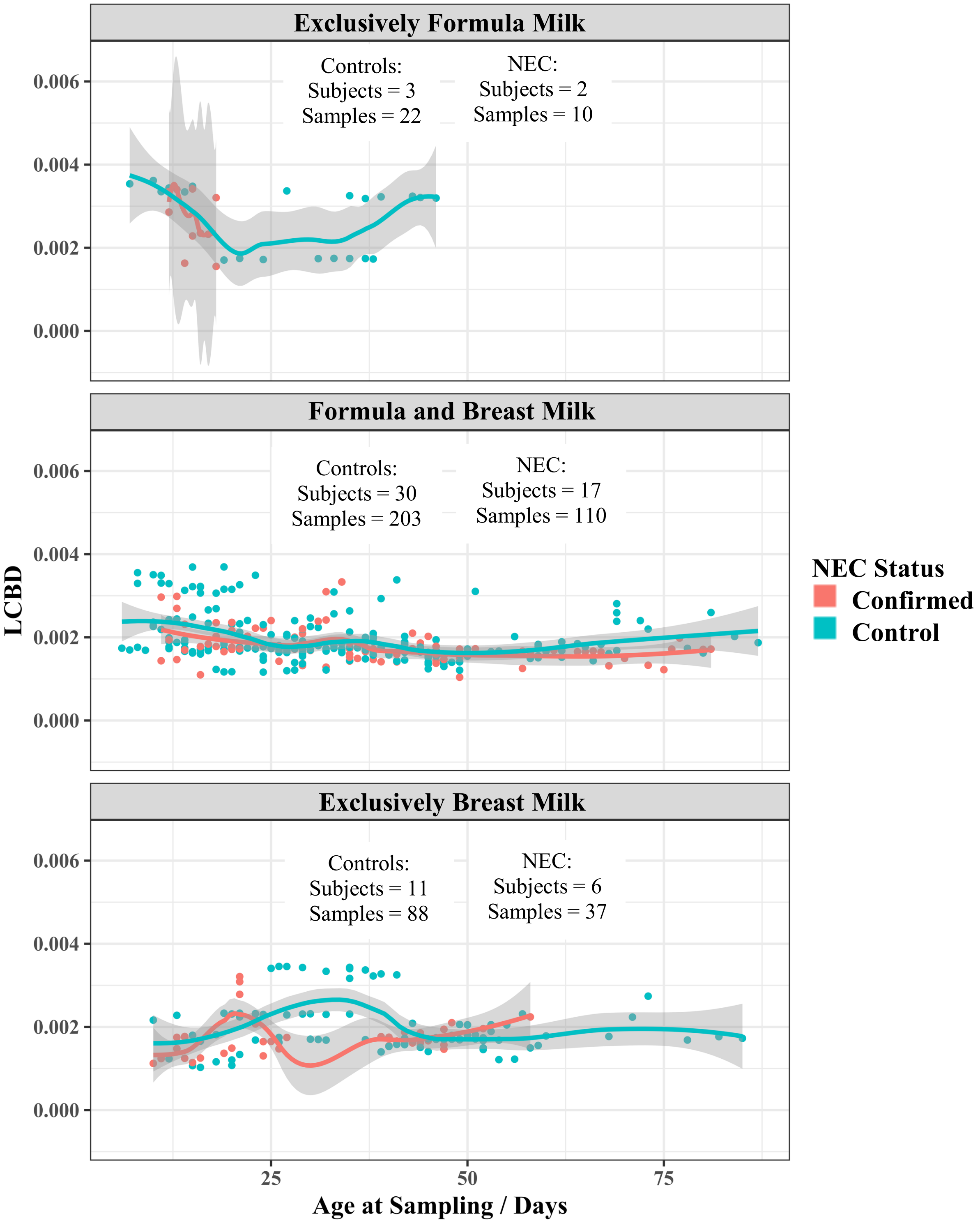


Figure 10 Changes in beta-diversity (LCBD) in relation to age, feeding regime and NEC status.

There were clear differences in the trends of regression plots from infants fed both formula and breast milk compared to infants fed breast milk exclusively. The mean LCBD values of infants on a mixed feeding regime were stable over time, with NEC and control samples showing very similar values over the duration of sampling. This suggested that over time there was little evidence for a taxa altering the sample's community composition, relative to the other samples. For infants fed breast milk exclusively the initial LCBD values were greater for control samples, however NEC samples showed a steeper increase from day one. By day 10 NEC samples had a greater mean LCBD value. NEC samples subsequently declined from day 20 to day 30 but showed an increase up to day 40, after which they stabilised. Control samples, on average, showed a slower initial rate of increase up to ~38 days of age followed by a decline to day 40. After day 40 all samples from both subsets were at similar LCBD values (~0.002).

As gestational duration was identified as a key factor within the subset regression analysis it was factored into the feeding regime subgroup analysis (Figure 11). Gestational is inherently associated with maturity and therefore if there is bias in the gestation of infants with the subgroups this could be create a bias in the comparisons made and distort the differences observed.

Control subjects born at later gestational durations were sampled at earlier ages regardless of the feeding regime of the subject. The gestational duration of infants with NEC were evenly distributed relative to the age at sampling, however average gestational age was lower relative to control infants within the subsets. Since NEC and control subject samples were observed to have the same trends in LCBD value changes relative to time across all subgroups these changes were considered independent of gestational duration.

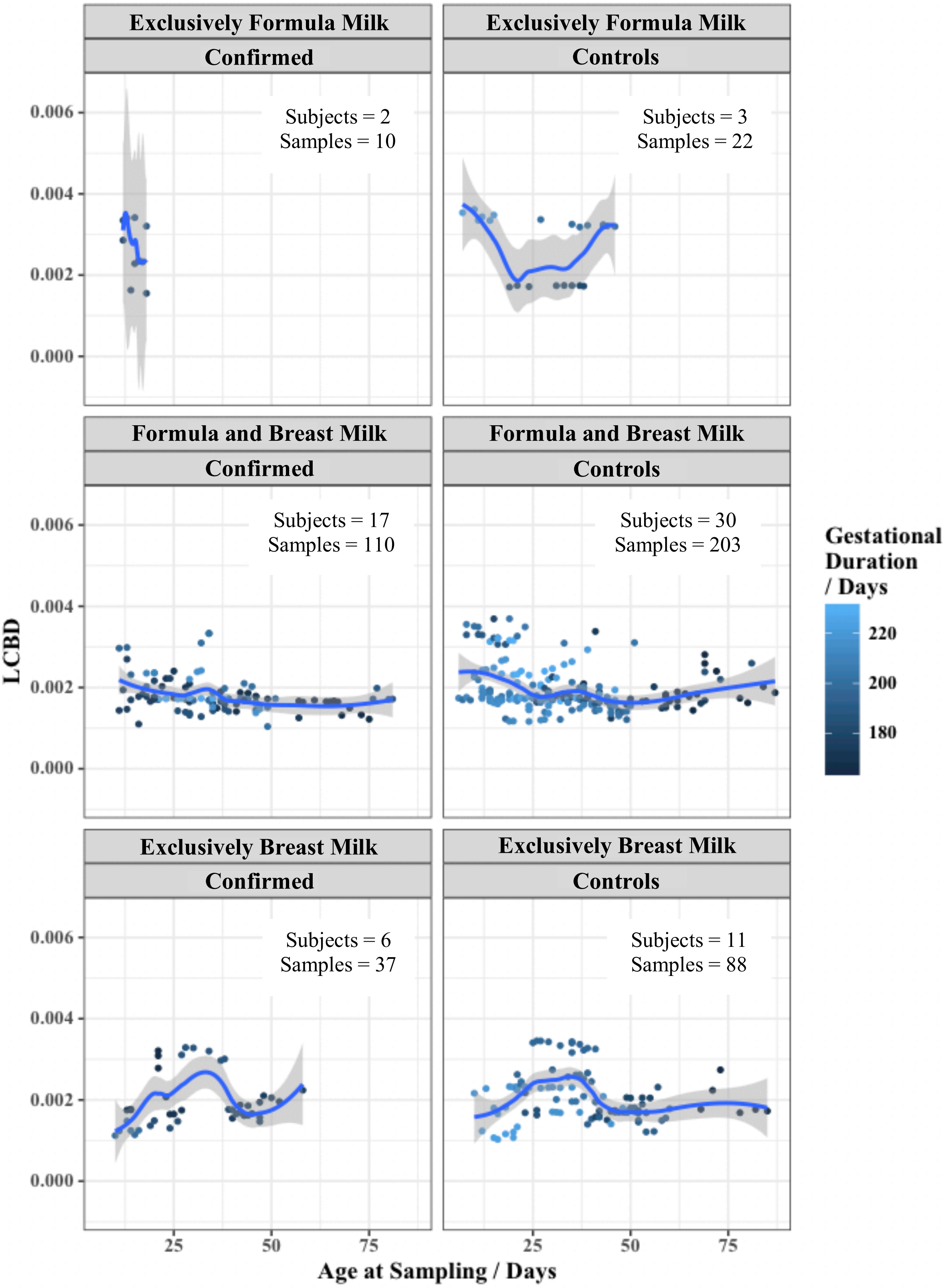


Figure 11 Subsets of feeding regime and NEC status relative to beta-diversity LCBD values and age at sampling.

### 4.2.3 Association of Delivery Method and Community Composition over Time

The importance of the delivery method was described by the increased variability accounted for in models that used this match factor (Second best model identified in Figure 9) therefore samples were subset by the mode of delivery to establish any visible trends (Figure 12). Compared to the differences observed for feeding regime (Section 4.2.2), each subset showed greater differences in the trends of LCBD values relative to the NEC status of the sample.

The initial sample LCBD values for vaginally delivered infants were observed to have the greatest difference between means from control and NEC subjects, but by 50 days of age there was a greater difference in caesarean delivered infant subgroup. This large difference in LCBD values between control and NEC subjects indicates that the community changes occurring are very different between the two groups. These results suggest that the initial colonising bacteria from the environment are much more similar between NEC and control samples from infants delivered by caesarean section, and there are relatively few changes occurring within these communities that are different between the two groups.

Both subsets appeared to have the same trend observed in infants fed natural breast milk. A peak in LCBD values at approximately 38 days followed by a decline to day 50 days. This showed that infants fed breast milk exclusively were highly influential on the average LCBD score within both delivery subsets.

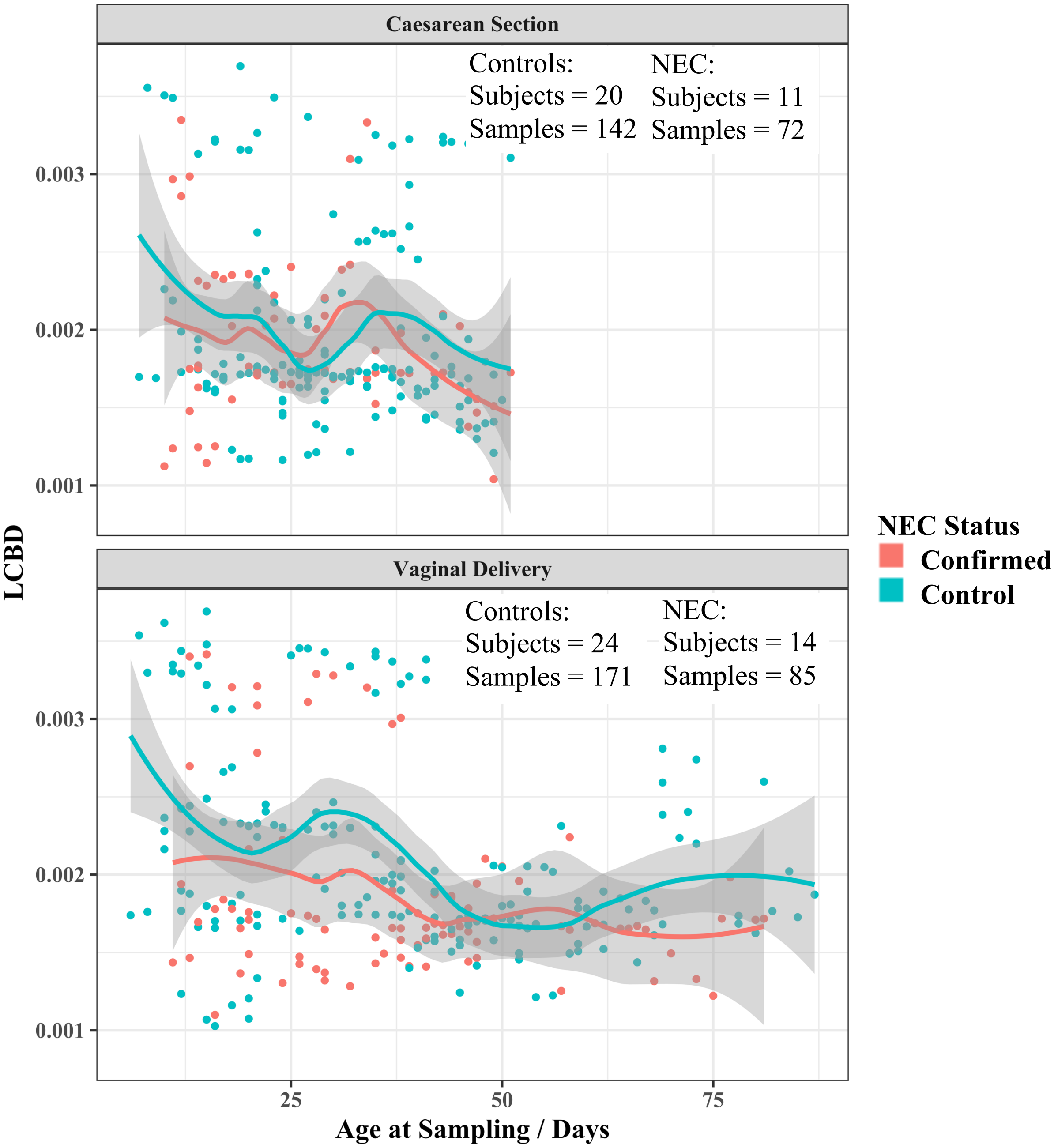


Figure 12 Changes in beta-diversity (LCBD) in relation to age, delivery method and NEC status.

### 4.2.4 Association of Feeding Regime, Mode of Delivery, Age and Gestation with Community Composition over time.

Subset regression analysis suggested that the model accounting for the most variance was a combination of feeding regime, mode of delivery, age and gestational duration. However, creating subsets based on a combination of mode of delivery and feeding regime limited the statistical strength of some subgroups due to the low subject and sampling densities.

Specifically, all the subgroups with infants exclusively fed formula milk and the subgroup of infants exclusively fed breast milk and delivered by caesarean section (Figure 15).

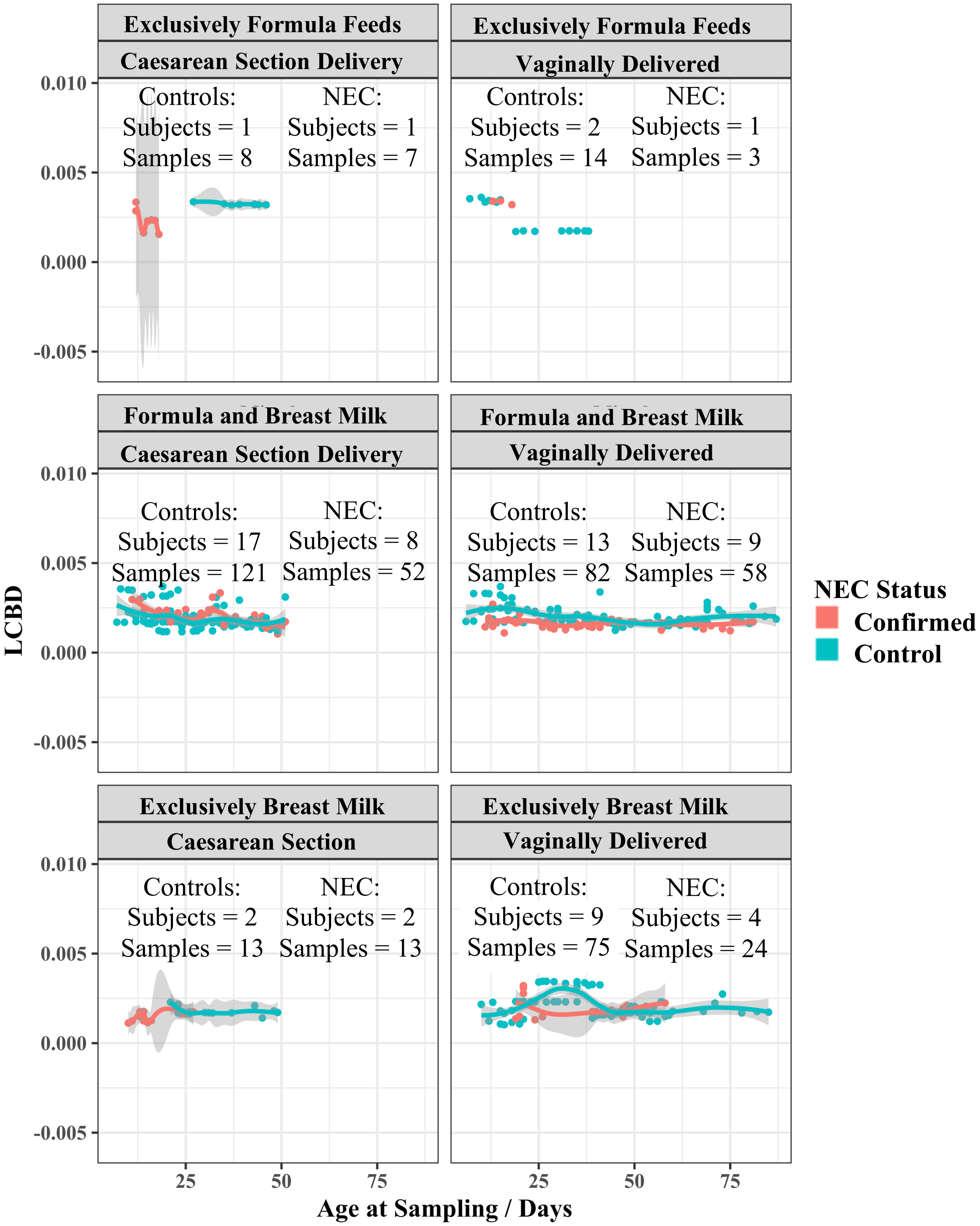


Figure 13 Changes in beta-diversity (LCBD) in relation to age, feeding regime, delivery method and NEC status.

Those subgroups that could be compared were vaginally delivered infants fed breast milk exclusively and both subgroups for infants fed a mixture of formula and breast milk. These samples showed very little difference with regards to NEC status and beta-diversity fluctuations. This suggested that the trends in community composition were less affiliated with the NEC status of the infant sampled and more associated with feeding regime and mode of delivery.

Both subsets for samples from infants on the mixed feeding regime were seen to have similar, stable, LCBD values over time, which differed from the samples of control infants fed exclusively breast milk. These samples showed a peak LCBD at ~30 days which declined up to day 40, after which all samples from all subgroups had similar LCBD values (~0.002).

## 4.3 Canonical-Correlation Analysis (CCA) of Match Factors Associated with Trends in Community Composition

Match factors were summarised using analysis of variance (ANOVA) with Bray-Curtis dissimilarity index[413] to establish associations with sample distribution and beta-diversity. Additionally, to understand the uniqueness of each community the subject ID was included in this analysis. By doing so it was possible to partition the distance matrices from sources of variation and to fit linear models to the distance matrix.

The results from the permutational ANOVA (PERMANOVA) test identified subject ID and the age at sampling as the only two factors significantly associated with beta-diversity. These accounted for 78% and 0.005% of the variance between samples respectively (Table 10).

Table 10 PERMANOVA analysis of match factors in relation to OTU abundance counts using Bray-Curtis dissimilarity index with 999 permutations. *** = p-value less than 1x10$^{-3}$.

| | Df | Sum Of Squares | Mean Squares | F Model | $R^2$ | Pr(>F) |
|---|---|---|---|---|---|---|
| Subject ID | 68 | 129.43 | 1.90 | 23.60 | 0.78 | 0.001 *** |
| Age at Sampling | 1 | 0.75 | 0.75 | 9.32 | 4.45x10$^{-3}$ | 0.001 *** |
| Residuals | 398 | 32.11 | 0.08 | | 0.19 | |
| Total | 469 | 167.15 | | | 1 | |

This test confirmed that the microbiome was unique to each individual and altered with age. While these methods accounted for the trends in the beta-diversity of samples they did not describe the differences in the taxa of those communities. In order to establish how similar the community compositions were within each subset non-metric multidimensional scaling (NMDS) analysis was performed and used to identify unique clusters associated with NEC and control subjects

## 4.4 NMDS Analysis of Community Composition

Multidimensional scaling enabled the visualisation of each taxa abundance per sample relative to all other sample taxonomic abundances in the population. This was performed using a set of related ordination techniques to display the OTU presence and abundance. NMDS analysis was used to identify both a non-parametric monotonic relationship between the dissimilarities in the abundance table and the Euclidean distances between samples.

NMDS depends only on a biologically meaningful view of the data, therefore the choice of standardisation, transformation and similarity coefficient appropriate to the hypothesis under investigation are important[414]. Bray-Curtis dissimilarity index was considered the most appropriate for the dataset as it is used to quantify compositional dissimilarity between samples using OTU counts from each sample relative to the rest.

Subset regression established feeding regime as the factor most significantly associated with LCBD value of the beta-diversity over time. Normalising for feeding regime provided the best means of identifying sample clustering with respect to the NEC status of the infant.

### 4.4.1 Analysis of Taxonomic Ranking and Discrimination of Necrotising Enterocolitis Status Based on the Sample Community Composition

It was important to establish which taxonomic level best described the NEC status of samples prior to analysis with demographic/medical information. To establish the most successful taxonomic rank in clustering samples according to NEC status, NMDS with Bray Curtis was performed at the Phylum, Class, Order, Family, Genus and OTU level (where OTUs were sequences defined as unique based on the QIIME parameters described in the methods section, Section 2.9.2).

NMDS analysis establishes the similarity between multi-dimensional objects in a dataset by representing each dimension on an axis. In this case the dimension is a taxonomy and the value of the axis is the normalised abundance of that taxon represented as an ordered position (non-metric) of similarity to the other objects. In this manner, objects with similar multi-dimensional profiles are clustered together, and this can be shown statistically using PERMANOVA analysis. In this case, NEC and control samples (NMDS objects) were represented at a given taxonomic rank, which represented the individual axis, and compared to significance using PERMANOVA. This was performed for all the taxonomic ranks individually.

The extent and significance of NEC clustering within the NMDS plots was assessed using permutational PERMANOVA. This technique did not assume parametric distributions and could be combined with the Bray-Curtis measure of dissimilarity to compare pairs of samples or variables[415].

The most significant association of sample clustering and NEC status was at the genus level (Table 11), however there was no distinct clustering of NEC samples within the NMDS plot even at this level. This showed that there was a greater level of complexity within the community structure relative to NEC status (Figure 14). The previously described trends in community structure with LCBD values showed significant associations with feeding regime and mode of delivery (Section 4.2). Further analysis focused on subsets based on feeding regime and mode of delivery at genus level.

Table 11 PERMANOVA analysis of community composition and NEC Status for all samples using different taxonomic ranks. *** = p-value less than $1x10^{-3}$, ** = p-value less than $1x10^{-2}$.

| | **Sums Of Squares** | **Mean Squares** | **F Model** | **R$^2$** | **Pr(>F)** | |
|---|---|---|---|---|---|---|
| Genus | 1.828 | 1.82789 | 7.1189 | 0.01461 | 0.001 | *** |
| Order | 1.082 | 1.08171 | 6.0817 | 0.01251 | 0.001 | *** |
| Class | 1.002 | 1.00163 | 5.9362 | 0.01222 | 0.001 | *** |
| Family | 1.017 | 1.01695 | 5.5493 | 0.01143 | 0.001 | *** |
| Species | 1.202 | 1.20183 | 3.26 | 0.00675 | 0.001 | *** |
| Phylum | 0.717 | 0.71657 | 4.4577 | 0.0092 | 0.009 | ** |

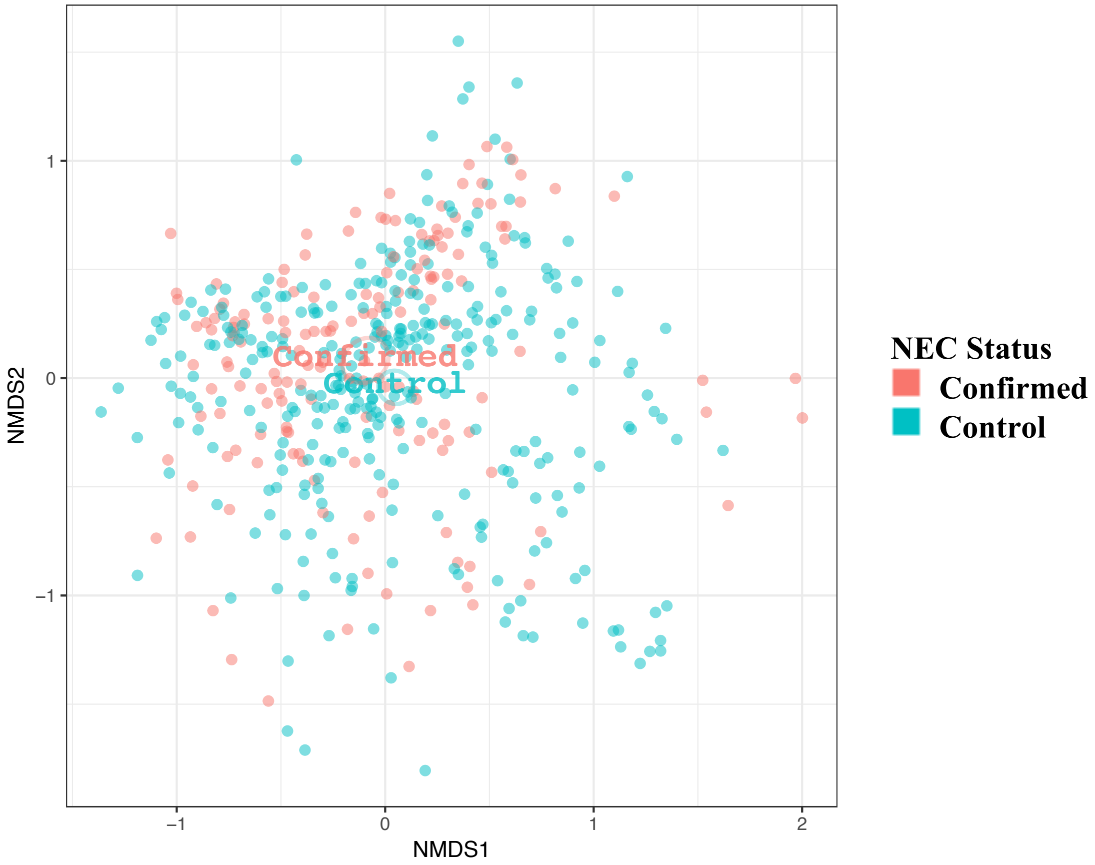


Figure 14 NMDS Plot for all samples with community composition described at the Genus Level with the whole, viable, cohort, using PERMANOVA.

### 4.4.2 Analysis of Community Composition and Necrotising Enterocolitis Status of Samples Subset by Feeding Regime

Subset regression analysis suggested that feeding regime was one of the most influential factors in the changes of beta-diversity as represented by LCBD values. NMDS plots were constructed at the genus level for each subgroup in order to establish whether there was sample clustering based on NEC status with respect to feeding regime.

The subgroups showed greater differentiation between sample NEC status and clustering relative to the entire cohort (Figure 15), as confirmed by the increased $R^2$ value (Table 12). However, there was a reduced significance for the sample distribution and association with NEC compared with the total cohort.

Samples from infants fed exclusively breast milk were seen to show greater separation with respect to NEC status, as indicated by the ellipses on the NMDS plot. These ellipses represent the centroid of the samples from each group. Infants on mixed feeding regimes were seen to cluster by NEC status less relative to the total case-control population. The lack of sampling density excluded formula fed infants from analysis.

Although there was an increase in the variance of the communities for all subgroups, with each observed to have a significant association between sample distribution and NEC status, these differences were less significant than PERMANOVA/NMDS analysis of the whole cohort. This was likely due to the influence of feeding regime on the community composition (as previously described) which would result in community structures that are more similar across all samples. As shown by the reduction in variance described in the $R^2$ scores.

Table 12 PERMANOVA analysis of community composition and NEC status for samples subset according to feeding regimes at the genus taxonomic rank. *** = p-value less than $1x10^{-3}$, ** = p-value less than $1x10^{-2}$. * = p-value less than 0.05.

| | **Sum Of Squares** | **Mean Squares** | **F Model** | **$R^2$** | **Pr(>F)** | |
|---|---|---|---|---|---|---|
| Formula Milk Only | 0.9935 | 0.9935 | 3.9242 | 0.11568 | 0.006 | ** |
| Formula & Breast Milk | 1.107 | 1.1068 | 4.6355 | 0.01469 | 0.007 | ** |
| Breast Milk Only | 0.861 | 0.86102 | 3.1728 | 0.02296 | 0.016 | * |
| **Cohort Genus Level** | **1.828** | **1.82789** | **7.1189** | **0.01461** | **0.001** | *** |

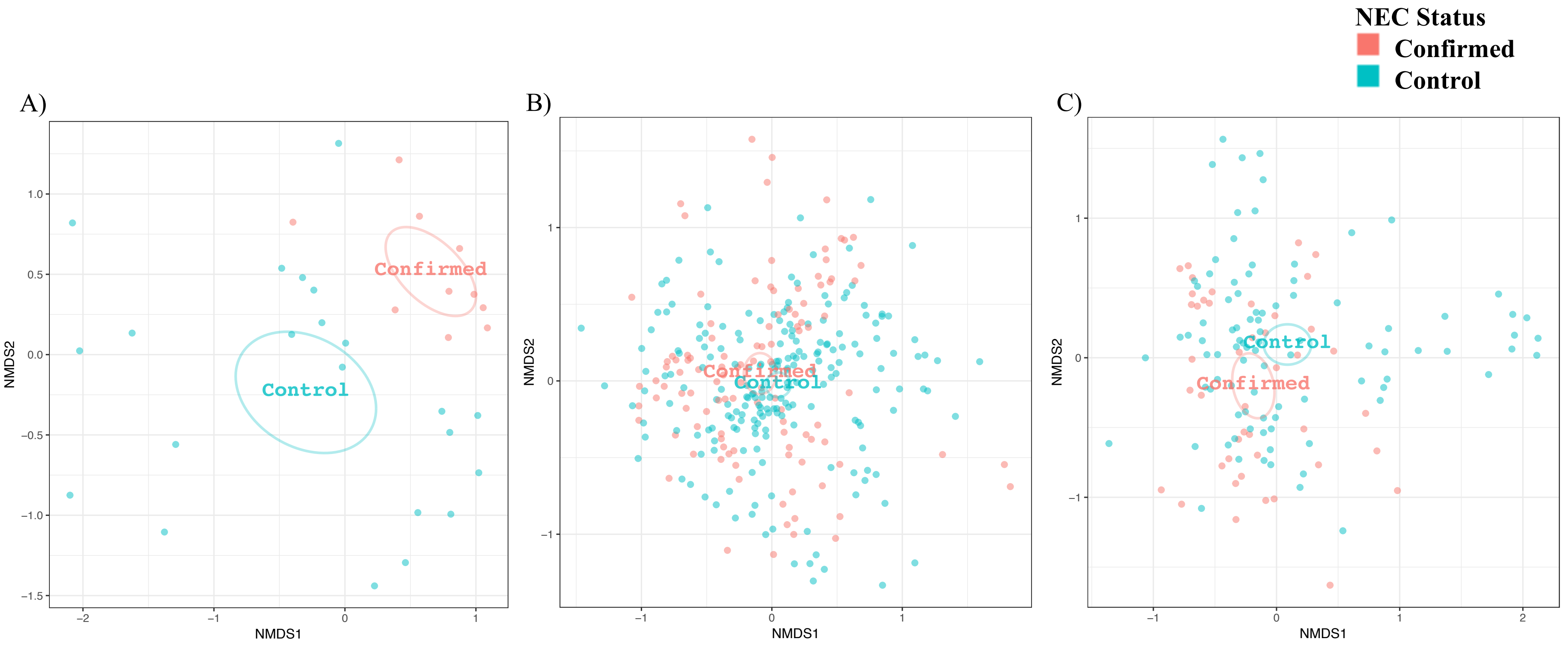


Figure 15 NMDS Plot for samples with community composition described at the Genus Level. A) Infants fed Exclusively Formula Milk. B) Infants on a mixture of formula and breast milk. C) Infants on a diet of exclusively breast milk.

### 4.4.3 Analysis of Sample Community Composition and Necrotising Enterocolitis Status for Infants Subset According to Delivery Method

Delivery method was the next most significant categorical factor in the subset regression analysis. Subgroups of samples were generated according to delivery method and NMDS plots were constructed at the genus level for each subset.

The clustering of samples from infants delivered vaginally just as significant with respect to their NEC status as seen in PERMANOVA analysis for the total cohort. This subset also accounted for a greater proportion of the variance within the sample distribution. The clustering of samples from infants delivered by caesarean section was less significant with respect to NEC, although the variance accounted for was similar to the entire case-control cohort sample population (Table 13).

This showed that differences in community composition of infants delivered vaginally was greater between NEC and control samples compared to those delivered by caesarean section. This separation was shown clearly in the NMDS plots, wherein samples from vaginally delivered infants showed greater clustering relative to their NEC status (Figure 16).

Table 13 PERMANOVA analysis of community composition and NEC status for samples subset according to delivery method at the genus taxonomic rank. *** = p-value less than 1x10$^{-3}$, ** = p-value less than 1x10$^{-2}$.

| | Sums Of Squares | Mean Squares | F Model | R$^2$ | Pr(>F) | |
|---|---|---|---|---|---|---|
| Vaginal Delivery | 3.755 | 3.7547 | 13.854 | 0.0504 | 0.001 | *** |
| Caesarean Delivery | 0.729 | 0.72923 | 3.4651 | 0.01572 | 0.007 | ** |
| **Total at Genus Level** | **1.828** | **1.82789** | **7.1189** | **0.01461** | **0.001** | *** |

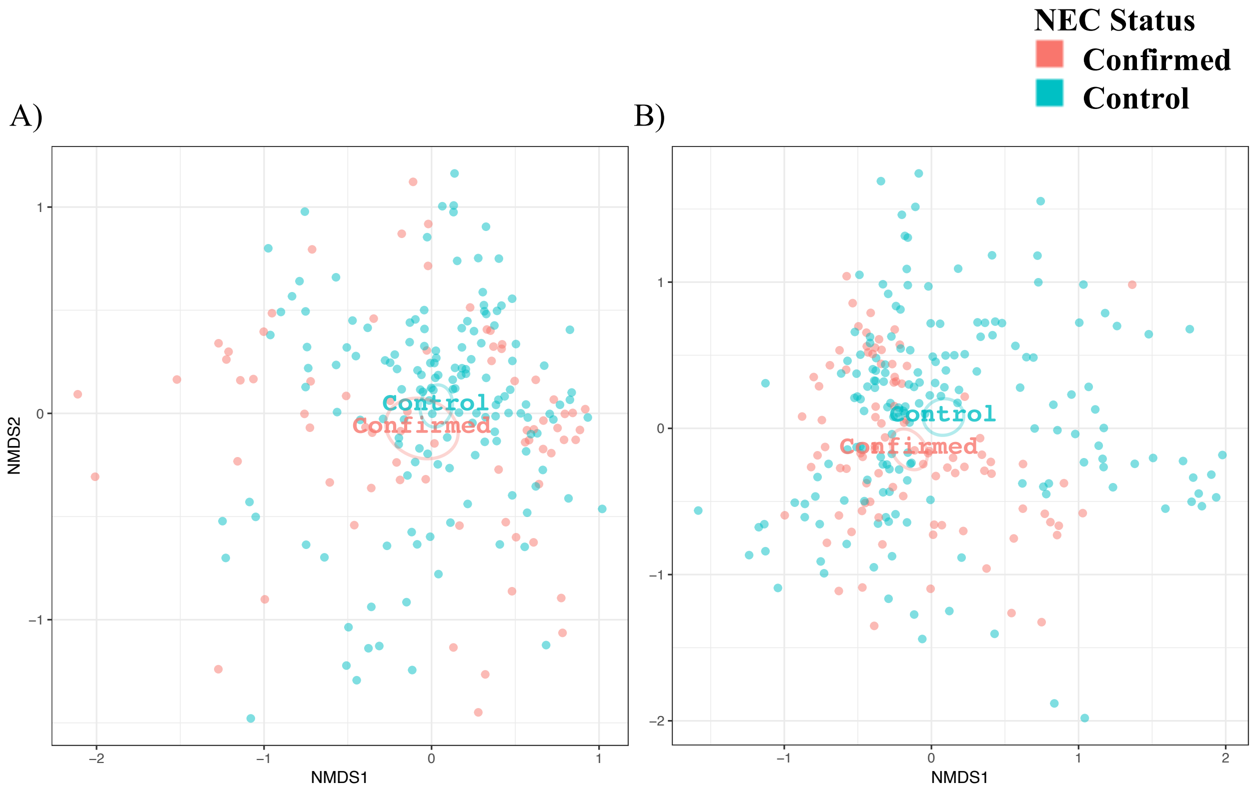


Figure 16 NMDS Plot for samples with community composition described at the Genus Level. A) Samples from infants delivered by caesarean section. B) Samples from infants delivered vaginally.

### 4.4.4 Analysis of Sample Community Composition and Necrotising Enterocolitis Status for Infants Grouped by Delivery Method and Feeding Regime

Infants fed on the mixed feeding regime and delivered vaginally were the most significantly associated with NEC status and sample distribution as well as accounting for the most variance. NEC status within this subset accounted for 10% of the total variance in sample distribution and the highest significance of association across all the subsets analysed ($p = 0.001$) (Table 14). This was clearly shown in the NMDS plot by NEC samples being distributed towards lower X-axis values within this subgroup (Figure 17: C). Community compositions from both mixed feed subgroups were seen be significantly associated with NEC status with the subset of infants delivered by caesarean section increasing to 3% of the variance accounted for (relative to the total cohort) (Figure 17: B).

Table 14 PERMANOVA analysis of community composition and NEC status for samples subset according to delivery method and feeding regime at the genus taxonomic rank. *** = p-value less than $1x10^{-3}$, ** = p-value less than $1x10^{-2}$. * = p-value less than 0.05.

| | Sums Of Squares | Mean Squares | F Model | $R^2$ | Pr(>F) | |
|---|---|---|---|---|---|---|
| Formula & Breast Milk Vaginal Delivery | 3.906 | 3.9056 | 16.011 | 0.10396 | 0.001 | *** |
| Formula Only Caesarean Delivery | 1.462 | 1.46198 | 12.518 | 0.49056 | 0.002 | ** |
| Formula & Breast Milk Caesarean Delivery | 1.019 | 1.0186 | 5.1139 | 0.02904 | 0.004 | ** |
| Breast Milk only Caesarean Delivery | 0.4526 | 0.45259 | 2.4369 | 0.07752 | 0.065 | |
| Breast Milk Only Vaginal Delivery | 0.6054 | 0.60543 | 2.1185 | 0.01996 | 0.077 | |
| Formula milk Only Vaginal Delivery | 0.4007 | 0.40067 | 1.5041 | 0.09113 | 0.242 | |
| **Total at Genus Level** | **1.828** | **1.82789** | **7.1189** | **0.01461** | **0.001** | *** |

Samples from infants delivered vaginally and fed exclusively breast milk were not observed to have a significant association in composition and NEC status (Figure 17: F). This subgroup showed a lower percentage of variance accounted for by NEC status compared to either the entire case-control cohort or the other subsets in Figure 17. While there is a lower level of clustering with respect to the NEC status of samples there is also a high level of separation between NEC and control samples.

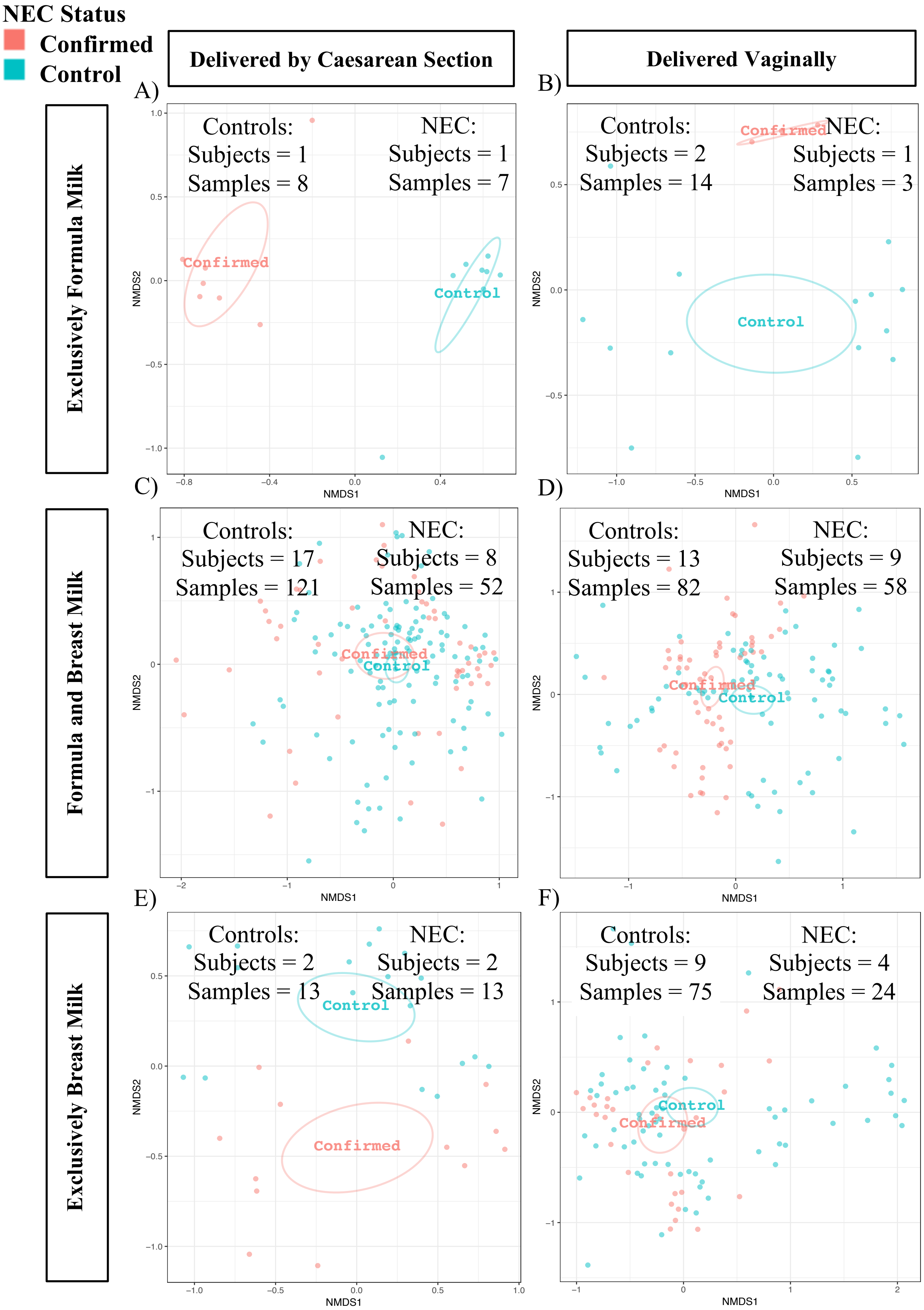


Figure 17 NMDS Plot for samples with community composition described at the Genus Level. A) Formula fed infants delivered by caesarean section. B) Formula fed infants delivered vaginally. C) Formula and breast milk fed infants delivered by caesarean section. D) Formula and breast milk fed infants delivered vaginally. E) Breast milk fed infants delivered by caesarean section. F) Breast milk fed infants delivered vaginally.

### 4.4.5 Weighting Gestation and Age in NMDS Analysis of Sample Community Composition with Respect to Necrotising Enterocolitis

The remaining continuous match factors (birthweight, gestational duration and age at sampling) were plotted as vectors over each NMDS plots to establish any association with sample distribution for the three subgroups with sufficient sampling depth. The vectors represented the strength of association a variable had with the distribution of samples within the NMDS plot. None of these factors showed strong association with the distribution of any subgroups any of the subgroups (Figure 18).

The age at sampling was weighted on the NMDS plot using 2-D smoothing due to the significant association with LCBD values (Section 4.3) and it's importance in statistical viability for case-control cohort analysis[412]. Additionally, this factor was considered important from a biological perspective due to large fluctuations in the infant gut microbiome during the early stages of development[311]. There was no improvement in clustering for infants born by caesarean section on a mixed feeding regime when age was weighted in the NMDS plots (Figure 19: A).

Infants fed exclusively breast milk and delivered vaginally showed some clustering (Figure 19: B) but this did not appear to be a clear improvement on the original NMDS plot (Figure 17: C). Samples from infants delivered vaginally and fed breast milk exclusively were seen to show poor clustering when age was weighted in the NMDS plot, although there was again marginally increased separation between NEC and control samples, relative to samples from caesarean delivered infants fed both formula and breast milk (Figure 19: C).

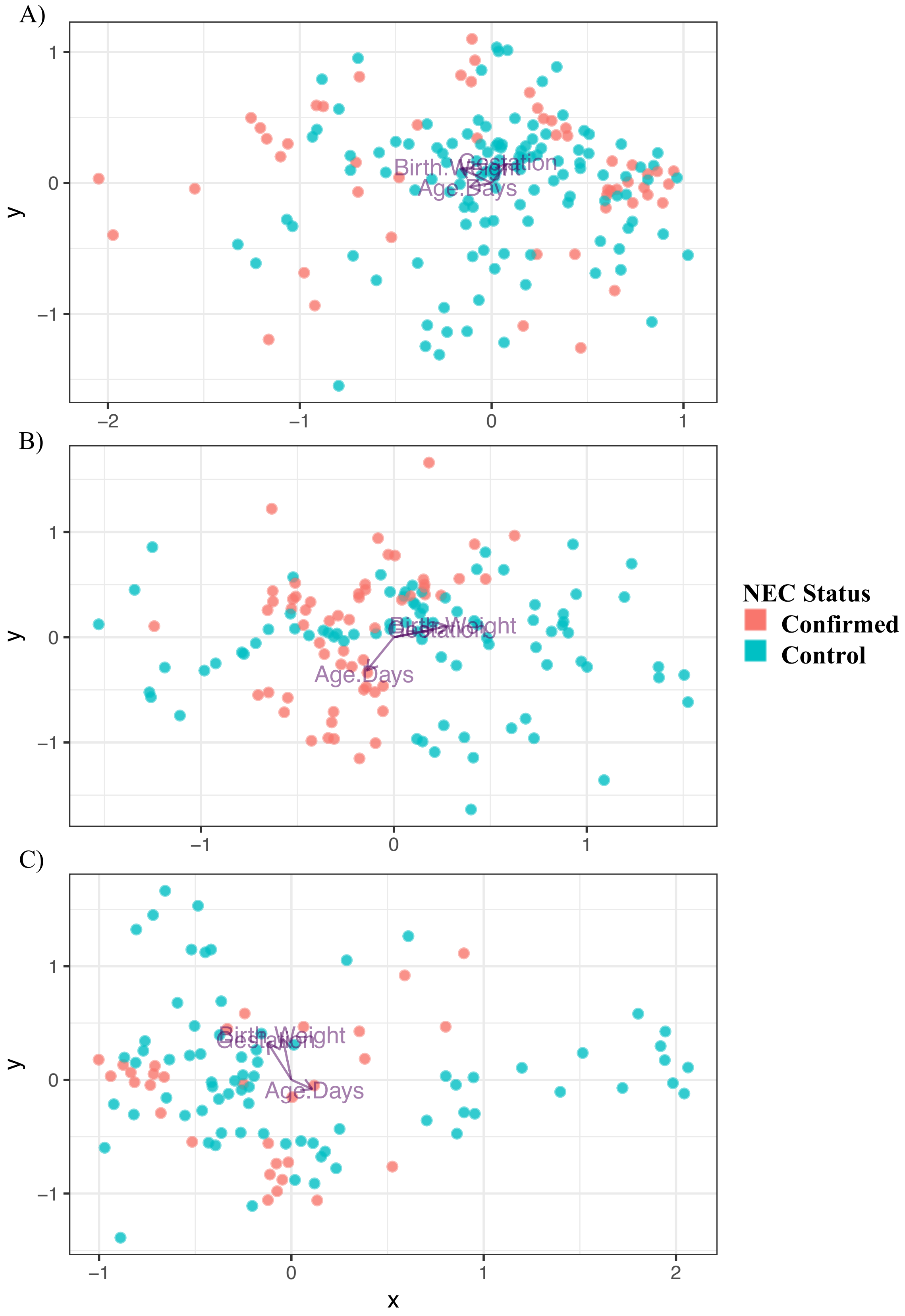


Figure 18 NMDS Plot for samples with community composition described at the Genus Level with vectors describing the association of sample distribution and gestational duration, birthweight (g) and age at sampling (days). A) Formula and breast milk fed infants delivered by caesarean section. B) Formula and breast milk fed infants delivered vaginally. C) Breast milk fed infants delivered vaginally.

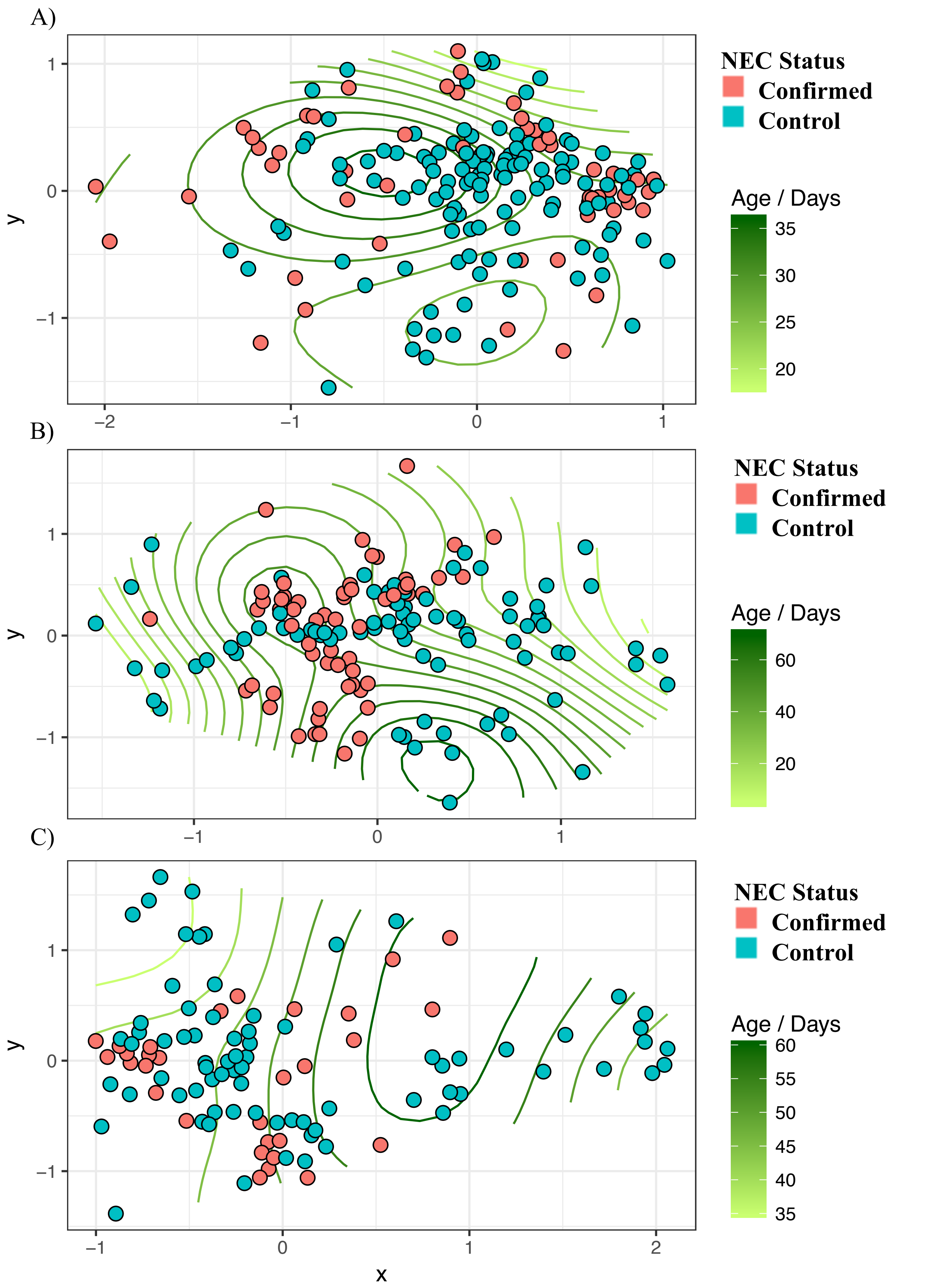


Figure 19 NMDS Plot for samples with community composition described at the Genus Level. Sample distribution was normalised using age at sampling. A) Formula and breast milk fed infants delivered by caesarean section. B) Formula and breast milk fed infants delivered vaginally. C) Breast milk fed infants delivered vaginally.

## 4.5 Discussion

This chapter aimed to identify match factors significantly associated with trends in the bacterial community of samples from NEC and control infants. This was quantified using the LCBD beta-diversity metric and tested using linear regression, subset regression and CCA. Once significant match factors were established with conserved trends in the community composition of samples, NMDS analysis was used to visualise how similar NEC and control infant gut communities were over the duration of sampling. PERMANOVA analysis was used on the NMDS plots to statistically tested the association of NEC and sample distribution.

### 4.5.1 Sample Preparation, Amplification and Sequencing

Sampling proved challenging due to the low quantity and inconsistency of faeces from infants diagnosed with NEC and their associated controls. This was likely due to NEC treatment often being in the form of nil-by-mouth in addition to the reduced passage of stool due to the symptoms of the disease. Low quantities of stool on the days where faeces were available also limited the success of DNA extraction, and by association, 16S V4 PCR amplification.

To ensure that the time series would accurately represent differences in the microbiome of infants with NEC, only samples from controls that were assigned to successfully sequenced NEC patients were used for microbiome analysis. In addition to this, infants within the control groups were selected based on the presence of at least three samples from each infant over the time series. This sought to ensure that each subject contributed to the time series with enough samples to provide an average and identify any outliers for a given subject.

### 4.5.2 Establishing Key Match Factors Associated with Different Community Structures

Subset regression identified that LCBD values were significantly associated with the age at sampling (Pr(>|t|) = $1.07 \times 10^{-3}$). This supported the evidence that changes in beta-diversity over time in the early infant life are attributed to a colonisation pattern primarily instigated by pioneering bacteria from the mother and environment[328,320]. Therefore, it was important to establish comparisons that were normalised for age, especially when identifying taxonomic differences between infant microbiomes as in the first year of life are typically characterised by low diversity and high instability[311,416].

Linear regression of LCBD values established age at sampling, gestational duration and feeding regime as being significantly associated with trends in the beta-diversity (Linear Regression: p = $1.33 \times 10^{-15}$).

Gestational duration has been shown to be correlated with *Enterobacteriaceae* and other potentially pathogenic bacteria such as *Clostridium difficile* or *Klebsiella pneumoniae*, which were found in greater abundances in preterm infants[417]. Full term infants were characterised by greater diversity and more of the common genera i.e. *Bifidobacterium*, *Lactobacillus* and *Streptococcus*[418]. Additionally, it has been shown that infants born after a shorter gestational duration were seen to have a delayed acquisition of *Bifidobacterium* and *Collinsella species*[419].

However, when sample distributions in NMDS plots were weighted by gestational duration there was no clear increase in the level of clustering of the samples with regards to their NEC status. This could have been due to all infants within this study being preterm and the influence of gestational duration not being as prominent as it has been observed between preterm and term infants.

The results from subset regression show that there was a clear association with the feeding regime and changes in the LCBD values of infant samples. Infants on a mixed feeding regime were seen to be the most significantly associated with trends in LCBD values ($Pr(>|t|) = 1.38x10^{-12}$), closely followed by infants exclusively fed breast milk (($Pr(>|t|) = 4.70x10^{-10}$). The association of feeding regime and the development of the infant microbiota has been explored extensively and believed to be highly influential in the composition and development of the microbiome in the first six months[314,317,318].

Analysis of LCBD values as a function of time suggested that infants fed a mixture of formula and breast milk were seen to have very stable LCBD values. This indicated that there was little change in the beta-diversity of individual samples compared to the sample population within this subset. This was contrary to those infants fed breast milk exclusively which appeared to have an increase in LCBD values initially before stabilising around 40 days of age. The greater proportion of breast milk within the diet maybe linked with an increased rate of acquisition as observed in literature[42], however the influence of formula milk on the development of the community cannot be discounted. Formula feeds could potentially reduce the colonisation rate of novel species within the community resulting in less differences between individual samples relative to the population, however it was not possible to observe this due to the lack of samples from infants fed exclusively formula milk.

Further analysis should establish the influence of formula milk and breast milk independently and combined on the development of the microbiota. There was no distinction between NEC and control samples for infants fed a mixed feeding regime. Infants fed exclusively breast milk were seen to differ between 28-40 days of age, however there were few NEC samples between these days and therefore the NEC mean assumed lacked statistical weight. Where there was sufficient sampling, i.e. before 28 days and after 40 days, case and control samples were seen to have very similar LCBD values, with overlapping standard errors.

Whilst mode of delivery was not observed to be significantly associated with LCBD values it was observed to be an influential factor in the subset regression (Figure 10). The lack of association with changes in the community composition may be due to the duration of sampling. Samples taken closer to the delivery date are more likely to be influenced by the mode of delivery, however over the course of sampling in the population this association may be non-significant across the population. Additionally, some infants would not have been sampled close to the date of birth and the influence of delivery method is reduced in those cases.

The increased significance of association between subset regression analysis and the mode of delivery is likely associated with infants sampled closer to the date of delivery. However, there is no negative influence with the inclusion of this factor for samples taken from infants at older ages. Therefore, mode of delivery can increase the predictability of LCBD values in some samples/subjects but this does not apply to all subjects in the population, only those sampled close to the date of birth.

Mode of delivery has been extensively explored in previous literature as well, with evidence that the microbiomes of infants delivered vaginally are closely associated with the maternal faecal microflora[156,309] and those delivered by caesarean section resemble the environmental and mother's skin microbiota[309]. Due to the proximity of sampling after delivery it was considered pertinent that community composition was assessed according to delivery method.

When LCBD values were plotted by delivery method and age there were similar trends for samples from both groups, these both resembled the trends seen for samples from infants exclusively fed breast milk. This confirmed that infant feeding regime was more influential than the mode of delivery over the duration of sampling. While samples subset by delivery method showed the greatest discrepancy between NEC and control infants the standard errors from case and controls were never distinct which indicated that there was substantial overlap in the LCBD values of the sampling populations.

Each subject was seen to have significantly specific bacterial compositions and trends. This was assessed through CCA analysis of LCBD values with the inclusion of subject ID. Subject ID was seen to account for 78% of the variance between sample LCBD values (Pr(<F) = 0.001). This inter-individual variability has been observed in infants[419,420] and is likely a culmination of both match factors and other factors that were not accounted for, for example antibiotic administration, staff and site microbiomes  This demonstrated the difficulty challenge of identifying differences between case and controls subject samples when subject microbiomes were each so unique.

Of all the factors that could not be accounted for antimicrobial administration was believed to be the most important. Premature infants are routinely administered antimicrobials as a preventative measure however NEC infants within the population were shown to be administered significantly more antimicrobials compared to non-NEC infants (Section 3.1.7, T-test: $_{0.95,\,99.3}$: t = 14.86, p-value < $2.00\text{x}10^{-16}$, Figure 5).

Research into antimicrobials has shown that there is an association with administration and reduced taxonomic richness, diversity and evenness of the community, but the extent of this change is variable among individuals[421,420]. The severe side-effects of antimicrobial treatment on the gut microbiota range include self-limiting “functional” diarrhoea and life threatening pseudomembranous colitis[422,423]. Chronic conditions such as asthma and atopic disease have also been linked to long-term use of antimicrobials in children in addition to an altered intestinal microbiota[424,425,426]. While it has been shown the composition of the gut community returns to a community of similar structure these findings are primarily based on low resolution techniques[421,427,428] and other studies have shown the effects of a single course of antimicrobials can persist for years[429,430,431].

Current literature reviewing the impact of antimicrobials is predominantly limited to a single type, class or regime[432]. Within the NICU infants are often on multiple antimicrobials which raises questions about the impact and lasting changes they are likely to incur. However, this also lowers the likelihood of finding appropriate control subjects when considering a disease such as NEC. Therefore, this should be of particular note in future research that this factor is a likely candidate for the inter-individual variation between subjects and could ultimately limit the ability to detect differences between NEC and control subjects.

### 4.5.3 Taxonomic Rank that Best Discriminated Between Necrotising Enterocolitis and Control Samples

Genera was identified as the most effective taxonomic level to discriminate between NEC and control samples using NMDS and PERMANOVA analysis of sample distributions (Pr(>F) < 0.001). Even when using genera only a very small percentage of the variance between NEC and control samples could be accounted for ($R^2 = 0.015$). Therefore, analysis of match factor subsets was used to identify distinct changes in the gut microbiome community of infants with NEC.

### 4.5.4 Differentiation of Samples Based on Necrotising Enterocolitis Status and Community Composition

As stated previously, existing literature has shown that samples from infants with NEC and their respective controls have been clustered independently[110] but this study was from a single location using only one sample per subject. This limited the conclusions that could be made and extrapolated from the data. More recent large scale studies identified differences in the rarefied proportional abundances of aggregated samples[46] but showed no evidence of independent clustering with NMDS analysis.

Following this analysis, it was not possible to identify distinct clusters separating samples from infants with NEC and controls. However, it was possible to increase this differentiation/clustering between samples for some subgroups established earlier in the chapter.

1.5% of the sample variance for the overall population when presented at the genus level in NMDS plots was explained by the distribution of NEC and control samples (PERMANOVA $p = 0.001$). When samples were subset by feeding regime infants fed breast milk exclusively increased to 2.9% but were considered non-significant ($p = 0.08$). Infants fed a mixture of formula and breast milk delivered vaginally or by caesarean were seen to be the most distinct (10% variance, $p = 0.001$ and 2.9%, $p = 0.004$, respectively). The lack significance for infants fed breast milk and delivered vaginally could be attributable to the low sample count or the lack of NEC samples over a large portion of sampling days.

The demonstrated that whilst inter-individual variation within the microbiome was a significant factor, and the composition of the microbiomes of NEC and controls were highly different, this difference could be reduced by accounting for the feeding regime and mode of delivery. By normalising for these match factors that are known to influence the composition and development of the microbiome it was possible to increase the clustering or separation of NEC samples from their controls. What cannot be accounted for within this analysis is the wide range of environmental influences that also contribute to the community composition such as antibiotic regime, health care worker microbiomes or the NICU microbiome.

Although clustering has been demonstrated in other studies subject samples were subset according to the week of life that a subject was sampled over[105]. This however was not considered appropriate within the dataset presented here due to the lack of information regarding the date of NEC onset. Instead it was considered more appropriate to monitor the trends over time. This enabled a direct comparison of samples over the duration of sampling ages as well as the identification of time periods that lacked sufficient sampling.

## 4.6 Conclusion

Sufficient sampling density was established and allowed for the analysis of the microbiome of premature infants regarding diagnostic or prognostic factors associated with NEC. Due to many subjects yielding fewer than three samples, these could not be included in many gold standard NEC groups, though this was an inherent complication of NEC and as such little could be done to improve sampling success.

The inter-individual variation between samples was the strongest factor associated with the community composition of samples, regardless of the infant's NEC status. However, match factors were observed to have significant associations with the community composition and this was demonstrated by subset regression and NMDS analysis. But by subsetting samples according to delivery method and feeding regime it was possible to increase the variance accounted for in the community composition. Two of these three groups were also considered to have a significant association in the sample distribution and NEC status. The one subset that was seen to be non-significant also showed increased separation but was limited by sample counts and potentially a lack of sampling for NEC subjects. Further analysis should describe the differences in taxonomy between these subsets to establish how NEC and control sample communities are different or whether there is an underlying pathogenic/probiotic taxa.

# 5 Taxonomic Differences between NEC and Control Infants

Subsets of samples using match factors that were most significantly associated with the altered beta-diversity trends did not clearly differentiate between NEC and control subjects (Section 4.3). However, it has been shown that differences could be associated with specific taxonomic groups (such as Clostridia[433,434], Bifidobacteria[435] and Lactobacillus[436]) rather than changes in the community composition.

## 5.1 Background Literature on the Taxonomic Differences Associated with Necrotising Enterocolitis

Sántulli *et al* were the first to suggest that bacterial colonisation was one of three critical factors in the occurrence of NEC[437,438] (alongside mucosal injury and enteral feeding regime). This implication was built upon by common findings of bacteraemia[439,440,441] and endotoxinaemia[442,443] in infants with NEC. Increased dominance of Escherichia species preceding NEC was observed in small scale twin studies,[110,444] and more recently in a large scale study by Warner *et al,* which suggested that time-by-necrotising-enterocolitis was positively associated with Gammaproteobacteria ($p = 1.00x10^{-3}$), and negatively associated with strictly anaerobic bacteria, specifically Negativicutes ($p = 1.90x10^{-3}$).

Morrow *et al* suggested that early and late onset NEC could be differentiated based on distinctive taxonomic differences relative to samples from control subjects. In infants with early onset NEC, the diagnostic markers were observed when the infant was between four to nine days of age, and consisted of a Firmicutes dominance, specifically *Bacillus*, *Staphylococcus* and *Enterococcus* genera. Late onset NEC was characterised by a Gram-negative Proteobacteria signature associated with Enterobacteriaceae of which the dominant genera were *Enterobacter* and *Escherichia*[102].

Whilst no single pathogenic strain has consistently been associated with the gut microbiome of infants with NEC, these communities are very complex and highly unique to individuals; therefore identifying a specific pathogen can be challenging without multiple large scale studies and consideration of environmental factors that influence the gut microbiome composition.

As discussed in Section 1.9.2, the human gut is exposed to a vast number of foreign microorganisms following birth, and must be able to distinguish those harmless, dietary microbial antigens from the potential pathogens. The neonatal immune system is dependent on external stimuli to develop the mature immune competence required in the formation of a stable, healthy gut microbiome. Breast milk and indigenous intestinal microbiotas are considered the most appropriate sources of maturational stimuli[445]. A critical paper by Sudo *et al* in 1997 demonstrated that key members of the intestinal community, *Bifidobacterium infantis*, restored oral tolerance to germ-free neonatal mice through the modulation of the IgE response to allergens[446]. In the same paper, other primary members of the murine gut microbiome, *Bacteroides*, were also shown to correct defective immune maturation.

Previous evidence has shown that *Bifidobacteria* are one of the dominant strains in term infancy. In conjunction with *Lactobacilli, Bifidobacteria* are known to promote indigenous lactic-acid bacterial (bifidogenic effect) through the production of short-chain fatty acids[292, 447,448]. The liberation of short-chain fatty acids is an important energy source for the intestinal mucosa, as well as modulating the immune response and tumour genesis in the gut[6].

There is also evidence within gut microbiome studies of bacteria contributing beneficial effects to the host, i.e. probiotic bacterial species. It has been suggested that an absence or reduced presence of beneficial bacteria is responsible for the occurrence of NEC. One of the first investigations into the efficacy of probiotics was by Hoyos in 1999, who showed that daily administration of *Lactobacillus* and *Bifidobacteria* could be associated with a decrease in the mortality rates of NEC when administered as a probiotic[436]. Two further large-scale studies identified that the use of *Bifidobacteria* and *Lactobacillus* in combination with breast milk significantly reduced NEC incidence[155,449]. These studies were based on evidence that term infants have a higher frequency of *Bifidobacteria* and *Lactobacillus* species[450] relative to preterm infants. Evidence has also suggested an absence of *Propionibacterium* in the first week could be implicated in the onset of NEC[102]. *Propionibacterium* is a genus of gram-positive, rod shaped bacteria named for their ability to synthesise propionic acid using transcarboxylase enzymes[451].

While being much simpler than the adult gut community the premature infant gut microbiome is still highly complex. Therefore, the interactions and contributions of community members are still poorly understood at the community level, meaning that the beneficial or pathogenic impact a given strain has on the host is difficult to establish. There are contradicting reports between probiotic clinical trials of NEC in relation to the efficacy of the *Bifidobacteria* genus in preventing NEC morbidity, for example, *Bifidobacterium breve* BBG-001 showed no significant difference in the rates of outcomes for infants and NEC when administered as a probiotic in a large scale, randomised clinical study[452].

The literature described in this introduction shows that the presence of pathogenic taxa or the absence of beneficial, probiotic, bacteria are implicated in the occurrence and mortality rate of NEC. However, neither has been shown to be consistently responsible across studies, although there is an argument that Gammaproteobacteria are widely observed to be associated with NEC and that some *Bifidobacteria* have a demonstrably beneficial impact on the outcome of NEC. What is clear across studies is that the preterm infant microbiome is a volatile environment which is influenced by delivery method, NICU, antibiotic administration, and feeding regimes[309,155,390]. This results in a highly unique community specific to each individual infant, which was shown to be true in the CCA analysis of Chapter 4 Section 3. As such, comparisons between subjects are challenging, both as a community level and at the taxonomic level, as the microbiomes are rarely consistent between twin pairs and clearly become more disparate within population level studies.

This places greater importance on the statistical methods in which changes in abundances of taxa are measured and compared between groups. The industry standard has been to rarefy highly variable count data produced using high-throughput sequencing technology to an appropriate level, using the rarefying technique developed in 1968 by Howard Sanders[453]. Rarefaction is based on the construction of rarefaction curves. These plot the number of taxa as a function of the number of samples. Usually the initial curve of the rarefaction plot is steep as the most common taxa are readily found, but the curve will plateau off as only the rare taxa remain to be sampled. The rarefaction curves are generated by randomly re-sampling the pool of samples multiple times and plotting the average number of taxa found in each sample.

Whilst this methodology is straight forward and easy to understand, make comparisons and visualise complex datasets produced from high-throughput sequencing, it also simplifies to the lowest sample count within the population and adds noise within the data through the random subsampling[454].

The DESeq2 R package[378] allowed for negative binomial normalisation for abundance data without the requirement of rarefaction. This enabled the comparison of taxonomic abundances between two categorical groups without the addition of noise or omission of valid data.

Log2 fold analysis identified those OTUs that were significantly elevated or decreased relative to NEC status. The results of this output were then used as a training set for Random Forest modelling to identify those OTUs that had the greatest ability to discriminate between NEC and controls. Random Forest results were summarised by Mean Decreased Accuracy and the Mean Decreased Gini Index; in both metrics, the greater the value, the more important the taxa were in the model.

The Mean Decreased Accuracy is computed from permuting the "out of the bag" (OOB) data. For each tree, the prediction error on the OOB portion of the data is recorded (error rate). The same is then done after permuting each predictor variable. The difference between the two are then averaged over all the trees and normalised by standard deviation of the differences. If the standard deviation of the differences is equal to 0 for a variable, then the division is not done. This represents the mean decrease in accuracy of a model without this variable included, and effectively calculates the accuracy of the predictive OTU.

The Gini Index is the total decrease in node impurities from splitting on the variable, averaged over all the trees. It is used to describe the overall explanatory power of the variables, i.e. are they all equally important or does one have greater explanatory value. This yields an overall sum of the explanatory relationships between the variables selected.

The combination of these strategies enabled the identification of taxa that were consistently observed to be significantly different between control and NEC samples, without the loss or distortion of count data, and to establish predictive capability of those taxa using Random Forest machine learning. Analysis focused at the genus level as this taxonomic rank was shown by PERMANOVA to be have the most significant difference between NEC and control sample community compositions. Both subset regression and NMDS analysis indicated that feeding regime and method of delivery were key criteria in the distinction of beta-diversity trends and community compositions for NEC and control subject samples. Further analysis into these differences focused on subgroups consisting of more than three subjects per group; all infants fed both formula and breast milk, regardless of delivery method, in addition to infants fed exclusively breast milk and delivered vaginally. These methods were used initially at the genus level but analysis was also performed at the species level to establish potential pathogenic or probiotic species consistent with the occurrence or absence of NEC. Analysis was compared between subsets and the population of samples to identify whether there was an improvement in the detection of NEC based on taxonomic genera/species relative to samples from all the subsets investigated.

## 5.2 Infants Fed Formula and Breast Milk Delivered by Caesarean Section

### 5.2.1 Significantly Different Genera Between Necrotising Enterocolitis and Control Samples

When sample taxonomic abundances were normalised and summarised by log2 fold change, that is an increase of 100% of the original abundance observed. In total 10 genera were seen to have significant differences between NEC and control subjects. Three of these differences had high mean abundances (Figure 20); these were associated with *Bifidobacterium*, *Streptococcus* and *Proteus* (Table 15).

*Bifidobacterium* was seen to have a much greater mean abundance (2950) relative to all other genera that were significantly different. *Proteus* and *Streptococcus* were also some of the most abundant genera observed in samples, with mean abundances of 390 and 477 respectively. All other significantly different genera were considered to be minor community members with mean abundances lower than 80.

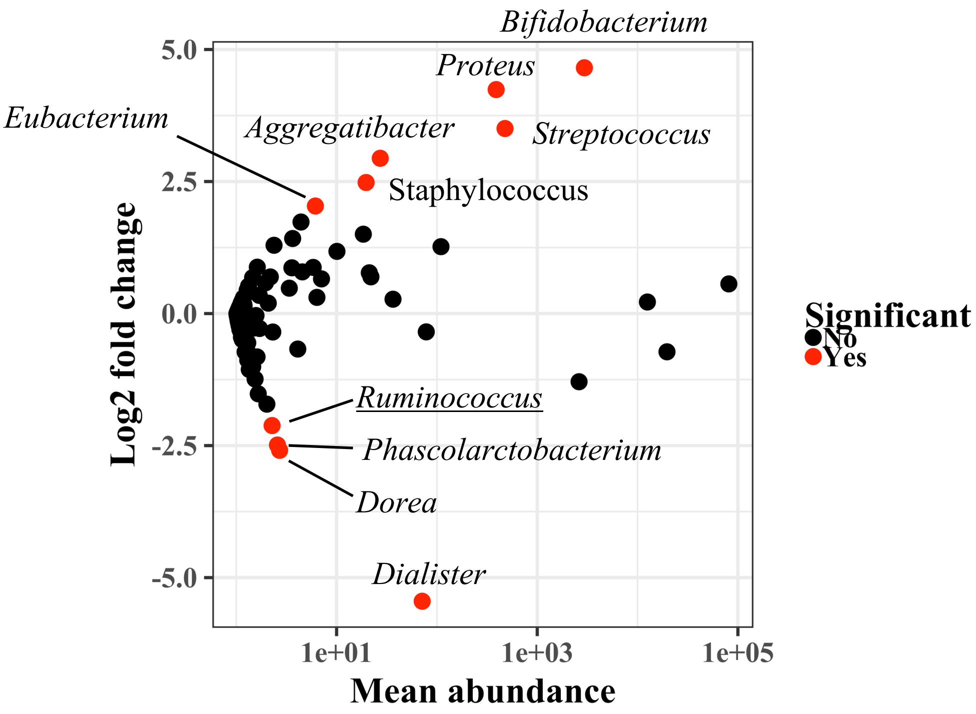


Figure 20 Genera with log2 fold differences between samples from NEC and control subjects delivered by caesarean section fed formula and breast milk.

*Dialister* was the most significantly different genus between NEC and control subject samples and was seen to be elevated in NEC samples (P-adjusted = $2.685\text{x}10^{-33}$) (Figure 21). *Bifidobacterium* (P-Adjusted = $6.62\text{x}10^{-17}$), *Streptococcus* (P-Adjusted = $2.56\text{x}10^{-11}$) and *Staphylococcus* (P-Adjusted = $1.54\text{x}10^{-9}$) had significantly greater abundances in control samples. Proteus appeared to have very similar mean abundances but with a subset of control samples with greater abundances relative to the rest of the population. This resulted in a significant P-adjusted value that suggests the populations are significantly different, however the majority of samples appeared to have similar abundances between NEC and control subjects.

These P-adjusted values were calculated as a population without factoring in the age at sampling, which is known to be a key factor in the development and colonisation of the gut microbiota. Therefore, further assessment focused on the changes occurring relative to the age that an infant was sampled.

Table 15 Summary of all genera with log2 fold significant differences for normalised abundances of NEC and control subject samples.

| **Genera** | **Base Mean** | **Log2 fold Change** | **P-Adjusted** |
|---|---|---|---|
| *Dialister* | 71.32 | -5.45 | $2.69 \times 10^{-33}$ |
| *Dorea* | 2.71 | -2.59 | $1.13 \times 10^{-18}$ |
| *Bifidobacterium* | 2949.49 | 4.65 | $6.62 \times 10^{-17}$ |
| *Phascolarctobacterium* | 2.57 | -2.49 | $9.85 \times 10^{-17}$ |
| *Ruminococcus* | 2.27 | -2.12 | $4.42 \times 10^{-13}$ |
| *Proteus* | 389.54 | 4.24 | $3.43 \times 10^{-12}$ |
| *Streptococcus* | 477.23 | 3.50 | $2.56 \times 10^{-11}$ |
| *Staphylococcus* | 19.69 | 2.48 | $1.54 \times 10^{-9}$ |
| *Aggregatibacter* | 27.24 | 2.94 | $5.78 \times 10^{-9}$ |
| *Eubacterium* | 6.14 | 2.04 | $2.05 \times 10^{-6}$ |

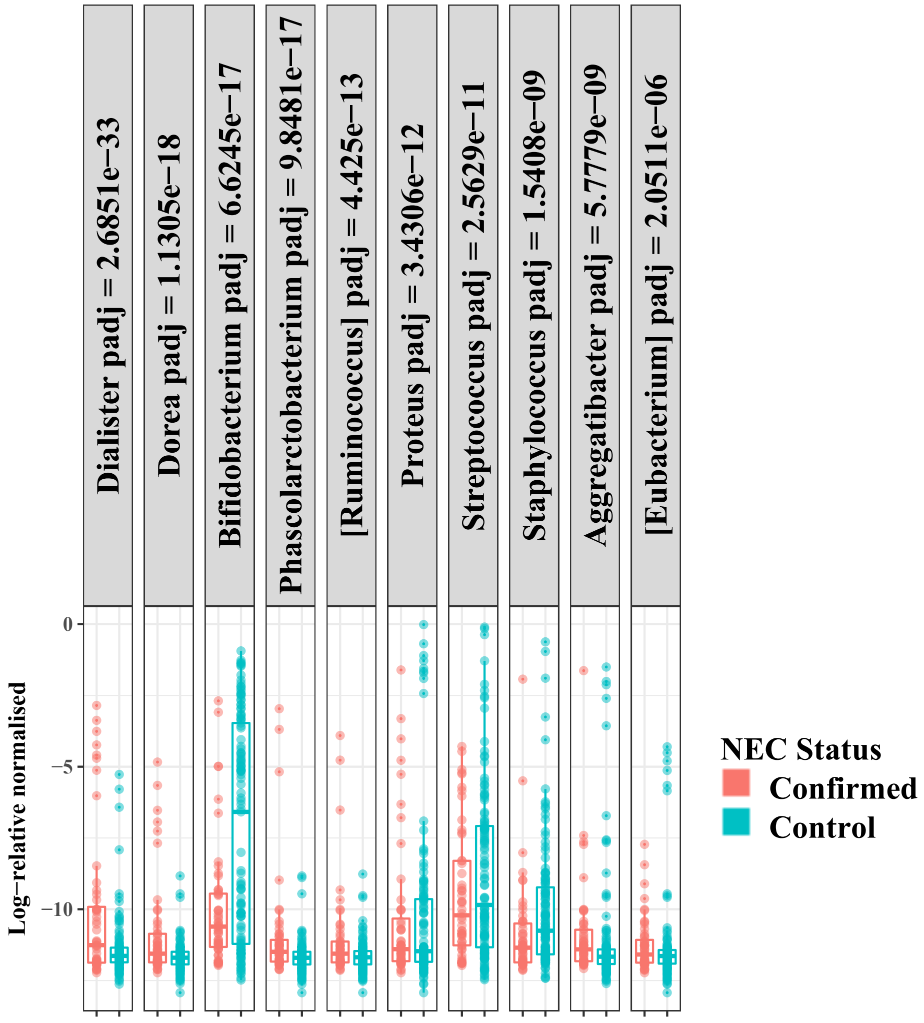


Figure 21 Genera with log2 fold differences between samples from NEC and control subjects fed formula and breast milk and delivered by caesarean section.

### 5.2.2 The presence of Different Genera between Necrotising Enterocolitis and Control Samples over Time

Initially, *Aggregatibacter* was seen to have similar mean abundances at earlier ages in both control and case samples over time. After 30 days of age controls, samples maintained a lower mean abundance of *Aggregatibacter* compared to NEC samples (Figure 22: A). However, across the sampling period the standard error consistently showed an overlap between the two groups, indicating that the overall the abundances were similar.

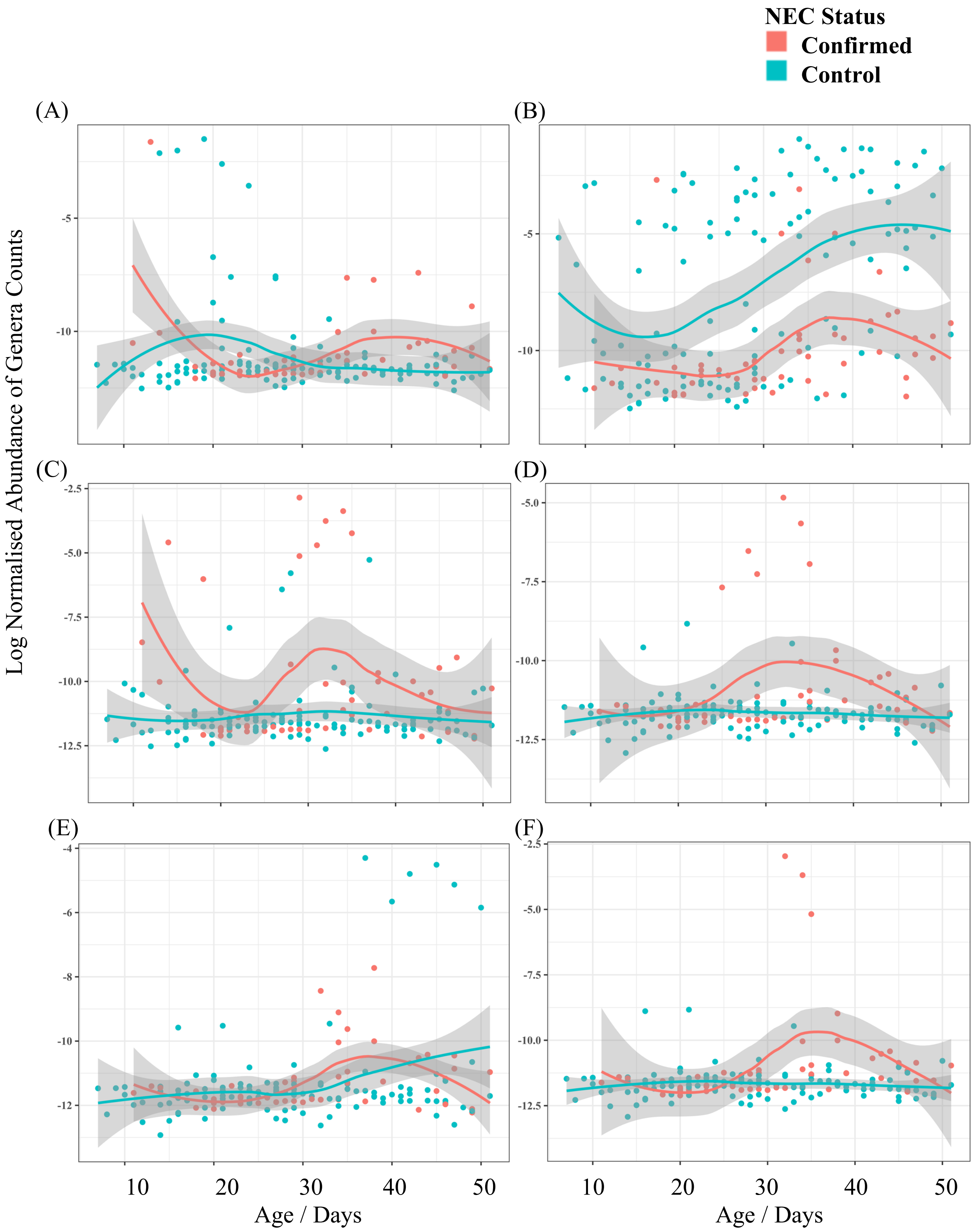
*Bifidobacterium* was the only genus observed to have greater normalised abundances in samples from control subjects relative to NEC subjects. In both case and control samples there was a positive correlation with age and the abundance of this genus. The increase in abundance occurred at an earlier age in control subjects (~15 days) compared to NEC subjects (~28 days). NEC samples were seen to decline in abundance after ~35 days of age (Figure 22: B).

*Dialister* was seen to fluctuate in NEC samples over time, with an initially high abundance at earlier ages which declined up to ~25 days of age. Between 25 - 35 days, NEC samples increased in mean abundance before returning to similar levels to those observed in controls by day 50. In contrast, controls maintained a very low, stable abundance over the duration of sampling (Figure 22: C).

The mean abundance of *Dorea* in NEC samples showed a continued increase from the start of sampling up to ~35 days of age, after which there was a decline to similar levels observed in control samples. Control samples maintained a very low, stable abundance similar to that observed for *Dialister* (Figure 22: D).

The mean abundance of *Eubacterium* in both control and NEC samples was similar over the duration of sampling, with overlapping standard errors. The normalised abundances in samples from case and controls were seen to increase gradually over time, although there was some evidence that NEC samples declined in abundance after 40 days of age (Figure 22: E).

Up to ~25 days of age, the mean abundance of *Phascolarctobacterium* was seen to be very low and stable in both case and control samples. After 25 days of age, NEC samples showed an increase in mean abundance up to ~35 days of age, after which there was a decline to similar levels observed in control samples. Control samples maintained the same level of abundance over the duration of sampling (Figure 22: F).



[Figure continues onto the next page]

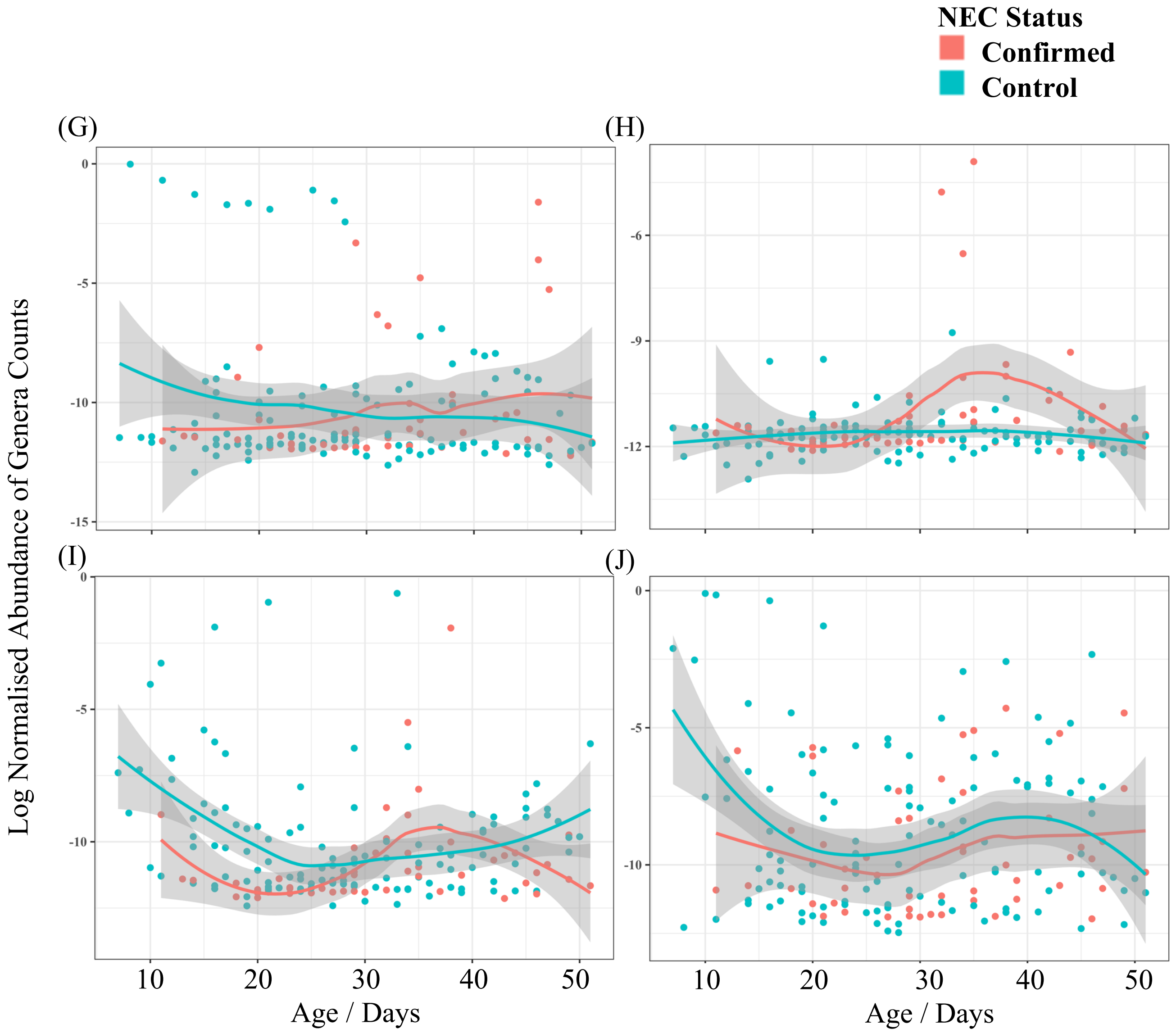


Figure 22 Nonparametric regression analysis for normalised log-abundances for genera observed to be significantly different between NEC and control samples, plotted against the age at sampling and coloured by NEC status. (A) *Aggregatibacter*. (B) *Bifidobacterium*. (C) *Dialister*. (D) *Dorea*. (E) *Eubacterium*. (F) *Phascolarctobacterium*. (G) Proteus. (H) *Ruminococcus*. (I) *Staphylococcus*. (J) *Streptococcus*.

The mean abundance levels for Proteus were observed to maintain a stable level over the duration of sampling for both NEC and control samples. There was some evidence of a slow decline in abundance for controls, and a similarly slow increase in abundance for NEC samples. However, the overlap of standard error over the duration of sampling suggested that these differences would not differentiate between the case and controls (Figure 22: G).

The normalised abundances for *Ruminococcus* in both NEC and control samples taken at early ages were similar up to 20 days of age. After 20 days, NEC samples saw an increase in the abundance of *Ruminococcus* up to ~35 days of age; returning to similar levels observed in controls by day 50. In contrast, control samples maintained the same normalised abundance level over the duration of sampling (Figure 22: H)

The mean abundance of *Staphylococcus* appeared to be greater in controls at earlier ages, but both case and control samples showed a similar decline in abundance after the start of sampling. NEC samples were seen to increase in abundance after 20 days of age up to ~38 days, before declining from then on. *Staphylococcus* in control samples were seen to increase in abundance after 25 days of age but at a slower rate relative to NEC samples. The standard errors between the case and control samples overlapped for much of the sampling duration, again suggesting that as a population there was little differentiation between the samples (Figure 22: I).

Over the duration of sampling the changes in mean abundance of *Streptococcus* in both case and control samples exhibited similar trends, with a decline in abundance up to 30 days of age followed by a gradual increase in abundance. There was some evidence for a decline in control samples after 40 days of age. The mean abundances for both groups were similar over the duration of sampling and the overlap in standard error indicated that this genus would not provide a clear means of differentiation (Figure 22:I).

*Bifidobacterium* was the only genus observed to be significantly elevated in control samples over the duration of sampling and maintained a high mean abundance, as demonstrated by the lack of overlapping standard error in the graph. *Dialister*, *Dorea*, *Phascolarctobacterium* and

*Ruminococcus* were all observed to show distinct peaks in NEC samples at later ages however they were also heavily influenced by a small number of samples with high abundances.

### 5.2.3 Random Forest Analysis of Genera Predictive of Necrotising Enterocolitis

Using machine learning techniques it was possible to identify complex patterns in the different abundances of bacteria that would otherwise be extremely time consuming and prohibitively difficult to perform manually. Using Random Forest it was possible to attribute a schema for the different bacterial abundances and a success rate in the identification of NEC subjects.

Confusion matrices for Random Forest models were generated using the ten significantly different genera and compared against models using all genera (total number of genera = 221) observed within the samples. The model that used significantly different genera had a 9% improvement in correctly identifying samples from NEC subjects compared to models utilising all the genera observed within this subset of infants (Table 16).

Table 16 Summary Tables for Random Forest model predictions of the NEC status of infants using the genera composition of samples. 1,500 trees were constructed for both models. (A) Using only genera identified as being significantly different between NEC and controls, three variables tried at each split. (B) Using all genera identified within the communities, twelve variables tried at each split

A)

| **Confusion matrix** | **Confirmed** | **Control** | **Class Error** |
|---|---|---|---|
| Confirmed | 35 | 17 | 0.33 |
| Control | 5 | 116 | 0.04 |
| OOB estimate of error rate | 0.13% | | |

B)

| **Confusion matrix** | **Confirmed** | **Control** | **Class Error** |
|---|---|---|---|
| Confirmed | 30 | 22 | 0.42 |
| Control | 8 | 113 | 0.07 |
| OOB estimate of error rate | 0.17% | | |

All genus with log2 fold differences were seen to be important in the Random Forest model generation, however, *Bifidobacterium* demonstrated the most association with increased model prediction in both the accuracy and the Gini Index scores (Figure 23). *Bifidobacterium*, which was associated with increased abundances in control samples, reduced the model accuracy by ~45%. *Dialister,* which was shown to be associated with NEC subjects, was of similar importance and the second most influential genus in the model generation. The mean decreased Gini Index scores also indicated that *Staphylococcus* was important in the prediction accuracy of the model.

These scores were reinforced when the model tree structure was represented graphically. Samples with normalised abundances of *Bifidobacterium* > -9.28 and *Staphylococcus* > -11.54 were observed to be associated with control subjects only. Infants that had *Dorea* abundances of < -11.67, even when *Staphylococcus* was in abundance < -11.54, were also observed not to be associated with NEC. Even if infants were seen to have low abundances of *Bifidobacterium* (< -9.28), if they also maintained low abundance of *Aggregatibacter* (< -11.97) and high abundances of *Staphylococcus* (>-12.08), they were not associated with NEC (Figure 24).

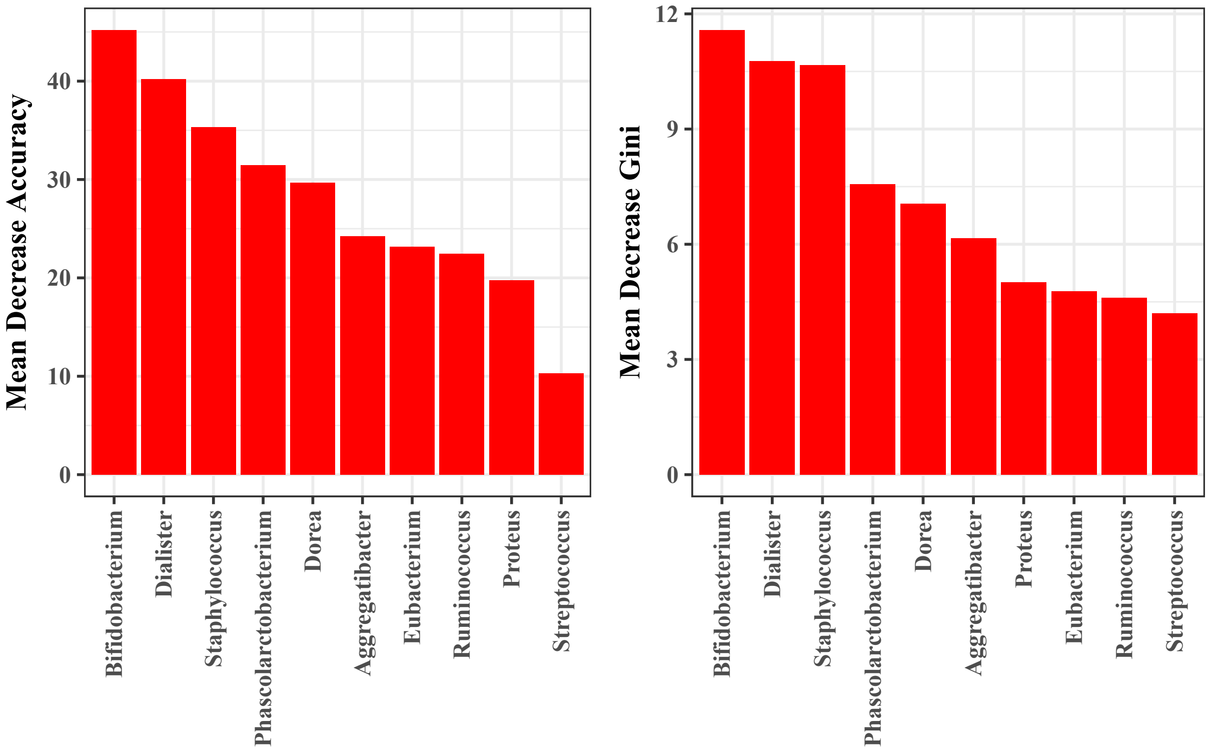


Figure 23 Random Forest summary graphs describing the ability of genera that were seen to be significantly different between NEC and control samples in formula and breast milk delivered by caesarean section to predict the NEC status of an infant. (A) Mean Decreased Accuracy for phyla with log-2 fold. (B) Mean Decrease Gini Index scores.

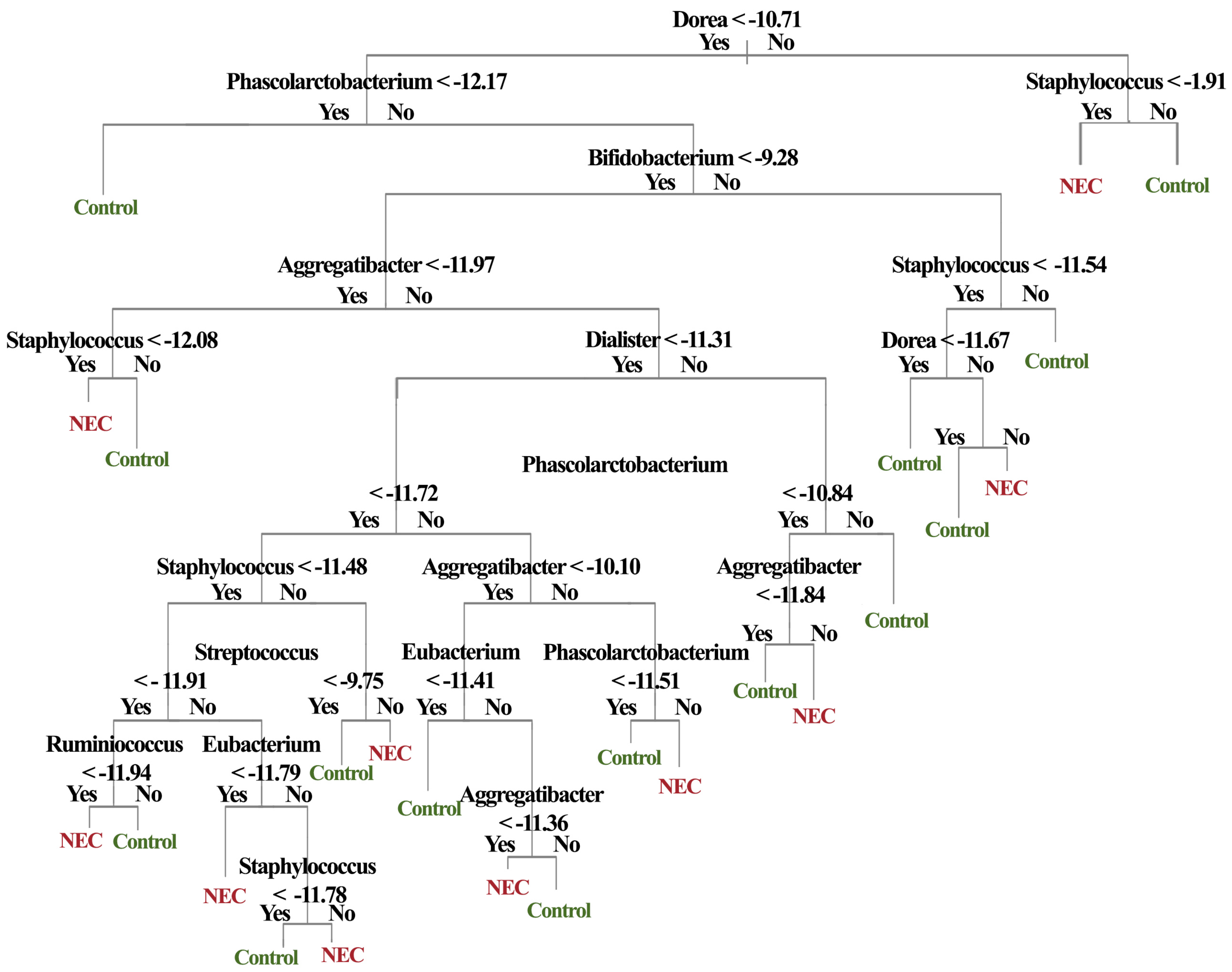


Figure 24
Tree representation of Random Forest decision based prediction of NEC status using the log-normalised genera abundance of taxa observed to be significantly different in abundance between case and control samples. Values in the tree represent the log normalised value that a given decision is made upon. There is a complex hierarchy of decisions in the prediction of NEC status with an error rate of 0.33. The importance of *Bifidobacterium* is represented with three of four forks associated with log scores greater then -9.28 resulting in a control diagnosis. This diagnosis is also associated with a high abundance of *Staphylococcus*.

### 5.2.4 Diagnostic Species

Many studies focus analysis at the genus level, however very few give reasons for doing so. Using the Bray-Curtis measure of dissimilarity, PERMANOVA analysis of the NMDS for the community compositions at the different taxonomic levels showed that the greatest distinction between NEC and control samples was observed at the genus level. This would provide the best level to identify taxa that discriminated between NEC and control samples, therefore it was the focus of further analysis. However, there is evidence that this leads to over simplification and that at the sub-genus level there have been observed associations, particularly with *Prevotella* and *Bacteroides*, to specific dietary patterns[455].

This subgroup was analysed for diagnostic species that could be indicative of NEC or associated with the genera that were seen to be significantly different between NEC and control samples. Three species were observed to be significantly different between NEC and control subjects (Table 17), however, only *Clostridium butyricum* was observed to differ in normalised abundance when age at sampling was factored in (Figure 25). This species was seen to be elevated in NEC subjects, though this was specifically associated with the one subject UHCW062 and therefore not representative of the whole NEC population.

Table 17 Summary of all species with log2 fold significant differences for normalised abundances of NEC and control subject samples.

| Species | Base Mean | Log2 fold Change | P-Adjusted |
|---|---|---|---|
| *Clostridium butyricum* | 11.98 | -4.56 | $2.69 \times 10^{-35}$ |
| *Streptococcus anginosus* | 126.51 | 6.58 | $7.17 \times 10^{-34}$ |
| *Eubacterium dolichum* | 6.11 | 2.19 | $5.67 \times 10^{-7}$ |

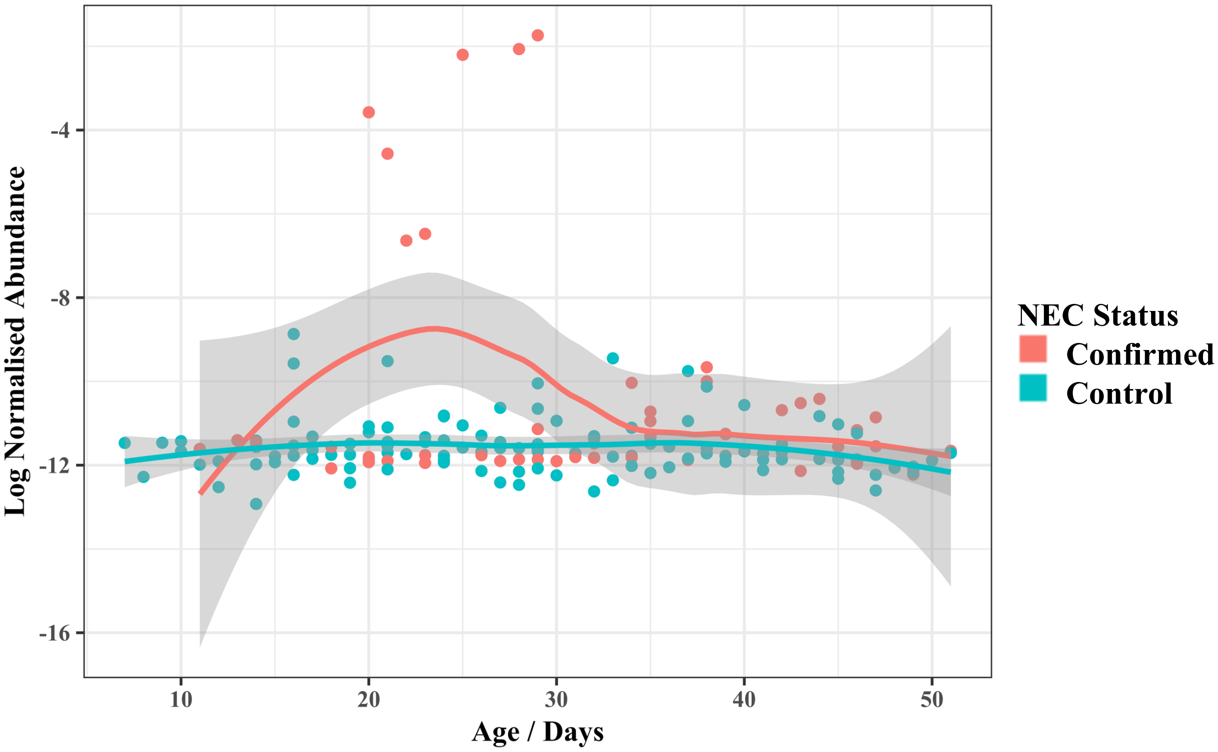


Figure 25 Nonparametric regression analysis for normalised log-abundances for *Clostridium butyricum* observed to be significantly different between NEC and control samples, plotted against the age at sampling and coloured by NEC status.

Using the species that were significantly different between NEC and control samples in the Random Forest model showed that there was a reduction in the ability to identify samples from NEC subjects by 9% compared to the model that used genera that were significantly different (Table 18: A). The error rate of the Random Forest model was increased to 48% when utilising all the species observed within samples (Table 18: B). This suggested that species' differences were not as useful in discriminating between NEC and control samples compared to the predictive capabilities of models using genera.

Table 18 Summary Tables for Random Forest model predictions of the NEC status of infants using the species composition of samples. 1,500 trees were constructed for both models. (A) Using only species identified as being significantly different between NEC and controls, one variable tried at each split. (B) Using all species identified within the communities, ten variables tried at each split.

(A)

| **Confusion matrix** | **Confirmed** | **Control** | **Class Error** |
|---|---|---|---|
| Confirmed | 30 | 22 | 0.42 |
| Control | 17 | 104 | 0.14 |
| OOB estimate of error rate | | 22.54% | |

(B)

| **Confusion matrix** | **Confirmed** | **Control** | **Class Error** |
|---|---|---|---|
| Confirmed | 27 | 25 | 0.48 |
| Control | 9 | 112 | 0.07 |
| OOB estimate of error rate | | 19.65% | |

## 5.3 Infants Fed Formula and Breast Milk Delivered Vaginally

### 5.3.1 Significantly Different Genera Between Necrotising Enterocolitis and Control

NEC and control samples from infants that were delivered vaginally and fed both formula and breast milk were seen to have four significantly different genera when all samples were compared (Figure 26). *Bifidobacterium* and *Bacteroides* had the highest mean abundances (8678 and 6652, respectively) and log2 fold changes (9.36 and 7.90, respectively) of all genera. *Veillonella* was also seen to have a high mean abundance (6971) but with the lowest difference represented by log2 fold change (2.22). *Megamonas* was seen to be the third most significant genus of the four but had the lowest mean abundance of all the significant genera (16) (Table 19).

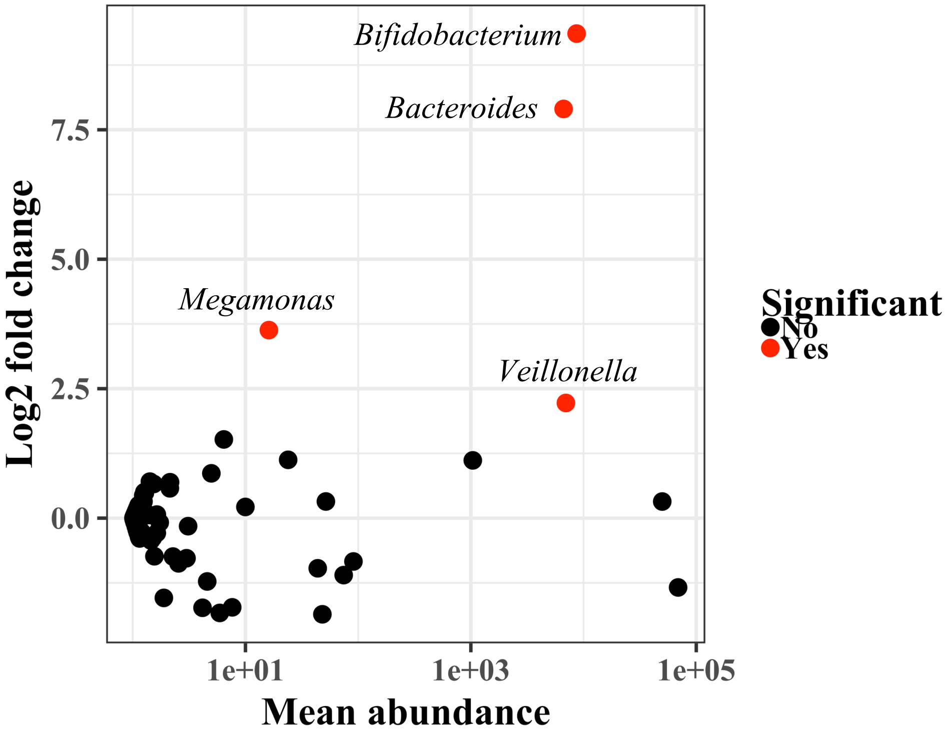


Figure 26 Genera with log2 fold differences between samples from NEC and control subjects fed formula and breast milk and delivered vaginally.

Table 19 Summary of all genera with log2 fold significant differences in normalised abundances between

| Genera | Base Mean | Log2 fold Change | P-Adjusted |
|---|---:|---:|---:|
| *Bifidobacterium* | 8678.09 | 9.36 | $1.72 \times 10^{-86}$ |
| *Bacteroides* | 6652.56 | 7.90 | $9.17 \times 10^{-54}$ |
| *Megamonas* | 16.14 | 3.63 | $6.25 \times 10^{-17}$ |
| *Veillonella* | 6971.39 | 2.22 | $9.00 \times 10^{-6}$ |

When time was not factored into the analysis, samples from control subjects were seen to have greater means for log normalised abundances in *Bifidobacterium* and *Veillonella*. NEC subjects were seen to have a greater mean abundance for *Bacteroides. Megamonas* had very similar means for both case and control samples, however a small subset of control samples were seen to have elevated abundances relative to the rest of the cohort, this led to the significance observed in the P-adjusted score. Overall, the majority of samples had similar abundances of *Megamonas* (Figure 27).

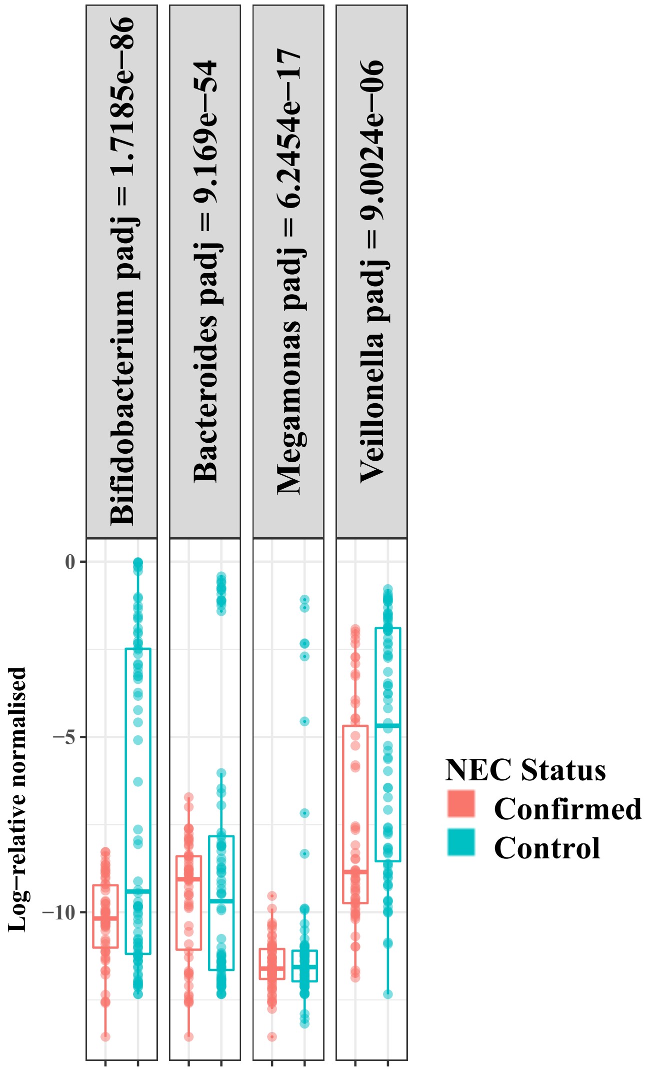


Figure 27 Genera with log2 fold differences between samples from NEC and control subjects delivered vaginally fed formula and breast milk.

### 5.3.2 Significantly Different Genera between Necrotising Enterocolitis and Control Samples over Time

The normalised abundances of *Bacteroides* in both NEC and control samples showed very similar trends and normalised abundances over time, as characterised by the overlapping standard errors. However, an increase in abundance was seen for control infants in samples taken after 50 days of age. NEC samples did not show this increase in abundance and had similar levels on day 75 as on day 50 (Figure 28:A).

Overall, *Bifidobacterium* was observed to have consistently greater normalised abundances in control samples for the duration of sampling and exhibited an increase from the start of sampling up to ~25 days of age, before declining to similar values observed in NEC samples. At day 50, the abundance of *Bifidobacterium* in control samples rose again. By day 75, samples had the greatest mean abundance observed for control samples over the course of sampling (Figure 28:B). Samples from NEC subjects showed little to no change over time aside from a slight decrease at ~40 days of age. The standard error for the mean abundance of *Bifidobacterium* was clearly distinct aside from days 30-40. This suggests that the initial abundance and abundances at later ages differed between case and controls, although there were some control samples with abundances of *Bifidobacterium* that were as low or lower than those observed in NEC samples.

*Megamonas* showed no difference in the trends or abundance of NEC and control samples over the duration of sampling, with the exception of an elevated initial abundance in control samples at the start of sampling (Figure 28:C). This initially high abundance value was skewed by samples from UHCW078 only. UHCW078 had a higher than average gestation duration (31 weeks) and birthweight (1,880g) relative to the control cohort (27 weeks and 1,419g) which could have accounted for the initial high abundance of *Megamonas*.

*Veillonella* maintained a higher abundance in control samples and showed a clear, steady increase over time (Figure 28:D). NEC samples fluctuated over the course of sampling, with an initial increase up to ~30 days of age, after which there was a dramatic decline up to day 50. The mean abundance for *Veillonella* appeared to stabilise after this time point. After ~38 days of age the standard error for each group showed little crossover, suggesting that as a population, control subjects maintained a higher normalised abundance of *Veillonella* at later ages compared to NEC subjects.

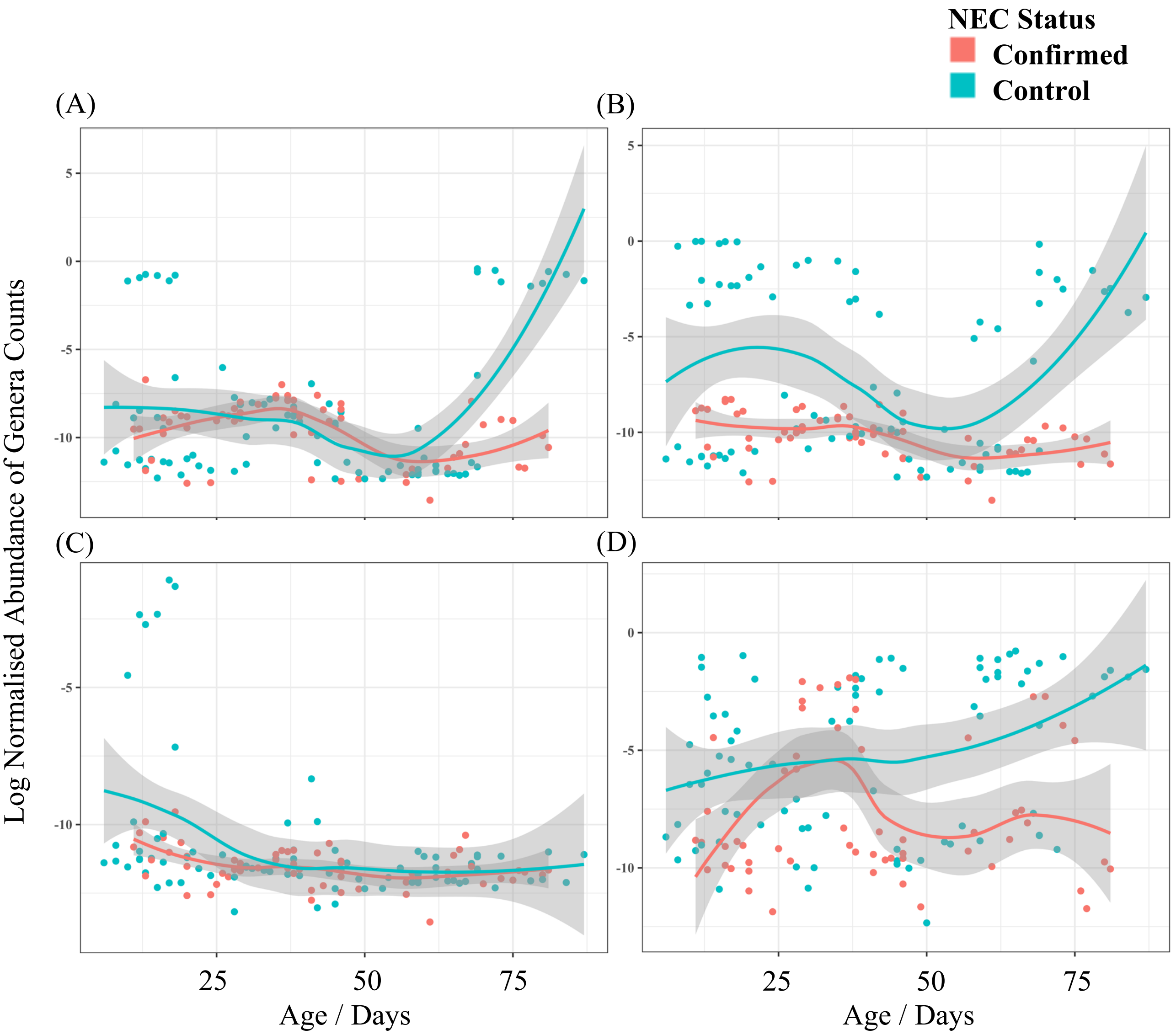


Figure 28 Nonparametric regression analysis for normalised log-abundances for genera observed to be significantly different between NEC and control samples, plotted against the age at sampling and coloured by NEC status. (A) *Bacteroides*. (B) *Bifidobacterium*. (C) *Megamonas*. (D) *Veillonella.*

### 5.3.3 Random Forest Analysis of Genera Predictive of Necrotising Enterocolitis

Confusion matrices for Random Forest models were generated using the four significantly different genera and compared against models using all genera observed within the samples (total number of genera = 166). The model using significantly different genera was observed to have a much greater error rate when predicting NEC samples (24%) when compared to the model using all genera observed in the samples (12%). Both models from samples within this subset could discern samples from NEC subjects with lower class error scores than those observed for infants delivered by caesarean section on the same feeding regime.

Table 20 Summary Tables for Random Forest model predictions of the NEC status of infants using the genera composition of samples. 1,500 trees were constructed for both models. (A) Using only genera identified as being significantly different between NEC and controls, three variables tried at each split. (B) Using all genera identified within the communities, twelve variables tried at each split.

(A)

| **Confusion matrix:** | **Confirmed** | **Control** | **Class Error** |
|---|---|---|---|
| Confirmed | 44 | 14 | 0.24 |
| Control | 13 | 69 | 0.16 |
| OOB estimate of error rate | | 19.29 | |

(B)

| **Confusion matrix:** | **Confirmed** | **Control** | **Class Error** |
|---|---|---|---|
| Confirmed | 51 | 7 | 0.12 |
| Control | 7 | 75 | 0.09 |
| OOB estimate of error rate | | 10.00 | |

Bifidobacterium was the most important genus in the distinction of NEC and control samples; this is seen through both the mean decreased accuracy and mean decreased Gini scores for the model constructed from significantly different genera (Figure 29). This genus was seen to have almost double the mean decreased accuracy score relative to *Veillonella*, the next most important genera in the model decision tree.

Further investigation found that an unknown genus was observed to be the most important in the Random Forest tree construction when all genera were considered for the model, however, no further information was available in terms of taxonomic annotation. Additionally, the genus *Pantoea* was seen to be the second most significant taxa in discriminating between NEC and control samples, followed by *Bifidobacteria*.

The importance of *Bifidobacterium* was further reinforced when the Random Forest tree was visualised (Figure 30). Samples with *Bifidobacterium* abundances greater than -8.17 were highly likely to be assigned as controls. After *Bifidobacterium*, *Veillonella* and *Bacteroides* were seen to be important in the assignment of NEC status, specifically, low *Veillonella* (< -1.99) and high *Bacteroides* (> -7.85) abundance.

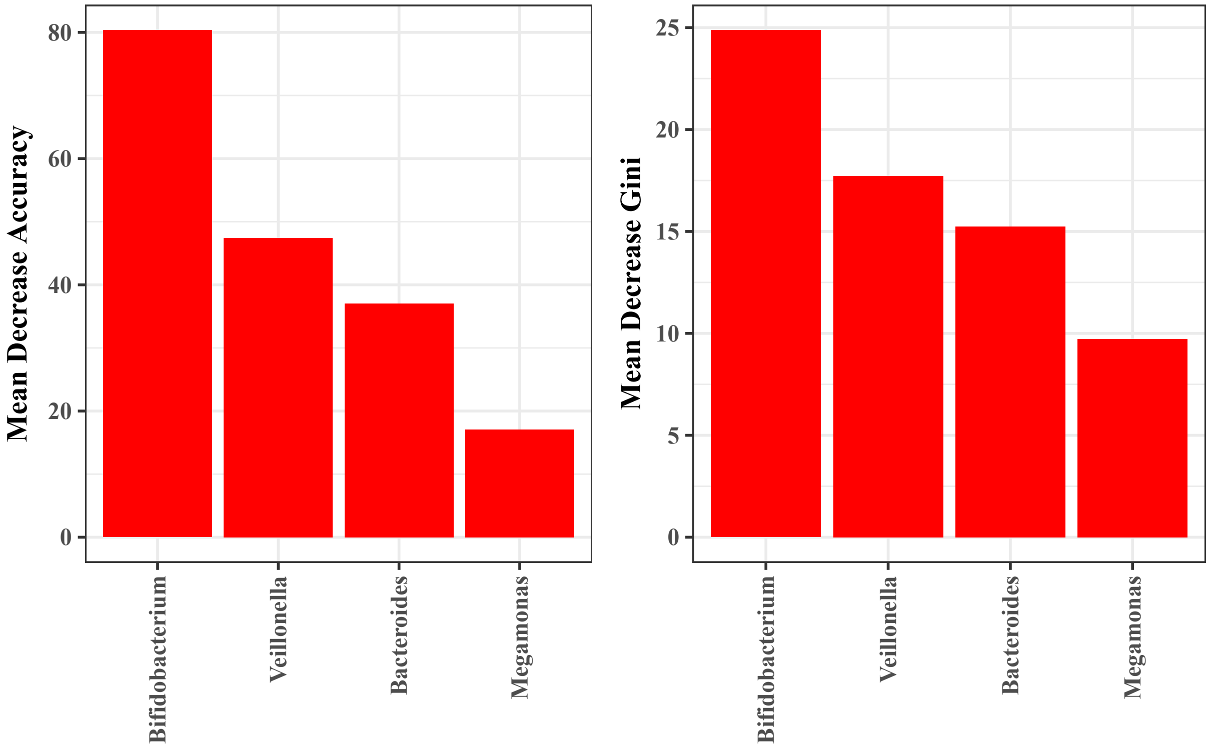


Figure 29 Random Forest Summary graphs describing the ability of genera that were seen to be significantly different between NEC and control samples in formula and breast milk delivered vaginally to predict the NEC status of an infant. (A) Mean Decreased Accuracy for phyla with log-2 fold. (B) Mean Decrease Gini Index scores.

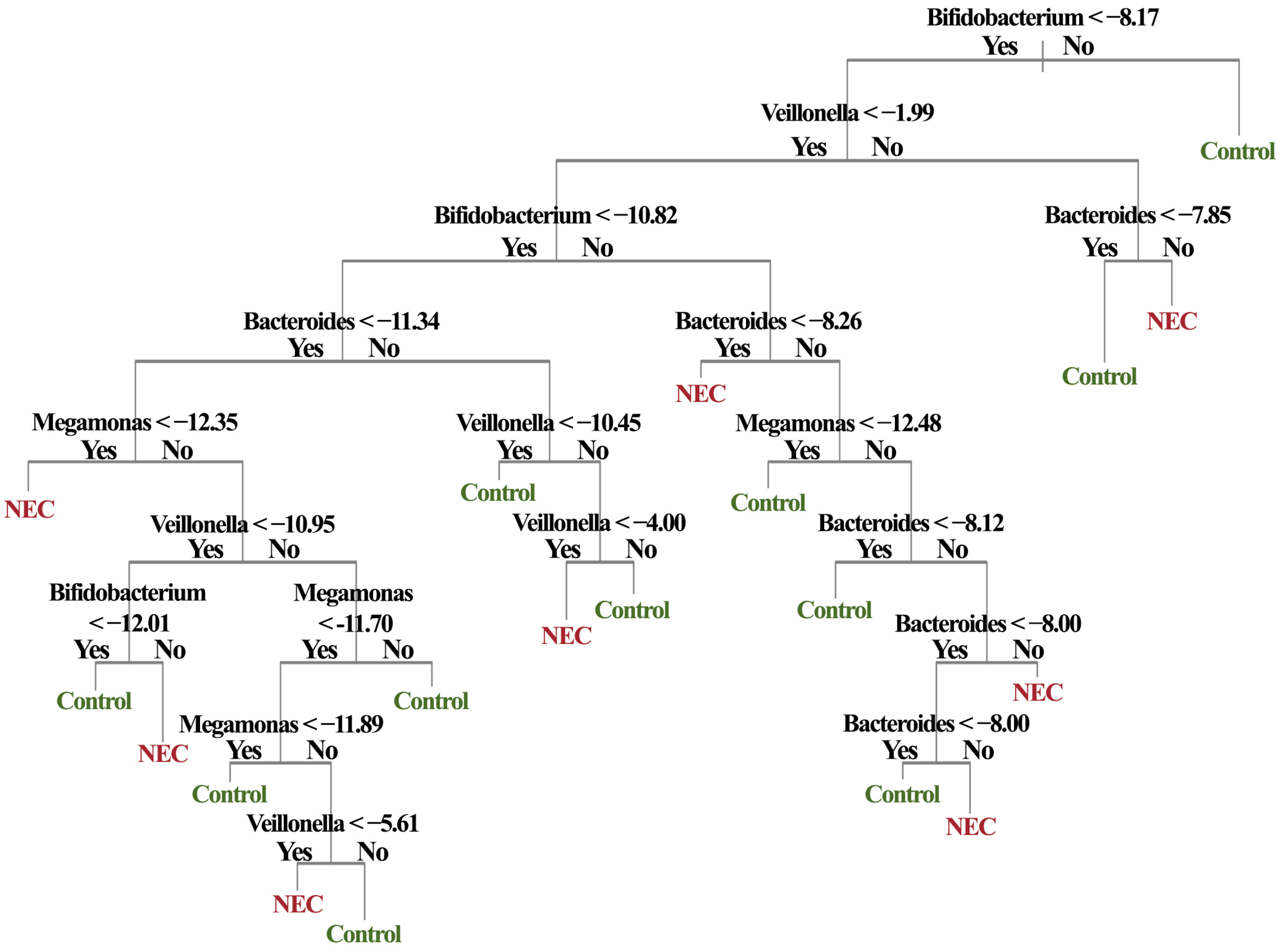


Figure 30 Tree representation of Random Forest decision based prediction of NEC status using the log-normalised genera abundance of taxa observed to be significantly different in abundance between case and control samples. Values in the tree represent the log normalised value that a given decision is made upon. There is a complex hierarchy of decisions in the prediction of NEC status with an error rate of 0.24. Increased abundance of *Bifidobacterium* associated with controls when values were greater than -8.17. High abundances of *Veillonella* (>-1.99) and *Bacteroides* (>-7.85) were associated with NEC samples

### 5.3.4 Diagnostic Species

Three species were identified as being significantly different between NEC and control samples when the age at sampling was not considered in the analysis (Table 21). Only *Streptococcus agalactiae* showed a clear diagnostic signal when plotted over time, however, the standard error and trend lines for the NEC samples were heavily skewed by the NEC subject BWH221 (Figure 31). BWH221 was the only subject to have normalised sample abundances between -2.5 and -7.5 and therefore did not reflect the NEC population accurately. The species was therefore not considered to be a diagnostic signal.

Table 21 Summary of all species with log2 fold significant differences for normalised abundances of NEC and control subject samples.

| Species | Base Mean | Log2 fold Change | P-Adjusted |
|---|---|---|---|
| *Bacteroides fragilis* | 6626.32 | 8.40 | $5.94 \times 10^{-54}$ |
| *Streptococcus agalactiae* | 11.80 | -4.16 | $1.28 \times 10^{-29}$ |
| *Clostridium perfringens* | 809.04 | 2.85 | $2.84 \times 10^{-8}$ |

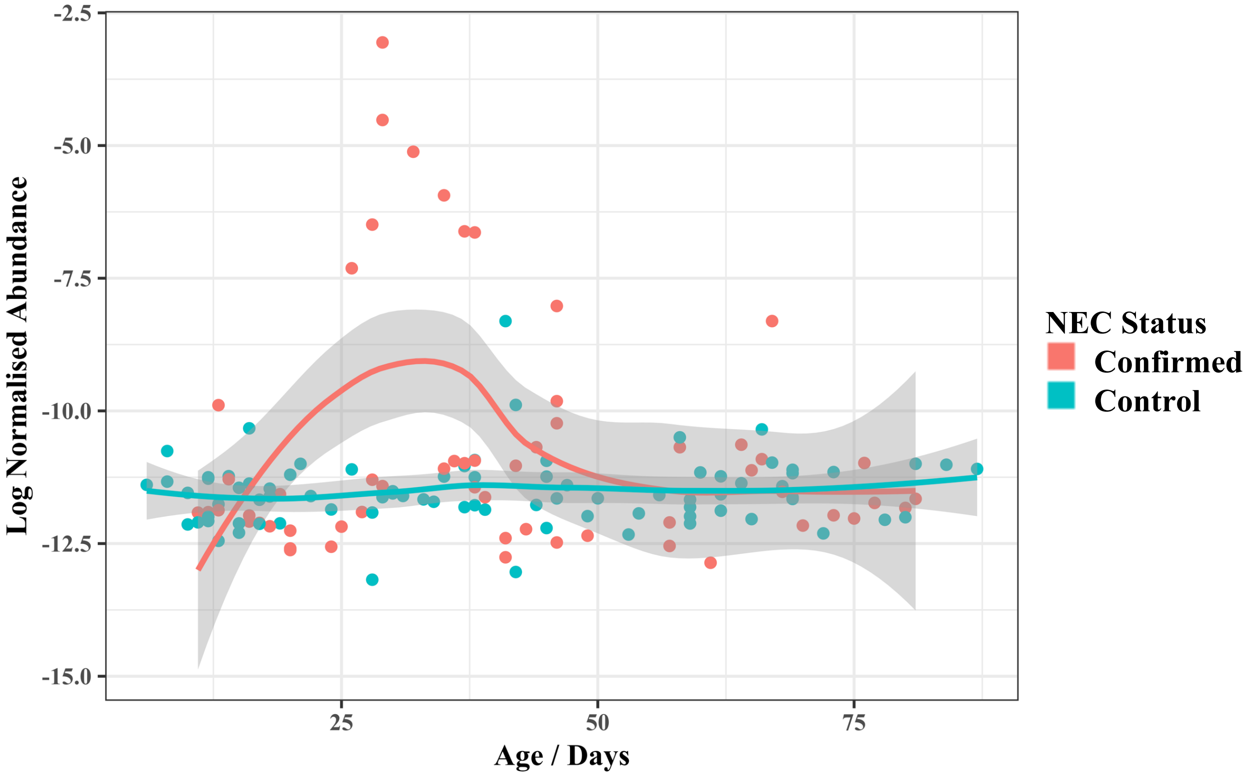


Figure 31 Nonparametric regression analysis for normalised log-abundances for *Streptococcus agalactiae* observed to be significantly different between NEC and control samples, plotted against the age at sampling and coloured by NEC status.

The class error for NEC sample identification was the same when using significantly different species as when using significantly different genera (24%, Table 22: A), however, this error rate was increased when all genera were used in the model construction (Table 22: B). In contrast, when all genera were used to construct a Random Forest model, the error rate was substantially reduced to 12%.

Table 22 Summary Tables for Random Forest model predictions of the NEC status of infants using the species composition of samples. 1,500 trees were constructed for both models. (A) Using only species identified as being significantly different between NEC and controls, one variable tried at each split. (B) Using all species identified within the communities, seven variables tried at each split.

(A)

| Confusion matrix | Confirmed | Control | Class Error |
|---|---|---|---|
| Confirmed | 44 | 14 | 0.24 |
| Control | 15 | 67 | 0.18 |
| OOB estimate of error rate | 20.71% | | |

(B)

| Confusion matrix | Confirmed | Control | Class Error |
|---|---|---|---|
| Confirmed | 43 | 15 | 0.26 |
| Control | 10 | 72 | 0.12 |
| OOB estimate of error rate | 17.86% | | |

## 5.4 Infants Exclusively Fed Breast Milk and Delivered Vaginally

### 5.4.1 Significantly Different Genera Between Necrotising Enterocolitis and Control Samples

Of the five genera observed to have significantly different abundances between NEC and control samples in infants delivered vaginally and fed exclusively breast milk, three were seen to have high log2 fold changes and mean abundances (Figure 32).

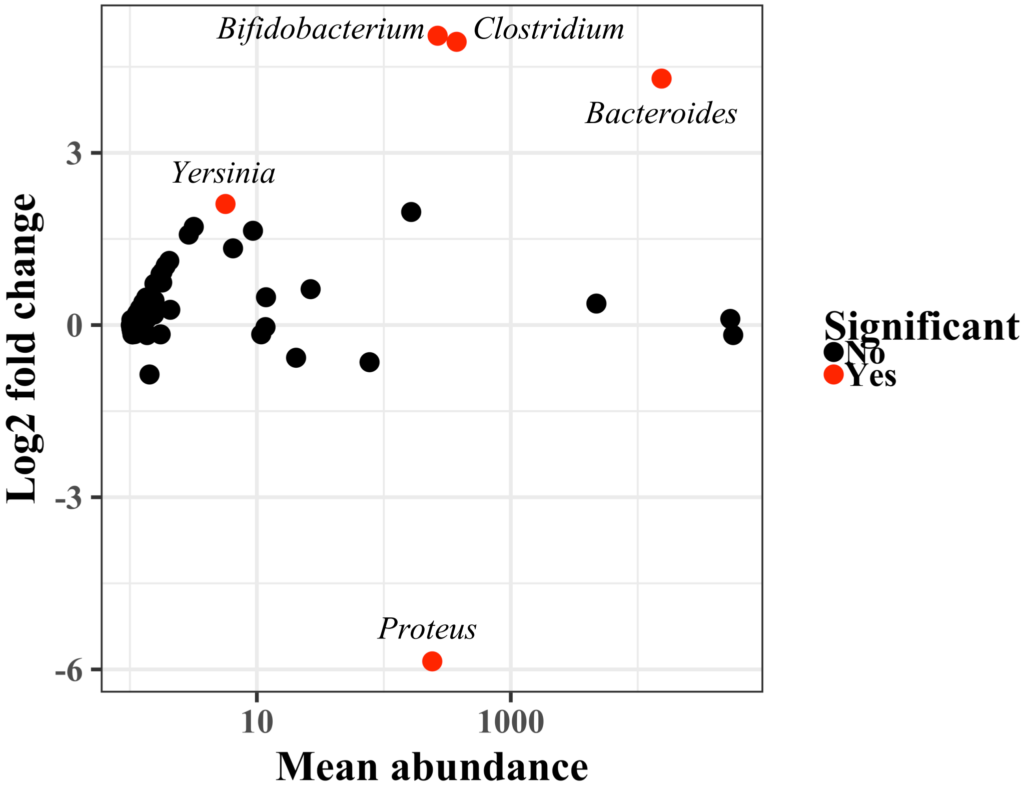


Figure 32 Genera with log2 fold differences between samples from NEC and control subjects exclusively fed breast milk and delivered vaginally. In total four genera were observed to be significantly different between case and control samples within this subset.

*Proteus* was the most significantly different genera between NEC and control subjects when age was not factored into the analysis (P-adjusted = 3.46x10$^{-30}$), but *Bacteroidetes* had the greatest mean abundances across all samples (15403). *Clostridium* and *Bifidobacterium* were also seen to be significantly different and maintain high mean abundances (374 and 265, respectively). *Yersinia* was the genus with the lowest significant difference (5.59x10$^{-5}$), base mean abundance (6), and log2 fold change (2.11) of all significantly different genera (Table 23).

Table 23 Summary of all genera with log2 fold significant differences in normalised abundances between NEC and control samples for all infants fed breast milk exclusively and delivered vaginally. Ordered by adjusted P-value scores.

| **Genera** | **Base Mean** | **Log2 fold Change** | **P-Adjusted** |
|---|---|---|---|
| *Proteus* | 240.72 | -5.86 | $3.46 \times 10^{-30}$ |
| *Clostridium* | 374.43 | 4.93 | $2.88 \times 10^{-18}$ |
| *Bifidobacterium* | 265.33 | 5.04 | $1.73 \times 10^{-17}$ |
| *Bacteroides* | 15402.50 | 4.29 | $3.30 \times 10^{-10}$ |
| *Yersinia* | 5.66 | 2.11 | $5.59 \times 10^{-5}$ |

*Proteus* displayed similar means for both case and control sample normalised abundances, with a marginally greater mean in NEC samples (Figure 33). *Clostridium* appeared to have the greatest difference when samples were analysed as a whole, with a greater mean abundance in control samples. *Bifidobacterium* and *Bacteroides* also had greater mean abundances in control samples, whereas *Yersinia* showed negligible differences, although control samples did appear to have a higher abundance in the 3rd quantile.

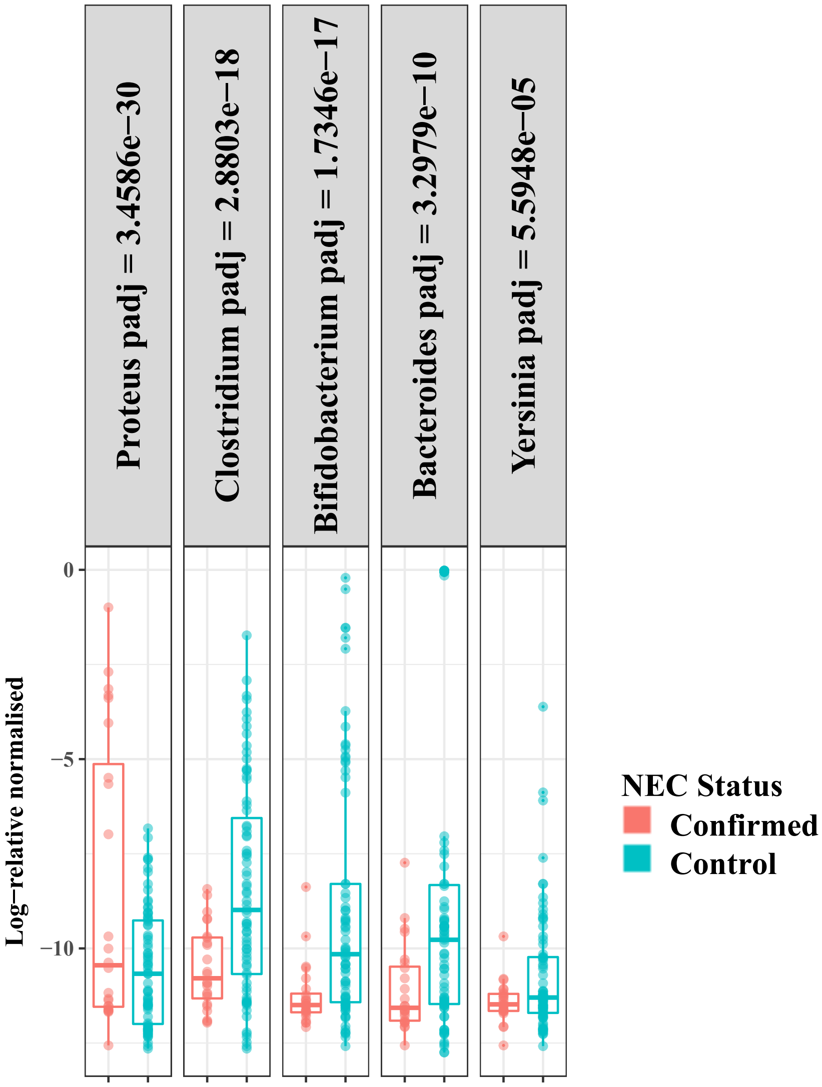


Figure 33 Genera with log2 fold differences between samples from NEC and control subjects exclusively fed breast milk and delivered vaginally.

### 5.4.2 Significantly Different Genera between Necrotising Enterocolitis and Control Samples over Time

When plotted against time, *Bacteroides* showed an increase in abundance for samples from control infants, up to approximately 40 days of age (Figure 34: A). However, this was skewed by one subject (BWH204) which had very high abundances relative to the other controls. BWH204 had a lower than average gestational duration of 28 weeks compared to the control cohort average of 30, though the birthweight of 1,200g was similar to that observed for most control subjects (mean = 1,419g). The only other medical information that could have differentiated this subject was that the vaginal delivery was a breach. If BWH204 was not considered, the abundances in NEC and control samples would be very similar over the duration of sampling.

On average, *Bifidobacterium* abundance was greater in control samples than NEC samples over the duration of sampling. In both control and NEC samples *Bifidobacterium* maintained a stable abundance throughout the sampling period. However, the difference in abundance was not enough to discriminate between NEC and control samples due to the overlapping standard errors (Figure 34: B).

The mean abundance of *Clostridium* was seen to increase in control samples from an early age, up to day 30. After this the mean abundance declined but was relatively stable up to 60 days of age. Within NEC samples there was little change in the abundance of *Clostridium* over time, though it was seen to be lower than the mean normalised abundance in control samples. Whilst there was little overlap in the standard errors of the groups, there were samples from the control group that showed lower abundances than NEC samples at the same time period (Figure 34:C).

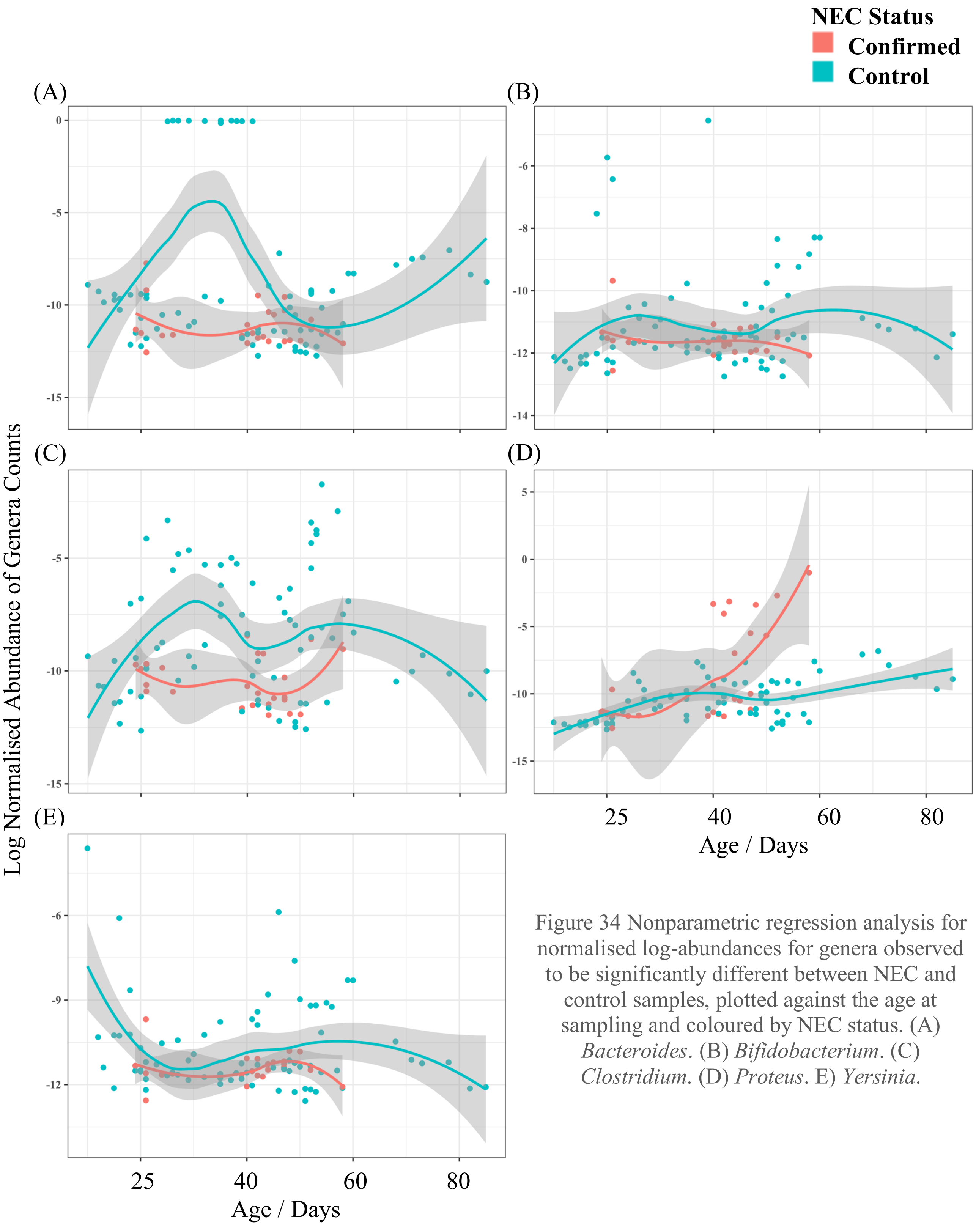


Figure 34 Nonparametric regression analysis for normalised log-abundances for genera observed to be significantly different between NEC and control samples, plotted against the age at sampling and coloured by NEC status. (A) *Bacteroides*. (B) *Bifidobacterium*. (C) *Clostridium*. (D) *Proteus*. E) *Yersinia*.

*Proteus* was seen to increase in both NEC and control samples over the duration of sampling, however the rate of increase was much greater in the NEC samples (Figure 34:D). The initial abundance of *Yersinia* was observed to be much greater in control samples than in NEC samples, however, from the age of 20 days it appeared to maintain a stable abundance at similar levels for samples from both subject groups (Figure 34:E).

Trends or changes observed in samples after 60 days of age were considered inappropriate to analyse due to the lack of comparable NEC samples. This subgroup was limited in the conclusions that could be made when considering the sampling density over time due to the sparsity of samples. This was particularly notable for NEC samples over ~28 and ~35 days of age.

### 5.4.3 Random Forest Analysis of Genera Predictive of Necrotising Enterocolitis

The success rate of Random Forest model predictions for NEC samples was the lowest of all the subgroups analysed (58%). However, this was substantially greater than the accuracy for the model using all genera observed (total number of genera = 154) in the samples for this subgroup (42%) (Table 24).

Table 24 Summary Tables for Random Forest model predictions of the NEC status of infants using the genera composition of samples. 1,500 trees were constructed for both models. (A) Using only genera identified as being significantly different between NEC and controls, two variables tried at each split. (B) Using all genera identified within the communities, twelve variables tried at each split.

A)

| Confusion matrix | Confirmed | Control | Class Error |
|---|---|---|---|
| Confirmed | 14 | 10 | 0.42 |
| Control | 5 | 70 | 0.07 |
| OOB estimate of error rate | 15.15% | | |

B)

| Confusion matrix | Confirmed | Control | Class Error |
|---|---|---|---|
| Confirmed | 10 | 14 | 0.58 |
| Control | 1 | 74 | 0.01 |
| OOB estimate of error rate | 15.15% | | |

*Proteus* was observed to be the most significant genus within the Random Forest models, accounting for more than a 35% decrease in accuracy when removed (Figure 35). *Clostridium* was the next most significant genus in distinguishing NEC from control samples, accounting for an approximate decrease in accuracy of 30% when removed. Whilst *Bifidobacterium* was not observed to account for as much accuracy in the model (~20%), it was seen to have a highly similar mean decreased Gini score, suggesting that it was of similar importance in partitioning the data as *Clostridium*.

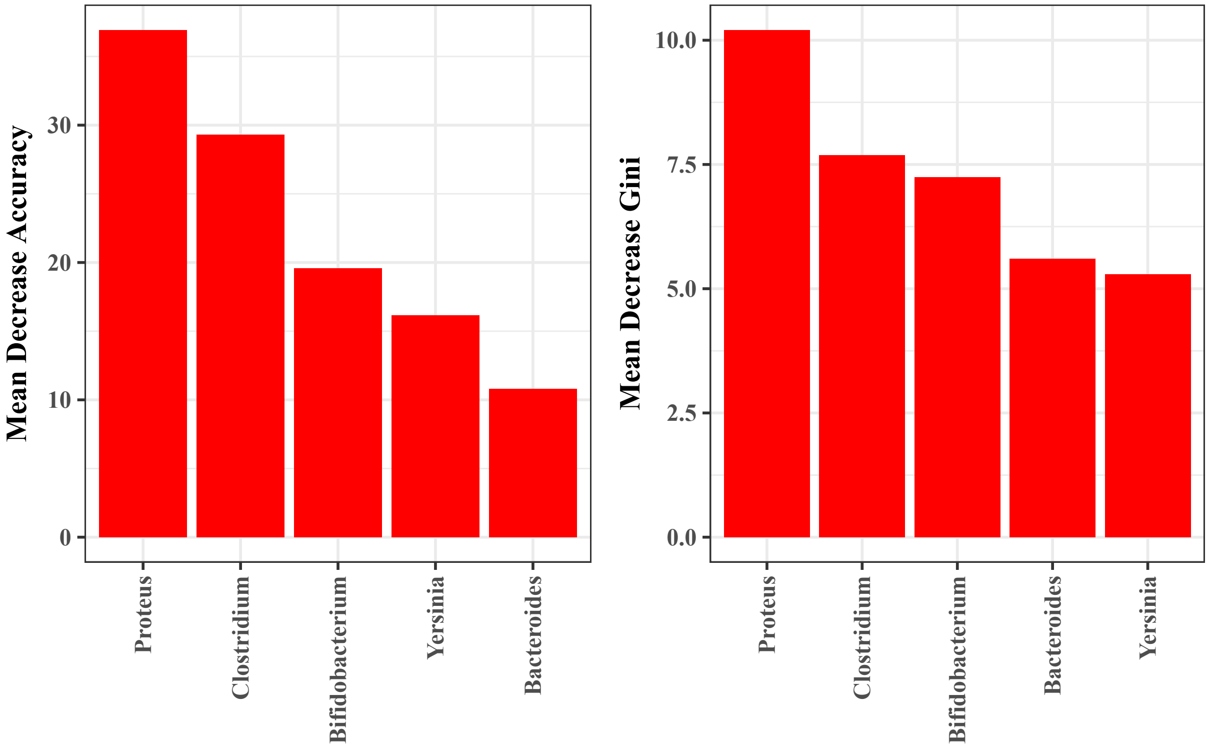


Figure 35 Random Forest Summary graphs describing the ability of genera that were seen to be significantly different between NEC and control samples in infants fed exclusively breast milk delivered and delivered vaginally to predict the NEC status of an infant. (A) Mean Decreased Accuracy for phyla with log-2 fold. (B) Mean Decrease Gini Index scores.

The tree representation of the model highlights the importance of *Bifidobacterium* in partitioning the data, with high abundances being associated with the prediction of control samples. Low abundances of *Bacteroides*, *Clostridium,* and *Proteus* were mostly associated with the prediction of control samples (Figure 36).

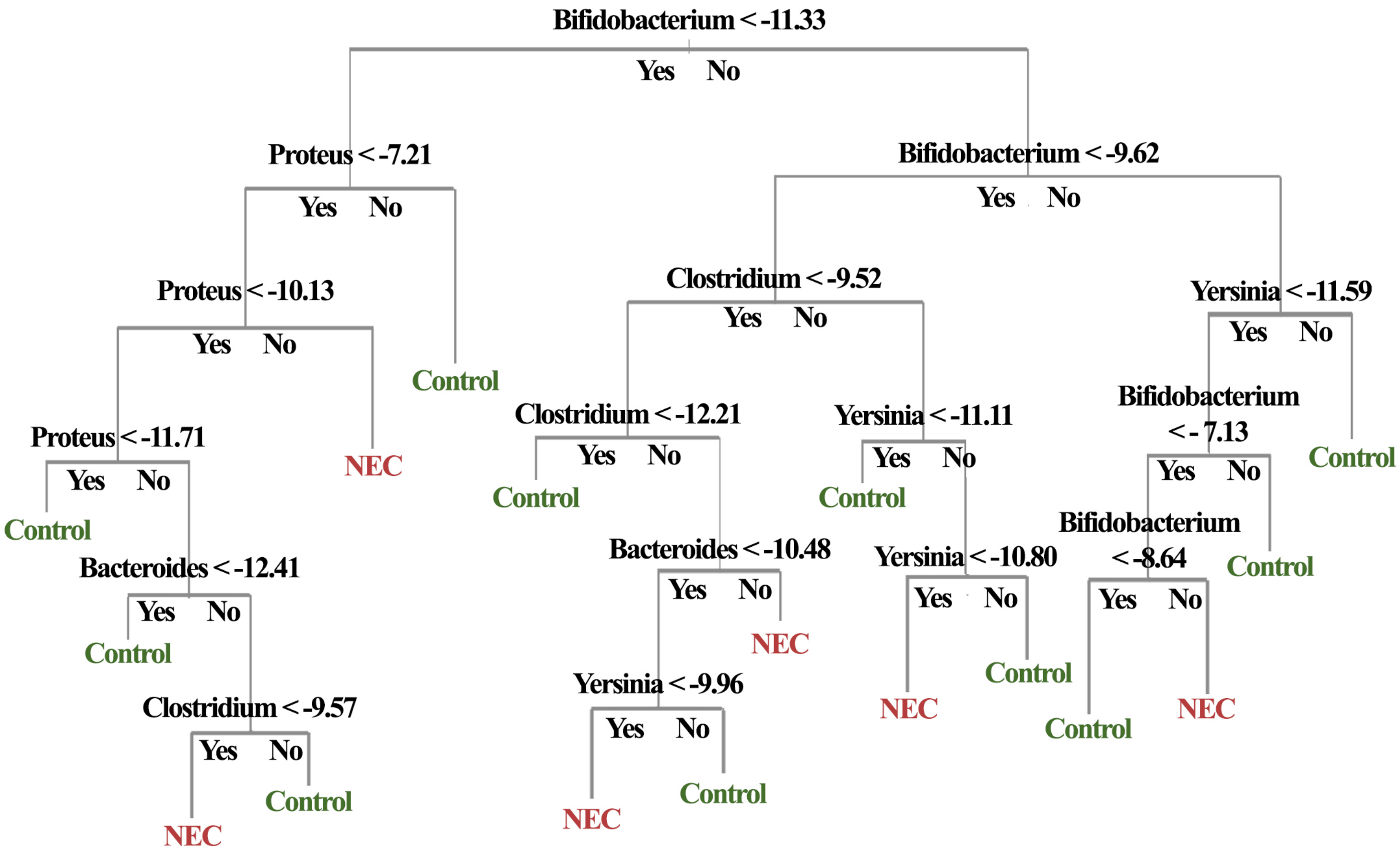


Figure 36 Tree representation of Random Forest decision based prediction of NEC status for samples using the log-normalised genera abundance of taxa observed to be significantly different in abundance between case and control samples. Values in the tree represent the log normalised value that a given decision is made upon. *Bifidobacterium* was seen to be highly influential in partitioning the data however it did not contribute as much to the accuracy of the model as Proteus or *Clostridium*. Generally, low abundance of *Proteus* and *Clostridium* were associated with control samples whilst high abundances of *Bacteroides* were seen to be associated with NEC samples

### 5.4.4 Diagnostic Species

Six species were seen to be significantly different in normalised abundances of NEC and control samples (Table 25). *Bacteroides fragilis* and *Corynebacterium kroppenstedtii* were the only two species to show different trends in abundance when factored with the age at sampling (Figure 37). *Bacteroides fragilis* was disproportionately represented by a BWH204 whose samples all had normalised abundance values of 0. This distorted the trends in the species over time and made it an unlikely candidate in discerning NEC and control samples.

*Corynebacterium kroppenstedtii* was not overrepresented by any single NEC or control sample over the duration of time and could be a viable candidate for a probiotic species of bacteria. However, where there is a rise in the abundance of *Corynebacterium kroppenstedtii,* there is also a lack of NEC samples, suggesting that further investigation into the influence of this species on the microbiota would be essential before any conclusions could be drawn.

Table 25 Summary of all species with log2 fold significant differences for normalised abundances of NEC and control subject samples. All samples were from infants delivered vaginally fed breast milk exclusively. The table is ordered by P-adjusted values.

| **Species** | **Base Mean** | **Log2 fold Change** | **P-Adjusted** |
|---|---|---|---|
| *Bacteroides fragilis* | 15473.701 | 8.63 | $6.95x10^{-28}$ |
| *Clostridium perfringens* | 375.68 | 5.86 | $1.44x10^{-20}$ |
| *Corynebacterium kroppenstedtii* | 4.59 | 2.28 | $7.80x10^{-6}$ |
| *Pantoea agglomerans* | 141.36 | 2.27 | $3.35x10^{-4}$ |
| *Acinetobacter rhizosphaerae* | 3.90 | 2.07 | $3.40x10^{-4}$ |
| *Streptococcus agalactiae* | 20.31 | -2.33 | $4.39x10^{-4}$ |

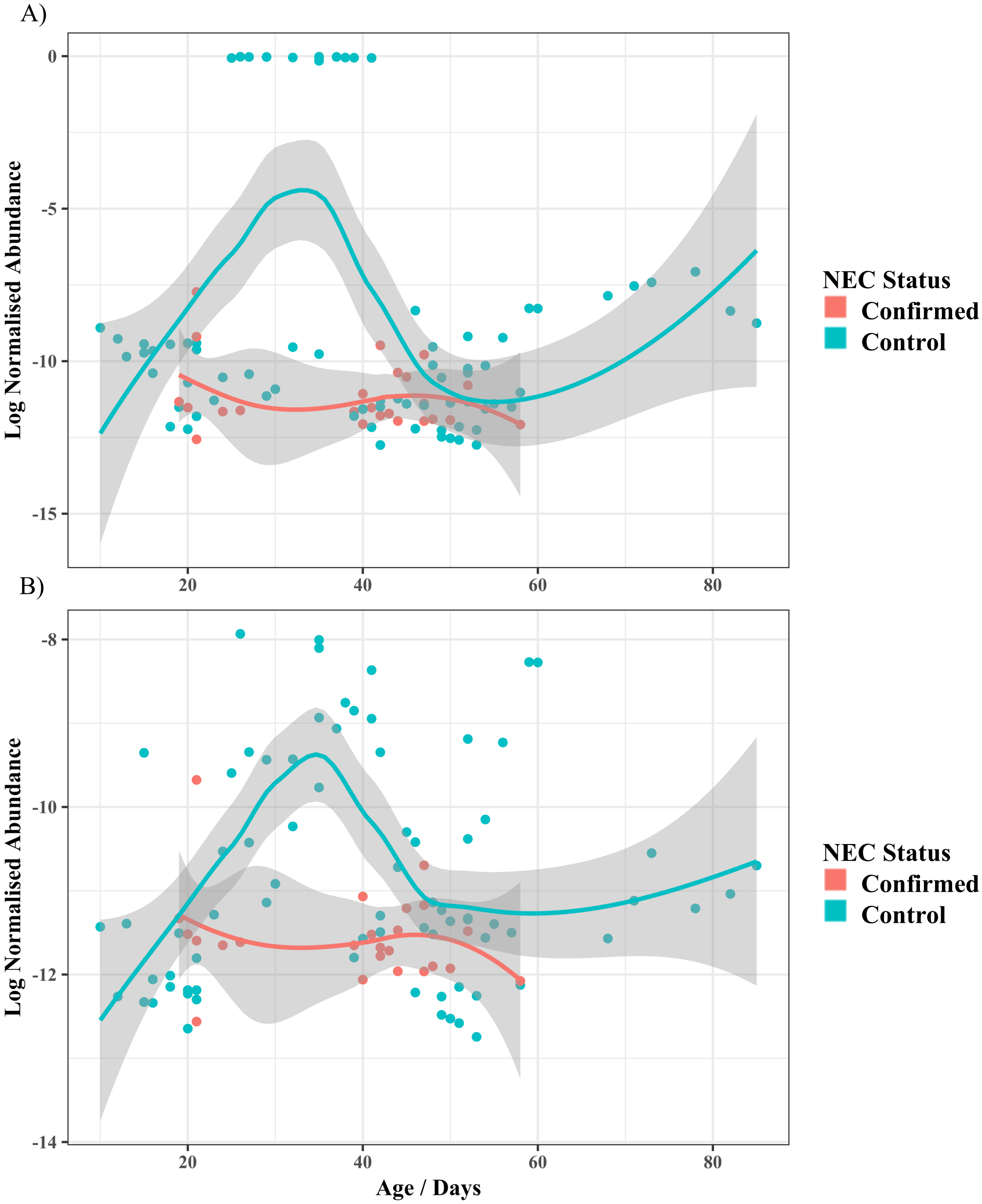


Figure 37 Nonparametric regression analysis for normalised log-abundances for species observed to be significantly different between NEC and control samples, plotted against the age at sampling and coloured by NEC status. (A) *Bacteroides fragilis*. (B) *Corynebacterium kroppenstedtii*. The elevated level of *Bacteroides fragilis* was influenced by control subject BWH204. BWH204 was the only subject to have samples with a value of 0 and significantly skewed the trend line and standard errors for control samples. No single subject was observed to represent upper or lower log-normalised abundance values for *Corynebacterium kroppenstedtii*.

Random Forest models using significantly different species had an increased error rate in the identification of NEC samples (50%) relative to the model using significantly different genera (42%) (Table 26:A). This further increased to 58% when using all observed species; the same error rate as the model using all genera observed in the samples (Table 26:B).

Table 26 Summary Tables for Random Forest model predictions of the NEC status of infants using the species composition of samples. 1,500 trees were constructed for both models. (A) Using only species identified as being significantly different between NEC and controls, two variables tried at each split. (B) Using all species identified within the communities, seven variables tried at each split.

(A)

| Confusion matrix | Confirmed | Control | Class Error |
|---|---|---|---|
| Confirmed | 12 | 12 | 0.5 |
| Control | 6 | 69 | 0.08 |
| OOB estimate of error rate | | 18.18% | |

(B)

| Confusion matrix | Confirmed | Control | Class Error |
|---|---|---|---|
| Confirmed | 10 | 14 | 0.58 |
| Control | 2 | 73 | 0.03 |
| OOB estimate of error rate | | 16.16% | |

## 5.5 All Viable Subgroups: All infants Fed both Formula and Breast Milk & Infants Exclusively Fed Breast Milk Delivered Vaginally

Following individual analysis of these subgroups, elevated occurrence of *Bifidobacterium* appeared to be consistent in control samples from all groups. With the exception of *Bacteroides*, which showed significant differences between NEC and control samples collectively but little difference in abundance when plotted over time, there was a lack of consistent trends across all NEC samples for these subgroups. Further analysis aimed to establish whether this was a correct assumption by performing analysis on all of these samples collectively and ascertaining whether the subgroup analysis was an improvement on the population as a whole in discerning NEC samples from controls.

### 5.5.1 Significantly Different Genera Between Necrotising Enterocolitis and Control Samples

Four taxa displayed significant differences in normalised abundances of samples from NEC and control infants; *Bifidobacterium*, *Bacteroides*, *Dialister*, and *Megamonas*. Two of these taxa were observed to have high mean abundances and a high level of log2 fold changes (Figure 38). These were associated with *Bifidobacterium*, which had the highest base mean abundance (4986) and level of significance (P-adjusted = $8.64 \times 10^{-145}$) relative to all taxa that showed significantly different abundances between case and control samples. *Bacteroides* was the only one of the four taxa to maintain a high mean abundance (1422) and level of significance ($3.94 \times 10^{-91}$) (Table 27).

Compared to the subgroup analyses previously described, *Bifidobacterium* and *Bacteroides* were seen to have the greatest significant differences between case and control samples of all genera across all subgroups. *Dialister* and *Megamonas* were also significantly different, these were the same genera, in the same order, within the subgroup of infants delivered vaginally and fed both formula and breast milk.

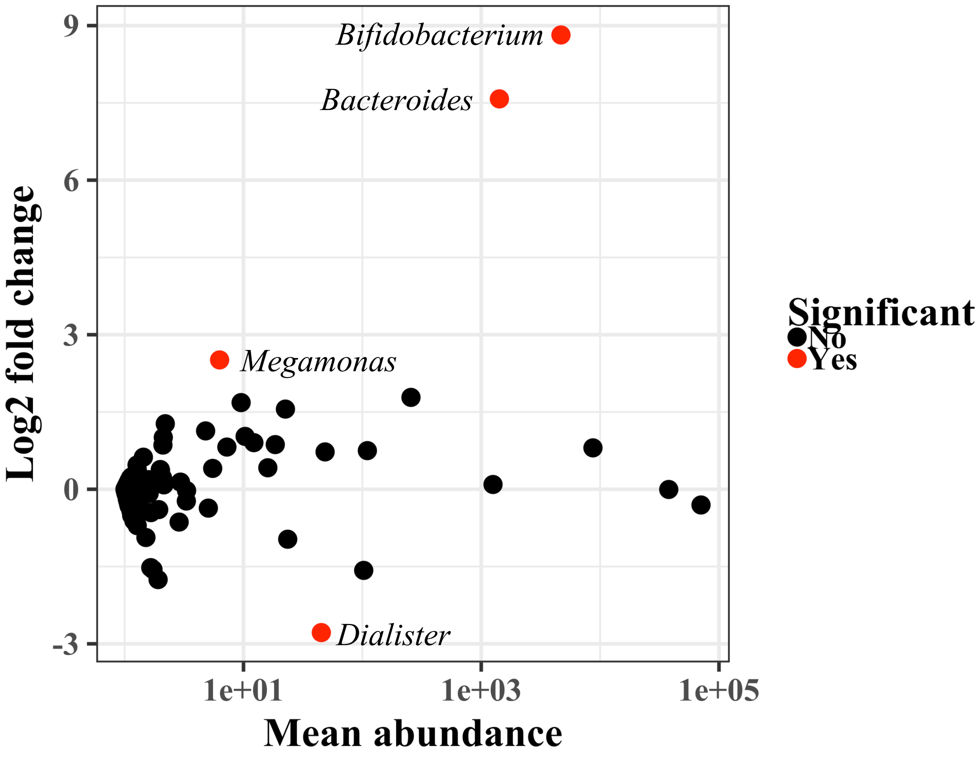


Figure 38 Genera with log2 fold differences between samples from all NEC and control subjects within the three subgroups previously analysed.

Table 27 Summary of all genera with log2 fold significant differences in normalised abundances between all NEC and control samples in the three subgroups previously analysed.

| Genera | Base Mean | Log2 fold Change | P-Adjusted |
|---|---|---|---|
| *Bifidobacterium* | 4685.91 | 8.82 | $8.64x10^{-145}$ |
| *Bacteroides* | 1421.93 | 7.58 | $3.94x10^{-91}$ |
| *Dialister* | 45.16 | -2.78 | $2.72x10^{-18}$ |
| *Megamonas* | 6.30 | 2.51 | $4.01x10^{-21}$ |

On the whole, *Bifidobacterium* and *Bacteroides* were elevated in control samples, although the difference was considerably more for *Bifidobacterium*. Abundances of *Dialister* were seen to be greater in NEC samples on average, whilst *Megamonas* showed little difference between case and control samples (Figure 39).

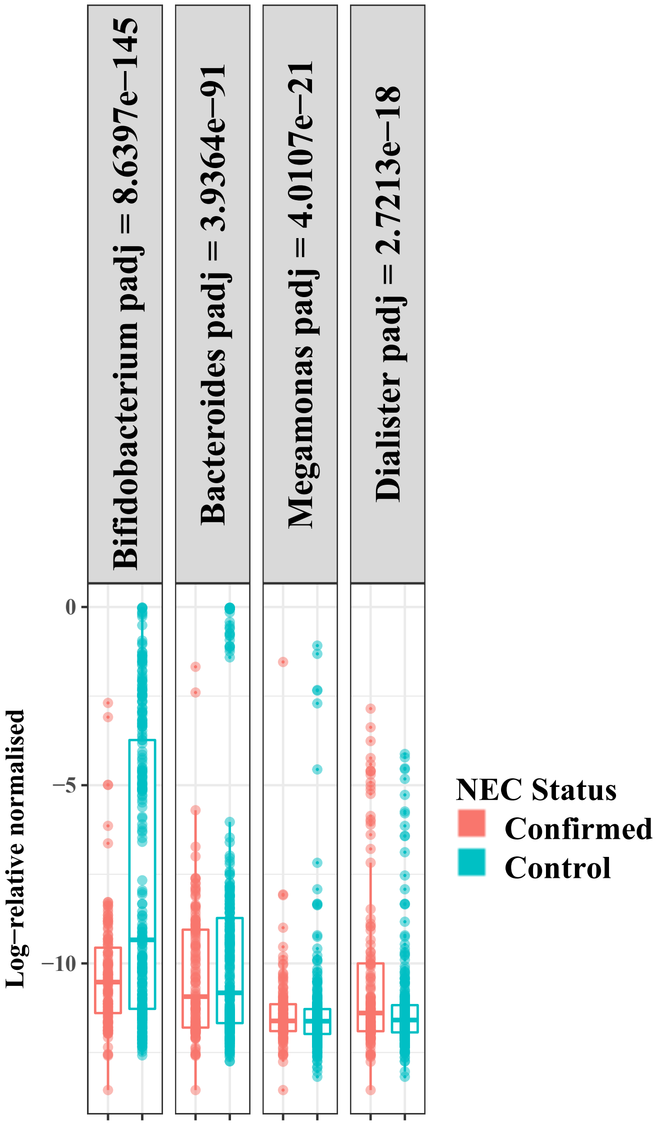


Figure 39 Genera with log2 fold differences between samples from all NEC and control subjects in the three subgroups previously analysed in this chapter.

### 5.5.2 Significantly Different Genera between Necrotising Enterocolitis and Control Samples over Time

Overall, *Bacteroides* was seen to maintain a higher abundance in control samples when observed as a function of time. NEC samples showed similar trends in the changes of *Bacteroides* abundance, however, these samples always maintained an average lower than that observed in control samples (Figure 40).

*Bifidobacterium* showed the greatest difference between case and controls. Controls maintained a greater mean abundance and displayed a tendency to increase their abundance of *Bifidobacteria* at older ages. NEC subjects showed the same trend in the normalised abundance of samples up to day 50. However, these samples had significantly lower mean abundances and after 50 days of age there was no evidence to support an increase in abundance.

*Dialister* was the only genus to show a clear association with NEC samples; the mean abundance increased up to approximately 30 days of age, after which it declined before returning to similar levels observed in control samples by day 50.

There was some evidence for an increase in the mean abundance of *Megamonas* up to ~40 days of age in NEC samples, whilst control samples were seen to decrease in abundance over the same time period. However, the normalised abundances were not clearly distinct as indicated by the high crossover in the standard errors of the two groups.

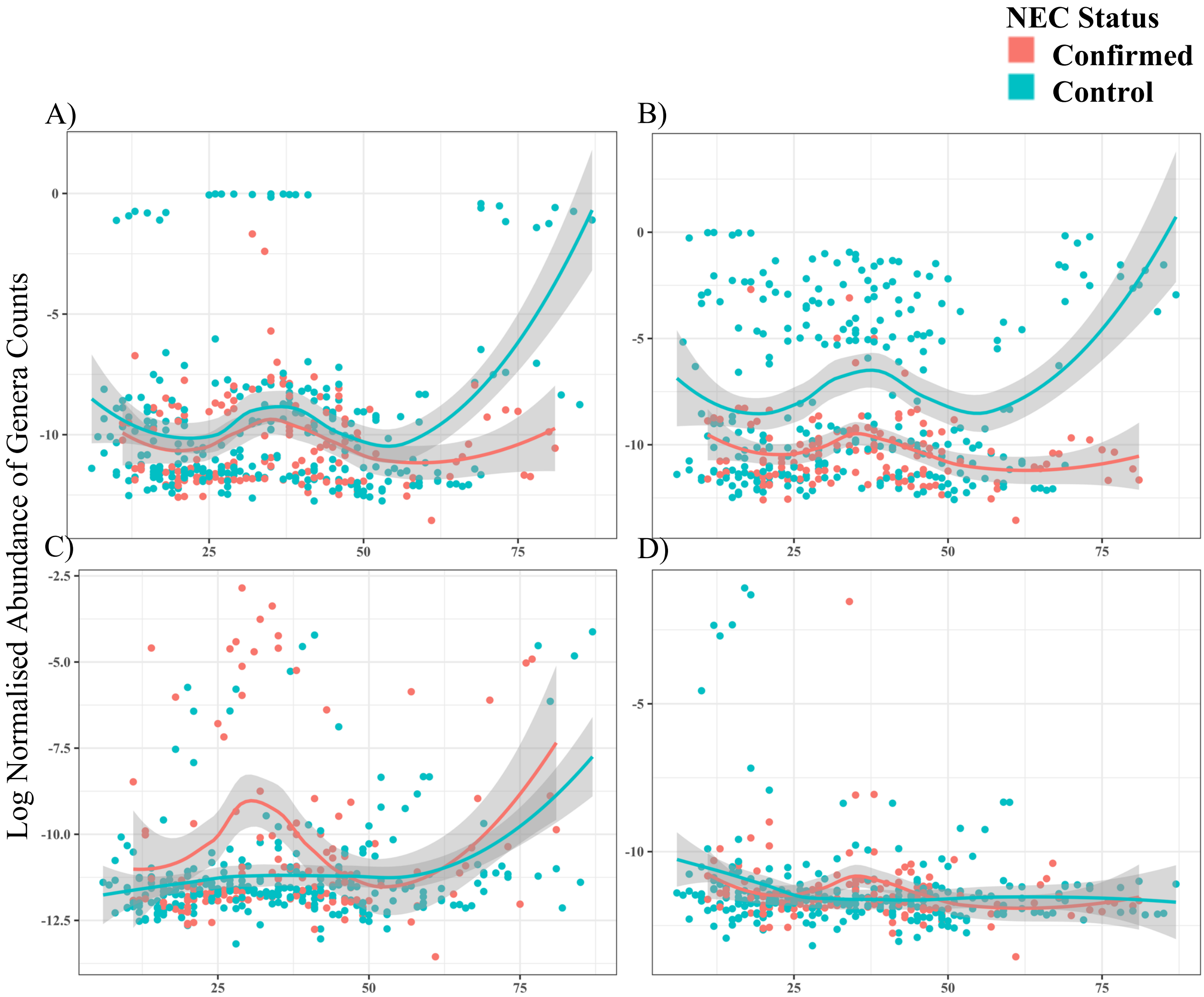


Figure 40 Nonparametric regression analysis for normalised log-abundances for genera observed to be significantly different between NEC and control samples, plotted against the age at sampling and coloured by NEC status. (A) *Bacteroides*. (B) *Bifidobacterium*. (C) *Dialister*. (D) *Megamonas*.

*Bifidobacterium* was the only genus consistently found in each subgroup and in the overall population with a significantly elevated mean abundance in control samples. This difference in normalised abundance was maintained over the duration of sampling for all subgroups. Controls delivered by caesarean section and fed formula and breast milk showed a consistent increase over time in the mean abundance of *Bifidobacterium* up to 50 days of age. All controls delivered vaginally saw an increase in the mean abundance of *Bifidobacterium* but also exhibited a decline at later ages; this occurred earlier in infants fed formula and breast milk relative to those fed breast milk exclusively.

*Bacteroides* displayed significantly different abundances for both subgroups in which infants were delivered vaginally, however, these subgroups showed distinct trends from each other. Samples from infants fed a mixture of formula and breast milk were only seen to diverge in mean abundance relative to NEC status after ~60 days of age, where the sampling density was much lower. The difference observed in samples from infants fed exclusively breast milk was also seen to be associated with lower sampling density and especially high abundances in one subject (BWH206). When all samples were analysed as a population, the trends in *Bacteroides* mean abundances were very similar between NEC and control infants. It is likely that the observed increased mean abundance for control samples was influenced by BWH206 and reduced NEC sample densities.

*Dialister* was only associated with infants delivered by caesarean section and fed both formula and breast milk. The observed peak corresponds to high abundance in samples from two infants with NEC and three controls.

Significant differences in the mean abundance of *Megamonas* were only seen in infants delivered vaginally and fed both formula and breast milk. When plotted over time there was no clear distinction between NEC and control samples aside from control samples having a greater initial abundance relative to NEC samples. In the population analysis however, there appeared to be a peak associated with NEC subjects at ~30-40 days of age.

### 5.5.3 Random Forest Analysis of Genera Predictive of Necrotising Enterocolitis

When using significantly different genera, the probability of the Random Forest algorithm to correctly identify samples from NEC subjects when all subgroups were assessed as a single cohort was decreased relative to all previously analyses. When all samples were analysed, the Random Forest had a 47% error rate in detecting NEC. In contrast, samples from infants delivered by caesarean section and fed both formula and breast milk had a 33% error rate; samples from infants delivered vaginally and fed both formula and breast milk had a 24% error rate; and infants delivered vaginally and fed breast milk exclusively had a 41% error rate.

When all genera observed in all the samples were used, the class error rate in correctly identifying NEC samples from all subgroups was 56%. This was lower when compared to samples from infants fed formula and breast milk and delivered both by caesarean section (42%) and vaginally (12%). However, this was an improvement on the error rate for samples from infants delivered vaginally and fed breast milk exclusively (58%).

Table 28 Summary Tables for Random Forest model predictions of the NEC status of infants using the genera composition of samples. 1,500 trees were constructed for both models. (A) Using only genera identified as being significantly different between NEC and controls, two variables tried at each split. (B) Using all genera identified within the communities, twelve variables tried at each split.

A)

| Confusion matrix | Confirmed | Control | Class Error |
|---|---|---|---|
| Confirmed | 71 | 63 | 0.47 |
| Control | 52 | 226 | 0.19 |
| OOB estimate of error rate | | 27.91% | |

B)

| Confusion matrix | Confirmed | Control | Class Error |
|---|---|---|---|
| Confirmed | 59 | 75 | 0.56 |
| Control | 16 | 262 | 0.06 |
| OOB estimate of error rate | | 22.09% | |

The importance of *Bifidobacterium* in the distinction between NEC and control samples was highlighted in the mean decreased accuracy (~80) and mean decreased Gini scores (>60) relative to the other genera observed to have significantly different abundances between NEC and control samples (Figure 41). This mean decreased accuracy accounted for approximately twice the accuracy in the model relative to the next most important genera, *Dialister*. *Dialister*, which also had the second highest mean decreased Gini score, was only observed to be significant for samples from caesarean delivered infants fed formula and breast milk. *Megamonas* and *Bacteroides* were seen to be similarly important in the model construction, both in terms of the mean decreased accuracy and Gini scores. Whilst a tree representation of this model was constructed, the complexity and size of the figure meant that it provided no useful information in the prediction of NEC nor was there a feasible method of incorporating it on paper.

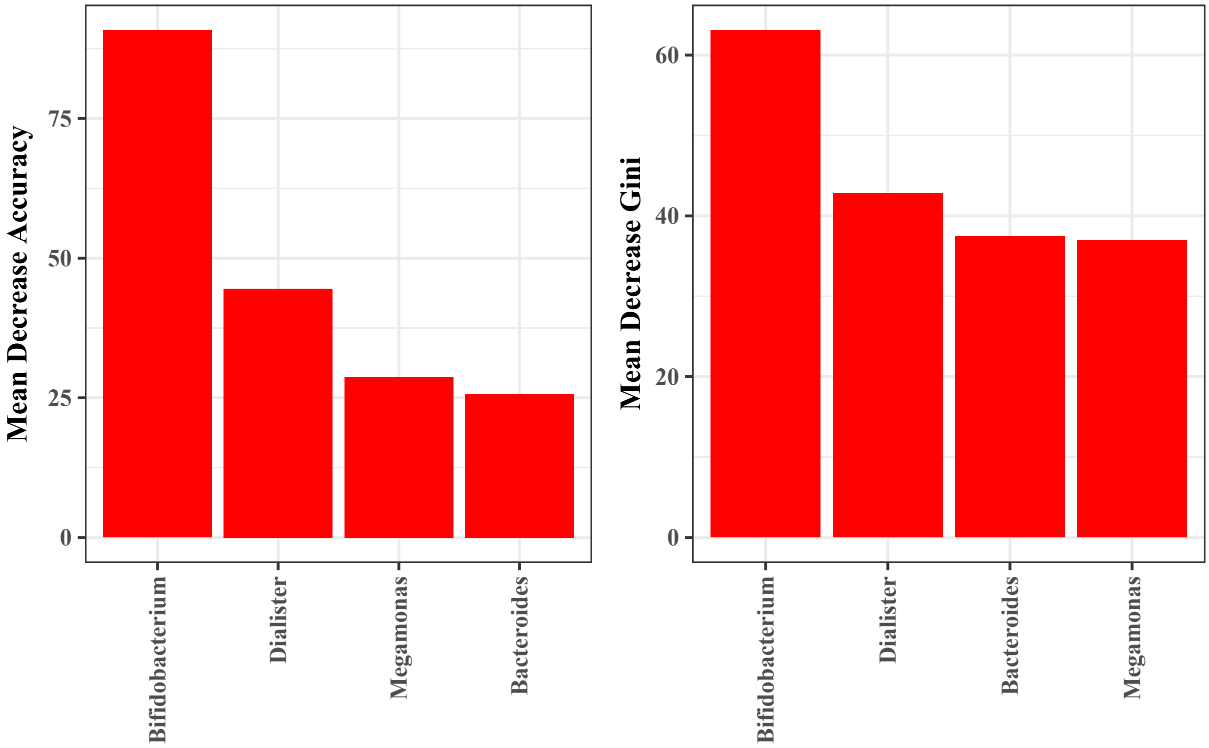


Figure 41 Random Forest Summary graphs describing the ability of genera that were seen to be significantly different between all NEC and control samples from subgroups previously analysed in this chapter in the prediction of an infant's NEC status (A) Mean Decreased Accuracy for phyla with log-2 fold. (B) Mean Decrease Gini Index scores.

### 5.5.4 Diagnostic Species

Four species were observed to have significantly different normalised abundances when the age at sampling was not factored into the analysis (Table 29). Only two of these species demonstrated different trends in abundance over time; *S. agalactiae* and *C. butyricum*. However, these differences were localised to a small number of subjects and were not consistent for the majority of samples (Figure 42). Samples with log normalised abundance between -7.5 and 0 for *S. agalactiae* were seen to be represented by three subjects; two controls BWH204 and STH051, and the NEC subject RSH114. The elevated abundances for *C. butyricum* were associated with NEC subject UHCW062 only.

Table 29 Summary of all species with log2 fold significant differences for normalised abundances of NEC and control subject samples. Samples were from all infants subgroups analysed within this chapter.

| Species | Base Mean | Log2 fold Change | P-Adjusted |
|---|---|---|---|
| *Bacteroides fragilis* | 1416.29 | 8.074 | $5.67\text{x}10^{-98}$ |
| *Streptococcus agalactiae* | 11.09 | -4.57 | $2.20\text{x}10^{-88}$ |
| *Clostridium butyricum* | 5.97 | -3.35 | $4.29\text{x}10^{-52}$ |
| *Streptococcus anginosus* | 32.56 | 4.17 | $2.72\text{x}10^{-40}$ |

Both Random Forest models showed an increase in the error rate for classifying NEC samples compared to the models utilising genera. An increase of 6% was observed for the model using significantly different species to identify NEC samples, and an increase of 11% was seen for the model using all species (relative to the models using genera) (Table 30). This indicated that models using species were less effective in identifying samples from NEC subjects.

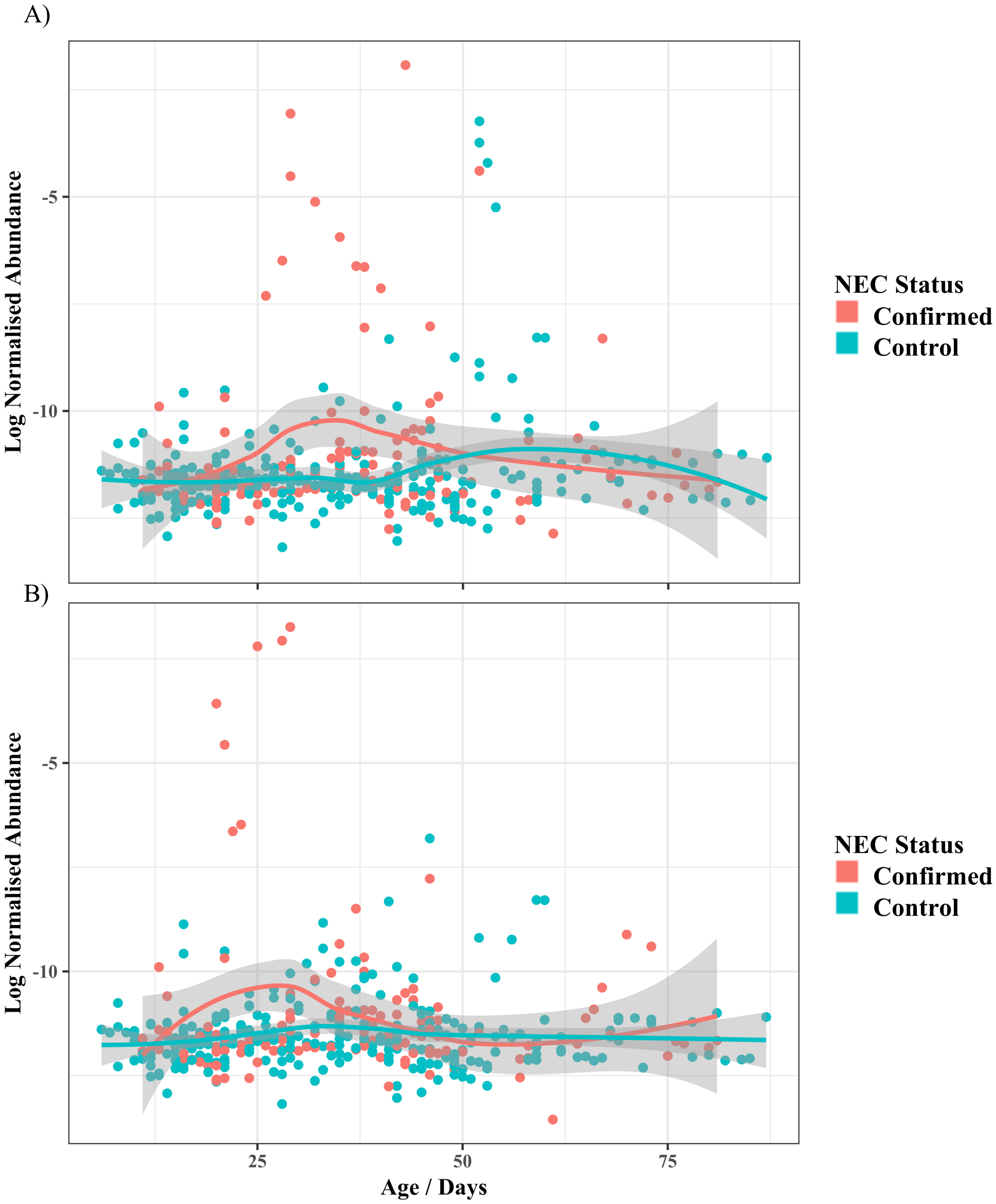


Figure 42 Nonparametric regression analysis for normalised log-abundances for species observed to be significantly different between NEC and control samples, plotted against the age at sampling and coloured by NEC status. (A) *Streptococcus agalactiae*. (B) *Clostridium butyricum*.

Table 30 Summary Tables for Random Forest model predictions of the NEC status of infants using the species composition of samples. 1,500 trees were constructed for both models. (A) Using only species identified as being significantly different between NEC and controls, two variables tried at each split. (B) Using all species identified within the communities, eleven variables tried at each split.

(A)

| **Confusion matrix** | **Confirmed** | **Control** | **Class Error** |
|---|---|---|---|
| Confirmed | 63 | 71 | 0.53 |
| Control | 50 | 228 | 0.18 |
| OOB estimate of error rate | | 29.37% | |

(B)

| **Confusion matrix** | **Confirmed** | **Control** | **Class Error** |
|---|---|---|---|
| Confirmed | 45 | 89 | 0.66 |
| Control | 25 | 253 | 0.09 |
| OOB estimate of error rate | | 27.67% | |

## 5.6 Discussion

As stated in the introduction, Chapter 1.10.1, throughout the first year of life the infant gut is colonised by blooms of bacteria which are predominantly anaerobic. Initial colonisation is thought to be mainly by facultative anaerobes such as *Staphylococcus*, *Streptococcus*, *Escherichia*, and *Enterobacteria*. Whilst the maternal vaginal microbiome is predominantly *Lactobacillus*, these do not actively colonise the infant gut, instead it is the maternal faecal bacteria (*Enterobacteriaceae* and *Bifidobacteria*) that have been associated with the colonisation of infants delivered vaginally[156]. In contrast, caesarean infants were observed to harbour bacterial communities similar to the skin microbiota of the mother[309]. This influence was reinforced by the subset regression and NMDS analysis in Chapter 4.

In addition to the method of delivery, the feeding regime was also identified as a significant factor in the trends of beta-diversity. This  variable had previously been reported to influence the microbiota composition and development, especially in the first six months of life[314,315]. Therefore, it was considered pertinent to ensure that subgroups based on feeding regime and means of birth were established and compared independently to isolate differences in how the microbiota changed over time in NEC and control infants.

Chapter 4 also indicated that the best taxonomic level to differentiate between NEC and control samples was at the genus level. This was used as the primary comparison level, however, to ensure that there were not any common pathogenic species, further analysis was performed at the species level in the same manner.

### 5.6.1 Infants Delivered by Caesarean Section and Fed Formula Feeds & Breast Milk

Analysis of infants delivered by caesarean section and fed a mixture of formula and breast milk identified 10 unique genera that were significantly different between samples from NEC and control infants. Three genera were seen to have elevated normalised mean abundances within the samples: *Bifidobacterium, Streptococcus,* and *Proteus.* This suggested that these genera constituted a significant proportion of the community composition in addition to being significantly different between NEC and control samples. *Bifidobacterium* had the greatest abundance and this was significantly elevated in control samples. *Dialister* was the genus observed to be most significantly elevated in NEC samples, but it had a lower mean sample abundance relative to the other genera that were significantly different.

When plotted as a function of time, the only genera that was seen to be consistently different between NEC and control samples was *Bifidobacterium*. This genus was seen to have consistently greater mean abundances in control samples over the duration of sampling. However, there were control subjects with lower normalised sample abundances relative to NEC subjects sampled at the same age. These were predominantly samples taken prior to 30 days of age. There was some evidence of genera that were seen to peak in NEC subject samples, however these were exclusively associated with individuals and not conserved in the population.

Random Forest model prediction of NEC samples from the population had a reasonable accuracy when using the genera observed to be in significantly different abundances between NEC and control subjects (33%). This accuracy was reduced when all genera were incorporated into the model predictions (42%). This decrease in accuracy could be indicative of the complexity and specificity of each individual infant's microbiome; by including all genera the uniqueness of each microbiome decreases the accuracy of successful NEC identification.

The importance of *Bifidobacterium* was clearly established in the mean decreased accuracy and Gini index scores, as well as the tree representation of the model. Additionally, *Dialister* and *Staphylococcus* were identified as being important in model predictions, with elevated normalised abundances of *Dialister* and *Staphylococcus* being associated with NEC samples. However, the tree structure in the model predictions demonstrated how complex and specific each prediction pathway was, which had an overall error rate of 33%.

Analysis at the species level did not identify any species associated with genera that were significantly different within this subgroup of infants. Of the three species identified (*Clostridium butyricum, Streptococcus anginosus,* and *Eubacterium dolichum*), only *C. butyricum* was observed to show differences over time, and when further analysis was performed on this species (which appeared to be elevated in NEC samples) it was clear that a single subject (UHCW062) had a bloom over the duration of sampling and that this was not representative of the NEC sample population. This also highlighted the importance of high sample counts with respect to understanding the population trends associated with microbiota changes at the community level. Our sampling for NEC subjects could be considered limited even at this level due to the skew in normalised abundances introduced by a single infant.

The reduced accuracy of the Random Forest model at the species level also lends itself to the theory that the infants' highly specific composition is unique and at higher resolutions introduces more noise than clarity. This theory was somewhat supported by the CCA analysis and high attribution of variation associated with trends in beta-diversity (LCBD values) to the subject ID (Chapter 4.3).

*Bifidobacterium* has been shown to naturally occur in many niches that are associated with the animal gastrointestinal tract. This genus has metabolic abilities and genetic traits that aid in the evasion of the host adaptive immune system and colonisation through specific appendages[456]. These metabolic abilities include the production of acetate and lactate which contributes to the lower of the gastrointestinal tract pH via short-chain fatty acid production. This improves the availability of caesium, magnesium and inhibition for potentially pathogenic bacteria[457]. *Bifidobacterium* dominate the gut population in healthy breast-fed infants[290,292,458]. On average, more than 12% of the annotated open reading frames within the *Bifidobacterial* genome are predicted to encode carbohydrate metabolic enzymes[459]. Additionally, Milani *et al* demonstrated that *Bifidobacterium* species have a wide range of carbohydrate genes, greater than that observed in other known members of the gut microbiota – specifically, glycan-degrading abilities of *Bifidobacteria* which are believed to reflect the available carbon sources in the human gut[459]. Transcript profiling of the genomes also provided evidence for the involvement of various chromosomal loci in glycan metabolism. This supports the hypothesis that saccharidic resource sharing among *Bifidobacteria* – through species-specific metabolic specialising and cross-feeding – results in trophic relationships between members of the gut microbiota.

Human isolates of *Dialister* species are reported to produce propionate[460,461] and their abundances have been implicated in disease states such as obesity, IBD, and Crohn's disease. Within a normal microbiota, *Dialister* has been categorised as a member of the low gene count microbiomes due to its relatively low community abundance[462,285]. This concept of low gene count members has been implicated in the health and disease states of individuals with obesity, wherein low gene count members harboured a higher proportion of pro-inflammatory bacteria such as *Bacteroides* and *Ruminococcus gnavus*, both of which are known to be associated with IBD[463].

The genus *Dialister* has very little characterisation in literature with regards to its functional metabolic pathways within the gut microbiome community. This could be due to its low abundance within the community, although there does appear to be evidence that it is elevated in communities that seem to be associated with disease states such as Crohn's Disease[270].

During vaginal delivery, *Staphylococcus* (among other facultative anaerobic species) colonise the infant gut in the first few days of life, this allows strict anaerobes such as *Bacteroides* and *Bifidobacterium* to establish a presence[464]. Additionally, breast milk is dominated by *Staphylococcus* and genera such as *Bifidobacterium* and *Lactobacillus*[465]. Using culture and strain level discrimination, *Lactobacillus* has also been shown to transfer to the gut after feeding on breast milk[466,467]. Whilst *Staphylococcus* has been shown to be present in the gut colonisation of vaginally delivered infants, they are dominant in infants delivered by caesarean section[309], with the latter's communities showing a closer resemblance to the human skin microbiota.

*Staphylococcus* was initially elevated in control infant samples relative to NEC samples but returned to similar levels to those observed in NEC infants by the age of 30 days. Whilst this is a limited difference, it could suggest that NEC infants delivered by caesarean section showed an abnormal colonisation pattern. Very few control samples were seen to have abundances of *Streptococcus* lower than those observed in NEC samples in the early sampling duration.

However, it is important to consider that the difference in *Bifidobacteria* and *Staphylococcus* abundances could be associated with differences in feeding regimes. NEC infants are often restricted in their feeding intake and struggle to take on feeds. Evidence has clearly been shown that associated these genera with breast milk, therefore the reduced abundances could in fact not be associated with a gut dysbiosis but rather a reduced feeding regime relative to controls. This would be more an effect of the treatment, which would occur after the detection of NEC, than a cause of the disease.

### 5.6.2 Infants Delivered Vaginally Fed Formula and Breast Milk

*Bifidobacterium* was again observed to be a genus that is significantly different between NEC and control samples within the vaginally delivered infants on a mixed feeding regime. However, *Bacteroides, Megamonas,* and *Veillonella* were the other genera observed to be significantly different, none of which were seen to be significantly different in caesarean delivered infants on a mixed feeding regime.

*Bifidobacterium* and *Bacteroides* were seen to have the greatest mean abundances and log2 fold changes; while *Veillonella* had a high normalised mean abundance, it also had the lowest difference between NEC and control samples. *Megamonas* had the lowest abundance of all the significantly different genera and a marginally greater difference between NEC and controls relative to *Veillonella,* suggesting it was a low abundance member of the community.

*Bifidobacterium* and *Veillonella* had greater mean log-relative abundances in control samples overall, while *Bacteroides* had a greater mean log-relative abundance in NEC samples. *Megamonas* was seen to have very similar normalised abundances between case and control samples.

As a function of time, *Bacteroides* were seen to have similar normalised abundances until 50 days of age, after which control samples showed an increase in abundance which was not observed in NEC subjects. Of the 133 samples within this group, 33 were taken after 50 days of age (NEC = 14, Controls = 19). *Bifidobacterium* was shown to have consistently greater normalised abundances in control samples over the duration of sampling. Normalised abundances of *Megamonas* showed no clear differences in trends for the duration of sampling. *Veillonella* maintained a higher abundance in control samples, and a clear steady increase over time. In contrast, NEC subjects showed a sharp increase in abundance of *Veillonella* at early ages which dropped off after ~30 days of age. These results suggested that when time was factored, the only clear differences observed were between *Bifidobacterium* and *Veillonella.*

This was further confirmed by the Random Forest analysis which showed that *Bifidobacterium* and *Veillonella* had the greatest mean decreased accuracy and Gini index scores as part of the models that utilised the genera that were significantly different. The error rate for models using the significantly different genera were seen to be much higher (24%) than when the model utilised all the genera observed within the samples (12%) in the identification of NEC samples from the population. However, both of these values were lower than the best error rate observed for infants delivered by caesarean section.

This suggests that the detailed composition of the microbiome from infants on mixed feeding regimes who were delivered vaginally provided a more accurate means of identifying the NEC status of a given sample. Therefore, the influence of the genera that were seen to be significantly different, whilst being better predictors than those observed in caesarean delivered infants, were not as good at predicting the NEC status of samples compared to models that incorporated the entire community composition at the genus level. However, importantly, high abundances of *Bifidobacterium* were clearly assigned as controls within the tree structure of the Random Forest model, suggested that *Bifidobacterium* was highly associated with a reduced risk of NEC.

*Veillonella* have been primarily observed as a core member of the oral microbiome[468] but they have also been characterised throughout the digestive tract in differing abundances[469,470]. Whilst there is little to no intake of non-digestible carbohydrates in the premature infant diet – which is predominantly associated with proteins, fats and sugars in the form of human milk oligosaccharides[471]– the *Veillonella* species are known to metabolise propionate from non-digestible carbohydrates. It could be that this genus was introduced into the infant microbiome from the maternal faecal microbiome.

Propionate has been shown to have hypophagic and hypocholesterolemic properties in many studies, with potentially health-promoting implications such as anti-lipogenic, cholesterol reducing, anti-inflammatory and anti-carcinogenic effects[472,473]. The supplementation of propionate in white bread has been shown to improve blood-glucose responsiveness, increased faecal mass, and microflora, specifically *Bifidobacteria*. There appeared to be evidence that propionate enhanced satiety, however the implications of this for infants are unknown[473]. The elevated abundance of *Veillonella* could increase levels of propionate in the microbiome and by association the abundance of *Bifidobacteria*, which have been shown to be negatively correlated with the incidence of NEC within this cohort.

In addition, *Veillonella parvula* has been shown to gain additional energy from succinate in the presence of lactate as the main growth substrate[474]. This may explain the elevated abundance in infants more than any other factor, as a high proportion of breast milk is composed of lactate.

Of the two species that showed increasing abundance up to ~45 days old (*Bacteroides fragilis* and *Corynebacterium kroppenstedtii*), significant differences were only observed where NEC subjects were poorly sampled in terms of age. *Bacteroides fragilis* was disproportionately represented by a single control subject with a very high abundance relative to all other infants within the cohort. Where sampling was sufficient for both controls and NEC infants, between 40 and 50 days there was little difference between samples for both species. This indicated that there was no clear beneficial or pathogenic species associated with control or NEC infants within this subset of the cohort. The reduction in the accuracy of NEC sample identification in the Random Forest models generated at the species level relative to those generated at the genus level further confirms this.

### 5.6.3 Infants Delivered Vaginally and Fed Breast Milk Exclusively

Infants delivered vaginally and fed breast milk exclusively were the smallest viable subgroup of the three derived from the cohort and as such there were limits in the comparisons that could be made with respect to sampling age. Specifically, prior to 40 days of age there were very few NEC samples to compare against control samples. This limited comparisons within this subset to an age range between 40 and 60 days. This highlights the importance of plotting the data over time. A genus that may be seen to be significantly different with respect to the population of samples, based on normalised abundance, might not be comparable if there was limited sampling or if within the time frame where comparisons could be made there was no observable difference.

For infants delivered vaginally and fed breast milk exclusively *Proteus* was the genus with the greatest difference between NEC and control samples, this translated with time as infants with NEC had greater normalised abundances on average between 40 and 60 days of age. Although this was not the case for all NEC samples there was a clear difference as indicated by the error margins, which showed no overlap between the case and control samples after 50 days of age.

*Clostridium* was also seen to be significantly different between controls and NEC subjects, with a decreased normalised abundance in NEC subjects. As with Proteus this was not seen to be exclusively lower in NEC subjects, with some control samples having comparable abundances, but on average there was a much lower abundance with very little overlapping of error margins.

All other genera (*Bacteroides*, *Bifidobacterium* and *Yersinia*) had overlapping error margins, and similar trends in mean abundances between 40 and 60 days of age. Of note was *Bifidobacterium*, which was observed to be consistently different between the two previous subsets of infants and appeared to contribute the greatest role within their respective Random Forest models.

In contrast, and as expected based on the regression plots and log-2 fold significance results, *Proteus* and *Clostridium* were the most significant taxa within this subset to contribute to the Random Forest models. However, the error rates for both models were seen to be greater than those observed within the previous two subgroups.

None of the species was identified as being significantly different between case and control subjects when factored with time. The error rates of both Random Forest models constructed were greater than their respective models generated at the genus level; 50% and 58%, respectively.

The genus *Proteus* includes proteolytic rod shaped Gram-negative facultative anaerobic heterotrophs that can be opportunistic human pathogens[475]. *Proteus* is a highly adaptable genus that has been isolated from multiple human and environmental environments. Within the human environment they have been characterised as both commensals and pathogens. Their prevalence in human infections has been extensively characterised and attributable to the O-antigen variability and virulence factors. They have been associated with infections of the urinary tract, wound and burns, respiratory tract, bacteraemia, meningitis, intestine (diarrhoea) and nosocomial[476,477,478,479,480,481,482,483]. These infections are primarily associated with people with impaired immune systems. However, within our study *Proteus* was associated with control subjects that were considered otherwise healthy.

It is believed that *Proteus* is part of the naturally occurring faecal microflora in a proportion of the population even though it has been considered an opportunistic pathogen. It has been observed that there are differences in the *Proteeae* members within health and diarrhoeal patients and it was speculated that *Proteus* becomes pathogenic when the onset of illness is attributable to another genera[484],[485]. It has been suggested that whilst some strains are associated with urinary tract infections these are isolated from an intestinal reservoir[486], where they were more frequently isolated [487]. Of interest was the evidence of high fat diets within animal models and increased numbers of *P. mirabilis* in a study into the association of the gut microbiome of rats and obesity. A significant positive correlation was found between the abundance of *P. mirabilis* and all ten of the metabolic parameters associated with obesity[488].

The *Clostridium* genus is composed of gram-positive, rod-shaped bacteria from the Firmicutes phylum. Those that inhabit the gut microbiome are predominantly represented by the fusiform-shaped bacteria from cluster XIVa and IV (10-40%) of the total bacteria within the gut microbiome[489,490,491]. *Clostridia* have bene shown to colonise the human intestine of infants fed breast milk within the first month of life[492]. This was clearly observed within the control samples which showed an increase up to 40 days of age, after which the abundance of *Clostridium* stabilised. There was a reduced abundance at later ages (>60 days) but these were represented by a single subject and as such this limited the conclusions that could be made at these ages. Within this study there as little evidence for increased abundances in *Clostridium* with NEC samples.

Mouse models *Clostridium* have been shown to colonise specific regions of the intestinal mucosa. 20% of the sequences from the interfold region have been classified as members of the *Clostridium* cluster XIVa group. In contrast only 3% of were represented by the genus in the digestive region[493,494]. Regions of the central lumen consisted primarily of *Bacteroidaceae*, *Enterococcaceae* and *Lactobacillaceae*[495]. It is possible that the commensal species of *Clostridia* are associated with specific regions within the intestinal mucosa in order to maintain a close spatial arrangement with host gut cells required to perform physiological co-operative functions in an optimal manner[496].

The location of the *Clostridium* species in the mucosa implies an impact on normal intestinal structure and physiology and an association with functions relating to mucus production, water retention, epithelial cell cycles and peristalsis. All of these functions have been shown to be abnormal within germ-free mouse models[497]. Importantly, *Clostridia* play a role in the maintenance of colonocytes by releasing butyrate as an end-product of fermentation[498]. In another study, samples seen to have a high abundance of *Bifidobacterium* had a low abundance of *Clostridium*[312]. A similar pattern was observed within this cohort. Infants delivered vaginally and fed exclusively breast milk were seen to have a lower abundance of *Bifidobacterium* (base mean =265) relative to the other two groups (2949, 8678), and was also the only group to show a significant difference in the abundance of *Clostridia* in NEC and control samples. *Clostridium* was also identified as a significant genus in the Random Forest Analysis.

This short chain fatty acid, along with acetate and propionate, appears to be a primary energy source for colonocytes[499,500]. Additionally, these have been shown to have an important role in the health of the colon[501,502]. Butyrate is not detectable in the portal blood vessels although the colonic mucosa absorbs 95% of it, this provides an indication of the speed at which butyrate is utilised. Butyrate has also been implicated in gene expression through hyperacetylation of chromatin by its action as a non-competitive inhibitor of histone deacetylases[503]. Butyrate appears to have an impact of pro-inflammatory cytokines by inhibiting the activation of the transcription factor NF-kB[504,505]. The concentration of butyrate has been implicated in the cell growth and differentiation and in the prevention or reduction of conditions such as ulcerative colitis[501,502,503,506,507]. Conditions which lead to a reduction in the energy supply of colonocytes (70% of which is provided by butyrate) can lead to colitis, with additional implications in colorectal cancer and IBD[503,506,507].

Butyrate production is common across many *Clostridium* species, especially those of the XIVa and IV phylum clusters related to *Roseburia* and *F. prausnitzii* which express Butyryl CoA: acetate CoA transferase activity[508]. The majority of this metabolic information is sourced from industrial experimentation of solventogenic *Clostridia* and the validity of this data within the gut microbiome environment should not be assumed to be correct. Additionally much of the information previously described about Clostridia within the microbiome environment within this discussion is based on animal models. Whilst there is clear evidence of Clostridia's location within the human gut microbiome the complexity of the community may result in differing actions within different hosts, however the wealth of evidence suggests that Clostridia does have a beneficial, commensal impact on the host that appears to limit the onset of colitis among other conditions. The importance of Clostridia is in accordance with the data of La Rosa *et al* that demonstrated the progression of the microbiome in term infant gut microbiomes from Bacilli to Gammaproteobacteria to Clostridia, irrespective of the initial colonising conditions[108].

Clostridia have also been implicated in priming the host immune system through the promotion of αβ T-cell receptor intraepithelial lymphocyte (IEL) and immunoglobulin A producing cells[509]. Evidence has suggested that IEL, IG-A producing cells and intestinal epithelial cells are vital to the nature of the immune response to antigens or pathogens ingested. Reduced numbers observed within Germ-free mouse models are associated with a low Thy-1 expression and a low cytolytic activity of IEL[510,511]. This was built up by transplantation of 46 strains of Clostridia that were cultured from wild-type to Germ free mice which lead to an increase in the ratio of CD4$^-$ CD8$^+$ cells to CD4$^+$ CD8$^-$ in αβIEL within the large intestine[509]. Additionally, species of *Clostridium* associated with clusters XIV and IV have been associated with strong induction of colonic T regulatory  cell accumulation[512], placing further importance on their role with the host immune system. It has been suggested that the high proportion of  T regulatory cells within the intestine, wherein they are present in greater numbers than any other location, may be responsible for the development of the microbiome as Foxp3+ T regulatory cells have been observed to markedly influence the community[513]. This would suggest that infants with lower abundances of Clostridia are likely to have impaired immune responses to bacterial colonisation events or upon exposure to potential pathogens. Inappropriate immune responses can lead to excessive inflammation which is a known symptom of NEC.

There is evidence that Clostridia are influential in the development and maintenance of both the response to bacteria colonising the intestine and the colonocytes utilising metabolites produced by commensals within the gut microbiome. The limited number of samples within this subset from our study inhibits our ability to form conclusions. Specifically, consistent sampling across a larger age range would help to clarify whether the microbiome of NEC subjects fails to cultivate sufficient numbers of Clostridia to establish a balanced community and maintain metabolic pathways vital for the host intestinal cells.

### 5.6.4 All Subgroups Analysed as a Population

When all samples from each subgroup of the cohort were analysed a whole *Bacteroides*, *Bifidobacterium*, *Dialister* and *Megamonas* were observed to be significantly different between NEC and control samples. When factored with time *Bifidobacterium* was consistently observed to be maintained at significantly elevated normalised abundances in control samples relative to NEC samples. This genus was also the most distinctly separated between case and control samples with no overlapped standard errors over the duration of sampling. Although many control samples were observed to have lower normalised abundances relative to NEC samples, as a population this genus was the most differentiated between case and control samples.

Whilst *Bacteroides* were considered one of the most significantly different genera based on the normalised mean abundance and without factoring for time, there was little difference between NEC and control samples over the duration of sampling. Whilst there was a consistently elevated abundance in control samples over the duration of sampling and evidence of a late increase that was distinct to control samples after 50 days of age, there was little distinction between case and controls, as indicated overlapping standard errors.

*Dialister* was the only genus to be positively associated with NEC samples. Early sampling showed a trend for increased abundances in NEC samples up to the age of ~35 days, after which levels returned to that observed in control samples by 50 days of age.

The Random Forest model predictions were worse than all subgroups models when performed on the combined sample population. The best model was produced when using only those genera that were seen to be significantly different between NEC and was able to correctly identify NEC samples 47% of the time. *Bifidobacterium* was the most significant genus to contribute to successful predictions within the models, this was approximately twice the accuracy in the model compared to the next most important genus, *Dialister*.

The model also demonstrated the strength of the association between *Dialister* and NEC subjects. However, this was isolated to those infants delivered by caesarean section, fed formula and breast milk, as these were the only NEC subjects observed to have a clear association with this genus over time.

Of the four species observed to be significantly different only two were seen have different trends when factored with time; *Streptococcus agalactiae* and *Clostridium butyricum*. These were seen to have trends for increasing normalised abundances in NEC samples up to the age of ~35 days of age. Controls were not observed to have such trends, although both case and control samples had similar abundances by 50 days of age. These trends were limited to a small number of samples and subjects and were not consistently observed in all NEC samples.

*Streptococcus agalactiae* is a group B streptococcus and implicated as a major pathogen causing a wide variety of problems. It has been found to be a member of the gut community in 10-30% of healthy adults[514,515,516] and it has also been linked to severe pneumonia, sepsis and meningitis in neonates[517]. There are two distinct developments from the group B Streptococci infection. Early onset of the disease is usually accompanied by sepsis and pneumonia from day 0 to 7, whereas the late-onset presentation is often associated with meningitis from day seven until three months of age. Maternal colonisation by *S. algalactiae* is a known risk factor in the development of this disease[518],[519] and often associated with vertical transmission from the faecal microbiota[515]. As these infants all went through vigorous diagnostic criteria the idea that sepsis instigated by *S. algalactiae* was misdiagnosed as NEC is unlikely, however whether some patients went on to have further complications and develop full NEC after infection could indicate either a causative agent or a subset of infants that develop NEC in a unique manner. However, there were no data regarding positive bacterial cultures and as such it was not possible to know which infants had infections.

*Clostridium butyricum* is a strictly anaerobic spore-forming bacillus and a common member of the human and animal gut microbiota. As described earlier the infant microbiome is progressively colonised by increasing proportions of strictly anaerobic bacteria, including *C. butyricum*[520] and by 33 weeks 44% of asymptomatic infants positive for its presence in faecal samples[521]. Whilst this bacterium has been considered beneficial and even used as a probiotic, predominantly in Asia[522], is has also been implicated in botulism and NEC more recently.

Botulism is an acute paralytic disease caused by the botulinum neurotoxin which is secreted by *Clostridium botulinum*[523]. In 1976 Pickett *et al* described the first instance of botulism occurring in an infant in America[524]; in 1986 it was first described in Italy and since has been observed in other countries (China, India, Japan, Ireland and the USA)[525]. This occurrence of the botulinum neurotoxin gene across multiple *Clostridium* species appears to be linked to horizontal transmission mediated by plasmids or phage[526].

As stated in the introduction, no one species has been consistently linked to NEC, however *Clostridium* is one of the most commonly associated with its occurrence[434]. The first association of NEC and *C. butyricum* was in 1977 when it was observed in nine out of ten faecal samples from preterm neonates during an outbreak of NEC[181]. Further studies indicated that this could be due to the introduction of the species by medical staff after staff tested positive for the presence of *C. butyricum* on their hands during an NEC outbreak[433], and preventative measures were seen to control such outbreaks[51].

A more recent, large scale study identified *C. butyricum* specifically in NEC samples using 16S rRNA sequencing and culture based methods, indicating a possible toxigenic mechanism alongside dysbiosis, with an oxidised, acid, and poorly diversified gut microbiota[527]. Four genes have been identified as potential homologues of haemolysins associated with swine dysentery[522]. Of these haemolysins the pore-forming β-haemolysin has been implicated in causing enterocyte necrotic lesions via the culture supernatant[528,51]. Experimental models in animals have also proven that *C. butyricum* can cause NEC-like lesions[529,434]. *C. butyricum* has been implicated in the pathogenesis of NEC via the fermentation of carbohydrate products, however this was dependent on the lactase deficiency of preterm neonates[530,531].

### 5.6.5 Conclusion

All subgroups were seen to have significantly elevated abundances of *Bifidobacterium* when samples were analysed at the genera level and without considering time as a factor. When time was factored into the analysis, only the vaginally delivered control infants exclusively fed breast milk were not seen to be clearly divergent from NEC infants with respect to *Bifidobacterium* abundances. When analysed at the species level no subgroup showed significant differences in the normalised mean abundances of *Bifidobacterium* species. Whilst the genus is important in the discrimination between NEC and control samples there appeared to be no clear probiotic species. This indicated that multiple *Bifidobacterium* species were associated with the community structure of control subjects and less likely to be found in high abundances in NEC subjects. However, the species of *Bifidobacterium* that were observed across control subjects varied per infant and were not consistently present in all subjects.

The influence of subject specific variation was observed multiple times and across multiple subgroups. This was most evident for the subgroup of infants delivered vaginally and fed breast milk exclusively, this highlighted the importance of sampling density and cohort size. The subset of infants vaginally delivered and fed exclusively breast milk was inhibited by the lack of sampling for NEC subjects between 25 and 39 days of age. This severely limited the conclusions that could be made with regards to colonisation patterns as infants were seen to have spikes in taxonomic abundances or a subject specific dominance of taxa that distorted the overall trends of the population.

Random Forest tress were effective at identifying the NEC samples from controls in both subgroups fed a mixture of formula and breast milk. This was likely to be due to the increased sampling density of these subgroups. However, consistently observed across all models was an increased abundance of *Bifidobacterium* negatively associated with NEC samples. Subsets showed a marked improvement in the model predictions and the discrimination of samples, especially with respect to the genus *Bifidobacterium*. This was clearly shown by the reduced class error in mode/feed subset models relative to the model predictions for all samples from each of these subsets being analysed as a whole.

# 6 Discussion

The aim of this project was to establish the taxonomic changes in the microbiome of premature infants that could be used prognostically to identify subjects who develop NEC. By doing so it could aid in the early introduction of treatments known to reduce the incidence or severity of NEC. Previous studies had demonstrated an association with the developing microbiome and the onset of NEC however very few characterised consistent changes between studies or across large datasets. By leveraging the low cost and high throughput potential of modern sequencing technologies it was possible to perform a large scale, prospective cohort analysis by sequencing the V4 region of the 16S ribosomal subunit. In doing so it was possible to assess the presence of trends in the community (LCBD values), similar community structures (NMDS) and unique taxonomic signals (Random Forest) associated with NEC in relation to relevant medical information. Through high sampling rates and careful case-control cohort assignment the trends and taxa that associated with NEC, one of the most important neonatal diseases currently afflicting premature infants, were established.

## 6.1 Cohort Demographics

The cohort demographics were seen to have the same significant associations with NEC incidence previously described in literature[532], namely gestational duration[419], birthweight[533] and the number of antimicrobials administered[534]. In addition to mode of delivery gestational duration and birthweight were metrics or events that occurred prior to the onset of NEC with which a clear cause-effect hypothesis could be established. One major limitation with this project was the inability to make associations with antimicrobial regimes and changes in the microbiome. This was due to the lack of administration timing and dosage being included in the patient metadata.

Whilst there was a great effort invested in obtaining the antimicrobial administration dates for each subject they were not obtained for the majority of enrolled infants in this project, or for most of the subjects assigned in case-control cohort. Antimicrobials have been shown to have an influential impact both in the short term and long term composition of the gut microbiome community[429,432]. It was considered especially important that these could be accounted for when comparing NEC and control infant community profiles, however without the date and duration an infant was administered the antimicrobial it was impossible to make any conclusive statements about the influence they had on the microbiome. However, the validity of antimicrobial administration analysis is debatable as many of the infants were administered multiple antimicrobials over their sampling duration. Being able to discriminate the influence of combined antimicrobial regimes on the microbiome of an infant could be beyond the capabilities of even the most regimented and strictly controlled studies.

This time-of-effect also limited the conclusions that could be made from different feeding regimes. Since no metadata was provided that described when an infant was administered a specific feed it was not possible to establish the effect feeds had on the microbiota. However, it was possible to identify conserved trends over time for generalised feeding regimes, as well as the differences between those regimes, and this was sufficient in demonstrating significant taxonomic differences during the development of infant microbiome, and whether NEC had a greater association to a given feeding regime.

## 6.2 Case-Control Cohort Selection

It was important that match factors were identified to appropriately assign a case-control cohort due to the scale of prospective enrolment in this project. The weighting of match factors was done with the assistance of medical staff who had extensive experience in the NICU environment and medical research alongside an experienced statistician. This made it possible to incorporate many factors and prioritise them accordingly in order to identify the best possible candidates to analyse. The sampling density from the best candidates was then assessed and those with sufficient sampling depth at the appropriate ages were analysed.

Across the population all NEC subjects were observed to be significantly different from infants that did not develop NEC for all the match factors. Following control assignment this was still observed but the difference was less than that of the general population with regards to key risk factors. However, by combining all the match factors into a single weighting score it limited the ability to normalised controls for each match factor.

While some factors were observed to increase in significance difference due to the lower weighting assigned to them (gender, location and feeding regime), these were not seen to be significantly association to NEC. However, risk factors that were associated with NEC (gestation, birthweight, mode of delivery) were seen to be less different between case and control subjects relative to the whole cohort.

Sampling was a limiting factor within this project. Premature infants do not pass faeces as often as full-term infants and, additionally, premature infants produce less stool when they do pass. This greatly reduced the number of eligible NEC infants that could be used within this study and limited control assignment and comparisons across infants based on the age at sampling. This issue was more pronounced when subjects were subset according to important match factors in Chapters 4 and 5. It was possible to identify those groups that had sufficient sampling to provide statistically viable results whilst spanning a sufficient number of days to establish a time series.

This dearth in sampling was associated with NEC infants who are often treated with nil-by-mouth or have symptoms that lead to a reduction in the frequency and quantity of stools produced. Unfortunately, this is a limitation in studying the disease that cannot be avoided in a non-invasive analysis of the premature microbiome done in such a manner. However, the aggregation of this sequence data, through community databases, with future research could help to complement future projects on NEC.

Stool samples are also limited in how accurately they can describe the true profile of the intestinal microbiome *in vivo*. The intestinal tract has multiple communities at various stages which currently cannot be discriminated in each stool sample analysed, regardless of the quantity[535]. Whether the stool sample accurately depicts these communities in the premature infants within this study was not a prerogative, although it would aid in understanding the functions and pathways that may be different between case and control subjects. Stool samples were used because they provided a non-invasive technique of sampling close to, and intimately related with, events happening in the intestine, close to the known disease sites.

## 6.3 Match Factors Associated with Trends in a Developing Microbiome

CCA analysis demonstrated that LCBD values were most associated with the individual and showed high inter-individual variation across the sample population. This has been observed in infants previously[419,420] and is unsurprising due to the high number of unique factors that are likely to influence the development of the microbiome e.g. environment, visitors, antimicrobial regimes, SOPs etc.

Linear regression and subset regression identified significant associations with the LCBD values in infant samples and the age at sampling, gestational duration and feeding regime. This confirmed that the age at which an infant was sampled was an important factor to normalise for as the taxonomic differences between infant microbiomes over the first year of life are great. Therefore, comparisons were made between NEC and controls that were of similar ages when sampled.

Gestational duration was an important factor and has been correlated with increased abundances of pathogenic taxa as well are reduced diversity and lower abundances of the more common, communal, microbiota taxa[419,536]. However, these differences described in literature were observed between preterm and term infants and further analysis of samples based on gestation showed little association with changes in LCBD values over time. This could be considered unexpected as gestational duration is correlated to maturation and as discussed in the introduction of the thesis, the maturity of the intestinal tract has important health implications for premature infants. However, this is likely to be due to the cohort consisting exclusively of premature infants thus reducing the range in gestational duration sampled within the cohort. Therefore, the differences in the LCBD values describing the individual sample species diversity relative to all the collective sample diversity is likely to be much less than if the samples were compared over a larger gestational range or with full term infant samples. An additional point to consider is that the data were analysed as a time series and other environmental factors will play a more significant role in influence changes in diversity the further from the date of delivery an infant is analysed.

Infants either on a mixed feeding regime or fed exclusively breast milk were seen to have a significant association with trends in LCBD values over time.  This association of feeding regimes and changes in the microbiome has been explored extensively and it was suggested to be vital in the development of the microbiome during the first six months[314]. For example infants fed breast milk, relative to those fed formula milk, are seen to acquire more *Bifidobacteria* and *Lactobacillus*[318].

The data presented here show that infants fed a mixture of formula and breast milk were seen to have stable LCBD values over the duration of sampling, regardless of the NEC status of the infant. In contrast, infants fed breast milk exclusively were observed to have an increase in LCBD scores before stabilising at ~40 days of age. These results show that there are different trends in the colonisation of the microbiota depending on which feeding regime the infant was administered.

Mode of delivery was not observed to be significantly associated to LCBD trends over time, however it was considered an important match factor in the case-control cohort assignment. Additionally, there is evidence that vaginally delivered infants showed more similarity to the maternal faecal microflora than those delivered by caesarean section, who had a greater similarity to the maternal skin flora and the microbiome of the NICU[156,309]. Because of this evidence and the close duration between birth and sampling it was considered important to factor into the analysis.

Similar trends were observed for sample LCBD values over time, irrespective of the delivery method. Both groups exhibited similar trends seen for samples from infants fed exclusively breast milk. This indicated that feeding regime, specifically those exclusively fed breast milk, was the main factor influencing the LCBD values within these sample subsets.

## 6.4 Comparison of Community Compositions of Samples from Infants with Necrotising Enterocolitis and Assigned Controls

Literature has described clustering associated with the taxonomic composition of samples and the NEC status of infants in NMDS plots[110]. Even though this study only described a single site and small number of subjects it provided evidence that the communities of NEC subjects might be more similar to each other than to controls, indicating an underlying, consistent signal for NEC. This warranted investigation with the larger dataset presented here.

The clustering of microbiota community compositions does not appear in larger populations from multiple sites. One of the factors that would influence these results was the high inter-individual variability. The results presented within this thesis showed no evidence of discrete clustering of NEC samples or control samples. The overall population showed the lowest discrimination between case and control samples (1.5%) which was improved upon in certain subsets, namely samples from infants fed breast milk exclusively (2.2%).

The significance of clustering, established with PERMANOVA, showed that samples subset by feeding regime and mode of delivery increased the discrimination of NEC and control samples. However only samples from infants fed mixed feeding regimes delivered either by caesarean section or vaginally were significantly clustered. Further analysis that attempted to incorporate gestation and the age at sampling failed to improve or contribute to the clustering in any meaningful way.

Overall NMDS analysis with feeding regimes and mode of delivery provided the most effective clustering, however it was not possible to re-create the clustering demonstrated by Richmond *et al*. This implied that the community composition was not a significant factor in descriminating NEC and controls samples. Further investigation focused on identifying unique taxonomic signatures that could be associated with increased or decreased susceptibility to NEC.

## 6.5 Taxonomic Differences Between Subjects with Necrotising Enterocolitis and Assigned Controls

The data showed that there are significant differences in the abundance of genera between infants who developed NEC and those who did not. The taxonomic abundances of genera were initially analysed within subgroups shown to either be significantly associated with the trends in the microbiome (feeding regime) or important in the initial colonisation of the microbiome (mode of delivery). Following this, samples from these subsets were analysed as a population to establish whether there were conserved differences between NEC and control sample taxonomic abundances in the population relative to subsets of samples. In total, 16 NEC subjects had between three and eight samples pre-NEC (mean = 5.36) with a total of 86 samples, two more subjects had a single sample each pre-NEC diagnosis. Overall the pre-diagnosis samples represented 58% of all samples from NEC subjects.

*Bifidobacterium* was observed to be significantly different over time for samples taken from infants delivered by caesarean section and fed breast milk and formula feeds. This genus had greater normalised mean abundances in control samples consistently over the duration of sampling. The error rate for Random Forest predictions of NEC for samples from infants with NEC was 33% when using significantly different genera between NEC and control samples and irrespective of time. Other genera seen to be significant within the model generation were *Dialister* and *Staphylococcus*. There has been no clear research suggesting characteristics shared between these genera in relation to the gut microbiota disorders, however there is some evidence that they are co-associated in the oral microbiota of patients suffering from a range of disorders including chemomechanical preparation[537], persistent root canal infections[538] and irreversible pulpitis[539].

*Bifidobacterium* was also observed to have significantly different normalised abundances between NEC and control samples taken from infants delivered vaginally and fed both formula milk and breast milk. *Veillonella* was the only other genus seen to be significantly different in normalised mean abundance for this subset when factored with time. These genera also the greatest influence in the model generation (as shown by the mean decreased accuracy and Gini Index scores) and were seen to be elevated in control samples. Error rates for models produced with this subset of samples were much lower than those for caesarean delivered infants on mixed feeds (12%). This suggested that the differences between samples based on genera were clearer within this group of infants.

Whilst there was evidence of differences for genera within the subgroup that were delivered vaginally and fed breast milk exclusively this group lacked consistent sampling for NEC subjects when factored with time. This resulted in a high degree of bias from control subjects and limited the conclusions that could be made. However, there was evidence that increased abundances of *Bifidobacterium*, *Bacteroides* and *Clostridium* in control samples and *Proteus* had greater normalised mean abundances in NEC samples where the sampling density was acceptable (between 40-60 days of age). The limitations of this subgroup were further confirmed with the high error rate in predicting the NEC status of subjects.

Further support for the positive association of *Bifidobacterium* within the microbiome of healthy subjects was seen when all samples from these subsets were combined and assessed as a population. *Bacteroides* was also seen to be significantly associated with control samples within this population analysis, but was not conserved through all subsets.

There were no clear signs of species that presented as either probiotic or pathogenic in either subsets or population analysis. This demonstrated that there is no single causative agent associated NEC in this dataset. Interestingly there were no annotations for *Bifidobacterium* species and only three OTUs assigned to the genus in total. This meant that it was not possible to observe whether a particular species was significantly elevated in any subset or across the population despite the consistent association of the genus with controls samples. This could suggest that multiple species contribute in small proportions to the microbiome of health subjects.

# 7 Conclusion

This project demonstrated that the unique nature of the microbiome and the high degree of inter-individual variation within the community makes direct comparisons challenging. In testing match factors defined by leading medical personnel for significant association with NEC infants and establishing subsets based on these results it was possible to increase the ability to differentiate between NEC and control samples at the community level.

Further analysis with non-reductive normalisation of abundance data in conjunction with machine learning methodologies enabled the detection of taxonomy significantly associated with control infant microbiomes. These results support the current interest in utilising *Bifidobacterium* as a probiotic, however there was no evidence for a specific species from this genus that would provide the scaffold for a community associated with healthy outcomes.

Additionally, the results from this project indicate that there is no clear pathogenic species associated across large populations and NEC. There was limited evidence for taxa at any taxonomic level being associated with NEC microbiomes, although further research into the influence of *Dialister* and *Proteus* could prove enlightening.

This project was able to utilise a large cohort and high sample counts to generate subsets of samples based on key match factors with sufficient sampling depth that would otherwise have been difficult. However, case-control matching would have been better implemented with a smaller number of match factors that were more closely associated with NEC. This could have improved the similarity between NEC and control subjects with respect to more important risk factors.

Additionally, the lack of antimicrobial administration information severely limited the use of this information. This was considered one of the most important factors to influence the microbiome of premature infants and being able to match case and controls based on the same administration regimes at the time of sampling could have provided a wealth of information with regards to the impact on the microbiome and any association with NEC.

The initial premise of this project was to identify diagnostic or prognostic factors associated with the microbiome and the development of NEC. Whilst it was possible to identify key taxa associated with non-NEC microbiomes it was not possible to establish whether these established a presence prior to the onset of NEC. This was due to the large time frame over which patients developed NEC. As it was not possible to account for the extremely active colonisation process during the early stages of life it would have been inappropriate to compare NEC or control subjects relative to the date of diagnosis irrespective of the age at sampling.

Further research should aim to establish the impact the antimicrobials have on the premature infant microbiome and their association with the development of NEC. Additionally, feeding regimes and probiotic trials should look to establish the underlying changes being made to the microbiome and how these are associated with patient health or whether they can be used to reduce the impact of antimicrobial administration on the health of the microbiome.

Overall in NEC subjects there is a clear trend for lower abundances of bacteria known to be beneficial within the gut microbiome, principally *Bifidobacterium*. It was possible to increase the discrimination between NEC and control subject gut microbiotas by factoring in large scale sampling and establishing subsets according to feeding regime and the mode of delivery. Further work should aim to establish the impact of antimicrobial regimes on the high inter-individuality of subject microbiomes. Whilst each microbiome will be unique it is important to understand how the trends in the establishment of stable, healthy infant microbiomes are influenced by outside factors and to what extent these have microbiome associated diseases.

# 8 References

1. Luig, M. & Lui, K. Epidemiology of necrotizing enterocolitis--Part I: Changing regional trends in extremely preterm infants over 14 years. *J. Paediatr. Child Health* **41,** 169–73 (2005).

2. Gagliardi, Ã. L., Bellu, R., Cardilli, V. & Curtis, M. De. Necrotising Enterocolitis in Very Low Birth Weight Infants in Italy : Incidence and Non-nutritional Risk Factors. *J. Pediatr. Gastroenterol. Nutr.* **47,** 206–210 (2008).

3. Arumugam, M. *et al.* Europe PMC Funders Group Enterotypes of the human gut microbiome. **473,** 174–180 (2013).

4. Rajilić-Stojanović, M., Heilig, H. G. H. J., Tims, S., Zoetendal, E. G. & De Vos, W. M. Long-term monitoring of the human intestinal microbiota composition. *Environ. Microbiol.* **15,** 1146–1159 (2013).

5. Voreades, N., Kozil, A. & Weir, T. L. Diet and the development of the human intestinal microbiome. *Front. Microbiol.* **5,** 1–9 (2014).

6. Cho, I. & Blaser, M. J. The human microbiome: at the interface of health and disease. *Nat. Rev. Genet.* **13,** 260–70 (2012).

7. Niemarkt, H. J. *et al.* Necrotizing enterocolitis: a clinical review on diagnostic biomarkers and the role of the intestinal microbiota. *Inflamm. Bowel Dis.* **21,** 436–44 (2015).

8. Jacob, J. A. In Infants With Necrotizing Enterocolitis, Gut Dysbiosis Precedes Disease. 1–2 (2016). doi:10.1001/jama.2016.4341

9. Baucells, B. J., Hally, M. M., Sánchez, A. T. Á. & Aloy, J. F. Probiotic associations in the prevention of necrotising enterocolitis and the reduction of late-onset sepsis and neonatal mortality in preterm infants under 1500g. A systematic review. *An. Pediatría (English Ed.* (2015). doi:10.1016/j.anpede.2015.07.021

10. Peter, C. S. *et al.* Necrotising enterocolitis: is there a relationship to specific pathogens? *Eur. J. Pediatr.* **158,** 67–70 (1999).

11. Lopetuso, L. R. *et al.* Gut Microbiota in Health, Diverticular Disease, Irritable Bowel Syndrome, and Inflammatory Bowel Diseases: Time for Microbial Marker of Gastrointestinal Disorders? *Dig. Dis.* **5,** 134–7 (2017).

12. Böhn, L. *et al.* Diet Low in FODMAPs Reduces Symptoms of Irritable Bowel Syndrome as Well as Traditional Dietary Advice: A Randomized Controlled Trial. *Gastroenterology* **149,** 1399–1407.e2 (2015).

13. Hrnčířová, L., Krejsek, J., Šplíchal, I. & Hrnčíř, T. Crohn's Disease: A Role of Gut Microbiota and NOD2 Gene Polymorphisms in Disease Pathogenesis. *Acta Medica Cordoba.* 89–96 (2014). doi:10.14712/18059694.2014.46

14. CM, B. *Traité des maladies des enfants nouvea*. (1928).

15. Obladen, M. Necrotizing enterocolitis--150 years of fruitless search for the cause. *Neonatology* **96,** 203–210 (2009).

16. QUAISER, K. [A specially severe form of enteritis in newborn, enterocolitis ulcerosa necroticans. II. Clinical studies]. *Osterr. Z. Kinderheilkd. Kinderfuersorge.* **8,** 136–52 (1952).

17. SCHMID, K. O. [A specially severe form of enteritis in newborn, enterocolitis ulcerosa necroticans. I. Pathological anatomy]. *Osterr. Z. Kinderheilkd. Kinderfuersorge.* **8,** 114–35 (1952).

18. O'Shea, T. M., Klinepeter, K. L., Goldstein, D. J., Jackson, B. W. & Dillard, R. G. Survival and developmental disability in infants with birth weights of 501 to 800 grams, born between 1979 and 1994. *Pediatrics* **100,** 982–6 (1997).

19. Doyle, L. W. *et al.* Changing mortality and causes of death in infants 23-27 weeks' gestational age. *J. Paediatr. Child Health* **35,** 255–259 (1999).

20. Hack, M., Friedman, H. & Fanaroff, A. A. Outcomes of extremely low birth weight infants. *Pediatrics* **98,** 931–7 (1996).

21. Lorenz, J. M. Survival of the extremely preterm infant in North America in the 1990s. *Clin. Perinatol.* **27,** 255–262 (2000).

22. Touloukian, R. J., Berdon, W. E., Amoury, R. A. & Santulli, T. V. Surgical experience with necrotizing enterocolitis in the infant. *J. Pediatr. Surg.* **2,** 389–401 (1967).

23. Agerty, H. A., Ziserman, A. J. & Shollenberger, C. L. A case of perforation of the ileum in a newborninfant with operation and recovery. *J. Pediatr.* **22,** 233–238 (1943).

24. O'Neill Jr, J. A. & Holcomb Jr, G. W. Surgical experience with neonatal necrotizing enterocolitis (NNE). *Ann. Surg.* **189,** 612–619 (1979).

25. Bell, M. J. *et al.* Neonatal necrotizing enterocolitis. Therapeutic decisions based upon clinical staging. *Ann. Surg.* **187,** 1–7 (1978).

26. Richmond, J. A. & Mikity, V. Benign form of necrotizing enterocolitis. *Am. J. Roentgenol. Radium Ther. Nucl. Med.* **123,** 301–6 (1975).

27. Wilson, R. *et al.* Age at onset of necrotizing enterocolitis: an epidemiologic analysis. *Pediatr. Res.* **16,** 82–5 (1982).

28. Han, V. K., Sayed, H., Chance, G. W., Brabyn, D. G. & Shaheed, W. A. An outbreak of Clostridium difficile necrotizing enterocolitis: a case for oral vancomycin therapy? *Pediatrics* **71,** 935–941 (1983).

29. Popoff, M. R. & Ravisse, P. Lesions produced by Clostridium butyricum strain CB 1002 in ligated intestinal loops in guinea pigs. *J. Med. Microbiol.* **19,** 351–7 (1985).

30. Cushing, A. H. Necrotizing enterocolitis with Escherichia coli heat-labile enterotoxin. *Pediatrics* **71,** 626–630 (1983).

31. Blakey, J. L. *et al.* Enteric colonization in sporadic neonatal necrotizing enterocolitis. *J. Pediatr. Gastroenterol. Nutr.* **4,** 591–5 (1985).

32. Walsh, M. C. & Kliegman, R. M. Necrotizing enterocolitis: treatment based on staging criteria. *Pediatr Clin North Am* **33,** 179–201 (1986).

33. Ballance, W. A., Dahms, B. B., Shenker, N. & Kliegman, R. M. Pathology of neonatal necrotizing enterocolitis: A ten-year experience. *J. Pediatr.* **117,** (1990).

34. Neu, J. Necrotizing enterocolitis: the search for a unifying pathogenic theory leading to prevention. *Pediatr. Clin. North Am.* **43,** 409–32 (1996).

35. Kosloske, a M. A unifying hypothesis for pathogenesis and prevention of necrotizing enterocolitis. *J. Pediatr.* **117,** S68-74 (1990).

36. Quigley, M. & Mcguire, W. Formula versus donor breast milk for feeding preterm or

low birth weight infants ( Review ). *Cochrane Libr.* 1–92 (2014). doi:10.1002/14651858.CD002971.pub3.www.cochranelibrary.com

37. Berseth, C. L. Effect of early feeding on maturation of the preterm infant's small intestine. *J. Pediatr.* **120,** 947–953 (1992).

38. Hooper, L. V. Bacterial contributions to mammalian gut development. *Trends Microbiol.* **12,** 129–134 (2004).

39. Weisburg, W. G., Barns, S. M., Pelletier, D. A. & Lane, D. J. 16S ribosomal DNA amplification for phylogenetic study. *J. Bacteriol.* **173,** 697–703 (1991).

40. Voelkerding, K. V., Dames, S. A. & Durtschi, J. D. Next-generation sequencing:from basic research to diagnostics. *Clin. Chem.* **55,** 641–658 (2009).

41. Parameswaran, P. *et al.* A pyrosequencing-tailored nucleotide barcode design unveils opportunities for large-scale sample multiplexing. *Nucleic Acids Res.* **35,** 1–9 (2007).

42. Song, S. J., Dominguez-Bello, M. G. & Knight, R. How delivery mode and feeding can shape the bacterial community in the infant gut. *Cmaj* **185,** 373–374 (2013).

43. Fanaro, S., Chierici, R., Guerrini, P. & Vigi, V. Intestinal microflora in early infancy: composition and development. *Acta paediatr Suppl* **92,** 48–55 (2003).

44. Turnbaugh, P. J. *et al.* Human gut microbiome viewed across age and geography. *Nature* **457,** 222–227 (2009).

45. Deshpande, G., Rao, S., Patole, S. & Bulsara, M. Updated meta-analysis of probiotics for preventing necrotizing enterocolitis in preterm neonates. *Pediatrics* **125,** 921–30 (2010).

46. Warner, B. B. *et al.* Gut bacteria dysbiosis and necrotising enterocolitis in very low birthweight infants: a prospective case-control study. *Lancet (London, England)* **387,** 1928–36 (2016).

47. Aschner, J. L., Deluga, K. S., Metlay, L. A., Emmens, R. W. & Hendricks-Munoz, K. D. Spontaneous focal gastrointestinal perforation in very low birth weight infants. *J. Pediatr.* **113,** 364–7 (1988).

48. Mintz, A. C. & Applebaum, H. Focal gastrointestinal perforations not associated with necrotizing enterocolitis in very low birth weight neonates. *J. Pediatr. Surg.* **28,** 857–860 (1993).

49. Donta, S. T., Stuppy, M. S. & Myers, M. G. Neonatal antibiotic-associated colitis. *Am. J. Dis. Child.* **135,** 181–2 (1981).

50. O'Neill, J. a, Stahlman, M. T. & Meng, H. C. Necrotizing enterocolitis in the newborn: operative indications. *Ann. Surg.* **182,** 274–9 (1975).

51. Hutto, D. L. & Wannemuehler, M. J. A comparison of the morphologic effects of Serpulina hyodysenteriae or its beta-hemolysin on the murine cecal mucosa. *Vet. Pathol.* **36,** 412–22 (1999).

52. Kliegman, R. M. Neonatal necrotizing enterocolitis: implications for an infectious disease. *Pediatr. Clin. North Am.* **26,** 327–344 (1979).

53. Moomjian, A. S. *et al.* 1000 NECROTIZING ENTEROCOLITIS-ENDEMIC VS. EPIDEMIC FORM. *Pediatr. Res.* **12,** 530–530 (1978).

54. Hsueh, W. *et al.* Neonatal necrotizing enterocolitis: Clinical considerations and

pathogenetic concepts. *Pediatr. Dev. Pathol.* **6,** 6–23 (2003).

55. Behrman, R. E. *et al.* Epidemiology of necrotizing enterocolitis: A case control study. *J. Pediatr.* **96,** 447–451 (1980).

56. Llanos, A. R. *et al.* Epidemiology of neonatal necrotising enterocolitis: a population-based study. *Paediatr. Perinat. Epidemiol.* **16,** 342–9 (2002).

57. Guthrie, S. O. *et al.* Necrotizing enterocolitis among neonates in the United States. *J. Perinatol.* **23,** 278–285 (2003).

58. Rees, C. M., Eaton, S. & Pierro, A. National prospective surveillance study of necrotizing enterocolitis in neonatal intensive care units. *J. Pediatr. Surg.* **45,** 1391–1397 (2010).

59. Boccia, D., Stolfi, I., Lana, S. & Moro, M. L. Nosocomial necrotising enterocolitis outbreaks: epidemiology and control measures. *Eur. J. Pediatr.* **160,** 385–91 (2001).

60. Battersby, C., Santhalingam, T., Costeloe, K. & Modi, N. Incidence of neonatal necrotising enterocolitis in high-income countries: A systematic review. *Arch. Dis. Child. Fetal Neonatal Ed.* **103,** F182–F189 (2018).

61. Gale, C., Morris, I. & Neonatal Data Analysis Unit (NDAU) Steering Board. The UK National Neonatal Research Database: using neonatal data for research, quality improvement and more. *Arch. Dis. Child. Educ. Pract. Ed.* **101,** 216–8 (2016).

62. Battersby, C., Longford, N., Costeloe, K., Modi, N. & UK Neonatal Collaborative Necrotising Enterocolitis Study Group. Development of a Gestational Age-Specific Case Definition for Neonatal Necrotizing Enterocolitis. *JAMA Pediatr.* **171,** 256–263 (2017).

63. Collado, M. C. *et al.* Factors Influencing Gastrointestinal Tract and Microbiota Immune Interaction in Preterm Infants. *Pediatr. Res.* **77,** 726–731 (2015).

64. Gewolb, I. H., Schwalbe, R. S., Taciak, V. L., Harrison, T. S. & Panigrahi, P. Stool microflora in extremely low birthweight infants. *Arch. Dis. Child. Fetal Neonatal Ed.* **80,** F167-73 (1999).

65. Goldmann, D. a, Leclair, J. & Macone, a. Bacterial colonization of neonates admitted to an intensive care environment. *J. Pediatr.* **93,** 288–93 (1978).

66. Jacquot, A. *et al.* Dynamics and clinical evolution of bacterial gut microflora in extremely premature patients. *J. Pediatr.* **158,** 390–6 (2011).

67. Arboleya, S., Sol??s, G., Fern??ndez, N., de los Reyes-Gavil??n, C. G. & Gueimonde, M. Facultative to strict anaerobes ratio in the preterm infant microbiota: A target for intervention? *Gut Microbes* **3,** 583–588 (2012).

68. Anand, R. J., Leaphart, C. L., Mollen, K. P. & Hackam, D. J. The role of the intestinal barrier in the pathogenesis of necrotizing enterocolitis. *Shock* **27,** 124–33 (2007).

69. Musemeche, C. A., Kosloske, A. M., Bartow, S. A. & Umland, E. T. Comparative efects of ischemia, bacteria, and substrate on the pathogenesis of intestinal necrosis. *J. Pediatr. Surg.* **21,** 536–538 (1986).

70. Choi, Y. Y. Necrotizing enterocolitis in newborns: Update in pathophysiology and newly emerging therapeutic strategies. *Korean J. Pediatr.* **57,** 505–513 (2014).

71. Stoll, B. J. *et al.* Neonatal outcomes of extremely preterm infants from the NICHD Neonatal Research Network. *Pediatrics* **126,** 443–56 (2010).

72. Abu-Shaweesh, J. M. & Martin, R. J. Neonatal apnea: What's new? *Pediatr. Pulmonol.* **43,** 937–944 (2008).

73. Wohlferd, M., Goldberg, J. & Golbus, M. Two thirds of spontaneous abortion/fetal deaths after genetic midtrimester amniocentesis are the result of a pre-existing subclinical inflammatory process of the amniotic cavity. *Am. J. Obstet. Gynecol.* **172,** 261 (1995).

74. Combs, A., Garite, T. J. & Lapidus, J. 14: Amniotic fluid glucose and interleukin-6 as independent markers of intraamniotic infection in preterm labor. *Am. J. Obstet. Gynecol.* **216,** S10–S11 (2017).

75. Wenstrom, K. D. *et al.* Elevated amniotic fluid interleukin-6 levels at genetic amniocentesis predict subsequent pregnancy loss. *Am. J. Obstet. Gynecol.* **175,** 830–833 (1996).

76. Downard, C. D. *et al.* Treatment of necrotizing enterocolitis: an American Pediatric Surgical Association Outcomes and Clinical Trials Committee systematic review. *J. Pediatr. Surg.* **47,** 2111–2122 (2012).

77. Ho, M. Y. & Yen, Y. H. Trend of Nutritional Support in Preterm Infants. *Pediatr. Neonatol.* **57,** 365–370 (2016).

78. Zhang, H. Y., Wang, F. & Feng, J. X. Intestinal microcirculatory dysfunction and neonatal necrotizing enterocolitis. *Chin. Med. J. (Engl).* **126,** 1771–1778 (2013).

79. Neal, M. D. *et al.* A critical role for TLR4 induction of autophagy in the regulation of enterocyte migration and the pathogenesis of necrotizing enterocolitis. *J. Immunol.* **190,** 3541–51 (2013).

80. Richter, J. M. *et al.* LPS-binding protein enables intestinal epithelial restitution despite LPS exposure. *J. Pediatr. Gastroenterol. Nutr.* **54,** 639–44 (2012).

81. Terrin, G., Scipione, A. & De Curtis, M. Update in Pathogenesis and Prospective in Treatment of Necrotizing Enterocolitis. *Biomed Res. Int.* **2014,** (2014).

82. Akira, S. & Takeda, K. Toll-like receptor signalling. *Nat. Rev. Immunol.* **4,** 499–511 (2004).

83. Hoffmann, J. a, Kafatos, F. C., Janeway, C. a & Ezekowitz, R. a. Phylogenetic perspectives in innate immunity. *Science* **284,** 1313–1318 (1999).

84. Leaphart, C. L. *et al.* A critical role for TLR4 in the pathogenesis of necrotizing enterocolitis by modulating intestinal injury and repair. *J. Immunol.* **179,** 4808–4820 (2007).

85. Richardson, W. M. *et al.* Nucleotide-binding oligomerization domain-2 inhibits toll-like receptor-4 signaling in the intestinal epithelium. *Gastroenterology* **139,** 904–17, 917.e1–6 (2010).

86. Sodhi, C. P. *et al.* Toll-like receptor-4 inhibits enterocyte proliferation via impaired beta-catenin signaling in necrotizing enterocolitis. *Gastroenterology* **138,** 185–96 (2010).

87. Sodhi, C. P. *et al.* Intestinal epithelial Toll-like receptor 4 regulates goblet cell development and is required for necrotizing enterocolitis in mice. *Gastroenterology* **143,** 708-18.e1–5 (2012).

88. Martin, N. A. *et al.* Active transport of bile acids decreases mucin 2 in neonatal ileum:

Implications for development of necrotizing enterocolitis. *PLoS One* **6,** (2011).

89. McDonough, H. & Patterson, C. CHIP: a link between the chaperone and proteasome systems. *Cell Stress Chaperones* **8,** 303–308 (2003).

90. Afrazi, A. *et al.* Intracellular heat shock protein-70 negatively regulates TLR4 signaling in the newborn intestinal epithelium. *J. Immunol.* **188,** 4543–57 (2012).

91. Good, M. *et al.* Amniotic fluid inhibits Toll-like receptor 4 signaling in the fetal and neonatal intestinal epithelium. *Proc. Natl. Acad. Sci.* **109,** 11330–11335 (2012).

92. Soliman, A. *et al.* Platelet-activating factor induces TLR4 expression in intestinal epithelial cells: Implication for the pathogenesis of necrotizing enterocolitis. *PLoS One* **5,** 1–8 (2010).

93. Claud, E. C. & Walker, W. A. Hypothesis: inappropriate colonization of the premature intestine can cause neonatal necrotizing enterocolitis. *FASEB J. Off. Publ. Fed. Am. Soc. Exp. Biol.* **15,** 1398–1403 (2001).

94. Walker, A. W. *et al.* Dominant and diet-responsive groups of bacteria within the human colonic microbiota. *ISME J.* **5,** 220–30 (2011).

95. LeBlanc, J. G. *et al.* Bacteria as vitamin suppliers to their host: A gut microbiota perspective. *Curr. Opin. Biotechnol.* **24,** 160–168 (2013).

96. Honda, K. & Littman, D. R. The microbiota in adaptive immune homeostasis and disease. *Nature* **535,** 75–84 (2016).

97. Sommer, F. & Bäckhed, F. The gut microbiota--masters of host development and physiology. *Nat. Rev. Microbiol.* **11,** 227–38 (2013).

98. Wopereis, H., Oozeer, R., Knipping, K., Belzer, C. & Knol, J. The first thousand days - intestinal microbiology of early life: establishing a symbiosis. *Pediatr. Allergy Immunol.* 1–11 (2014). doi:10.1111/pai.12232

99. Raveh-Sadka, T. *et al.* Gut bacteria are rarely shared by co-hospitalized premature infants, regardless of necrotizing enterocolitis development. *Elife* **2015,** 1–25 (2015).

100. Claud, E. C. *et al.* Bacterial community structure and functional contributions to emergence of health or necrotizing enterocolitis in preterm infants. *Microbiome* **1,** 20 (2013).

101. Normann, E., Fahlén, A., Engstrand, L. & Lilja, H. E. Intestinal microbial profiles in extremely preterm infants with and without necrotizing enterocolitis. *Acta Paediatr. Int. J. Paediatr.* **102,** 129–136 (2013).

102. Morrow, A. L. *et al.* Early microbial and metabolomic signatures predict later onset of necrotizing enterocolitis in preterm infants. *Microbiome* **1,** 13 (2013).

103. Torrazza, R. M. *et al.* Intestinal microbial ecology and environmental factors affecting necrotizing enterocolitis. *PLoS One* **8,** 1–9 (2013).

104. Mai, V. *et al.* Fecal microbiota in premature infants prior to necrotizing enterocolitis. *PLoS One* **6,** e20647 (2011).

105. Zhou, Y. *et al.* Longitudinal analysis of the premature infant intestinal microbiome prior to necrotizing enterocolitis: a case-control study. *PLoS One* **10,** e0118632 (2015).

106. Sim, K. *et al.* Dysbiosis anticipating necrotizing enterocolitis in very premature infants. *Clin. Infect. Dis.* **60,** 389–397 (2015).

107. Neu, J., Chen, M. & Beierle, E. Intestinal innate immunity: How does it relate to the pathogenesis of necrotizing enterocolitis. *Semin. Pediatr. Surg.* **14,** 137–144 (2005).

108. La Rosa, P. S. *et al.* Patterned progression of bacterial populations in the premature infant gut. *Proc. Natl. Acad. Sci.* **111,** 6–11 (2014).

109. Lu, P., Sodhi, C. P. & Hackam, D. J. Toll-like receptor regulation of intestinal development and inflammation in the pathogenesis of necrotizing enterocolitis. *Pathophysiol. Off. J. Int. Soc. Pathophysiol.* **21,** 81–93 (2014).

110. Wang, Y. *et al.* 16S rRNA gene-based analysis of fecal microbiota from preterm infants with and without necrotizing enterocolitis. *ISME J.* **3,** 944–954 (2009).

111. Coates, E. W., Karlowicz, M. G., Croitoru, D. P. & Buescher, E. S. Distinctive distribution of pathogens associated with peritonitis in neonates with focal intestinal perforation compared with necrotizing enterocolitis. *Pediatrics* **116,** e241-6 (2005).

112. Bury, R. G. & Tudehope, D. Enteral antibiotics for preventing necrotizing enterocolitis in low birthweight or preterm infants. *Cochrane database Syst. Rev.* CD000405 (2001). doi:10.1002/14651858.CD000405

113. Krediet, T. *et al.* Microbiological factors associated with neonatal necrotizing enterocolitis: protective effect of early antibiotic treatment. *Acta Paediatr.* **92,** 1180–1182 (2007).

114. Patel, R. M. *et al.* Probiotic bacteria induce maturation of intestinal claudin 3 expression and barrier function. *Am. J. Pathol.* **180,** 626–635 (2012).

115. Weaver, L. T., Laker, M. F. & Nelson, R. Intestinal permeability in the newborn. *Arch. Dis. Child.* **59,** 236–241 (1984).

116. van Elburg, R. M., Fetter, W. P. F., Bunkers, C. M. & Heymans, H. S. A. Intestinal permeability in relation to birth weight and gestational and postnatal age. *Arch. Dis. Child. Fetal Neonatal Ed.* **88,** F52-5 (2003).

117. Dvorak, B. Milk epidermal growth factor and gut protection. *J. Pediatr.* **156,** S31-5 (2010).

118. Henning, S. J. Development of the gastrointestinal tract. *Proc. Nutr. Soc.* **45,** 39–44 (1986).

119. Goldman, A. S. Modulation of the gastrointestinal tract of infants by human milk. Interfaces and interactions. An evolutionary perspective. *J. Nutr.* **130,** 426S–431S (2000).

120. Colomé, G. *et al.* Intestinal permeability in different feedings in infancy. *Acta Paediatr. Int. J. Paediatr.* **96,** 69–72 (2007).

121. Weaver, L. T., Laker, M. F., Nelson, R. & Lucas, A. Milk feeding and changes in intestinal permeability and morphology in the newborn. *J. Pediatr. Gastroenterol. Nutr.* **6,** 351–8

122. Rakoff-Nahoum, S., Paglino, J., Eslami-Varzaneh, F., Edberg, S. & Medzhitov, R. Recognition of commensal microflora by toll-like receptors is required for intestinal homeostasis. *Cell* **118,** 229–41 (2004).

123. Hooper, L. V. *et al.* Molecular analysis of commensal host-microbial relationships in the intestine. *Science* **291,** 881–4 (2001).

124. Piena-Spoel, M., Albers, M. J. I. J., Ten Kate, J. & Tibboel, D. Intestinal permeability

in newborns with necrotizing enterocolitis and controls: Does the sugar absorption test provide guidelines for the time to (Re-)introduce enteral nutrition? *J. Pediatr. Surg.* **36,** 587–592 (2001).

125. Kuppala, V. S., Meinzen-Derr, J., Morrow, A. L. & Schibler, K. R. Prolonged initial empirical antibiotic treatment is associated with adverse outcomes in premature infants. *J. Pediatr.* **159,** 720–725 (2011).

126. Robinson, J. Cochrane in context: probiotics for prevention of necrotizing enterocolitis in preterm infants. *Evid Based Child Heal.* **9,** 672–674 (2014).

127. Clark, J. A. *et al.* Intestinal barrier failure during experimental necrotizing enterocolitis: Protective effect of EGF treatment. *Am. J. Physiol. - Gastrointest. Liver Physiol.* **291,** 938–949 (2006).

128. Han, X., Fink, M. P. & Delude, R. L. Proinflammatory cytokines cause NO*-dependent and -independent changes in expression and localization of tight junction proteins in intestinal epithelial cells. *Shock* **19,** 229–37 (2003).

129. Shiou, S. R. *et al.* Erythropoietin protects intestinal epithelial barrier function and lowers the incidence of experimental neonatal necrotizing enterocolitis. *J. Biol. Chem.* **286,** 12123–12132 (2011).

130. Beumer, C. *et al.* Calf intestinal alkaline phosphatase, a novel therapeutic drug for lipopolysaccharide (LPS)-mediated diseases, attenuates LPS toxicity in mice and piglets. *J. Pharmacol. Exp. Ther.* **307,** 737–744 (2003).

131. Koyama, I., Matsunaga, T., Harada, T., Hokari, S. & Komoda, T. Alkaline phosphatases reduce toxicity of lipopolysaccharides in vivo and in vitro through dephosphorylation. *Clin. Biochem.* **35,** 455–461 (2002).

132. Tuin, A., Huizinga-Van der Vlag, A., van Loenen-Weemaes, A.-M. M. A., Meijer, D. K. F. & Poelstra, K. On the role and fate of LPS-dephosphorylating activity in the rat liver. *Am. J. Physiol. Gastrointest. Liver Physiol.* **290,** G377-85 (2006).

133. van Veen, S. Q. *et al.* Bovine intestinal alkaline phosphatase attenuates the inflammatory response in secondary peritonitis in mice. *Infect. Immun.* **73,** 4309–14 (2005).

134. O'Boyle, C. J. *et al.* Microbiology of bacterial translocation in humans. *Gut* **42,** 29–35 (1998).

135. Balzan, S., de Almeida Quadros, C., de Cleva, R., Zilberstein, B. & Cecconello, I. Bacterial translocation: overview of mechanisms and clinical impact. *J. Gastroenterol. Hepatol.* **22,** 464–471 (2007).

136. Souza, D. G. *et al.* The Essential Role of the Intestinal Microbiota in Facilitating Acute Inflammatory Responses. *J. Immunol.* **173,** 4137–4146 (2004).

137. Blencowe, H. *et al.* National, regional, and worldwide estimates of preterm birth rates in the year 2010 with time trends since 1990 for selected countries: A systematic analysis and implications. *Lancet* **379,** 2162–2172 (2012).

138. Kramer, M. S. *et al.* Secular trends in preterm birth: a hospital-based cohort study. *Jama* **280,** 1849–54 (1998).

139. Howson, C. P., Kinney, M. V, McDougall, L. & Lawn, J. E. Born too soon: preterm birth matters. *Reprod. Health* **10 Suppl 1,** S1 (2013).

140. Office for National Statistics. Child mortality in England and Wales: 2016. (2016).

141. Rees, C. M., Eaton, S. & Pierro, A. Trends in infant mortality from necrotising enterocolitis in England and Wales and the USA. *Arch. Dis. Child. Fetal Neonatal Ed.* **93,** F395-6 (2008).

142. Tucker, J. M. *et al.* Etiologies of Preterm Birth in an Indigent Population: Is Prevention a Logical Expectation? *Obstet. Gynecol.* **77,** 343–347 (1991).

143. Ananth, C. V & Vintzileos, A. M. Epidemiology of preterm birth and its clinical subtypes. *J Matern Fetal Neonatal Med* **19,** 773–782 (2006).

144. Diseases, C. Surveillance of necrotising enterocolitis, 1981-2. *Br. Med. J. (Clin. Res. Ed).* **287,** 824–6 (1983).

145. Berrington, J. E., Hearn, R. I., Bythell, M., Wright, C. & Embleton, N. D. Deaths in preterm infants: Changing pathology over 2 decades. *J. Pediatr.* **160,** 49–53.e1 (2012).

146. Hooper, L. V, Midtvedt, T. & Gordon, J. I. How host-microbial interactions shape the nutrient environment of the mammalian intestine. *Annu. Rev. Nutr.* **22,** 283–307 (2002).

147. Hooper, L. V, Stappenbeck, T. S., Hong, C. V & Gordon, J. I. Angiogenins: a new class of microbicidal proteins involved in innate immunity. *Nat. Immunol.* **4,** 269–273 (2003).

148. Stappenbeck, T. S., Hooper, L. V & Gordon, J. I. Developmental regulation of intestinal angiogenesis by indigenous microbes via Paneth cells. *Proc. Natl. Acad. Sci. U. S. A.* **99,** 15451–15455 (2002).

149. Mshvildadze, M. *et al.* Intestinal microbial ecology in premature infants assessed with non-culture-based techniques. *J. Pediatr.* **156,** 20–5 (2010).

150. Claud, E. C. *et al.* Developmentally regulated IkappaB expression in intestinal epithelium and susceptibility to flagellin-induced inflammation. *Proc. Natl. Acad. Sci. U. S. A.* **101,** 7404–8 (2004).

151. Son, M., Grobman, W. A. & Miller, E. S. Is mode of delivery associated with the risk of necrotizing enterocolitis? *Am. J. Obstet. Gynecol.* **214,** S204-s205 (2016).

152. El Aidy, S. *et al.* Temporal and spatial interplay of microbiota and intestinal mucosa drive establishment of immune homeostasis in conventionalized mice. *Mucosal Immunol.* **5,** 567–579 (2012).

153. Hooper, L. V, Littman, D. R. & Macpherson, A. J. Interactions between the microbiota and the immune system. *Science* **336,** 1268–73 (2012).

154. Martin, R. *et al.* Early life: Gut microbiota and immune development in infancy. *Benef. Microbes* **1,** 367–382 (2010).

155. Gronlund, M. M., Lehtonen, O. P., Eerola, E. & Kero, P. Fecal microflora in healthy infants born by different methods of delivery: permanent changes in intestinal flora after cesarean delivery. *J Pediatr Gastroenterol Nutr* **28,** 19–25 (1999).

156. Tannock, G. W., Fuller, R., Smith, S. L. & Hall, M. A. Plasmid profiling of members of the family Enterobacteriaceae, lactobacilli, and bifidobacteria to study the transmission of bacteria from mother to infant. *J. Clin. Microbiol.* **28,** 1225–1228 (1990).

157. Long, S. S. & Swenson, R. M. Development of anaerobic fecal flora in healthy

newborn infants. *J. Pediatr.* **91,** 298–301 (1977).

158. Bennet, R. & Nord, C. E. Development of the faecal anaerobic microflora after caesarean section and treatment with antibiotics in newborn infants. *Infection* **15,** 332–336 (1987).

159. Grand, R. J., Watkins, J. B. & Torti, F. M. Development of the human gastrointestinal tract. A review. *Gastroenterology* **70,** 790–810 (1976).

160. Guilloteau, P., Biernat, M., Woliński, J. & Zabielski, R. Chapter 11 Gut regulatory peptides and hormones of the small intestine. *Biol. Grow. Anim.* **1,** 325–362 (2002).

161. Book, L. S., Herbst, J. J., Atherton, S. O. & Jung, A. L. Necrotizing enterocolitis in low-birth-weight infants fed an elemental formula. *J. Pediatr.* **87,** 602–605 (1975).

162. Agostoni, C. *et al.* Enteral nutrient supply for preterm infants: commentary from the European Society of Paediatric Gastroenterology, Hepatology and Nutrition Committee on Nutrition. *J. Pediatr. Gastroenterol. Nutr.* **50,** 85–91 (2010).

163. Schanler, R. J., Shulman, R. J. & Lau, C. Feeding strategies for premature infants: beneficial outcomes of feeding fortified human milk versus preterm formula. *Pediatrics* **103,** 1150–1157 (1999).

164. Lucas, A. & Cole, T. J. Breast milk and neonatal necrotising enterocolitis. *Lancet (London, England)* **336,** 1519–1523 (1990).

165. Herrmann, K. & Carroll, K. An exclusively human milk diet reduces necrotizing enterocolitis. *Breastfeed. Med.* **9,** 184–90 (2014).

166. Uauy, R. D. *et al.* Necrotizing enterocolitis in very low birth weight infants: biodemographic and clinical correlates. National Institute of Child Health and Human Development Neonatal Research Network. *J. Pediatr.* **119,** 630–638 (1991).

167. Rromano-Keeler, J. *et al.* Early life establishment of site-specific microbial communities in the gut. *Gut Microbes* **5,** 192–201 (2014).

168. Clark, R. H. *et al.* Characteristics of patients who die of necrotizing enterocolitis. *J. Perinatol.* **32,** 199–204 (2012).

169. Bizzarro, M. J., Ehrenkranz, R. A. & Gallagher, P. G. Concurrent bloodstream infections in infants with necrotizing enterocolitis. *J. Pediatr.* **164,** 61–6 (2014).

170. Ullrich, T. *et al.* Absence of gastrointestinal pathogens in ileum tissue resected for necrotizing enterocolitis. *Pediatr. Infect. Dis. J.* **31,** 413–4 (2012).

171. Cotten, C. M. *et al.* Prolonged Duration of Initial Empirical Antibiotic Treatment Is Associated With Increased Rates of Necrotizing Enterocolitis and Death for Extremely Low Birth Weight Infants. *Pediatrics* **123,** 58–66 (2009).

172. Alexander, V. N., Northrup, V. & Bizzarro, M. J. Antibiotic exposure in the newborn intensive care unit and the risk of necrotizing enterocolitis. *J. Pediatr.* **159,** 392–7 (2011).

173. Coggins, S. A., Wynn, J. L. & Weitkamp, J.-H. Infectious causes of necrotizing enterocolitis. *Clin. Perinatol.* **42,** 133–54, ix (2015).

174. Bagci, S. *et al.* Detection of Astrovirus in Premature Infants With Necrotizing Enterocolitis. *Pediatr. Infect. Dis. J.* **27,** 347–350 (2008).

175. Bagci, S. *et al.* Clinical characteristics of viral intestinal infection in preterm and term

neonates. *Eur. J. Clin. Microbiol. Infect. Dis.* **29,** 1079–1084 (2010).

176. Chappé, C. *et al.* Astrovirus and digestive disorders in neonatal units. *Acta Paediatr. Int. J. Paediatr.* **101,** 208–212 (2012).

177. Stewart, C. J. *et al.* Bacterial and fungal viability in the preterm gut: NEC and sepsis. *Arch. Dis. Child. Fetal Neonatal Ed.* **98,** F298-303 (2013).

178. Parra-Herran, C. E., Pelaez, L., Sola, J. E., Urbiztondo, A. K. & Rodriguez, M. M. Intestinal candidiasis: an uncommon cause of necrotizing enterocolitis (NEC) in neonates. *Fetal Pediatr. Pathol.* **29,** 172–180 (2010).

179. Smith, S. D. *et al.* The hidden mortality in surgically treated necrotizing enterocolitis: fungal sepsis. *J. Pediatr. Surg.* **25,** 1030–3 (1990).

180. Mitchell, R. G., Etches, P. C. & Day, D. G. Non-toxigenic clostridia in babies. *J. Clin. Pathol.* **34,** 217–20 (1981).

181. Howard, F. M., Flynn, D. M., Bradley, J. M., Noone, P. & Szawatkowski, M. Outbreak of necrotising enterocolitis caused by Clostridium butyricum. *Lancet (London, England)* **2,** 1099–1102 (1977).

182. Sturm, R., Staneck, J. L., Stauffer, L. R. & Neblett, W. W. 3rd. Neonatal necrotizing enterocolitis associated with penicillin-resistant, toxigenic Clostridium butyricum. *Pediatrics* **66,** 928–931 (1980).

183. Tengsupakul, S. *et al.* Asymptomatic DNAemia heralds CMV-associated NEC: case report, review, and rationale for preemption. *Pediatrics* **132,** e1428-34 (2013).

184. Gessler, P., Bischoff, G. A., Wiegand, D., Essers, B. & Bossart, W. Cytomegalovirus-associated necrotizing enterocolitis in a preterm twin after breastfeeding. *J. Perinatol.* **24,** 124–6 (2004).

185. Tran, L. *et al.* Necrotizing enterocolitis and cytomegalovirus infection in a premature infant. *Pediatrics* **131,** e318-22 (2013).

186. Al Jumaili, I. J., Shibley, M., Lishman, A. H. & Record, C. O. Incidence and origin of Clostridium difficile in neonates. *J. Clin. Microbiol.* **19,** 77–78 (1984).

187. Mathew, O. P., Bhatia, J. S. & Richardson, C. J. An outbreak of Clostridium difficile necrotizing enterocolitis. *Pediatrics* **73,** 265–266 (1984).

188. Chany, C., Moscovici, O., Lebon, P. & Rousset, S. Association of coronavirus infection with neonatal necrotizing enterocolitis. *Pediatrics* **69,** 209–214 (1982).

189. Dittmar, E. *et al.* Necrotizing enterocolitis of the neonate with Clostridium perfringens: Diagnosis, clinical course, and role of alpha toxin. *Eur. J. Pediatr.* **167,** 891–895 (2008).

190. Schlapbach, L. J., Ahrens, O., Klimek, P., Berger, S. & Kessler, U. Clostridium perfringens and necrotizing enterocolitis. *J. Pediatr.* **157,** 175 (2010).

191. de la Cochetiere, M.-F. *et al.* Early intestinal bacterial colonization and necrotizing enterocolitis in premature infants: the putative role of Clostridium. *Pediatr. Res.* **56,** 366–70 (2004).

192. Lake, A. M., Lauer, B. A., Clark, J. C., Wesenberg, R. L. & McIntosh, K. Enterovirus infections in neonates. *J. Pediatr.* **89,** 787–791 (1976).

193. Johnson, F. E., Crnic, D. M., Simmons, M. a & Lilly, J. R. Association of fatal

Coxsackie B2 viral infection and necrotizing enterocolitis. *Arch. Dis. Child.* **52,** 802–4 (1977).

194. Birenbaum, E. *et al.* Echovirus type 22 outbreak associated with gastro-intestinal disease in a neonatal intensive care unit. *Am. J. Perinatol.* **14,** 469–473 (1997).

195. Stoll, B. J., Hansen, N., Fanaroff, A. A. & Lemons, J. A. Enterobacter sakazakii is a rare cause of neonatal septicemia or meningitis in VLBW infants. *J. Pediatr.* **144,** 821–823 (2004).

196. van Acker, J. *et al.* Outbreak of necrotizing enterocolitis associated with Enterobacter sakazakii in powdered milk formula. *J. Clin. Microbiol.* **39,** 293–7 (2001).

197. Townsend, S., Hurrell, E. & Forsythe, S. Virulence studies of Enterobacter sakazakii isolates associated with a neonatal intensive care unit outbreak. *BMC Microbiol.* **8,** 2180–2189 (2008).

198. Hunter, C. J. *et al.* Enterobacter sakazakii enhances epithelial cell injury by inducing apoptosis in a rat model of necrotizing enterocolitis. *J. Infect. Dis.* **198,** 586–93 (2008).

199. Desfrere, L. *et al.* Increased incidence of necrotizing enterocolitis in premature infants born to HIV-positive mothers. *AIDS* **19,** 1487–1493 (2005).

200. Schmitz, T., Weizsaecker, K., Feiterna-Sperling, C., Eilers, E. & Obladen, M. Exposure to HIV and antiretroviral medication as a potential cause of necrotizing enterocolitis in term neonates. *AIDS* **20,** 1082–3 (2006).

201. Van Der Meulen, E. F., Bergman, K. A. & Kamps, A. W. A. Necrotising enterocolitis in a term neonate with trisomy 21 exposed to maternal HIV and antiretroviral medication. *Eur. J. Pediatr.* **168,** 113–114 (2009).

202. Raskind, C. H., Dembry, L.-M. & Gallagher, P. G. Vancomycin-resistant enterococcal bacteremia and necrotizing enterocolitis in a preterm neonate. *Pediatr. Infect. Dis. J.* **24,** 943–4 (2005).

203. Stuart, R. L. *et al.* An outbreak of necrotizing enterocolitis associated with norovirus genotype GII.3. *Pediatr. Infect. Dis. J.* **29,** 644–647 (2010).

204. Pelizzo, G. *et al.* Isolated colon ischemia with norovirus infection in preterm babies: a case series. *J. Med. Case Rep.* **7,** 108 (2013).

205. Turcios-Ruiz, R. M. *et al.* Outbreak of Necrotizing Enterocolitis Caused by Norovirus in a Neonatal Intensive Care Unit. *J. Pediatr.* **153,** 339–344 (2008).

206. Armbrust, S., Kramer, A., Olbertz, D., Zimmermann, K. & Fusch, C. Norovirus infections in preterm infants: wide variety of clinical courses. *BMC Res. Notes* **2,** 96 (2009).

207. Bell, M. J., Feigin, R. D., Ternberg, J. L. & Brotherton, T. Evaluation of gastrointestinal microflora in necrotizing enterocolitis. *J. Pediatr.* **92,** 589–592 (1978).

208. Speer, M. E. *et al.* Fulminant neonatal sepsis and necrotizing enterocolitis associated with a ‘nonenteropathogenic’ strain of Escherichia coli. *J. Pediatr.* **89,** 91–95 (1976).

209. Guner, Y. S., Malhotra, A., Ford, H. R., Stein, J. E. & Kelly, L. K. Association of Escherichia coli O157:H7 with necrotizing enterocolitis in a full-term infant. *Pediatr. Surg. Int.* **25,** 459–463 (2009).

210. Sharma, R. *et al.* Rotavirus-Associated Necrotizing Enterocolitis: An Insight into a Potentially Preventable Disease? *J. Pediatr. Surg.* **39,** 453–457 (2004).

211. Rotbart, H. A. *et al.* Neonatal rotavirus-associated necrotizing enterocolitis: case control study and prospective surveillance during an outbreak. *J. Pediatr.* **112,** 87–93 (1988).

212. Keller, K. M., Schmidt, H., Wirth, S., Queisser-Luft, A. & Schumacher, R. Differences in the clinical and radiologic patterns of rotavirus and non-rotavirus necrotizing enterocolitis. *Pediatr. Infect. Dis. J.* **10,** 734–738 (1991).

213. Gregersen, N. *et al.* Klebsiella pneumoniae with extended spectrum beta-lactamase activity associated with a necrotizing enterocolitis outbreak. *Pediatr. Infect. Dis. J.* **18,** 963–967 (1999).

214. Hill, H. R., Hunt, C. E. & Matsen, J. M. Nosocomial colonization with Klebsiella, type 26, in a neonatal intensive-care unit associated with an outbreak of sepsis, meningitis, and necrotizing enterocolitis. *J. Pediatr.* **85,** 415–419 (1974).

215. Stone, H. H., Kolb, L. D. & Geheber, C. E. Bacteriologic considerations in perforated necrotizing enterocolitis. *South. Med. J.* **72,** 1540–4 (1979).

216. Vaucher, Y. E. *et al.* Pleomorphic, enveloped, virus-like particles associated with gastrointestinal illness in neonates. *J. Infect. Dis.* **145,** 27–36 (1982).

217. Lodha, A., de Silva, N., Petric, M. & Moore, A. M. Human torovirus: a new virus associated with neonatal necrotizing enterocolitis. *Acta Paediatr.* **94,** 1085–1088 (2005).

218. Cheng, Y.-L., Lee, H.-C., Yeung, C.-Y. & Chan, W.-T. Clinical significance in previously healthy children of Pseudomonas aeruginosa in the stool. *Pediatr. Neonatol.* **50,** 13–7 (2009).

219. Henderson, A., Maclaurin, J. & Scott, J. M. Pseudomonas in a Glasgow baby unit. *Lancet (London, England)* **2,** 316–7 (1969).

220. Leigh, L., Stoll, B. J., Rahman, M. & McGowan, J. Pseudomonas aeruginosa infection in very low birth weight infants: a case-control study. *Pediatr. Infect. Dis. J.* **14,** 367–71 (1995).

221. Pumberger, W. & Novak, W. Fatal neonatal Salmonella enteritidis sepsis. *J. Perinatol.* **20,** 54–56 (2000).

222. Stein, H., Beck, J., Solomon, A. & Schmaman, A. Gastroenteritis with Necrotizing Enterocolitis in Premature Babies. *J. Investig. Dermatology Lancet Br. Med. J.* **38,** 202–775 (1968).

223. Overturf, G. D., Sherman, M. P., Scheifele, D. W. & Wong, L. C. Neonatal necrotizing enterocolitis associated with delta toxin-producing methicillin-resistant Staphylococcus aureus. *Pediatr.Infect.Dis.J.* **9,** 88–91 (1990).

224. Stewart, C. J. *et al.* The preterm gut microbiota: changes associated with necrotizing enterocolitis and infection. *Acta Paediatr.* **101,** 1121–7 (2012).

225. Ng, P. C. *et al.* Bacterial contaminated breast milk and necrotizing enterocolitis in preterm twins. *J. Hosp. Infect.* **31,** 105–10 (1995).

226. Okogbule-Wonodi, A. C. *et al.* Necrotizing enterocolitis is associated with ureaplasma colonization in preterm infants. *Pediatr. Res.* **69,** 442–447 (2011).

227. Perzigian, R. W. *et al.* Ureaplasma urealyticum and chronic lung disease in very low birth weight infants during the exogenous surfactant era. *Pediatr. Infect. Dis. J.* **17,**

620–5 (1998).

228. Engum, S. A. & Grosfeld, J. L. Necrotizing enterocolitis. *Curr. Opin. Pediatr.* **10,** 123–30 (1998).

229. Atkinson, S. D., Tuggle, D. W. & Tunell, W. P. Hypoalbuminemia may predispose infants to necrotizing enterocolitis. *J. Pediatr. Surg.* **24,** 674–676 (1989).

230. Tapia-Rombo, C. A., Velasco-Lavin, M. R. & Nieto-Caldelas, A. Risk factors of necrotizing enterocolitis. *Bol. Med. Hosp. Infant. Mex.* **50,** 650–654 (1993).

231. Richard J Schanler, M. Clinical features and diagnosis of necrotizing enterocolitis in newborns. *UpToDate* (2017). Available at: https://www-uptodate-com.liverpool.idm.oclc.org/contents/clinical-features-and-diagnosis-of-necrotizing-enterocolitis-in-newborns#H6. (Accessed: 20th March 2017)

232. Yu, V. Y., Tudehope, D. I. & Gill, G. J. Neonatal necrotizing enterocolitis: 1. Clinical aspects. *Med. J. Aust.* **1,** 685–688 (1977).

233. Sharma, R. *et al.* Portal venous gas and surgical outcome of neonatal necrotizing enterocolitis. *J. Pediatr. Surg.* **40,** 371–6 (2005).

234. Bömelburg, T. & von Lengerke, H. J. Sonographic findings in infants with suspected necrotizing enterocolitis. *Eur. J. Radiol.* **15,** 149–153 (1992).

235. Silva, C. T. *et al.* Correlation of sonographic findings and outcome in necrotizing enterocolitis. *Pediatr. Radiol.* **37,** 274–282 (2007).

236. Bohnhorst, B. Usefulness of abdominal ultrasound in diagnosing necrotising enterocolitis. *Arch. Dis. Child. Fetal Neonatal Ed.* **98,** F445-50 (2013).

237. Muchantef, K. *et al.* Sonographic and radiographic imaging features of the neonate with necrotizing enterocolitis: Correlating findings with outcomes. *Pediatr. Radiol.* **43,** 1444–1452 (2013).

238. Garbi-Goutel, A. *et al.* Prognostic value of abdominal sonography in necrotizing enterocolitis of premature infants born before 33 weeks gestational age. *J. Pediatr. Surg.* **49,** 508–513 (2014).

239. Pierro, A. The surgical management of necrotising enterocolitis. *Early Hum. Dev.* **81,** 79–85 (2005).

240. Neu, J. & Walker, W. A. Necrotizing enterocolitis. *N. Engl. J. Med.* **364,** 255–64 (2011).

241. Stoll, B. J. Epidemiology of necrotizing enterocolitis. *Clin. Perinatol.* **21,** 205–218 (1994).

242. Zani, A. *et al.* International survey on the management of necrotizing enterocolitis. *Eur J Pediatr Surg* **25,** 27–33 (2015).

243. Henry, M. C. W. & Moss, R. L. Current issues in the management of necrotizing enterocolitis. *Semin. Perinatol.* **28,** 221–233 (2004).

244. Ein, S. H., Marshall, D. G. & Girvan, D. Peritoneal drainage under local anesthesia for perforations from necrotizing enterocolitis. *J. Pediatr. Surg.* **12,** 963–7 (1977).

245. Moss, R. L. *et al.* Laparotomy versus peritoneal drainage for necrotizing enterocolitis and perforation. *N. Engl. J. Med.* **354,** 2225–2234 (2006).

246. Ramani, M. & Ambalavanan, N. Feeding practices and necrotizing enterocolitis. *Clin.*

*Perinatol.* **40,** 1–10 (2013).

247. Hans, D. M., Pylipow, M., Long, J. D., Thureen, P. J. & Georgieff, M. K. Nutritional practices in the neonatal intensive care unit: analysis of a 2006 neonatal nutrition survey. *Pediatrics* **123,** 51–57 (2009).

248. Fenton, T. R. *et al.* Validating the weight gain of preterm infants between the reference growth curve of the fetus and the term infant. *BMC Pediatr.* **13,** 1–10 (2013).

249. Viehmann, L. *et al.* Breastfeeding and the use of human milk. *Pediatrics* **129,** e827-41 (2012).

250. Pollack, P. F. *et al.* Effects of enterally fed epidermal growth factor on the small and large intestine of the suckling rat. *Regul. Pept.* **17,** 121–132 (1987).

251. Dvorak, B. *et al.* Milk-borne epidermal growth factor modulates intestinal transforming growth factor-alpha levels in neonatal rats. *Pediatr. Res.* **47,** 194–200 (2000).

252. Dvorak, B. *et al.* Epidermal growth factor reduces the development of necrotizing enterocolitis in a neonatal rat model. *Am J Physiol Gastrointest Liver Physiol* **282,** G156-64 (2002).

253. Book, L. S., Herbst, J. J. & Jung, A. L. Comparison of fast- and slow-feeding rate schedules to the development of necrotizing enterocolitis. *J. Pediatr.* **89,** 463–6 (1976).

254. Neu, J., Mshvildadze, M. & Mai, V. A roadmap for understanding and preventing necrotizing enterocolitis. *Curr. Gastroenterol. Rep.* **10,** 450–457 (2008).

255. Tarnow-Mordi, W. O., Wilkinson, D., Trivedi, A. & Brok, J. Probiotics reduce all-cause mortality and necrotizing enterocolitis: it is time to change practice. *Pediatrics* **125,** 1068–1070 (2010).

256. Lau, J. *et al.* Chorioamnionitis with a fetal inflammatory response is associated with higher neonatal mortality, morbidity, and resource use than chorioamnionitis displaying a maternal inflammatory response only. *Am. J. Obstet. Gynecol.* **193,** 708–713 (2005).

257. Seliga-Siwecka, J. P. & Kornacka, M. K. Neonatal outcome of preterm infants born to mothers with abnormal genital tract colonisation and chorioamnionitis: A cohort study. *Early Hum. Dev.* **89,** 271–275 (2013).

258. Been, J. V, Lievense, S., Zimmermann, L. J. I., Kramer, B. W. & Wolfs, T. G. A. M. Chorioamnionitis as a Risk Factor for Necrotizing Enterocolitis: A Systematic Review and Meta-Analysis. *J. Pediatr.* **162,** 236–242.e2 (2013).

259. Hammerman, C. *et al.* Oral probiotics reduce the incidence and severity of necrotizing enterocolitis in very low birth weight infants Randomized , controlled trial of slow versus rapid feeding volume advancement in preterm infants. 2005 (2005).

260. Gregory, S. G. *et al.* The DNA sequence and biological annotation of human chromosome 1. *Nature* **441,** 315–321 (2006).

261. Sender, R., Fuchs, S. & Milo, R. Revised Estimates for the Number of Human and Bacteria Cells in the Body. *PLoS Biol.* **14,** 1–14 (2016).

262. Yang, X., Xie, L., Li, Y. & Wei, C. More than 9,000,000 unique genes in human gut bacterial community: Estimating gene numbers inside a human body. *PLoS One* **4,** 0–7

(2009).

263. Lederberg, J. Infectious history. *Science* **288,** 287–293 (2000).

264. Turnbaugh, P. J. *et al.* Feature The Human Microbiome Project. *Nature* **449,** 804–810 (2007).

265. Chu, H. *et al.* Gene-microbiota interactions contribute to the pathogenesis of inflammatory bowel disease. *Science* **352,** 1116–20 (2016).

266. Park, J. S., Seo, J. H. & Youn, H.-S. Gut Microbiota and Clinical Disease: Obesity and Nonalcoholic Fatty Liver Disease. *Pediatr. Gastroenterol. Hepatol. Nutr.* **16,** 22–27 (2013).

267. Lakshmanan, V., Selvaraj, G. & Bais, H. P. Functional soil microbiome: belowground solutions to an aboveground problem. *Plant Physiol.* **166,** 689–700 (2014).

268. Cummings, J. H., Hill, M. J., Bone, E. S., Branch, W. J. & Jenkins, D. J. The effect of meat protein and dietary fiber on colonic function and metabolism. II. Bacterial metabolites in feces and urine. *Am. J. Clin. Nutr.* **32,** 2094–101 (1979).

269. Qin, J. *et al.* A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464,** 59–65 (2010).

270. Joossens, M. *et al.* Dysbiosis of the faecal microbiota in patients with Crohn's disease and their unaffected relatives. *Gut* **60,** 631–637 (2011).

271. O'Hara, A. M. & Shanahan, F. The gut flora as a forgotten organ. *EMBO Rep.* **7,** 688–693 (2006).

272. Macfarlane, G. T. & Macfarlane, S. Fermentation in the human large intestine: its physiologic consequences and the potential contribution of prebiotics. *J. Clin. Gastroenterol.* **45l,** S120–S127 (2011).

273. Roediger, W. E. & Nance, S. Metabolic induction of experimental ulcerative colitis by inhibition of fatty acid oxidation. *Br. J. Exp. Pathol.* **67,** 773–82 (1986).

274. Macfarlane, S. Polysaccharide degradation by human intestinal bacteria during growth under multi-substrate limiting conditions in a three-stage continuous culture system. *FEMS Microbiol. Ecol.* **26,** 231–243 (1998).

275. Macfarlane, S. & Macfarlane, G. T. Regulation of short-chain fatty acid production. *Proc. Nutr. Soc.* **62,** 67–72 (2003).

276. Sartor, R. B. Microbial influences in inflammatory bowel diseases. *Gastroenterology* **134,** 577–594 (2008).

277. Magwira, C. A. *et al.* Diversity of faecal oxalate-degrading bacteria in black and white South African study groups: insights into understanding the rarity of urolithiasis in the black group. *J. Appl. Microbiol.* **113,** 418–28 (2012).

278. Thomas, C. M. *et al.* Histamine derived from probiotic Lactobacillus reuteri suppresses TNF via modulation of PKA and ERK signaling. *PLoS One* **7,** e31951 (2012).

279. De Biase, D. & Pennacchietti, E. Glutamate decarboxylase-dependent acid resistance in orally acquired bacteria: function, distribution and biomedical implications of the gadBC operon. *Mol. Microbiol.* **86,** 770–86 (2012).

280. Baddini Feitoza, A., Fernandes Pereira, A., Ferreira da Costa, N. & Gonçalves Ribeiro,

B. Conjugated linoleic acid (CLA): effect modulation of body composition and lipid profile. *Nutr. Hosp.* **24,** 422–8 (2009).

281. Devillard, E., McIntosh, F. M., Duncan, S. H. & Wallace, R. J. Metabolism of linoleic acid by human gut bacteria: Different routes for biosynthesis of conjugated linoleic acid. *J. Bacteriol.* **189,** 2566–2570 (2007).

282. Devillard, E. *et al.* Differences between human subjects in the composition of the faecal bacterial community and faecal metabolism of linoleic acid. *Microbiology* **155,** 513–20 (2009).

283. Velagapudi, V. R. *et al.* The gut microbiota modulates host energy and lipid metabolism in mice. *J. Lipid Res.* **51,** 1101–12 (2010).

284. Marín, L., Miguélez, E. M., Villar, C. J. & Lombó, F. Bioavailability of dietary polyphenols and gut microbiota metabolism: antimicrobial properties. *Biomed Res. Int.* **2015,** 905215 (2015).

285. Jandhyala, S. M. *et al.* Role of the normal gut microbiota. *World J. Gastroenterol.* **21,** 8836–8847 (2015).

286. Perry, G. H. *et al.* Diet and the evolution of human amylase gene copy number variation. *Nat. Genet.* **39,** 1256–60 (2007).

287. Hehemann, J. H. *et al.* Transfer of carbohydrate-active enzymes from marine bacteria to Japanese gut microbiota. *Nature* **464,** 908–912 (2010).

288. Turnbaugh, P. J. *et al.* A core gut microbiome in obese and lean twins. *Nature* **457,** 480–4 (2009).

289. Wold, A. E. & Adlerberth, I. Breast feeding and the intestinal microflora of the infant - Implications for protection against infectious diseases. *Adv. Exp. Med. Biol.* **478,** 77–93 (2000).

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

291. Penders, J. *et al.* Factors Influencing the Composition of the Intestinal Microbiota in Early Infancy. *Pediatrics* **118,** 511–521 (2006).

292. Harmsen, H. J. *et al.* Analysis of intestinal flora development in breast-fed and formula-fed infants by using molecular identification and detection methods. *J. Pediatr. Gastroenterol. Nutr.* **30,** 61–7 (2000).

293. Vos, Q., Lees, a, Wu, Z. Q., Snapper, C. M. & Mond, J. J. B-cell activation by T-cell-independent type 2 antigens as an integral part of the humoral immune response to pathogenic microorganisms. *Immunol. Rev.* **176,** 154–170 (2000).

294. Fox, C. J., Hammerman, P. S. & Thompson, C. B. Fuel feeds function: energy metabolism and the T-cell response. *Nat. Rev. Immunol.* **5,** 844–52 (2005).

295. Michalek, R. D. & Rathmell, J. C. The metabolic life and times of a T cell. *Immunol. Rev.* **236,** 190–202 (2011).

296. Lupton, J. R. Diet Induced Changes in the Colonic Environment and Colorectal Cancer. *J. Nutr* **134,** 479–482 (2004).

297. Peng, L., He, Z., Chen, W., Holzman, I. R. & Lin, J. Effects of butyrate on intestinal barrier function in a caco-2 cell monolayer model of intestinal barrier. *Pediatr. Res.*

**61,** 37–41 (2007).

298. Depestel, D. D. & Aronoff, D. M. Epidemiology of Clostridium difficile infection. *J. Pharm. Pract.* **26,** 464–75 (2013).

299. Rohlke, F. & Stollman, N. Fecal microbiota transplantation in relapsing Clostridium difficile infection. *Therap. Adv. Gastroenterol.* **5,** 403–420 (2012).

300. Langdon, A., Crook, N. & Dantas, G. The effects of antibiotics on the microbiome throughout development and alternative approaches for therapeutic modulation. *Genome Med.* **8,** (2016).

301. Jiménez, E. *et al.* Is meconium from healthy newborns actually sterile? *Res. Microbiol.* **159,** 187–193 (2008).

302. Ardissone, A. N. *et al.* Meconium microbiome analysis identifies bacteria correlated with premature birth. *PLoS One* **9,** 1–8 (2014).

303. Klebanoff, M. & Searle, K. The role of inflammation in preterm birth - Focus on periodontitis. *BJOG An Int. J. Obstet. Gynaecol.* **113,** 43–45 (2006).

304. Stout, M. J. *et al.* Identification of intracellular bacteria in the basal plate of the human placenta in term and preterm gestations. *Am. J. Obstet. Gynecol.* **208,** 1–14 (2013).

305. DiGiulio, D. B. *et al.* Microbial prevalence, diversity and abundance in amniotic fluid during preterm labor: A molecular and culture-based investigation. *PLoS One* **3,** 1–10 (2008).

306. Han, Y. W., Shen, T., Chung, P., Buhimschi, I. A. & Buhimschi, C. S. Uncultivated bacteria as etiologic agents of intra-amniotic inflammation leading to preterm birth. *J. Clin. Microbiol.* **47,** 38–47 (2009).

307. Aagaard, K. *et al.* The placenta harbors a unique microbiome. *Sci. Transl. Med.* **6,** 237ra65 (2014).

308. Van De Wijgert, J. H. H. M. *et al.* The vaginal microbiota: What have we learned after a decade of molecular characterization? *PLoS One* **9,** (2014).

309. Dominguez-Bello, M. G. *et al.* Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 11971–5 (2010).

310. Redondo-Lopez, V., Cook, R. L. & Sobel, J. D. Emerging role of lactobacilli in the control and maintenance of the vaginal bacterial microflora. *Rev. Infect. Dis.* **12,** 856–872 (1990).

311. Palmer, C., Bik, E. M., DiGiulio, D. B., Relman, D. a & Brown, P. O. Development of the human infant intestinal microbiota. *PLoS Biol.* **5,** e177 (2007).

312. Jost, T., Lacroix, C., Braegger, C. P. & Chassard, C. New insights in gut microbiota establishment in healthy breast fed neonates. *PLoS One* **7,** e44595 (2012).

313. Turroni, F. *et al.* Diversity of bifidobacteria within the infant gut microbiota. *PLoS One* **7,** 20–24 (2012).

314. Arumugam, M. *et al.* Enterotypes of the human gut microbiome. *Nature* **473,** 174–80 (2011).

315. WHO. Infant and young child feeding. Available at: http://www.who.int/mediacentre/factsheets/fs342/en/. (Accessed: 22nd March 2017)

316. David, L. A. *et al.* Diet rapidly and reproducibly alters the human gut microbiome. *Nature* **505,** 559–63 (2014).

317. Bezirtzoglou, E., Tsiotsias, A. & Welling, G. W. Microbiota profile in feces of breast- and formula-fed newborns by using fluorescence in situ hybridization (FISH). *Anaerobe* **17,** 478–482 (2011).

318. Kau, A. L., Ahern, P. P., Griffin, N. W., Goodman, A. L. & Gordon, J. I. Human nutrition, the gut microbiome and the immune system. *Nature* **474,** 327–36 (2011).

319. Weng, M. & Walker, W. A. The role of gut microbiota in programming the immune phenotype. *J. Dev. Orig. Health Dis.* **4,** 203–14 (2013).

320. Bergström, A. *et al.* Establishment of intestinal microbiota during early life: A longitudinal, explorative study of a large cohort of Danish infants. *Appl. Environ. Microbiol.* **80,** 2889–2900 (2014).

321. Fallani, M. *et al.* Determinants of the human infant intestinal microbiota after the introduction of first complementary foods in infant samples from five European centres. *Microbiology* **157,** 1385–1392 (2011).

322. Fallani, M. *et al.* Intestinal microbiota of 6-week-old infants across Europe: geographic influence beyond delivery mode, breast-feeding, and antibiotics. *J. Pediatr. Gastroenterol. Nutr.* **51,** 77–84 (2010).

323. Marques, T. M. *et al.* Programming infant gut microbiota: Influence of dietary and environmental factors. *Curr. Opin. Biotechnol.* **21,** 149–156 (2010).

324. Oozeer, R. *et al.* Intestinal microbiology in early life: Specific prebiotics can have similar functionalities as human-milk oligosaccharides. *Am. J. Clin. Nutr.* **98,** (2013).

325. Knol, J. *et al.* Increase of faecal bifidobacteria due to dietary oligosaccharides induces a reduction of clinically relevant pathogen germs in the faeces of formula-fed preterm infants. *Acta Paediatr. Suppl.* **94,** 31–3 (2005).

326. Cabrera-Rubio, R. *et al.* The human milk microbiome changes over lactation and is shaped by maternal weight and mode of delivery. *Am. J. Clin. Nutr.* **96,** 544–51 (2012).

327. De Filippo, C. *et al.* Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 14691–6 (2010).

328. Koenig, J. E. *et al.* Succession of microbial consortia in the developing infant gut microbiome. *Proc. Natl. Acad. Sci. U. S. A.* **108 Suppl,** 4578–4585 (2011).

329. Eckburg, P. B. *et al.* Diversity of the human intestinal microbial flora. *Science* **308,** 1635–8 (2005).

330. Faith, J. J. *et al.* The long-term stability of the human gut microbiota. *Science* **341,** 1237439 (2013).

331. Claesson, M. J. *et al.* Composition, variability, and temporal stability of the intestinal microbiota of the elderly. *Proc. Natl. Acad. Sci. U. S. A.* **108,** 4586–4591 (2011).

332. Bartosch, S., Fite, A., Macfarlane, G. T. & McMurdo, M. E. T. Characterization of bacterial communities in feces from healthy elderly volunteers and hospitalized elderly patients by using real-time PCR and effects of antibiotic treatment on the fecal microbiota. *Appl. Environ. Microbiol.* **70,** 3575–81 (2004).

333. Woodmansey, E. J. Intestinal bacteria and ageing. *J. Appl. Microbiol.* **102,** 1178–1186 (2007).

334. Claesson, M. J. *et al.* Gut microbiota composition correlates with diet and health in the elderly. *Nature* **488,** 178–184 (2012).

335. Qato, D. M. & Johnson, B. Use of Prescription and Over-the-counter Medications and Dietary Supplements Among Older Adults in the United States. *NIH Public Access* **300,** 2867–2878 (2009).

336. Turnbaugh, P. J. *et al.* The effect of diet on the human gut microbiome: a metagenomic analysis in humanized gnotobiotic mice. *Sci. Transl. Med.* **1,** 6ra14 (2009).

337. Ley, R., Turnbaugh, P., Klein, S. & Gordon, J. Microbial ecology: human gut microbes associated with obesity. *Nature* **444,** 1022–3 (2006).

338. Turnbaugh, P. J. *et al.* An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* **444,** 1027–31 (2006).

339. Koeth, R. a *et al.* Intestinal microbiota metabolism of L-carnitine, a nutrient in red meat, promotes atherosclerosis. *Nat. Med.* **19,** 576–85 (2013).

340. Tang, W. W. *et al.* Intestinal Microbial Metabolism of Phosphatidylcholine and Cardiovascular Risk. *N. Engl. J. Med.* **368,** 1575–1584 (2013).

341. Simrén, M. *et al.* Intestinal microbiota in functional bowel disorders: a Rome foundation report. *Gut* **62,** 159–76 (2013).

342. Halmos, E. P. *et al.* Diets that differ in their FODMAP content alter the colonic luminal microenvironment. *Gut* **64,** 93–100 (2015).

343. Shen, E. P. & Surawicz, C. M. Current Treatment Options for Severe Clostridium difficile-associated Disease. *Gastroenterol. Hepatol. (N. Y).* **4,** 134–9 (2008).

344. Rao, K. & Young, V. B. Fecal Microbiota Transplantation for the Management of Clostridium difficile Infection. *Infect. Dis. Clin. North Am.* **29,** 109–122 (2015).

345. Theriot, C. M. *et al.* Antibiotic-induced shifts in the mouse gut microbiome and metabolome increase susceptibility to Clostridium difficile infection. *Nat. Commun.* **5,** 3114 (2014).

346. Hampton-Marcell, J. T., Lopez, J. V. & Gilbert, J. A. The human microbiome: an emerging tool in forensics. *Microb. Biotechnol.* **10,** 228–230 (2017).

347. Dickson, R. P. *et al.* Analysis of culture-dependent versus culture-independent techniques for identification of bacteria in clinically obtained bronchoalveolar lavage fluid. *J. Clin. Microbiol.* **52,** 3605–3613 (2014).

348. Dethlefsen, L., McFall-Ngai, M. & Relman, D. A. An ecological and evolutionary perspective on human-microbe mutualism and disease. *Nature* **449,** 811–818 (2007).

349. Lozupone, C. A., Stombaugh, J. I., Gordon, J. I., Jansson, J. K. & Knight, R. Diversity, stability and resilience of the human gut microbiota. *Nature* **489,** 220–30 (2012).

350. Fedurco, M., Romieu, A., Williams, S., Lawrence, I. & Turcatti, G. BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. *Nucleic Acids Res.* **34,** (2006).

351. Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456,** 53–9 (2008).

352. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437,** 376–80 (2005).

353. Turcatti, G., Romieu, A., Fedurco, M. & Tairi, A. P. A new class of cleavable fluorescent nucleotides: Synthesis and optimization as reversible terminators for DNA sequencing by synthesis. *Nucleic Acids Res.* **36,** (2008).

354. Balasubramanian, S. Sequencing nucleic acids: from chemistry to medicine. *Chem. Commun. (Camb).* **47,** 7281–6 (2011).

355. Woese, C. R. Bacterial evolution. *Microbiol. Rev.* **51,** 221–71 (1987).

356. Klindworth, A. *et al.* Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Res.* **41,** 1–11 (2013).

357. Tap, J. *et al.* Towards the human intestinal microbiota phylogenetic core. *Environ. Microbiol.* **11,** 2574–2584 (2009).

358. Suau, A. *et al.* Direct analysis of genes encoding 16S rRNA from complex communities reveals many novel molecular species within the human gut. *Appl Env. Microbiol* **65,** 4799–4807 (1999).

359. Hold, G. L., Pryde, S. E., Russell, V. J., Furrie, E. & Flint, H. J. Assessment of microbial diversity in human colonic samples by 16S rDNA sequence analysis. *FEMS Microbiol. Ecol.* **39,** 33–39 (2002).

360. Van de Peer, Y., Chapelle, S. & De Wachter, R. A quantitative map of nucleotide substitution rates in bacterial rRNA. *Nucleic Acids Res.* **24,** 3381–3391 (1996).

361. Ghyselinck, J., Pfeiffer, S., Heylen, K., Sessitsch, A. & De Vos, P. The effect of primer choice and short read sequences on the outcome of 16S rRNA gene based diversity studies. *PLoS One* **8,** e71360 (2013).

362. D'Amore, R. *et al.* A comprehensive benchmarking study of protocols and sequencing platforms for 16S rRNA community profiling. *BMC Genomics* **17,** 55 (2016).

363. Yang, B., Wang, Y. & Qian, P.-Y. Sensitivity and correlation of hypervariable regions in 16S rRNA genes in phylogenetic analysis. *BMC Bioinformatics* **17,** 135 (2016).

364. Rintala, A. *et al.* Gut microbiota analysis results are highly dependent on the 16s rRNA gene target region, whereas the impact of DNA extraction is minor. *J. Biomol. Tech.* **28,** 19–30 (2017).

365. Grimes, D. A. & Schulz, K. F. Cohort studies: Marching towards outcomes. *Lancet* **359,** 341–345 (2002).

366. Embleton, N. D. *et al.* Mechanisms Affecting the Gut of Preterm Infants in Enteral Feeding Trials. *Front. Nutr.* **4,** 14 (2017).

367. Her Magesties Government. Human Tissue Act 2004. *HTA Website* 1–63 (2004). doi:10.1258/rsmmlj.72.4.148

368. Choo, J. M., Leong, L. E. X. & Rogers, G. B. Sample storage conditions significantly influence faecal microbiome profiles. *Sci. Rep.* **5,** 16350 (2015).

369. Ariefdjohan, M. W., Savaiano, D. A. & Nakatsu, C. H. Comparison of DNA extraction kits for PCR-DGGE analysis of human intestinal microbial communities from fecal specimens. 1–8 (2010).

370. Patel, R. K. & Jain, M. NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. *PLoS One* **7,** e30619 (2012).

371. Masella, A. P., Bartram, A. K., Truszkowski, J. M., Brown, D. G. & Neufeld, J. D. PANDAseq: paired-end assembler for illumina sequences. *BMC Bioinformatics* **13,** 1–7 (2012).

372. Bankevich, A. *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19,** 455–77 (2012).

373. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–10 (1990).

374. Caporaso, J. G. *et al.* NIH Public Access. **7,** 335–336 (2011).

375. Caporaso, J. G. *et al.* PyNAST: A flexible tool for aligning sequences to a template alignment. *Bioinformatics* **26,** 266–267 (2010).

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

377. A, L. & M, W. Classification and Regression by randomForest. *R News* **2,** 18–22 (2002).

378. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15,** 550 (2014).

379. Torondel, B. *et al.* Assessment of the influence of intrinsic environmental and geographical factors on the bacterial ecology of pit latrines. *Microb. Biotechnol.* **9,** 209–223 (2016).

380. Ijaz, U. Z. http://userweb.eng.gla.ac.uk/umer.ijaz/bioinformatics/ecological.html. (2017).

381. Alfa, M. J. *et al.* An outbreak of necrotizing enterocolitis associated with a novel clostridium species in a neonatal intensive care unit. *Clin. Infect. Dis.* **35,** S101–S105 (2002).

382. Sari, F. N. *et al.* Oral probiotics: Lactobacillus sporogenes for prevention of necrotizing enterocolitis in very low-birth weight infants: a randomized, controlled trial. *Eur. J. Clin. Nutr.* **65,** 434–9 (2011).

383. Lin, H.-Y., Chang, J. H., Chung, M.-Y. & Lin, H.-C. Prevention of necrotizing enterocolitis in preterm very low birth weight infants: Is it feasible? *J. Formos. Med. Assoc.* 1–8 (2013). doi:10.1016/j.jfma.2013.03.010

384. Adlerberth, I. & Wold, A. E. Establishment of the gut microbiota in Western infants. *Acta Paediatr. Int. J. Paediatr.* **98,** 229–238 (2009).

385. Stoll, BJ; Gordon, T; Korones, S. B. Late-onset sepsis in very low birth weight neonates: a report from the National Institute of Child Health and Human Development Neonatal Research Network. *J. Pediatr.* **129,** 63–71 (1996).

386. Berseth, C. L. Gestational evolution of small intestine motility in preterm and term infants. *J Pediatr* **115,** 646–651 (1989).

387. Kafetzis, D. A., Skevaki, C. & Costalos, C. Neonatal necrotizing enterocolitis: an overview. *Curr. Opin. Infect. Dis.* **16,** 349–355 (2003).

388. Glinianaia, S. V., Skjaerven, R. & Magnus, P. Birthweight percentiles by gestational

age in multiple births. A population-based study of Norwegian twins and triplets. *Acta Obstet. Gynecol. Scand.* **79,** 450–8 (2000).

389. Bokulich, N. A. *et al.* Antibiotics, birth mode, and diet shape microbiome maturation during early life. *Sci. Transl. Med.* **8,** 1–14 (2016).

390. Biasucci, G. *et al.* Mode of delivery affects the bacterial community in the newborn gut. *Early Hum. Dev.* **86,** 13–15 (2010).

391. Thavagnanam, S., Fleming, J., Bromley, A., Shields, M. D. & Cardwell, C. R. A meta-analysis of the association between Caesarean section and childhood asthma. *Clin. Exp. Allergy* **38,** 629–33 (2008).

392. Bager, P., Wohlfahrt, J. & Westergaard, T. Caesarean delivery and risk of atopy and allergic disesase: Meta-analyses. *Clin. Exp. Allergy* **38,** 634–642 (2008).

393. Cardwell, C. R. *et al.* Caesarean section is associated with an increased risk of childhood-onset type 1 diabetes mellitus: a meta-analysis of observational studies. *Diabetologia* **51,** 726–35 (2008).

394. Huh, S. Y. *et al.* Delivery by caesarean section and risk of obesity in preschool age children: a prospective cohort study. *Arch. Dis. Child.* **97,** 610–6 (2012).

395. Patel, a L. *et al.* Reducing necrotizing enterocolitis in very low birth weight infants using quality-improvement methods. *J. Perinatol.* 1–8 (2014). doi:10.1038/jp.2014.123

396. Dvorak, B. *et al.* Maternal milk reduces severity of necrotizing enterocolitis and increases intestinal IL-10 in a neonatal rat model. *Pediatr. Res.* **53,** 426–433 (2003).

397. Martin, C., Ling, P.-R. & Blackburn, G. Review of Infant Feeding: Key Features of Breast Milk and Infant Formula. *Nutrients* **8,** 279 (2016).

398. Rashid, M. U., Weintraub, A. & Nord, C. E. Effect of new antimicrobial agents on the ecological balance of human microflora. *Anaerobe* **18,** 249–253 (2012).

399. Willing, B. P., Russell, S. L. & Finlay, B. B. Shifting the balance: Antibiotic effects on host-microbiota mutualism. *Nat. Rev. Microbiol.* **9,** 233–243 (2011).

400. Fouhy, F. *et al.* High-throughput sequencing reveals the incomplete, short-term recovery of infant gut microbiota following parenteral antibiotic treatment with ampicillin and gentamicin. *Antimicrob. Agents Chemother.* **56,** 5811–20 (2012).

401. Kliegman, R. M., Pittard, W. B. & Fanaroff, A. A. Necrotizing enterocolitis in neonates fed human milk. *J. Pediatr.* **95,** 450–453 (1979).

402. Stey, a. *et al.* Outcomes and Costs of Surgical Treatments of Necrotizing Enterocolitis. *Pediatrics* **135,** e1190–e1197 (2015).

403. Thyoka, M. *et al.* Advanced necrotizing enterocolitis part 1: Mortality. *Eur. J. Pediatr. Surg.* **22,** 8–12 (2012).

404. Sullivan, S. *et al.* An Exclusively Human Milk-Based Diet Is Associated with a Lower Rate of Necrotizing Enterocolitis than a Diet of Human Milk and Bovine Milk-Based Products. *J. Pediatr.* **156,** (2010).

405. Carlisle, E. M., Poroyko, V., Caplan, M. S., Alverdy, J. A. & Liu, D. Gram negative bacteria are associated with the early stages of necrotizing enterocolitis. *PLoS One* **6,** 1–7 (2011).

406. LaTuga, M. S. *et al.* Beyond bacteria: A study of the enteric microbial consortium in extremely low birth weight infants. *PLoS One* **6,** 1–10 (2011).

407. Martín, R. *et al.* Human milk is a source of lactic acid bacteria for the infant gut. *J. Pediatr.* **143,** 754–758 (2003).

408. Makino, H. *et al.* Transmission of intestinal Bifidobacterium longum subsp. longum strains from mother to infant, determined by multilocus sequencing typing and amplified fragment length polymorphism. *Appl. Environ. Microbiol.* **77,** 6788–6793 (2011).

409. Matsumiya, Y., Kato, N., Watanabe, K. & Kato, H. Molecular epidemiological study of vertical transmission of vaginal Lactobacillus species from mothers to newborn infants in Japanese, by arbitrarily primed polymerase chain reaction. *J. Infect. Chemother. Off. J. Japan Soc. Chemother.* **8,** 43–49 (2002).

410. Flint, H. J., Scott, K. P., Louis, P. & Duncan, S. H. The role of the gut microbiota in nutrition and health. *Nat. Rev. Gastroenterol. Hepatol.* **9,** 577–589 (2012).

411. Legendre, P. & De Cáceres, M. Beta diversity as the variance of community data: Dissimilarity coefficients and partitioning. *Ecol. Lett.* **16,** 951–963 (2013).

412. Pearce, N. Analysis of matched case-control studies. *BMJ* **352,** i969 (2016).

413. Bray, J. R. & Curtis, J. T. An Ordination of the Upland Forest Communities of Southern Wisconsin. *Source Ecol. Monogr.* **27,** 326–349 (1957).

414. Clarke, K. R. Non-parametric multivariate analyses of changes in community structure. *Aust. J. Ecol.* **18,** 117–143 (1993).

415. Kelly, B. J. *et al.* Power and sample-size estimation for microbiome studies using pairwise distances and PERMANOVA. *Bioinformatics* **31,** 2461–2468 (2015).

416. Spor, A., Koren, O. & Ley, R. Unravelling the effects of the environment and host genotype on the gut microbiome. *Nat. Rev. Microbiol.* **9,** 279–290 (2011).

417. Arboleya, S. *et al.* Establishment and development of intestinal microbiota in preterm neonates. *FEMS Microbiol. Ecol.* **79,** 763–72 (2012).

418. Arboleya, S. *et al.* Deep 16S rRNA metagenomics and quantitative PCR analyses of the premature infant fecal microbiota. *Anaerobe* **18,** 378–380 (2012).

419. Dogra, S. *et al.* Dynamics of infant gut microbiota are influenced by delivery mode and gestational duration and are associated with subsequent adiposity. *MBio* **6,** 1–9 (2015).

420. Dethlefsen, L. & Relman, D. A. Incomplete recovery and individualized responses of the human distal gut microbiota to repeated antibiotic perturbation. *Proc. Natl. Acad. Sci. U. S. A.* **108 Suppl,** 4554–61 (2011).

421. Huse, S. M. *et al.* Exploring microbial diversity and taxonomy using SSU rRNA hypervariable tag sequencing. *PLoS Genet.* **4,** e1000255 (2008).

422. Beaugerie, L. & Petit, J.-C. Microbial-gut interactions in health and disease. Antibiotic-associated diarrhoea. *Best Pract. Res. Clin. Gastroenterol.* **18,** 337–52 (2004).

423. Wilcox, M. H. Gastrointestinal disorders and the critically ill. Clostridium difficile infection and pseudomembranous colitis. *Best Pract. Res. Clin. Gastroenterol.* **17,** 475–93 (2003).

424. Marra, F. *et al.* Does antibiotic exposure during infancy lead to development of asthma? a systematic review and metaanalysis. *Chest* **129,** 610–8 (2006).

425. Noverr, M. C. & Huffnagle, G. B. The 'microflora hypothesis' of allergic diseases. *Clin. Exp. Allergy* **35,** 1511–20 (2005).

426. Prioult, G. & Nagler-Anderson, C. Mucosal immunity and allergic responses: lack of regulation and/or lack of microbial stimulation? *Immunol. Rev.* **206,** 204–18 (2005).

427. Lode, H., Von der Hoh, N., Ziege, S., Borner, K. & Nord, C. E. Ecological effects of linezolid versus amoxicillin/clavulanic acid on the normal intestinal microflora. *Scand. J. Infect. Dis.* **33,** 899–903 (2001).

428. Donskey, C. J. *et al.* Use of denaturing gradient gel electrophoresis for analysis of the stool microbiota of hospitalized patients. *J. Microbiol. Methods* **54,** 249–56 (2003).

429. Jernberg, C., Löfmark, S., Edlund, C. & Jansson, J. K. Long-term ecological impacts of antibiotic administration on the human intestinal microbiota. *ISME J.* **1,** 56–66 (2007).

430. Löfmark, S., Jernberg, C., Billström, H., Andersson, D. I. & Edlund, C. Restored fitness leads to long-term persistence of resistant Bacteroides strains in the human intestine. *Anaerobe* **14,** 157–60 (2008).

431. Sjölund, M., Wreiber, K., Andersson, D. I., Blaser, M. J. & Engstrand, L. Long-term persistence of resistant Enterococcus species after antibiotics to eradicate Helicobacter pylori. *Ann. Intern. Med.* **139,** 483–7 (2003).

432. Kim, S., Covington, A. & Pamer, E. G. The intestinal microbiota: Antibiotics, colonization resistance, and enteric pathogens. *Immunol. Rev.* **279,** 90–105 (2017).

433. Gorham, P., Millar, M. & Godwin, P. G. Clostridial hand-carriage and neonatal necrotising enterocolitis. *J. Hosp. Infect.* **12,** 139–41 (1988).

434. Waligora-Dupriet, A. J., Dugay, A., Auzeil, N., Huerre, M. & Butel, M. J. Evidence for clostridial implication in necrotizing enterocolitis through bacterial fermentation in a gnotobiotic quail model. *Pediatr. Res.* **58,** 629–635 (2005).

435. Lin, H.-C. *et al.* Oral probiotics reduce the incidence and severity of necrotizing enterocolitis in very low birth weight infants. *Pediatrics* **115,** 1–4 (2005).

436. Hoyos, A. B. Reduced incidence of necrotizing enterocolitis associated with enteral administration of Lactobacillus acidophilus and Bifidobacterium infantis to neonates in an intensive care unit. *Int. J. Infect. Dis.* **3,** 197–202 (1999).

437. Sántulli, T. V *et al.* Acute necrotizing enterocolitis in infancy: a review of 64 cases. *Pediatrics* **55,** 376–87 (1975).

438. Schullinger, J. N., Mollitt, D. L., Vinocur, C. D., Santulli, T. V & Driscoll, J. M. J. Neonatal necrotizing enterocolitis. Survival, management, and complications: a 25-year study. *Am. J. Dis. Child.* **135,** 612–614 (1981).

439. Cordero, L., Rau, R., Taylor, D. & Ayers, L. W. Enteric gram-negative bacilli bloodstream infections: 17 Years' experience in a neonatal intensive care unit. *Am. J. Infect. Control* **32,** 189–195 (2004).

440. Palmer, S. R., Biffin, A. & Gamsu, H. R. Outcome of neonatal necrotising enterocolitis: results of the BAPM/CDSC surveillance study, 1981-84. *Arch. Dis. Child.* **64,** 388–94 (1989).

441. Noel, G. J., Laufer, D. A. & Edelson, P. J. Anaerobic bacteremia in a neonatal intensive care unit: an eighteen-year experience. *Pediatr. Infect. Dis. J.* **7,** 858–862 (1988).

442. Sharma, R. *et al.* Neonatal gut barrier and multiple organ failure: role of endotoxin and proinflammatory cytokines in sepsis and necrotizing enterocolitis. *J. Pediatr. Surg.* **42,** 454–461 (2007).

443. Scheifele, D. W., Olsen, E. M. & Pendray, M. R. Endotoxinemia and thrombocytopenia during neonatal necrotizing enterocolitis. *Am. J. Clin. Pathol.* **83,** 227–229 (1985).

444. Stewart, C. J. *et al.* Development of the preterm gut microbiome in twins at risk of necrotising enterocolitis and sepsis. *PLoS One* **8,** e73465 (2013).

445. Holt, P. G. & Jones, C. a. The development of the immune system during pregnancy and early life. *Allergy* **55,** 688–97 (2000).

446. Sudo, N. *et al.* The requirement of intestinal bacterial flora for the development of an IgE production system fully susceptible to oral tolerance induction. *J. Immunol.* **159,** 1739–1745 (1997).

447. Deshpande, G. C., Rao, S. C., Keil, A. D. & Patole, S. K. Evidence-based guidelines for use of probiotics in preterm neonates. *BMC Med.* **9,** 92 (2011).

448. Ohashi, Y. & Ushida, K. Health-beneficial effects of probiotics: Its mode of action. *Anim. Sci. J.* **80,** 361–371 (2009).

449. Lin, H. C. *et al.* Oral Probiotics Prevent Necrotizing Enterocolitis in Very Low Birth Weight Preterm Infants: A Multicenter, Randomized, Controlled Trial. *Pediatrics* **122,** 693–700 (2008).

450. Kleessen, B., Bunke, H., Tovar, K., Noack, J. & Sawatzki, G. Influence of two infant formulas and human milk on the development of the faecal flora in newborn infants. *Acta Paediatr.* **84,** 1347–56 (1995).

451. Cheung, Y. F., Fung, C. H. & Walsh, C. Stereochemistry of propionyl-coenzyme A and pyruvate carboxylations catalyzed by transcarboxylase. *Biochemistry* **14,** 2981–2986 (1975).

452. Costeloe, K. *et al.* Bifidobacterium breve BBG-001 in very preterm infants: a randomised controlled phase 3 trial. *Lancet (London, England)* **387,** 649–60 (2016).

453. Sanders, H. L. Marine Benthic Diversity: A Comparative Study. *Am. Nat.* **102,** 243–282 (1968).

454. McMurdie, P. J. & Holmes, S. Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible. *PLoS Comput. Biol.* **10,** (2014).

455. De Filippis, F., Pellegrini, N., Laghi, L., Gobbetti, M. & Ercolini, D. Unusual sub-genus associations of faecal Prevotella and Bacteroides with specific dietary patterns. *Microbiome* **4,** 57 (2016).

456. O'Callaghan, A. & van Sinderen, D. Bifidobacteria and their role as members of the human gut microbiota. *Front. Microbiol.* **7,** (2016).

457. Teitelbaum, J. E. & Walker, W. A. Nutritional impact of pre- and probiotics as protective gastrointestinal organisms. *Annu. Rev. Nutr.* **22,** 107–38 (2002).

458. Leahy, S. C., Higgins, D. G., Fitzgerald, G. F. & Van Sinderen, D. Getting better with

bifidobacteria. *J. Appl. Microbiol.* **98,** 1303–1315 (2005).

459. Milani, C. *et al.* Bifidobacteria exhibit social behavior through carbohydrate resource sharing in the gut. *Sci. Rep.* **5,** 15782 (2015).

460. Downes, J. Dialister invisus sp. nov., isolated from the human oral cavity. *Int. J. Syst. Evol. Microbiol.* **53,** 1937–1940 (2003).

461. Morotomi, M., Nagai, F., Sakon, H. & Tanaka, R. Dialister succinatiphilus sp. nov. and Barnesiella intestinihominis sp. nov., isolated from human faeces. *Int. J. Syst. Evol. Microbiol.* **58,** 2716–2720 (2008).

462. Le Chatelier, E. *et al.* Richness of human gut microbiome correlates with metabolic markers. *Nature* **500,** 541–6 (2013).

463. Swidsinski, A., Weber, J., Loening-baucke, V., Hale, L. P. & Lochs, H. Spatial Organization and Composition of the Mucosal Flora in Patients with Inflammatory Bowel Disease Spatial Organization and Composition of the Mucosal Flora in Patients with Inflammatory Bowel Disease. *J. Clin. Microbiol.* **43,** 3380–3389 (2005).

464. Pantoja-Feliciano, I. G. *et al.* Biphasic assembly of the murine intestinal microbiota during early development. *ISME J.* **7,** 1112–1115 (2013).

465. Hunt, K. M. *et al.* Characterization of the diversity and temporal stability of bacterial communities in human milk. *PLoS One* **6,** 1–8 (2011).

466. Martín, R., Heilig, G. H. J., Zoetendal, E. G., Smidt, H. & Rodríguez, J. M. Diversity of the Lactobacillus group in breast milk and vagina of healthy women and potential role in the colonization of the infant gut. *J. Appl. Microbiol.* **103,** 2638–2644 (2007).

467. Solís, G., de los Reyes-Gavilan, C. G., Fernández, N., Margolles, A. & Gueimonde, M. Establishment and development of lactic acid bacteria and bifidobacteria microbiota in breast-milk and the infant gut. *Anaerobe* **16,** 307–310 (2010).

468. Zaura, E., Keijser, B. J., Huse, S. M. & Crielaard, W. Defining the healthy 'core microbiome' of oral microbial communities. *BMC Microbiol.* **9,** 259 (2009).

469. Segata, N. *et al.* Composition of the adult digestive tract bacterial microbiome based on seven mouth surfaces, tonsils, throat and stool samples. *Genome Biol.* **13,** R42 (2012).

470. Arif, N., Sheehy, E. C., Do, T. & Beighton, D. Diversity of Veillonella spp. from sound and carious sites in children. *J. Dent. Res.* **87,** 278–82 (2008).

471. Ballard, O. & Morrow, A. L. Human milk composition: nutrients and bioactive factors. *Pediatr. Clin. North Am.* **60,** 49–74 (2013).

472. Hosseini, E., Grootaert, C., Verstraete, W. & Van de Wiele, T. Propionate as a health-promoting microbial metabolite in the human gut. *Nutr. Rev.* **69,** 245–258 (2011).

473. Vinolo, M. A. R., Rodrigues, H. G., Nachbar, R. T. & Curi, R. Regulation of inflammation by short chain fatty acids. *Nutrients* **3,** 858–876 (2011).

474. Janssen, P. H. Growth yield increase and ATP formation linked to succinate decarboxylation in Veillonella parvula. *Arch. Microbiol.* **157,** 442–445 (1992).

475. Drzewiecka, D. Significance and Roles of Proteus spp. Bacteria in Natural Environments. *Microb. Ecol.* **72,** 741–758 (2016).

476. Palusiak, A. Immunochemical properties of Proteus penneri lipopolysaccharides - One

of the major Proteus sp. virulence factors Dedicated to Professor Yuriy A. Knirel in recognition of his outstanding contribution to the field. *Carbohydr. Res.* **380,** 16–22 (2013).

477. Różalski, A. *et al.* Proteus sp. – an opportunistic bacterial pathogen – classification, swarming growth, clinical significance and virulence factors. *Folia Biol. Oecologica* **8,** 1–17 (2012).

478. Arbatsky, N. P. *et al.* Structure of a Kdo-containing O polysaccharide representing Proteus O79, a newly described serogroup for some clinical Proteus genomospecies isolates from Poland. *Carbohydr. Res.* **379,** 100–105 (2013).

479. Armbruster, C. E. & Mobley, H. L. T. Merging mythology and morphology: the multifaceted lifestyle of Proteus mirabilis. *Nat. Rev. Microbiol.* **10,** 743–54 (2012).

480. Drzewiecka, D. *et al.* Structural and serological studies of the O-polysaccharide of strains from a newly created Proteus O78 serogroup prevalent in Polish patients. *FEMS Immunol. Med. Microbiol.* **58,** 269–276 (2010).

481. Drzewiecka, D., Zych, K. & Sidorczyk, Z. Characterization and serological classification of a collection of Proteus penneri clinical strains. *Arch. Immunol. Ther. Exp. (Warsz).* **52,** 121–128 (2004).

482. Siwińska, M. *et al.* Classification of a Proteus penneri clinical isolate with a unique O-antigen structure to a new Proteus serogroup, O80. *Carbohydr. Res.* **407,** 131–136 (2015).

483. Wang, Y. *et al.* An outbreak of Proteus mirabilis food poisoning associated with eating stewed pork balls in brown sauce, Beijing. *Food Control* **21,** 302–305 (2010).

484. Muller, H. E. Occurrence and pathogenic role of Morganella-Proteus-Providencia group bacteria in human feces. *J. Clin. Microbiol.* **23,** 404–405 (1986).

485. Muller, H. E. The role of Proteae in diarrhea. *Zentralbl. Bakteriol.* **272,** 30–35 (1989).

486. Peerbooms, P. G., Verweij, A. M., Oe, P. L. & MacLaren, D. M. Urinary pathogenicity of Proteus mirabilis strains isolated from faeces or urine. *Antonie Van Leeuwenhoek* **52,** 53–62 (1986).

487. Senior, B. W. & Leslie, D. L. Rare occurrence of Proteus vulgaris in faeces: a reason for its rare association with urinary tract infections. *J. Med. Microbiol.* **21,** 139–44 (1986).

488. de Louvois, J. Serotyping and the Dienes reaction on Proteus mirabilis from hospital infections. *J. Clin. Pathol.* **22,** 263–268 (1969).

489. Hold, G. L., Pryde, S. E., Russell, V. J., Furrie, E. & Flint, H. J. Assessment of microbial diversity in human colonic samples by 16S rDNA sequence analysis. *FEMS Microbiol. Ecol.* **39,** 33–9 (2002).

490. Frank, D. N. *et al.* Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. *Proc. Natl. Acad. Sci. U. S. A.* **104,** 13780–5 (2007).

491. Manson, J. M., Rauch, M. & Gilmore, M. S. The commensal microbiology of the gastrointestinal tract. *Adv. Exp. Med. Biol.* **635,** 15–28 (2008).

492. Roberts, A. K. *et al.* Supplementation of an adapted formula with bovine lactoferrin: 1. Effect on the infant faecal flora. *Acta Paediatr.* **81,** 119–24 (1992).

493. Collins, M. D. *et al.* The phylogeny of the genus Clostridium: proposal of five new genera and eleven new species combinations. *Int. J. Syst. Bacteriol.* **44,** 812–826 (1994).

494. Rainey, F. a & Janssen, P. H. Phylogenetic analysis by 16S ribosomal DNA sequence comparison reveals two unrelated groups of species within the genus Ruminococcus. *FEMS Microbiol. Lett.* **129,** 69–73 (1995).

495. Nava, G. M., Friedrichsen, H. J. & Stappenbeck, T. S. Spatial organization of intestinal microbiota in the mouse ascending colon. *ISME J.* **5,** 627–638 (2011).

496. Lopetuso, L. R., Scaldaferri, F., Petito, V. & Gasbarrini, A. Commensal Clostridia: leading players in the maintenance of gut homeostasis. *Gut Pathog.* **5,** 23 (2013).

497. Tlaskalová-Hogenová, H. *et al.* Commensal bacteria (normal microflora), mucosal immunity and chronic inflammatory and autoimmune diseases. *Immunol. Lett.* **93,** 97–108 (2004).

498. Pryde, S. E., Duncan, S. H., Hold, G. L., Stewart, C. S. & Flint, H. J. The microbiology of butyrate formation in the human colon. *FEMS Microbiol. Lett.* **217,** 133–139 (2002).

499. Clausen, M. R. & Mortensen, P. B. Kinetic studies on colonocyte metabolism of short chain fatty acids and glucose in ulcerative colitis. *Gut* **37,** 684–689 (1995).

500. Ritzhaupt, A., Ellis, A., Hosie, K. B. & Shirazi-Beechey, S. P. The characterization of butyrate transport across pig and human colonic luminal membrane. *J. Physiol.* **507,** 819–830 (1998).

501. Scheppach, W., Luehrs, H. & Menzel, T. Beneficial health effects of low-digestible carbohydrate consumption. *Br. J. Nutr.* **85 Suppl 1,** S23–S30 (2001).

502. Mortensen, P. B. & Clausen, M. R. Short-chain fatty acids in the human colon: relation to gastrointestinal health and disease. *Scand. J. Gastroenterol. Suppl.* **216,** 132–148 (1996).

503. Csordas, A. Butyrate, aspirin and colorectal cancer. *European Journal of Cancer Prevention* **5,** 221–231 (1996).

504. Segain, J. P. *et al.* Butyrate inhibits inflammatory responses through NFkappaB inhibition: implications for Crohn's disease. *Gut* **47,** 397–403 (2000).

505. Lührs, H. *et al.* Cytokine-activated degradation of inhibitory κB protein α is inhibited by the short-chain fatty acid butyrate. *Int. J. Colorectal Dis.* **16,** 195–201 (2001).

506. Wächtershäuser, A. & Stein, J. Rationale for the luminal provision of butyrate in intestinal diseases. *Eur. J. Nutr.* **39,** 164–171 (2000).

507. Topping, D. L. & Clifton, P. M. Short-chain fatty acids and human colonic function: roles of resistant starch and nonstarch polysaccharides. *Physiol Rev* **81,** 1031–1064 (2001).

508. Duncan, S. H., Barcenilla, A., Stewart, C. S., Pryde, S. E. & Flint, H. J. Acetate utilization and butyryl coenzyme A (CoA): Acetate-CoA transferase in butyrate-producing bacteria from the human large intestine. *Appl. Environ. Microbiol.* **68,** 5186–5190 (2002).

509. Umesaki, Y., Setoyama, H., Matsumoto, S., Imaoka, A. & Itoh, K. Differential roles of segmented filamentous bacteria and clostridia in development of the intestinal immune

system. *Infect. Immun.* **67,** 3504–11 (1999).

510. Lefrancois, L. & Goodman, T. In vivo modulation of cytolytic activity and Thy-1 expression in TCR-gamma delta+ intraepithelial lymphocytes. *Science* **243,** 1716–8 (1989).

511. Umesaki, Y., Okada, Y., Matsumoto, S., Imaoka, A. & Setoyama, H. Segmented Filamentous Bacteria Are Indigenous Intestinal Bacteria That Activate Intraepithelial Lymphocytes and Induce MHC Class II Molecules and Fucosyl Asialo GM1 Glycolipids on the Small Intestinal Epithelial Cells in the Ex-Germ-Free Mouse. *Microbiol. Immunol.* **39,** 555–562 (1995).

512. Atarashi, K. *et al.* Induction of colonic regulatory T cells by indigenous Clostridium species. *Science* **331,** 337–41 (2011).

513. Feuerer, M. *et al.* Genomic definition of multiple ex vivo regulatory T cell subphenotypes. *Proc Natl Acad Sci U S A* **107,** 5919–5924 (2010).

514. Dillon, H. C. J., Gray, E., Pass, M. A. & Gray, B. M. Anorectal and vaginal carriage of group B streptococci during pregnancy. *J. Infect. Dis.* **145,** 794–799 (1982).

515. Hickman, M. E., Rench, M. a, Ferrieri, P. & Baker, C. J. Changing epidemiology of group B streptococcal colonization. *Pediatrics* **104,** 203–209 (1999).

516. Manning, S. D. *et al.* Prevalence of group B streptococcus colonization and potential for transmission by casual contact in healthy young men and women. *Clin. Infect. Dis.* **39,** 380–388 (2004).

517. Brimil, N. *et al.* Epidemiology of Streptococcus agalactiae colonization in Germany. *Int. J. Med. Microbiol.* **296,** 39–44 (2006).

518. Schuchat, A. & Wenger, J. D. Epidemiology of group B streptococcal disease. Risk factors, prevention strategies, and vaccine development. *Epidemiol. Rev.* **16,** 374–402 (1994).

519. Boyer, K. M. & Gotoff, S. P. Strategies for chemoprophylaxis of GBS early-onset infections. *Antibiot. Chemother.* **35,** 267–280 (1985).

520. Mountzouris, K. C., McCartney, A. L. & Gibson, G. R. Intestinal microflora of human infants and current trends for its nutritional modulation. *Br J Nutr* **87,** 405–420 (2002).

521. Ferraris, L., Butel, M.-J. & Aires, J. Antimicrobial susceptibility and resistance determinants of Clostridium butyricum isolates from preterm infants. *Int. J. Antimicrob. Agents* **36,** 420–423 (2010).

522. Cassir, N., Benamar, S. & La Scola, B. Clostridium butyricum: From beneficial to a new emerging pathogen. *Clin. Microbiol. Infect.* **22,** 37–45 (2016).

523. Fenicia, L., Anniballi, F. & Aureli, P. Intestinal toxemia botulism in Italy, 1984-2005. *Eur. J. Clin. Microbiol. Infect. Dis.* **26,** 385–394 (2007).

524. Pickett, J., Berg, B., Chaplin, E. & Brunstetter-Shafer, M. A. Syndrome of botulism in infancy: clinical and electrophysiologic study. *N. Engl. J. Med.* **295,** 770–2 (1976).

525. Dykes, J. K., Lúquez, C., Raphael, B. H., McCroskey, L. & Maslanka, S. E. Laboratory investigation of the first case of botulism caused by clostridium butyricum type e toxin in the United States. *J. Clin. Microbiol.* **53,** 3363–3365 (2015).

526. Hill, K. K. *et al.* Recombination and insertion events involving the botulinum neurotoxin complex genes in Clostridium botulinum types A, B, E and F and

Clostridium butyricum type E strains. *BMC Biol.* **7,** 66 (2009).

527. Cassir, N. *et al.* Clostridium butyricum Strains and Dysbiosis Linked to Necrotizing Enterocolitis in Preterm Neonates. *Clin. Infect. Dis.* **61,** 1107–1115 (2015).

528. Hsu, T., Hutto, D. L., Minion, F. C., Zuerner, R. L. & Wannemuehler, M. J. Cloning of a beta-hemolysin gene of Brachyspira (Serpulina) hyodysenteriae and its expression in Escherichia coli. *Infect. Immun.* **69,** 706–11 (2001).

529. Popoff, M. R., Szylit, O., Ravisse, P., Dabard, J. & Ohayon, H. Experimental cecitis in gnotoxenic chickens monoassociated with Clostridium butyricum strains isolated from patients with neonatal necrotizing enterocolitis. *Infect. Immun.* **47,** 697–703 (1985).

530. Szylit, O. *et al.* An experimental model of necrotising enterocolitis. *Lancet* **350,** 33–34 (1997).

531. Thymann, T. *et al.* Carbohydrate maldigestion induces necrotizing enterocolitis in preterm pigs. *Am. J. Physiol. Gastrointest. Liver Physiol.* **297,** G1115-25 (2009).

532. Rasiah, V., Yajamanyam, P. K. & Ewer, A. K. Necrotizing enterocolitis: current perspectives. *Res. Reports Neonatol.* 31–42 (2014). doi:10.2147/RRN.S36576

533. Damodaram, M., Story, L., Kulinskaya, E., Rutherford, M. & Kumar, S. Early adverse perinatal complications in preterm growth-restricted fetuses. *Aust. New Zeal. J. Obstet. Gynaecol.* **51,** 204–209 (2011).

534. Yee, W. H. *et al.* Incidence and timing of presentation of necrotizing enterocolitis in preterm infants. *Pediatrics* **129,** e298-304 (2012).

535. Tedjo, D. I. *et al.* The effect of sampling and storage on the fecal microbiota composition in healthy and diseased subjects. *PLoS One* **10,** e0126685 (2015).

536. McGeachie, M. J. *et al.* Longitudinal Prediction of the Infant Gut Microbiome with Dynamic Bayesian Networks. *Sci. Rep.* **6,** 20359 (2016).

537. Gomes, B. P. F. A., Berber, V. B., Kokaras, A. S., Chen, T. & Paster, B. J. Microbiomes of Endodontic-Periodontal Lesions before and after Chemomechanical Preparation. *J. Endod.* **41,** 1975–84 (2015).

538. Murad, C. F. *et al.* Microbial diversity in persistent root canal infections investigated by checkerboard DNA-DNA hybridization. *J. Endod.* **40,** 899–906 (2014).

539. Rôças, I. N. *et al.* Advanced Caries Microbiota in Teeth with Irreversible Pulpitis. *J. Endod.* **41,** 1450–5 (2015).

# 9 Figure Index

# 10 Table Index

# 11 Supplementary Tables

Supp. Table 1 The bacterial species detected in amniotic fluid by both culture and DNA based methods; * Culture in findings reported as 'mixed anaerobes' are not listed. ** Per culture, one case was positive for *Prevotella melaniogenica*, † in 4 cases, *Strep. agalactiae* was identified by DNA when culture detected Group B Streptococcus, ‡ in 2 cases, culture identified *Fusobacterium spp.*, while in two other cases, both culture and DNA based methods detected *Fusobacterium nucleatum*. §in two cases, *Ureaplasma urealyticum* was identified by culture only, while DNA analysis resulted in U. parvum. ¶ in two cases, G. vaginalis was identified by culture only, while DNA analysis resulted in *Sneathia sanguinegens*, †† in two cases, *Shigella* was identified by culture, while DNA analysis resulted in E. coli

| Species | Detected by DiGiulio[305] | Species* | Detected by Han[306] |
|---|---|---|---|
| Gram-positive Firmicutes | | | |
| *Streptococcus mitis* | DNA only | | |
| *Streptococcus agalactiae* | Culture and DNA | *Strep. Agalactiae* | DNA only |
| | | Streptococcus Group B | DNA only |
| Lactobacillus sp. | Culture and DNA | | |
| Bacillus sp. (not anthracis) | Culture only | | |
| Coagulase-negative Staphylococcus sp. | Culture only | | |
| Peptostreptococcus sp. | Culture only | Peptostreptococcus sp. | |
| | | *Clostridiales bacterium* | |
| *Peptostreptococcus asaccharolyticus* | Culture Onl | | |
| Fusobacteria | | | |
| *Fusobacterium nucleatum* | Culture and DNA, culture only | *Fus. nucleatum* | DNA only, culture‡ |
| Uncultivated Fusobacterium | DNA only | | |
| *Sneathia* (formerly Leptotrichia) *sanguinegens* | DNA only | *Sneathia* (formerly Leptotrichia) *sanguinegens* | DNA only |
| *Leptotrichia amnionii* | DNA only | *Leptotrichia amnionii* | DNA only |
| Tenericutes | | | |
| *Mycoplasma hominis* | Culture and DNA | *Myc. hominis* | Culture only |
| *Ureaplasma urealyticum* | Culture only | *U. urealyticum* | Culture only§) |
| Ureaplasma sp. | DNA only | *Ureaplasma parvum* | DNA only§ |
| Actinobacteria | | | |
| *Gardnerella vaginalis* | Culture only | | |
| Gram-negative Bacterioidetes | | | |
| Uncultivated Bacteroidetes bacterium | DNA Only | *Bacteroides ureolyticus* | DNA only, DNA and culture |
| | | *Bacteroides fragilis* | DNA only |
| Prevotella sp. | DNA onle, NDA and Culture** | *Prevotella bivia* | Culture only, culture and DNA |

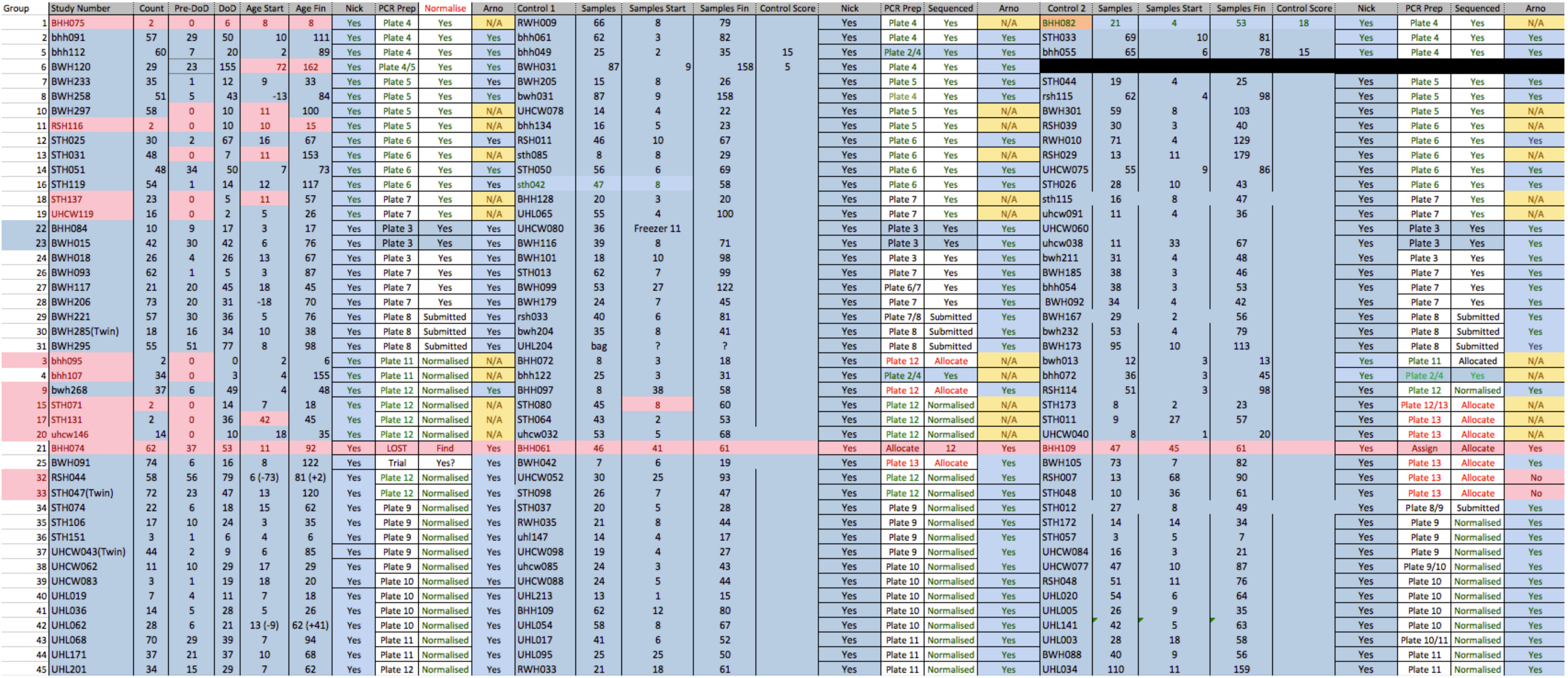
Supp. Table 1 continued

| Species | Detected by DiGiulio | Species* | Detected by Han |
|---|---|---|---|
| Gram-negative Proteobacteria | | Bergeyella sp. | DNA only |
| *Delftia acidovorans* | DNA Only | | |
| *Neisseria cinerea* | DNA Only | | |
| | | *Citrobacter koseri* | Culture and DNA |
| | | *Klebsiella pneumoniae* | Culture only |
| | | Shigella spp. | DNA only†† |
| | | *Escherichia coli* | Culture only†† |
| | | *Eikenella corrodens* | Culture only |

Supp. Table 2 Species which the V4 region of the 16S rRNA subunit universal primers are difficult to identify accurately

| Species |
|---|
| *Chlorobium phaeovibrioides* DSM 265 |
| Desulfovibriodesulfuricans ATCC 27774 |
| *Rhodospirillum rubrum* ATCC 11170 |
| *Salinispora arenicola* CNS-205 |
| *Shewanella baltica* OS185 |
| Sulfitobacter sp. NAS-14.1 |
| *Sulfolobus tokodaii* 7(S311) |
| Sulfurihydrogenibium yellowstonense SS-5 |
| *Thermotoga neapolitana* DSM 4359 |
| *Acidobacterium capsulatum* ATCC 51196 |
| *Burkholderia xenovorans* LB400 |
| *Dictyoglomus turgidum* DSM 6724 |
| *Gemmatimonas aurantiaca* T-27T |
| *Nitrosomonas europaea* ATCC 19718 |
| *Thermotoga petrophila* RKU-1 |
| *Treponema denticola* ATCC 35405 |
| *Zymomonasmobilis mobilis* ZM4 |

Supp. Table 3 The original 45 gold standard confirmed NEC subjects and their associated control selection. This was reduced to 42 NEC subjects and 81 controls after sample assessment and control selection



| Group | Study Number | Count | Pre-DoD | DoD | Age Start | Age Fin | Nick | PCR Prep | Normalise | Arno | Control 1 | Samples | Samples Start | Samples Fin | Control Score | Nick | PCR Prep | Sequenced | Arno | Control 2 | Samples | Samples Start | Samples Fin | Control Score | Nick | PCR Prep | Sequenced | Arno |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | BHH075 | 2 | 0 | 6 | 8 | 8 | Yes | Plate 4 | Yes | N/A | RWH009 | 66 | 8 | 79 |  | Yes | Plate 4 | Yes | N/A | BHH082 | 21 | 4 | 53 | 18 | Yes | Plate 4 | Yes | N/A |
| 2 | bhh091 | 57 | 29 | 50 | 10 | 111 | Yes | Plate 4 | Yes | Yes | bhh061 | 62 | 3 | 82 |  | Yes | Plate 4 | Yes | Yes | STH033 | 69 | 10 | 81 |  | Yes | Plate 4 | Yes | Yes |
| 5 | bhh112 | 60 | 7 | 20 | 2 | 89 | Yes | Plate 4 | Yes | Yes | bhh049 | 25 | 2 | 35 | 15 | Yes | Plate 2/4 | Yes | Yes | bhh055 | 65 | 6 | 78 | 15 | Yes | Plate 4 | Yes | Yes |
| 6 | BWH120 | 29 | 23 | 155 | 72 | 162 | Yes | Plate 4/5 | Yes | Yes | BWH031 | 87 | 9 | 158 | 5 | Yes | Plate 4 | Yes | Yes | [redacted] | [redacted] | [redacted] | [redacted] | [redacted] | [redacted] | [redacted] | [redacted] | [redacted] |
| 7 | BWH233 | 35 | 1 | 12 | 9 | 33 | Yes | Plate 5 | Yes | Yes | BWH205 | 15 | 8 | 26 |  | Yes | Plate 5 | Yes | Yes | STH044 | 19 | 4 | 25 |  | Yes | Plate 5 | Yes | Yes |
| 8 | BWH258 | 51 | 5 | 43 | -13 | 84 | Yes | Plate 5 | Yes | Yes | bwh031 | 87 | 9 | 158 |  | Yes | Plate 4 | Yes | Yes | rsh115 | 62 | 4 | 98 |  | Yes | Plate 5 | Yes | Yes |
| 10 | BWH297 | 58 | 0 | 10 | 11 | 100 | Yes | Plate 5 | Yes | N/A | UHCW078 | 14 | 4 | 22 |  | Yes | Plate 5 | Yes | N/A | BWH301 | 59 | 8 | 103 |  | Yes | Plate 5 | Yes | N/A |
| 11 | RSH116 | 2 | 0 | 10 | 10 | 15 | Yes | Plate 5 | Yes | N/A | bhh134 | 16 | 5 | 23 |  | Yes | Plate 5 | Yes | N/A | RSH039 | 30 | 3 | 40 |  | Yes | Plate 6 | Yes | N/A |
| 12 | STH025 | 30 | 2 | 67 | 16 | 67 | Yes | Plate 6 | Yes | Yes | RSH011 | 46 | 10 | 67 |  | Yes | Plate 6 | Yes | Yes | RWH010 | 71 | 4 | 129 |  | Yes | Plate 6 | Yes | Yes |
| 13 | STH031 | 48 | 0 | 7 | 11 | 153 | Yes | Plate 6 | Yes | N/A | sth085 | 8 | 8 | 29 |  | Yes | Plate 6 | Yes | N/A | RSH029 | 13 | 11 | 179 |  | Yes | Plate 6 | Yes | N/A |
| 14 | STH051 | 48 | 34 | 50 | 7 | 73 | Yes | Plate 6 | Yes | Yes | STH050 | 56 | 6 | 69 |  | Yes | Plate 6 | Yes | Yes | UHCW075 | 55 | 9 | 86 |  | Yes | Plate 6 | Yes | Yes |
| 16 | STH119 | 54 | 1 | 14 | 12 | 117 | Yes | Plate 6 | Yes | Yes | sth042 | 47 | 8 | 58 |  | Yes | Plate 6 | Yes | Yes | STH026 | 28 | 10 | 43 |  | Yes | Plate 6 | Yes | Yes |
| 18 | STH137 | 23 | 0 | 5 | 11 | 57 | Yes | Plate 7 | Yes | N/A | BHH128 | 20 | 3 | 20 |  | Yes | Plate 7 | Yes | N/A | sth115 | 16 | 8 | 47 |  | Yes | Plate 7 | Yes | N/A |
| 19 | UHCW119 | 16 | 0 | 2 | 5 | 26 | Yes | Plate 7 | Yes | N/A | UHL065 | 55 | 4 | 100 |  | Yes | Plate 7 | Yes | N/A | uhcw091 | 11 | 4 | 36 |  | Yes | Plate 7 | Yes | N/A |
| 22 | BHH084 | 10 | 9 | 17 | 3 | 17 | Yes | Plate 3 | Yes | Yes | UHCW080 | 36 | Freezer 11 |  |  | Yes | Plate 3 | Yes | Yes | UHCW060 |  |  |  |  | Yes | Plate 3 | Yes | Yes |
| 23 | BWH015 | 42 | 30 | 42 | 6 | 76 | Yes | Plate 3 | Yes | Yes | BWH116 | 39 | 8 | 71 |  | Yes | Plate 3 | Yes | Yes | uhcw038 | 11 | 33 | 67 |  | Yes | Plate 3 | Yes | Yes |
| 24 | BWH018 | 26 | 4 | 26 | 13 | 67 | Yes | Plate 3 | Yes | Yes | BWH101 | 18 | 10 | 98 |  | Yes | Plate 3 | Yes | Yes | bwh211 | 31 | 4 | 48 |  | Yes | Plate 3 | Yes | Yes |
| 26 | BWH093 | 62 | 1 | 5 | 3 | 87 | Yes | Plate 7 | Yes | Yes | STH013 | 62 | 7 | 99 |  | Yes | Plate 7 | Yes | Yes | BWH185 | 38 | 3 | 46 |  | Yes | Plate 7 | Yes | Yes |
| 27 | BWH117 | 21 | 20 | 45 | 18 | 45 | Yes | Plate 7 | Yes | Yes | BWH099 | 53 | 27 | 122 |  | Yes | Plate 6/7 | Yes | Yes | bhh054 | 38 | 3 | 53 |  | Yes | Plate 7 | Yes | Yes |
| 28 | BWH206 | 73 | 20 | 31 | -18 | 70 | Yes | Plate 7 | Yes | Yes | BWH179 | 24 | 7 | 45 |  | Yes | Plate 7 | Yes | Yes | BWH092 | 34 | 4 | 42 |  | Yes | Plate 7 | Yes | Yes |
| 29 | BWH221 | 57 | 30 | 36 | 5 | 76 | Yes | Plate 8 | Submitted | Yes | rsh033 | 40 | 6 | 81 |  | Yes | Plate 7/8 | Submitted | Yes | BWH167 | 29 | 2 | 56 |  | Yes | Plate 8 | Submitted | Yes |
| 30 | BWH285(Twin) | 18 | 16 | 34 | 10 | 38 | Yes | Plate 8 | Submitted | Yes | bwh204 | 35 | 8 | 41 |  | Yes | Plate 8 | Submitted | Yes | bwh232 | 53 | 4 | 79 |  | Yes | Plate 8 | Submitted | Yes |
| 31 | BWH295 | 55 | 51 | 77 | 8 | 98 | Yes | Plate 8 | Submitted | Yes | UHL204 | bag | ? | ? |  | Yes | Plate 8 | Submitted | Yes | BWH173 | 95 | 10 | 113 |  | Yes | Plate 8 | Submitted | Yes |
| 3 | bhh095 | 2 | 0 | 0 | 2 | 6 | Yes | Plate 11 | Normalised | N/A | BHH072 | 8 | 3 | 18 |  | Yes | Plate 12 | Allocate | N/A | bwh013 | 12 | 3 | 13 |  | Yes | Plate 11 | Allocated | N/A |
| 4 | bhh107 | 34 | 0 | 3 | 4 | 155 | Yes | Plate 11 | Normalised | N/A | bhh122 | 25 | 3 | 31 |  | Yes | Plate 2/4 | Yes | N/A | bhh072 | 36 | 3 | 45 |  | Yes | Plate 2/4 | Yes | N/A |
| 9 | bwh268 | 37 | 6 | 49 | 4 | 48 | Yes | Plate 12 | Normalised | Yes | BHH097 | 8 | 38 | 58 |  | Yes | Plate 12 | Allocate | Yes | RSH114 | 51 | 3 | 98 |  | Yes | Plate 12 | Normalised | Yes |
| 15 | STH071 | 2 | 0 | 14 | 7 | 18 | Yes | Plate 12 | Normalised | N/A | STH080 | 45 | 8 | 60 |  | Yes | Plate 12 | Normalised | N/A | STH173 | 8 | 2 | 23 |  | Yes | Plate 12/13 | Allocate | N/A |
| 17 | STH131 | 2 | 0 | 36 | 42 | 45 | Yes | Plate 12 | Normalised | N/A | STH064 | 43 | 2 | 53 |  | Yes | Plate 12 | Normalised | N/A | STH011 | 9 | 27 | 57 |  | Yes | Plate 13 | Allocate | N/A |
| 20 | uhcw146 | 14 | 0 | 10 | 18 | 35 | Yes | Plate 12 | Normalised | N/A | uhcw032 | 53 | 5 | 68 |  | Yes | Plate 12 | Normalised | N/A | UHCW040 | 8 | 1 | 20 |  | Yes | Plate 13 | Allocate | N/A |
| 21 | BHH074 | 62 | 37 | 53 | 11 | 92 | Yes | LOST | Find | Yes | BHH061 | 46 | 41 | 61 |  | Yes | Allocate | 12 | Yes | BHH109 | 47 | 45 | 61 |  | Yes | Assign | Allocate | Yes |
| 25 | BWH091 | 74 | 6 | 16 | 8 | 122 | Yes | Trial | Yes? | Yes | BWH042 | 7 | 6 | 19 |  | Yes | Plate 13 | Allocate | Yes | BWH105 | 73 | 7 | 82 |  | Yes | Plate 13 | Allocate | Yes |
| 32 | RSH044 | 58 | 56 | 79 | 6 (-73) | 81 (+2) | Yes | Plate 12 | Normalised | Yes | UHCW052 | 30 | 25 | 93 |  | Yes | Plate 12 | Normalised | Yes | RSH007 | 13 | 68 | 90 |  | Yes | Plate 13 | Allocate | No |
| 33 | STH047(Twin) | 72 | 23 | 47 | 13 | 120 | Yes | Plate 12 | Normalised | Yes | STH098 | 26 | 7 | 47 |  | Yes | Plate 12 | Normalised | Yes | STH048 | 10 | 36 | 61 |  | Yes | Plate 13 | Allocate | No |
| 34 | STH074 | 22 | 6 | 18 | 15 | 62 | Yes | Plate 9 | Normalised | Yes | STH037 | 20 | 5 | 28 |  | Yes | Plate 9 | Normalised | Yes | STH012 | 27 | 8 | 49 |  | Yes | Plate 8/9 | Submitted | Yes |
| 35 | STH106 | 17 | 10 | 24 | 3 | 35 | Yes | Plate 9 | Normalised | Yes | RWH035 | 21 | 8 | 44 |  | Yes | Plate 9 | Normalised | Yes | STH172 | 14 | 14 | 34 |  | Yes | Plate 9 | Normalised | Yes |
| 36 | STH151 | 3 | 1 | 6 | 4 | 6 | Yes | Plate 9 | Normalised | Yes | uhl147 | 14 | 4 | 17 |  | Yes | Plate 9 | Normalised | Yes | STH057 | 3 | 5 | 7 |  | Yes | Plate 9 | Normalised | Yes |
| 37 | UHCW043(Twin) | 44 | 2 | 9 | 6 | 85 | Yes | Plate 9 | Normalised | Yes | UHCW098 | 19 | 4 | 27 |  | Yes | Plate 9 | Normalised | Yes | UHCW084 | 16 | 3 | 21 |  | Yes | Plate 9 | Normalised | Yes |
| 38 | UHCW062 | 11 | 10 | 29 | 17 | 29 | Yes | Plate 9 | Normalised | Yes | uhcw085 | 24 | 3 | 43 |  | Yes | Plate 10 | Normalised | Yes | UHCW077 | 47 | 10 | 87 |  | Yes | Plate 9/10 | Normalised | Yes |
| 39 | UHCW083 | 3 | 1 | 19 | 18 | 20 | Yes | Plate 10 | Normalised | Yes | UHCW088 | 24 | 5 | 44 |  | Yes | Plate 10 | Normalised | Yes | RSH048 | 51 | 11 | 76 |  | Yes | Plate 10 | Normalised | Yes |
| 40 | UHL019 | 7 | 4 | 11 | 7 | 18 | Yes | Plate 10 | Normalised | Yes | UHL213 | 13 | 1 | 15 |  | Yes | Plate 10 | Normalised | Yes | UHL020 | 54 | 6 | 64 |  | Yes | Plate 10 | Normalised | Yes |
| 41 | UHL036 | 14 | 5 | 28 | 5 | 26 | Yes | Plate 10 | Normalised | Yes | BHH109 | 62 | 12 | 80 |  | Yes | Plate 10 | Normalised | Yes | UHL005 | 26 | 9 | 35 |  | Yes | Plate 10 | Normalised | Yes |
| 42 | UHL062 | 28 | 6 | 21 | 13 (-9) | 62 (+41) | Yes | Plate 10 | Normalised | Yes | UHL054 | 58 | 8 | 67 |  | Yes | Plate 10 | Normalised | Yes | UHL141 | 42 | 5 | 63 |  | Yes | Plate 10 | Normalised | Yes |
| 43 | UHL068 | 70 | 29 | 39 | 7 | 94 | Yes | Plate 11 | Normalised | Yes | UHL017 | 41 | 6 | 52 |  | Yes | Plate 11 | Normalised | Yes | UHL003 | 28 | 18 | 58 |  | Yes | Plate 10/11 | Normalised | Yes |
| 44 | UHL171 | 37 | 21 | 37 | 10 | 68 | Yes | Plate 11 | Normalised | Yes | UHL095 | 25 | 25 | 50 |  | Yes | Plate 11 | Normalised | Yes | BWH088 | 40 | 9 | 56 |  | Yes | Plate 11 | Normalised | Yes |
| 45 | UHL201 | 34 | 15 | 29 | 7 | 62 | Yes | Plate 12 | Normalised | Yes | RWH033 | 21 | 18 | 61 |  | Yes | Plate 11 | Normalised | Yes | UHL034 | 110 | 11 | 159 |  | Yes | Plate 11 | Normalised | Yes |

Supp. Table 4 16S V4 rDNA Primer Sequences. These sequences were identified by D'Amore et al[362]. as the best performing universal primer of the V1-V9 variables regions that target the 16S subunit

| Name | Sequence |
|---|---|
| 515Fw | CTACACTCTTTCCCTACACGACGCTCTTCCGATCTGTGCCAGCMGCCGCGGTAA |
| 806Rv | GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGGACTACHVGGGTWTCTAAT |

Supp. Table 5 3' and 5' Dual Index barcode primer sequences used for multiplexing samples on the Illumina MiSeq apparatus. 3' and 5' primer sequences are in tables (A) and (B) respectively.

(A)

| Name | 3' adapter | i7 index | pad/linker |
|---|---|---|---|
| DI_N701Rev | CAAGCAGAAGACGGCATACGAGAT | TCGCCTTA | GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT |
| DI_N702Rev | CAAGCAGAAGACGGCATACGAGAT | CTAGTACG | GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT |
| DI_N703Rev | CAAGCAGAAGACGGCATACGAGAT | TTCTGCCT | GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT |
| DI_N704Rev | CAAGCAGAAGACGGCATACGAGAT | GCTCAGGA | GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT |
| DI_N705Rev | CAAGCAGAAGACGGCATACGAGAT | AGGAGTCC | GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT |
| DI_N706Rev | CAAGCAGAAGACGGCATACGAGAT | CATGCCTA | GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT |
| DI_N707Rev | CAAGCAGAAGACGGCATACGAGAT | GTAGAGAG | GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT |
| DI_N708Rev | CAAGCAGAAGACGGCATACGAGAT | CCTCTCTG | GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT |
| DI_N709Rev | CAAGCAGAAGACGGCATACGAGAT | AGCGTAGC | GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT |
| DI_N710Rev | CAAGCAGAAGACGGCATACGAGAT | CAGCCTCG | GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT |
| DI_N711Rev | CAAGCAGAAGACGGCATACGAGAT | TGCCTCTT | GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT |
| DI_N712Rev | CAAGCAGAAGACGGCATACGAGAT | TCCTCTAC | GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT |

[Supp. Table 5 continued]

(B)

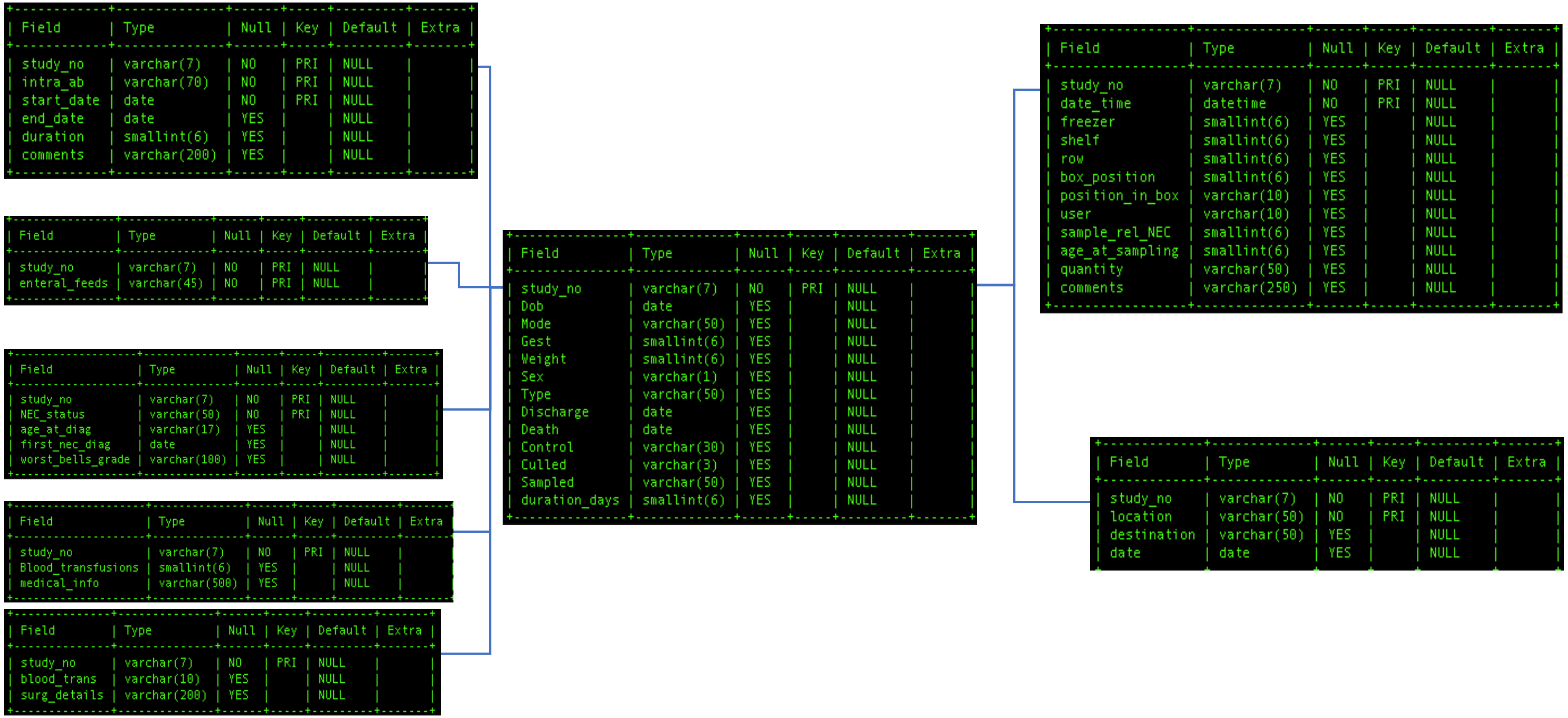
| Name | 5' adapter | i5 index | pad/linker |
|---|---|---|---|
| DI_N501For | ATTGATACGGCGACCACCGAGATCTACAC | TAGATCGC | ACACTCTTTCCCTACACGACG |
| DI_N502For | ATTGATACGGCGACCACCGAGATCTACAC | CTCTCTAT | ACACTCTTTCCCTACACGACG |
| DI_N503For | ATTGATACGGCGACCACCGAGATCTACAC | TATCCTCT | ACACTCTTTCCCTACACGACG |
| DI_N504For | ATTGATACGGCGACCACCGAGATCTACAC | AGAGTAGA | ACACTCTTTCCCTACACGACG |
| DI_N505For | ATTGATACGGCGACCACCGAGATCTACAC | GTAAGGAG | ACACTCTTTCCCTACACGACG |
| DI_N506For | ATTGATACGGCGACCACCGAGATCTACAC | ACTGCATA | ACACTCTTTCCCTACACGACG |
| DI_N507For | ATTGATACGGCGACCACCGAGATCTACAC | AAGGAGTA | ACACTCTTTCCCTACACGACG |
| DI_N508For | ATTGATACGGCGACCACCGAGATCTACAC | CTAAGCCT | ACACTCTTTCCCTACACGACG |

Supp. Table 6 Marascuilo procedural analysis of Chi-squared results for proportions of NEC between NICUs. NICU populations were limited to confirmed and non-NEC infants. No significant difference was observed between the proportion of NEC infants between any two given NICUs.

| Comparison | Difference in Proportions | Critical Value |
|---|---|---|
| BHH-BWH | 0.075 | 0.703 |
| BHH-LWH | 0.056 | 0.702 |
| BHH-RSH | 0.036 | 0.694 |
| BHH-RWH | 0.040 | 0.733 |
| BHH-STH | 0.021 | 0.715 |
| BHH-UHCW | 0.013 | 0.714 |
| BHH-UHL | 0.017 | 0.670 |
| BWH-LWH | 0.131 | 0.670 |
| BWH-RSH | 0.111 | 0.661 |
| BWH-RWH | 0.115 | 0.706 |
| BWH-STH | 0.054 | 0.685 |
| BWH-UHCW | 0.087 | 0.684 |
| BWH-UHL | 0.092 | 0.634 |
| LWH-RSH | 0.021 | 0.660 |
| LWH-RWH | 0.016 | 0.704 |
| LWH-STH | 0.077 | 0.684 |
| LWH-UHCW | 0.044 | 0.683 |
| LWH-UHL | 0.039 | 0.632 |
| RSH-RWH | 0.004 | 0.697 |
| RSH-STH | 0.057 | 0.675 |
| RSH-UHCW | 0.023 | 0.674 |
| RSH-UHL | 0.018 | 0.621 |
| RWH-STH | 0.033 | 0.717 |
| RWH-UHCW | 0.027 | 0.716 |
| RWH-UHL | 0.023 | 0.673 |
| STH-UHCW | 0.033 | 0.697 |
| STH-UHL | 0.038 | 0.649 |
| UHCW-UHL | 0.005 | 0.648 |

# 12 Supplementary Figures

Supp. Figure 1 Schematic view of the MySQL database structure used to maintain and associated patient medical information and sample details.



```
+------------+-------------+------+-----+---------+-------+
| Field      | Type        | Null | Key | Default | Extra |
+------------+-------------+------+-----+---------+-------+
| study_no   | varchar(7)  | NO   | PRI | NULL    |       |
| intra_ab   | varchar(70) | NO   | PRI | NULL    |       |
| start_date | date        | NO   | PRI | NULL    |       |
| end_date   | date        | YES  |     | NULL    |       |
| duration   | smallint(6) | YES  |     | NULL    |       |
| comments   | varchar(200)| YES  |     | NULL    |       |
+------------+-------------+------+-----+---------+-------+
```

```
+--------------+-------------+------+-----+---------+-------+
| Field        | Type        | Null | Key | Default | Extra |
+--------------+-------------+------+-----+---------+-------+
| study_no     | varchar(7)  | NO   | PRI | NULL    |       |
| enteral_feeds| varchar(45) | NO   | PRI | NULL    |       |
+--------------+-------------+------+-----+---------+-------+
```

```
+-------------------+--------------+------+-----+---------+-------+
| Field             | Type         | Null | Key | Default | Extra |
+-------------------+--------------+------+-----+---------+-------+
| study_no          | varchar(7)   | NO   | PRI | NULL    |       |
| NEC_status        | varchar(50)  | NO   | PRI | NULL    |       |
| age_at_diag       | varchar(17)  | YES  |     | NULL    |       |
| first_nec_diag    | date         | YES  |     | NULL    |       |
| worst_bells_grade | varchar(100) | YES  |     | NULL    |       |
+-------------------+--------------+------+-----+---------+-------+
```

```
+--------------------+--------------+------+-----+---------+-------
| Field              | Type         | Null | Key | Default | Extra
+--------------------+--------------+------+-----+---------+-------
| study_no           | varchar(7)   | NO   | PRI | NULL    |
| Blood_transfusions | smallint(6)  | YES  |     | NULL    |
| medical_info       | varchar(500) | YES  |     | NULL    |
+--------------------+--------------+------+-----+---------+-------
```

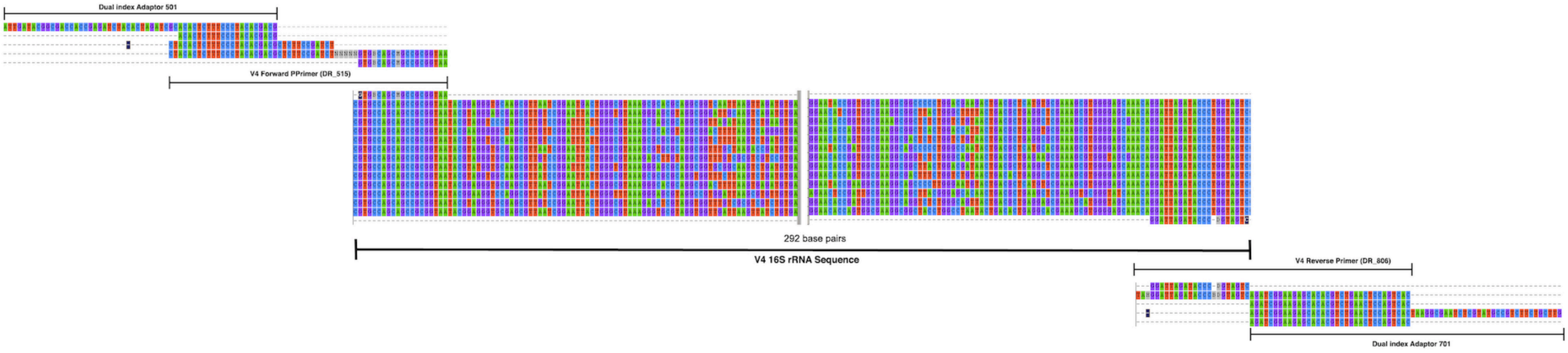
```
+--------------+--------------+------+-----+---------+-------+
| Field        | Type         | Null | Key | Default | Extra |
+--------------+--------------+------+-----+---------+-------+
| study_no     | varchar(7)   | NO   | PRI | NULL    |       |
| blood_trans  | varchar(10)  | YES  |     | NULL    |       |
| surg_details | varchar(200) | YES  |     | NULL    |       |
+--------------+--------------+------+-----+---------+-------+
```

```
+---------------+-------------+------+-----+---------+-------+
| Field         | Type        | Null | Key | Default | Extra |
+---------------+-------------+------+-----+---------+-------+
| study_no      | varchar(7)  | NO   | PRI | NULL    |       |
| Dob           | date        | YES  |     | NULL    |       |
| Mode          | varchar(50) | YES  |     | NULL    |       |
| Gest          | smallint(6) | YES  |     | NULL    |       |
| Weight        | smallint(6) | YES  |     | NULL    |       |
| Sex           | varchar(1)  | YES  |     | NULL    |       |
| Type          | varchar(50) | YES  |     | NULL    |       |
| Discharge     | date        | YES  |     | NULL    |       |
| Death         | date        | YES  |     | NULL    |       |
| Control       | varchar(30) | YES  |     | NULL    |       |
| Culled        | varchar(3)  | YES  |     | NULL    |       |
| Sampled       | varchar(50) | YES  |     | NULL    |       |
| duration_days | smallint(6) | YES  |     | NULL    |       |
+---------------+-------------+------+-----+---------+-------+
```

```
+-----------------+--------------+------+-----+---------+-------+
| Field           | Type         | Null | Key | Default | Extra |
+-----------------+--------------+------+-----+---------+-------+
| study_no        | varchar(7)   | NO   | PRI | NULL    |       |
| date_time       | datetime     | NO   | PRI | NULL    |       |
| freezer         | smallint(6)  | YES  |     | NULL    |       |
| shelf           | smallint(6)  | YES  |     | NULL    |       |
| row             | smallint(6)  | YES  |     | NULL    |       |
| box_position    | smallint(6)  | YES  |     | NULL    |       |
| position_in_box | varchar(10)  | YES  |     | NULL    |       |
| user            | varchar(10)  | YES  |     | NULL    |       |
| sample_rel_NEC  | smallint(6)  | YES  |     | NULL    |       |
| age_at_sampling | smallint(6)  | YES  |     | NULL    |       |
| quantity        | varchar(50)  | YES  |     | NULL    |       |
| comments        | varchar(250) | YES  |     | NULL    |       |
+-----------------+--------------+------+-----+---------+-------+
```

```
+-------------+-------------+------+-----+---------+-------+
| Field       | Type        | Null | Key | Default | Extra |
+-------------+-------------+------+-----+---------+-------+
| study_no    | varchar(7)  | NO   | PRI | NULL    |       |
| location    | varchar(50) | NO   | PRI | NULL    |       |
| destination | varchar(50) | YES  |     | NULL    |       |
| date        | date        | YES  |     | NULL    |       |
```

Supp. Figure 2 Diagram of dual index, nested, V4 PCR. Barcode primers are found on the far outside of the diagram, linked by adaptor sequences to the V4 primers, which in turn bind to the V4 region of the 16S rRNA sequence.



# 13 Supplementary Table Index

# 14 Supplementary Figures Index

# 15 Appendix

## 15.1 Bash Script for Pipeline of Sequence Data QC, Assembly and Error Correction

```
#set default directory:
dir=$(dirname -- $(readlink -fn -- "$0"))
## Creating folder structure
echo Creating directories...
mkdir Analysis
mkdir Assembly
mkdir Assembly/reads
mkdir Assembly/Panda_assembled/
mkdir Assembly/Spades
mkdir Analysis/QC
mkdir Analysis/QIIME
mkdir Assembly/QC
mkdir Analysis/PhiX

##Moving reads from trimmed folder
echo Copying reads to Analysis/reads/ folder...
cp  ./Trimmed/*/*.fastq.gz ./Assembly/reads/

#unzip reads
echo Unzipping reads...
gunzip Assembly/reads/*.gz

#Renaming if they those that have numbers
echo Removing number prefixes
cd Assembly/reads/
for f in *.fastq; do fh=$(ls $f | cut -d '-' -f 2-); mv $f $fh; done;
cd $dir

##fastqc on raw reads
echo Establishing Read Quality...
cat Assembly/reads/*.fastq > Assembly/reads/All_Raw_Reads.fastq
fastqc Assembly/reads/All_Raw_Reads.fastq -o Assembly/QC/
rm Assembly/reads/All_Raw_Reads.fastq

#read counts
echo Calculating read counts...
for f in Assembly/reads/*.fastq; do count=$(grep -c '@' $f); echo  "$f
 $count" >> Assembly/QC/Read_Counts.txt;done;
cut -d '/' -f 3 Assembly/QC/Read_Counts.txt >
 Assembly/QC/Raw_Read_Count.txt #cleanup list
rm Assembly/QC/Read_Counts.txt
```

```
##Bayes Hammer Error Correction
echo Running SPAdes Bayes Hammer error correction only...
cd Assembly/reads/
fastq=$(pwd)
ls *.fastq > files.txt

perl ~/scripts/Perl/yaml.pl files.txt all_samples.yaml ./

python ~/bin/SPAdes-3.0.0-Linux/bin/spades.py \
 --only-error-correction \
 -o ../Spades \
 --dataset all_samples.yaml

echo Changing to Spades directory
cd $dir/Assembly/Spades/corrected/
gunzip *.gz
for fh in ' *R1_001*.fastq'  ; do
          for sample in $fh ; do
                        i=$( echo $sample | cut -d _ -f 1 )

                        BC=$( echo $sample  | cut -d _ -f 2 )

                        for r in 1 2 ; do
echo /Assembly/reads/${i}_*_R${r}_001.fastq
awk -- '{if(ARGIND==1)bc[$1]=$2; else if(ARGIND==2){if($1 in bc)printf "%s
 %s\n",$0,bc[$1]; else printf "%s\n",$0; } }' \
<(paste - - - - <$dir/Assembly/reads/${i}_*_R${r}_001.fastq)
 $dir/Assembly/Spades/corrected/${i}_*_R${r}_001.*.fastq \
>${i}_${BC}_R${r}_Spades_001.fastq
done; done; done;

###Post Spades Trimming Overlap counts
echo Counting contigs left post Spades error correction...
for f in *.fastq; do count=$(grep -c '@' $f); echo  "$f $count" >>
 $dir/Analysis/QC/Spades_Count.txt; done;
cd $dir
echo Calculating quality scores for spades corrected contigs...
fastqc Assembly/Spades/corrected/All_spades.fastq -o $dir/Assembly/QC/
rm Assembly/Spades/corrected/All_spades.fastq

PandaSeq Step
echo Overlapping reads with PandaSeq, min length 240, max length 260...
cd Assembly/Spades/corrected/
for fh in *R1*.fastq ; do
          for sample in $fh ; do
            i=$( echo $sample | cut -d _ -f 1 )
            echo Sample $i processing ;

            BC=$( echo $sample  | cut -d _ -f 2 )
            echo Sample Barcode =  $BC ;

            pandaseq \
            -f ${i}_${BC}_R1*.fastq \
            -r ${i}_${BC}_R2*.fastq \
            -F \
            -T 24 \
            > ../../Panda_assembled/${i}_${BC}_spades_panda_assembly.fastq
      done;
done;

cd $dir/Assembly/Panda_assembled/

Panda Assembly Overlap counts
echo Counting Panda assemblies, check Assembled/QC/Panda_Overlap_Counts.txt
 for details.
for f in *.fastq; do
 count=$(grep -c '@' $f); echo  "$f $count" \
 >> ../QC/Panda_Overlap_Counts.txt;
done;

cd $dir

echo Establishing Panda Assembly quality...
cat Assembly/Panda_assembled/*.fastq >
 Assembly/Panda_assembled/All_Panda_overlaps.fastq
fastqc Assembly/Panda_assembled/All_Panda_overlaps.fastq -o Assembly/QC/
rm Assembly/Panda_assembled/All_Panda_overlaps.fastq

#Converting to FASTA
echo Converting Spades assembly from FASTQ to FASTA and moving to
 Analysis/Assembly/Spades_fasta...
mkdir Analysis/Spades_fasta
cd Assembly/Panda_assembled/
for f in *.fastq; do
 s=$(ls $f | cut -d '_' -f 1 );
 bc=$(ls $f | cut -d '_' -f 2);
 fastq2fasta < $f > $dir/Analysis/Spades_fasta/$s'_'$bc'_assembled.fasta';
done;
cd $dir

# Blast against PhiX DB
cp Analysis/Spades_fasta/*.fasta Analysis/PhiX
cd Analysis/PhiX/
for f in *.fasta; do
 i=$(ls $f | cut -d '_' -f 1 );
 BC=$(ls $f | cut -d '_' -f 2);

 formatdb -i "$i"_"$BC"_assembled.fasta -p F;
 blastall -p blastn -d "$i"_"$BC"_assembled.fasta -i
 /pub40/nfellaby/db/phiX.fa -f 2 | sort | uniq > "$i".phiX.list;
 cat "$i"_"$BC"_assembled.fasta | \
 perl -ne '$_=~s/^(>.+?)\n/$1\t/; print $_;' | \
 grep -vf "$i".phiX.list | \
 perl -ne '$_=~s/\t/\n/; print $_;' \
 > "$i"_"$BC"_phiXfiltered.fa;
done;

rm Analysis/PhiX/*assembled.fasta

cd $dir

cd Analysis/PhiX/
for f in *.fasta; do
 i=$(ls $f | cut -d '_' -f 1 );
 BC=$(ls $f | cut -d '_' -f 2);
 cat "$i"_"$BC"*phiXfiltered.fa | \
 perl -ne 'if($_=~m/^>/) {$x++; $_=~s/^>(\S+)\s+/>'"$i"'_$x
 orig_bc='"$BC"'\n/} print $_;' \
 > ../QIIME/"$i"."$BC".assembled.phiXfiltered.qiimeHeaders.fa;
done;

cd ../QIIME/
rm All.PandaSpades*.fa

cat *.assembled.phiXfiltered.qiimeHeaders.fa | \
perl -ne '$_=~s/^(>.*?)\n/$1\t/; print $_;' | \
sort -k6 | \
perl -ne '$_=~s/\t/\n/; print $_;' \
> All_seqs.fna

f=$(grep -c '>' All_seqs.fna); echo -e "All_seqs_count $f"
 >>../QC/All_seqs_counts.txt

perl ~/scripts/Perl/Seq_filter.pl \
 -i All_seqs.fna \
 -min 250 \
 -max 350
mv sequences_ok.fas All_seqs_filtered.fna
mv sequences_too_long.fas ../QC/
mv sequences_too_short.fas ../QC/
f=$(grep -c '>' All_seqs_filtered.fna); echo -e "All_seqs_filtered $f"
 >>../QC/All_seqs_counts.txt
cd $dir
echo Extracting lengths for assemblies from each sample...
rm Analysis/QC/Final_Lengths.txt
for FILE in Analysis/QIIME/*.fna; do
 grep -v '^>' $FILE | perl -ne 'print "'$FILE'"."\t".length($_)."\n";';
done >> Analysis/QC/Final_Lengths_tmp.txt
cat Analysis/QC/Final_Lengths_tmp.txt | cut -d '/' -f 3  >
 Analysis/QC/Final_Lengths.txt
rm Analysis/QC/Final_Lengths_tmp.txt
cd Analysis/QC/
Rscript ~/scripts/Length_Boxplot.R
cd $dir
echo Finished Assembly, QC and filtering.
```

### 15.2 R Script for LCBD Value Calculation

```
library(phyloseq)
library(vegan)
library(ggplot2)
library(plyr)
library(phangorn)
library(ape)
library(grid)
setwd("/LCBD/")
physeq<-import_biom("20_sorted_otu_table.biom")
abund_table<-otu_table(physeq)
abund_table<-t(abund_table)
OTU_taxonomy<-as.data.frame(tax_table(physeq))
colnames(OTU_taxonomy)<-
 c("Kingdom","Phylum","Class","Order","Family","Genus","Species")
#Ensure that all columns of OTU_taxonomy are character and not factors
OTU_taxonomy[] <- lapply(OTU_taxonomy, function(x) as.character(x))
OTU_taxonomy[is.na(OTU_taxonomy)]<-""
OTU_taxonomy[]<-lapply(OTU_taxonomy,function(x)
 gsub("k__|p__|c__|o__|f__|g__|s__","",x))
OTU_tree <-read.tree("16_pfiltered_pynast_aligned_rep_set_tree.txt")
meta_table<-read.csv("metadata.tsv",row.names=1,header=T,sep="\t")
#Get rid of N/As
meta_table[meta_table=="N/A"]<-NA
meta_table[meta_table=="No data"]<-NA
abund_table<-abund_table[rownames(meta_table),]
abund_table<-abund_table[,colSums(abund_table)>0]
OTU_taxonomy<-OTU_taxonomy[colnames(abund_table),]
#DEFINE GROUPING AND THE TAXONOMIC LEVEL & DISTANCE MEASURE##############
meta_table$Groups<-
 as.factor(paste(meta_table$NEC.status,meta_table$Study_No))
which_level<-"Genus" #Phylum Class Order Family Genus Otus
#COLLATE OTUS AT A PARTICULAR LEVEL#######################################
new_abund_table<-NULL
if(which_level=="Genus"){ new_abund_table<-abund_table} else {list<-
 unique(OTU_taxonomy[,which_level])new_abund_table<-NULL for(i in list){
 tmp<-data.frame( rowSums(
 abund_table[,rownames(OTU_taxonomy)[OTU_taxonomy[,which_level]==i],drop=F]
 ))    if(i==""){colnames(tmp)<-c("__Unknowns__")} else {colnames(tmp)<-
 paste("",i,sep="")} if(is.null(new_abund_table)){new_abund_table<-tmp}
 else {new_abund_table<-cbind(tmp,new_abund_table)}}}

new_abund_table<-as.data.frame(as(new_abund_table,"matrix"))
abund_table<-new_abund_table
#Convert the data to phyloseq format
OTU = otu_table(as.matrix(abund_table), taxa_are_rows = FALSE)
TAX = tax_table(as.matrix(OTU_taxonomy))
SAM = sample_data(meta_table)
physeq<-NULL
```

```
if(which_level=="Otus"){physeq<-merge_phyloseq(phyloseq(OTU, TAX), SAM,
 midpoint(OTU_tree))} else {physeq<-merge_phyloseq(phyloseq(OTU),SAM)}
beta.div <- function(Y, method="hellinger", sqrt.D=FALSE, samp=TRUE,
 nperm=999, save.D=FALSE, clock=FALSE)
### Internal functions
{  centre <- function(D,n)
# Centre a square matrix D by matrix algebra
# mat.cen = (I - 11'/n) D (I - 11'/n)
  {  One <- matrix(1,n,n)
     mat <- diag(n) - One/n
     mat.cen <- mat %*% D %*% mat
  }
   BD.group1 <- function(Y, method, save.D, per, n)
  { if(method=="profiles") Y = decostand(Y, "total")
    if(method=="hellinger") Y = decostand(Y, "hellinger")
    if(method=="chord") Y = decostand(Y, "norm")
    if(method=="chisquare") Y = decostand(Y, "chi.square")
    s <- scale(Y, center=TRUE, scale=FALSE)^2   # eq. 1
    SStotal <- sum(s)          # eq. 2
    BDtotal <- SStotal/(n-1)   # eq. 3
    if(!per) { SCBD<-apply(s,2,sum)/SStotal }else{ SCBD<-NA }  # eqs. 4a
 and 4b
    LCBD <- apply(s, 1, sum)/SStotal  # eqs. 5a and 5b
    #
    D <- NA
    if(!per & save.D)   D <- dist(Y)
    #
    out <- list(SStotal_BDtotal=c(SStotal,BDtotal), SCBD = SCBD, LCBD=LCBD,
 method = method, D=D)
  }
  BD.group2 <- function(Y, method, sqrt.D, n)
  {
    if(method == "divergence") {
      D = D11(Y)
     } else if(any(method ==
                    c("jaccard","sorensen","ochiai")))
    {
      if(method=="jaccard") D = dist.binary(Y, method=1) # ade4 takes
 sqrt(D)
      if(method=="sorensen")  D = dist.binary(Y, method=5) #ade4 takes
 sqrt(D)
      if(method=="ochiai") D = dist.binary(Y, method=7) # ade4 takes
 sqrt(D)

    } else if(any(method ==
                    c("manhattan","canberra","whittaker","%difference","ruz
icka","wishart")))
    {
      if(method=="manhattan") D = vegdist(Y, "manhattan")
      if(method=="canberra")  D = vegdist(Y, "canberra")
      if(method=="whittaker") D = vegdist(decostand(Y,"total"),
 "manhattan")/2
```

```
      if(method=="%difference") D = vegdist(Y, "bray")
      if(method=="ruzicka")   D = RuzickaD(Y)
      if(method=="wishart")   D = WishartD(Y)
    } else {
      if(method=="modmeanchardiff") D = D19(Y)
      if(method=="kulczynski")  D = vegdist(Y, "kulczynski")
      if(method=="ab.jaccard")  D = chao(Y, coeff="Jaccard", samp=samp)
      if(method=="ab.sorensen") D = chao(Y, coeff="Sorensen", samp=samp)
      if(method=="ab.ochiai")   D = chao(Y, coeff="Ochiai", samp=samp)
      if(method=="ab.simpson")  D = chao(Y, coeff="Simpson", samp=samp)
    }
    if(sqrt.D) D = sqrt(D)
    SStotal <- sum(D^2)/n      # eq. 8
    BDtotal <- SStotal/(n-1)   # eq. 3
    delta1 <- centre(as.matrix(-0.5*D^2), n)   # eq. 9
    LCBD <- diag(delta1)/SStotal               # eq. 10b
    out <- list(SStotal_BDtotal=c(SStotal,BDtotal), LCBD=LCBD,
                method=method, D=D)
  }
  epsilon <- sqrt(.Machine$double.eps)
  method <- match.arg(method, c("euclidean", "manhattan",
"modmeanchardiff", "profiles", "hellinger", "chord", "chisquare",
"divergence", "canberra", "whittaker", "%difference", "ruzicka",
"wishart", "kulczynski", "ab.jaccard",
"ab.sorensen","ab.ochiai","ab.simpson","jaccard","sorensen","ochiai","none
"))
  if(any(method == c("profiles", "hellinger", "chord", "chisquare",
"manhattan", "modmeanchardiff", "divergence", "canberra", "whittaker",
"%difference", "kulczynski"))) require(vegan)
  if(any(method == c("jaccard","sorensen","ochiai"))) require(ade4)
  if(is.table(Y)) Y <- Y[1:nrow(Y),1:ncol(Y)]    # In case class(Y) is
"table"
  n <- nrow(Y)
  if((n==2)&(dist(Y)[1]<epsilon)) stop("Y contains two identical rows,
hence BDtotal = 0")
  aa <- system.time({
    if(any(method ==
             c("euclidean", "profiles", "hellinger", "chord",
"chisquare","none"))) {note <- "Info -- This coefficient is Euclidean"
      res <- BD.group1(Y, method, save.D, per=FALSE, n)
      # Permutation test for LCBD indices, distances group 1
      if(nperm>0) {  p <- ncol(Y)
        nGE.L = rep(1,n)
        for(iperm in 1:nperm) {
          Y.perm = apply(Y,2,sample)
          res.p <- BD.group1(Y.perm, method, save.D, per=TRUE, n)
          ge <- which(res.p$LCBD+epsilon >= res$LCBD)
          nGE.L[ge] <- nGE.L[ge] + 1
        }
        p.LCBD <- nGE.L/(nperm+1)
      } else { p.LCBD <- NA }
      if(save.D) { D <- res$D } else { D <- NA }
      out <- list(SStotal_BDtotal=res$SStotal_BDtotal, SCBD=res$SCBD,
LCBD=res$LCBD, p.LCBD=p.LCBD, method=method, note=note, D=D)
    } else {
       if(method == "divergence") {
        note = "Info -- This coefficient is Euclidean"
      } else if(any(method == c("jaccard","sorensen","ochiai"))) {
        note = c("Info -- This coefficient is Euclidean because dist.binary
","of ade4 computes it as sqrt(D). Use beta.div with option sqrt.D=FALSE")
      } else if(any(method == c("manhattan", "canberra", "whittaker",
"%difference", "ruzicka", "wishart"))) {
        if(sqrt.D) {
          note = "Info -- In the form sqrt(D), this coefficient, is
Euclidean"
        } else {
          note = c("Info -- For this coefficient, sqrt(D) would be
Euclidean", "Use is.euclid(D) of ade4 to check Euclideanarity of this D
matrix")
        }
      } else { note = c("Info -- This coefficient is not Euclidean", "Use
is.euclid(D) of ade4 to check Euclideanarity of this D matrix")
      }
      res <- BD.group2(Y, method, sqrt.D, n)
      # Permutation test for LCBD indices, distances group 2
      if(nperm>0) {
        nGE.L = rep(1,n)
        for(iperm in 1:nperm) {
          Y.perm = apply(Y,2,sample)
          res.p <- BD.group2(Y.perm, method, sqrt.D, n)
          ge <- which(res.p$LCBD+epsilon >= res$LCBD)
          nGE.L[ge] <- nGE.L[ge] + 1
        }
        p.LCBD <- nGE.L/(nperm+1)
      } else { p.LCBD <- NA }
      if(sqrt.D) note.sqrt.D<-"sqrt.D=TRUE"  else  note.sqrt.D<-
"sqrt.D=FALSE"
      if(save.D) { D <- res$D } else { D <- NA }
      out <- list(SStotal_BDtotal=res$SStotal_BDtotal, LCBD=res$LCBD,
p.LCBD=p.LCBD, method=c(method,note.sqrt.D), note=note, D=D)
    }
  })
  aa[3] <- sprintf("%2f",aa[3])
  if(clock) cat("Time for computation =",aa[3]," sec\n")
  class(out) <- "beta.div"
  out
}
RuzickaD <- function(Y # Compute the Ruzicka dissimilarity = (B+C)/(A+B+C)
(quantitative form of Jaccard).
{
  n = nrow(Y)
  mat.sq = matrix(0, n, n)
```

```
# A = W = sum of minima in among-site comparisons B = sum_site.1 - W = K[1]
- W   # sum of differences for sp(site1) > sp(site2) C = sum_site.2 - W =
K[2] - W   # sum of differences for sp(site2) > sp(site1)
 W <- matrix(0,n,n) # matrix that will receive the sums of minima (A)
 K <- apply(Y,1,sum)  # row sums: (A+B) or (A+C)
 for(i in 2:n) for(j in 1:(i-1)) W[i,j] <- sum(pmin(Y[i,], Y[j,])) # sums
of minima (A)
 for(i in 2:n) {
  for(j in 1:(i-1)) {
   mat.sq[i,j]<-(K[i]+K[j]-2*W[i,j])/(K[i]+K[j]-W[i,j]) } #
(B+C)/(A+B+C)
 }
 mat = as.dist(mat.sq)
}
D11 <- function(Y, algo=1)
 # Compute Clark's coefficient of divergence. This is coefficient D11 in
Legendre and Legendre (2012, eq. 7.51).License: GPL-2 Author:: Pierre
Legendre, April 2011
{ Y <- as.matrix(Y)
 n <- nrow(Y)
 p <- ncol(Y)
 # Prepare to divide by pp = (p-d) = no. species present at both sites
 Y.ap <- 1 - decostand(Y, "pa")
 d <- Y.ap %*% t(Y.ap)
 pp <- p-d   # n. species present at the two compared sites
 if(algo==1) {   # Faster algorithm
  D <- matrix(0, n, n)
  for(i in 2:n) {
   for(j in 1:(i-1)) {
    num <- (Y[i,]-Y[j,])
    den <- (Y[i,]+Y[j,])
    sel <- which(den > 0)
    D[i,j] = sqrt(sum((num[sel]/den[sel])^2)/pp[i,j])
   }
  }
 } else {   # Slower algorithm
  D <- matrix(0, n, n)
  for(i in 2:n) {
   for(j in 1:(i-1)) {
    temp = 0
    for(p2 in 1:p) {
     den = Y[i,p2] + Y[j,p2]
     if(den > 0) {
      temp = temp + ((Y[i,p2] - Y[j,p2])/den)^2
     }
    }
    D[i,j] = sqrt(temp/pp[i,j])
   }
  }
 }
 DD <- as.dist(D)
}
```

```
D19 <- function(Y)
  # Compute the Modified mean character difference. This is coefficient D19
 in Legendre and Legendre (2012, eq. 7.46). Division is by pp = number of
 species present at the two compared sites License: GPL-2 Author:: Pierre
 Legendre, April 2011
{  Y <- as.matrix(Y)
  n <- nrow(Y)
  p <- ncol(Y) # Prepare to divide by pp = (p-d) = n. species present at
 both sites
  Y.ap <- 1 - decostand(Y, "pa")
  d <- Y.ap %*% t(Y.ap)
  pp <- p-d # n. species present at the two compared sites
  D <- vegdist(Y, "manhattan")
  DD <- as.dist(as.matrix(D)/pp)
}
WishartD <- function(Y)
# Compute dissimilarity = (1 - Wishart similarity ratio) (Wishart 1969).
 License: GPL-2. Author:: Pierre Legendre, August 2012
{
  CP = crossprod(t(Y))
  SS = apply(Y^2,1,sum)
  n = nrow(Y)
  mat.sq = matrix(0, n, n)
  for(i in 2:n) {
    for(j in 1:(i-1)) { mat.sq[i,j] = CP[i,j]/(SS[i] + SS[j] - CP[i,j]) }
  }
  mat = 1 - as.dist(mat.sq)
}
chao <- function(mat, coeff="Jaccard", samp=TRUE)
{ require(vegan)
  nn = nrow(mat)
  res = matrix(0,nn,nn)
  if(samp) {   # First for sample data
    for(k in 2:nn) {
      for(j in 1:(k-1)) { #cat("k =",k,"  j =",j,"\n")
        v1 = mat[j,]   # Vector 1
        v2 = mat[k,]   # Vector 2
        v1.pa = decostand(v1,"pa")   # Vector 1 in presence-absence form
        v2.pa = decostand(v2,"pa")   # Vector 2 in presence-absence form
        N.j = sum(v1)   # Sum of abundances in vector 1
        N.k = sum(v2)   # Sum of abundances in vector 2
        shared.sp = v1.pa * v2.pa   # Vector of shared species ("pa")
        if(sum(shared.sp) == 0) {
          res[k,j] = 1
        } else {
          C.j = sum(shared.sp * v1)   # Sum of shared sp. abundances in v1
          C.k = sum(shared.sp * v2)   # Sum of shared sp. abundances in v2
          # a1.j = sum(shared.sp * v1.pa)
          # a1.k = sum(shared.sp * v2.pa)
          a1.j = length(which((shared.sp * v2) == 1)) # Singletons in v2
          a1.k = length(which((shared.sp * v1) == 1)) # Singletons in v1
          a2.j = length(which((shared.sp * v2) == 2)) # Doubletons in v2
          if(a2.j == 0) a2.j <- 1
          a2.k = length(which((shared.sp * v1) == 2)) # Doubletons in v1
          if(a2.k == 0) a2.k <- 1
          # S.j = sum(v1[which(v2 == 1)]) #Sum abund. in v1 for singletons
in v2
          # S.k = sum(v2[which(v1 == 1)]) # Sum abund. in v2 for singletons
in v1
          sel2 = which(v2 == 1)
          sel1 = which(v1 == 1)
          if(length(sel2)>0) S.j = sum(v1[sel2]) else S.j = 0
          if(length(sel1)>0) S.k = sum(v2[sel1]) else S.k = 0
          U.j = (C.j/N.j) + ((N.k-1)/N.k) * (a1.j/(2*a2.j)) * (S.j/N.j) #
Eq. 11
          if(U.j > 1) U.j <- 1
          U.k = (C.k/N.k) + ((N.j-1)/N.j) * (a1.k/(2*a2.k)) * (S.k/N.k) #
Eq. 12
          if(U.k > 1) U.k <- 1
          if(coeff == "Jaccard") {                     # "Jaccard"
            res[k,j] = 1 - (U.j*U.k/(U.j + U.k - U.j*U.k))
          } else if(coeff == "Sorensen") {         # "Sorensen"
            res[k,j] = 1 - (2*U.j*U.k/(U.j + U.k))
          } else if(coeff == "Ochiai") {           # "Ochiai"
            res[k,j] = 1 - (sqrt(U.j*U.k))
          } else if(coeff == "Simpson") {
            # Simpson (1943), or Lennon et al. (2001) in Chao et al. (2006)
            res[k,j] = 1 -
              (U.j*U.k/(U.j*U.k+min((U.j-U.j*U.k),(U.k-U.j*U.k))))
          } else { #
            stop("Incorrect coefficient name")}}}} else {   # Now for
complete population data
    for(k in 2:nn) {
      for(j in 1:(k-1)) {
        v1 = mat[j,]   # Vector 1
        v2 = mat[k,]   # Vector 2
        v1.pa = decostand(v1,"pa")   # Vector 1 in presence-absence form
        v2.pa = decostand(v2,"pa")   # Vector 2 in presence-absence form
        shared.sp = v1.pa * v2.pa     # Vector of shared species ("pa")
        if(sum(shared.sp) == 0) {
          res[k,j] = 1
        } else {
          N1 = sum(v1)   # Sum of abundances in vector 1
          N2 = sum(v2)   # Sum of abundances in vector 2
          U = sum(shared.sp * v1)/N1   # Sum of shared sp. abundances in v1
          V = sum(shared.sp * v2)/N2   # Sum of shared sp. abundances in v2

          if(coeff == "Jaccard") {                     # "Jaccard"
            res[k,j] = 1 - (U*V/(U + V - U*V))
          } else if(coeff == "Sorensen") {         # "Sorensen"
            res[k,j] = 1 - (2*U*V/(U + V))
          } else if(coeff == "Ochiai") {           # "Ochiai"
            res[k,j] = 1 - (sqrt(U*V))
          } else if(coeff == "Simpson") { # "Simpson"
```

```
            res[k,j] = 1 - (U*V/(U*V+min((U-U*V),(V-U*V)))) # Eq. ?
          } else { #
            stop("Incorrect coefficient name")}}}}}res <- as.dist(res)}
######## End of beta.div function
beta_div<-
 beta.div(otu_table(physeq),method="hellinger",sqrt.D=F,samp=T,nperm=999)
#meta_table<-
 read.csv("../../data/All_gold_standard_case_control_metadata.tsv",row.name
 s=1,header=T, sep="\t")
meta_table<-
 data.frame(meta_table,data.frame(LCBD=beta_div$LCBD,p.LCBD=beta_div$p.LCBD
 ))
df_LCBD<-
 data.frame(Sample=names(beta_div$LCBD),LCBD=beta_div$LCBD,p.LCBD=beta_div$
 p.LCBD)
df_LCBD<-
 data.frame(df_LCBD,Type=meta_table[rownames(df_LCBD),"Groups",drop=F])
names(meta_table)
#Apply proportion normalisation
x<-otu_table(physeq)/rowSums(otu_table(physeq))
x<-x[,order(colSums(x),decreasing=TRUE)]

#Extract list of top N Taxa
N=0
if(which_level=="Otus"){N<-21} else {N<-22}
taxa_list<-colnames(x)[1:N]
#remove "__Unknown__" and add it to others
taxa_list<-taxa_list[!grepl("Unknown",taxa_list)]
N<-length(taxa_list)
#Generate a new table with everything added to Others
new_x<-data.frame(x[,colnames(x) %in%
 taxa_list],Others=rowSums(x[,!colnames(x) %in% taxa_list]))
if(which_level=="Otus"){
  colnames(new_x)<-c(paste(colnames(new_x)[-(N+1)],sapply(colnames(new_x)[-
 (N+1)],function(x)
 gsub(".*;","",gsub(";+$","",paste(sapply(OTU_taxonomy[x,],as.character),co
 llapse=";"))))),"Others")}
```

### 15.3 R Script for CCA Analysis of Risk Factors and LCBD Value Association

```
library(ggplot2)
library(vegan)
library(grid)
library(phyloseq)
setwd("/CCA analysis/")
physeq<-import_biom("20_sorted_otu_table.biom")
abund_table<-otu_table(physeq)
abund_table<-t(abund_table)
meta_table<-read.csv("metadata.csv",
#Convert to relative frequencies
abund_table<-abund_table/rowSums(abund_table)
```

```
#Use adonis to find significant environmental variables
abund_table.adonis <- adonis(abund_table ~ ., data=meta_table)
# # adonis(formula = abund_table ~ ., data = meta_table)
## remove LCBD and p.LCBD columns
meta_table <- subset(meta_table, select = -LCBD )
meta_table <- subset(meta_table, select = -p.LCBD )
abund_table.adonis <- adonis(abund_table ~ ., data=meta_table, method =
 "bray",)
abund_table.adonis
#Extract the best variables
bestEnvVariables<-
 rownames(abund_table.adonis$aov.tab)[abund_table.adonis$aov.tab$"Pr(>F)"<=
 0.01]
  #We are now going to use only those environmental variables in cca that
 were found significant
eval(parse(text=paste("sol <- cca(abund_table ~
 ",do.call(paste,c(as.list(bestEnvVariables),sep=" +
 ")),",data=meta_table)",sep="")))
#Use the following to use all the environmental variables
scores(sol, display=c("sp"))[1:5,]
scores(sol, display=c("wa"))[1:5,]
scores(sol, display=c("lc"))[1:5,]
scores(sol, display=c("bp"))[1:5,]
scores(sol, display=c("cn"))[1:5,]
scrs<-scores(sol,display=c("sp","wa","lc","bp","cn"))
summary(scrs$sites)
#Check the attributes
## Rename CCA1 and CC2 to x and y (axis)
colnames(scrs$sites)<-c("x","y")
#Extract site data first
names(meta_table)
df_sites<-data.frame(scrs$sites,as.data.frame(meta_table))
df_sites[1:5,]
#Draw sites
p<-ggplot()
p<-p+geom_point(data=df_sites,aes(x,y,colour= Study_No))
p<-p#+guides(col=guide_legend(ncol=1))
#Draw biplots
multiplier <- vegan:::ordiArrowMul(scrs$biplot)
summary(meta_table$Age.Days)
df_arrows<- scrs$biplot*multiplier
colnames(df_arrows)<-c("x","y")
df_arrows=as.data.frame(df_arrows)
p<-p+geom_segment(data=df_arrows, aes(x = 0, y = 0, xend = x, yend = y),
                 arrow = arrow(length = unit(0.2,
 "cm")),color="#808080",alpha=0.5)
p<-p+geom_text(data=as.data.frame(df_arrows*1.1),aes(x, y, label =
 rownames(df_arrows)),color="#808080",alpha=0.5)
p<-p+guides(col=guide_legend(ncol=1))
# Draw species
df_species<- as.data.frame(scrs$species)
colnames(df_species)<-c("X","Y")
```

```
names(df_species)
rownames(df_species)
p<-p+theme_bw()+theme(text=element_text(family="Times New Roman",
 face="bold", size=11))+labs(x="X", y="Y", colour="Subject ID")
png("CCA.png", width=12, height=6.5, units="in", pointsize=1, res=1200)
print(p)
dev.off()
```

## 15.4 R Script for NMDS Analysis

```
library(phyloseq)
library(vegan)
library(ggplot2)
library(ape)
library(phangorn)
physeq<-import_biom("20_sorted_otu_table.biom")
abund_table<-otu_table(physeq)
abund_table<-t(abund_table)
OTU_taxonomy<-as.data.frame(tax_table(physeq))
colnames(OTU_taxonomy)<-
 c("Kingdom","Phylum","Class","Order","Family","Genus","Species")
#Ensure that all columns of OTU_taxonomy are character and not factors
OTU_taxonomy[] <- lapply(OTU_taxonomy, function(x) as.character(x))
OTU_taxonomy[is.na(OTU_taxonomy)]<-""
OTU_taxonomy[]<-lapply(OTU_taxonomy,function(x)
 gsub("k__|p__|c__|o__|f__|g__|s__","",x))
#Load the tree using ape package
OTU_tree <- read.tree("16_pfiltered_pynast_aligned_rep_set_tree.txt")
meta_table<-read.csv("metadata.csv",row.names=1,header=T,sep=",")
### Subsetting cohort using meta_table and delivery method
attach(meta_table)
meta_table<-meta_table[ which(Delivery=='Vaginal' & Feeds =="Natural"),]
detach(meta_table)
abund_table<-abund_table[rownames(meta_table),]
abund_table<-abund_table[,colSums(abund_table)>0]
OTU_taxonomy<-OTU_taxonomy[colnames(abund_table),]
#DEFINE GROUPING AND THE TAXONOMIC LEVEL & DISTANCE MEASURE##############
meta_table$Groups<-as.factor(paste(meta_table$NEC.status))
which_level<-"Genus" #Phylum Class Order Family Genus Otus
which_distance<-"bray" #bray unifrac wunifrac
#COLLATE OTUS AT A PARTICULAR LEVEL#######################################
new_abund_table<-NULL
if(which_level=="Otus"){new_abund_table<-abund_table} else { list<-
 unique(OTU_taxonomy[,which_level]) new_abund_table<-NULL for(i in
 list){tmp<-data.frame (rowSums (abund_table[,rownames (OTU_taxonomy)
 [OTU_taxonomy [,which_level]==i],drop=F])) if(i==""){colnames(tmp)<-
 c("__Unknowns__")} else {colnames(tmp)<-paste("",i,sep="")}
 if(is.null(new_abund_table)){new_abund_table<-tmp} else {new_abund_table<-
 cbind(tmp,new_abund_table)}}}
new_abund_table<-as.data.frame(as(new_abund_table,"matrix"))
abund_table<-new_abund_table
```

```
#Convert the data to phyloseq format
OTU = otu_table(as.matrix(abund_table), taxa_are_rows = FALSE)
TAX = tax_table(as.matrix(OTU_taxonomy))
SAM = sample_data(meta_table)
physeq<-NULL
if(which_level=="Otus"){ physeq<-merge_phyloseq(phyloseq(OTU,
 TAX),SAM,midpoint(OTU_tree))} else {physeq<-
 merge_phyloseq(phyloseq(OTU),SAM)}

#Plotting Elipses
veganCovEllipse<-function (cov, center = c(0, 0), scale = 1, npoints =
 100)
{ theta <- (0:npoints) * 2 * pi/npoints
  Circle <- cbind(cos(theta), sin(theta))
  t(center + scale * t(Circle %*% chol(cov)))}
#coloring function
gg_color_hue<-function(n){
  hues=seq(15,375,length=n+1)
  hcl(h=hues,l=65,c=100)[1:n]}
sol<-NULL
if(which_distance=="bray"){
  sol<-ordinate(physeq, "NMDS",distance="bray")
} else if(which_distance=="wunifrac" & which_level=="Otus") {
  sol<-ordinate(physeq, "NMDS",distance="wunifrac")
} else if(which_distance=="unifrac" & which_level=="Otus"){
  sol<-ordinate(physeq, "NMDS",distance="unifrac")
}
if(!is.null(sol)){
  NMDS=data.frame(x=sol$points[,1],y=sol$points[,2],meta_table)

  plot.new()
  ord<-ordiellipse(sol, meta_table$Groups,display = "sites", kind ="se",
 conf = 0.95, label = T)
  dev.off()

 #Generate ellipse points
 df_ell <- data.frame()
 for(g in levels(NMDS$Groups)){
   if(g!="" && (g %in% names(ord))){
     if(sum(NMDS$Groups==g)>2){
       tryCatch(df_ell <- rbind(df_ell,
 cbind(as.data.frame(with(NMDS[NMDS$Groups==g,],
veganCovEllipse(ord[[g]]$cov,ord[[g]]$center,ord[[g]]$scale))),Groups=g)),e
 rror=function(e) NULL)}}}
  colnames(df_ell)<-c("x","y","Groups")
  #Generate mean values from NMDS plot grouped on
  NMDS.mean=aggregate(NMDS[,1:2],list(group=NMDS$Groups),mean)
  cols=gg_color_hue(length(unique(NMDS$Groups)))
  p<-ggplot(data=NMDS,aes(x,y,colour=Groups))
  p<-p + geom_point(alpha=0.5,size = 2)
  p<-p+theme_bw()
  p<-p+
 annotate("text",x=NMDS.mean$x,y=NMDS.mean$y,label=NMDS.mean$group,size=6,c
 olour=cols,family="Courier",fontface="bold",alpha=0.8,vjust=0.3)
  p<-p+ geom_path(data=df_ell, aes(x=x, y=y), size=1, linetype=1,alpha=0.3)
  p<-p+xlab("NMDS1")+ylab("NMDS2")
  pdf(paste("NMDS_",which_distance,"_",which_level,"_Grouping1",".pdf",sep=
 ""),width=7,height=6)
  print(p)
  dev.off()
  dist<-phyloseq::distance(physeq,which_distance)
  capture.output(adonis(dist ~ Groups,
 data=meta_table[rownames(otu_table(physeq)),]),file=paste("ADONIS_",which_
 distance,"_",which_level,"_Grouping1",".txt",sep=""))
}
```

## 15.5 R Script for Differential and Random Forest Analysis of Taxa

```
library(phyloseq)
library(vegan)
library(ggplot2)
library(plyr)
library(DESeq2)
library(extrafont)
font_import()
loadfonts(device="win")
library(reshape)
fonts()
library(phangorn)
library(randomForest)
which_level<-"Species" #Phylum Class Order Family Genus Otus
which_distance<-"bray" #bray unifrac wunifrac
sig = 0.05
fold = 2
physeq<-import_biom("20_sorted_otu_table.biom")
abund_table<-otu_table(physeq)
abund_table<-t(abund_table)
OTU_taxonomy<-as.data.frame(tax_table(physeq))
colnames(OTU_taxonomy)<-
 c("Kingdom","Phylum","Class","Order","Family","Genus","Species")
#Ensure that all columns of OTU_taxonomy are character and not factors
OTU_taxonomy[] <- lapply(OTU_taxonomy, function(x) as.character(x))
OTU_taxonomy[is.na(OTU_taxonomy)]<-""
OTU_taxonomy[]<-lapply(OTU_taxonomy,function(x)
 gsub("k__|p__|c__|o__|f__|g__|s__","",x))
OTU_tree <- read.tree("16_pfiltered_pynast_aligned_rep_set_tree.txt")
meta_table<-read.csv("meta_data.csv", row.names=1, header=T)
# attach(meta_table) # Subsetting if necessary
# meta_table<-meta_table[ which (Delivery =='Caesarean' & Feeds
 =="Mixed"),]
```

```
# detach(meta_table)
abund_table<-abund_table[rownames(meta_table),]
abund_table<-abund_table[,colSums(abund_table)>0]
OTU_taxonomy<-OTU_taxonomy[colnames(abund_table),]
#CHANGE THE GROUPING COLUMN AS YOU DESIRE##################################
#Hypothesis 1: Taxonomic abundances are significantly different between NEC
 and controls
meta_table$Groups<-as.factor(as.character(meta_table$NEC.status))
new_abund_table<-NULL if(which_level=="Otus"){new_abund_table<-abund_table
} else {list<-unique(OTU_taxonomy[,which_level]) new_abund_table<-NULL
 for(i in list){ tmp<-data.frame (rowSums (abund_table [,rownames
 (OTU_taxonomy) [OTU_taxonomy[,which_level]==i],drop=F]))
 if(i==""){colnames(tmp)<-c("__Unknowns__")} else {colnames(tmp)<-
 paste("",i,sep="")} if(is.null(new_abund_table)){new_abund_table<-tmp}
 else {new_abund_table<-cbind(tmp,new_abund_table)}}}
new_abund_table<-as.data.frame(as(new_abund_table,"matrix"))
abund_table<-new_abund_table
#We will convert our table to DESeqDataSet object
countData = round(as(abund_table, "matrix"), digits = 0)
countData<-(t(countData+1))
dds <- DESeqDataSetFromMatrix(countData, meta_table, as.formula(~ Groups))
data_deseq_test = DESeq(dds)

## Extract the results
res = results(data_deseq_test, cooksCutoff = FALSE)
res_tax = cbind(as.data.frame(res), as.matrix(countData[rownames(res), ]),
 OTU = rownames(res))
plot.point.size = 2
label=F
tax.display = NULL
tax.aggregate = "OTU"
res_tax_sig = subset(res_tax, padj < sig & fold < abs(log2FoldChange))
res_tax_sig <- res_tax_sig[order(res_tax_sig$padj),]
res_tax$Significant <- ifelse(rownames(res_tax) %in% rownames(res_tax_sig)
 , "Yes", "No")
res_tax$Significant[is.na(res_tax$Significant)] <- "No"
res_tax_sig_abund = cbind(as.data.frame(countData[rownames(res_tax_sig),
 ]), OTU = rownames(res_tax_sig), padj =
 res_tax[rownames(res_tax_sig),"padj"])

#Apply normalisation (either use relative or log-relative transformation)
data<-log((abund_table+1)/(rowSums(abund_table)+dim(abund_table)[2]))
data<-as.data.frame(data)
normalised_counts<-data

## Log Fold Differences in Taxa
### MA plot
res_tax$Significant <- ifelse(rownames(res_tax) %in% rownames(res_tax_sig)
 , "Yes", "No")
res_tax$Significant[is.na(res_tax$Significant)] <- "No"
temp<-res_tax[res_tax$Significant == "Yes",]
dim(temp)
```

```
### Table assocaited with NB_MA and NB_significant graphs
temp<-temp[,c(1,2,6),]
temp[with(temp, order(-baseMean)), ]
p1 <- ggplot(data = res_tax, aes(x = baseMean, y = log2FoldChange, color =
 Significant)) + geom_point(size = plot.point.size) + scale_x_log10() +
 scale_color_manual(values=c("black", "red")) + labs(x = "Mean abundance",
 y = "Log2 fold change")+theme_bw()+theme(text=element_text(family="Times
 New Roman", face="bold", size=12)) if(label == T){ if
 (!is.null(tax.display)){rlab <- data.frame(res_tax, Display =
 apply(res_tax[,c(tax.display, tax.aggregate)], 1, paste, collapse="; "))}
 else { rlab <- data.frame(res_tax, Display = res_tax[,tax.aggregate])}
  p1 <- p1 + geom_text(data = subset(rlab, Significant == "Yes"), aes(label
 = Display), size = 4, vjust = 1)}

#### Log 2-Fold Summary, scatter
png(paste("NB_MA_",paste(levels(meta_table$Groups),collapse="_"),"_",which_
 level,".png",sep=""), width=4.2, height=3.2, units="in", pointsize=1,
 res=1200)
print(p1)
dev.off()
res_tax_sig_abund = cbind(as.data.frame(countData[rownames(res_tax_sig),
 ]), OTU = rownames(res_tax_sig), padj =
 res_tax[rownames(res_tax_sig),"padj"])

#Apply normalisation
data<-log((abund_table+1)/(rowSums(abund_table)+dim(abund_table)[2]))
data<-as.data.frame(data)
df<-NULL #Now we plot taxa significantly different between the categories
sig_otus<-res_tax[rownames(res_tax_sig),"OTU"]
for(i in sig_otus){ tmp<-NULL if(which_level=="Otus"){ tmp<-
 data.frame(data[,i],meta_table$Groups,rep(paste(paste(i,gsub(".*;","",gsub
 (";+$","",paste(sapply(OTU_taxonomy[i,],as.character),collapse=";")))),"
 padj = ",sprintf("%.5g",res_tax[i,"padj"]),sep=""),dim(data)[1]))
  } else { tmp<-data.frame(data[,i],meta_table$Groups,rep(paste(i," padj =
 ",sprintf("%.5g",res_tax[i,"padj"]),sep=""),dim(data)[1]))}
  if(is.null(df)){df<-tmp} else { df<-rbind(df,tmp)}}
colnames(df)<-c("Value","Type","Taxa")
summary(df$Taxa)

#### Log 2-Fold Significant Taxa, barplot
p<-ggplot(df,aes(Type,Value,colour=Type))+ylab("Log-relative normalised")
p<-p+geom_boxplot(outlier.size = 0)+geom_jitter(position =
 position_jitter(height = 0, width=0),alpha=0.5,outlier.colour =
 NULL)+theme_bw()+
  facet_wrap( ~ Taxa , scales="free_x",nrow=1)
p<-p + theme (axis.text.x=element_text (angle=90,hjust=1,vjust=0.5)) +
 theme(strip.text.x = element_text(size = 16, colour = "black", angle =
 90))+theme(text=element_text( face="bold", size=12, family="Times New
 Roman"))
```

```
pdf(paste("NB_significant_",paste(levels(meta_table$Groups),collapse="_"),"
_",which_level,".pdf",sep=""),width=ceiling((length(sig_otus)*60/200)+2.6)
,height=8)
print(p)
dev.off()

# #Now we will use Breiman's random forest algorithm that can be used in
 unsupervised mode for assessing proximities
# #among data points
# #Apply normalisation (log-relative transformation)
abund_table_2<-abund_table
names(abund_table_2) <- gsub(x = names(abund_table_2), pattern = "\\[|\\]",
 replacement = "")
res_tax_sig_2<-res_tax_sig
rownames(res_tax_sig_2)<- gsub(x = rownames(res_tax_sig_2), pattern =
 "\\[|\\]", replacement = "")

data<-log((abund_table_2+1)/(rowSums(abund_table_2)+dim(abund_table_2)[2]))
data<-as.data.frame(data)
subset.data<-data[,as.character(res_tax[rownames(res_tax_sig_2),"OTU"])]

############ Linear Graphs for Significant Taxa
#### Log Normalised Abundance for Genus of Interest over time relative to
 NEC status
normalised_meta<-merge(normalised_counts, meta_table, by=0)
names(normalised_meta)
p<-ggplot(normalised_meta, aes(x=Age.Days, y= butyricum, colour=Study_No,
 shape=NEC.status))+geom_point()+geom_smooth(method="loess")+labs(x="Age /
 Days", y="Log Normalised Abundance", colour="NEC
 Status")+theme_bw()+theme(text=element_text(family="Times New Roman",
 face="bold", size=11))+geom_text(aes(label=Study_No),hjust=0, vjust=0)
png(filename="4_Regression_Norm_Abund_bray_butyricum.png", width=6.4,
 height=4, units="in", pointsize=1, res=1200)
p
dev.off()

### Random Forest Analaysis
#Now we plot taxa significantly different between the categories
res_tax_sig_abund = cbind(as.data.frame(countData[rownames(res_tax_sig),
 ]), OTU = rownames(res_tax_sig), padj =
 res_tax[rownames(res_tax_sig),"padj"])
#Apply normalisation (either use relative or log-relative transformation)
data<-log((abund_table+1)/(rowSums(abund_table)+dim(abund_table)[2]))
data<-as.data.frame(data)
df<-NULL
sig_otus<-res_tax[rownames(res_tax_sig),"OTU"]
for(i in sig_otus){tmp<-NULL if(which_level=="Otus"){ tmp<-
 data.frame(data[,i],meta_table$Groups,rep(paste(paste(i,gsub(".*;","",gsub
 (";+$","",paste(sapply(OTU_taxonomy[i,],as.character),collapse=";")))),"
 padj = ",sprintf("%.5g",res_tax[i,"padj"]),sep=""),dim(data)[1]))} else {
 tmp<-data.frame(data[,i],meta_table$Groups,rep(paste(i," padj =
 ",sprintf("%.5g",res_tax[i,"padj"]),sep=""),dim(data)[1]))}
 if(is.null(df)){df<-tmp} else { df<-rbind(df,tmp)}}
colnames(df)<-c("Value","Type","Taxa")

subset.data<-data[,as.character(res_tax[rownames(res_tax_sig),"OTU"])]
names(subset.data) <-gsub("\\[","",colnames(subset.data))
names(subset.data) <-gsub("\\]","",colnames(subset.data))
IDs_map<-data.frame (row.names=gsub (";.*$","",colnames(subset.data)),
 To=colnames(subset.data))
names(subset.data)<-paste("X",names(subset.data),sep="")

val<-randomForest(meta_table$Groups ~ ., data=subset.data, importance=T,
 proximity=T, ntree=1500, keep.forest=F)
rownames(val$importance)<-gsub("^X","",rownames(val$importance))
imp<-importance(val)
df_accuracy<-data.frame (row.names=NULL, Sample=rownames(imp),
 Value=abs(as.numeric(imp[,"MeanDecreaseAccuracy"])),Index=rep("Mean
 Decrease Accuracy",dim(imp)[1]))

r<-ggplot(data=df_accuracy, aes(x= reorder(Sample, -Value),Value))
r<-r+geom_bar(fill=I("red"),stat="identity")+theme_bw()
r<-r+theme(axis.text.x=element_text(angle=90,hjust=1,vjust=0.5))+ylab("Mean
 Decrease Accuracy")
r<-r+theme(axis.title.x=element_blank(),text=element_text(family="Times New
 Roman", face="bold", size=11))
pdf(paste("RF_MDA_",paste(levels(meta_table$Groups),collapse="_"),"_",which
 _level,".pdf",sep=""),width=5,height=5)
print(r)
dev.off()

df_gini<-data.frame (row.names=NULL, Sample=rownames(imp), Value=as.numeric
 (imp[,"MeanDecreaseGini"]),Index=rep("Mean Decrease Gini",dim(imp)[1]))


s<-ggplot(data=df_gini,aes(x= reorder(Sample, -Value),Value))
s<-s+geom_bar(fill=I("red"),stat="identity")+theme_bw()
s<-s+theme(axis.text.x=element_text(angle=90,hjust=1,vjust=0.5))+ylab("Mean
 Decrease Gini")
s<-s+theme(axis.title.x=element_blank(),text=element_text(family="Times New
 Roman", face="bold", size=11))

pdf(paste("RF_MDG_",paste(levels(meta_table$Groups),collapse="_"),"_",which
 _level,".pdf",sep=""),width=5,height=5)
print(s)
dev.off()

######### Visualising Random Forest Output
options(repos='http://cran.rstudio.org')
have.packages <- installed.packages()
cran.packages <- c('devtools','plotrix','randomForest','tree')
to.install <- setdiff(cran.packages, have.packages[,1])
if(length(to.install)>0) install.packages(to.install)
```

```
library(devtools)
if(!('reprtree' %in% installed.packages())){
 install_github('araastat/reprtree')} for(p in c(cran.packages,
 'reprtree')) eval(substitute(library(pkg), list(pkg=p)))
library(reprtree)
tree<-randomForest(meta_table$Groups ~ ., data=subset.data, importance=T,
 proximity=T,ntree=1500,keep.forest=T)
subset.data.tree<-signif(subset.data, 4)
reptree <- ReprTree(tree, subset.data.tree, metric="d2")
temp<-getTree(tree, k=1, labelVar=TRUE)
realtree<-reprtree:::as.tree(temp, tree)
pdf("4_RF_Natural_Vag_Annotated.pdf",width=100,height=100, fonts='Times New
 Roman')
plot(T, cex=0.25, cex.lab=0.5)
dev.off()
```

[ END ]