Strathclyde Institute of Pharmacy & Biomedical Sciences

2022

The Development and Application of Routine Gastrointestinal Microbiome Characterisation in the Laboratory Mouse using Next Generation Sequencing.

Julian D. Phipps, BSc. (Hons.), MSc.
GlaxoSmithKline, Stevenage
Hertfordshire, SG1 2NY

Academic Supervisor: Professor Paul Hoskisson, University of Strathclyde, Glasgow.

Industrial Supervisor: Doctor Peter Clements, GlaxoSmithKline, Ware.

Declaration

This project report is submitted as part of the requirements for a PhD. awarded by the Strathclyde Institute of Pharmacy & Biomedical Sciences at the University of Strathclyde, Glasgow. Except where otherwise stated, I declare that this work is that of the author. Sources of other information have been appropriately acknowledged.

Signed:

Date: 8th August 2022

Acknowledgements

I would like to thank Paul Hoskisson, Kumar Changani, Greg Whelan, Helen Murphy, Peter Clements, Paul Heron, Lindsay Hall, Kerry Theaker, Mark Lennon, Thomas Leary, Stuart Wallace, Lesley Hoyles, Niall McMullan, Jon Savage, Clement Coxsone Dodd, Joy Mercer, Cal, Rainy, Bones, Tezz, and Scott Poynter for their help, support, or inspiration in this endeavour.

For Daisy & Poppy

We are here face to face with the crucial paradox of knowledge. Year by year we devise more precise instruments with which to observe nature with more fineness, and when we look at the observations they are as uncertain as ever. We seem to be running after a goal which lurches away from us to infinity every time, we come within sight of it.

Jacob Bronowski (1973)

Abstract

Despite >60% of UK *in vivo* studies being conducted upon mice, routine characterisation of its gastrointestinal (GI) microbiome during drug discovery is not conducted despite the understanding of its role in the health of the host. Outsourced 16S rRNA gene-based experiments were conducted to develop an ethical sampling strategy by assessing diversity along the GI tract, the effect of transport, and sex bias. The diversity of excreted faeces was identical to colonic digesta, while the stomach was highly populated in contrast to the small intestine in both sexes. During transit from a commercial supplier 29% of operational taxonomic units (OTUs) were lost from the GI tract, which was shown to last >3-months in subsequent analysis. A comparison of GI diversity using 16S rRNA gene analysis, DNA metagenomics, and RNA-seq was conducted using the MG-RAST analysis server. No taxonomic agreement was evident between methods or databases while slow processing deterred further use. An in-house, hybrid, species-level method of characterising the 16S rRNA gene was then developed by coupling the freely accessible RefSeq database and licenced Lasergene alignment software. This enabled the improved illustration of the effect of disease progression upon prokaryotic communities in both dextran sulfate sodium (DSS) and CD4+ adoptive transfer mouse models of irritable bowel disease (IBD). DSS transiently effected a range of prokaryotes in a dose dependant manner, while the GI microbiome of immunocompromised mice remained unaffected by both transport and the inflammatory response initiated by CD4+ transfer. Using this evaluative tool

has increased the scope of routine health monitoring and the understanding of induced disease progression in the laboratory mouse.

List of Tables

List of Figures

Table of Contents

Chapter 1: Introducing the microbiome

1.1 The microbiome

The microbiome is defined as the entire breadth of microbial species (the microbiota) and their genetic content, in relation to the habitat in which they are found. Environmental habitats play a role in shaping microbiomes but are conversely affected by community members, their functions, and perturbations (Cullen *et al*, 2020). These integrated microbial communities are now detectable by advances in DNA/RNA analysis (Marchesi & Ravel, 2015). However, much of the research in this area has focused on what can be achieved by these advances, not what can be routinely applied (Cullen *et al,* 2020).

1.2 Use of animals in research

Screening potential therapeutic agents utilising *in vivo* models is a necessary step in the pathway from drug discovery to validation, bridging the safety assessment gap between rational compound design and clinical trials in humans. Current European legislation demands pharmaceutical compound safety testing on one non-rodent and one rodent species (ICH, 2009). Common species used by the pharmaceutical industry to investigate compound activity, metabolism, and toxicity include mice, rats, dogs, pigs, and primates. Of the 3.06 million experimental procedures conducted on animals in Britain during 2021, 934,200 were specifically conducted upon mice, while a further 1.15 million were conducted for the development or maintenance of genetically altered mouse models (HMHO, 2022). Although the use of *in vivo* work has fallen since 2004, these figures show a 6% increase in usage on

2020, due to the lifting of workplace restrictions imposed during the SARS-CoV-2 pandemic (HMHO, 2022; Wu *et al*, 2020).

1.3 Current health monitoring of *in vivo* models

Healthy animal models are an absolute prerequisite to all science founded upon their use in research. In Britain, all animals used in scientific procedures are protected from significant pain and discomfort by the Animals (Scientific Procedures) Act 1986 (HMHO, 1986). This act also ensures the provision of a certain level of veterinary care. However, more specific care and welfare considerations are detailed in the pursuant 'Code of Practice' (HMSO, 1989) to the 1986 Act which stipulates the provision of routine animal health monitoring. Specific guidelines regarding the methodology and frequency of this microbiological monitoring of *in vivo* models are provided in the Federation of European Laboratory Animal Science Associations (FELASA) recommendations (Mähler *et al*, 2014). Currently, this activity uses a range of microbiological assays; direct examination, serology, microscopy, culture, and polymerase chain reaction (PCR). The results of these analyses are knitted together to confirm the absence or presence of a list of prescribed deleterious microorganisms. However, the importance of the endogenous commensal microbiota has been ignored in this field, due to the lack of a routine methods to isolate all microorganisms in an ecological niche simultaneously (Lederberg, 2000). Next generation sequencing (NGS) has circumvented all barriers to this aim and offers the ability to understand what species of prokaryotes are present (16S rRNA amplicon analysis), what potential gene complement there is in an environmental sample (DNA metagenomics), and what genes a given microorganism may be expressing at a given time (RNA-seq; Izard & Rivera, 2015). Applying these methods in a diagnostic setting would push microbiology from lone identifications to population wide illustrations of potential gene function and expression patterning, generating a step change in understanding the ecology and metabolic processes of a specific habitat or host (Izard & Rivera, 2015).

1.4 Impact of infections and outbreaks

Animal models have long been used to reproduce or mimic the developmental origins of disease. They are also used in studies of disease amelioration through the administration of novel drug compounds or as tools to understand disease progression and systemic influence in the search for human therapeutics. However, housing large rodent populations in necessarily confined areas opens animal models up to the transmission of natural infections. Disease outbreaks and subsequent depopulations are an accepted risk in experimental colonies (Mähler *et al*, 2014). This led to the development of an exclusion criteria by FELASA which is routinely applied to trace outbreaks and hinder their progression. This ever-increasing list of microorganisms includes overt pathogens, organisms which impact breeding, zoonotic agents, and opportunistic species. The list contains members of all classes of microorganism (viruses, fungi, parasites, and bacteria) which necessities the implementation of the range of diagnostic methods used in their detection. However, global prevalence data indicates that only a handful of microbial entities stimulate this complex and expensive activity (Pritchett-Corning *et al*, 2009). The most recent survey of their occurrence indicated that agents with the highest prevalence were Mouse norovirus (24%), *Helicobacter* spp. (6.6%), *Rodentibacter pneumotropica* (13%), *Staphylococcus aureus* (23%) and *Entamoeba muris* (8%; Pritchett-Corning *et al*, 2009). All these particular agents are shed in faeces and are detectable by PCR analysis, however, other members of the FELASA exclusion list may only be reliably detected by culture, microscopy, or serology due to a lack of specificity in nucleic databases. This comes at a time where only <10% of agents on the exclusion list are isolated due to the success of FELASA in promoting models with a standardised microbiota (Mähler *et al*, 2014). It is now understood is that the health of animal colonies and the reproducibility of research data based on their use is equally influenced by their commensal microbiota (Ericsson *et al*, 2017).

The use of NGS technology has not yet been applied to routine health monitoring. In addition to allowing researchers to gauge a baseline microbiome reading either prior to animal delivery or study commencement, its use has the potential to elucidate how microbial communities are affected during study progression in illustrative therapeutic areas such as IBD. The application of NGS in this field would allow better model selection, more relevant study outcomes, reduced animal use, reduced animal discomfort, and ultimately reduced attrition rates in drug discovery programmes.

1.5 Drug metabolism

Less than 0.3% of compounds progress from candidate selection to market, therefore, reducing this attrition rate during the drug development process is a major challenge for the pharmaceutical industry (Waring *et al*, 2016). The reasons for this disparity are difficult to attribute to a single factor, although the highest single driver in attrition is due to commercial or company restructuring decisions, cutting 28% of potential medicines from the pipeline. However, a further 30% of portfolio reductions are generated during *in vivo* studies and are due to an amalgam of safety issues, lack of efficacy, and low bioavailability (Waring *et al*, 2016). Of the fraction of compounds that reach the market, around 85% are developed to be administered orally. This is as much to do with the cost of production as it is drug compliance, but it means that any compound must contend with a range of GI environmental conditions and the activity of their associated microbial communities (Sousa *et al*, 2008).

It is already appreciated that >50 presently licensed drugs are biotransformed by undetermined members of the GI microbiota. It can, therefore, be predicted that a proportion of the failures in compound progression could be due to biotransformation by commensal members of the *in vivo* model host microbiota. The most common methods by which drug molecules are bio transformed and therefore exploited by members of GI microbiota are by reduction and hydrolysis (Figure 1-1; Spanogiannopoulos *et al*, 2016). Reduction sees bonds obtain electrons, which provide alternate acceptors for anaerobic respiration (Figure 1-1A). Hydrolysis is a simple method by which bacteria may cleave a glycosylated bond by the addition of water, freeing sugars needed for bacterial growth (Figure 1-1B).

These interactions are usually detrimental, reducing bioavailability where thresholds are key to efficacy (e.g., digoxin metabolism in ~50% of arrhythmia patients) or re-glycosylation and therefore re-activation of biliary excreted host-generated metabolites (e.g., SN-38 accumulation causing diarrhoea in colorectal cancer patients; Spanogiannopoulos *et al*, 2016). Biotransformation in the GI tract is not limited to reduction and hydrolysis. The metabolically active members of the microbiota exert many types of exploitative activity upon the ingested GI digesta. The pharmaceutical sector often screen potential compounds for biotransformation using human faeces in static or continuous culture systems (Sousa *et al*, 2008).

However, using *ex vivo* methods of assessing microbial biotransformation fails to illustrate the true dynamic properties of complex ecological communities containing coevolved networks of both auxotrophic and functionally redundant community members (Zengler & Zaramela, 2018). More importantly, these methods are conducted post-administration and attempt to illustrate the effect of administration rather than predicting it. A possible method of prediction would be the functional characterisation of the complete GI microbiome by metagenomic and transcriptomic sequence analysis (Garza *et al*, 2018). This would allow the potential for deleterious microbial activity to be screened for prior to administration as a method of screening out carrier or non-carrier models.

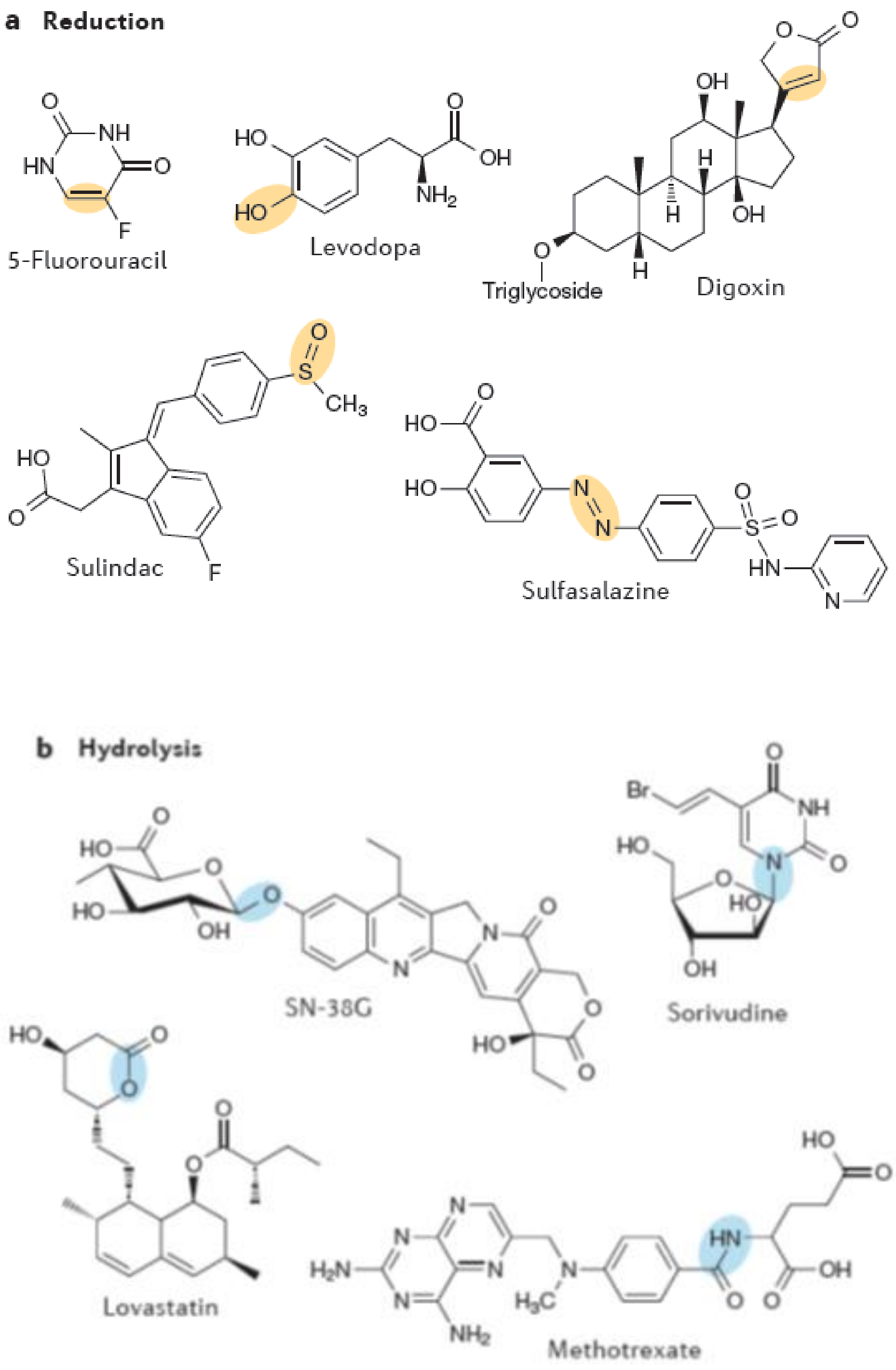


Figure 1-1: Two major classes of prokaryotic biotransformatory reaction. (A) reduction and (B) hydrolysis: with examples of drug compounds known to be acted upon by prokaryotic activity (Spanogiannopoulos *et al*, 2016).

Using characterised models may reduce animal usage and study attrition in the drug discovery process but it would also augment the current granularity of our knowledge of how integrated, inter-kingdom microbial pathways have developed and potentially what role individual community members contribute to the system and the host (Spanogiannopoulos *et al*, 2016).

1.6 Host-microbe interfaces

The varied geobiology of the matured surfaces generated during cephalocaudal folding creates numerous distinct but contiguous environments or niches, providing the opportunity for colonisation by diverse and coevolved microbial species and communities. Although a multitude of niches exist across the external surfaces of the host, these belong to four major environmental regions (Cho & Blaser, 2012), each with its own limiting and symbiotic ecology. The skin, where barrier function and aridity limits microbial diversity but prevents moisture loss (Grice & Serge, 2012), the female reproductive system, where protective diversity is observed to fluctuate with oestrus and pregnancy (Wallace *et al*, 2018). The oral cavity and lungs, where highly diverse communities are firstly found across the multiple irriguous structures (Wade, 2012) and secondly where a microbial community thrives in a once believed sterile environment (Dickson *et al*, 2016), and finally the GI tract, where a voluminous microbial milieu augments the host's catabolic and nutrient scavenging processes (Walter & Ley, 2011). A commonality across all these interfaces is the immunological peace or homeostasis which is generated and maintained between the host immune system and its symbiont partners (Scharschmindt *et al*, 2015). Nowhere is this more evident than the GI tract, where the high metabolic activity of faecal communities offers a continuously produced material for analysis at the site of biotransformation and disease impact in specific *in vivo* models.

1.7 Structure & function of the gastrointestinal tract

The mammalian gut is a dynamic environment and one that has afforded more attention in relation to its constitutive microbiome than any other, possibly due to the ubiquitous nature of its sample material. Its structure has coevolved with its colonising and mutable microbiota since the emergence of coalesced cellular structures with bilateral body forms billions of years ago (Hartenstein & Martinez, 2019). The GI tract is a series of defined structures which run from the anterior mouth to the posterior anus. Each structure allows the sequential breakdown, passage, and finally excretion of food material. Each step in this process allows maximal extraction and absorption of nutrients and energy from the digesta. Its conflicting absorbent and barrier functions necessitate an extremely strong and yet selectively penetrable structure allowing constitutive contact with the immune and circulatory systems (Walter & Ley, 2011). This open system allows the ingestion and colonisation of microorganisms.

It is thought that microorganisms found in the stomach reflect both those found in the oral cavity, being swallowed during mastication along with a highly specific local microbiota. The stomach is maintained at a pH of ~2.5, which enables proteases (e.g., pepsin) to function properly. Rapid transit time and the capacity of the stomach do little to temper the diversity of microorganisms that pass through it, allowing the colonisation of subsequent niches (Walter & Ley, 2011). The small intestine is a less harsh environment, but displays prokaryotic selectivity, seeing colonisation by members of the *Enterobacteriaceae* family and proteolytic anaerobes in the duodenum with a low pH, bile salts and a flood of agglutinating immunoglobulin (Ig)A molecules, inhibiting bacterial proliferation and preventing mucosal penetration (Pabst *et al*, 2016). Both mice and humans have evolved to support hindgut fermentation with the ileocecal valve serving as a barrier between the proximal and large intestine (Nguyen *et al*, 2015), separating the mass of competing bacteria from the initial site of carbohydrate harvest and metabolism in the small intestine giving the host an evolutionary advantage (Walter & Ley, 2011). The colon allows the greater numerical proliferation of prokaryotes by lowering immunological surveillance and therefore prescribed remodelling (Hill & Artis, 2010). A higher pH (~7) and slower transit time also allows increased species diversity and enables stepwise fermentation and more complete energy harvesting to take place. Manifest coevolution has resulted in an environment in which symbiont species prosper and the host benefits from innumerable microbial functions gained by community richness and diversity.

Changes in both geographical and seasonal diet for the host species and the complexity of both the microbiota and its mutable genetic functionality have generated a redundancy in this system where holistically they are interdependent and yet specifically independent (Davenport *et al*, 2014). It is known that gnotobiotic mice can survive in the absence of a formal microbiota but obtain an advantage in one’s presence (Chen *et al*, 2020). This symbiotic system allows each side of the relationship to thrive and multiply. However, each partner in this relationship is immunologically foreign to the other, with benefit only possible with structural separation and immunological surveillance, without which both partners perish.

This functionality is only possible by the presence of multiple specialist intestinal epithelial cell (IEC) types and structures. Almost 80% of the surface of the GI tract is made of enterocytes which have distinctive brushes or microvilli on their anterior surface which decrease in prevalence from the proximal gut to the rectum. These structures are supported by actin filaments, increasing surface area and therefore the absorptive potential of a highly invaginated surface (Figure 1-2A & B; Hill & Artis, 2010).

Enterocytes are arranged with tight junctions preventing paracellular ingress of microbes and express both major histocompatibility complex (MHC)I and MHCII on their posterior surfaces which allow the presentation of antigenic material, macromolecules, and microorganisms for monitoring by the host immune system from within the lamina propria which is measuredly primed by limited mucus penetration and colonisation of the villi (Weseman & Nagler, 2016).

Goblet cells constitute ~5% of the small intestine surface and ~15% of the intestine surface areas. These cells generate mucins which are branched oligosaccharides with hygroscopic and hydrophobic chains.

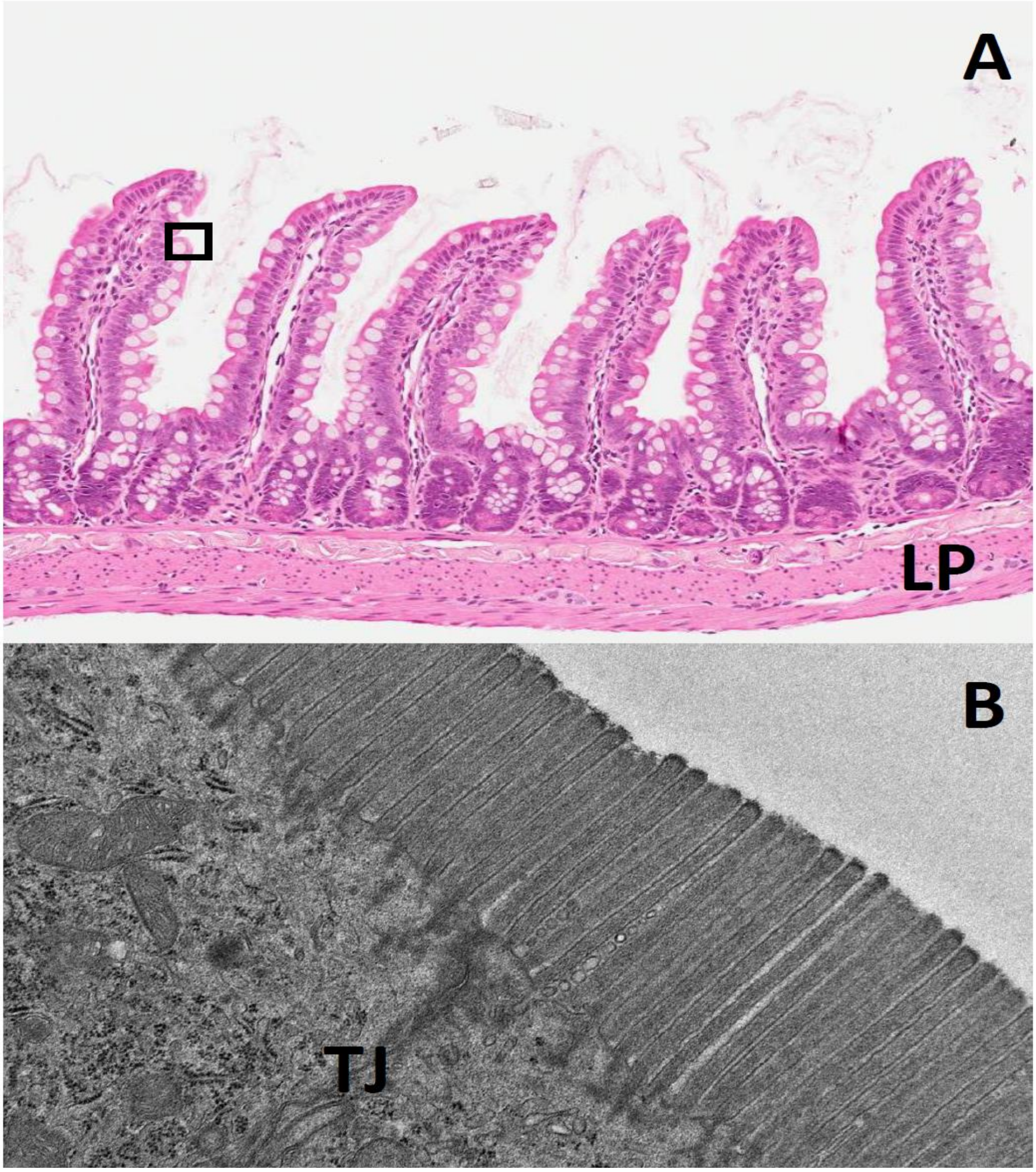


Figure 1-2: Intestinal villi and microvilli of the mouse. (A) haematoxylin and eosin stain image indicating the lamina propria (LP) and (B) an electron microscopy (EM) enlargement of the area indicated by the black square of enterocyte microvilli showing the tight junction (TJ) between cells. Section image supplied by UK Pathology team (GSK) and EM image generated by Ultrastructural & Cellular Bioimaging team (GSK).

This material limits direct contact between host and microbiota. In the densely populated colon, a layer of permissive colonisation is present above a sterile area, where no bacteria are present. Paneth, enteroendocrine and tuft cells each represent about 1% of the GI surface area. Paneth cells are responsible for the secretion of antimicrobial proteins (AMPs) such as α-defensins and lysozyme. These may be constitutionally secreted or stored for triggered release (Allaire *et al*, 2018). Enteroendocrine cells are present in multiple forms and are responsible for the secretion of hormones which control peristalsis, cell proliferation, and mucus secretion. Tuft cells secrete endogenous opioids and are associated with protozoal and helminth detection and reduction (Gerbe & Jay, 2016).

Although the GI tract represents an effective and formidable barrier, this activity in chronic exposure to antigenic and toxic materials requires the constant renewal of the cells described above (Weseman & Nagler, 2016). The complete GI tract is renewed every five days and is done so by the presence of stem cells in the crypts which generate each of the constitutive cell types which subsequently pushes developed cells up into contact with digesta and the native microbiota (Fler & Clevers, 2009).

1.8 Microbial acquisition

It is accepted that exposure to and acquisition of a thriving microbiota begins with parturition, physically driving maternal microbes into the foetus when it passes, most commonly facing the maternal spine, through the vagina. This immediate vertical transfer from the mother establishes an oral and therefore GI microbiota in the neonate which represents the mother's vaginal, skin, and gut communities (Spor *et al*, 2011). Subsequent and prolonged physical contact between offspring and mother continues to play a role in growing external and internal microbial diversity. Neonates display low colonisation resistance, strategically opening them up to both growing diversity and infection (Pabst *et al*, 2016). Although a study on mice indicated that host genotype can impact the abundance of community members, composition depended upon parenthood and occurs via vertical transmission, suckling proximity, and later shared diets (Ley *et al* 2006).

After the initial colonisation with vaginal community members, neonates are rapidly colonised by facilitative anaerobes (*Pseudomonadota*) which rapidly deplete lumen oxygen levels allowing the establishment of obligate anaerobes (e.g., *Bifidobacterium* spp.) which break down milk-based oligosaccharides and drive GI hypoxia and useful colonisation with members of the *Bacillota* and *Bacteroidota* phyla, leading to homeostasis (Byndloss *et al*, 2018). Although these complex carbohydrate polymers provide a substrate on which *Bifidobacterium* spp. may flourish, they also function as soluble decoy receptors, preventing mucosal attachment by all classes of pathogen (Bode, 2012). Recent work shows that the presence of *Bifidobacterium breve* in neonatal mice directly effects the transcriptome of IECs causing the up regulation of genes related to gap junction, tight junction, integrin, and cadherin expression (Kiu *et al,* 2020), Associated cell specificity work indicted that *B. breve* acted upon IEC stem cells of all types, providing early functional training to integral barrier function. Although the secondary pioneers such as *Bifidobacterium* spp. form a core GI microbiota they are eventually replaced or augmented by members from the greater environment with exposure increasing diversity after weaning and into adulthood (Spor *et al*, 2011).

1.8.1 Theories of acquisition

Three theories have been put forward regarding the construction and diversity of community members. Firstly, the deterministic build-up of the microbial communities is thought to occur by the principle that local conditions (pH, water availability and retention time) and innate local biogeography forces environmental pressures on these immigrants (Walter & Ley, 2011). Experiments involving the transplantation of gut microbiota indicate that the donated microbiota rebounds to one representing the usual, native community over time, supporting this deterministic acquisition theory (Rawls *et al*, 2006). Although deterministic acquisition may occur, colonising microorganisms must run a gauntlet of harsh environments to find themselves proximal to a suitable niche along a pyramid of growing diversity descending the GI tract, with the highest recorded population density of microorganisms at its base in the colon (Grice & Serge, 2012). Successful community members must express multiple attributes which allow adhesion, evasion, appeasement, and genetic adaptability. They must also survive the transfer to new hosts at expedient moments such as parturition or excretion (Lay *et al,* 2006). Secondly, it is postulated that a historical build-up of microorganism's shapes community membership, with the first to arrive in a naïve niche repelling subsequent immigrant species. These founder species are thought to hold more sway over community composition than the niche environment by preferentially altering local physiochemical properties. The third theory is neutral, with community membership being completely stochastic. Diversity, in this case, is obtained by random events and therefore may change rapidly over time (Walter & Ley, 2011).

Examples of each theory include *Helicobacter pylori* successfully surviving the local conditions of the stomach but failing to thrive in the colon (Walter & Ley, 2011), evidence of stable founder communities which resist subsequent integration events (Oh *et al*, 2013), and the use of antibiotics throughout life perturbs community membership stochastically with transient and permanent outcomes (Dethlefsen & Relman, 2011). Therefore, laying acquisitional responsibility to just one method of acquisition is unsound as evidence suggests all three are at work throughout the life of a host species but what controls the impact of each type of immigration and subsequent maintenance is a more complex process in which both autochthonous and allochthonous community members fluctuate, effect, and interact with the host, the environment and each other.

The stochastic, historical, and deterministic elements to the acquisition of a mature microbiome are not one-sided. Recent mathematical profiling of community assembly shows that strong dependencies between species may inhibit assembly, and independence and competition drive colonisation (Coyte *et al,* 2021). Assembly is made possible by host feeding, an example of bidirectional contact and evolution with the environment which is in itself dependent on host physiology, genotype, and existing health. Nevertheless, these interactions play as much a part in the evolution and survival of the microbial species as in that of the host (Davenport *et al*, 2017).

1.8.2 Autochthonous & allochthonous

Once established, microorganisms fall into two broad categories, autochthonous and allochthonous. Autochthonous community members are those which are acquired and remain *in situ* at a given niche for significant periods. These organisms may be said to be endogenous i.e., they are commonly found in most host species. Autochthonous species are not readily found in the external environment and are dependent on their specific host for carriage, nutrients, replication, and dispersal among family groups, yet their role may be commensal or parasitic. Commensals may be further split into neutral (one having no effect on the other), beneficial (one species benefits at no detriment to the other) or mutualistic (where both species benefit; Zengler & Zaramela, 2018). Parasitic interactions are usually considered detrimental to one side of the relationship; however, this may be in only one aspect of a relationship. For example, the detrimental loss of nutrients in the GI tract to a parasitic species may be countered by the low level stimulation of the humoral immune system and the subsequent avoidance of an allergic march which is of benefit to the host (Weseman & Nagler, 2016).

Allochthonous community members originate from external sources and are not found in considerable numbers in most hosts. As they are taken up from the environment, they may lack niche specificity but cannot form part of a stable community over prolonged periods of time. At best, these species are transient parasites, giving no benefit to the host, at worst they are overtly pathogenic causing acute disease which leads to their clearance by the immune system or self-induced ejection from the host to continue to infect subsequent hosts (Ley *et al*, 2006). Although these definitions may be of value to define a characteristic at a specific moment, like actors, microorganisms may play many roles in a host's lifetime with pathogenicity or commensalism being merely contextual states (Belkaid & Hand, 2014).

1.9 Function of the gastrointestinal microbiome

The GI microbiota plays an essential role in host health and survival. Three broad and yet overlapping functional roles have been found; dietary, where community members provide a nutritional benefit to the host, physical, where the number and type of microorganism augment the barrier between microbiota and host and finally, in the priming, calibration, and regulation of the immune system throughout the lifetime of the host (Walter & Ley, 2011).

1.9.1 Dietary impact

The leading role of both the neonatal and adult GI microbiomes is the exclusive breakdown of complex carbohydrates (Figure 1-3). Indigestible poly-saccharides (e.g., resistant starches, cellulose, pectins, pentosans, and hexosans), oligosaccharides (e.g., raffinose and lactose) and sugar alcohols (e.g., sorbitol) in the lower GI tract are all acted upon by a prokaryotic consortium (Flint *et al,* 2012). Degradation of these substrates depends heavily upon solubility, polymer linkage and degree of branching or enzymic accessibility. The lack of a host apparatus for this complex activity suggests a long symbiotic coevolution (Haller, 2018).

The initial breakdown of polysaccharides and secondary oligosaccharides depends upon multiple families of glycoside hydrolases, polysaccharide lyases, glycosyltransferases and carbohydrate esterases. More than eighty families of these carbohydrate-active enzymes (CAZymes) have been identified (Bhattacharya *et al*, 2015). The GI microbiome and the *Bacteroidetes* phyla in particular, contains an over-representation of genes associated with this activity, with an inverse number of genes needed for alternative harvesting pathways.

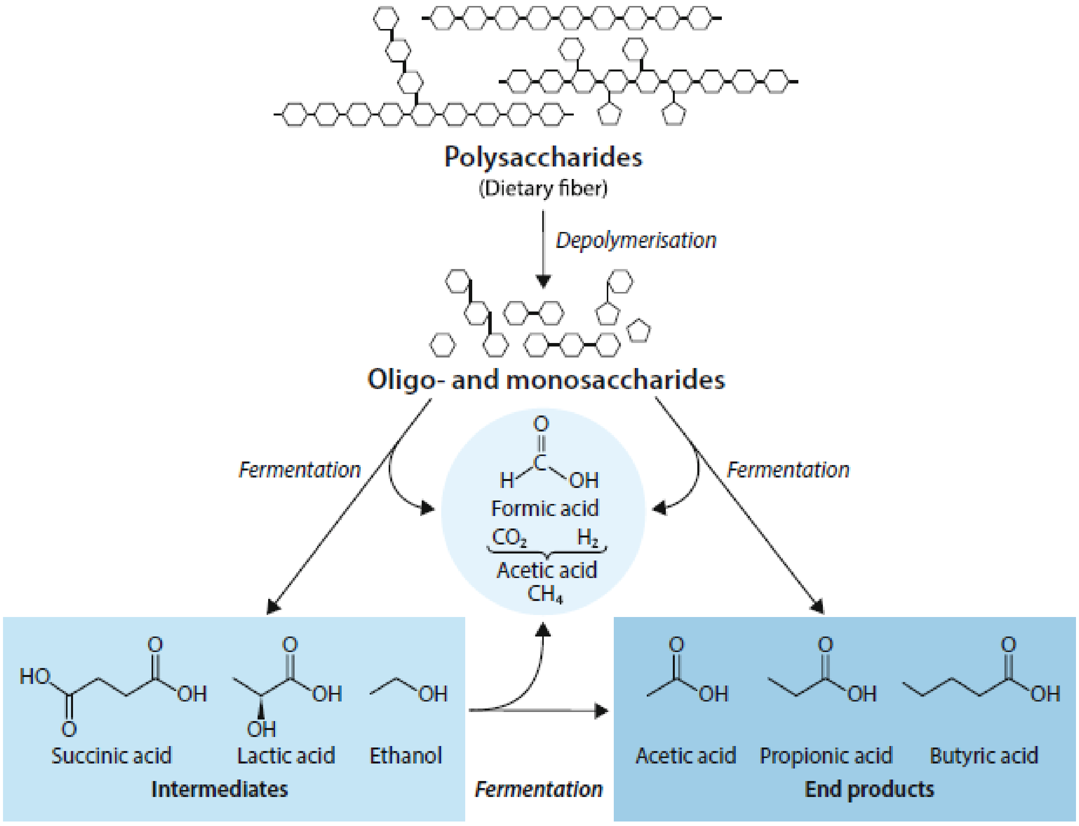


Figure 1-3: The breakdown of complex carbohydrates to short chain fatty acids. Prokaryotes work in concert to depolymerise and ferment complex polysaccharides into short chain fatty acids in the GI tract (Haller, 2018).

The phyla *Actinomycetota* including all *Bifidobacterium* spp. break down oligosaccharides found in breast milk for example (Kiu *et al,* 2020), while members of the *Bacillota* phylum (*Lachnospira* spp., *Ruminococcus* spp., and *Lactobacillus* spp.) break down monosaccharides into short chain fatty acids (SCFAs; Flint *et al,* 2012). Therefore, a diverse bacterial community is able to liberate a wider range of monosaccharides which are further fermented into intermediate molecules such as lactic acid or directly generating the SCFA acetate, propionate, and butyrate (Figure 1-3).

The final rounds of fermentative action towards SCFA generation cyclically decreases local pH which is tolerated by members of the *Bacillota* phylum (Flint *et al,* 2012). However, this view of cooperative networks maybe simplistic with ecological modelling indicating that stability possibly arises from competition, which allows species loss or variability without loss of function (Coyte *et al,* 2015).

Of the total SCFA generated by the GI microbiota, 95% is used by enterocytes, with 70% of their adenosine triphosphate (ATP) requirement being obtained by butyrate-sensor peroxisome proliferator activated receptor (PPAR-γ) activated mitochondrial β-oxidation of microbially derived SCFA (butyrate). This activity consumes oxygen, rendering the epithelial surface hypoxic, driving localised anaerobiosis ensuring obligate anaerobic species can proliferate, converting further complex carbohydrates to SCFA and suppressing inflammation by promoting the expansion of thymus derived regulatory cells (TREGS) (Byndloss *et al,* 2018; Figure 1-4).

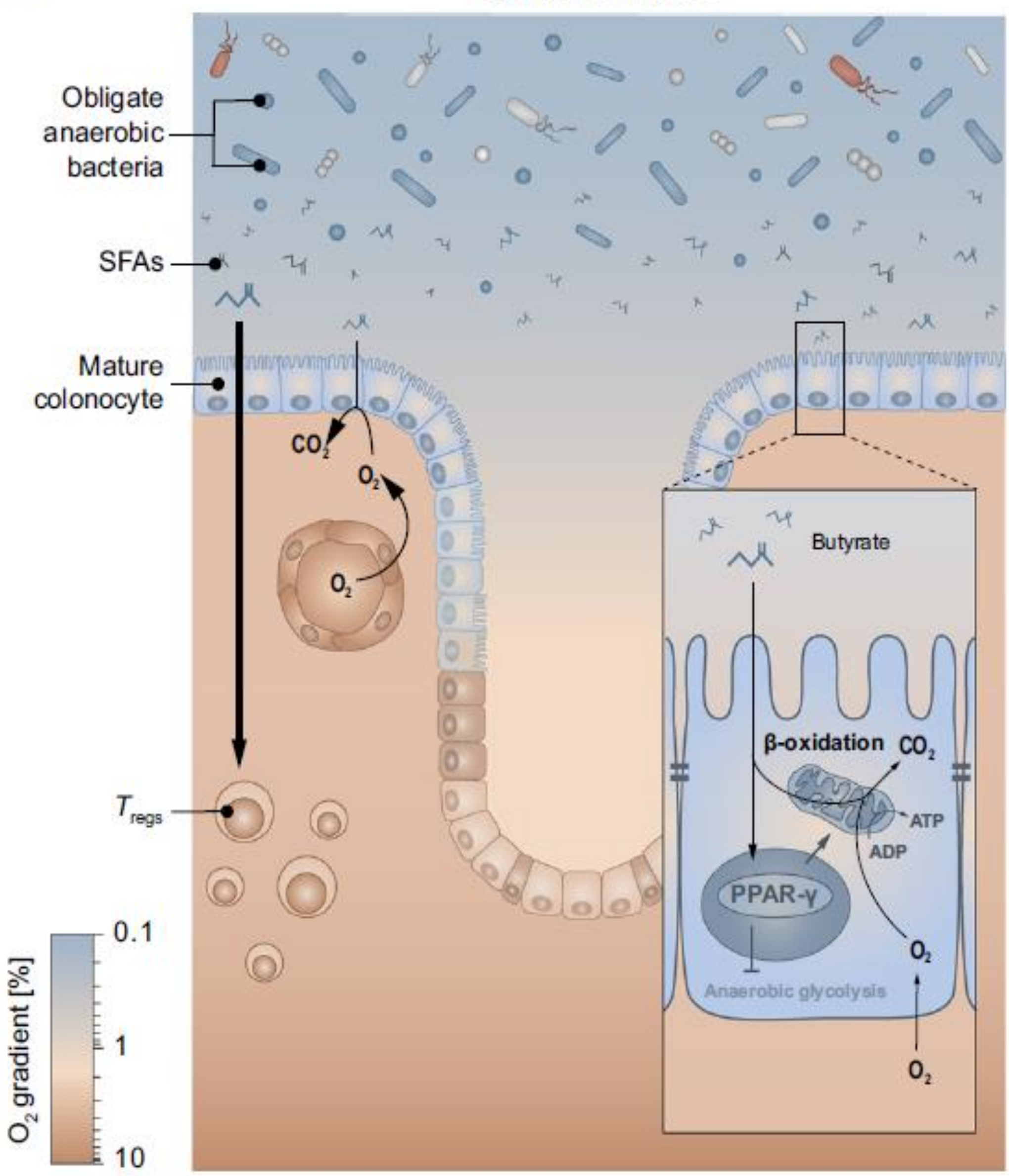


Figure 1-4: Anaerobic homeostasis in the colon. This is achieved by PPAR-γ activated mitochondrial β-oxidation of microbially derived short chain fatty acids produced in the lower GI tract (Byndloss *et al*, 2018).

Prokaryotic energy harvesting by the degradation of complex carbohydrates generates 15% of daily energy requirement for the host (Marchesi *et al*, 2016) while SCFA have been shown to increase solubility of calcium increasing bone heath (Flint *et al*, 2012).

Some gut microbes are overtly responsible for the generation of B-vitamins; thiamine (B1), riboflavin (B2), niacin (B3), pantothenate (B5), pyridoxine (B6), biotin (B7), folate (B9) and cobalamin (B12) and vitamin K. These essential factors are responsible for the catabolism of sugars and amino acids, vitamin activation, multiple metabolic processes, gluconeogenesis, cell growth, DNA synthesis, and blood clotting (Magnusdottir *et al*, 2015).

Many GI bacteria can synthesise complete vitamins, but many depend on the production of precursors for pathway completion. This suggests mutualism among bacterial genera and even phyla. However, evidence that plasma levels of vitamin K are not reflected in their abundance in the GI tract via microbial synthesis suggests a system in which dietary sources can be selectively complemented by microbial production (Karl *et al*, 2017).

The initial immunological selection of what colonises the small intestine shapes community range and overall biomass, allowing the host species to benefit from easily absorbed carbohydrate sources, leaving the digestion of more complex carbohydrates to the fermentative organisms found in the colon (Walter & Ley, 2011). This mutualistic relationship is further entwined with the link between adequate nutritional intake and the functionality of the immune system. It is shown that the mammalian target of rapamycin (mTOR) cell cycling, and proliferation pathway is detrimentally affected by reduced nutrition, which in turn impacts both the innate and acquired immune systems reducing dendritic cell (DC) maturation, T-cell differentiation, memory T-cell formation, and CD8+ regulation and trafficking (Kau *et al,* 2011).

Carbon dioxide and hydrogen are generated by many prokaryotes such as members of the *Enterobacteriaceae* family and the *Bacillota* phylum which are used by other bacteria to generate acetate. A sizable proportion of dietary protein reaches the lower GI tract along with host-derived proteins such as Ig molecules and inactivated proteases. These all represent alternative sources of energy and biosynthetic molecules. It is easier for most prokaryotes found in the colon to gain energy from carbohydrates, but multiple species are capable of proteolysis and therefore nitrogen and carbohydrate harvesting. The step wise breakdown of proteins generates amino acids that are fermented generating the SCFAs, carbon dioxide and hydrogen is often achieved by consortia of species (Haller, 2018). Although bile is not a major source of energy, it is a selective antibacterial agent which may be used by *Lactobacillus* spp., *Clostridium* spp. and *Bifidobacterium* spp. which possess bile salt hydrolases which liberate carbon and nitrogen. Along with structural carbohydrates, ingested vegetable matter contains metabolites which are used by GI bacteria (Haller, 2018). Isoflavones are used by *Adlecreutzia* spp. generating equol which has been linked to cancer prevention (Maruo *et al*, 2008).

1.9.2 Barrier integrity

As ingested food is only present in the stomach for a brief time (<1hr) the bacterial species found in this niche are considered to play no significant role in the digestion of complex dietary matter. They may, however, play a role in mucosal barrier function during the first prolonged contact the host has with food matter and allochthonous organisms (Walter & Ley, 2011). Complex coevolutionary elements become clear with the first point at which bacterial species aid barrier integrity and pathogen exclusion (Bode, 2012). Bifidobacteria acquired during parturition are selected for or fed by the provision of milk oligosaccharides and become dominant during the suckling period in mammals. These polymers constitute one third of milk and specifically and cyclically promote the growth of certain members of the *Actinomycetota* phylum which contain the catabolic gene products necessary for the breakdown of milk oligosaccharides (Yamada *et al*, 2017). This symbiotic relationship reduces the probability of pathogenic infection, diarrhoea and therefore morbidity during the period of initial immune priming and exposure in early life. The same genetic equipment allows the selective breakdown of mucus glycoproteins and augments epithelial adhesion via extracellular polysaccharides and auto-aggregation, further excluding unwanted ingress and generating immune tolerance (Hiipala *et al*, 2016).

Specific sculpting of prokaryotic species and niche community compartmentalisation is made possible by the glycosylation of mucin glycoprotein family members (MUC1 to 21). This wide group of host molecules are expressed in varying numbers at specific regions of the GI tract. They are divided into secreted, or membrane associated mucins. Secreted mucins play a role in the physical barrier and mobility of the GI tract whereas membrane associated members contribute to the rich cellular carbohydrate display which directly interacts with prokaryotic communities. Their preferential binding via a diverse range of carbohydrate binding molecules and other lectin-associated motifs acts as a highly evolved system of ensuring symbiotic preference along the GI tract throughout the life of the host (Corfield, 2018). This interaction provides barrier protection via weight of numbers, only if exposure and subsequent feeding occurs (Coyte *et al*, 2015), linking diet, barrier, immunity, and homeostasis.

The post-weaning increase in diversity is partly due to a widening of diet and the removal of milk oligosaccharides. This does not exclude the need for bacterial protection from deleterious microorganisms during primary enzymic energy recovery in the small intestine (Walter & Ley, 2011). The immediate niches below the stomach are maintained with low pH and are the site of primary innate immune activity. Here, lactobacilli breakdown simple sugars forming lactic acid which further promotes an acidic, anti-microbial environment (Porter & Martens, 2017). This niche is a fast moving, liquid environment which allows the absorption of exposed nutrients and minerals from the diet via the numerous microvilli found upon local IECs. Non-motile bacteria such as lactobacilli are adapted to adhere to mucin (Nishiyama *et al,* 2016). As fast as it is generated and passes through the area lactobacilli replicate and colonise new mucus from the goblet cells. Specific species of these bacillota contain the genes for specific adhesive capabilities, allowing comprehensive and diverse colonisation during life and dietary alterations, maintaining barrier function (Porter & Martens, 2017). These adhesion factors can be either cell-wall anchored (e.g., microtubule associated proteins), or multifunctional (e.g., glyceraldehyde 3-phosphate dehydrogenase) factors. In addition to this species variability in adhesion, *Lactobacillus* spp. are able to rapidly switch transcription pathways in the presence of alternate carbon sources (Nishiyama *et al*, 2016). These factors allow continued host pre-eminence via the activity of selected bacterial species.

1.9.3 Immune conflict, calibration, and control

Surveillance of this barrier is imperative in the maintenance of autochthonous species and the eradication of allochthonous pathogens for host survival (Hill & Artis, 2010). Along with microbial signatures, the host immune system has to content with responding to a plethora of xenobiotic signals from ingested diet. This balance is only possible with the combined activity of the innate and acquired immune systems (Artis, 2008). The complexity of this surveillance and control mechanism necessitates 70% of host immune system is focused on the GI tract, constituting the gut associated lymphoid tissue (GALT). The cells of the GI tract directly recognise microbial ingress by pattern recognition receptor (PRR) binding to innumerable non-host pathogen-associated molecular patterns (PAMPS). These include lipopolysaccharide (LPS), lipoproteins, peptidoglycan, flagella, and dsRNA. PRRs such as Toll-like receptors (TLRs) initiate downstream cellular and systemic responses to microbial interactions (Artis, 2008). Enterocytes also initiate such responses indirectly by detecting damage to the cellular matrix. Microfold or M-cells present microbial and dietary antigens to macrophages and dendritic cells which reside below the epithelial layer in the lamina propria. Along with their own expression of pro-inflammatory signals, these phagocytic, antigen presenting cells prime the acquired immune system via transfer to the mesenteric lymph nodes where they interact with T-cell populations leading to classical T_h1 cell-mediated or T_h2 humoral responses (Ost & Round, 2018). This archetypal, combative immune ramping must be kept in check at the interface of host and microbiota for both to survive by attaining homeostasis. Immune tolerance of symbiotic prokaryotes is driven from both sides. Not only do commensal species provide nutrients and aid barrier maintenance, but they also paradoxically ameliorate immune responses by close physical contact and release of metabolites (Artis, 2008).

Anaerobic colonic bacteria ferment non-digestible plant carbohydrates (e.g., cellulose, xylans, starch) to generate non-carbohydrate SCFAs such as acetate, propionate, and butyrate. Along with SCFAs providing energy to the enterocytes via neoglucogenesis which is the ubiquitous process of glucose synthesis from non-carbohydrate substrates (LeBlanc *et al*, 2017), they are shown to suppress a wide range of specific inflammatory elements and pathways. Acetate stimulates the proliferation of cell generation in GI crypts, inhibit nuclear factor (NF)-κB transcription cascades in colonic enterocytes where it also suppresses pro-inflammatory interleukin (IL)-6 expression. Acetate also serves as an intermediate substrate for bacteria to breakdown into butyrate. Propionate also inhibits NF-κB via guanosine triphosphate (GTP)-binding proteins in immune cells, inhibits LPS-induced tumour necrosis factor (TNF)α production of pro-inflammatory cytokines and is also used for butyrate generation (Flint *et al*, 2012).

Butyrate is the core energy source for enterocytes but also reduces macrophage IL-8 expression, stimulates mucin production in goblet cells, effects tight junction proteins affecting barrier function and permeability. Butyrate is also responsible for TREG proliferation and therefore downregulation of effector T-cell populations (LeBlanc *et al,* 2017). Along with water-soluble SCFAs, commensal bacteria express a range of membrane-bound effector molecules. These include *Bacteroides* spp*.* polysaccharide-A which promotes tolerance by activating TREGs, which suppress T_h17 responses, and *Roseburia* spp. flagellin which induce the upregulation of anti-inflammatory IL-22 expression and suppression of pro-inflammatory IFNγ and IL-17 (Pandiyan *et al,* 2019). These activities rely upon colonisation and penetration of the mucus layer so that close contact can be made between the host and the commensal species. At this juncture, immunoglobulins become pivotal in the control provided by the acquired immune system. IgA is the primary serotype secreted in the GI tract (Bunker & Bendelac, 2018). It is expressed from plasma cells throughout the lamina propria along with bile from the hepatic portal. Its full function here in controlling the microbial biomass is not completely understood. However, IgA selectively binds to a range of microbial species in the mucosal layer and this strain specific immobilisation may help with the removal of unwanted species and conservation of beneficial commensals (Pabst *et al*, 2016). IgA selective pressure on bacteria is found to drive surface antigen diversity in commensal species, augmenting resistance by pushing the perpetual expression of alternate surface epitopes (Peterson *et al,* 2007). This exquisite control by extensive elements of the immune system does not dampen responses blindly but calibrates pro and anti-inflammatory processes according to the state of multiple factors. Without this control, host and commensals would jointly fall, its success again points to a long coevolutionary mutualism.

1.10 Dysbiosis, perturbation, and disease

The elegant balance of symbiotic functionality and limited immunogenesis is a flexible relationship which necessarily concedes change (e.g., host age, diet, and microbial diversity) over time without penalty. However, chronic, and acute changes to the local environment can affect the harmony and subsequently influence the resilience of the host and the microbiota. These malformed and often less heterogenous community structures are said to be in dysbiosis. While this terminology is often contested, this remains qualitatively and quantitatively demonstratable (Byndloss *et al*, 2018; Levy *et al*, 2017). Localised anaerobiosis is generated by the uptake of SCFA by enterocytes.

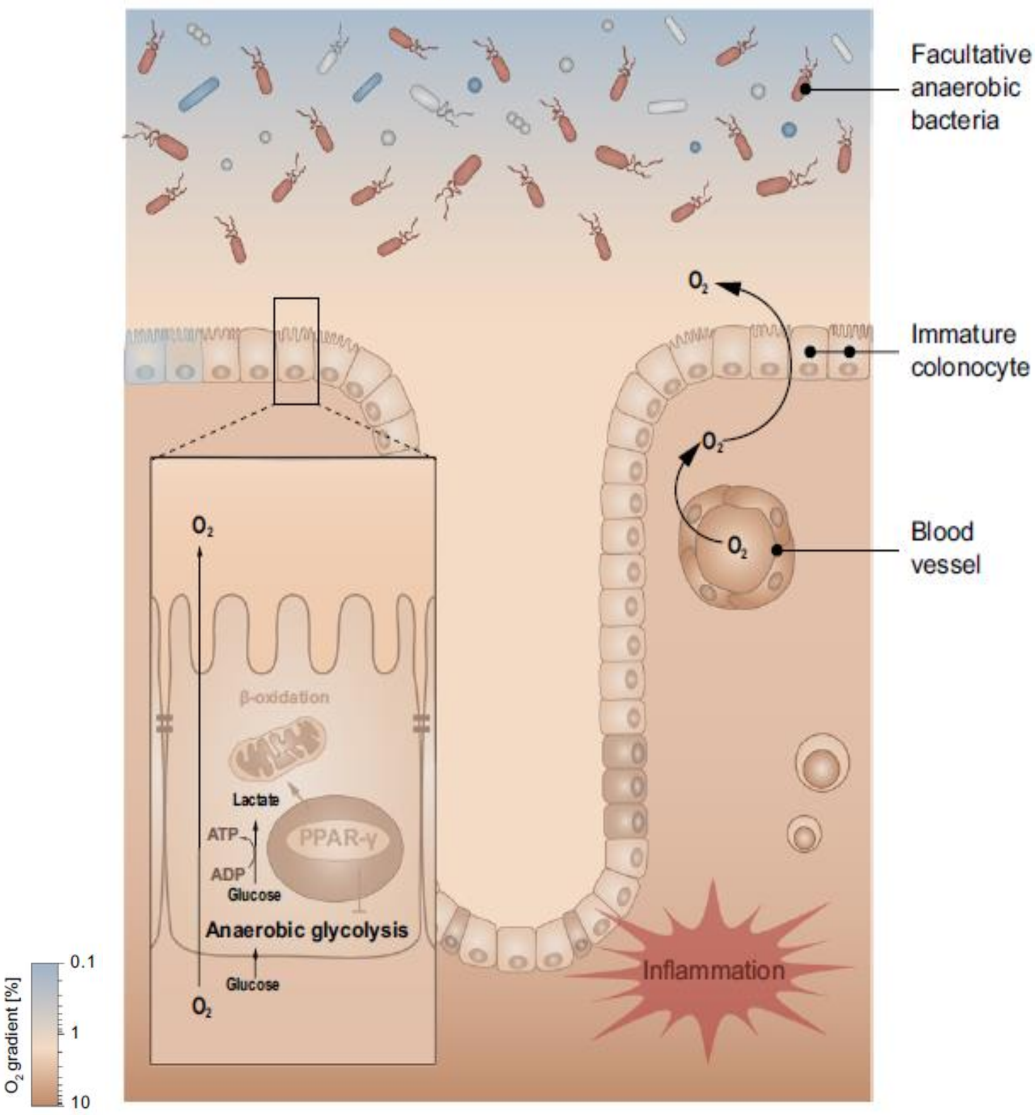


Figure 1-5: Dysregulation of anaerobic homeostasis. The dysregulation of PPAR-γ activated mitochondrial β-oxidation of microbially derived short chain fatty acids produces dysbiosis in the lower GI tract (Byndloss *et al*, 2018).

In the absence of complex carbohydrates, levels of SCFA are reduced, resulting in the proliferation of facultative anaerobes such as the *Enterobacteriaceae* causing anaerobic glycolysis in the enterocytes releasing oxygen, further preventing the growth of catabolic species, increasing local inflammation, and reducing barrier integrity (Byndloss *et al*, 2018) (Figure 1-5).

Conversely, increased carbohydrate intake, over time is strongly associated with weight gain in Western communities when compared to rural communities on a more restrictive diet (Martinez *et al*, 2017), likewise feeding a polysaccharide-rich diet to *ob/ob* mice sees shift in community structure towards dominance by *Bacillota*, the core carbohydrate harvesting gastrointestinal phyla, away from the *Bacteroidota* which are often associated with host leanness (Ley *et al*, 2005).

Overt chemical sculpting of prokaryotic communities by the administration of antibiotics often allows colonisation of niches by non-native species due to bystander expansion. The longevity of compositional alterations is shown to differ widely between individuals, with just two doses of ciprofloxacin temporarily altering diversity for ~8wks in some and permanently changing composition in others (Dethlefsen & Relman, 2011). Inter-kingdom, off-target consequences can be seen with the rise of resistance genes encoded in the genomes of bacteriophages post-administration which are not limited to the class of antibiotic used, indicating an ever-widening ripple in diversity and gene activity caused by a single antibiotic intervention (Modi *et al*, 2014). Once basal community structures are chemically altered, keystone symbiotic processes or control mechanisms are interrupted. *Clostridioides difficile* infections (CDI) display this effect, being the highest cause of healthcare associated infections, leading to ~30,00 deaths per annum in the USA (Peng *et al,* 2017).

Those at highest risk from CDI are in-patients, >65yrs old which have already undergone antibiotic treatment. *C. difficile* is able to survive this primary chemical intervention as it is a spore former, which allows its prolonged dissemination in an enclosed environment such as hospital. Secondly, it expresses biofilm associated proteins when presented with antibiotics, but *C. difficile* strains are often 100% resistant to first, second and third generation cephalosporins and fluoroquinolones. Primary treatment for known CDIs is metronidazole and vancomycin but by 2012, ~15% of strains were resistant to metronidazole creating an antibiotic-driven, therapeutic dead-end as protective members of the *Bacillota* and *Bacteroidota* phyla are removed (Peng *et al,* 2017). *C. difficile* strains may be toxigenic or non-toxigenic. Pathogenicity is dependent on the presence of one or both toxins possibly expressed by *C. difficile* (TcdA and TcdB), both found to be expressed in reaction to high levels of SCFAs (Gregory *et al*, 2021). Both of these toxins inactivate GTPases through glucosylation of a specific threonine residue, leading to polymerisation of actin and cell death. The subsequent inflammatory response increases tissue damage leading to diarrhoea and pseudomembranous colitis (Burke & Lamont, 2014).

In the IBD states ulcerative colitis (UC) and Crohn’s disease (CD), genetically susceptible individuals experience episodic manifestations of diarrhoea and discomfort and weight loss. Attacks are attributed to autoimmune responses to commensal bacteria which lead to population eradications and persistent translocations of maladapted species from inflamed niches.

Atopic diseases such as food allergy, asthma, and dermatitis which stem from hyper-reactivity, rather than mis-activity, at mucosal surfaces, originate from alterations in normal tolerance levels driven by a chronic decline in microbial diversity (Haller, 2018).

Typically, interactions between host and the commensal microbiota from birth, aid neurodevelopment with germ-free mice developing fewer neurons than their diversely populated counterparts, subsequently displaying reduced sociability and increased anxiety-like behaviours. A critical contact window exists for the codevelopment of a symbiotic microbiota alongside neuronal restructuring immediately after birth. During this period, normal functioning can be restored by faecal transplant. Normal functioning involves endocrine, metabolic, and immune pathways, directly influencing disease progression, behaviour, and longevity (Warner, 2018). Dysbiosis and disruption of this gut-brain axis also result in structural changes to multiple sites in the brain, indirectly influencing the behaviour and survival of the host (Vuong *et al,* 2017). Classical dysbiosis caused by the administration of single pathogenic bacteria (e.g., *Campylobacter* spp. or *Citrobacter rodentium*) rapidly induces anxiety-like responses in mice due to direct signalling via microbial peptides or activation of host receptors in the brain (Bravo *et al*, 2012). Conversely, anxiety can be reduced in autistic patients (Haller, 2018) and exploratory drive can be increased in mice (Bravo *et al*, 2012) upon the administration of antibiotics removing specific classes of bacteria from the microbiota.

If the modulation of specific members of the GI microbiota results in physical and psychological changes in the host, it stands that these changes and therefore disease states may have diagnostic microbial signatures. A meta-analysis of faecal metagenomic data of colon cancer patients indicated an increased prevalence of prokaryotic genes involved in protein and mucin catabolism and a corresponding decrease in genes needed for carbohydrate degradation suggesting that a diet rich in fat and protein tips the microbiota into dysbiosis, increasing the risk of this specific pathology (Wirbel *et al*, 2019). The integrated mechanisms necessary for health and survival are clearly derailed in many diseases. Not only are the effects of dysbiotic community structures implicated, but they can also now be mapped and used as diagnostic markers. Prokaryotes may be the biochemical drivers behind altered processes, but the full strata of the microbiota are both implicated in dysbiosis but affected by it.

1.11 Microbial strata

Although the microbiota can be divided into allochthonous and autochthonous, and there appears to be common acquisitional and anchoring processes in mammals, the human microbiome differs more between individual than within an individual over time (Gilbert *et al*, 2018). Within individuals, each ecological niche is inhabited by a complex mix of microorganisms across all kingdoms of life which have coevolved. The microbial strata, whose gene compliment equates to the full microbiome includes prokaryotes (bacteria and archaea), eukaryotes (fungi and protozoa) and viruses (of both eukaryotic and prokaryotic organisms). It is this full complement microorganisms which has coevolved with the mammalian host, driving health and disease equally (Cullen *et al*, 2020; Walter & Ley, 2011).

1.11.1 Eukaryotes

The domain of the *Eukaryota* is divided into four kingdoms: *Plantae*, *Fungi*, *Protista,* and *Animalia*. It is in this last group that the class *Mammalia* are found. The mammalian GI tract plays host to members of all four eukaryotic kingdoms along with those of the bacterial and archaeal domains. Often the role of the host is to allow forward transmission of transient species. However, microbiologically, the GI tract of mice may constitutionally contain parasitic invertebrates, protists, and yeasts identifiable by coprological examination, filtration, and microscopy depending on the stage of their life cycle (Theinpont *et al*, 1986). However, just as with the prokaryotes, rRNA gene specific PCR can be used to identify this class of organism, by amplifying and analysing the 18S rRNA gene (Woese & Fox, 1977). Fungal species may also be haphazardly observed by culture and microscopy but their role in the microbiota in anything, but acute disease states is little appreciated due to the lack of a parallel method of their joint isolation. However, evidence gleaned from studying patients with primary immunodeficiencies that bacterial, protist and fungal populations (and therefore interactions) are intrinsically linked by the immune status of the host (Oh *et al*, 2013).

1.11.2 Prokaryotes

The classification of organisms according to their primeval rRNA subunit gene sequences in the 1970s saw the division of life on Earth into three domains, the *Bacteria*, the *Archaea* and the *Eukaryota* (Woese & Fox, 1977). This distinction has pervaded phylogenic organisation and diagnostic techniques ever since. The use of full-16S rRNA gene specific PCR primers to distinguish prokaryotic species has become the blueprint for their discovery (Fox *et al*, 1999). This method of identifying or classifying bacterial species continued into the first work of the Human Microbiome Project Consortium (HMPC) (NIH, 2009) and ignited the cataloguing of bacterial communities. This method is ideal for understanding which species are present in a sample but the use of 16S rRNA gene primers can be biased, necessitating the development of many universal oligonucleotides to account for the variability of this gene target (Eaton *et al*, 1996). The work of the HMPC has formed the basis of the theoretical understanding of the scope of the microbiome and the technical and formal basis for its investigation (Cho & Blaser, 2012). Although the bacterial domain dominates the microbiota of the mammalian host, the continued application of taxonomic studies by 16S rRNA-designation possibly promoted a non-inclusive view of the inter-Kingdom reality of the microbiome, lessening the understanding of the complex nature of all ecological niches (Handley *et al*, 2012). Members of the *Archaea* constitute the other free-living prokaryote domain but until relatively recently the general view was that these organisms were largely unculturable extremophiles (Robertson *et al*, 2005). Culture independent methods now show a much wider distribution, comprising

~10% of total GI microbiota, existing in syntrophic relationships with other microorganisms in anaerobic niches, which supports their hydrogen-based energy metabolism. The archaea are not thought to be linked to overt disease in man but may play a role in multifactorial diseases such as periodontal disease and therefore are associated with endocarditis, stroke, atherosclerosis, and preterm delivery of infants. The severity of these diseases has been associated to the relative abundance of archaea in the oral cavity (Lepp *et al*, 2004). The presence of methanogens and methane generation has also been linked to pathology in UC and CD but their presence in patients may be more closely related to retention time of excreta than a result of a disease state (Aminov, 2013).

1.11.3 Eukaryotic viruses

The most overtly allochthonous strata of the mammalian microbiota is the virome. With no metabolic activity, this group falls outside the classical kingdoms of life (Woese & Fox, 1977). They are classified by the Baltimore system which assigns them to one of seven groups according to the various means by which they synthesise mRNA for eventual replication (Baltimore, 1971). This system mirrors their varied genomic structures, being RNA or DNA, single stranded or double stranded, and positive or negative sense. Additionally, they may possess a lipid envelope obtained from their last host or be non-enveloped, expressing just a protein capsid at cellular exit (Kudesia & Wreghitt, 2009). Viruses must straddle intracellular, extracellular, and external environments in order to replicate in a permissive host and pass through the external environment to gain a foothold in the next host (Dennehy, 2014). The lack of proof-reading inherent in viral RNA-dependant RNA polymerase and retroviral reverse transcriptase drives mutation and evolution in RNA viruses (Smith, 2017), while DNA viruses obtain and use host genes in a less mutable framework necessary for host gene mimicry (Tortorella *et al*, 2000). This genomic mosaicism and morphological variability indicates that they possess no common genetic element or framework by which to align and distinguish these microorganisms such as the 16S rRNA gene in prokaryotes. However, DNA and RNA metagenomics can be applied to interrogate their diversity or richness in an ecological niche (Zuo *et al*, 2021). As with fungal infections, their presence and diversity often reflects the immune status and presiding health of their host, which is shown to strongly affect community membership (Handley *et al,* 2012; Zuo *et al*, 2021).

This transient, subclinical flux of viruses in immunocompetent hosts masks their ubiquitous presence with diagnosis or isolation of viral nucleic acid previously depending upon serendipitous discovery (Karst *et al*, 2003). This has made the appreciation of the virome difficult, however, DNA and RNA metagenomic sequencing has increased the number of viral genomes submitted to databases such as GenBank which has concurrently increased the potential for their inclusion in studies of the full microbiome (Norman *et al,* 2014). By routinely considering the virome, their role in complex ecological niches will become clearer and the dark regions of the microbiota which may have obscured the etiological agent of ~40% of cases of human diarrhoea and significant mortality may be uncovered (Finkbeiner *et al*, 2008).

1.11.4 Prokaryotic viruses

The final strata of the microbiome are the least understood, that of the prokaryotic phages. They are the most numerous biological entities on Earth and yet until recently little thought of their impact on prokaryote populations has been considered again due to the lack of genome references (Shkoporov & Hill, 2019). The role of the phages in the development of sequencing and microbiology is paradoxical. Phages ØMS2 and ØX174 were the first complete genome sequences to be fully characterised in the mid-1970s (Fiers *et al,* 1976; Sanger *et al*, 1977), however, the focus of sequencing moved swiftly to whole bacterial genomes and then complete communities. Phage diagnosis continues to be used although it is biased towards those phages which cause lysis and the culture of permissive bacterial species. Phages also exist in non-lytic forms, becoming stable additions to the host genetic material for many generations by plasmid formation or integration. They are found in various structural forms pleomorphic, as filamentous rods or icosahedral with diagnostically indicative tails. However, morphology is shown not to be related to genomic phylogeny (Shkoporov & Hill, 2019). They can be identified by electron microscopy but as with microscopic examination for protists and parasites, phage density in a sample can affect their detection. To complicate matters more, both bacteria and archaea play host to divergent species of infective phage (Abedon, 2008). The lack of sequence data and therefore a framework on which to hang phage genomes has until recently been the block to investigating these microorganisms in isolation or as part of strata wide metagenomic studies. There were <500 phage genome sequences available in 2008, indicating the resurgence of interest in phage genomics in line with the increased application of NGS (Shkoporov & Hill, 2019). However, there are now thousands of phage genome sequences available for mapping and comparison on the IMG/VR database (Paez-Espino *et al*, 2017), making their inclusion into true microbiome studies now possible. Their numerical superiority in the environment and the ubiquitous nature of their hosts means they have the potential to alter whole ecosystems and necessitating their inclusion in characterisation studies if possible (Norman *et al*, 2014). This periodic alteration in niche community membership is an example of a stochastic selection event affecting historic community members, making space for new immigrants or the expansion of previously repressed member species which are resistant to infection and destruction. The examination of equine faeces using plaque formation indicated the presence of sixty-nine distinct phage species (Letarov & Kulikov, 2009), while none were found in captive murine samples (Kasman, 2005). These incongruous data indicate the need for culture-independent investigations into phage diversity in vertebrates. No attempt has yet been made to integrate phage detection into routine microbiology, perhaps due to the lack of a unifying method for their detection or classification but also perhaps that until recently it was difficult to appreciate or disentangle the complex genetic storm which revolves around diverse phage genomics and the relationships, they have with their hosts which drives antibiotic resistance and exemplifies horizontal gene transfer events (Shkoporov & Hill, 2019).

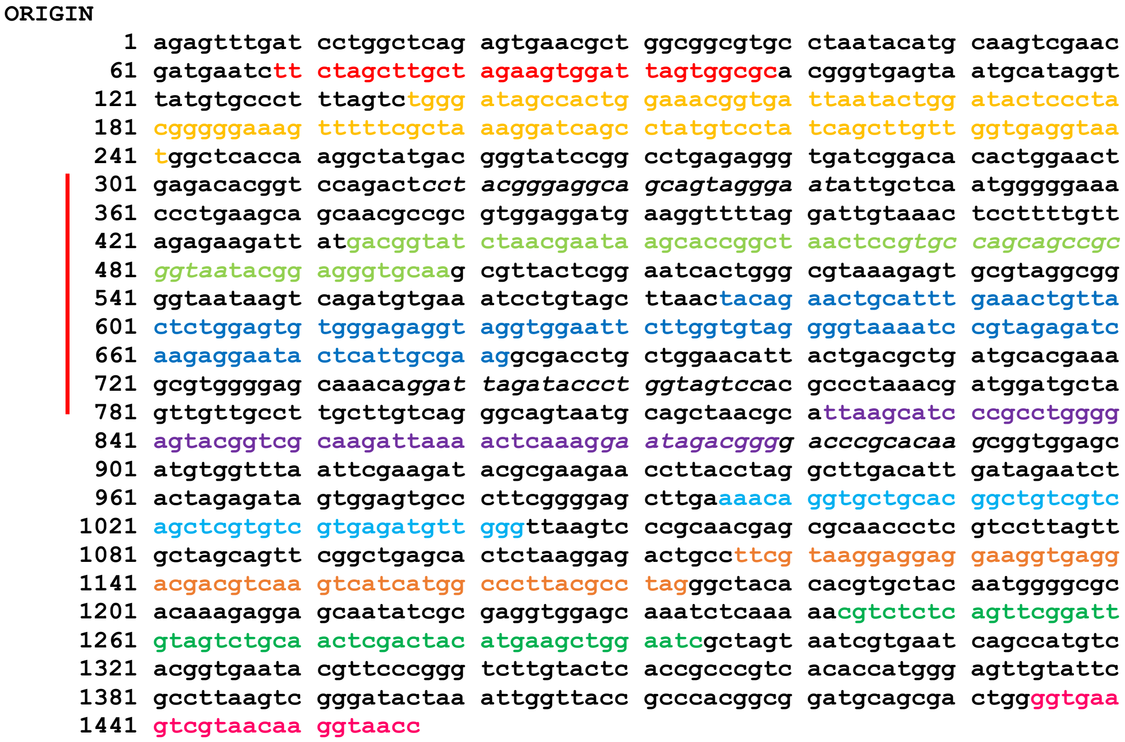
1.12 Characterisation

Originally described by van Leeuwenhoek by visual phenotype then later by the likes of Pasteur by biochemical properties, prokaryotic taxonomy has historically been driven by phenotype as illustrated in *Bergey's Manual of Systematic Bacteriology* (1994) and *Cowen & Steel's Manual for the Identification of Medical Bacteria* (1993). Concurrent with earlier editions of these texts, was published the first genetic description of the phylogeny of life based on dideoxy-characterisation of rRNA sequences (Woese & Fox, 1977). This was made possible by employing Sanger sequencing, which maps the random incorporation of individually labelled chain-terminating, dideoxynucleoside triphosphates (ddNTPs) in an extending product up to ~500bp (Sanger *et al*, 1977). Oligos are visualised along a capillary gel where each terminating base is detected, and a sequence strand mapped across the multiple channels. Initially it was employed to sequence short phage genomes, however, the development of PCR (Mullis *et al*, 1986) allowed the *in vitro* amplification of specific sequences of DNA, which accelerated the scope and speed of sequencing. This combined use of Sanger sequencing and PCR enabled the reciprocal development of the International Nucleotide Sequence Database Collection (INSDC) member databases; National Centre Biotechnology Information (NCBI), European Molecular Biology Laboratory

(EMBL), and DNA Databank of Japan (DDBJ). The 16S rRNA gene of *E. coli* was first sequenced and deposited in the NCBI database in 1983 (J01695). One third of all subsequent NCBI accessions are prokaryotic 16S rRNA gene sequences. However, only ~20% of these sequences are full 1500bp reads (Schloss *et al,* 2016). It is this weight of work that has propelled the 16S rRNA gene as the central biomarker in a post-phenotype taxonomic system, providing a more accurate illustration of evolutionary association and even of genetic mobility. The 16S rRNA gene is ~1540bp in length, with nine variable regions (V1-9) flanked by conserved areas. It is found in various copy numbers across all prokaryotic genomes, with up to fifteen copies per cell (Klappenbach *et al*, 2001). The 16S rRNA gene has a 67% base-pairing potential which allows an essential secondary structure to form in complex with a small number of scaffold proteins. This generates the small ribosomal subunit which then coalesces with the 23S ribosomal subunit, to create the functional ribosome, essential for protein synthesis or translation of mRNA into polypeptides in the cytoplasm. Its essentialness in bacterial replication, endows the 16S rRNA gene with a functional stability across time with a higher mutation rate in the less essential variable protrusions (Yarza *et al*, 2014). It is from these mutating regions that evolutionary relatedness can be traced, and phylogeny mapped. This explosion of genetic information led to the eventual adoption of genomic divergence as the foundation of prokaryotic classification.

The inherent specificity of this PCR-based work still made ecological profiling impossible. However, these small steps opened the path to advances in massively parallel sequencing or NGS (Yarza *et al*, 2014). This novel method co-amplifies material of multiple origins at once, repeatedly recording millions of nucleotide incorporation and base-calling events. Multiple NGS chemistries exist (ION Torrent, Pyrosequencing, Nanopore etc.), however, bridge amplification has proved the most successful method of sequencing. This method developed in 1997 (Glaxo Wellcome, 1998) has been adopted by Illumina as Sequencing by Synthesis (SBS) which presently accounts for ~90% of sequencing activity (Illumina, 2016). This hybridisation method employs index linkers to attach amplicons to a flow cell coated in a forest of complementary oligos. The free end of the tethered DNA strand then loops over forming a bridge to the alternate oligo. Enzymic extension occurs over this bridge using fluorescent labelled dNTPs. Single base incorporations are captured at hundreds of millions of points across the cell until extension reaches the 3’ end of the bridge, at which point the template is cleaved and the process begins again (Glaxo Wellcome, 1998). The commercial availability of NGS at the beginning of the 21st century, quickly saw the 16S rRNA-amplicon based approach enflower phylogenetic trees, allowing all prokaryotic species to be distinguished, mostly to genera level, in an environmental sample and was quickly applied to gauge the composition of bacterial populations from every conceivable environmental niche (Joval *et al,* 2016). From the study of the taxonomic relationships based on 16S rRNA gene similarity it has been possible to quantify what constitutes each level of the taxonomic ladder. Phyla can be differentiated by >75% dissimilarity, whereas class is defined by 78.5%, order 82%, family 86.5%, genera 94.5% and species by >98.7% similarity (Yarza *et al*, 2014). By this reckoning, 16S rRNA gene based NGS experiments should be able to define all prokaryotic entities and unravel previously hidden community structures. NGS is capable of hugely increasing the illustrative resolution when measuring diversity but what it fails to do is achieve this separation in conjunction with a universal discriminatory power or focus. This is because millions of short, variable-region reads, trade clarity with numeric superiority. The choice of what variable region or fragment (or combination) to use is key in the clarity of the resulting data (Klindworth *et al,* 2013). These fragments are the nine hypervariable regions (HVR) in the prokaryotic 16S rRNA gene, designated V1-9 (Figure 1-6). These are used to taxonomically categorise isolates, either singly or jointly. However, no single region can differentiate all bacteria and archaea to a sufficient level of identification (either genus or species). It is shown that V1 can distinguish the streptococci and staphylococci well; while V2, V3 and V6 can differentiate most species apart from members of the *Enterobacteriaceae* family. Regions V4, V5, V7 and V8 fail to discriminate singly any species (Chakravorty *et al*, 2007).

Each HVR is flanked by conserved regions which are used for universal PCR primer lift-off points (Eaton *et al,* 1996). The absence of a single diagnostic region in the rRNA gene necessitates the use of multiple HVRs to increase specificity. Unfortunately, this short-read length extrapolation may introduce bias (Sharpton, 2014).



```
ORIGIN
        1 agagtttgat cctggctcag agtgaacgct ggcggcgtgc ctaatacatg caagtcgaac
       61 gatgaatctt ctagcttgct agaagtggat tagtggcgca cgggtgagta atgcataggt
      121 tatgtgccct ttagtctggg atagccactg gaaacggtga ttaatactgg atactcccta
      181 cgggggaaag tttttcgcta aaggatcagc ctatgtccta tcagcttgtt ggtgaggtaa
      241 tggctcacca aggctatgac gggtatccgg cctgagaggg tgatcggaca cactggaact
      301 gagacacggt ccagactcct acgggaggca gcagtaggga atattgctca atgggggaaa
      361 ccctgaagca gcaacgccgc gtggaggatg aaggttttag gattgtaaac tcctttttgtt
      421 agagaagatt atgacggtat ctaacgaata agcaccggct aactccgtgc cagcagccgc
      481 ggtaatacgg agggtgcaag cgttactcgg aatcactggg cgtaaagagt gcgtaggcgg
      541 ggtaataagt cagatgtgaa atcctgtagc ttaactacag aactgcattt gaaactgtta
      601 ctctggagtg tgggagaggt aggtggaatt cttggtgtag gggtaaaatc cgtagagatc
      661 aagaggaata ctcattgcga aggcgacctg ctggaacatt actgacgctg atgcacgaaa
      721 gcgtggggag caaacaggat tagataccct ggtagtccac gccctaaacg atggatgcta
      781 gttgttgcct tgcttgtcag ggcagtaatg cagctaacgc attaagcatc ccgcctgggg
      841 agtacggtcg caagattaaa actcaaagga atagacgggg acccgcacaa gcggtggagc
      901 atgtggttta attcgaagat acgcgaagaa ccttacctag gcttgacatt gatagaatct
      961 actagagata gtggagtgcc cttcggggag cttgaaaaca ggtgctgcac ggctgtcgtc
     1021 agctcgtgtc gtgagatgtt gggttaagtc ccgcaacgag cgcaaccctc gtccttagtt
     1081 gctagcagtt cggctgagca ctctaaggag actgccttcg taaggaggag gaaggtgagg
     1141 acgacgtcaa gtcatcatgg cccttacgcc tagggctaca cacgtgctac aatggggcgc
     1201 acaaagagga gcaatatcgc gaggtggagc aaatctcaaa aacgtctctc agttcggatt
     1261 gtagtctgca actcgactac atgaagctgg aatcgctagt aatcgtgaat cagccatgtc
     1321 acggtgaata cgttcccggg tcttgtactc accgcccgtc acaccatggg agttgtattc
     1381 gccttaagtc gggatactaa attggttacc gcccacggcg gatgcagcga ctggggtgaa
     1441 gtcgtaacaa ggtaacc
```

Figure 1-6: The nine variable regions of the prokaryotic 16S rRNA gene. *Helicobacter hepaticus* (GenBank accession L39122.1) 16S rRNA gene sequence is shown indicating each variable region defined across 1457 bases. The span of V3/V4 amplicon used here for taxonomic designation (318-756) is indicated by the red bar.

The most widely used region used is V3/V4. However, this is more to do with segment length (438bp) than diagnostic accuracy as the Illumina MiSeq generates paired-end reads of 2x 300-350bp which after trimming should easily incorporate the complete V3/V4 HRVs (Wang *et al*, 2016).

This combination has its drawbacks but does create read data in a manner which can be repeated and therefore compared across studies. Although 16S rRNA gene based NGS analysis has known limitations, it has become the cornerstone to our understanding of the prokaryotic world. It currently represents a cost effective and computationally economical method to comparatively gauge the diversity of complex samples.

DNA metagenomic sequencing also employs parallelised amplification of ligated DNA libraries and records multiple, concurrent base incorporation events. However, initial material consists of enzymatically or mechanically fragmented genomic material rather than amplicon-specific products. Sequence data is then mapped and characterised by post-sequencing assembly and comparison to known genome scaffolds. Along with the ability to construct more granular phylogeny this technique allows the perception of potential gene function from the same sample. Understandably, this technique generates much more data and is more computationally challenging than that of 16S rRNA gene analysis (Izard & Rivera, 2015).

Transcriptomic sequencing uses the same sequencing methodology preceded by a reverse transcriptase step to generate DNA fragments which can be sequenced from RNA transcripts (or genomes) in a sample prior to analysis. This method refines the illumination of potential gene activity to the real-time process and can allow pathway mapping with further analysis (Izard & Rivera, 2015). Transcriptomic sequencing or RNA-seq, again, allows nonbiased expression profiling of complex biological samples. It relies upon the construction of cDNA molecules from extracted RNA strands. However, template quality is reliant on sample handling and the method of RNA extraction employed. Although granular activity catalogues can be constructed, RNA-seq can struggle to differentiate strand polarity and alternate, reverse, or non-coding features in both eukaryotic and prokaryotic genomes. As with all NGS workflows, multiple protocols can generate highly divergent data sets (Levin *et al*, 2010). Although the transcriptome of complex, niche communities will be dominated by prokaryotic genes, there will be a portion that represents the eukaryotes and a smaller one which reflects the virome. This deceivingly modest portion may represent multiple endogenous, phage or host specific viral elements. The low density of these viral transcripts, even considering their multitudinous presence, may necessitate specific fractionation or filtration steps prior to successful or representative sequence analysis and understanding (Marston *et al*, 2013).

These three broad sequencing approaches offer the researcher a key to characterising complex microbial communities and their temporal activities and auxotrophic relationships. The dogma of NGS has seemingly become *who, what, and when*, discernible by the step wise application of these workflows to a complex single microbiological sample. Applying these three approaches to characterise microbial communities is not a simple task but the use of these three methods in a diagnostic setting may allow the detection of all levels of the microbial strata which may allow not only the health status of an *in vivo* model be gauged but the effect of disease be measured by a new perimeter (Jovel *et al*, 2016).

1.13 Applying NGS analysis in a disease model

The ultimate goal of employing NGS here is to extend routine health monitoring from the FELASA exclusion list to a tool applicable to the characterisation of the GI microbiota of mice before and during drug discovery studies. This greater temporal understanding will hopefully contribute to study outcomes. Focusing this effort on research into IBD is an ideal pairing of applied NGS and a relevant biological model (Tindemans *et al*, 2020). Inflammatory bowel diseases (IBD) are multifactorial disorders characterised by chronic-progressive and relapsing intestinal inflammation. The two clinically defined forms of IBD are UC and CD. UC is a superficial ulcerative disease of the colon, whereas CD can be transmural and affecting the entire GI tract, while both conditions are associated with an increased risk of colon cancer (Lee & Chang, 2021). Another commonality between these states is CD4+ lymphocyte infiltration of the intestinal tissue. A subset of these cellular populations are the memory T-cells, which make up ~60% of the total GI lymphocyte population and are essential in mounting rapid immune responses to pathogens and as such, their action is normally highly regulated. Loss of control or the lack of pathogenic targets are thought to contribute to IBD progression (Tindemans *et al*, 2020). However, the exact aetiology of IBD is poorly understood but what was thought to be solely a genetic disease has now become one which the environment plays a significant role in disease development. Differential responses in monozygotic twins and the development of IBD in immigrants to Western counties characterised by urbanised housing and refined diets have been found to be associated with IBD (Kaser *et al*, 2010). However, IBD must be driven by a combination of genetic loci, the environment, and a myriad of further factors which themselves are comprised of multiple nuances such as the use of certain medications (e.g., non-steroidal anti-inflammatory drugs and antibiotics), exposure to cigarette smoke (protective against UC but detrimental to CD), environmental pollution, exercise, sleep levels, psychological factors, hygiene, and stress. However, combinations of these factors occur in unaffected populations making direct causality impossible (Kaser *et al*, 2010).

To attempt to clarify risk factors in individuals, genome-wide association studies (GWAS) have been used to identify specific loci for IBD. More than two hundred genetic risk loci have been associated with IBD. Of this number, causal variants involved in microbial sensing e.g., interleukin-23 receptor (IL23R), regulation of inflammatory responses e.g., nucleotide binding oligomerisation domain-containing protein 2 (NOD2), and regulation of autophagy e.g., autophagy 16-like 1 protein (ATG16L1) have been identified. Although there are discrete genetic factors for both UC and CD, some genetic associations are shared between the two states and many are linked to other multifactorial immune related disorders such as diabetes indicating that genetic markers are as hard to pinpoint as life style markers (Uniken *et al*, 2017).

It is understood that the microbiota plays a role in the regulation of immune responses in the GI tract, and that dysbiosis driven by environmental factors plays a key role in disease progression. The primary method of influencing the microbiota is diet. This has been found to influence susceptibility to IBD, with intake of dietary fibre, zinc, and vitamin D providing protection while a westernised diet high in processed food, refined sugar, and saturated fat has been implicated in changes to the diversity of GI microbiota seen in IBD. However, it is still unclear whether shifts in diversity are a result of disease progression or a driver behind it. Both changes in microbial composition and altered localisation of bacteria due to impaired barrier function can induce immune responses to non-pathogenic organisms which may drive dysregulation of an immune response (Khalili *et al*, 2018). Although CD4+ cells play a pivotal role in both CD and UC, dysregulation of both innate and adaptive arms of the immune system are implicated in aberrant behaviour. In the innate immune system reduced production of anti-microbial peptides in CD patients have reduced production of α-defensins, which is more pronounced in patients with NOD2 mutations. An early sign of intestinal inflammation is neutrophil infiltration which can contribute to pathogenesis through multiple mechanisms, including impairment of barrier function, tissue damage and secretion of soluble mediators which can amplify of inflammation.

Macrophages produce increased levels of IL-23, IL-6 and TNFα which can influence T-cell polarisation. Dendritic cells show an activated phenotype, expressing increased levels of TLR2, TLR4 and CD40 and secreting increased levels of IL-12 and IL-6. They also express higher levels of the chemokine receptor CCR7, aiding their migration to and retention in the colon (Souza & Fiocch, 2016).

In the adaptive immune system T-cells from CD patients produce IFNγ and IL-17A, suggesting that pathogenic T-cells polarise into T_h1/T_h17 cells. In CD, T_h17 cells also produce IL-21 and IL-22, driving further IFNγ production. In contrast, IL-22 is reduced in inflamed tissue from UC patients and lymphocytes from UC patients can display an atypical T_h2 response, with increased IL-5 and IL-13 expression, but low IL-4 production. Additionally, some UC patients harbour IL-9-producting T_h9 cells. Whilst TREG numbers are decreased in the blood, they are increased in lamina propria (Moschen *et al*, 2019). This suggests that these cells are impaired in their suppressive capacity as despite increased cell numbers they fail to control inflammation. In CD, T-cells are resistant to apoptosis, suggesting that accumulation of activated T-cells may contribute to disease pathogenesis. Anti-neutrophil cytoplasmic antibodies (ANCAs) are present in IBD, with increased prevalence in UC compared to CD. In addition, antibodies to microbial components are also present. Whilst these immunoglobulins do not contribute to the disease pathogenesis, they are elevated in active disease and highlight the importance of microbial antigens in driving disease (Moschen *et al*, 2019).

These dysregulations of the innate and acquired immune systems and the site of pathology points towards the GI microbiota as the most immediate environmental factor. The previously described immune priming afforded by the GI microbiota indicated the symbiotic relationship between microbial populations and the development of a healthy gut in neonates (Byndloss *et al*, 2018). It makes sense those disruptions to the microbial populations generated externally by antibiotics, diet and internally or physiologically, via the vagus nerve due to stress, work in the opposite way, creating an inverse microbial phenotype, classically observed by Ley *et al* (2006). Altered microbial patterns are associated with specific disease states or disease models but are often phyla-level and although quantified in studies, are in no way causal (Kaser *et al*, 2010). This points to a requirement to use culture independent tools capable of discerning changes at a much higher level of classification (genera and species) to understand specific roles or identify key marker species. The multifactorial element of IBD and its sometimes-distant relationship to genes which may or may not come into play according to GI microbiota and possible immune dysregulation are confounded by research in another species. Genetic loci can be easily altered in the mouse and specific targets studied which as generated much of the knowledge in this field (Kaser *et al,* 2010). More than sixty distinct animal models have been established to study IBD, which are classified primarily into transgenic, chemically induced, infection induced, cell-transfer models. These IBD models have provided significant contributions to not only dissect the mechanism but also develop novel therapeutic strategies for IBD. However, despite the many different methods by which colitis can be induced *in vivo*, there is no single model which completely mirrors the human disease. Therefore, the choice of disease model is dependent on the aspect of disease that is subject to investigation (Mizoguchi, 2012).

In the infection colitis model, the pathogenic Gram-negative *Citrobacter rodentium* is orally administered to naive immunocompetent mice resulting in epithelial damage, diversity alterations, weight loss and diarrhoea. Epithelial damage is associated with immune cell infiltration and loss of barrier integrity. This measurable element along with bacterial load and histological evidence make this understood model widely used although not in the UK (Bhinder *et al,* 2013).

The chemically induced model of colitis is generated by the administration of dextrin sulfate sodium (DSS) This was first described in 1990 and has been extensively used to understand the pathophysiology of IBD, the contribution of genes of interest to disease progression, and to evaluate the therapeutic efficacy of novel interventions (Kjellev *et al*, 2006). In this model, colitis is induced through the administration of DSS in the drinking water. DSS penetrates the intestinal mucosa, causing epithelial damage and barrier dysfunction. Dissemination of gut bacteria into the intestinal wall drives the recruitment of immune cells, resulting in an inflammatory response. The clinical signs of DSS-induced colitis include body weight loss, diarrhoea, and blood in the faeces. Histopathological changes are visible during microscopic evaluation of colon tissue (Eichele *et al*, 2017).

The T-cell transfer model of experimental colitis is generated by the transfer of a defined number of mouse naïve CD4+CD45RB high T-cells into immunodeficient mice. The recipient mice develop chronic colitis due to these T-cells homing to the intestinal mucosa and lack of functional TREGs in the host which allows the development of a T_h1/T_h17 adaptive immune response to antigens derived from intestinal bacteria. As the host animals lack TREGs, chronic colitis develops, mimicking human disease. The absence of TREGs is crucial for disease onset, as co-transfer of TREGS with naïve CD4+ T cells ablates colonic inflammation (Ostenin *et al*, 2008). Along with robust wild-type or immunodeficient strains of mice, spontaneous mutant and genetically altered models of disease offer insights to specific pathways and the effects of specific genetic loci expression. Both congenital and transgenic constructs may be used in conjunction with the cell transfer, DSS and infection systems of IBD induction. These methods of elemental presentation have become invaluable tools in this field of research (Prattis & Jurjus, 2015).

These experimental models have become useful tools in the prediction of clinical outcomes of studies involving biological entities. However, the chemical and cell transfer models currently used in the UK have associated advantages and disadvantages. Advantages of the DSS model over the T-cell transfer model are that wild-type animals used can generate lymphocytes, allowing interrogation of B-cell and CD8+ T-cell mechanisms and therapeutics. Additionally, cessation of DSS administration leads to the resolution of disease, allowing mechanisms of epithelial repair to be studied. Application of multiple cycles of DSS interspersed with normal drinking water imitates the development of remission and relapse, mimicking the chronic disease seen in patients. Disadvantages of DSS-model over the cell transfer model are that the induction of inflammation through chemically induced intestinal damage is less physiological. The variability and reproducibility of the model is often influenced by multiple factors, including microbiota, mouse strain and protocol used (Eichele *et al*, 2017). Comparison of transcriptomic changes in the colon suggests that the cell transfer model most closely reflects gene expression changes seen in IBD patients (Acera *et al,* 2021). Prediction of clinical efficacy may depend on the use of either the acute or chronic model.

1.14 Pharmaceutical translatability

The translatability of *in vivo* studies to the human patient is key in their use in the development of novel therapies. However, in any model, sometimes only a caricature can only be perceived, this especially true of the *in vivo* models used in pharmaceutical research (Lederberg, 2000).

Although the mouse is the most used model for human disease, this has come about due to its economic and reproductive advantages rather than its recent genomic characterisation. Along the path of drug discovery, it is hoped that data gleaned from the use of induced models of disease is applicable to human patients. This previously glaucomic activity was brought into sharper focus and translatability with the advent of transgenic rodent models which were created to mirror specific molecular aspects of human disease to evaluate for pharmaceutical alleviation (Hickman & Davis, 2005).

Most microbiome studies conducted on mice are reductionist and mechanistic, tending to use the mouse to answer discovery questions involving single organisms (Brugiroux *et al*, 2017). Studies attempting to understand the tool itself and how its innate qualities may be used to translate data lag behind similar efforts in the human (Kim *et al*, 2021).

Mouse models used in research today originate from fancy European and Asian mice generated around 100 years ago creating the commonly used C57BL/6, BALB/C, 129, and C3H inbred strains (Hugenholtz & de Vos, 2017). The benefit of inbreeding is genetic similarity creating a comparative tool. Today there are >400 described inbred strains. These inbred strains may be bred in sterile conditions or rederived by caesarean section or embryo transfer creating germ-free models for microbiome research (Hugenholtz & de Vos, 2017). Gnotobiotic strains may be germ-free or have been administered with a defined microbiota (e.g., altered Schaedler flora). Humanised, germ-free mouse models are seeded with defined consortia of human-derived bacteria (e.g., Oligo-Mouse-Microbiota) to gain functional understanding of causality in human disease such as colonisation resistance testing for single organisms (Brugiroux *et al*, 2017).

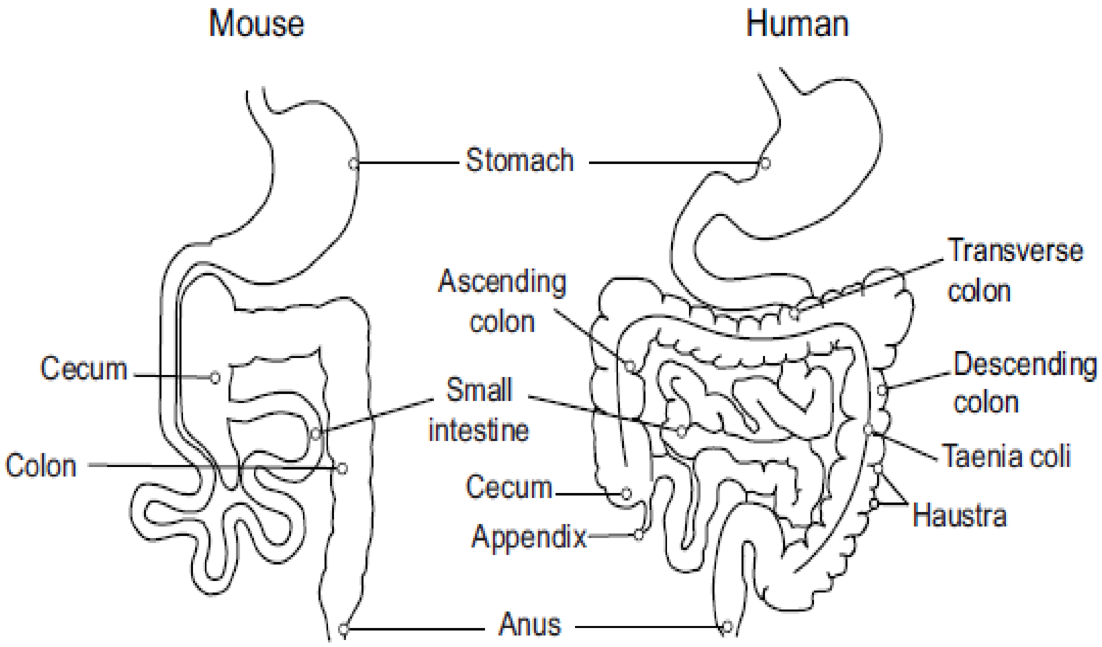


Figure 1-7: Comparison of murine and human gastrointestinal tracts. This shows the differential size of the caeca and the haustration of the human colon indicating the primary location of fermentative activity in each species (Nguyen *et al*, 2015).

The murine and human GI tracts are physiologically divergent (Nguyen *et al*, 2015), with similar metabolic processes occurring at different niches, for example with the most active site of bacterial fermentation in the mouse being the caecum and in humans this being the colon (Ley *et al*, 2006) (Figure 1-7). However, both humans and mice immunologically sculpt the microbiota of the small intestine to give priority to host carbohydrate harvesting (Santaolalla *et al,* 2012).

Characterising microbial communities and their activities at structural, metabolic or drug absorption sites along the murine GI tract may be used to understand the potential of drug entities to be metabolised, the correct microbiota for specific studies and mapping disease progression or alleviation. These attributes may reduce the number of animals used and reduce drug attrition by conducting more refined experiments. However, it has been found that ~80% of genera found in the mouse are present in the human GI microbiome but 70% of this common microbiota share <40% of core gene content (Kim *et al*, 2021). Furthermore, there is only a 10% overlap in taxa at the species level confirming the significant divergence between the model and the target species (Klieser *et al,* 2022). The mutable nature of prokaryotes and their use of auxotrophic networks for nutrient harvesting clearly indicates that appointment of taxonomic designations across host species is not translatable. The mouse is a tool for research, but it is a significantly blunt tool if its use continues without understanding its taxonomic and functional microbial repertoire. Improving our understanding of the murine microbiome represents the same movement as that in human trails by bringing microbiome data into the clinical realm (Guthrie & Kelly, 2019). Developing an accurate diagnostic tool to be routinely applied in microbial characterisation of the models used in drug development represents a step towards integrating this type of analysis in real-world scenarios (Cullen *et al*, 2020). A partnership was created with the applied immunity groups at GlaxoSmithKline (GSK) so that faecal samples could be taken during their studies and processed independently as a proof of concept in helping understand the effects of administration upon *in vivo* models and their microbiota. The following sections detail each study plan and background.

1.15 Hypothesis & experimental aims

Routine NGS characterisation of the murine GI microbiome is not currently undertaken during health monitoring or during *in vivo* studies. Consequently, the hypothesis that underpins the research presented in this thesis is that applying this evaluative tool would help characterise the microbiome, extending health monitoring beyond an exclusion criteria, aiding appropriate model selection, and allowing the tracking of disease progression in specific areas of research.

The hypothesis was tested by investigating potential workflows in a technologically conservative research space in order to develop the most accurate process by which this technique could be routinely applied in drug development.

Key objectives of this work are:

- To design a representative, repeatable and ethical sampling strategy which will be applied to the subsequent microbiome characterisation studies.
- To use this method to run pilot 16S rRNA gene analysis studies to assess the effect of host sex, geolocation, transport, and acclimatisation upon prokaryotic community diversity across multiple ecological niches of the murine GI tract to pinpoint a suitable sampling material for future studies.
- To assess and compare 16S rRNA gene analysis, DNA metagenomics, and RNA-seq methods in the description of both taxonomy and gene

function along the GI tract of a murine model using a single, openly available bioinformatic resource, to enable routine implementation.

- To improve 16S rRNA gene sequencing data generation and analysis methods and to coalesce these into a standard operating procedure for future use.
- To investigate and develop a method of producing standardised, high quality nucleic acid from faecal material to form the input material for the above workflow.
- To evaluate this complete standard operating procedure on the faecal material generated by mice throughout disease progression in two classes of IBD study.

Chapter 2: Materials & methods

2.1 Primary 16S rRNA gene analysis experiments (Chap. 3)

2.1.1 Mouse Models – niche, sex, and location study

C57(Jax) mice, from Area 52 at the Charles River Laboratories (CRL) breeding facility (Maidstone, Kent) were chosen for the geolocation, niche, and sex studies due to the length of uninterrupted colony maintenance (Table 2-1). Animals were group housed in autoclaved Tecniplast 1292N cage cages containing Datesand ECO7D, softwood flake and maintained at an ambient temperature of 21.0 +/- 1°C and relative humidity of 55% +/- 10%, on a 7am to 7pm light-dark cycle, with free access to food (VRF1 SDS Pelleted diet) and softened, filtered, UV treated and chlorinated water in bottles. These animals were provided from the specified pathogen free unit with a health surveillance report indicating the absence of all agents according to FELASA quarterly screening recommendations (Mähler *et al*, 2014). C57(Jax) mice for screening at GSK were packed into transport boxes with Clear-H_2O hydrating gel packs (South Portland, USA) at around midday on the day preceding shipment and delivered at ~8.30am on the day of transit. CRL is 109 miles from GSK, Stevenage, a journey which takes ~3hrs depending on the number of deliveries to universities and institutions *en route* across SE England. These animals were housed in autoclaved Techniplast GM500 cages containing IPS Lignocel BK8/15 bedding, Datesand Paper Shaving nesting material, a red

Perspex dome home, a cardboard fun tunnel, and a wooden chew block. Animals were maintained at an ambient temperature of 20.5 to 23.5°C with relative humidity of 39% to 61%, on a 6am to 6pm light-dark cycle, with free access to food (Labdiet expanded and irradiated 5LF2 Maintenance) and double reverse osmosis animal grade drinking water in bottles.

2.1.2 Mouse models – faecal/colon comparison study

Wildtype C57(Trim) mice, were bred in the GSK Transgenic Production Facility at Stevenage and were chosen to provide material for comparisons between faeces and colonic contents sequencing methods as they were unnecessary for the program for which they were originally generated (Table 2-2). Animals were housed in autoclaved Techniplast GM500 cages containing IPS Lignocel BK8/15 bedding with Datesand Paper Shaving nesting material, a red Perspex dome home, a cardboard fun tunnel and a wooden chew block. Animals were maintained at an ambient temperature of 20.5 to 23.5°C with relative humidity of 39% to 61%, on a 6am to 6pm light-dark cycle, with free access to food (Labdiet expanded and irradiated 5LF2 Maintenance) and double reverse osmosis animal grade drinking water in bottles. They were housed with study, long-term breeding, and stock animals but underwent no regulated procedures.

Table 2-1: Sample ID, sex, sampling location and niche of C57(Jax) digesta samples. A comparison of source, sex, and niche was conducted on these samples to gauge major effectors of prokaryotic diversity.

ID	Sex	Sampling location	Niche
JP25-1	M	CRL	Caecum
JP26-2	M	CRL	Caecum
JP27-3	F	CRL	Caecum
JP28-4	F	CRL	Caecum
JP29-5	M	GSK (at del.)	Caecum
JP30-6	M	GSK (at del.)	Caecum
JP31-7	F	GSK (at del.)	Caecum
JP32-8	F	GSK (at del.)	Caecum
JP33-9	M	GSK (2wks)	Caecum
JP34-10	M	GSK (2wks)	Caecum
JP35-11	F	GSK (2wks)	Caecum
JP36-12	F	GSK (2wks)	Caecum
JP37-13	M	CRL	Stomach
JP38-14	M	CRL	Stomach
JP39-15	F	CRL	Stomach
JP40-16	F	CRL	Stomach
JP41-17	M	CRL	Jejunum
JP42-18	M	CRL	Jejunum
JP43-19	F	CRL	Jejunum
JP44-20	F	CRL	Jejunum
JP45-21	M	CRL	Ileum
JP46-22	M	CRL	Ileum
JP47-23	F	CRL	Ileum
JP48-24	F	CRL	Ileum

Table 2-2: Sample ID, sex, and niche of C57(Trim) digesta and faecal samples. A comparison was connected here to ascertain whether faeces samples can be used to gauge prokaryotic diversity in the colon.

ID	Sex	Niche
JP01-1F	M	Faeces
JP02-1C	M	Colon
JP03-2F	M	Faeces
JP04-2C	M	Colon
JP05-3F	M	Faeces
JP06-3C	M	Colon
JP07-4F	M	Faeces
JP08-4C	M	Colon
JP09-5F	M	Faeces
JP10-5C	M	Colon

2.1.3 Mouse models – four-month longitudinal study

NOD severe combined immunodeficiency disease (NOD-SCID) immunocompromised mice for longitudinal study were obtained from Area 50 of CRL where they are housed in flexible isolators (consumables treated with chlorine dioxide at 250ppm for 10mins) and provided with VRF1 (SDS) diet (gamma irradiated), softened, filtered, UV treated and chlorinated water in bottles and kept on Nepco aspen bedding with Kleenex tissues and a cardboard tunnel (all gamma irradiated) on a 7am to 7pm light-dark cycle. These animals were provided from the specified pathogen-free unit with a health surveillance report indicating the absence of all agents according to FELASA quarterly screening recommendations conducted by PCR only (Mähler *et al,* 2014).

Upon arrival at GSK, they were housed in autoclaved Techniplast individual ventilated cages (IVCs) containing autoclaved sawdust, golden shavings, and a cardboard tunnel. Animals were maintained at an ambient temperature of 21.0 +/- 1°C with a relative humidity of 55% +/- 10%, on a 6am to 6pm light-dark cycle, with free access to food (5LF2 extruded diet) and autoclaved animal grade drinking water in bottles. These animals were provided from the specified pathogen-free unit with a health surveillance report indicating the absence of all agents according to FELASA quarterly screening recommendations (Mähler *et al,* 2014). During this study faecal pellets were sampled at delivery and the subsequent monthly anniversary for three months.

2.1.4 Terminal and cage sampling

The C57(Jax) study was designed to sample three groups of two males and two females at source (A), at delivery to GSK, Stevenage (B) and after 2wks acclimatisation in the experimental animal facility at Stevenage (C). Therefore, the acclimatised group were delivered to GSK at 5-6wks of age. Male and female groups from each sampling date had been either delivered or housed together for the period of transport or acclimatisation. All mice were all terminally sampled at 8wks of age, with sampling of group A taking place in the necropsy suite at CRL, Maidstone, while groups B & C were sampled in the microbiology necropsy suite at Stevenage.

All animals sampled were kept in a Scantainer (Scanbur, Denmark) to isolate them from sight or smell during *post mortem* procedures. Euthanasia was conducted using rising concentration of CO_2. After loss of the righting reflex, the ventral surface of each mouse was exposed upon a downdraft table and wiped with 70% ethanol. A cut was then made in the skin from the lower abdomen, following up through the rib cage, exposing the heart and lungs at which point the heart was removed as a secondary confirmation of death.

Individually packed and steam sterilised scissors and tweezers were employed for each mouse sampled. The gastrointestinal tract was then removed from the abdominal cavity by firstly cutting the oesophagus above the stomach and severing connective tissue to the liver and then gently pulling up and away, finally cutting at the lower colon just before the anus. The excised tract was then laid out on a sterile 90mm square petri dish (Sterilin, UK). Each GI tract was then cut into five sections from anterior to posterior (stomach, jejunum, ileum, caecum, and colon) to minimise the potential for contamination from areas with higher bacterial populations. and each section of the GI tract was opened using a new sterile scalpel blade which was then used to gently apply pressure to the exterior to push out digesta for collection into a PCR clean 1.5µl tube (Eppendorf, UK).

All sectional digesta samples were collected into sterile 2ml Sarstedt tubes (Numbrecht, Germany) on dry ice and were then stored at -80˚C until processing. Dedicated PPE was employed while conducting euthanasia and *post mortem* sampling of all animals used. All sampling events were conducted between 8.30-10.30am to minimise any possible temporal variations in diversity. All consumables were disinfected in 1.5% Virkon (Antec International, UK), prior to autoclaving and off-site incineration in accordance with current company waste-stream regulations.

Excreted cage faeces were obtained for the colon comparison and the NSG study by aseptically picking pellets from home cages at base changing intervals with sterile forceps into sterile 2ml Sarstedt tubes (Numbrecht, Germany). Collected samples were again stored at -80˚C until submission.

2.1.5 External material processing & NGS

DNA extractions, NGS library preparations, and Illumina MiSeq sequencing were conducted at Genewiz, Inc. DNA was extracted from submitted samples using DNeasy Powersoil kit (Qiagen, UK). Resulting DNA was quantified using a Qubit 2.0 Fluorometer (Invitrogen, USA). 30-50 ng DNA was used to generate amplicons using a MetaVx™ Library Preparation kit. V3, V4, and V5 hypervariable regions of prokaryotic 16S rRNA gene DNA were selected for generating amplicons and following taxonomy analysis. Genewiz designed a panel of proprietary primers aimed at conserved regions bordering the V3 and V4 hypervariable regions of bacteria and Archaea 16S rRNA gene DNA. First round PCR products were used as templates for second round amplicon enrichment PCR. At the same time, indexed adapters were added to the ends of the 16S rRNA DNA amplicons to generate indexed libraries ready for downstream NGS sequencing on Illumina MiSeq (San Diego, USA). DNA libraries were validated by Agilent 2100 Bioanalyzer (Palo Alto, USA), and quantified by Qubit 2.0 Fluorometer. DNA libraries were multiplexed and loaded on an Illumina MiSeq instrument according to manufacturer's instructions. Sequencing was performed using a 2x300/250 paired-end (PE) configuration; image analysis and base calling were conducted by the MiSeq Control Software (MCS) embedded in the MiSeq instrument. The cost per sample was £150.

2.1.6 External QIIME data analysis

The QIIME data analysis package was used for 16S rRNA gene OTU table generation. The forward and reverse reads were joined and assigned to samples based on barcode and truncated by cutting off the barcode and primer sequence. Quality filtering on joined sequences was performed and sequence which did not fulfil the following criteria were discarded: sequence length

<200bp, no ambiguous bases, mean quality score >= 20. Then the sequences were compared with the RDP reference database (www.rdp.cme.msu.edu) using UCHIME algorithm to detect chimeric sequence (UCHIME Home Page drive5.com), and then the chimeric sequences were removed. The effective sequences were used in the final analysis. Sequences were grouped into operational taxonomic units (OTUs) using the clustering program VSEARCH (1.9.6) against the Silva 119 database pre-clustered at 97% sequence identity. The Ribosomal Database Program (RDP) classifier was used to assign a taxonomic category to all OTUs at a confidence threshold of 0.8. The RDP classifier uses the Silva 119 database which has taxonomic categories predicted to the species level.

2.1.7 Statistical analysis of OTU tables

The use of the Shannon Index (Shannon & Weaver, 1964) was applied here using the following equation in Excel (Microsoft, USA), where the individual read number is divided by the sum of the reads per sample multiplied by the natural log of the individual read count multiplied by the sum of the reads per sample.

$$\text{Shannon Index (H)} = - \sum_{i=1}^{s} p_i \ln p_i$$

The raw p-values and false discovery rates (FDR) were calculated (Verhoeven *et al*, 2005) using Array Studio (Qiagen, UK) from OTU read counts by applying the log+1 of each count with the assistance of the GSK Development Statistics group.

2.2 NGS method comparison using MG-RAST server (Chap. 4)

2.2.1 Mouse model – niche comparison

Wildtype C57(Trim) mice, were bred in the Transgenic Production Facility at Stevenage and were chosen to provide material for comparisons between 16S rRNA, shotgun and RNA-seq sequencing methods as they were unnecessary for the program for which they were originally generated (Table 2-3). Animals were housed as those in section 2.1.2.

These mice were housed in a room with study, long-term breeding, and stock animals but underwent no regulated procedures. Two cages were sampled, each containing three 10-12wk old mice. The first cage housed three littermates (432, 435 & 436). The second cage house two littermates (645 & 642) and an unrelated mouse of the same age (495) (Table 2-3). These animals were transferred to an approved necropsy area and euthanised by rising concentration of CO_2 with confirmation of death being made by the removal of heart. Samples were collected aseptically into PCR-clean 1.5ml Eppendorf Safe-Lock tubes on dry ice and then stored at –80°C until processing.

Table 2-3: Sample ID and cage origin of digesta samples (1-14) obtained from C57(Trim) mice. These different materials were used to compare niche diversities using 16S rRNA gene analysis, DNA metagenomic, and RNA-seq NGS methods.

	C57(Trim)					
Cage #	Cage 1			Cage 2		
Animal	432-	435-	436-	495-	645-	642-
Stomach	1			8		
Jejunum	2			9		
Ileum	3			10		
Caecum	4			11		
Colon	5	6	7	12	13	14

2.2.2 External material processing & NGS

Digesta samples from the C57(Trim) mice were shipped on dry ice to Genewiz (Plainfield, USA). Nucleic acids were co-extracted using PowerViral DNA/RNA kit (Qiagen, UK).

For the 16S rRNA gene analysis workflow, resulting DNA was quantified using a Qubit 2.0 Fluorometer (Invitrogen, USA). 30-50 ng DNA was used to generate amplicons using a MetaVx™ Library Preparation kit (Genewiz, Inc., USA). V3, V4, and V5 hypervariable regions of prokaryotic 16S rRNA gene were selected for generating amplicons and following taxonomy analysis. Genewiz designed a panel of proprietary primers aimed at conserved regions bordering the V3 and V4 hypervariable regions of bacteria and Archaea 16S rRNA gene. The V3 and V4 regions were amplified using forward primer (5'-CCTACGGRRBGCASCAGKVRVGAAT) and reverse primer (5'-GGACTA CNVGGGTWTCTAATCC). The V4 and V5 regions were amplified using forward primer (5'-GTGYCAGCMGCCGCGGTAA) and reverse primer (5'-CTTGTGCGGKCCCCCGYCAATTC). First round PCR products were used as templates for second round amplicon enrichment PCR. At the same time, indexed adapters were added to the ends of the 16S rRNA gene DNA amplicons to generate indexed libraries ready for downstream NGS sequencing on Illumina MiSeq. DNA libraries were validated by Agilent 2100 Bioanalyzer (Agilent Technologies, Palo Alto, CA, USA), and quantified by Qubit 2.0 Fluorometer. DNA libraries were multiplexed and loaded on an Illumina MiSeq instrument according to manufacturer's instructions. Sequencing was performed using a 2x300/250 paired-end (PE) configuration;

image analysis and base calling were conducted by the MiSeq Control Software (MCS) embedded in the MiSeq instrument.

For the RNA-seq workflow, rRNA depletion was performed using Illumina Ribozero rRNA Removal Kit and TruSeq Stranded Total RNA library Prep kit following manufacturer's protocol (Illumina, Cat# RS-122-2101). Briefly, rRNA was depleted with Ribp-Zero rRNA Removal Kit, rRNA depleted RNAs were fragmented for 8 minutes at 94 °C. First strand and second strand cDNA were subsequently synthesised. The second strand of cDNA was marked by incorporating dUTP during the synthesis. cDNA fragments were adenylated at 3'ends, and indexed adapter was ligated to cDNA fragments. Limited cycle PCR was used for library enrichment. The incorporated dUTP in second strand cDNA quenched the amplification of second strand, which helped to preserve the strand specificity. Sequencing libraries were validated using DNA Analysis Screen Tape on the Agilent 2200 TapeStation (Agilent Technologies, Palo Alto, CA, USA), and quantified by using Qubit 2.0 Fluorometer (Invitrogen, Carlsbad, CA) as well as by quantitative PCR (KAPA Biosystems, Wilmington, MA, USA). The cost per sample was ~£500.

For the metagenomic workflow, DNA library preparations and sequencing reactions were conducted at Genewiz, Inc. NEB NextUltra DNA Library Preparation kit was used following the manufacturer's recommendations. Briefly, the genomic DNA was fragmented by acoustic shearing with a Covaris S220 instrument. The DNA was end repaired and adenylated. Adapters were ligated after adenylation of the 3’ends. Adapter-ligated DNA was indexed and enriched by limited cycle PCR. The DNA library was validated using TapeStation (Agilent Technologies, Palo Alto, CA, USA), and was quantified using Qubit 2.0 Fluorometer. The pooled libraries were clustered and loaded on the Illumina HiSeq instrument (4000 or equivalent) according to manufacturer’s instructions and sequenced using a 2x150bp Paired End (PE) configuration. Image analysis and base calling were conducted by the HiSeq Control Software (HCS). Raw sequence data (.bcl files) generated from Illumina HiSeq was converted into fastq files and de-multiplexed using Illumina's bcl2fastq 2.17 software. The cost per sample was ~£500.

2.2.3 External data analysis

Although no data analysis was conducted by Genewiz on the metagenomic or RNA-seq files, the QIIME data analysis package was used for 16S rRNA OTU table generation as described in 2.1.6.

2.2.4 Data delivery & security

The C57(Trim) digesta samples generated 84 raw .fastq sequence files (42 paired end) generated during these experiments which were uploaded by Genewiz onto the DNAnexus (https://www.dnanexus.com) cloud-based server (Mountain View, USA) for secure transfer to an internal network.

2.2.5 Internal data analysis

Files were subsequently downloaded from DNAnexus using a Virgin Media (Reading, UK) 100Mps domestic package. Downloaded compressed files (.fastq) were extracted using Winzip (Ottawa, Canada) and stored on a Western Digital (San Jose, USA) 4Tb external hard drive. The same domestic package with an upload speed of <10Mps was used to transfer these .fasta files to the MG-RAST DNA metagenomics analysis server for analysis (https://www.mg-rast.org/mgmain.html.; Meyer *et al,* 2008). Assemblies and comparative analysis were made on the MG-RAST server between all available 16S rRNA databases, those being Greengenes (GG), Ribosomal Database Project (RDP), NCBI Reference Sequences (RefSeq) SILVA small and sub-unit databases (SSU & LSU). Functional analysis of DNA metagenomics and RNA-seq files was conducted using the Kyoto Encyclopedia of Genes and Genomes (KEGG) Orthologue (KO) database (https://www.genome.jp/kegg/) via the MG-RAST server.

2.3 Hybrid 16S rRNA gene characterisation method (Chap. 5)

2.3.1 Sample data

Initial bioinformatic research used the raw MiSeq .fastq files generated for 16S rRNA gene comparison in the previous experiment for manipulation and analysis.

2.3.2 Improved 16S rRNA gene analysis

All file sharing and analysis was carried using a HP Z-book G5 laptop with an 8-core Intel® i9-9880H processor, 2x32Gb DDR4 2666 RAM and 2x 2Tb solid state hard drives for read-write performance running Windows 10. Analysis was conducted on .fasta files using Lasergene NGS suite (https://www.dnastar .com/software/genomics), a software package comprising of multiple tools for genome and metagenomic sequence assembly and analysis developed by DNAstar (Madison, USA). A bespoke 16S rRNA gene reference database was downloaded from the NCBI Nucleotide website (https://www.ncbi.nlm.nih .gov/nucleotide/) defined .fasta files, by selecting ‘bacteria’ and ‘archaea,’ ‘RefSeq’ and ‘rRNA’ tick boxes. This novel reference database contained 21,762 complete RefSeq 16S rRNA genes, defined to the strain level.

2.3.3 Application of Python script

Python script (https://www.python.org) was used for the generation of an annot.txt file, used to fully annotate the OTU tables generated during assembly phase of the novel bioinformatic pipeline in Array Star (Lasergene). Saving the reference database (.annot) in the same folder as script allows automatic proximity running. The script shown below was written with assistance of Lasergene (Madison, USA):

```
Import glob
File1=”db.fas”
With open(file1,”r”) as oFile:
With open(“annot.txt”,”w”) as wFile:
wFile.write(“REF\tGI\tNotes\n”)
rLines=oFile.readlines()
for x in rLines:
if x. startswith(“>”):
xSplit=x.strip() .split(“|”)
print xSplit
#’>gi’, ‘219722938’, ‘ref’, ‘NR_024570.1’, Escherichia coli
strain U 5/41 16S ribosomal RNA, partial sequence\n’
wFile.write(x.split[3]+”\t”+xSplit[1]+”t\”+xSplit[4]+”\n”)
```

2.3.4 External DNA extraction comparison

Stock C57BL/6 mice were chosen for the DNA extraction experiments. Animals were housed at the Stevenage animal facility in autoclaved Techniplast GM500 cages containing IPS Lignocel BK8/15 bedding with Dates and Paper Shaving nesting material, a red Perspex dome home, a cardboard fun tunnel, and a wooden chew block. Animals were maintained at an ambient temperature of 20.5 to 23.5°C with relative humidity of 39% to 61%, on a 6am to 6pm light-dark cycle, with free access to food (Labdiet expanded and irradiated 5LF2 Maintenance) and double reverse osmosis animal grade drinking water in bottles. They were housed with long-term breeding, and stock animals and had undergone no regulated procedures. The faecal pellets from four stock cages containing six individual mice were taken aseptically into a single 20ml Sterilin tube, mixed and six were then removed aseptically into PCR-clean 1.5ml Safe-Lock tubes (Eppendorf, UK). These randomised samples were stored at –80°C until in-house DNA extraction and/or submission to four contract research organisations (CROs): Genewiz, Eurofins, Qiagen and Charles River Laboratories (CRL). DNA was extracted in-house using the DNeasy PowerSoil Pro kit (Qiagen, UK) according to manufacturer's instructions but using a 75µl final elution volume. Extraction took place on the day before submission to all external providers and stored overnight at -20°C. Two samples of in-house extracted DNA and two Eppendorf tubes containing faecal pellets were submitted to the four CROs for extraction and 16S rRNA library preparation followed by MiSeq sequencing and raw .fasta file generation.

Eurofins employed a Kingfisher magnetic extraction kit, Genewiz used DNeasy PowerSoil (Qiagen, UK), Qiagen used QIAmp PowerFecal Pro DNA Kit (Qiagen, UK) and CRL used Powersoil kit (Qiagen, UK). Raw .fasta files representing the four samples were downloaded from each CRO and analysed by the method developed here. Counts for each data entry on the reference database were compared and a Spearman rank correlation matrix was generated to compare the pairs of samples and gauge extraction technique quality and reproducibility.

2.4 Application of hybrid 16S rRNA gene analysis method (Chap. 6)

2.4.1 Description of IBD studies and animals

For the adoptive CD4+ transfer model of colitis, C.B-17/IcrHsd-PrkcdSCID mice were used to receive donated CD4+ T-helper cells harvested from Balb/c mice (Table 2-4). Excreted faecal samples were taken at delivery, during acclimatisation, and throughout the study until termination. Groups 1A and 3A (each having six mice) were given 100µl phosphate buffered saline (PBS; intraperitoneal) as controls, while groups 2A, 2B, 4A and 4B (seven mice in each group) received 5x10^4 naïve CD4+ T-cells in 100µl PBS (intraperitoneal) on Day 0. Groups 1A, 2A & 2B were given non-sterile water and food (non-sterile group), while groups 3A, 4A & 4B were given double reverse osmosis water and irradiated diet (sterile group) (Table 2-4). Body weights were taken every day. Tail bleeds were conducted on all animals on Day 8 and Day 22, and all animals underwent endoscopy on Day 21. Oral sham dosing with 10ml/kg 1% methylcellulose vehicle took place every day from Day 25 until study termination on Day 38. These mice were housed as those in section 2.1.2 Every three weeks, 50% of bedding was removed and replaced with new along with more regular latrine area cleaning. All manipulations took placed inside a changing station.

For the dextran sodium sulphate (DSS) dose response study 32 12wk old female C57BL/6 mice were split into four dose groups (0%, 2%, 3% and 4% DSS) housed between two cages due to stocking density (Table 2-4). Mice were acclimatised for fourteen days post-arrival and given Hydrogel along with tap water from date of arrival. Hydrogel was removed at Day 0, when DSS and water was administered. DSS in water was replaced by tap water and Hydrogel at Day 5. All animals were terminated on Day 8. These C57 mice were housed as those animals in section 2.1.2. Cages were changed weekly, but environmental enrichment remained constant for each cage throughout study. All manipulations took placed inside a changing station.

Table 2-4: Group ID, and study days of sampling for the CD4+ and DSS IBD studies. Material collection days are given for all samples analysed by the new 16S rRNA gene analysis methodology.

CD4+	Day -12	Day -4	Day 3	Day 10	Day 17	Day 24	Day 31	Day 38
Delivery	1-6							
1A		7	8	9	10	11	12	13
2A		14	15	16	17	18	19	20
2B		21	22	23	24	25	26	27
3A		28	29	30	31	32	33	34
4A		35	36	37	38	39	40	41
4B		42	43	44	45	46	47	48
DSS	Day -20	Day -16	Day -12	Day 0	Day 2	Day 4	Day 7	Day 8
Delivery	1							
Delivery	2							
4%		3	11	19	27	35	43	51
4%		4	12	20	28	36	44	52
H_2O		5	13	21	29	37	45	53
H_2O		6	14	22	30	38	46	54
2%		7	15	23	31	39	47	55
2%		8	16	24	32	40	48	56
3%		9	17	25	33	41	49	57
3%		10	18	26	34	42	50	58

2.4.2 Samples & storage

Faecal samples were obtained naturally during routine weighing and study interventions on the dates given (Table 2-4). Faecal material was immediately placed aseptically into PCR-clean 1.5ml Safe-Lock tubes (Eppendorf, UK) on dry ice and then stored at –80°C until extraction.

2.4.3 Internal DNA extraction

DNA was extracted onsite using the DNeasy PowerSoil Pro kit (Qiagen, UK) according to manufacturer's instructions using a 75µl final elution volume on the day before submission to CRL and stored overnight at -20°C.

2.4.4 New external sequencing partner

Frozen DNA was submitted to CRL and delivered within 48hrs to their sequencing laboratory in Wilmington, USA. Recovery yield and DNA quality was determined by fluorometric analysis (QuBit, ThermoFisher). DNA concentration was adjusted to specifications and amplified using broadly reactive 16S rRNA gene primers spanning the V3 and V4 regions. Resulting amplified PCR products were analysed for quantity and correct product size (Bioanalyzer, Agilent Technologies) then purified and amplified with primers containing unique sample nucleotide barcodes (Illumina). PCR products quality and quantity were further analysed by SYBR green qPCR (KAPA, Roche Biotechnologies). All samples were pooled and adjusted to a normalised concentration. The DNA library pool was denatured with sodium hydroxide, normalised to optimal loading concentration, and combined with PhiX control (Illumina). Extended read lengths up to 2 X 300 bp was used for cluster generation and sequencing on an Illumina MiSeq. Following the sequencing run, the sequence data was de-multiplexed based on the nucleotide barcode. Sequence. fastq files were finally uploaded onto One Codex file share application (app.onecodex.com). The cost per sample was £75.

2.4.5 Novel data analysis

Raw. fastq sequencing files were downloaded from One Codex on both GSK and domestic internet networks via Wi-Fi connection. All computational work was carried as described in 2.3.2 and 2.3.3.

2.4.6 Use of animals in this study

All animal studies were ethically reviewed and conducted in accordance with Animals Directive 2010/63/EEC and the GSK policy on Care, Welfare, and Treatment of Animals.

Chapter 3: Primary 16S rRNA gene analysis experiments

3.1 Introduction

The number of microbiome-based publications published annually reflects the number of sampling, processing, and analysis methods available to researchers. This figure reflects the limitations in applying a single method that will work in all circumstances. This point drives the organic nature and applications seen in the literature. Remarkably, the routine use of microbiome studies and NGS for *in vivo* model selection purposes is non-existent. The cost of such studies now makes it economic to embed these approaches early within the drug discovery and development pipeline. To create a routine sampling and sequencing method where none exists, proof of concept studies were necessary to answer key questions regarding variable input sources. These would be the geographical location of study animals, the acclimatisation period applied, the sex of the animal, and the GI niche sampled to provide the most useful information, taking advantage of the current provision or availability of the necessary techniques. At this point, 16S rRNA gene sequence analysis was used and all NGS work (extraction, library preparation, sequencing, and analysis) was outsourced to Genewiz (Plainfield, USA). These exploratory studies would provide understanding of the effect general animal handling and husbandry variables have upon data generated from the widest number of sampling conditions i.e., length of study, sex of the animals on study and from where they originate. This activity would facilitate a more robust experimental design going forward.

To align microbiome analysis with the 3Rs (reduce, replace, and refine animal use), all research experiments described here were conducted on surplus animals or those already being sacrificed at study termination. Adding microbiome analysis to existing study plans (rather than generating microbiome-specific investigations) reduced animal usage while still producing essential knowledge. It was intended that this approach would be continued, with deeper health monitoring being gleaned only from animals on study.

In order to develop an ethical sampling methodology, the communities found in excreted faeces and colonic digesta would be compared to assess whether microbiome analysis could be carried out on cage faeces, negating the need to sacrifice animals. In addition to an ethical strategy, it was intended that this work would generate a sampling method utilising the highest degree of aseptic technique to avoid contamination across multiple samples. This work method would go on to form the basis of a standard operating procedure for future investigations.

3.2 Results

3.2.1 Sampling strategy

The aseptic removal of the complete mouse GI tract was performed by sectioning the organ at the base of the oesophagus and gently pulling it away from the body cavity in one movement, bisecting the organ at the anus. Once the complete structure was laid on a petri dish, equidistant anterior/posterior sections could be cut using sterilised, single-use instruments producing uncontaminated samples of equal volume (Figure 3-1). Digesta was pushed from each section in an anterior-posterior direction using a sterile scalpel. Samples were all snap-frozen on dry-ice. This method of terminal sampling was used, unaltered, for all subsequent studies.

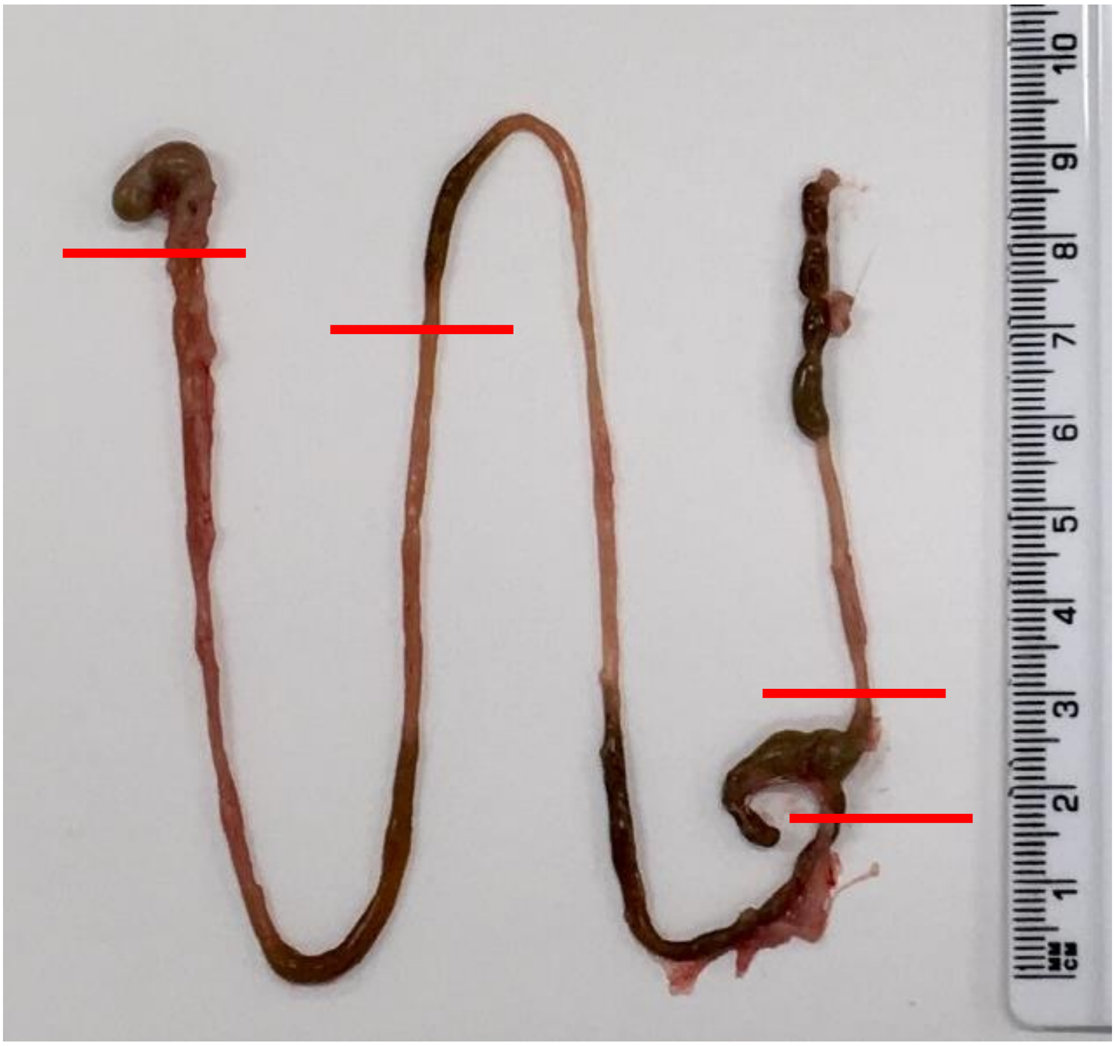


Figure 3-1: The complete GI tract of a C57BL/6 mouse. Scale and bars indicating the sequential sectioning points from left to right, representing the physiological environments deemed to represent the stomach, jejunum, ileum, caecum, and colon.

3.2.2 Effect of long-term storage of digesta upon DNA yield & OTUs

An assessment of the concentration of extracted DNA was conducted to gauge any effect of long-term freezer storage and the resulting values are shown for the niche comparison experiment (Table 3-1) and the excretion experiment (Table 3-2). Concentrations of DNA extracted from C57(Jax) mouse digesta indicated a high degree of variation between site-specific samples which was expected but a high degree of variability was seen in the results within each niche. The stomach samples had a range of 0.43-6.67ng/µl ($\bar{X}$ of 2.88ng/µl), the jejunum samples had a range of 0.24-1.44ng/µl ($\bar{X}$ of 0.72 ng/µl), the ileum samples had a range of 0.13-21.3ng/µl ($\bar{X}$ of 5.72ng/µl), and the caecum samples had a range of 3.23-36.7ng/µl ($\bar{X}$ of 21.68ng/µl). The DNA concentrations from C57(Trim) faeces and colonic digesta (Table 3-2) showed less inter-sample variation and a clear difference between materials with faeces generating a 30.6-68.0ng/µl range ($\bar{X}$ of 43.7ng/µl), while the colon contents generated a lower range of 12.6-48.8ng/µl ($\bar{X}$ of 39.4ng/µl).

.

Table 3-1: Sample ID, GI niche, and concentration (ng/µl) of DNA extracted from C57(Jax) mouse digesta. This indicates the variability of resulting DNA concentrations measured after 18 months storage at -80°C.

ID	Niche	DNA conc. (ng/µl)
JP25-1	Caecum	21.9
JP26-2	Caecum	16.5
JP27-3	Caecum	3.23
JP28-4	Caecum	9.13
JP29-5	Caecum	35.3
JP30-6	Caecum	22.7
JP31-7	Caecum	36.7
JP32-8	Caecum	6.8
JP33-9	Caecum	29.6
JP34-10	Caecum	28.3
JP35-11	Caecum	31.3
JP36-12	Caecum	18.7
JP37-13	Stomach	6.67
JP38-14	Stomach	0.43
JP39-15	Stomach	1.47
JP40-16	Stomach	2.96
JP41-17	Jejunum	1.44
JP42-18	Jejunum	0.24
JP43-19	Jejunum	0.54
JP44-20	Jejunum	0.69
JP45-21	Ileum	21.3
JP46-22	Ileum	0.13
JP47-23	Ileum	1.17
JP48-24	Ileum	0.29

The total number of OTUs obtained from C57(Jax) faeces was 304, while the C57(Trim) faeces generated 303. The total number of defined phylotypes described using the outsourced analysis pipeline in both studies was ninety-one. Only six phylotypes reached the species level designation, with twenty-six genera, twenty-four families, fifteen orders, thirteen classes and seven phyla being noted. The only disparity between C57(Jax) and C57(Trim) sampling and storage was the former were kept at -80°C for >18 months prior to analysis, while the latter were processed after ∼1 week at -80°C. Freeze thaw may have affected DNA concentrations but OTU and phylotype outputs were comparable

Table 3-2: Sample ID, niche, and concentration (ng/µl) of DNA extracted from C57(Trim) mouse faeces. This indicates the variability of resulting DNA concentrations (ng/µl) after storage at -80°C for <1 month.

ID	Niche	DNA conc. (ng/µl)
JP01-1F	Faeces	50.2
JP02-1C	Caecum	24.6
JP03-2F	Faeces	30.6
JP04-2C	Caecum	22.6
JP05-3F	Faeces	38.8
JP06-3C	Caecum	14.4
JP07-4F	Faeces	30.8
JP08-4C	Caecum	12.6
JP09-5F	Faeces	68
JP10-5C	Caecum	48.8

3.2.3 Effect of host sex upon GI community diversity

To assess whether the sex of the host influences community diversity, hierarchical clustering of the OTU counts from pairs of caecal samples taken at CRL, GSK at time of arrival, and after two weeks acclimatisation was conducted using ArrayStudio. This suggested that OTUs in male and female caecal samples taken at each site sampled, behaved independently from the sex of the host. When raw p-values were calculated between the sex variable for the log1 of each OTU count observed from C57(Jax) mouse faeces, it was found that sixteen OTUs had a p-value <0.05 (thirteen of which were bacillota). When the FDR was applied to p-values, no difference was found between the diversity of each sex, illustrating that the presence of OTUs in the caeca is not dependant on the sex of the host.

3.2.4 Effect of GI niche upon community diversity

A simple numerical count of OTUs derived from the C57(Jax) digesta was made at each ecological site evaluated at the phyla level. This showed that the stomach contains a numerous and diverse microbiota (224 OTUs), second only to the colon (277 OTUs), while the jejunum and ileum were found to have seventy-six and seventy-one OTUs respectively (Figure 3-2). The phyla level of *Bacteroidetes* is constant at all sites while numbers of those belonging to the *Bacillota*, which include the drivers of carbohydrate breakdown, were lowest in the jejunum and ileum. It is also clear that diversity ranges are independent of DNA extraction efficiency (Table 3-1 & 3-2).

A detailed microbiological assessment was made by studying the proportions of 250/300 identified bacteria across the taxonomic levels (Figure 3-3). These analyses indicate that two *Lactobacillus* OTUs dominate the acid environment of the stomach and small intestine, along with multiple S24-7 OTUs. Many of these enigmatic members of the GI microbiota are also found in the caecum, which represents a homogenous environment with no dominant OTU being obvious from these data. This indicates that the low-pH niches, where initial host absorption of carbohydrates occurs, are dominated by a small number of mucin-integrated, acidophilic species (Flint *et al*, 2012). These data also confirm that the caecum contains a wide range of OTUs possibly fulfilling multiple metabolic pathways in a competitive manner (Coyte *et al*, 2015). This diversity also likely indicates a functional microbiota mirroring that of humans (Flint *et al*, 2012), possibly created by immune sculpting (Zheng *et al* 2019).

The community diversities of colon contents and excreted faeces were compared to determine whether animals need to be sacrificed to provide representative samples for 16S rRNA gene analysis. It was thought that alterations in diversity could arise from subsequent home-cage contamination with external skin commensals or via post-excretion growth of facultative anaerobes and death of obligate species, however, this work indicates that this is not the case and that excreted faeces are comparable to colon contents in terms of its diagnostic potential.

The Shannon Index (Shannon & Weaver, 1963) was employed to allow all diversity analysis to measure the richness (diversity), evenness (frequency), and dominance where results <1.5 are considered low diversity, >1.5 & <2.5

are considered medium and >2.5 are considered high (Wagner *et al*, 2018). This approach was made to all subsequent analysis to provide a common approach. The results of this analysis (Figure 3-4A) confirm those found by numerical evaluation but aligns the richness of the stomach closer to that of the small intestine (Figure 3-2). This is due to dominance of fewer OTUs lowering the richness measurement, as opposed to the lack of OTU dominance in the caecum. The box and whisker plot (Figure 3-4B) indicate the median counts more aligned to numerical the values described in Figure 3-2, with remarkably similar spreads across the quartiles of each niche.

Principal component analysis (PCA) was conducted to examine the difference between the OTUs found in each niche (Figure 3-5). This again indicated the similarity in terms of OTUs from the jejunum and ileum samples and their high difference from that found in the caecum. It was found that the diversity of the stomach, not only numerically approached that of the caecum but its membership was more aligned than those of the small intestine.

In addition to these illustrative comparisons of diversity, the change in community diversity was also statistically interrogated; the fold change, raw p-value, and FDR being calculated from the log+1 of each taxonomically designated OTU count between the stomach, jejunum, ileum, and caecum which indicated that OTUs behave independently within a host, rather than across all samples. The changes in diversity or so many OTUs are more easily illustrated by plotting raw p-values against fold change (Figure 3-6A-C). A decrease shift is seen between the stomach and the small intestine environments (Figure 3-6A), whereas the jejunum and ileum displayed a low-level shift in both directions (Figure 3-6B), while a huge increase shift was seen between the ileum and caecum (Figure 3-6C).

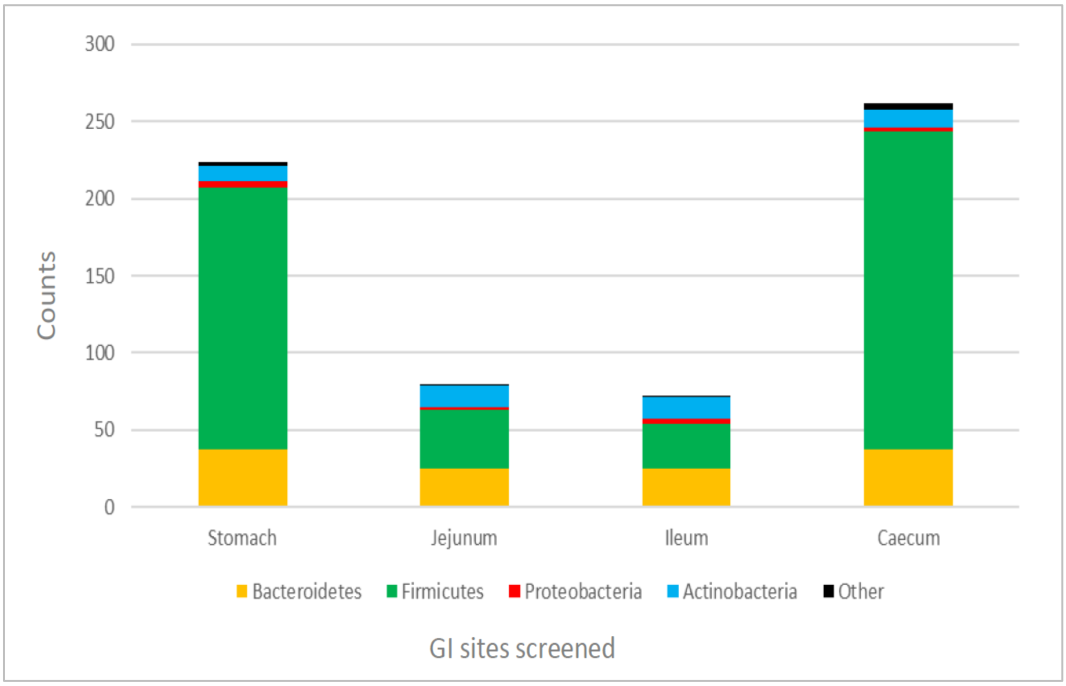


Figure 3-2: A numerical representation of phyla found in each environmental niche of the C57BL/6 mouse GI tract. This indicates the reduction of overall diversity and specifically the phyla *Bacillota* (synonym *Firmicutes*) in the small intestine (jejunum and ileum) compared to that of the stomach and the caecum.

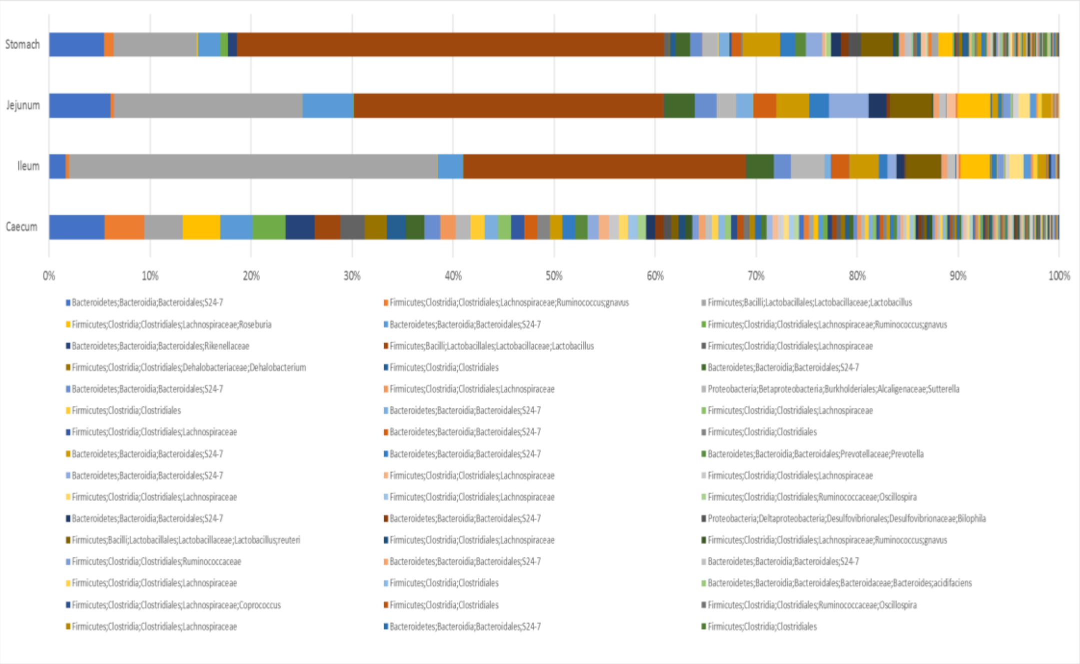


Figure 3-3: The relative abundance of OTUs identified in C57BL/6 mice at each given niche. This indicates a high prokaryotic diversity or richness in the caecum compared to the stomach, ileum, and jejunum where specific lactobacilli are seen to dominate the niche communities.

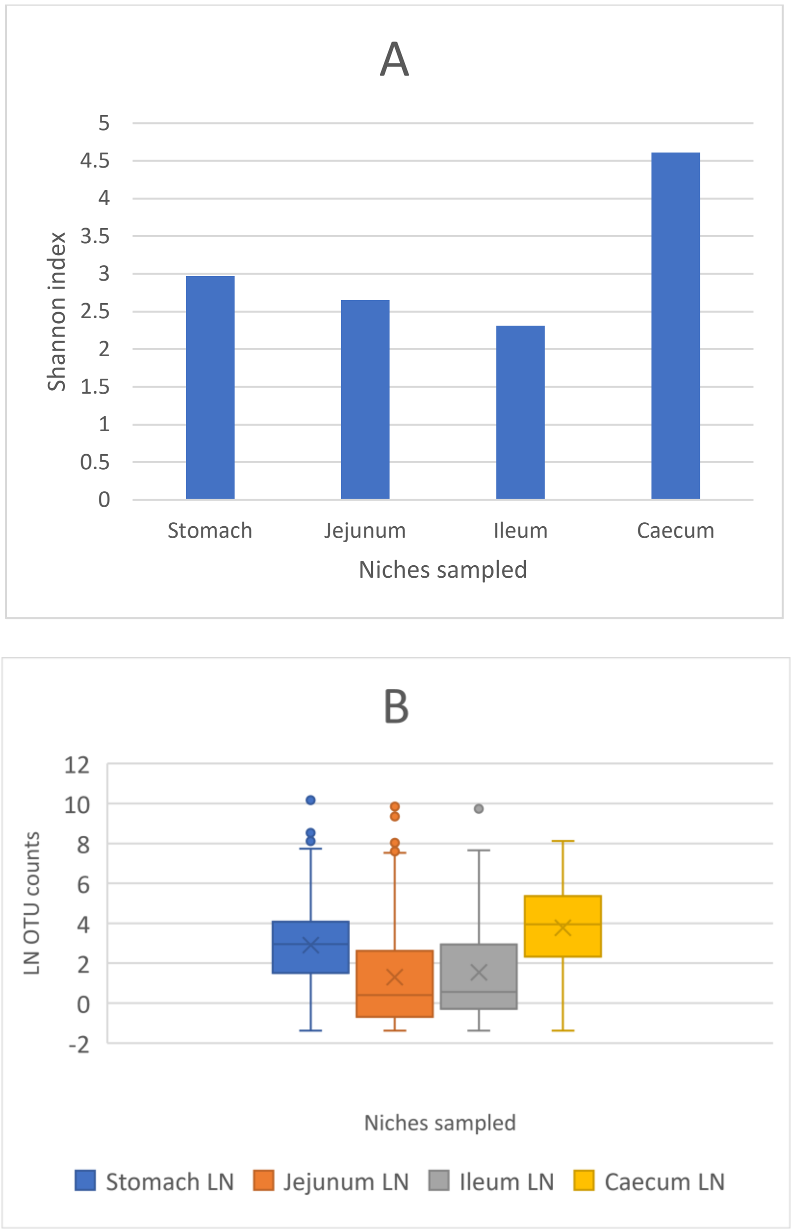


Figure 3-4: Alpha diversity analysis of each environmental niche of the GI tract in C57BL/6 mice. (A) Shannon Index indicating the higher richness of the caecum compared to all other sites analysed and (B) box and whisker plot showing the lower relative range of inter quartile values found in the small intestine compared to the stomach and the caecum.

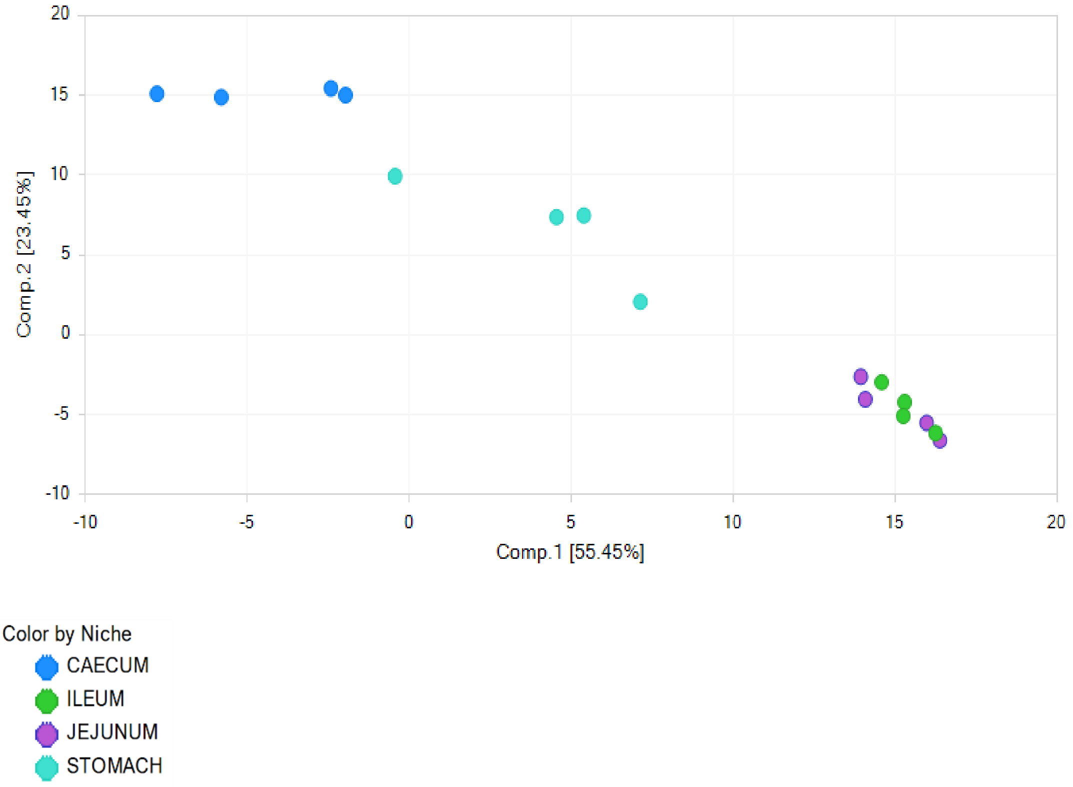


Figure 3-5: Principal component analysis of the OTUs identified at each GI niche in C57BL/6 mice. This shows the clustering of small intestine community structures together compared to those obtained from the stomach and caecum. PCA generated using Array Studio.

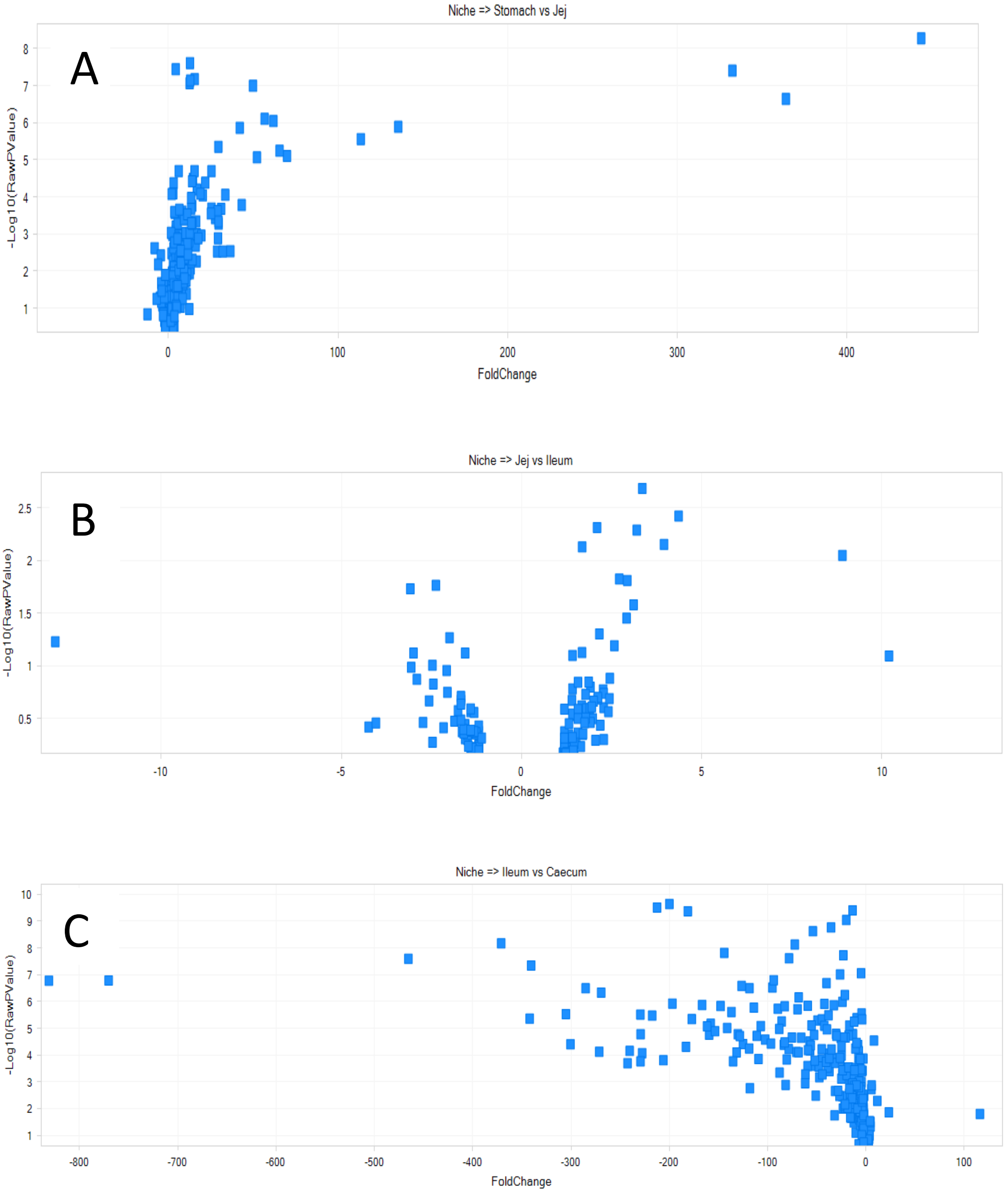


Figure 3-6: Volcano plots indicating changes in OTU numbers between GI niches. Plots generated in Array Studio by plotting -log10 raw p-value vs. fold change for stomach vs. jejunum (A), jejunum vs. ileum (B) & ileum vs. caecum (C) comparing counts in C57BL/6 mice samples taken at CRL only.

3.2.5 Effect of excretion upon community diversity

To determine if faecal material would be suitable as an indicator for colon diversity and therefore for long-term studies adhering to the 3R principles a relative abundance graph was generated comparing the OTUs obtained in the colon and from excreta (Figure 3-7). Using the mean count for each of the >300 OTUs identified in the five C57(Trim) mouse studied, no obvious difference was observed between the samples (raw p-value or FDR) when log+1 of counts were analysed for each OTU. This correlation is confirmed by Shannon diversity analysis (Figure 3-8A) and box and whisker comparison (Figure 3-8B). This similarity in excreted cage faeces and digesta from the terminal colon justifies the future use of excreted faecal material for microbiome studies revolving around the lower intestine rather than sacrificing animals.

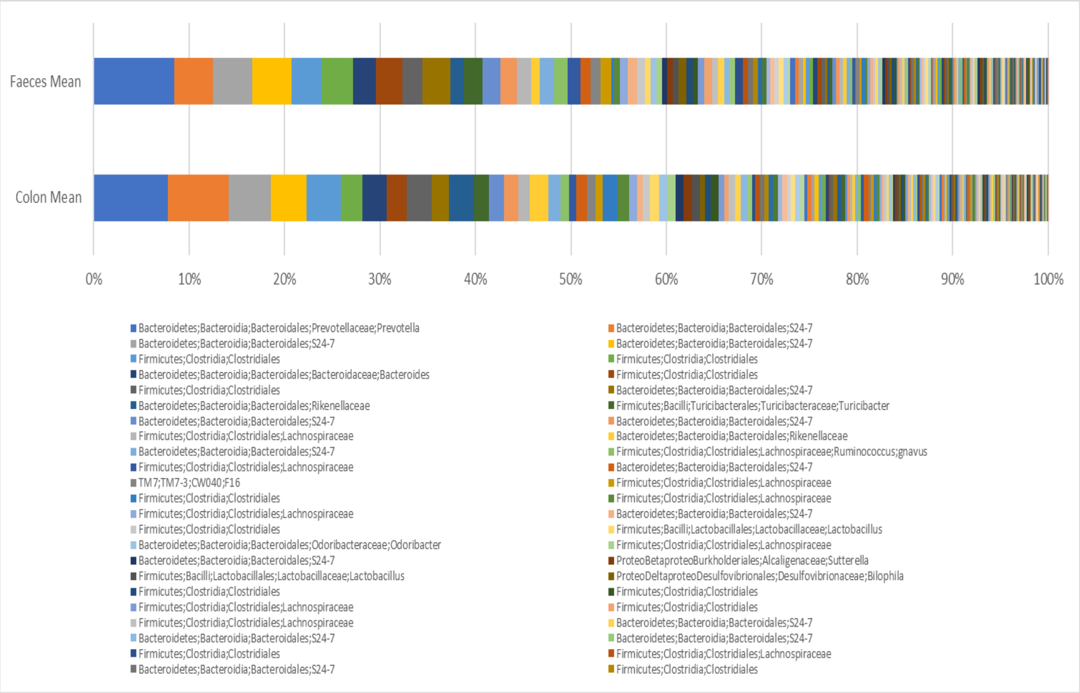


Figure 3-7: The relative abundance of OTUs identified in C57(Trim) colonic digesta and excreted faeces. This indicates a high level of community similarity between colonic material and that of excreted faeces.

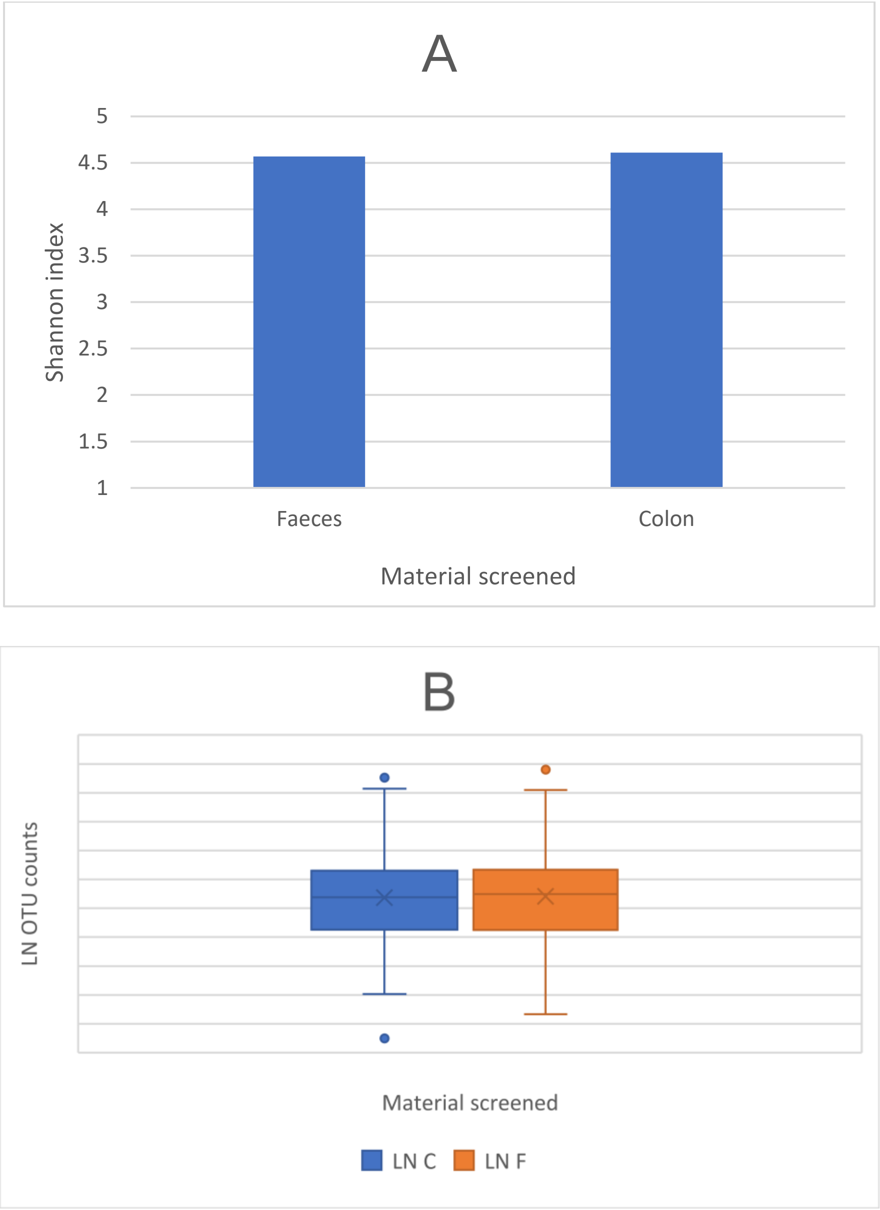


Figure 3-8: Alpha diversity analysis of colonic digest and excreted faeces in C57(Trim) mice. (A) Shannon Index indicating the similarity in richness between faeces and colonic digesta and (B) box and whisker plot indicating the high a slight increase in spread of data from excreted faeces over colonic digesta. LN C is natural log of colon counts and LN F is natural log of faecal counts.

3.2.6 Effect of sampling location and transport upon community diversity

To assess the effect of the location and the possible impact of transport between sites upon the microbial diversity and dynamics in mice prior to experimental procedures, identical samples were taken at the originating breeding facility in Maidstone (CRL, UK), upon arrival at GSK (Stevenage, UK) and after two weeks acclimatisation onsite at GSK. The raw p-values, fold changes and FDRs were calculated for each taxonomically designated OTU which changed following transport. It was found that 277 OTUs could be identified in caecal samples taken at CRL, while only 198 were identified in identical samples taken immediately post-arrival, seeing a loss of 28.6% of OTU diversity following a two hour journey. Two-week acclimatisation was not sufficient not for these OTUs to recover or rebound within samples. Of the OTUs that were lost during transport, thirty-three belonged to the *Bacteroidetes* phyla, suggesting that metabolic potential of the gut was altered in terms of complex carbohydrate harvesting (Haller, 2018), as it was shown in Figure 3-2 that this bacterial phyla is found with equal abundancy across GI niches, whereas it is the *Bacillota* which undergo immune sculpting (Santaolalla *et al,* 2012). Only five OTUs were found to change in population size during the two-week acclimatisation period. The C57(Jax) mouse GI tract hosted ~55-65% of its microbiota belonging to the *Bacillota* phyla, with ~25-30% belonging to the *Bacteroidota* across all GI tract niches. Post-delivery at GSK, the *Bacillota* were seen to rise proportionality within the GI tract to ~85-90%. Generation of a relative abundance graph, a homogenous caecal community, dominated by no single organism is evident at source in CRL

(Figure 3-9). The subsequent transport of these mice resulted in the permanent reduction of OTU diversity and succession of a dominant S24-7 phylotype in the caecum, along with an overall rise in OTUs belonging to the *Bacillota*. The dominant representative of the phylum *Bacillota* identified at CRL was a *Lactobacillus* sp. Numbers of this beneficial symbiont species were seen to fall upon transit with a potential negative impact to mucosal defence and the general wellbeing of the host (Nishiyama *et al*, 2016). The OTUs belonging to the *Bacillota* were observed to rise post-delivery, were all designated as members of the order *Clostridiales*.

Employing the calculation of the Shannon Index across the caecal samples it was found that the CRL samples exhibited high richness, but the transport of mice resulted in a decrease in microbial diversity post-arrival at GSK which did not rebound after two weeks acclimatisation (Figure 3-10A). The Box and whisker plot of these data illustrates a common median between counts from each site but a wider, and more dysbiotic spread in samples taken at GSK (Figure 3-10B).

Subsequent PCA analysis of the OTUs (Figure 3-11) obtained from sequencing pre, post-transport and after acclimatisation confirmed that GI microbiota following transit is highly divergent from the original diversity found in identical mice at source. To simplify analysis, the raw p-value was plotted against the fold changes (Figure 3-12) to illustrate the huge change in OTUs during transit (Figure 3-12A) and the relatively low level and even shifts in diversity between the two-week sampling occasions at GSK (Figure 3-12B).

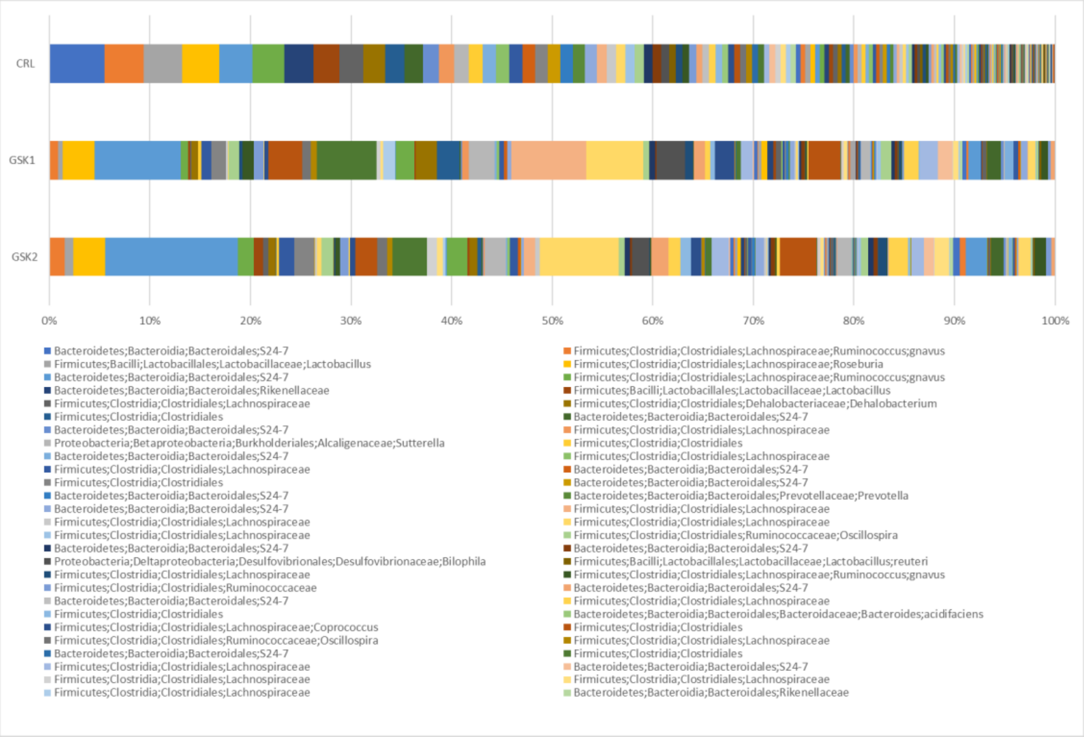


Figure 3-9: The relative abundance of OTUs identified in C57BL/6 mouse caeca taken at source, at delivery, and after 2wks acclimatisation. This indicates a decrease in prokaryotic diversity and increase in several dominant genera which did not return to its previous levels over the two-week housing period.

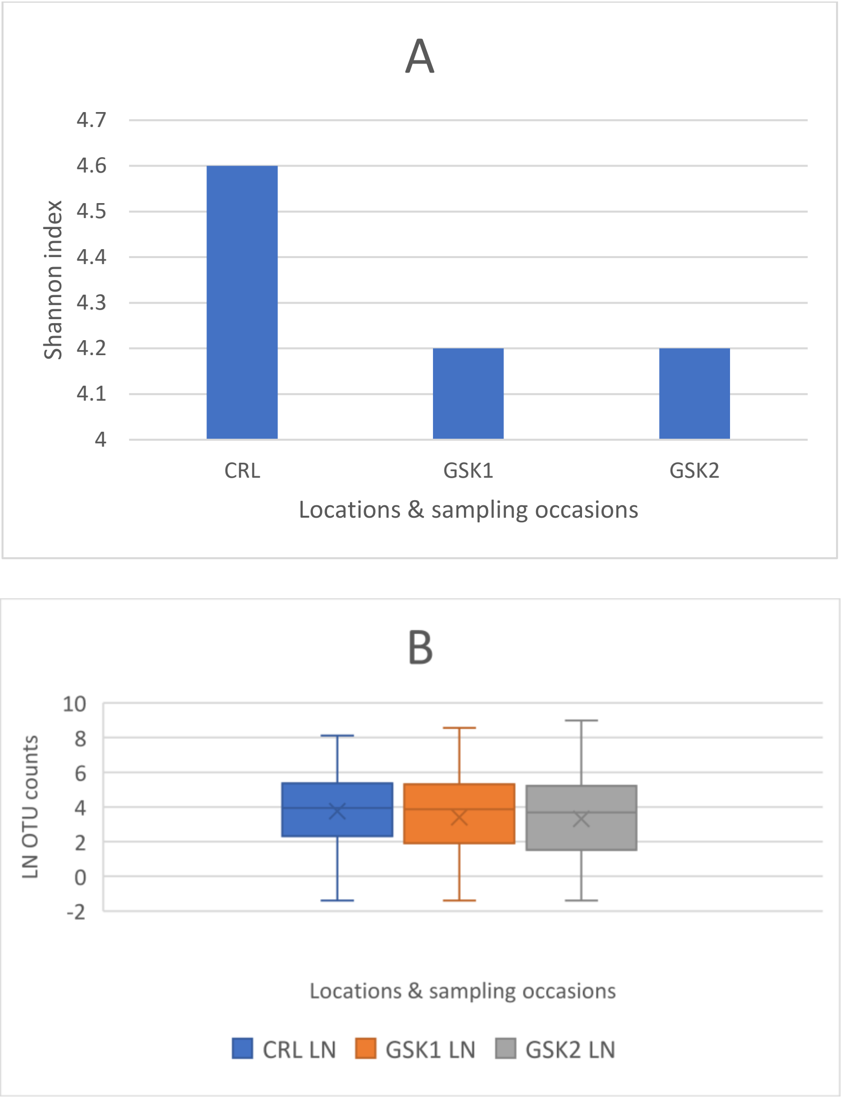


Figure 3-10: Alpha diversity analysis of caecal material taken at source, at delivery, and after 2wks acclimatisation. (A) Shannon Index indicating a reduction in prokaryotic richness and (B) box and whisker plot indicating the widening of data in the prokaryotic diversity of the caecum during the two post-arrival sampling opportunities compared to that derived from delivery boxes.

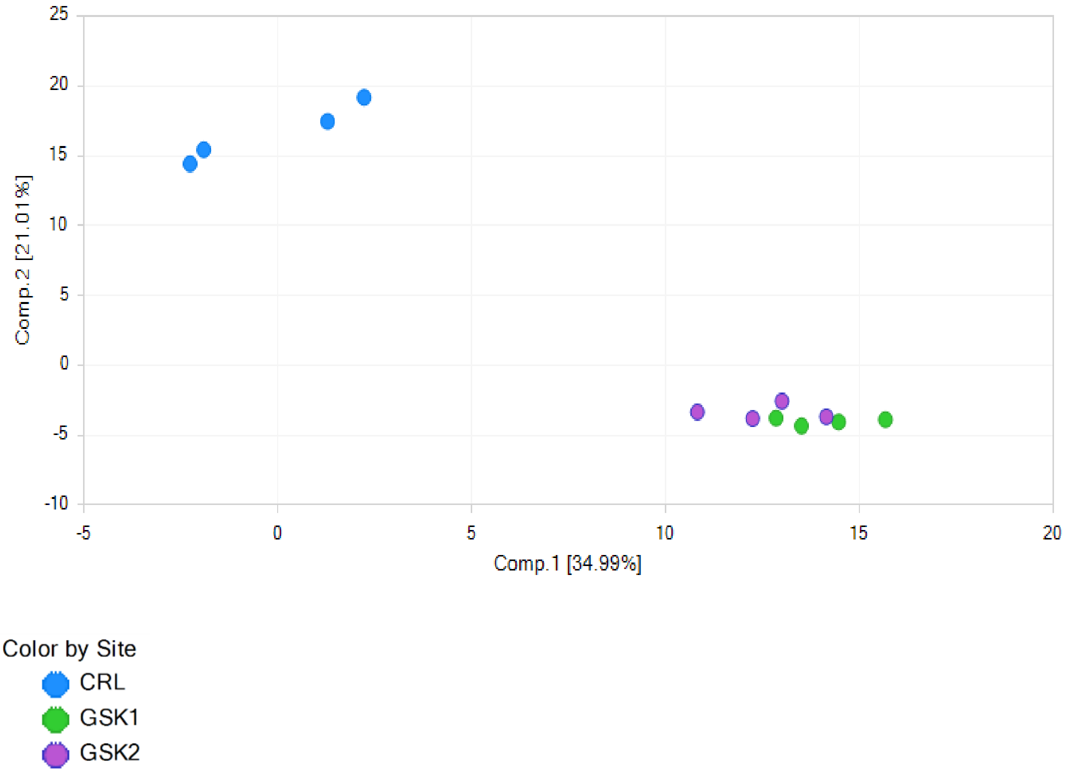


Figure 3-11: Principal component analysis of OTUs identified in caecal material taken at source, at delivery, and after 2wks acclimatisation. This indicates the clustering of results from material analysed post-arrival and the stability of the new microbiota over the following two weeks. PCA generated using Array Studio.

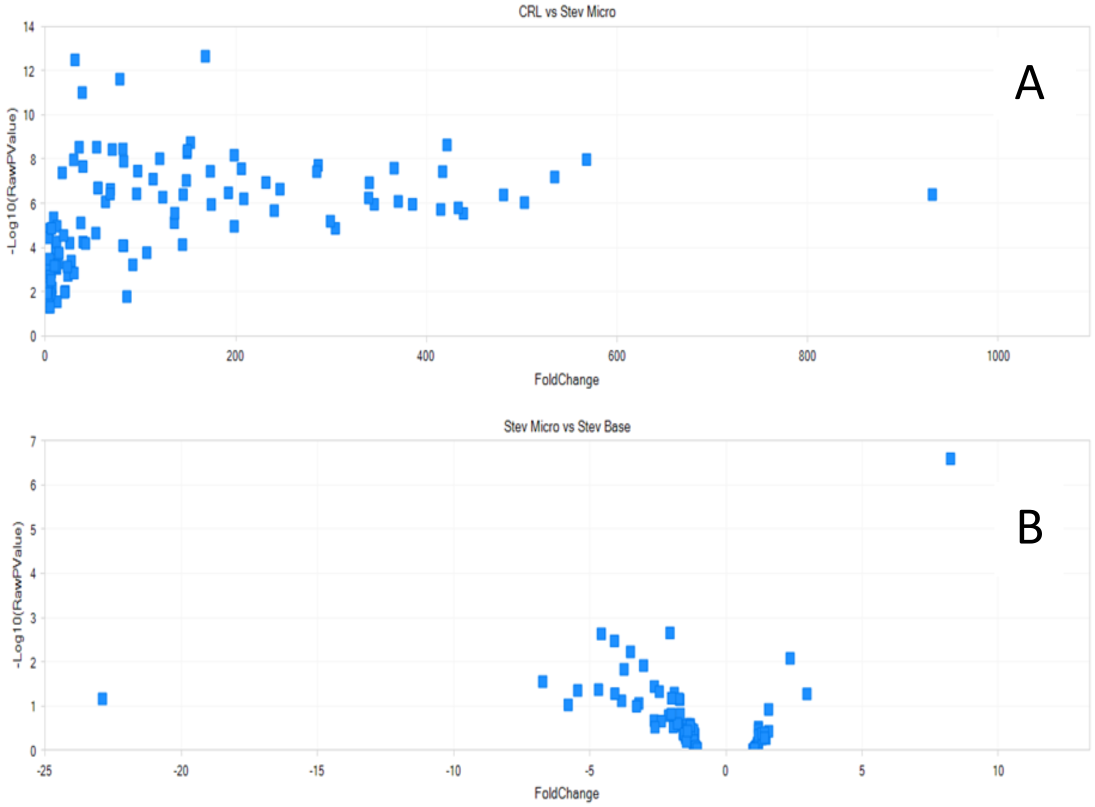


Figure 3-12: Volcano plots Volcano plots indicating changes in OTU numbers between sampling occasions. Plots generated in Array Studio by plotting -log10 raw p-value vs. fold change for CRL vs delivery at GSK (A), and between delivery and 2 weeks acclimatisation (B) comparing counts derived from C57(Jax) mice caecal samples taken at CRL & GSK.

3.2.7 Effect of four-month housing upon community diversity

To better assess post-delivery microbiota stability and to understand changes to the microbiota of mice during acclimatisation over a longer period, diversity was analysed from faecal samples obtained from cohoused, male NOD-SCID mice over a four-month period. In this study 1006 OTUs were identified by outsourced processing.

A change in diversity was illustrated between origin and the three subsequent sampling opportunities at GSK by creating a relative abundance graph (Figure 3-13). By analysing relative abundance graphs based on individual OTU counts from across the four-month period, the dominant S24-7 OTU was seen to increase from ~2% to ~10% of the populations, while the second and third most populous S24-7s fell in number along with a *Ruminococcus* sp. (Figure 3-13). In this study C57(Jax) and NOD-SCID samples were dominated by multiple S24-7 OTUs, while C57(Trim) samples were dominated by a single OTU belonging to the genus *Prevotella*.

However, Shannon index analysis indicated a similar richness across all sampling opportunities (Figure 3-14A) and box and whisker analysis indicates a slight reduction in diversity spread or a constriction of OTUs (Figure 3-14B). By conducting PCA analysis (Figure 3-15), the shift in diversity, was observed as a permanent change to a new defined microbiota once animals were received which could affect study outcome.

By analysing the p-values and FDR it was possible to observe a huge change in diversity affecting >200 OTUs between delivery in May to the first post-arrival screen in June. The shift between June and July showed twenty-two changes in OTU number (fifteen up and seven down), indicating a low-level change once mice are received. The same analysis between July and August at which point the study was terminated shows twelve significant changes which are all members of the *Bacillota* phylum indicting that the GI microbiota attained a new diversity during transit which did not change once animals were received.

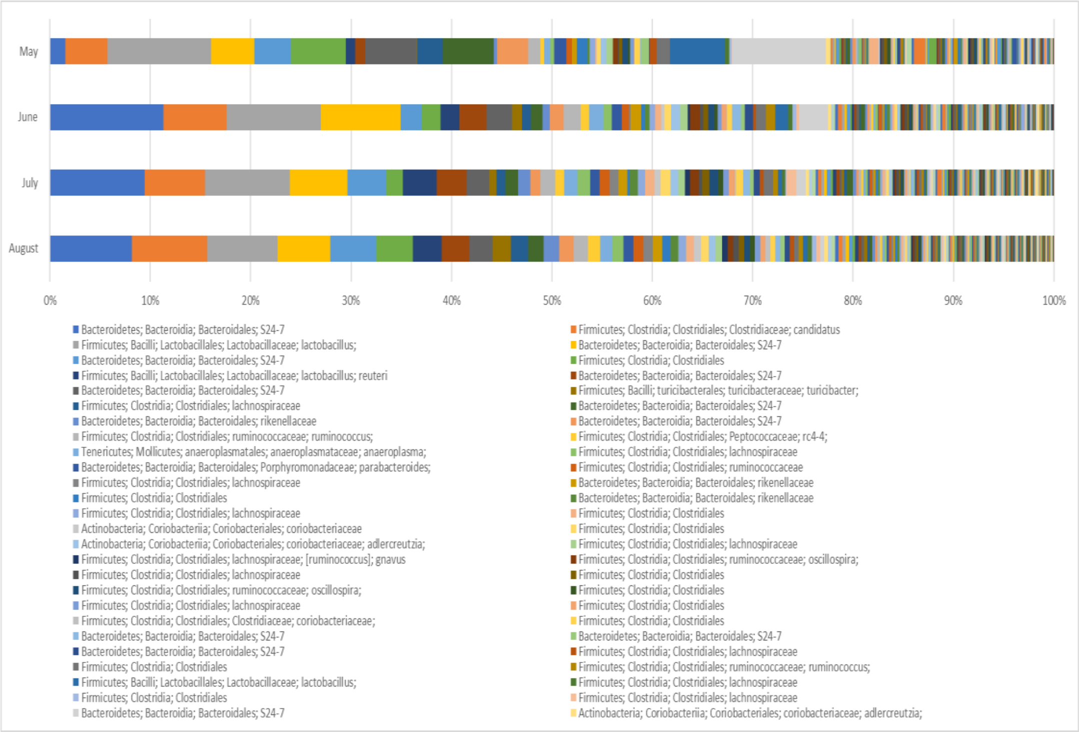


Figure 3-13: Relative abundance of OTUs identified in NOD-SCID faeces over a four-month period. This indicates the post-arrival changes in prokaryotic diversity which remained stable over the subsequent three months.

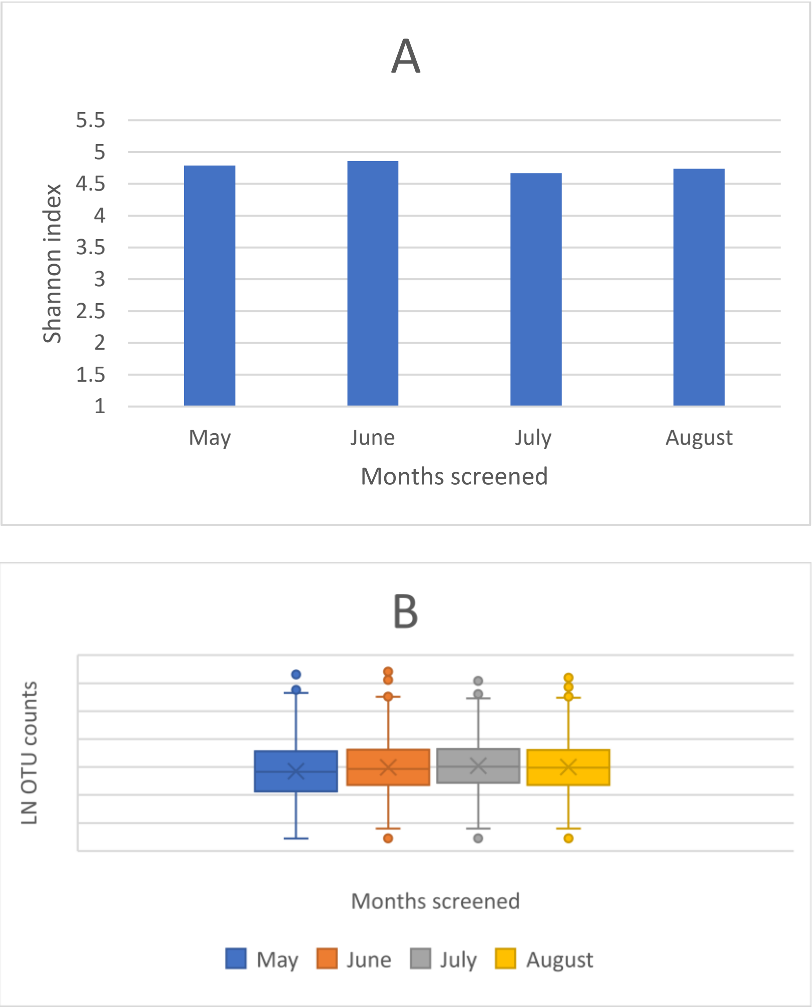


Figure 3-14: Alpha diversity analysis of NOD-SCID faeces over a four-month period. (A) Shannon Index indicating a similar richness across time and (B) box and whisker plot indicating the relative stability and tightening of data of prokaryotic diversities when this type of analysis is conducted.

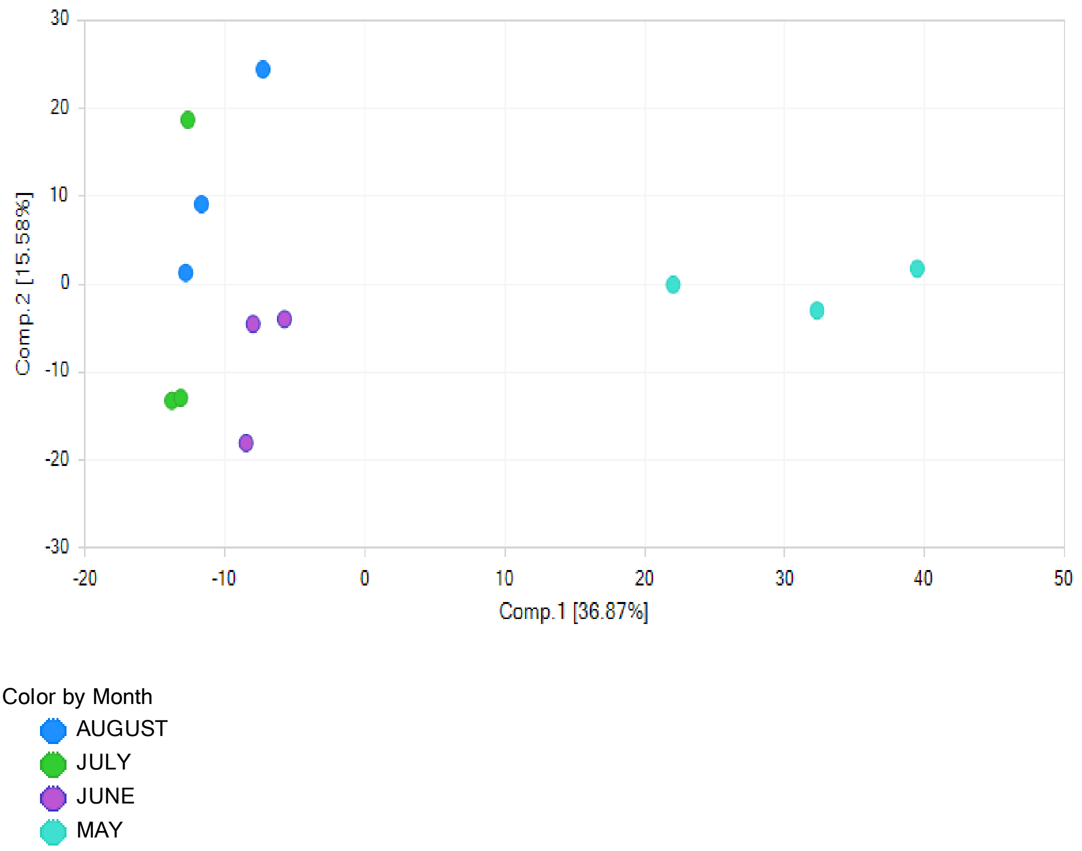


Figure 3-15: Principal component analysis of NOD-SCID faeces over a four-month period. This indicates a change in diversity post-arrival which remained stable for the subsequent three months. PCA generated using Array Studio.

3.2.8 Dynamic OTU behaviour during community disruption

The interpretation of OTU tables generated helped summarise changes in diversity across a large number of biological entities. Studying OTU behaviour at its numerical basis (individual NGS reads associated with a single OTU) can be informative during disruptive events rather than looking at proportional changes in OTUs or diversity within samples. Here, data illustrated how the majority of OTUs relating to a recently cultured bacterium, S24-7 (Ormerod *et al*, 2016), were lost during transit. However, by studying this data in greater detail (Table 3-3), OTU24 (green) was seen to be in low abundance at CRL but rose during transit, only to fall after 2-weeks acclimatisation where most S24-7 OTUs do not follow this pattern. OTU3 (yellow) was seen to rise dramatically during transit to become the most dominant community member. This dominance was seen to rise over the subsequent two weeks of acclimatisation. Through interrogation of these data at the level of sequencing reads, it was inferred that genera (or species as it was unclear which) within the S24-7 taxonomic family exhibit dynamic behaviour as the host animal underwent transport. Transport could have caused the host mice to experience stress which may have affected their microbiota including the S24/7 family. Table 3-3 shows that genera or species belonging to S24/7 react differently to these external influences, but the lack of diagnostic clarity prevents any understanding of which genera or species contract or proliferate under these specific conditions.

Table 3-3: Multiple S24-7 OTU designations and OTU counts for C57BL/6 mouse faeces taken at source, at delivery, and after 2wks acclimatisation. indicating the decrease in numbers of one OTU post-arrival (24) and the increase in another (3) over the same period indicating specific lower-classification level behaviour.

	Caecum											
OTU	CRL				GSK (delivery)				GSK (2wks acclimatisation)			
20	509	450	1473	1417	1	4	1	2	0	12	0	0
209	74	75	84	94	0	0	0	0	0	2	0	0
222	9	1	1	0	90	80	111	56	16	4	42	51
228	75	42	23	30	0	0	0	0	0	0	0	0
24	487	306	937	1054	1	1	0	0	0	13	3	0
243	143	110	586	588	1	2	0	0	0	0	1	0
25	1002	808	619	648	1	1	0	0	1	11	3	1
29	373	345	174	272	0	1	1	0	2	27	4	0
298	1	1	0	0	2	0	0	0	3	4	6	0
3	1043	1327	2824	2843	8333	5060	2460	4905	4182	6049	11220	10560
300	162	261	217	205	0	1	0	1	0	3	1	0
302	1	0	0	0	0	0	0	0	0	0	0	0

3.3 Summary

To investigate the suitability of an outsourced 16S rRNA amplicon sequencing approach for the routine elucidation of microbial diversity in the murine GI tract during drug discovery and validation studies, five comparative analyses were conducted from three *in vivo* sampling opportunities. This work enabled the design of a sampling strategy and allowed the analysis of data from a standardised 16S rRNA gene analysis NGS protocol. In addition to providing experience in data handling and the development of an easily implemented pipeline, it illustrated the effect several basic variables had upon resulting data which would lead to informed sampling decisions and strategies in the future.

Firstly, the effect host sex plays upon community diversity was assessed. This was to determine whether samples from both sexes would be representative of given murine population. Secondly, four sites along the murine GI tract were analysed to illustrate local community structure, indicating possible function. Following this, a comparison was made between digesta from the terminal colon and cage faeces to assess the affect elimination and environmental exposure has upon this material. The results of this determined whether samples could be generated in-life rather than only terminally acquired. The effect of a change in host location was then assessed and finally, the longevity of any community changes was explored.

It was found that community diversity was independent of host sex. This confirms the work of Wallace *et al* (2018), who also found that the GI microbiota did not alter through the oestrous cycle in C57(Jax) mice when analysed using a 16S rRNA gene analysis approach. Each GI niche was shown to contain an independent community, with richness and diversity being driven by, and also contributing to, the local environmental conditions. Here the increasing biomass/decreasing pH model descending the GI tract asserted by Walter & Ley (2011) was not observed as the stomach was found to host a numerous and diverse community, second only to that of the caecum. It was found that excreted faeces do not undergo any post-excretion changes during a period of up to 18hrs. This indicates that excreted material is representative of the lower GI tract, removing the need to sacrifice animals in pursuit of lower GI community characterisation. GI community disruption has been illustrated after long-term (5-day) transport across the USA (Montonye *et al,* 2018). However, no investigations have been conducted into the effect of moderate (~22hr) transit across the 'research triangle' in the Southeast of England. The eradication of nearly one third of OTUs identified (mostly Bacteroidetes) in caecal samples taken at CRL during transit would likely have an impact on the metabolic potential for community members and the host alike. The switch to a secondary community structure was subsequently found to continue for up to four months. These findings indicate that transport-induced changes in GI diversity, or dysbiosis may affect the functional role played by the prokaryotic population in the gut and this effect is long-lived. This may also contribute to the lack of reproducibility in *in vivo* studies conducted across distinct locations (Montonye *et al,* 2018).

In relation to the observation of S24-7, the use of the GreenGenes reference database failed to drop below the family level of classification, masking clear species or genera traits. It also fails to supply relevant taxonomic information. The culturing of S24-7 (Ormerod *et al*, 2016) would have done little to illuminate the breadth of this recently re-assigned family which has >650 species and is now known as the *Muribaculaceae* (Lagkouvardos *et al,* 2019). Utilising reference databases which are not updated as discoveries are made inhibits understanding and confuses research.

Although community structures were discernible during this work, the use of a standardised approach to taxonomic characterisation in conjunction with an out-of-date reference lacked granularity which hampered functional understanding and clinical significance. These attributes are essential in diagnostic microbiology and therefore a more refined analysis of the NGS data was still required.

Chapter 4: NGS Method Comparison using MG-RAST

4.1 Introduction

In the previous chapter 16S rRNA gene amplicon characterisation of excreted murine faeces was shown to be equally representative of colonic digesta in illustrating community diversity and dynamics in a range of conditions. However, the data generated during this fully outsourced process, using the QIIME analysis methodology (Kuczynski *et al,* 2011) lacked diagnostic clarity below the family level for this study. The Greengenes reference database which was employed to assign nomenclature was based on data from 2013 (McDonald *et al,* 2012). Although the breadth of microbiological information obtained was wider than that possible through classical culture, the failure to identify genera and species ultimately inhibited the understanding the clinical relevance or biological function of a member of the population. The lack of this information limits the usefulness of these data in terms of animal husbandry studies and the deployments of these methods within the specific Good Laboratory Practice and Veterinary landscapes of the pharmaceutical industry.

To address this, a study was designed to determine if the processing of identical starting material (cage faeces) through 16S rRNA amplicon sequencing, DNA metagenomic, and RNA-seq transcriptomic NGS workflows and subsequent analysis of the resulting data through a single open source bioinformatic tool (MG-RAST) would provide diagnostic data of an improved quality which could be compared between workflows and utilised in an expansive routine health monitoring regime (Meyer *et al,* 2008). This improvement would be firstly assessed by the diagnostic granularity of all outputs by employing a range of reference databases. Secondly, below species level, gene function and spatial and temporal control would be assessed by applying relevant functional databases to the analysis.

MG-RAST (Meyer *et al,* 2008) is a free-to-use tool that can be used by operators with only basic knowledge of bioinformatics. It provides the analysis of all types of sequencing data via a common interface, enabling parameterisation to be controlled. This would potentially enable easier deployment in a routine setting. Registered users can use publicly available datasets or upload their own data sets can be run concurrently through any a range of pipelines to obtain rapid interpretation of data for both taxonomy, abundance, and function. Once data is uploaded it is quality checked before analysis (.fasta and .fastq). A single, logical graphical user interface (GUI) makes this resource easy to understand and make successive, related analyses for comparison. Analysis takes seconds to complete, even if multiple database searches are requested. Major limitations on MG-RAST are that prokaryotic hits below the genera level and all eukaryotic results should be viewed as inaccurate (Meyer *et al*, 2008). The SILVA database (www.arb-silva.de; (Quast *et al,* 2013), is a manually curated compendium of bacterial, archaeal, and eukaryotic genetic information and suite of bioinformatic tools. Data is divided into small (16S rRNA/18S) and large (23S/28S) prokaryotic/eukaryotic ribosomal subunit sections (SSU & LSU). The 2012 release contains >3x10^6 prokaryotic and ~3x10^5 eukaryotic entries (Quast *et al,* 2012). The RDP database (www.rdp.cme.msu.edu) is based on 16S rRNA sequences from Bacteria, Archaea and Fungi (Eukarya) and tools. It contains

~$3x10^6$ 16S rRNA sequences and ~$7x10^5$ fungal 18S sequences. It is mostly made of incomplete sequences derived from PCR amplification (Cole *et al,* 2013). The Greengenes taxonomic database (www.greengenes). Second genome.com) is comprised of bacterial and archaeal 16S rRNA genes and contains ~$5x10^5$ full length 16S rRNA entries (McDonald *et al,* 2012). The NCBI taxonomic database (www.ncbi.nlm.nih.gov/refseq/) is manually curated and updated daily. It contains the names of all accession numbers or organisms associated with submissions to the NCBI sequence database (Pruitt *et al*, 2006). It represents the primary information source for the biotechnology community and is the hub for an integrated web of associated databases and tools. The Kyoto Encyclopedia of genes and genomes of KEGG server (www.genome.jp/kegg/) is a comprehensive, interlinked web of databases and tools for the interpretation molecular data at the functional level. It was instigated in 1995 as the need grew to understand gene function networks. Functions are stored in the KEGG Orthology (KO) database, where each is defined as an ortholog of other genes. Networks, relationships, and pathways are made visible and open further links and interactions.

Using data obtained from this study, a comparison of these databases was employed to identify which is most suitable for use in semi-automated analysis of microbiome data as part of routine monitoring of *in vivo* study data.

4.2 Results

4.2.1 NGS data handling

Fourteen physical samples of digesta were submitted to Genewiz for 16S rRNA gene analysis, DNA metagenomics, and RNA-seq NGS processing (utilising co-extraction of nucleic acids). Genewiz indicated that sample number 8 (stomach of 495-87.1 in Cage 2) was contaminated in the laboratory during library preparation. The data for this sample are included in the analysis but must be treated with caution, although this known contaminant could have been deleted. Each NGS method created twenty eight (paired end) read files, totalling eighty four final raw .fastq sequence files for download from the DNAnexus server (Appendix 9.2).

4.2.2 Comparison of taxonomic assemblies using MG-RAST.

To assess the accuracy of each taxonomic database available on MG-RAST, the 16S rRNA, DNA metagenomic, and RNA-seq data from sample number 1 (stomach of mouse 432-87.1, cage 1) was analysed using all options and then examined at the phylum and generic level. The 16S rRNA data cannot be processed through RefSeq on MG-RAST so taxonomic descriptions were found by running DNA metagenomic (12 x10^6 hits) or RNA-seq (0.7x10^6 hits) data sets through the suite. The data indicated *Bacteroidota*, *Bacillota* and *Verrucomircobiota* were the major phyla present, while the major genera identified were the *Bacteroides, Clostridium* and *Prevotella* (Figure 4-1). The

Greengenes database generated the most useful output from RNA-seq data (>2x10^6 hits), while DNA metagenomic data generated <10x10^4 hits. The 16S rRNA amplicon data generated ~4x10^5 hits, with the major phyla identified being *Bacteroidota, Bacillota* and *Pseudomonadota*, and the major genera identified as unclassified, *Barnsiella* or *Clostridium*. This was unexpected as Greengenes is a rRNA database. By using the Ribosomal Database Project database, the highest number of hits processing of RNA-seq data (>3x10^6 hits), while DNA metagenomics generated <10x10^4 hits. 16S rRNA data generated ~4.5x10^5 hits. The 16S rRNA data indicated that the major phyla were again *Bacteroidota, Bacillota* and *Pseudomonadota*, the major genera decerned were *Barnsiella, Clostridium* or unclassified. The SILVA SSU database again yielded the most hits from processing of RNA-seq data (>5.1x10^6 hits), while DNA metagenomics generated <10x10^4 hits. 16S rRNA data generated ~5x10^5 hits. The 16S rRNA data indicated that the major phyla were *Bacteroidota, Bacillota* and *Pseudomonadota*, while the major genera decerned were again unclassified, *Barnsiella* or *Clostridium*. While the SILVA LSU database cannot process 16S rRNA (small subunit data) but exceeds all other databases in the processing of RNA-seq data (>30x10^6 hits), while DNA metagenomics generated <5x10^5 hits. The data from RNA-seq indicated that the major phyla were *Bacteroidota, Bacillota* and *Chordata*, while the major genera decerned were *Homo*, *Clostridium* and *Flavobacterium*. Indicating that there was obvious human contamination in this sample.

Comparing the taxonomic data generated in Figs. 1 – 5 (above) illustrated the failure of MG-RAST to generate consensus between databases and datasets.

This failure meant that this service could not be employed as a universal platform that can be employed for rapid analysis of routine microbiome data. RefSeq and Silva LSU databases did not recognise 16S rRNA amplicon-based sequences and Greengenes, RDP and SILVA (LSU and SSU), which are RNA-based gene databases, failed to generate extensive results from metagenomic DNA datasets. Additionally, RefSeq, which is a DNA sequence database could not provide taxonomy from RNA-seq datasets which may be because MG-RAST cannot align 5' to 3' DNA reads against complimentary RNA sequence reads. Excluding RefSeq, processed RNA-seq data generates higher numbers of taxonomic designations than the more often cited method, DNA metagenomics. A metagenomic approach, in terms of read numbers generating designated hits fell between amplicon-based and transcriptomic methods. It therefore illustrated the general inconsistency between data generated through different analysis pipelines originating from different sequencing methods. This was possibly due to RNA-seq data arising from organisms that are actively growing in a particular niche versus the sequencing of total DNA which could come from a range of sources; actively growing organisms within that niche, dead cells, DNA from food consumed and contaminants from environmental sources or even rRNA copy number in different prokaryotic species (Klappenbach *et al*, 2001). Some of these approaches also limited further analysis such as Shannon Indexing or biogeographical assessment which would have been useful in terms of drug discovery and the routine deployment of these methods within the pharmaceutical industry.

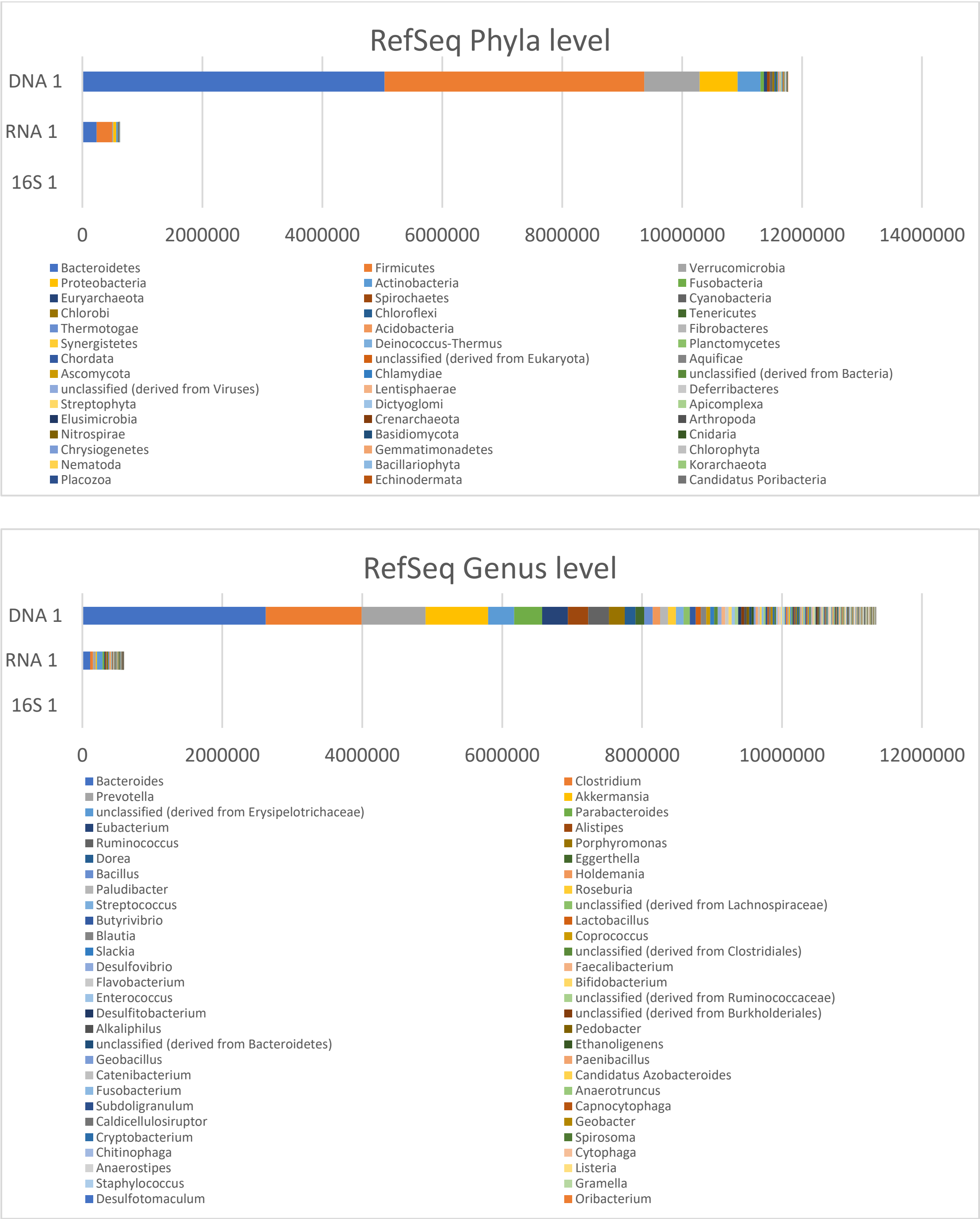


Figure 4-1: The phyla and genus level diversity of C57(Trim) stomach digesta using the RefSeq database. Generated by passing 16S rRNA, DNA metagenomic and RNA-seq data through the NCBI RefSeq database on MG-RAST.

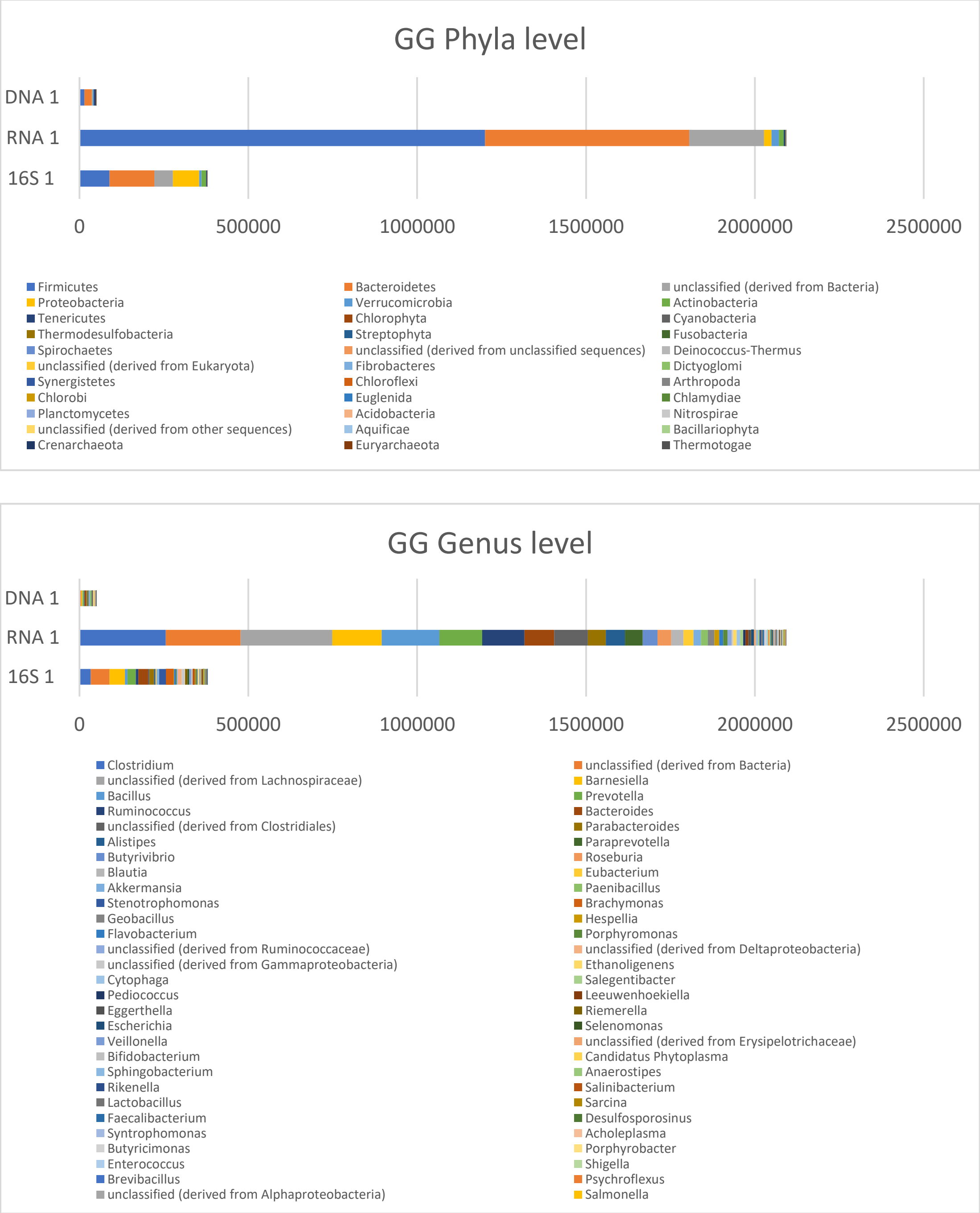


Figure 4-2: The phyla and genus level diversity of C57(Trim) stomach digesta using the GreenGenes database. Generated by passing 16S rRNA**,** DNA metagenomic and RNA-seq data through the Greengenes database on MG-RAST.

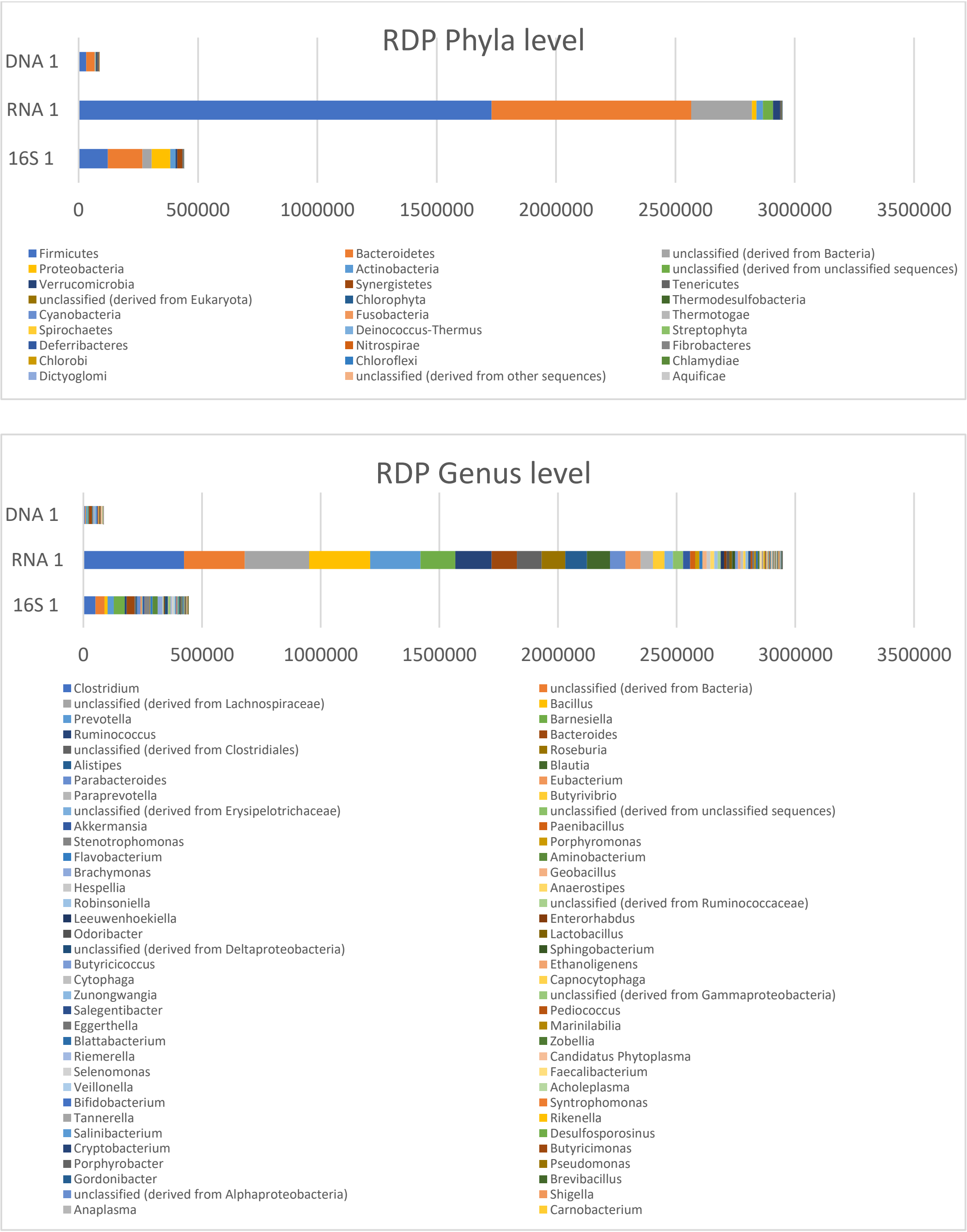


Figure 4-3: The phyla and genus level diversity of C57(Trim) stomach digesta using the Ribosomal Database Project database. Generated by passing 16S rRNA, DNA metagenomic and RNA-seq data through the RDP database on MG-RAST.

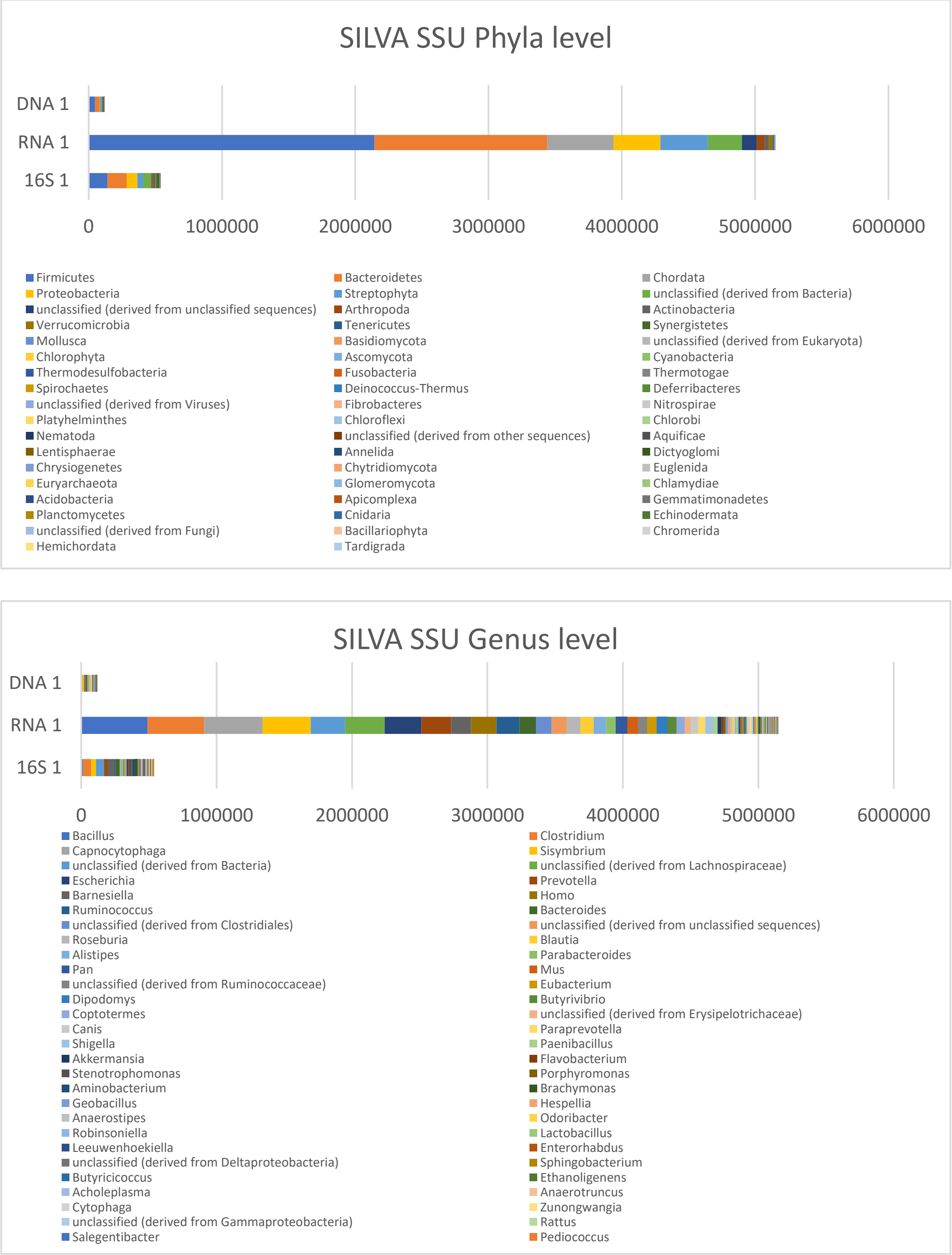


Figure 4-4: The phyla and genus level diversity of C57(Trim) stomach digesta using the SILVA SSU database. Generated by passing DNA metagenomic and RNA-seq data through the SILVA SSU database on MG-RAST.

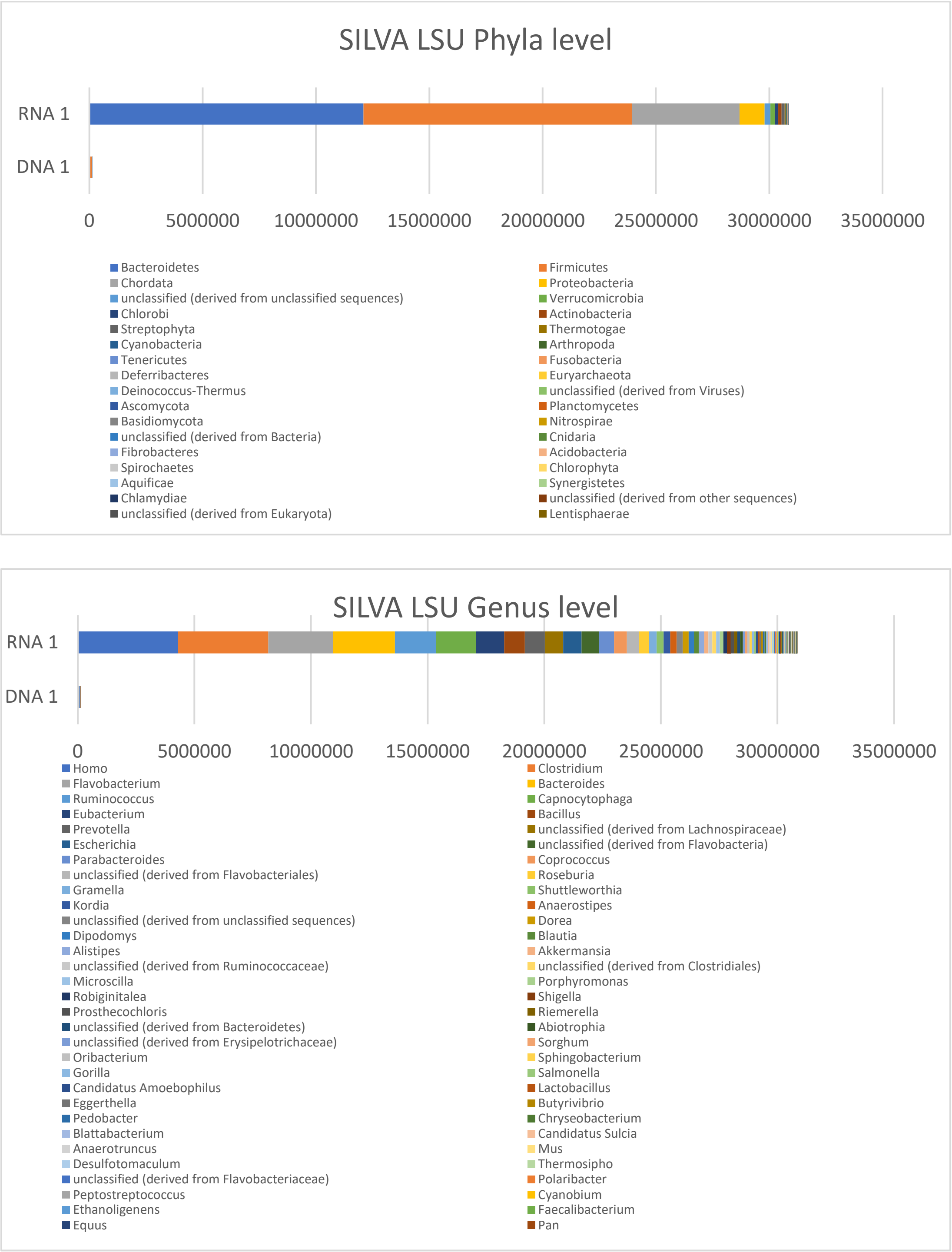
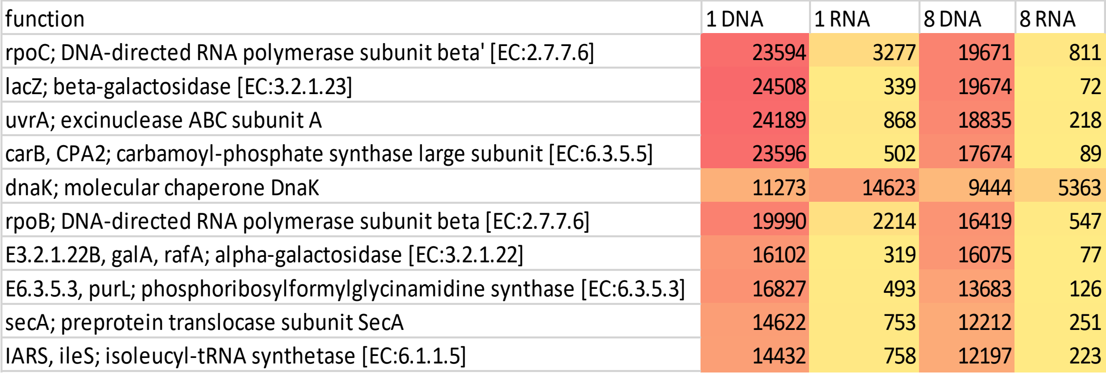


Figure 4-5: The phyla and genus level diversity of C57(Trim) stomach digesta using the SILVA LSU database. Generated by passing DNA metagenomic and RNA-seq data through the SILVA LSU database on MG-RAST.

4.2.3 Comparison of functional assemblies using MG-RAST.

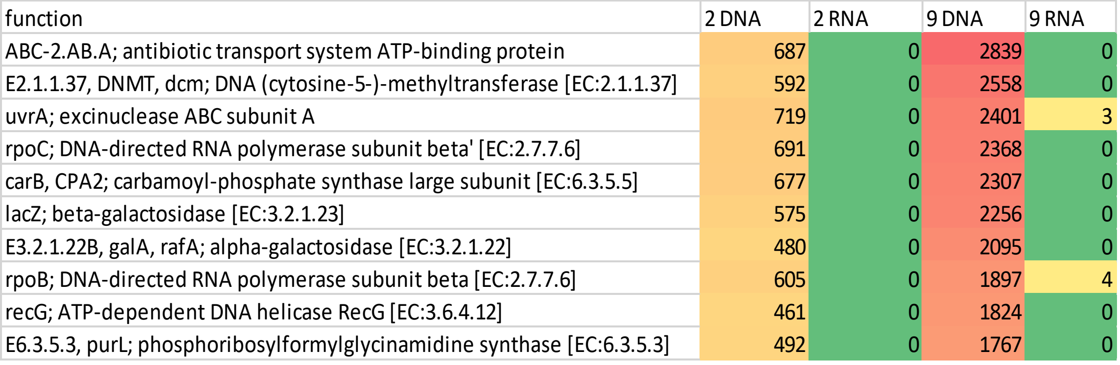
To rapidly assess the output from processing large (20-50Gb) DNA metagenomic, and RNA-seq raw sequencing files using the MG-RAST suite of tools, only the KO analysis option was chosen. Firstly, this was to overtly gauge prokaryotic gene activity (potential and actual) in each sample. Secondly, this was attempted to see whether the results for a single sample would produce a functional agreement between the methods of data generation. To illustrate the results, the highest ten gene associations (or counts) for stomach samples (1 and 8) were shown (Table 4-1; in descending order, along with the ten highest counts for jejunum samples 2 and 9 (Table 4-2). The counts for both DNA metagenomics and RNA-seq (RNA) analysis procedures are given for comparison. In the stomach samples, DNA counts are higher than associated RNA counts, suggesting overall transcriptional activity of organisms in the stomach may be lower. The remaining 2515 results show a similar numerical relationship. The jejunum samples showed a lower level of genomic counts but an almost complete absence of expressed RNA counts. This numerical divergence was seen in the remaining 5464 results. As this may have been due to functional suppression in the small intestine, raw sequence files for all sample types were passed through KO analysis. The lack of RNA counts (and therefore expression) was seen in all samples taken from below the pyloric sphincter. This failure to genetically characterise each sample by DNA and RNA databases prevented identity confirmation and negated any further numerical or functional comparison between DNA and RNA data sets.

Table 4-1: The highest (mean average) DNA and RNA functional counts derived from C57(Trim) stomach samples (1 & 8). Generated by passing metagenomic and RNA-seq data sets through the KO facility on the MG-RAST server.



function	1 DNA	1 RNA	8 DNA	8 RNA
rpoC; DNA-directed RNA polymerase subunit beta' [EC:2.7.7.6]	23594	3277	19671	811
lacZ; beta-galactosidase [EC:3.2.1.23]	24508	339	19674	72
uvrA; excinuclease ABC subunit A	24189	868	18835	218
carB, CPA2; carbamoyl-phosphate synthase large subunit [EC:6.3.5.5]	23596	502	17674	89
dnaK; molecular chaperone DnaK	11273	14623	9444	5363
rpoB; DNA-directed RNA polymerase subunit beta [EC:2.7.7.6]	19990	2214	16419	547
E3.2.1.22B, galA, rafA; alpha-galactosidase [EC:3.2.1.22]	16102	319	16075	77
E6.3.5.3, purL; phosphoribosylformylglycinamidine synthase [EC:6.3.5.3]	16827	493	13683	126
secA; preprotein translocase subunit SecA	14622	753	12212	251
IARS, ileS; isoleucyl-tRNA synthetase [EC:6.1.1.5]	14432	758	12197	223

Table 4-2: The highest (mean average) DNA and RNA functional counts derived from C57(Trim) jejunum samples (2 & 9). Generated by passing metagenomic and RNA-seq data sets through the KO facility on the MG-RAST server.



function	2 DNA	2 RNA	9 DNA	9 RNA
ABC-2.AB.A; antibiotic transport system ATP-binding protein	687	0	2839	0
E2.1.1.37, DNMT, dcm; DNA (cytosine-5-)-methyltransferase [EC:2.1.1.37]	592	0	2558	0
uvrA; excinuclease ABC subunit A	719	0	2401	3
rpoC; DNA-directed RNA polymerase subunit beta' [EC:2.7.7.6]	691	0	2368	0
carB, CPA2; carbamoyl-phosphate synthase large subunit [EC:6.3.5.5]	677	0	2307	0
lacZ; beta-galactosidase [EC:3.2.1.23]	575	0	2256	0
E3.2.1.22B, galA, rafA; alpha-galactosidase [EC:3.2.1.22]	480	0	2095	0
rpoB; DNA-directed RNA polymerase subunit beta [EC:2.7.7.6]	605	0	1897	4
recG; ATP-dependent DNA helicase RecG [EC:3.6.4.12]	461	0	1824	0
E6.3.5.3, purL; phosphoribosylformylglycinamidine synthase [EC:6.3.5.3]	492	0	1767	0

4.2.4 Assessment of DNA metagenomics analysis using MG-RAST.

Unlike RNA-seq data generated in parallel, DNA metagenomic data appeared to offer a robust (if not confirmed by a second method) characterisation of the potential identification of gene function of the samples analysed. The KO workflow allowed results to be exported as .xls files, enabling simple filtering and graph rendering in Excel allowing easy deployment in a laboratory setting. This gave a clear and comparable indication of the most common prokaryotic genes in each sample. The most common genes identified were unsurprising with many of these highly conserved across prokaryotes such as the RNA polymerase subunits (*rpoC* and *rpoB*), the genes that are essential for protein secretion across all domains of life such as *secA*, genes associated with DNA replication and repair (*uvrA and dnaK*) and genes associated with core purine (DNA/RNA synthesis) metabolism such as *purL* (Table 4-1 and 4-2). Within the MG-RAST tool kit (https://www.mg-rast.org/mgmain.html.) the identity and function of sequence entities can be visualised on Krona http plots or as members of KEGG database function maps (Meyer *et al,* 2008). These link across the multiple databases in KEGG, allowing an understanding of function, structure, and importance in disease of a single gene. One example of a metabolic gene is *lacZ*, encoding the beta-galactosidase (EC:3.2.1.23), which was most highly represented in the gene complement of the stomach samples analysed (Figure 4-6). The *lacZ* gene product, beta-galactosidase and it is unsurprising that this ubiquitous gene was found to be the highest ranking in the results. Due to the huge size of data generated during DNA metagenomics (>2.5x10^6 hits per sample), classes of genes were more easily captured and communicated by this activity. Functional analysis of these sequences by KEGG allows the potential identification of genes involved in metabolism of xenobiotics which could affect metabolism of drug molecules within trials (Figure 4-7).

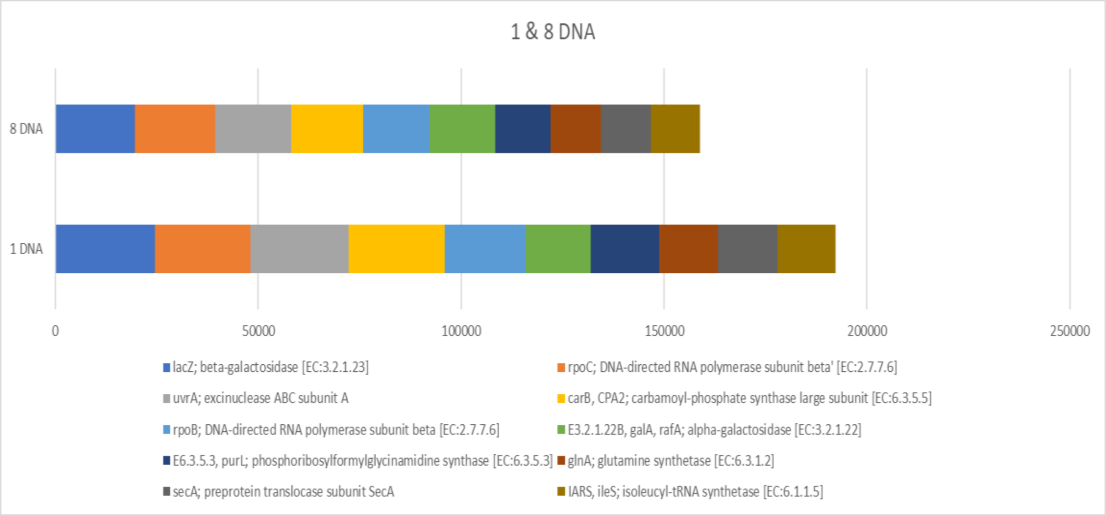


Figure 4-6: The relative abundance of prokaryotic gene counts and identities from C57(Trim) stomach digesta samples (1 & 8). Generated by passing metagenomic and RNA-seq data sets through KEGG database on the MG-RAST server suite indicating the robustness of this method**.**

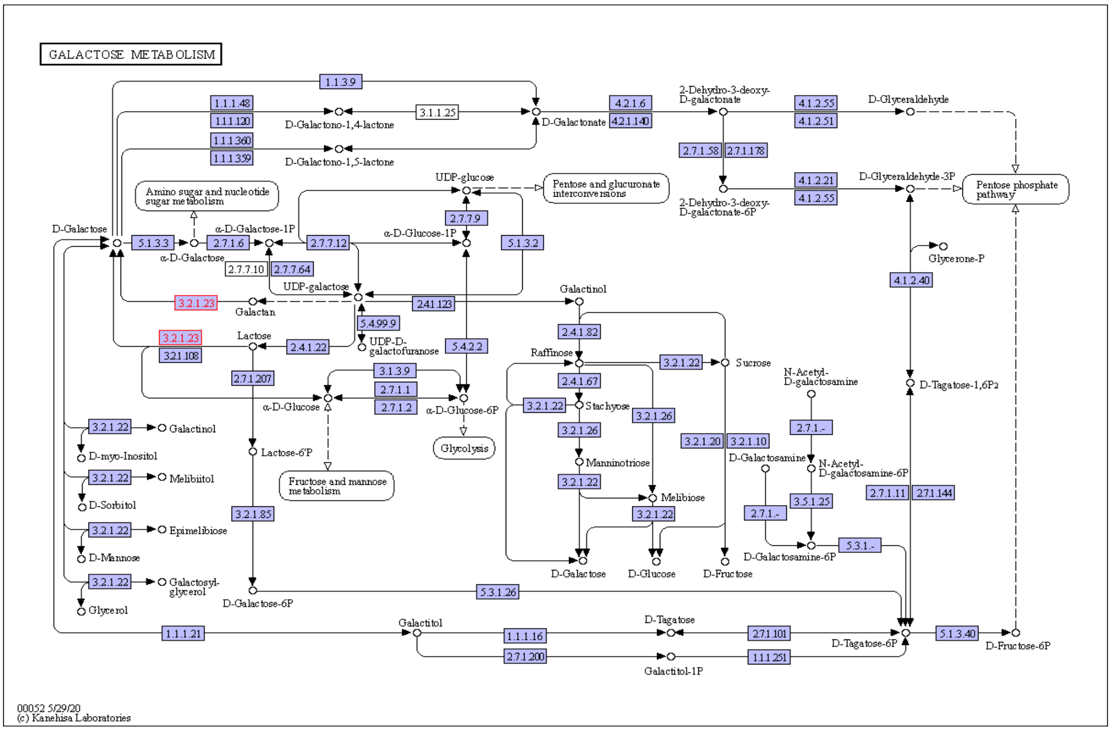


Figure 4-7: A KEGG Orthologue gene function map indicating the position of *lacZ* gene (EC.3.2.1.23). This map shows the catabolism of lactose, identified via the MG-RAST analysis of stomach samples (1 & 8) derived from DNA metagenomic sequence data indicating the potential for highly granular functional output.

4.3 Summary

In previous work, a significant shortfall was observed in the granularity of taxonomic data produced using an outsourced 16S rRNA approach to sample characterisation. This limited the diagnostic clarity which subsequently inhibited the functional understanding of a biological entities identified from digesta samples. This limitation was due to an outdated reference database being used in conjunction with an established NGS method. This highlighted the integrated nature of any sequencing work, where the quality of each step in the process effects the next step and the eventual results and interpretation. It was therefore hypothesised that by taking control of the analysis stage of NGS work, taxonomic data could be more descriptively assessed and bolstered by functional characterisations. Here, a small sampling exercise generated a range of digesta sample types, which were used to generate comparable 16S rRNA gene analysis, DNA metagenomic, and RNA-seq NGS data sets. These data would form the starting material used to assess the ease of working with large data sets in a highly conservative computational space and the applicability of MG-RAST in generating both taxonomic and functional profiles from defined GI niches.

This work illustrated that large NGS data sets can be easily generated. However, the movement, storage, and analysis of this level of output was problematic as described by Ding *et al,* (2008). The use of the MG-RAST server was shown to be fast and simple and requires no bioinformatic training. However, the simplicity of comparative analysis allowed the lack of consensus between reference databases, and the three NGS methods to be observed when taxonomic characterisation was attempted. When functional analysis was conducted using the MG-RAST server, DNA metagenomics generated detailed output but RNA-seq data generated no confirmatory output as counts were seldom higher than zero. This could have been due to high shipping temperature, incorrect handling of samples, or an inefficient method co-extraction of RNA and DNA during the outsourced segment of NGS data production. Paradoxically, this work showed that too much data is as inhibitory to understanding, as not enough. Additionally, the excessive cost of DNA metagenomic and RNA-seq processing, and the crippling time taken for data movement precluded this method to be used in routine functional characterisation of GI communities (Appendix 9.2).

Chapter 5: Development of hybrid 16S gene analysis method

5.1 Introduction

In the last chapter a common starting material was processed through 16S rRNA gene analysis, DNA metagenomic and RNA-seq transcriptomic workflows and the resulting data were analysed and compared using a single bioinformatic suite of tools. This was attempted to firstly, gauge whether this approach could improve diagnostic granularity and secondly, to bring into control an essential aspect of the NGS workflow to enable rapid routine analysis. Perhaps naïvely, it was also considered that although technically, highly divergent, each umbrella-NGS method would complement or confirm the others using a single alignment toolbox. This significant expansion of scope was made by using the openly available MG-RAST server. Although it was shown to be an agile and diverse tool kit, significant difficulty in uploading raw sequencing data packets, from multiple computational infrastructures was a hinderance to further use. MG-RAST's multiple taxonomic databases also failed to find any taxonomic consensus, leaving the outcome ambiguous rather than merely obscured as in the fully outsourced 16S rRNA gene analysis experiments. Processing DNA metagenomic and RNA-seq data through its functional databases worked extremely quickly and generated abundant data. However, the failure of any samples other than those obtained from above the pyloric sphincter, to generate any transcriptomic data not only removed an essential quality check on the validity of DNA metagenomic sequencing but indicated that Genewiz had employed a possibly inappropriate RNA/DNA co-extraction method. Ultimately, DNA metagenomic and RNA-seq are entirely different entities which warrant highly divergent sampling and processing methods to produce results of consistent quality. Their application is best suited to answering highly specific questions which warrants the investment rather than the desire to routinely characterise entire microbial populations.

Although seemingly a failure, this work indicated that taking control of outsourced processes improved the quality of results. Therefore, employing an improved 16S rRNA gene analysis approach to population characterisation may offer clear results which could then be extrapolated to gene content via data mining. This reduction in bioinformatic scope may also work better if it included a new, less burdensome method of bioinformatic analysis. To this end, a novel pipeline was developed, leaning on the experiences in Chapter 3 and 4, to produce a health monitoring methodology that could be routinely embedded within a drug discovery study workflow, allowing the visualisation of changes in prokaryotic populations during studies which may establish the role of microbiome and how this may influence the outcome of some studies or how it is affected by study design.

5.2 Results

5.2.1 Development of alternative analysis workflow

The disparity in both the taxa identified, and the relative ability of each database employed on the MG-RAST server to process each type of data generated in the previous experiments indicated that an alternate method of assembling raw data files should be employed. This idea was cemented by the convoluted processes by which raw data was received and managed to the point of analysis. It was evident from previous experiments that such a unifying method should be robust enough to work inside and outside a secure network, be computationally discreet and flexible in the data it could process and specific in the data it generated. The DNAstar Lasergene (Madison, USA) suite of bioinformatic tools was already widely used within the GSK network for PCR primer development and sequence alignments. Although not widely used at GSK, a stand-alone academic licence was acquired suite for high through-put next generation sequence analysis.

The Lasergene Genomics suite of tools firstly allows the assembly of raw sequence file reads (.fasta) in Lasergene ‘Ngen,’ forming contigs generated on uploaded scaffold databases such as Greengenes or SILVA. These databases are free to download from each project home page (e.g., www.arb-silva.de). Although specific to prokaryotes, the Greengenes database was considered too out of date to use for the initial assembly as it was last updated in 2013. Therefore, the first 16S rRNA assembly carried out employed the SILVA small subunit (SSU) database (Quast *et al,* 2013) which was downloaded to the laptop and used as a reference (www.arb-silva.de/no_cache/download/ archive/current/ARB_files /). This assembly was conducted using the 16S rRNA gene data generated in the previous study. The assembly took more than a week to process, and generated data littered with improbable eukaryotic members including *Plasmodium falciparum*. The computational size and taxonomic inclusivity of each of the established databases (e.g., Greengenes or SILVA) was not apparent when using the MG-RAST remote server. However, when running assemblies on a standalone laptop (however fast and powerful) bioinformatics at this scale became computationally slow and diagnostically inaccurate.

Due to these immediate issues with resulting data, quality checking assemblies using Lasergene ‘Seq-Man-Pro’ became an essential activity in assessing the quality of processed sequence data. ‘Seq-Man-Pro’ displays each alignment alongside its NCBI accession number so these can be matched checked by eye rapidly in the NCBI’s Nucleotide, Taxonomy and BLAST resources (https://www.ncbi.nlm.nih.gov). Equally, gaps, base substitutions or run insertions can be examined closely across assemblies using the same tools. While checking the quality of these alignments using NCBI BLAST and Taxonomy databases it became apparent that bespoke catalogues of specified categories of microorganism can be downloaded from the NCBI Nucleotide website (www.ncbi.nlm.gov/nuccore) as small and defined .fasta files. By selecting ‘bacteria’ and ‘archaea,’ ‘RefSeq’ and ‘rRNA,’ a .fasta file containing 21,762 complete RefSeq 16S rRNA genes, defined to the strain level could be generated. This bespoke reference database was exquisitely matched to all sequence data file types. The .fasta file generated by the above method was stored on the assembly laptop as a primary reference database for Lasergene Genomics software just as the SILVA database was. This reference was used again in conjunction with the 16S rRNA gene data generated in Chapter 4. The first assembly of 16S rRNA gene sequence data using this method took seven minutes to run to conclusion (compared to seven days with Greengenes). The assembly generated a .astr file which was viewed in grid format (counts and RefSeq accession number) in Lasergene 'Array Star.' This data was devoid of full taxonomy and minimally assigned NCBI accession number.

To create a full taxonomic description necessary to understand the biology behind the limited NCBI accession number output, an annot.txt file was created using Python script editor (https://www.python.org). This activity created a .txt file used to assign taxonomy in a subsequent annotation step in Array-Star by matching the accession numbers to display their full identity. This generated tables which show NCBI accession number, taxonomic nomenclature, and molecule type which were exported into Excel or Array-Studio for primary numerical and secondary statistical analysis. Although MG-RAST was unable to process amplicon-based sequence data (i.e., 16S rRNA gene) using the RefSeq database, this hybrid method allowed the use of the most comprehensive curated microbial gene catalogue presently available. Additionally, as the RefSeq database is updated with every new submission, newer versions of the bespoke database annotation file could be created by re-processing a new .fasta file through the Python script.

5.2.2 Array-Star variant assessment

Although proprietary, Array-Star software allows the end-user to alter many of the default settings. Most impactful changes are found by toggling percentage identity score, trim settings, and k-mer length prior to each assembly run. It was hoped that a standard approach to setting each parameter could be found so that data was comparable across samples and studies.

A widely accepted tenet is the setting a percentage identity score threshold for clustering at 97%, which is used to represent a defined cut-off for species similarity (Edgar, 2017). It is intended that by pragmatically setting this threshold, OTUs could be safely considered to represent defined biological entities. However, the accuracy of this assumption is a balance between the inclusion of error-driven calls and a strict cut off point for that inclusion. By lowering the percentage identity, more variants are considered as specific entities and conversely, by raising the percentage identity score sees a fall in counts but a rise in accuracy. This supposed improvement in quality ignores the differential degree of mutation in each V-region within many lineages of microorganism making true differentiation often impossible using V-region characterisation methods in isolation (Mysara *et al*, 2017). Single Illumina sequencing reads are most often <250 bases long, while paired end reads represent longer, linked spans but contain four strand extremities. Illumina base calling is conducted internally using CASAVA software, which assigns a Phred score to each call. These quality (Q) scores relate to the probability of the base being incorrect using the equation p=10$^{(-Q/10)}$. Strand extremities tend to have reduced Q scores. Inclusion of these incorrectly called bases will bias data and invalidate taxonomic assignments. It is important then to remove or trim these aberrant bases from assemblies prior to taxonomic assignment of binned sequences.

Trimming may be achieved by either correcting or eliminating incorrect base calls. Correcting depends on readjusting calls according to frequency of insertion, whereas elimination removes effected stretches of bases. Each has its drawbacks, both computationally and interpretationally. Trimming has therefore been widely adopted in metagenomic studies (Fabbo *et al*, 2013). Here, the effect of auto-trimming was made apparent by the highest inclusion of a bacteria found in shellfish (NR_043177.1, *Spiroplasma penaei*, strain SHRIMP). This was due to auto-trimming too strongly, generating a small contig which matched a highly conserved area of the 16S rRNA gene, effectively decreasing specificity. Once auto-trim was removed, results did not alter during subsequent changes to percentage identity and k-mer size in high-ranking diversity and reflected the biology of the niche screened.

Altering percentage identity scores from 75-100%, saw the resulting OTUs fall from 617 to 148. Changing the score again, did not affect the high-ranking entities in each niche, indicating high quality data. However, it is probable that by reducing OTUs to 148, many species with low V-region (specific) variation will be binned together, still creating a false diversity. Therefore, a high (97%)

score was employed which gave good quality output but still allowed some variation during binning.

Both increasing and decreasing the k-mer size (default 17) from 10 to 31 saw a slight reduction in resulting OTU numbers (277 to 220 and 211 respectively). Again, the highest-ranking entities were not affected and as both variations to the default settings generated decreases in numbers it was thought that the default setting was best left alone (data not shown). If a conservative approach to OTU binning is taken 16S rRNA gene based studies would benefit from increasing percentage scores to 97%, with no auto-trimming and a k-mer size of 17. However, it would be best practice to always quality check a range of high and low scoring identities for accuracy post-assembly.

5.2.3 Bespoke database workflow results

Simple analysis of 16S rRNA gene analysis in Excel indicated a site specific microbiota, most easily seen between the stomach samples (1 & 8) compared to those of the lower GI tract. These samples were dominated by the genera *Moraxella, Brevundimonas*, *Haemophilus,* and *Pseudomonas*. Other common species unique to this niche were *Staphylococcus capitis, S. aureus, S. caprae, S. epidermidis, and S. simiae.* While the lower GI tract was devoid of these species, *S. scuiri* was found to be the only staphylococci present. Other Gram-positive genera belonging to the *Actinomycetota* phyla: *Corynebacterium, Cutibacterium*, *Micrococcus* and *Kocuria* were only found in these samples, not in the lower GI tract. Also, *Clostridium scindens* was the only *Clostridia* found in this niche, while it and all others were found through the entire lower GI tract. *Enterorhabdus* spp., *Escherichia* spp., *Kineothrix* spp. and *Bacteroides* spp. were not found in this niche, but in all others. The upper GI tract was dominated by waterborne genera (*Brevundimonas* and *Pseudomonas*), the *Moraxella/Haemophilus* group and multiple Gram positive genera (*Corynebacterium, Staphylococcus* or *Micrococcus*) while being devoid of *Enterobacteriaceae* and most obligate anaerobes responsible for the fermentative breakdown of complex carbohydrates in the lower GI tract (*Clostridium* and *Bacteroides*). This suggested that a niche specific microbiota (down to the strain level) was discernible using this newly defined method of characterisation. It is also shown that simple comparative analysis of this dataset can be achieved using the conditional formatting, hide/unhide and sort/filter functions in Excel. However, to uncover any less obvious details and to use a less subjective method of analysis, this full data set was statistically processed with Array Studio software.

5.2.4 Alpha diversity analysis of 16S rRNA gene data

The new method of analysing GI diversity replicated variance seen in the primary niche experiments (3.2.4) generating 137 defined OTUs. Shannon index analysis (Figure 5-1A) showed a fall in richness in the small intestine from the stomach which was seen in Fig. 3-4 but also showed the rise in richness from the caecum to the colon, a site omitted in the original sampling opportunity. This further validated the use of faeces as a representation of colonic diversity and possible metabolic activity in mice. Box and whisker analysis of the same data again indicated the rise in diversity down the GI tract with the stomach showing the tightest spread of data (Figure 5-1B).

Once imported into Array Studio, this data was standardised by transforming to a +1-log scale to allow widely varying counts to be compared equally. Principle component analysis was conducted (Figure 5-2). This showed the wide difference between niches above and below the pyloric sphincter but also the clustering of all samples from below including the caecum this shows that there is a dramatic shift in microbial diversity between the stomach (dark blue) and that of the niches below the pyloric sphincter (jejunum, ileum, caecum, and colon), which was not evident in previous analysis (Figure 5-5).

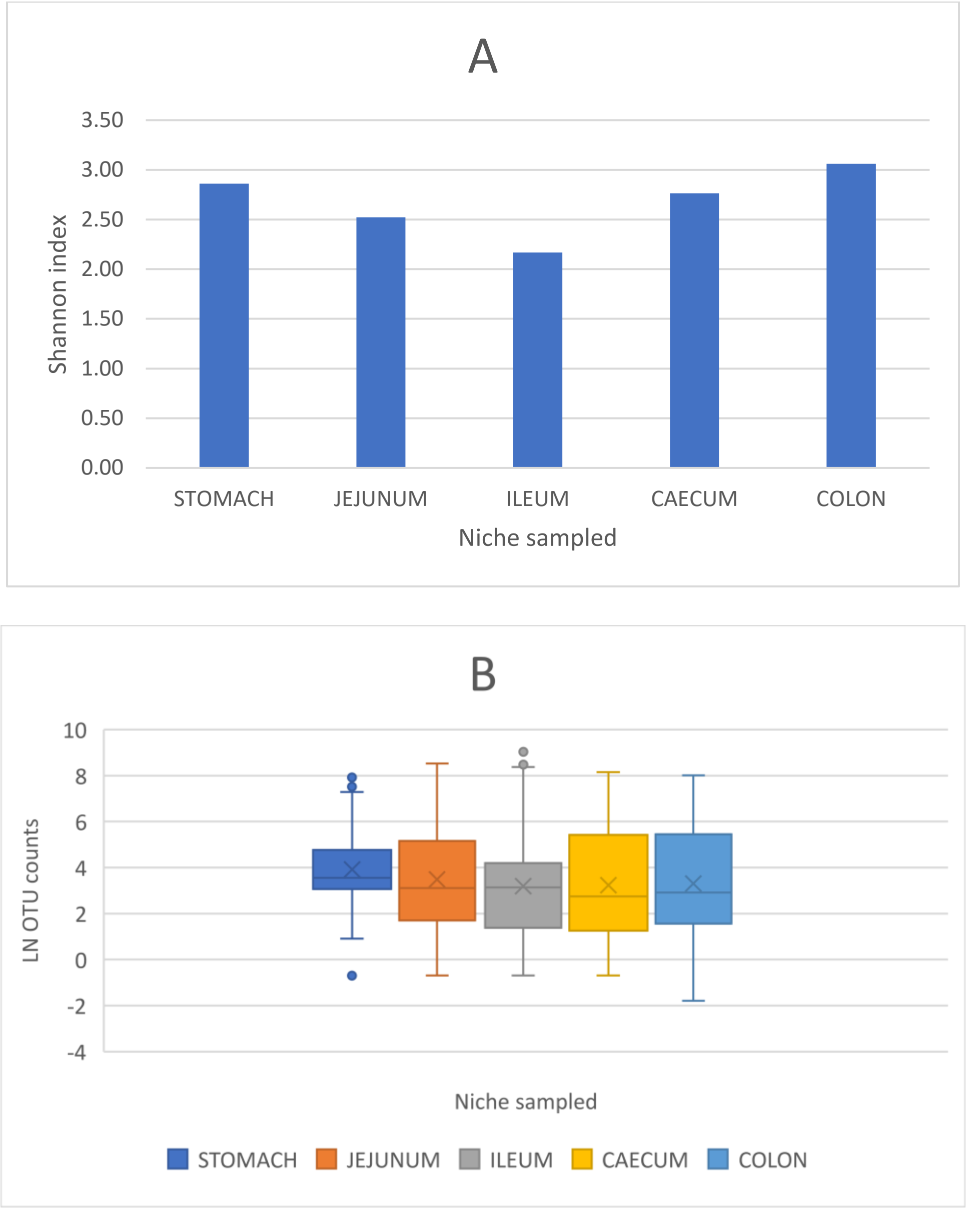


Figure 5-1: Alpha diversity analysis of C57(Trim) GI niches identified using the novel method of analysis. The Shannon index (A) indicates the broad change in bacterial diversity between the stomach and those niches of the upper GI tract and the increase in the sites of the lower GI tract while the box and whisker plot indicates a broadening of species down the GI tract.

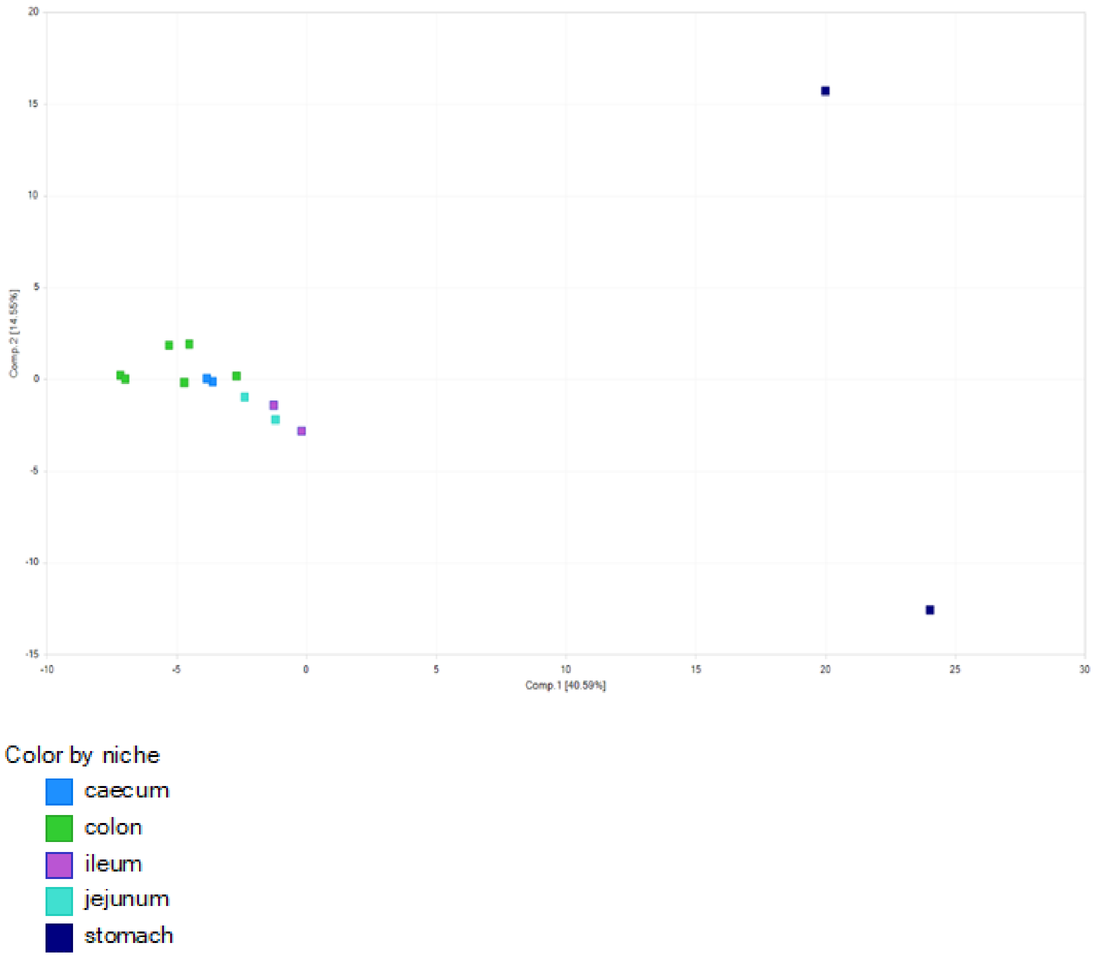


Figure 5-2: Principal component analysis of C57(Trim) GI niches identified using the novel method of analysis. This indicates the broad difference in species diversity between niches above and below the pyloric sphincter of mice. PCA generated using Array Studio.

5.2.5 DNA extraction and CRO NGS comparison

In the first outsourced studies, Genewiz used the DNeasy Power Soil kit (Qiagen-A, 2021) for the extraction of DNA from faecal disgesta for 16S rRNA analysis. In subsequent work, Genewiz employed the AllPrep Powerviral DNA/RNA kit (Qiagen-B, 2021) for the co-extraction of nucleic acids for 16S rRNA, metagenomic and RNA-seq workflows. The former kit is based upon non-biased physical (bead) material disruption, whereas the latter method uses physical and chemical (β-mercaptoethanol) disruption steps. Each kit uses different proprietary buffers and wash solutions (Qiagen A & B, 2021). This essentially generated vastly different starting material for each 16S rRNA study, which unfortunately created incomparable diversity data. The concurrent extraction of both DNA and RNA is an ineffective approach to microbiome analysis where good data is dependent on good quality nucleic acid. Inefficiencies at a critical stage in a costly process involving animals possibly terminally sampled, should not be tolerated. Therefore, a comparison was made between the existing method for PCR grade analysis DNA using the QIAmp DNeasy kit and the DNeasy Powersoil Pro kit using both suggested physical disruption devices (Genie centrifuge adapter and the TissueLyser II). The average concentration of extracted DNA using QIAmp was 5.3µg/µl, using Powersoil and Genie centrifuge was 144.8 µg/µl and using Powersoil and the TissueLyser was 263.6µg/µl. Therefore, a 4873% increase in non-biased DNA yield was attained by using the Powersoil kit in conjunction with the TissueLyser II (Table 5-1) although 260/280 ratio purity testing was not carried out.

Table 5-1: Comparison of DNA extraction methods and resulting DNA concentration (µg/µl) obtained from C57BL/6 mouse faeces. the Powersoil with Genie and TissueLyser II disruption techniques (as recommend) compared to that extracted using the enzymic QIAmp DNeasy method.

Method	Pellet (number)	Faeces (mg)	DNA conc. (ug/ul)
Genie	3	40	67
Genie	10	125	161
Genie	10	125	169
Genie	20	250	199
Genie	20	250	128
TL II	3	40	61
TL II	10	125	296
TL II	10	125	276
TL II	20	250	308
TL II	20	250	377
QIAmp	10	125	6
QIAmp	10	125	4
QIAmp	10	125	6

Samples sent for third party analysis were processed fastest by CRL (4-weeks) and slowest by Qiagen (3-months). Eurofins generated data by around 8 weeks. Genewiz left the samples for extraction and sequence analysis in a warehouse near Stanstead Airport for two weeks before processing. When samples eventually arrived at their sequencing facility in Plainfield, USA no dry ice was present. They conducted the contracted work and only communicated the issue after processing.

The resulting raw sequence files were analysed using the novel bioinformatic workflow and counts for each of the 21,762 possible entities were compared and a Spearman's rank correlation coefficient (Figure 5-3) was conducted which derives a similarity score between two variables. Here the variables were the company carrying out the DNA extractions and the number of resulting OTUs for each sequencing run. Matrix plots are given below, with response plots below the diagonal and Spearman ranks above which tends towards one, when samples compare more highly and reduces from one where correlation is less strong. The diagonal shows a histogram for each sample. The responses here are plotted on a log ten scale and 0.01 was added to all entries with zero counts after sequencing. Entities which scored zero counts across all samples were excluded prior to this analysis to remove a high number of data points. This left 6050 entities which are compared here. It was evident from this analysis that the in-house use of Powersoil kit generated higher scores and therefore better correlation than those samples extracted and run at Eurofins and Qiagen. Conversely, higher correlation was found between samples extracted and run at CRL although a single low ranking in-house sample being found was unfortunate. This work indicated that CRL offered the fastest turnaround time and although Qiagen generated higher correlation between samples, their turnaround time was the longest and they are not a GSK preferred supplier. Therefore, CRL were chosen for future outsourced sequencing work and in-house DNA extraction was the most reproducible method of submitting material for analysis.

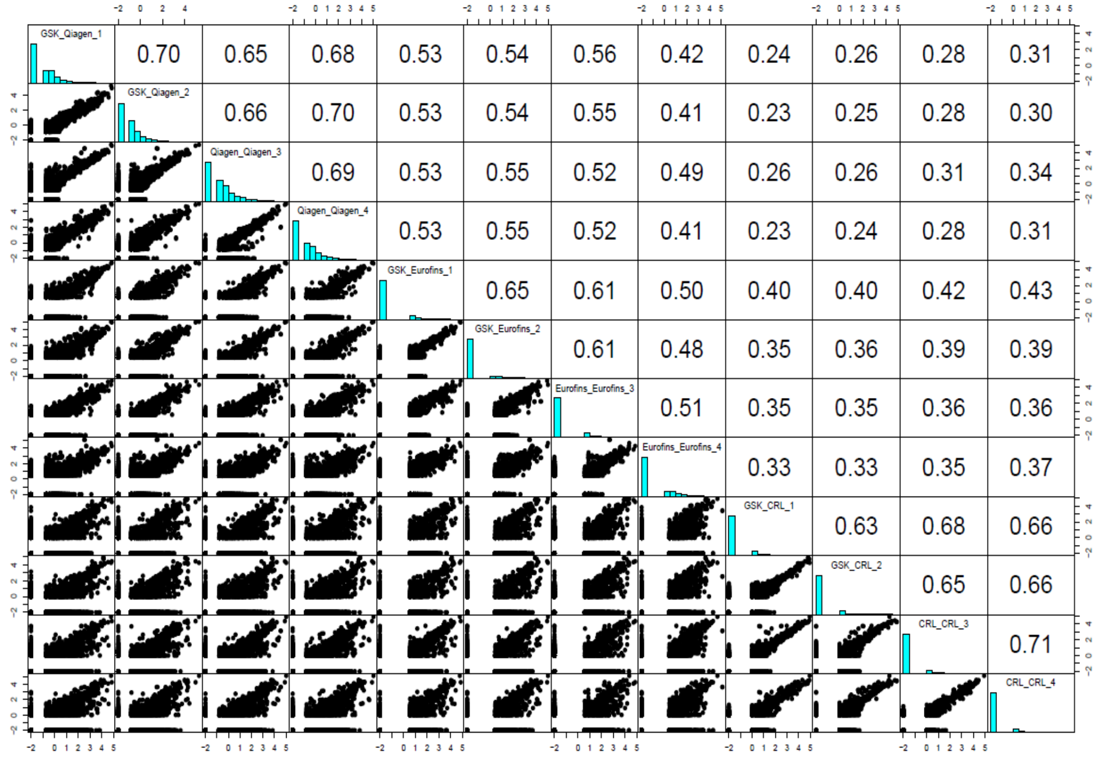


Figure 5-3: Comparison of external NGS provider quality using a Spearman rank plot. by Eurofins, Qiagen and CRL sequencing labs showing the relationships between DNA extraction and OTUs at each indicating highest ranking when both activities took place at CRL.

5.2.6 Development of SOP for 16S rRNA gene analysis

The driving force behind these experiments was to develop a standard operating procedure (SOP) to deploy microbiome analysis as standard methodology for animal studies. This SOP could then be employed to generate samples of comparable quality going forward into NGS processing, so that resulting datasets could be compared and conclusions drawn from these efforts. These combined bioinfomatic experiments were designed to firstly illuminte the biology behind each question asked, but secondly to generate a knowledge base form which considered changes could be made in the future as technology and processes develop and evolve. The flow diagram (Figure 5-4) illustrates the interlinked processes used to generate full taxonomic descriptions drawn from raw 16S rRNA gene sequence files.

During the development of this process several technical points became apparent. The complete dataset of all 21,762 entries was copied from Array Star into Excel. This ensured that all scores were included in subsequent analysis. The mean average was calculated across all samples for each taxa, and then the taxa present in numbers below that of the number of samples taken, was hidden from the analysis pipeline. This ensured that any graphs generated are not littered with low scores for organisms in the reference dataset making interpretation easier.

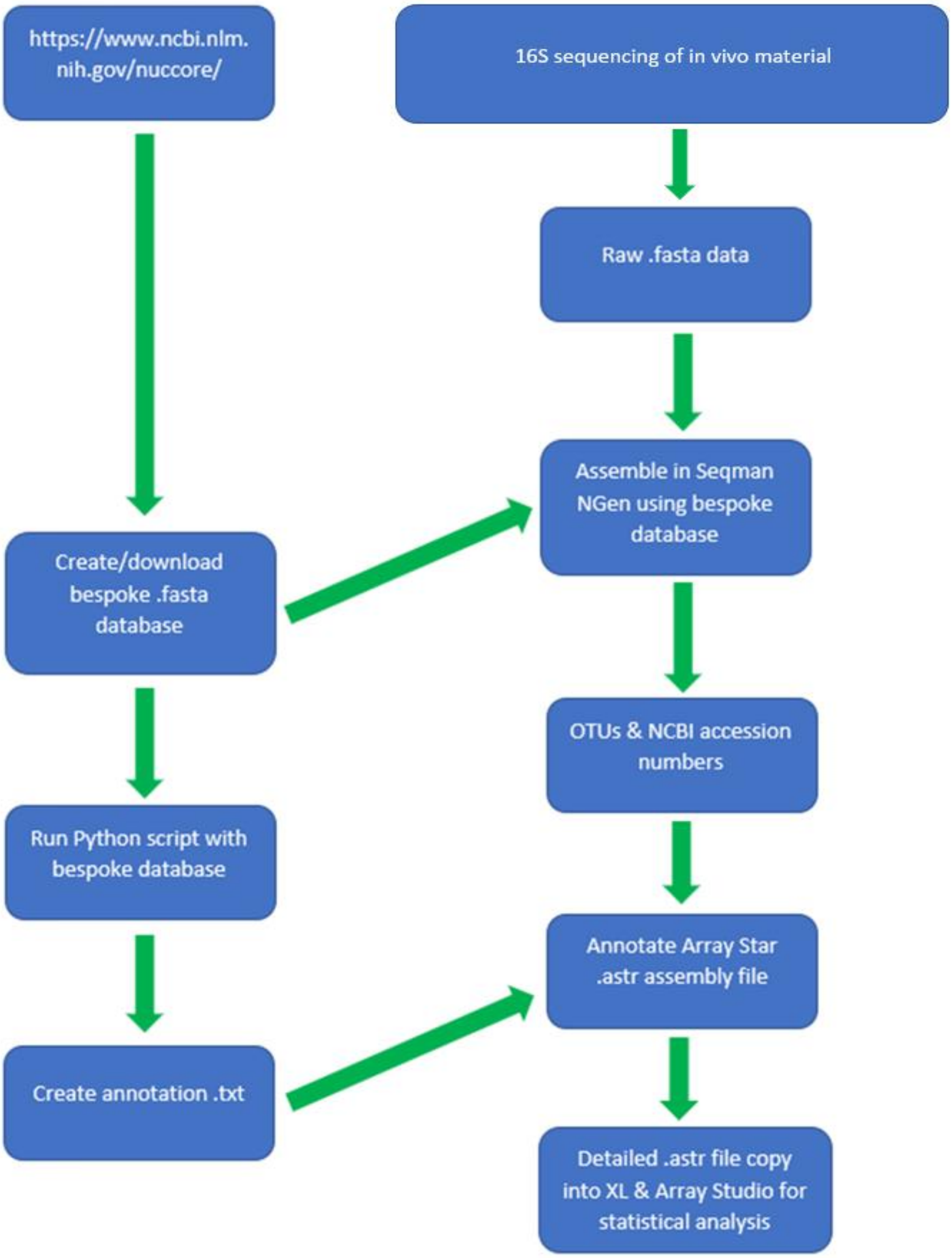

https://www.ncbi.nlm.nih.gov/nuccore/
16S sequencing of in vivo material
Raw .fasta data
Assemble in Seqman NGen using bespoke database
Create/download bespoke .fasta database
OTUs & NCBI accession numbers
Run Python script with bespoke database
Annotate Array Star .astr assembly file
Create annotation .txt
Detailed .astr file copy into XL & Array Studio for statistical analysis

Figure 5-4: A flow diagram indicating the novel 16S rRNA gene analysis workflow. Original .fasta format in conjunction the Lasergene suite and the use of Python script to create an annotation file thus generating detailed taxonomic characterisations from all classes of sequence data.

The resulting stacked graphs generated in Excel were conditionally formatted (high to low) so that the comparative biology of each sample became easier to understand highlighting the keystone species. This allowed diversity to be more readily appreciated. When the OTUs were arranged alphabetically it became apparent which species and therefore which phyla and genera are present in a sample. Grouping bacterial species like this allowed both high level and extremely granular data to be represented.

The issue of generic granularity due to 16S rRNA gene similarities still exists for the *Enterobacteriaceae*. However, although this method was unable to identify easily cultured bacteria, it was able to identify >500 other unculturable bacteria to the strain level. The drawback of culture-based techniques is that they only identify a limited number of species. By reassessing our view of what is possible and what is not by understanding the shortcomings of each technique indicates a depth of knowledge rather than a failure of the method employed. Once OTU tables are generated in this way, individual phyla, genera or families can be lifted from the mass and inspected across all samples in a study in isolation. Samples can be hidden in Excel allowing data from litter mates or cage mates be compared at the species/strain level.

5.3 Summary

The work described above represents a solution to the previous findings that fully and partly outsourced 16S rRNA gene analysis, DNA metagenomic, and RNA-seq methods of taxonomic characterisation were equally ill-suited to the routine production of accurate diagnostic data from the GI tract of the mouse.

This work involved the generation of a novel bioinformatic tool by pairing an openly available reference database and a proprietary software package to produce a functionally descriptive output. This key step forward in diagnostic granularity was augmented by an assessment of nucleic extraction techniques and sequencing partners. The reference database used here was a .fasta copy of the RefSeq, fully curated list of full 16S rRNA gene sequences. As such, it provided a prokaryotic-focused reference, aiding rapid and accurate alignments as an annotation using the Lasergene Array-Star suite of tools. This pairing was made possible by creating a bespoke computational device capable of running alignments as a network-free, standalone tool.

By taking experimental control of the input of nucleic acids and by applying a novel bioinformatic tool to the analysis phase, the previously observed lack of diagnostic clarity was removed. Additionally, by leaving only the sequencing of high-quality DNA to an expert provider of NGS, a repeatable and cost effective workflow was established and a formal SOP created (Appendix 9.1). This will allow detailed characterisation studies to take place in a comparable manner. This sets a standard for the application of this technique in drug development studies. By applying this analytical technique across multiple murine studies, an awareness of the models used and how their microbiota is modified by the administration of therapeutic agents or disease simulations will increase. It is hoped that from this increasing body of data, study-specific markers will become apparent which will aid understanding of the underlying biology and contribute to more informed decisions and study designs in the future which ultimately augment drug discovery.

Chapter 6: Application of hybrid 16S gene analysis method

6.1 Introduction

In the previous chapter both the bioinformatic pipeline process and the extraction of nucleic acids were removed from the initial, fully outsourced method of NGS characterisation. Firstly, a hybrid bioinformatic process was developed which worked within a conservative computational space, generated repeatable results of species-level granularity, and which could be readily updated. Secondly, a superior method of nucleic acid extraction was evaluated and introduced as a standardising feature of all pre-sequencing work. These advances, added to the sampling method previously described meant that only the sequencing runs were to be outsourced by a CRO. The comparative testing conducted between four CROs indicated that CRL supplied a consistently robust submission, delivery, sequencing, data sharing process and was a preferred supplier. Therefore, CRL was chosen as the external provider of NGS for subsequent studies. These advances enabled the development of an SOP which laid out every step of a routine health monitoring NGS process, enabling reproducibility in a divergent field. With this process in place, the diagnostic capabilities of the hybrid 16S rRNA gene analysis-based approach could now be focused upon relevant disease areas to help gain understanding of the impact of disease generation to members of the GI microbiota and to study the feasibility of deploying microbiome analysis routinely during *in vivo* studies. This real-world application would allow the detailed quantification of species-level diagnostics based on the analysis of ethical material throughout studies, allowing the visualisation of the fate of species generated by different classes of disease instigation and amelioration. This work may uncover both the effect of administration upon the microbiota and the microbiota upon the host and disease progression, a link which is seldom found (Fischbach, 2018).

6.1.1 CD4+ cellular induction model

A study was designed to evaluate the effect of two refinements to the husbandry of this IBD model but carry out the study as a direct mirror of a real investigation, employing normal study procedures (i.e., vehicle dosing, weighing, and handling). It was hoped that adverse effects (i.e., body weight loss and clinical signs) could be reduced without affecting the development of colitis while maintaining the robustness of the endpoints (i.e., colon density), body weight and clinical scores, measured between PBS controls and CD4+ recipients.

The success of this model of IBD is also thought to be dependent on the bacterial load of the GI tract (Guarner, 2003). As immunocompromised mice have an innately and environmentally altered prokaryotic diversity (Zheng *et al,* 2019), providing food and water containing low levels of these pathogens is thought to assist in the development of disease. By keeping PBS and CD4+ groups of C.B-17/IcrHsd-PrkcdSCID mice in non-sterile and sterile environments the strength of outcome could be observed according to the possible increase in bacterial burden in the non-sterile group. Secondly, the animals were to be given diet gel and soaked food from Day 18 as animals used in this model typically show signs of dehydration and weight loss due to the thickening of the colon as disease progresses towards the end of the study. This leaves the animal less able to absorb water resulting in loose faeces which is expected at around Day 30. Providing additional liquid in the diet could allow greater absorption, reducing visible signs of dehydration and weight loss, however, it may also increase levels of diarrhoea due to loss of normal functionality. By employing the 16S gene analysis to this study, the bacterial numbers in both sterile and non-sterile groups in a CD4+ study would be measured for the first time in addition to species level identifications through the study, unlike the reliance on subjective scores and terminal examination of the colon architecture.

6.1.2 DSS chemical induction model

In this model the severity of colitis is influenced by several factors, including the DSS concentration (typically 1-10%), the susceptibility of the mouse strain used, and the duration and switching of DSS administration. DSS-induced colitis can be either acute or chronic; acute colitis is achieved through administration of single cycle of DSS, whereas chronic colitis is achieved through the application of repeated cycles of DSS, interspersed with periods of normal drinking water administration. Whilst similar pathological changes are seen in both approaches, the chronic model of DSS induced colitis replicates some pathological changes seen in human disease that are not seen in the acute model. In addition, the withdrawal of DSS can lead to disease resolution, allowing investigation of the mechanisms controlling epithelial repair and resolution of inflammation (Munyaka *et al*, 2016). Signs of disease can appear as early as one day after DSS administration, with changes in the expression of epithelial cell tight junction proteins, which is thought to contribute to impaired barrier function. Altered expression of tight junction proteins have also been observed in human IBD. The cytokine profile observed after DSS administration has indicated that T_h1 and T_h2 pathways are both implicated in this model (Munyaka *et al*, 2016). The aim of this study was to observe the optimal concentration of DSS in conjunction with wild type C57BL/6 mice under present husbandry and facility confines. Again, it was hoped that measuring fluctuations in prokaryote species and number throughout a DSS study for the first time, a link between dose-related pathology and community structure could be made. Additionally, it was hoped that any physiological repair of the GI tract after administration could be linked to changes in the prokaryotic community structure.

6.2 Results

6.2.1 Description of inflammation in the CD4+ induction model of IBD

At the termination of the study (Day 38) eyelids and colons (as sites of inflammation) were removed and underwent histological preparation, haematoxylin and eosin staining, and scanning. Images of the colon representing all groups (sterile and non-sterile PBS and CD+) are shown (Figs. 6-1A & B and 6-2A & B). These indicate the lack of cellular infiltration in the PBS groups and the presence of proliferating lymphocytes in the mucosal surfaces of the colon indicating the successful development of the disease model in C.B-17/IcrHsd-PrkcdSCID mice. Although, moderate structural changes to the topography of the colon were formed by the number of mononuclear infiltrates causing hyperplasia and apoptosis in the CD4+ recipients, none was observed in the PBS groups. Neither was a perceptible difference observed between sterile and non-sterile groups. In addition to colonic images, histological examination of eyelids indicted the presence of high numbers of CD4+ infiltrates. This data provides cellular confirmation of the clinical signs (squinting) observed (Figure 6-3). This graph also indicates the difference in clinical signs observed between the CD4+ and PBS control groups which reflects the externally observable indications of disease.

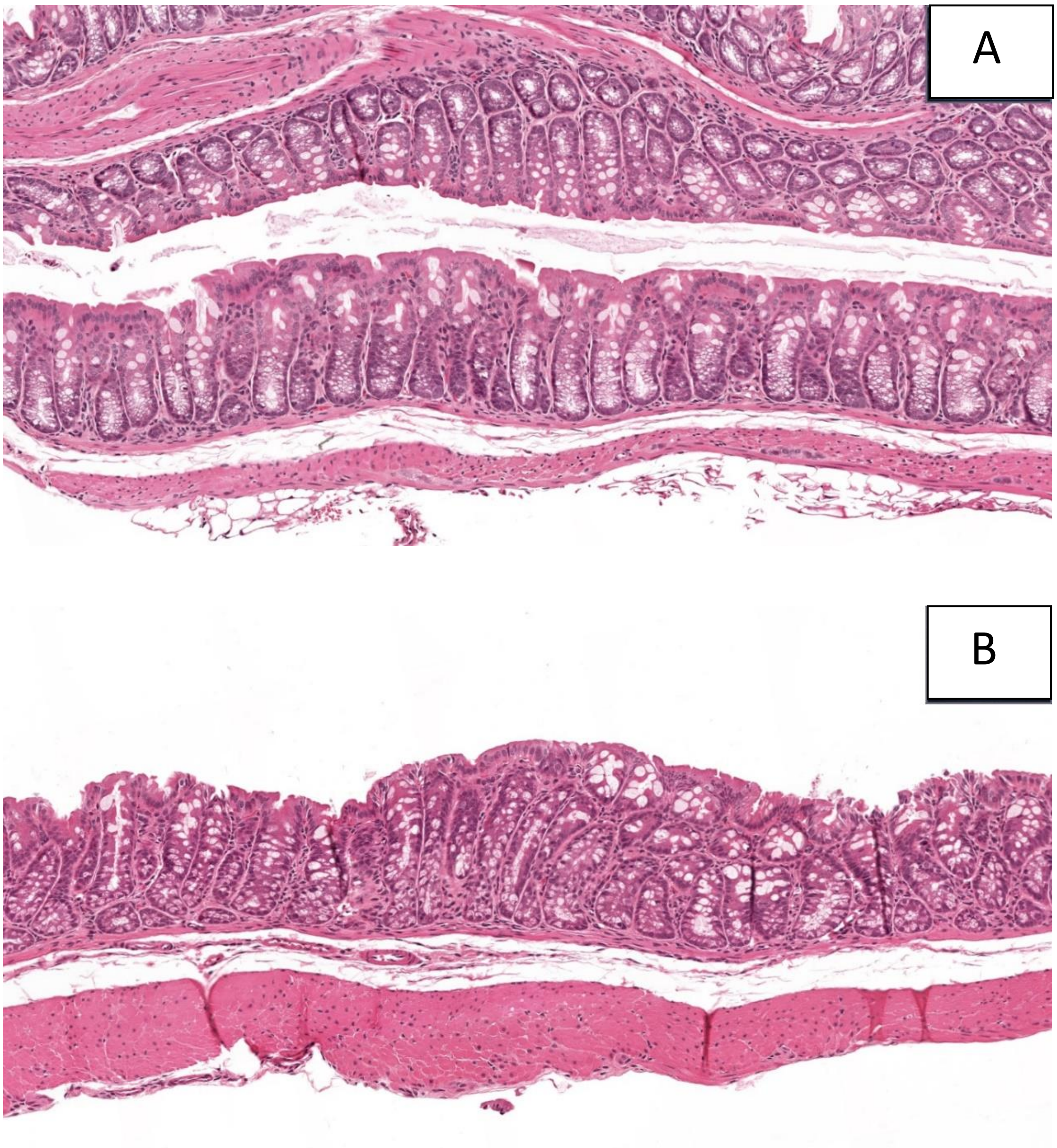


Figure 6-1: Haematoxylin and eosin-stained C.B-17/IcrHsd-PrkcdSCID colon at termination of CD4+ study. (A) representing the sterile PBS group and (B) representing the non-sterile PBS group in the CD4+ transfer model. Both images indicate the lack of lymphocyte infiltration (cells stained blue) in comparison to Fig. 6-2. Sections and images generated by the Non-clinical Histology and UK Pathology teams (GSK).

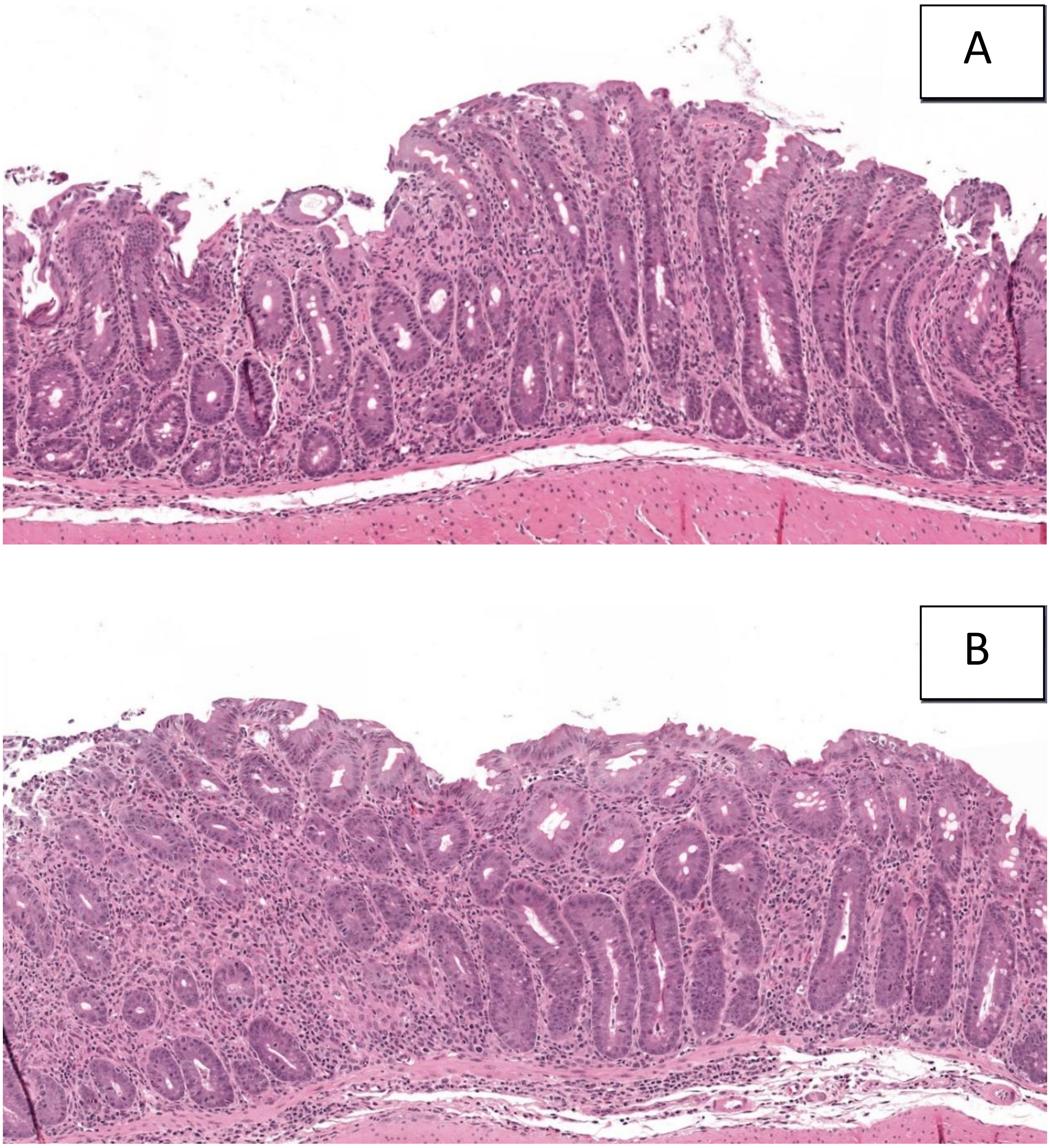


Figure 6-2: Haematoxylin and eosin-stained C.B-17/IcrHsd-PrkcdSCID colon at termination of CD4+ study. (A) representing the sterile CD4+ group and (B) representing the CD4+ non-sterile group in the CD4+ transfer model. Both images indicate the presence of large numbers of lymphocytes (cells stained blue) and successful generation of the disease model. Sections and images generated by the Non-clinical Histology and UK Pathology teams (GSK).

However, it also shows the subjective nature of visual observations as peaks and troughs seen here are common in all groups (going against disease trend). The dates with low observations (Days 20, 27 and 34) were weekends when staff unaccustomed to the study were required to record subjective events during overtime.

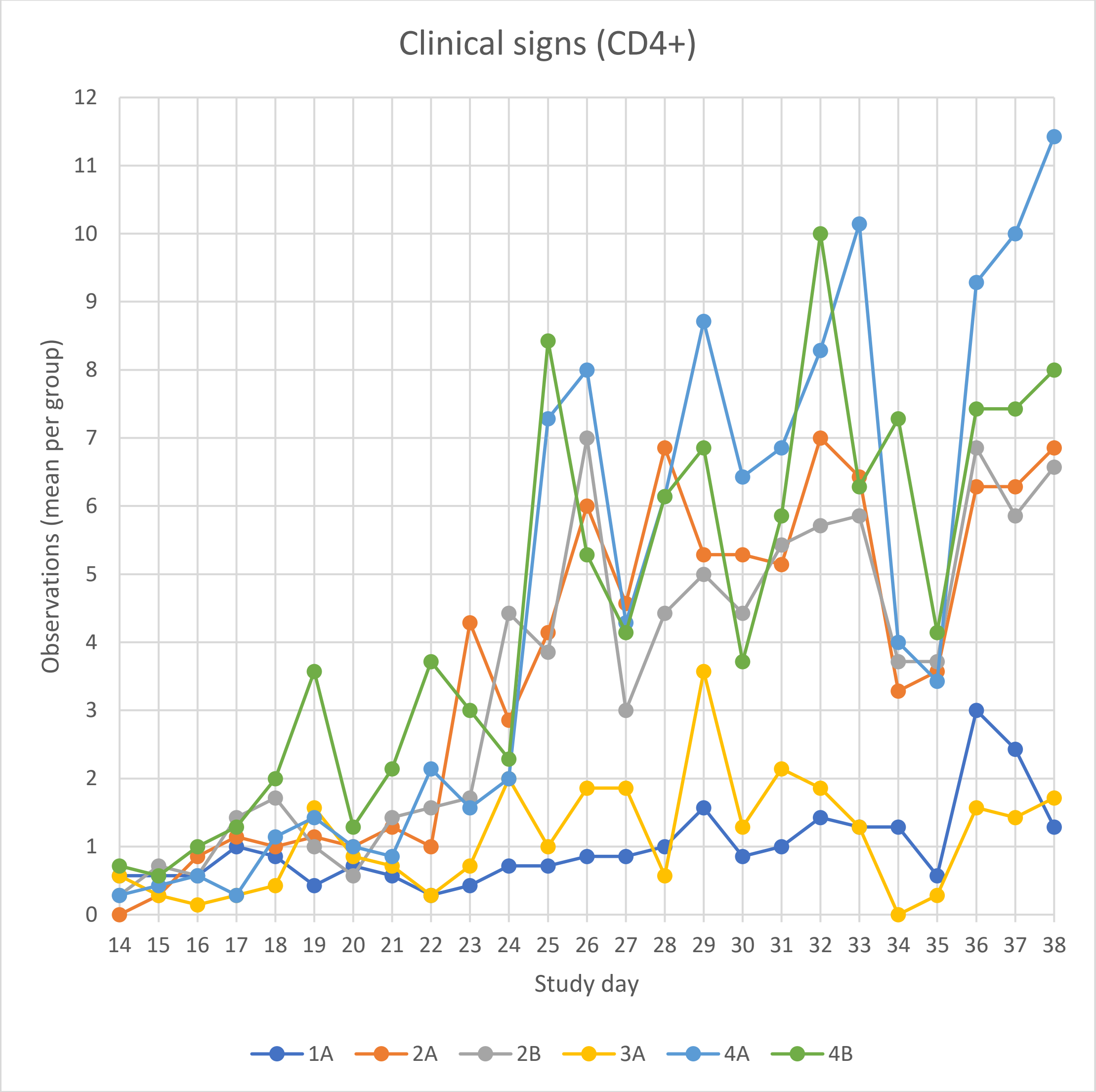


Figure 6-3: Total daily clinical observations made of C.B-17/IcrHsd-PrkcdSCID mice during daily welfare examinations throughout the CD4+ transfer study. These signs were hunching, lethagy, eye inflammation, squinting, anal swelling, sore feet, abnormal respiration, straining, reduced group interaction, and piloerection. 1A (non-sterile PBS), 2A/B (non-sterile CD4+), 3A (sterile PBS) and 4A/B (sterile CD4+).

In addition to clinical signs, daily body weights were recorded for all groups in the CD4+ transfer study. The CD4+ recipients' weight decreases throughout the study as disease progression takes place following CD4+ transfer on Day 0 (Figure 6-4). Clinical signs and weigh plateau or loss both begin to be observed at Day 16 post-transfer (Figs. 6-3 and 6-4). The PBS control groups exhibit a slowing of natural weight gain towards the end of the study while the CD4+ recipient groups all suffered weight loss after Day 23. The trends for each group are consistent from beginning to the termination of the study. However, there is evidence of non-random selection at study commencement with the CD4+ recipients all weighing more than the PBS control groups in the original weight measurements.

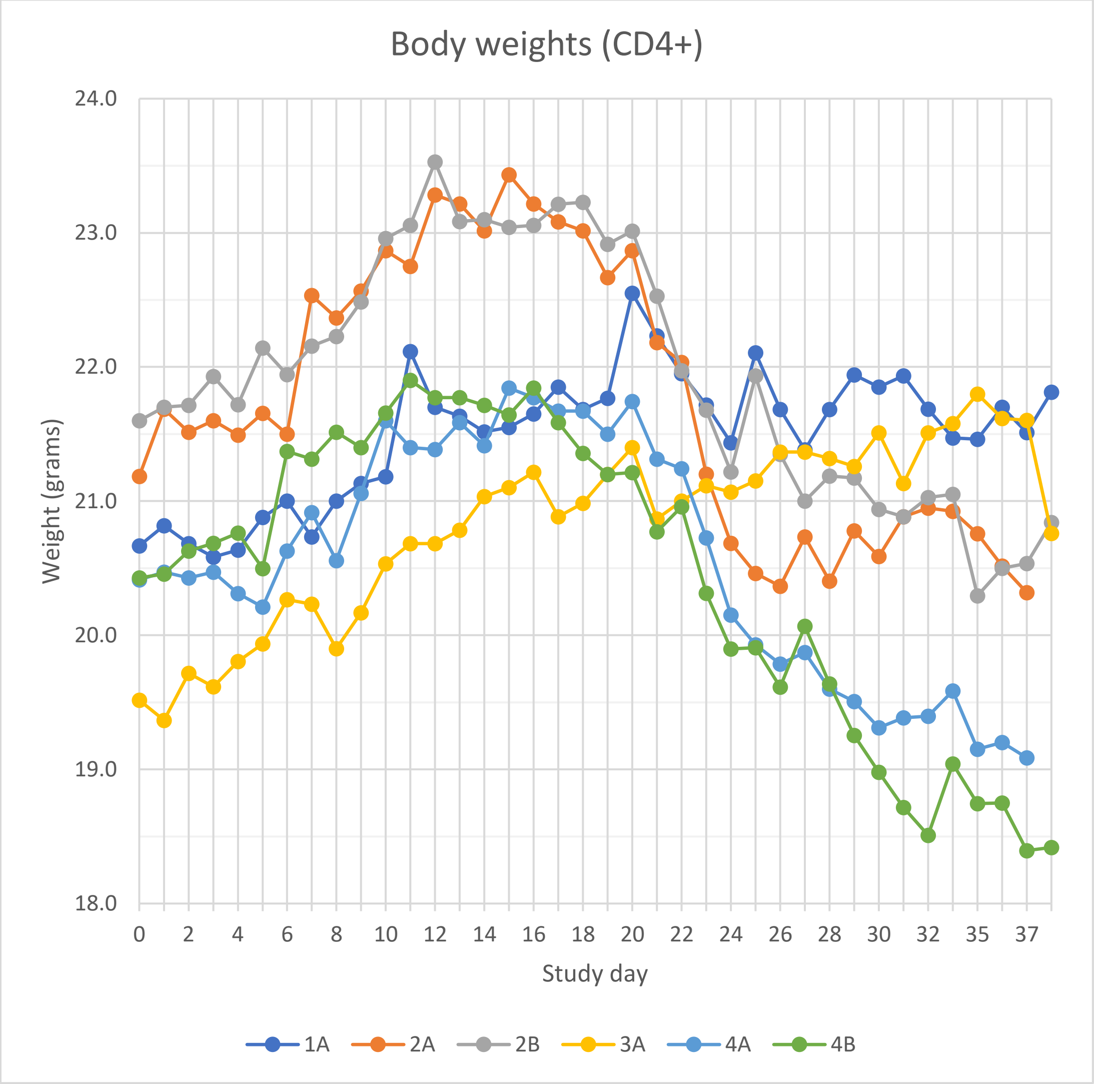


Figure 6-4: Daily body weights of C.B-17/IcrHsd-PrkcdSCID mice on each day of the CD4+ study. the adoptive CD4+ cell transfer study of IBD. Groups 1A (non-sterile) & 3A (sterile) are PBS controls and 2A, 2B (non-sterile), 4A & 4B (sterile) are CD4+ recipients.

6.2.2 Prokaryotic diversity in the CD4+ induction model of IBD

The cellular proliferation of naive CD4+CD45RBhigh lymphocytes when transferred into an immunocompromised mouse, is initiated by GI localisation, tissue infiltration and direct contact with commensal bacteria (Kjellev *et al*, 2006). As immunocompromised mice possess an environmentally and innately altered GI microbiota (Zhang *et al*, 2019), it was hypothesised that environmental microorganisms may augment proliferation. These could originate from the water, diet, or cage environment and so a study was designed to evaluate whether sterile or non-sterile conditions affected disease progression in the absence of a research antagonist.

The diversity of all pooled faecal samples was analysed by compiling a Shannon index to normalise quantitative OTU data between the sterile and non-sterile groups and between the CD4+ recipients and the PBS control groups (Figure 6-5). These analyses show that community diversity increased during the first week post-arrival in all groups but the sterile CD4+, which has not been found in other studies. No other clear pattern of community behaviour is seen in these plots. By comparing the counts across all sampling points and including replicates at arrival and the two cage samples provided by the CD4+ recipients (per sterility group) it was possible to generate p-values and subsequently calculate the FDR using Array Studio. This statistically showed that there was no difference in community composition between CD4+ and PBS or sterile and non-sterile.

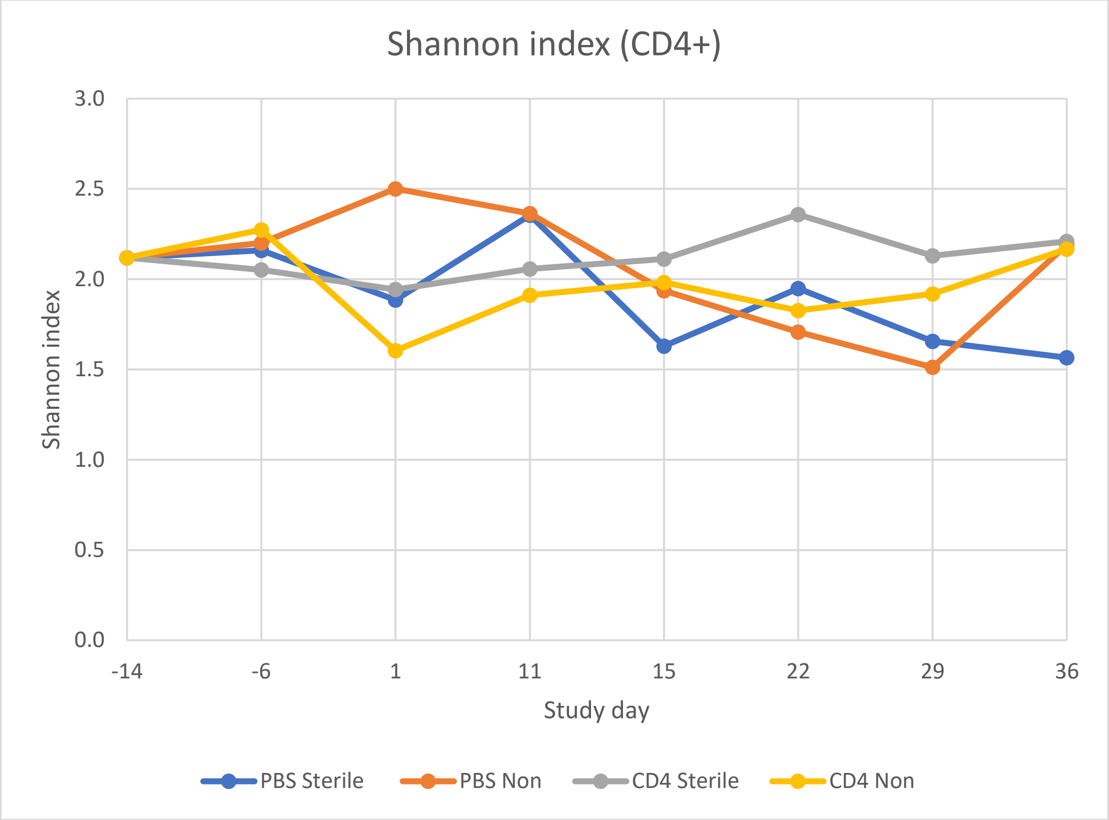


Figure 6-5: Alpha diversity analysis of C.B-17/IcrHsd-PrkcdSCID mouse faeces throughout the CD4+ study. The Shannon index of prokaryotic OTUs found in C.B-17/IcrHsd-PrkcdSCID mouse faecal output comparing diversity at delivery with the CD4+ recipients and PBS controls.

In addition to the use of Shannon indexing to describe changes in community diversity, taxonomic characterisation of each C.B-17/IcrHsd-PrkcdSCID mouse faecal sample was undertaken (Figure 6-6). Using this method of illustrating changes in diversity allowed the detailed understanding of how successful disease progression in the CD4+ model effected species dynamics. Of the original 322 OTUs named using the hybrid bioinformatic method described here, only thirty-six had a mean average of >48 in this analysis. The p-values and FDRs were calculated using the OTU output and showed no significant changes between the sterile, non-sterile, CD4+ and PBS groups.

However, when the thirty-six OTUs are illustrated using a stacked bar chart with a specific a colour code (Figure 6-6), it is possible to see that *Staphylococcus lentus* (Figure 6-7 top), *S. cohnii, S. vitulinus* and *Sporosarcina pasteurii* (Figure 6-7 bottom) all increased in numbers through the period of the study. The staphylococci are skin commensals may also inhabit the GI tract of healthy mice, while *Sporosarcina* (formally *Bacillus*) *pasteurii* is an environmental spore former which can inhabit inhabits nutritionally low, external surfaces. The rise in bacterial numbers in all CD+ and PBS groups regardless of sterility status, possibly shows environmental conditions post-delivery, (such as the use of Aspen wood chips instead of highly processed Uber-dry at source), which were more conducive to staphylococcal proliferation. Novel transmission from the new facility environment regardless of the engineering controls put in place using IVCs (e.g., enclosed change stations) could also be implicated here. The former represents a factor in the reproducibility of studies across facilities and the latter represents a short fall in the effectiveness of IVCs which if extrapolated to the transmission of viruses presents a husbandry technique unfit for purpose. Both staphylococci and *Sporosarcina pasteurii* express elevated levels of urease which breaks down urea resulting in the release of ammonia which is found in long-term studies where litter is not changed completely or regularly (Washington & Peyton, 2016). *S. pasteurii* is also able to precipitate calcite which is often found affixed to the bases of dirty rodent caging (Chou *et al,* 2008).

This taxonomic analysis confirms the statistical results indicting that there was no difference between sterile and non-sterile groups. It does however reveal the community structure at source and how it changes upon delivery. The transfer of CD4+ cells or the onset of weight loss and other clinical manifestations clearly noted during daily observations (hunching, lethagy, eye inflammation, squinting, anal swelling, sore feet, abnormal respiration, straining, reduced group interaction, and piloerection). It is possible that the hyperplasic but structurally unchanged topography of the GI tract in these cohorts allows the maintenance of a stable community structure with only marginal ingress by species made possible by changes in cage litter which are advantageous to environmental bacterial species capable of breaking down urea.Here it has been shown that disease progression may be driven by urease-positive, dermal, or environmental bacterial species rather than the expected GI microbiota which were found to not change during the study in any group. It may be that the push into pathology is driven by acquired species which cause additional discomfort to the animals on study.

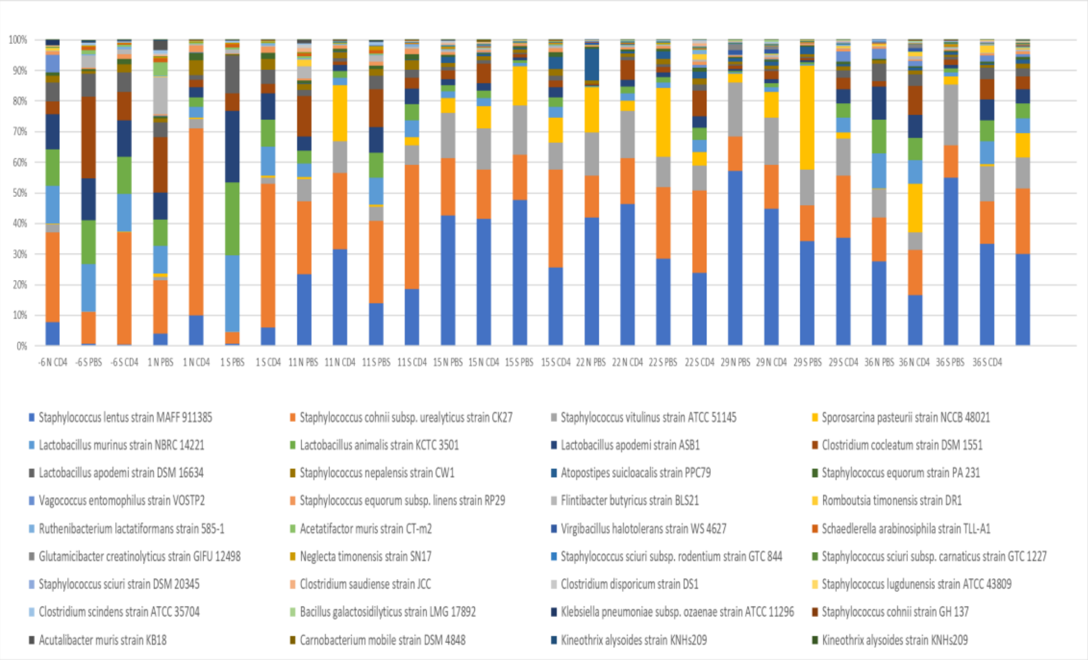
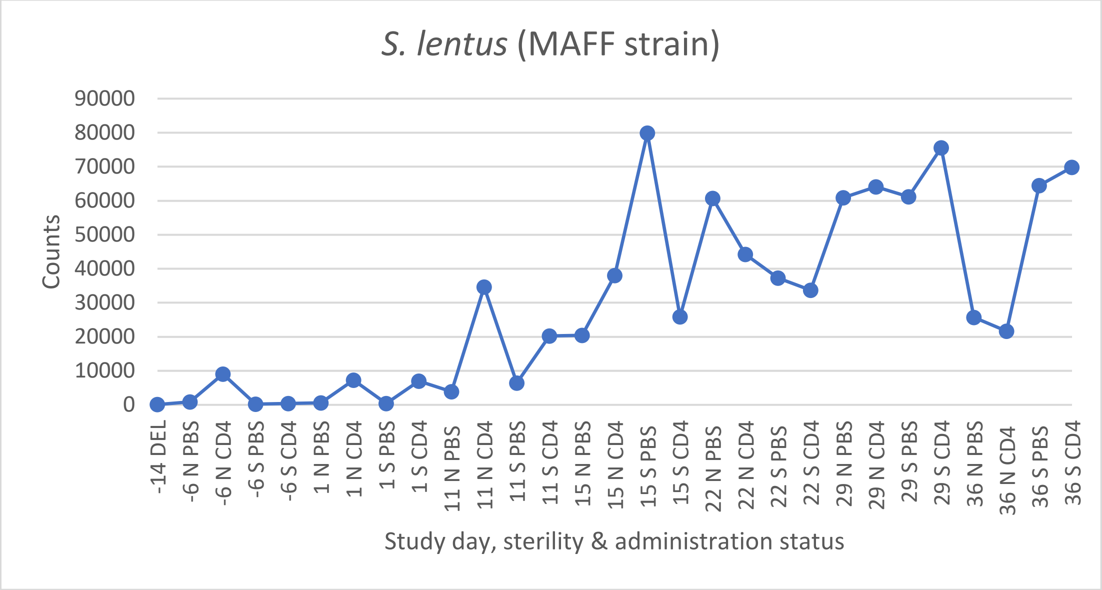


Figure 6-6: The relative abundance of OTUs obtained from C.B-17/IcrHsd-PrkcdSCID mouse faeces throughout the CD4+ study. This graph indicates the changes in prokaryotic diversity post-arrival and the subsequent changes throughout the study between the sterile and non-sterile groups. The day of study at which point samples were obtained (e.g., -6) is indicated, along with the sterility status of the group (e.g., Non or Sterile), and whether the group was a CD4+ or PBS recipient.



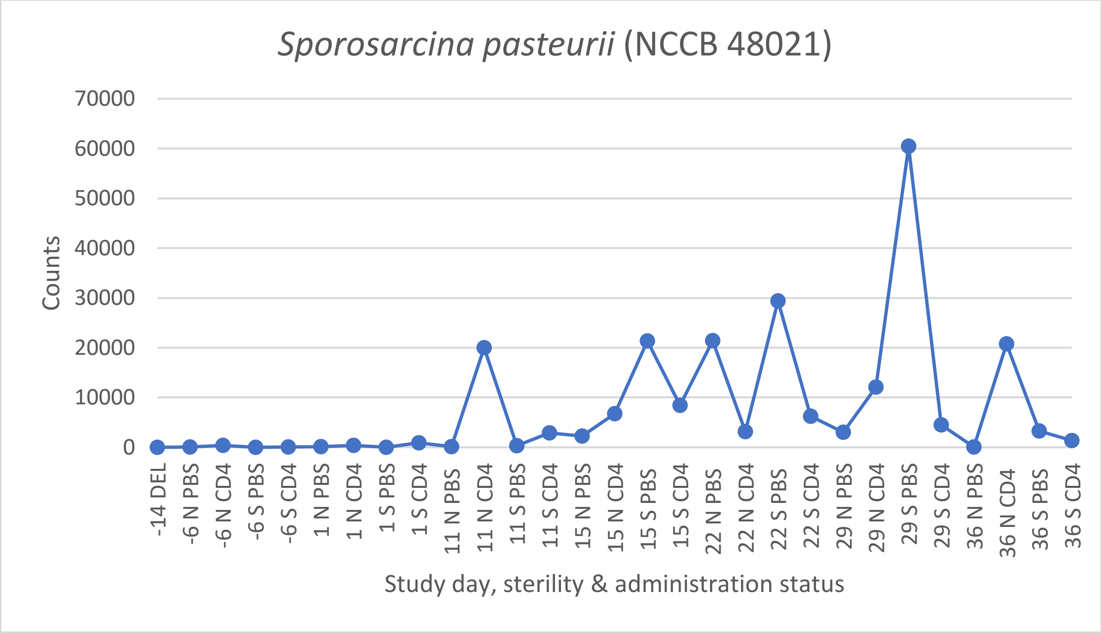


Figure 6-7: Two examples of bacterial species increasing in prevalence throughout the CD4+ study. *Staphylococcus lentus* (top) and *Sporosarcina pasteurii* (bottom) seen across all groups of C.B-17/IcrHsd-PrkcdSCID mice. The day of study at which point samples were obtained (e.g., -6) is indicated, along with the sterility status of the group (e.g., Non or Sterile), and whether the group was a CD4+ or PBS recipient .

6.2.3 Description of inflammation in the DSS model of colitis

Classically, the shortening and thickening, or tubularisation of the GI tract in mice during DSS administration is accompanied by rigidity and loss of function (Rieder & Fiocchi, 2008). Here, defining the GI topography by the examination of haematoxylin and eosin stained sections of the GI tract was made possible in all dose groups at study termination. This histological work illustrated the similarity in structure between those representing the H_2O control group (Figure 6-8A) and those illustrating a healthy GI tract (Figure 1-2). This structure, along with the lack of clinical signs, normal weight gain, and the absence of changes in prokaryotic diversity indicated a healthy GI tract. Sections from the 2% dose group (Figure 6-8B) shows a uniform change in GI structure at termination, with inflammatory ingression and hyperplasia. The section from the 2% dose group indicated further signs of microvilli architecture destruction and influx of inflammatory cells from the lamina propria (black arrow) along with a loss of normal glandular architecture possibly indicating previous ulceration and subsequent repair (Figure 6-9A). The section representing the 4% dose group is shown to be completely infiltrated with inflammatory cells and devoid of all functional achitecture (Figure 6-9B), which must have impacted normal absorbtion and functionality contributing to rapid weight loss. This analysis was useful in assessing the affect local architecture may have had upon the associated microbiota.

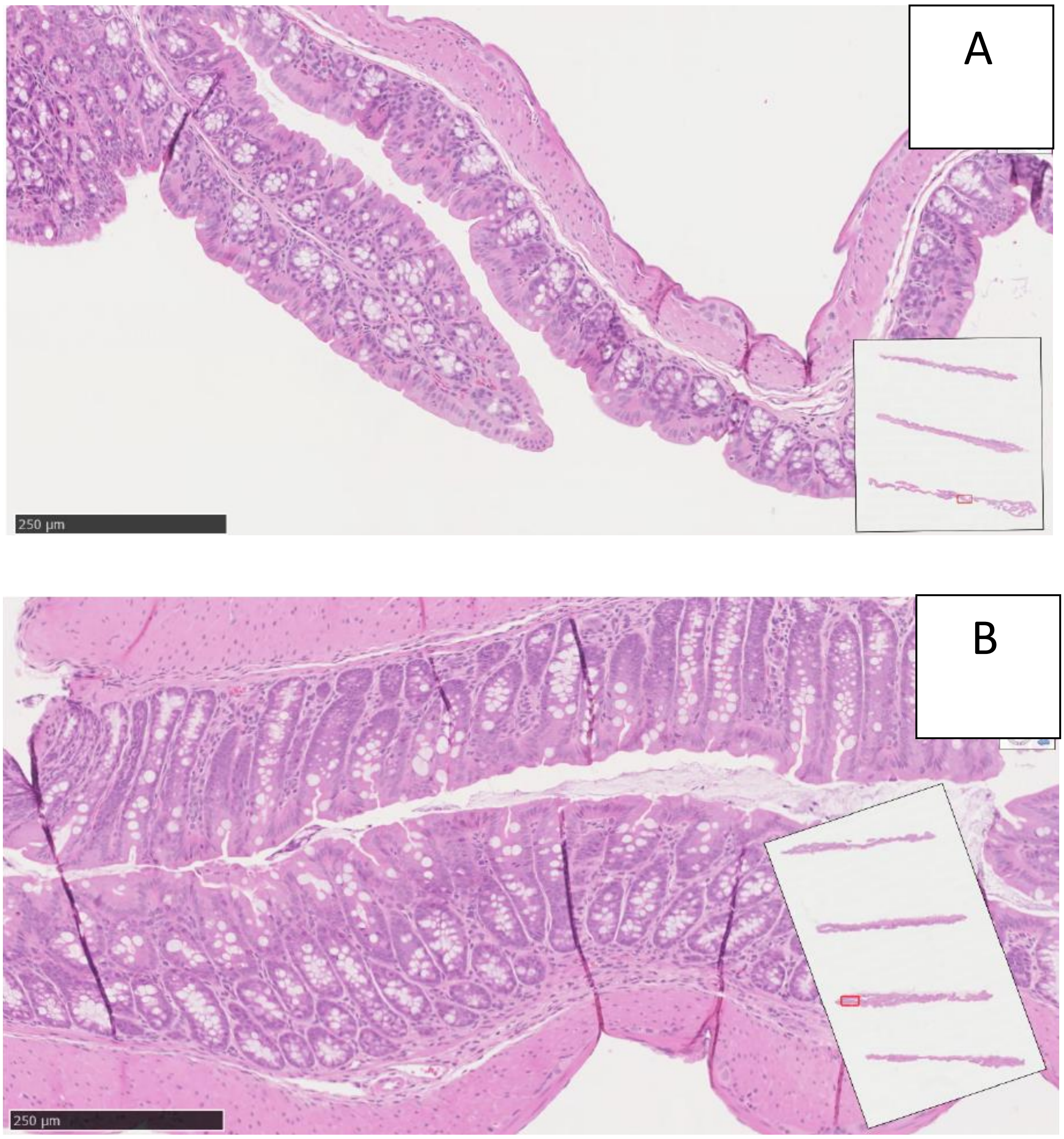


Figure 6-8: Haematoxylin and eosin stained C57BL/6 colon at termination of DSS study. (A) representing H_2O control group with no cellular changes in evidence and (B) indicating cellular proliferation (cells stained blue) and inflammatory swelling seen in the 2% DSS dose group. Sections and images generated by the Non-clinical Histology and UK Pathology teams (GSK).

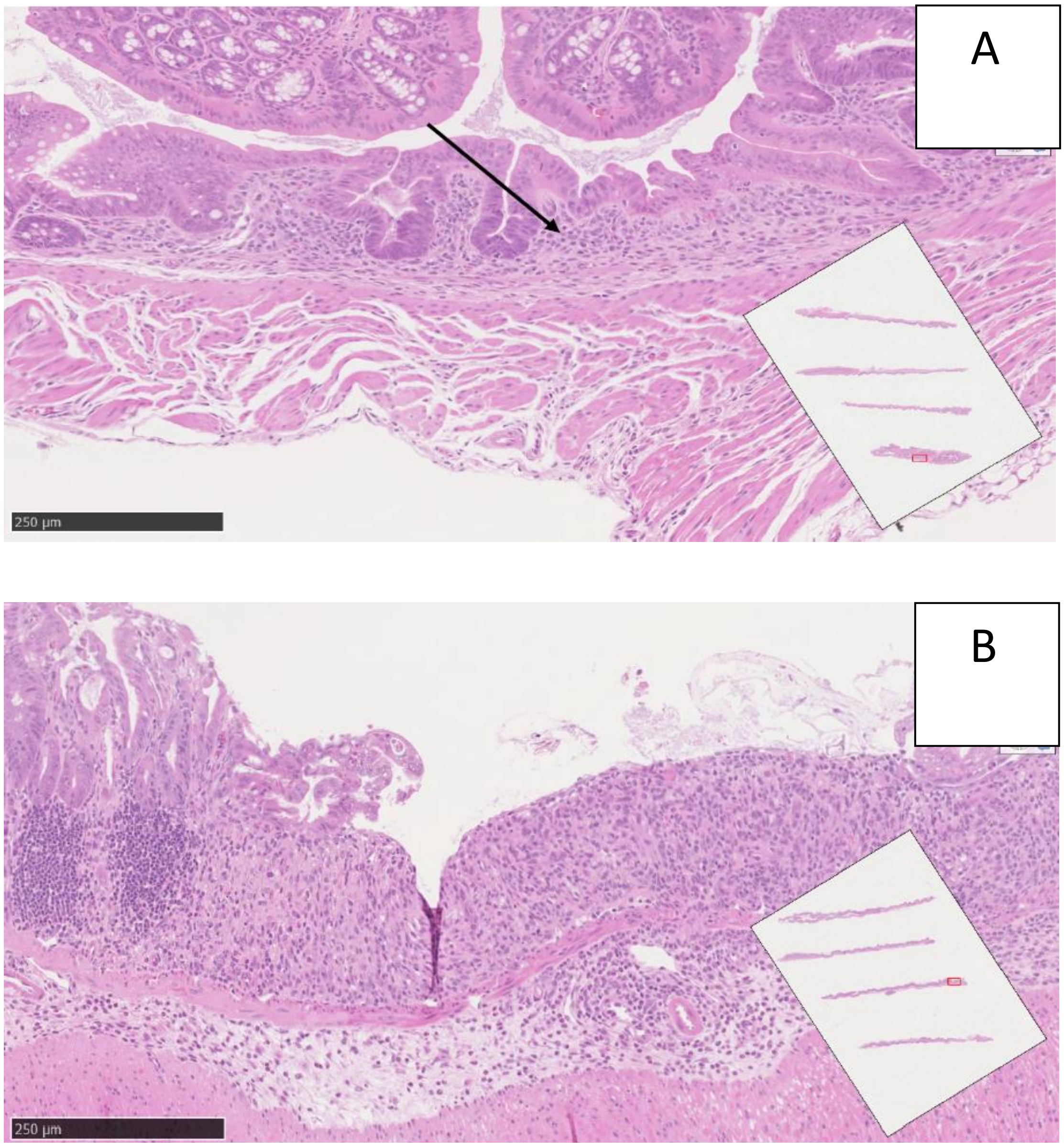


Figure 6-9: Haematoxylin and eosin stained C57BL/6 colon at termination of DSS study. Cellular infiltration (black arrow) by mucosal immune cells seen in the 2% DSS dose group (A) and (B) cell infiltrate (cells stained blue) within an ulcerated area of mucosa and loss of mucosal layer and functionality seen in the 4% DSS dose group**.** Sections and images generated by the Non-clinical Histology and UK Pathology teams (GSK).

Along with analysing the microscopic architecture by histology, a standard *post mortem* parameter used to measure the impact of DSS upon the physiology and functionality of the GI tract is colon density. This provides an organ-wide, macroscopic view of the altered environment in which the host's GI microbiota resides. A numerical density value is achieved by dividing length in mm (Figure 6-10A) by weight in grams (Figure 6-10B). This was carried out and is plotted in Figure 6-10C. The length of the GI is dose-linked with the 4% dose group's length being half of that found in the H_2O control group. However, organ weight was found to be uniformly increased in the 2%, 3% and 4% dose groups (170-260mg) when compared to the H_2O control group (145-195mg) (Fig. 6-10). This may indicate that water-driven hyperplasia and subsequent loss of epithelial cells is linked to cellular infiltration. Clustering of the eight values from each dose group indicated that all dose groups had a greater density value than the H_2O group, but are grouped together with no dose relationship being evident.

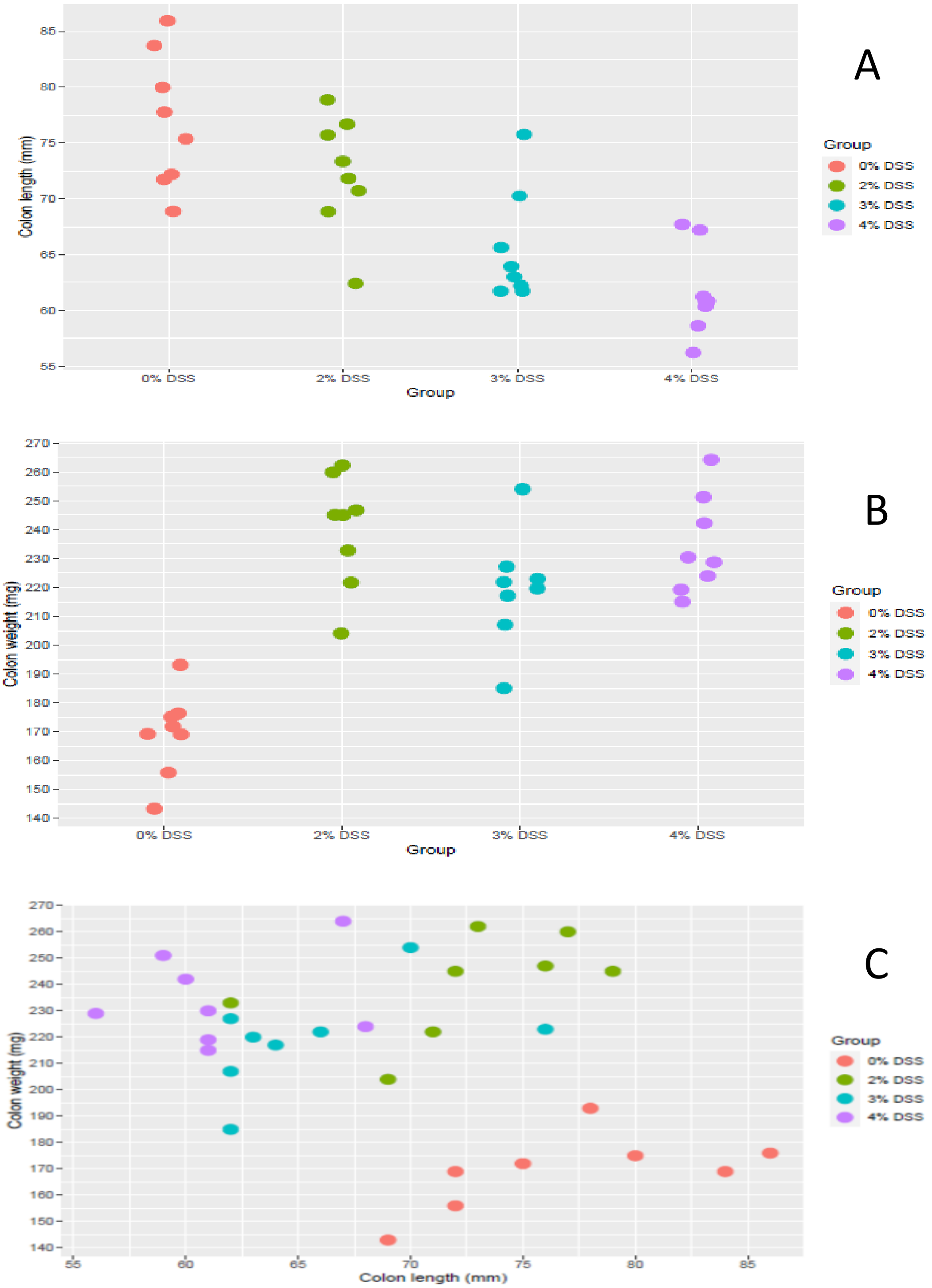


Figure 6-10: Comparisons of C57BL/6 colon attributes at termination of DSS study. The length (A), weight (B) and density (C) of the C57BL/6 colons at termination of the DSS study shown according to dose group.

As with the CD4+ study, in-life observations allowed disease progression or onset of welfare issues to be picked up during the study without the need to sacrifice the animal. Here, externally observed clinical signs were noted to give a numerical indictaion of disease progression and highlight animals which may require veterinary assistance or culling. These signs were hunching, lethagy, eye inflammation, squinting, anal swelling, sore feet, abnormal respiration, straining, reduced group interaction, and piloerection. The combined scores for each dose group show a range of clinical manifestations and the subjective nature of their observation (Figure 6-11), and the inherant varibility found in *in vivo* models is illustrated by a single finding in the H_2O control group. Six clinical signs were observed in the 2% DSS group of eight animals, while thirteen were observed in the 3% group. A total of thirty eight signs were observed in the 4% dose group by termination date. This reflects the decay in GI functionality in these high dose animals. Comparing this data with diversity is essential in understanding the effect of DSS upon the microbiota and the effect of DSS upon the host. It was found that clinical signs were only observed from Day 4 (Figure 6-11).

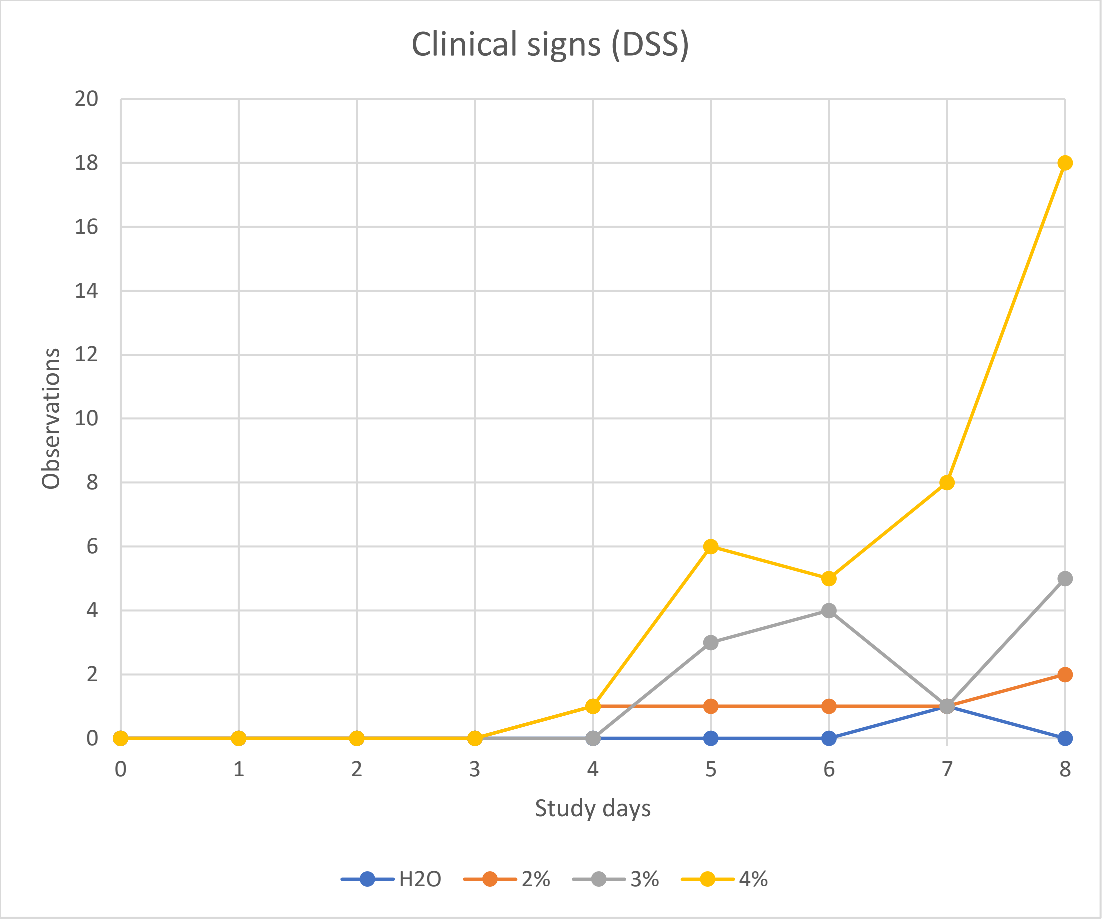


Figure 6-11: Total daily clinical observations made of C57BL/6 mice on each day of the DSS study. This graph shows data for the duration of the study according the dose groups indicating the difference in observations in the 4% DSS dose group compared to the the 3% and 2% groups. These signs were hunching, lethagy, eye inflammation, squinting, anal swelling, sore feet, abnormal respiration, straining, reduced group interaction, and piloerection.

DSS causes loss of membrane function and junction strength followed by cellular apoptosis and ultimately the destruction of the fine microvilli which reduces the host’s absorptive ability, removing and immediate host-microbiome interface, inhibiting faecal pellet formation (Anbazhagan *et al,* 2018). It is hypothesised that these three factors would affect prokaryotic diversity in the colon which has been shown to be measurable in faeces. Body weight was used as an in-life indicator of GI function (Figure 6-12) and here it was found that the H_2O control group continued to put on weight as healthy animals should throughout the study. This group was also seen to put on more weight after the provision of Hydrogel. The 2% dose group were shown to fall in weight through the study but regain this at DSS removal. The 3% dose group was seen to lose around 10% body weight but were also found to slightly gain weight in the last day of the study after DSS had been removed. All animals in the 4% dose group lost between 10-19% body weight by the termination day (Day 8). The loss of colonic function in this group continued to take effect after the removal of DSS and the provision of Hydrogel, which aided the lower dose groups to functionally recover. Therefore, in the 4% dose group, the chronic model of IBD was not found, rather the acute loss of GI function caused significant insult and injury to these animals. Theoretically, the lower doses (2% and 3%) where weight rebound was observed were more representative of chronic IBD which was an initial driver for this investigation.

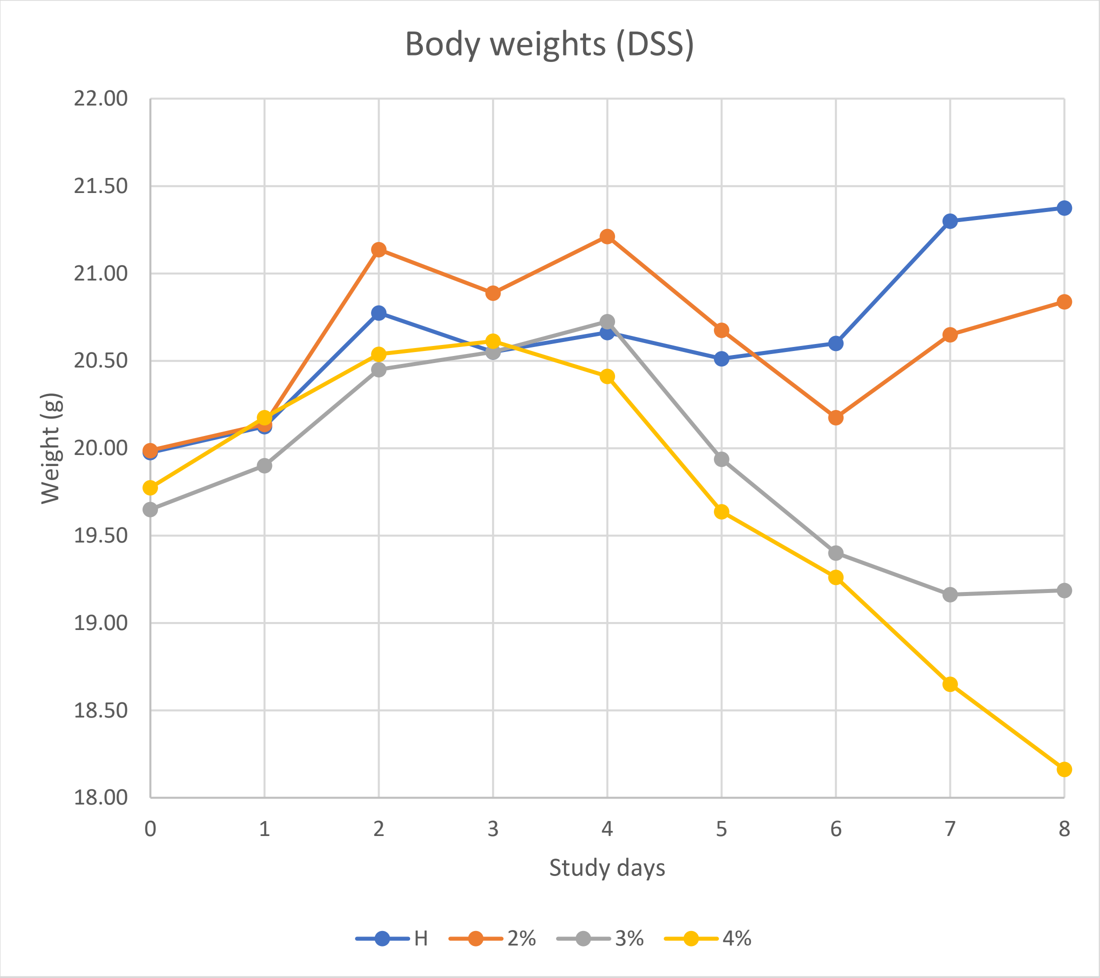


Figure 6-12: Daily mean averaged body weights of C57BL/6 mice in DSS dose groups. This graph shows the dose groups (H_2O, 2%, 3%, 4%) across the duration of the DSS study. It shows the normal weight gain of the H_2O group, the rebound of the 2% group, the plateau of the 3% group, and the decline of the 4% group.

A confounding factor in the DSS study was the administration of Hydrogel at delivery until DSS dosing (Day 0), and then from when dosing had finished (Day 5) until termination (Day 8). This material was provided at delivery to introduce the mice to this substrate and after dosing to augment recovery and weight gain. Hydrogel overtly provides 98% sterile water along with 0.07mg protein, 0.9mg carbohydrates, 0.9mg dietary fibre, 20.4mg calcium, 23.9mg phosphorus, 25.1mg potassium and 23.6mg sodium (www.clearh2o.com /resources). When water consumption was analysed during dosing and afterwards (Figure 6-13), it was found that mice in all dose groups prefer Hydrogel to the water provided. The added electrolytes and sources of carbohydrate and protein could affect the microbial diversity in the post-dose period. As no measurements of water consumption were taken prior to dosing, only changes in microbial diversity indicated this effect.

This factor is further confounded as it is known that DDS (a ribose sugar) is sweet to the pallet which drives successful self-administration of this caustic substance to the mice on study. All dose groups were found to increase intake of water when Hydrogel was removed (Day 0) and the only dose group which showed a decrease in DSS uptake (Figure 6-13; Day 4) was the 4% group where clinical signs were noted (Figure 6-11). When DSS was removed and replaced by water and Hydrogel, water consumption fell to ~15% of that in all dose groups during its withdrawal.

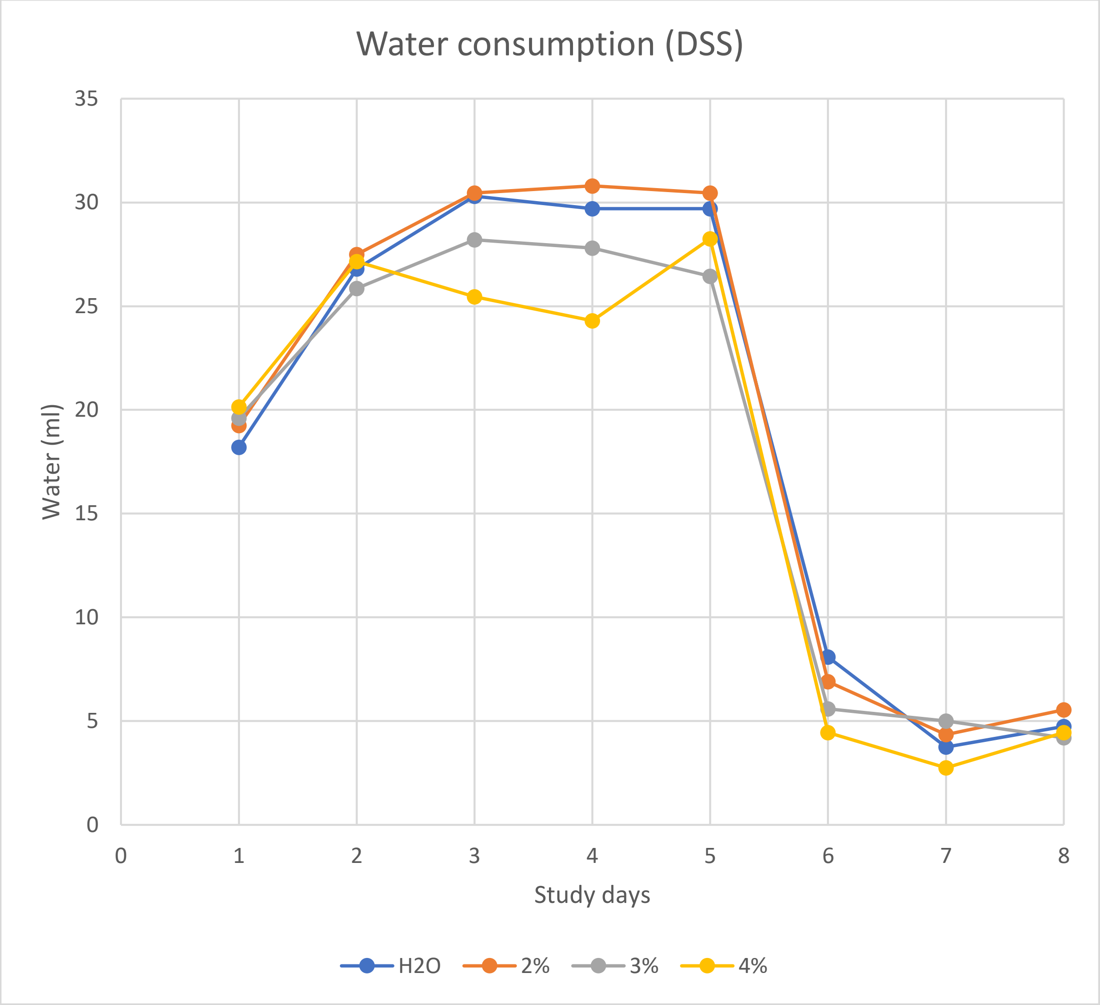


Figure 6-13: Daily mean averaged water consumption values in C57BL/6 mice in DSS dose groups. This graph indicates the period of DSS administration and subsequent replacement with water and Hydrogel. The fall in water consumption indicates the preference for Hydrogel over water when both were made available.

6.2.4 Prokaryotic diversity in the DSS mouse model of IBD

The loss of normal architecture is thought to massively impact the microbial community structure and function. Tracking prokaryotic community structures in the presence of GI injury was initially measured here by generating a Shannon index from the original OTU table (Figure 6-14). There was a fall in diversity at delivery which was seen in all earlier work apart from the CD4+ study, followed by only minimal fluctuations until dose administration on Day 0. Administration of DSS saw the reduction of diversity in all dose groups. However, the reduction in diversity was greatest in the 2% dose group, with the lowest change in overall diversity found in the 4% dose group. However, the diversity in the H_2O group also changed in concert with dose groups, while never being exposed to DSS.

Administration of DSS in drinking water and the removal of Hydrogel occurred on Day 0. It was shown by histology and colon density that the differing stages of injury resulted in dose-specific topography, in an equally dense tissue but in a dose-related length. Although hyperplasia was found at termination in the 2% group, ulceration indicated possible characteristics of repair by study termination. In-life observations indicated that clinical signs become apparent in the 4% dose group on Day 4 along with the 2% dose group, while the 3% dose group showed signs on Day 5. Weight loss was recorded in all groups from Day 4 (Figure 6-12).

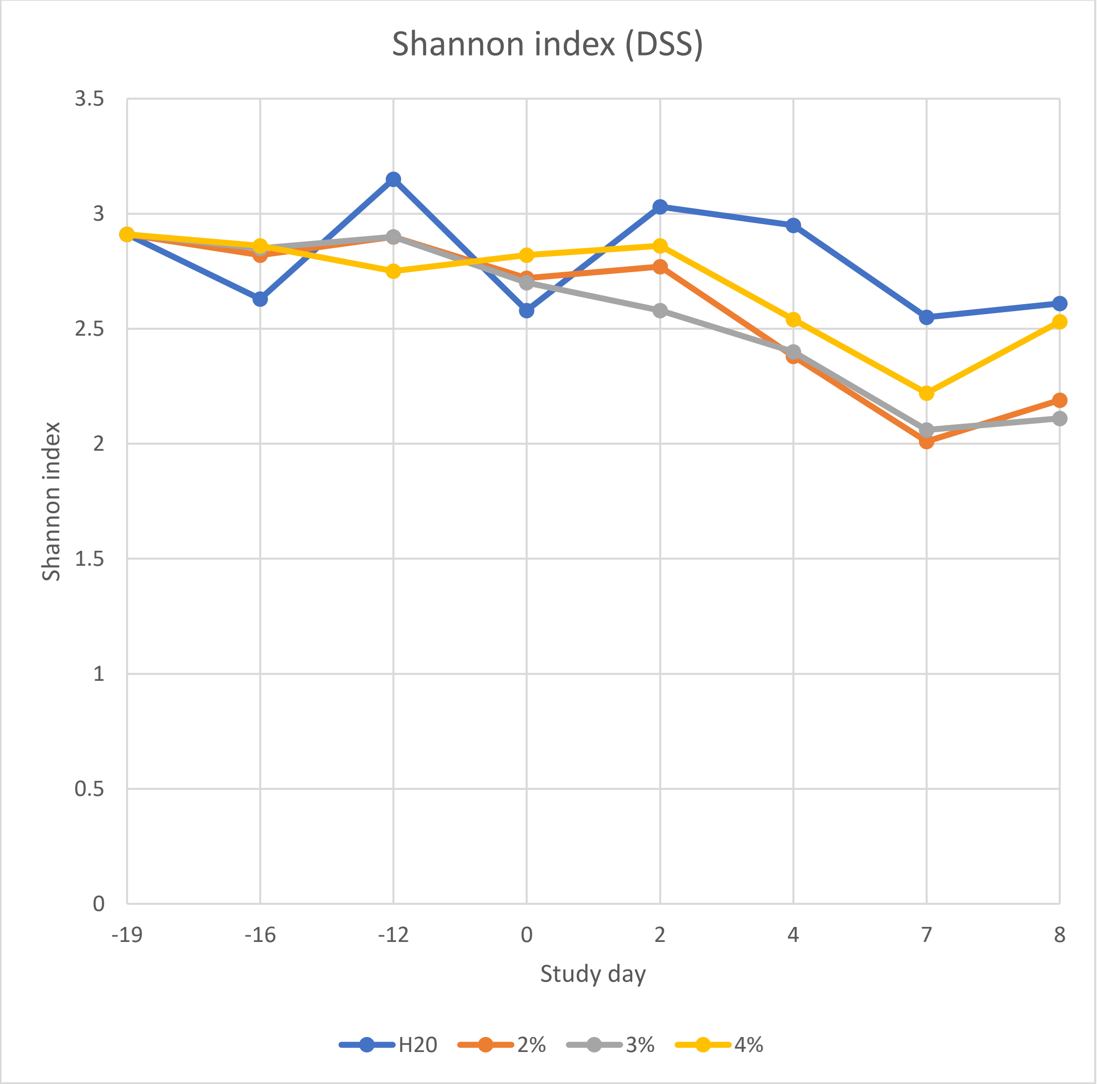


Figure 6-14: Alpha diversity analysis of C57BL/6 mouse faeces throughout the DSS study. This indicates the overall similarity in changes in diversity regardless of dose between 2%, 3% and 4% groups and the change in H_2O control group possibly due to changes in sources of hydration throughout the study.

Water consumption was seen to increase in all groups from DSS administration until it falls in the 4% group on Day 2 (Figure 6-13). These data show a lag in effects of DSS administration across all dose groups and a common period of observable measurements taking place between Day 1 and Day 5, at which point DSS was removed.

By looking at the Shannon index created for this period it is seen that diversity remains reduced in all dose groups until termination. However, more granular taxonomic structure of faecal samples across this same period (Figure 6-15). A lag is also evident, with no changes to bacterial species on Day 0 in any group but all (including the H_2O control group) had undergone a shift in microbial structure by Day 2. However, a rebound in bacterial species is seen in all groups from Day 4, when clinical assigns, weight loss and water consumption had all been affected and GI architecture altered.

Key bacterial species seen to fluctuate during the dosing period include *Ligilactobacillus apodemi* and *Akkermansia muciniphila* (Figure 6-16 top and bottom). *B. caecimuris* is only seen in small numbers until DSS administration (Day 0) at which point it increases >9-fold in all dose groups (Day 2) reducing to >4-fold (Day 4) remaining present higher than originally in dose groups until termination. The opposite effect is seen with *L. apodemi* which is seen to increase ~100-fold in the H_2O group during removal of Hydrogel (Day 2 to Day 4), returning to pre-dose levels until termination. *A. muciniphila* decreases in numbers by ~50% in all dose groups during initial administration (Day 0) of

DSS but rises in dose groups while DSS is still present to original levels but sees a 10-fold increase after the removal of DSS (Day 5)

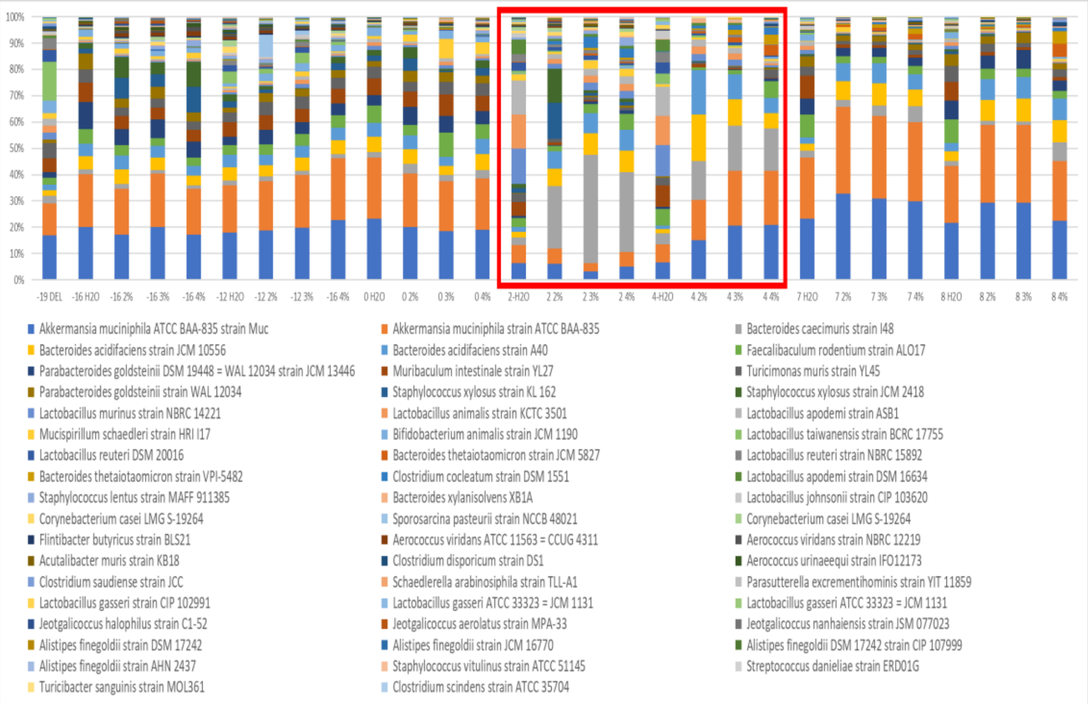
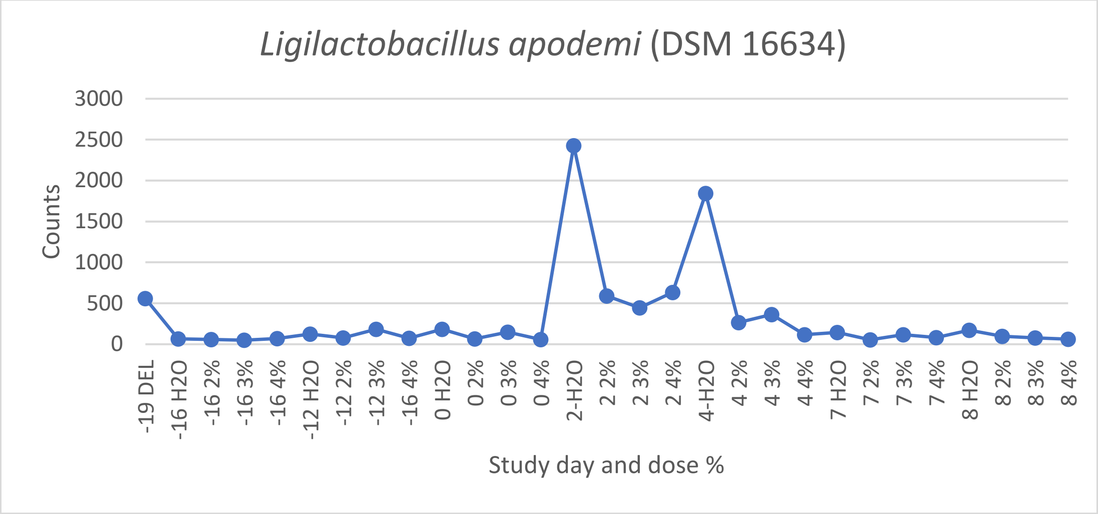


Figure 6-15: The relative abundance of OTUs obtained from C57BL/6 mouse faeces throughout the DSS study. This graph indicates the changes in prokaryotic diversity post-arrival and the subsequent changes throughout the study between all dose groups (H_2O, 2%, 3% & 4%) of C57BL/6 mice across the DSS study indicating the post-delivery change in diversity and that during DSS administration (red box) and replacement with water and Hydrogel. The day of study at which point samples were obtained (e.g., -6) is indicated, along with the dose percentage group (e.g., 4%).



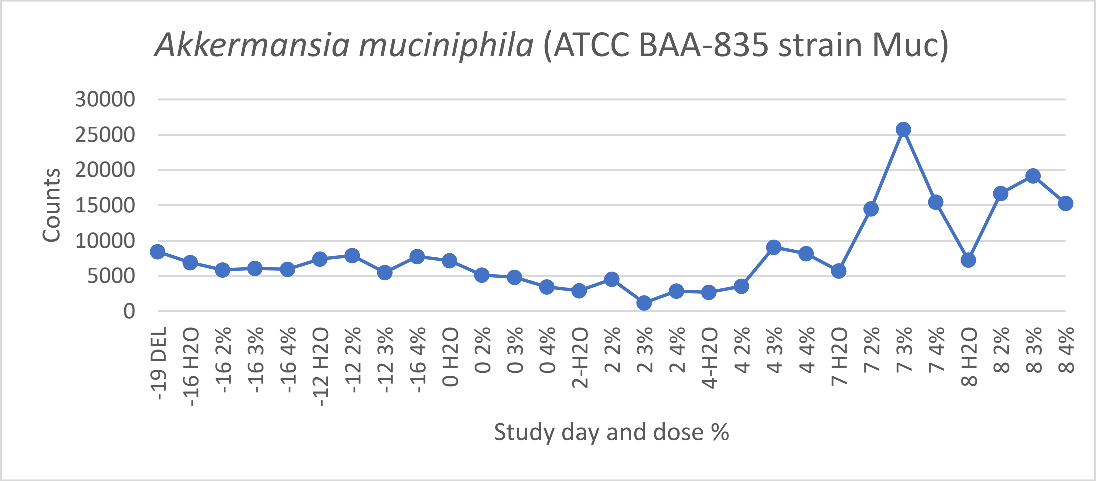


Figure 6-16: Two examples of DSS-driven bacterial prevalence throughout the DSS study. *Ligilactobacillus apodemi* (top) and *Akkermansia muciniphila* (bottom) seen across all groups of C57BL/6 mice in all dose groups (H_2O, 2%, 3% & 4%) throughout the DSS study.

Ligilactobacillus apodemi is a Gram-positive, non-spore former with tannase activity which uses many carbohydrates to generate acid (Osawa *et al*, 2006). As this species was seen to increase in numbers during DSS administration but only in the water groups it could be assumed that the concurrent removal of Hydrogel generates an environment which temporarily this species can acclimatise quickly. There is scarce literature regarding this bacterium but the genome sequence and functional inference catalogue of *L. apodemi* can be found at:

NCBI - https://www.ncbi.nlm.nih.gov/genome/?term=lactobacillus+apodemi or KEGG - https://www.genome.jp/entry/T07549.

Akkermansia muciniphila is a Gram-negative, cocci-bacillus which is known to degrade mucin and help reduce diabetes and obesity in mice (Zhai *et al,* 2019). *A. muciniphila* has also be shown to reduce pro-inflammatory cytokines e.g., IL6 and TNFα in a 2% DSS induced colitis mouse model (Bian *et al,* 2019). Here, *A. muciniphila* increased in numbers towards the end of the dosing regime (Figures 6-15 & 6-16).

Although stool consistancy is an established method of assesing disease severity in human patients and mouse models (Kim *et al*, 2012), all but one sample supplied for this work was normally formed and the experimental data for this measure was not supplied. The loss of the mucus layer is an early event in colitis in both human patients and mouse models with both a reduction of structural mucus protein expression and fall in numbers of goblet cells (Post *et al*, 2018). Both the host and the microbiota undergo transcriptional changes during this event. The host displays an altered transcriptional signature for activated macrophages and granulocytes as a defence while the microbiota upregulates genes involved in resistance to oxidative stress and nutrient deprivation as a survival mechanism (Ilott *et al,* 2016). The loss of mucus generating goblet cells during disease progression in this model may account for the initial fall of *A. muciniphila* in all DSS dose groups (Fig. 6-15), but not the recovery of this mucus utilising bacteria in all groups including the 4% where subsequent histologicasl analysis indicated the loss of normal villi structure and total loss of goblet cells. The post-dose rise in numbers possibly aids a reduction in inflammation and augments cellular and functional repair, illustrated by the return to weight gain seen in all groups apart from 4% DSS by the termination date. The terminal histological findings in the 4% dose group (Figure 6-9B) possibly indicate a structure devoid of crypt driven repair. The genome and functional inference gene data of *A. muciniphila* can be accessed at:

NCBI - https://www.ncbi.nlm.nih.gov/genome/?term=txid239935%5borgn%5d or KEGG - https://www.genome.jp/entry/T00736.

6.3 Summary

The SOP for routine microbiome analysis was successfully employed here in conjunction with two models of IBD. This method of characterising the prokaryotic diversity in faeces obtained non-invasively from mice increased understanding of how disease is instigated, and how the microbiome aids repair of the GI tract when specific community members are present. This method of characterisation negates pursuing costly DNA metagenomic and less robust RNA-seq methods attempted previously here (Chapter 4). By applying the hybrid method of generating species/strain level identities, it was possible to diagnose the presence of prokaryotes in a cost-effective manner from which gene complement can be extrapolated using freely available databases such as NCBI and KEGG. These curated and constantly updated resources provide the researcher with a stable and direct access library of gene function to study at no experimental cost. This wealth of information allows granular knowledge to be gained about the prokaryotes found in complex communities which would pass unknown if reliance on an exclusion-based health monitoring continued.

Currently, externally observable clinical signs such as weight loss, lethargy, piloerection, and faecal blood content analysis are the only in-life indication of disease progression in the CD4+ and DSS models of IBD. Blood analysis can supply an indication of lymphocyte proliferation but is highly invasive and is limited by sampling site injury and available volume. At termination, histological analysis of the GI tract can provide a single time point assessment of pathology. However, it was shown here that 16S rRNA gene characterisation of excreted faeces offered a non-invasive and temporal signal of disease progression in both models studied. Specific bacterial species were found to proliferate in the GI environment in each model (e.g., *S. lentus* in CD4+ and *A. muciniphila* in DSS). It may be possible to use these organisms as real-time biomarkers via quantitative PCR (QPCR) to measure disease progression while the study is running as a non-invasive method.

Ocular inflammation is an extraintestinal symptom of IBD in humans (Troncoso *et al,* 2017). It can affect all parts of the eye and can result in cataracts, optic neuritis, and oedema. It is thought to be initiated by the migration of microbial antigens to the periphery or misdirected immune responses to local self-antigens. It is also a known pathology in mice models of chronic IBD (Watts *et al*, 2013). Although no formal eye function analysis was conducted in these studies, it was found that squinting in the model became a significant contributing feature in the build-up of clinical signs in the CD4+ model. This may be because activated lymphocytes migrate to all mucosal membranes such as the mouth and the eyes. Here their numbers are fewer than in the gut but infiltration and interaction with commensal bacteria cause unregulated inflammation and probable distress to the animal. This effect may be exacerbated by the increasing numbers of staphylococci in the cage environment which colonised the mice on the CD4+ study. In this study, cage cleaning was kept to a minimum in order to present a bacterial burden to the animals. This work indicates that present husbandry practices are only accidently responsible for inflammation, and this is quite possibly off-target.

It has also shown that there is no difference in community structure between the sterile and non-sterile groups which provides evidence that the use of irradiated goods and water could be halted, saving money, and reducing cluttering from triple layered items in already cramped changing stations. This work has also shown that the water given to the mice on study does not play a role in inflammation as no waterborne bacteria such as *Ralstonia pickettii* or *Pseudomonas aeruginosa* were identified in faecal analysis. Presently, the idea that dirty water instigates CD4+ driven inflammation means that lower grade water is administered to animals on study. The use of multiple grades of water in a barriered animal facility is technically problematic but also provides a method of ingress and transmission to other models.

This work has provided and insight into bacterial dynamics in two mouse models of IBD. It has also suggested the use of more specific, non-invasive, in-life measurements. In addition to these scientific outcomes, it has provided information key to improving the welfare of mice during study design and which could positively affect the length and outcome of future studies. These positive impacts have fostered a tighter relationship with inflammation research groups and veterinary scientists which have seen the benefit of increasing the scope of health monitoring in line with the technology now available to the microbiologist.

Chapter 7: Discussion, future work, and conclusion

There is a need to embed microbiome analysis into routine diagnostic microbiology. Using NGS as an evaluative tool in this field would increase our understanding of the multiple aspects of the relationship between microbiome and host (Cullen *et al*. 2020). There is also a need to develop a greater understanding of the GI microbiota of the laboratory mouse to increase the relevance of its use in research (Kieser *et al*, 2022). This thesis aimed to evaluate which method of routinely characterising the murine microbiome would be best applied in the specific field of pharmaceutical development where no consideration, method, or infrastructure currently exists. This work would hopefully see the development of a novel approach to health monitoring, moving from a strict exclusion criteria (Mähler *et al*, 2014), to a full, in-life characterisation to aid model selection and gauge the effect of disease induction on the microbiome as a measure of the effect upon the *in vivo* model.

7.1 NGS data in a novel environment

Big data is defined as substantial amounts of computational output which cannot be managed using traditional software, hardware, and storage. It is said to grow in three dimensions; volume, velocity, and variety (Dash *et al,* 2019). Here, it was shown that with a limited number of samples and a reasonably small investment, all three aspects of big data growth were observed and presented immediate obstacles in the further use of this technology. The volume of data generated outstripped all physical storage boundaries, the velocity of its generation compounded this issue, and the range of data generated led to extended periods of analysis and a loss of scientific focus. Experiencing these impeding problems in a novel computational environment resulted in two outcomes. Firstly, once the breadth and wide applicability of NGS data was appreciated, the focus of subsequent work was regained by limiting the enquiry and correspondingly expanding control or improving of all aspects of data generation and analysis, reducing data analysis speeds, and improving specificity (Yin *et al,* 2017). Secondly, this deluge of data necessitated the permanent deletion of all but essential raw .fasta files and .astr alignment files. In a limited sense, these saved files become an accessible library, while the generated SOP represented the process by which the library may be accessed, and this thesis, the metadata needed to interpret the output. However, the permanent loss of all alignment files represented a block to quality assurance or fault-finding activities. These equally positive and negative outcomes mirror common bottlenecks in this field (Papageorgiou *et al,* 2018), which need to be overcome to allow successful and repeatable scientific enquiry (Eck, 2018). The activities described here, developed in an organic fashion, with progression only being made when locally acceptable workarounds were identified and subsequently demonstrated to be functional alternatives within the confines of the existing computational infrastructure or bounds of security policies. This widespread approach to problem solving in the face of quickly generated big data has created as many analysis concepts and workflows as groups researching the microbiome (Kulkarni & Frommolt, 2017). Each has had to deal with universal issues on an individual basis, with considerable effort to the detriment of evaluation of data (Tanjo *et al*, 2021). It would be preferential to submit data sets to an established and curated cloud storage server such as the European Nucleotide Archive for DNA Metagenomics (https://www.ebi.ac.uk/ena/browser/home) for open-access and utilisation (Harris *et al*, 2021).

However, in a conservative data environment such as the pharmaceutical industry, storing data on externally controlled cloud servers is not tolerated due to intellectual property and patent protection policies. The workaround achieved here was the storage of all data on external hard drives. This physically fragile method of holding data is open to deletion, damage, and criminal loss. Sharing data by this method is also not ideal as each copy will incur subtle compression changes. An ideal method of storage would be an onsite server-based source, from which identical libraries could be repeatedly accessed (Harris *et al*, 2021; Tanjo *et al,* 2021).

In establishments with pre-existing, low-level internet connectivity, the movement of data of this size was never considered (Eck, 2018). However, information is key in understanding problems, improving organisations, and driving developments (Dash *et al,* 2019). Without improvements to security, storage, policies, and investment in technology which would drive the controlled access to, and systematic use of data, nothing will be gained from its muddled generation (Toga & Dinov, 2015). An additional block to direct access and enquiry is that if this work is introduced to drug development studies rather than be used for health monitoring purposes then all sequence data would fall under the umbrella of a regulated study pack which would necessitate archiving along with all associated metadata. This represents a static repository, prohibiting access and enquiry. It seems that the generation of data which may have infinite enquiries made upon it, and which promises a degree of enlightenment by this inclusion (Coyte *et al,* 2021) may be permanently locked away.

7.2 Gene function: a bridge too far and a change in direction

With the fall in cost and rise in speed of NGS methods, has come a shift in experimental aim, from simple taxonomic form to potential and temporal gene function (Montonye *et al*, 2018). Community diversity and stability of the microbiome has been shown by ecological modelling to play a key role in functionality, with community structure being able to influence the ecosystem (GI) function, and the health of the host (Coyte *et al,* 2015). However, misdirected prokaryotic metabolism in the GI tract has been shown to reduce bioavailability and increase attrition in drug development (Spanogiannopoulos *et al*, 2016; Zimmerman *et al*, 2019). Routinely applying DNA metagenomic and RNA-seq methods to characterise the murine microbiome during drug development studies may facilitate an understanding of this potentially detrimental microbial activity. This approach could represent a monumental shift in health monitoring from a simple exclusion monitoring to one of cataloguing the total microbial diversity held within murine faeces and a snapshot of microbial gene expression (Beresford-Jones *et al*, 2022). This method would be easily translatable and as knowledge grows useful in the prediction of drug interactions from early phase *in vivo* experiments to late phase human trials. Extensive individual efforts have been made to catalogue the gene content of the murine microbiome indicating the possibility of generating this level of data (Kieser *et al*, 2022; Lagkouvardos *et al,* 2016; Xiao *et al,* 2015; Zhu *et a*l, 2020; Zimmerman *et al,* 2016). Some of these exploratory studies where augmented with an effort to culture and fully genome sequence previously unknown prokaryotes (Beresford-Jones *et al*, 2022; Lagier *et al,* 2018; Lui *et al*, 2020). However, few have described the routine application of this work to be possible using freely available bioinformatic pipelines (e.g., MG-RAST; Durrazi *et al,* 2021). Although it was hoped that DNA metagenomics (and RNA-seq) could illustrate all microbial domains (from phage to protist), the use of MG-RAST for the characterisation of eukaryotes is discouraged in the RAST handbook, as is describing with any certainty anything below the family level although this data is consistently generated (Meyer *et al*, 2017).

MG-RAST also allows the use of SILVA, RDP and Greengenes are RNA databases, which although permissive of 16S gene enquiries, contain both eukaryote and prokaryote references making searches inaccurate and slow, or they are taxonomically redundant. The inability to process 16S rRNA gene enquiries through the RefSeq database on the MG-RAST server was the catalyst behind moving towards the bespoke method of amplicon characterisation here as it was realised that this database represents the greatest breadth of filterable datasets or reference libraries (Camacho *et al,* 2009). The high-level curation of RefSeq (O'Leary *et al,* 2008) offers strain level identifications, rather than producing high level taxonomic designations resulting in diagnostic clarity.

Beresford-Jones *et al* (2022) showed that deep functional analysis of the murine microbiome was possible and that efforts to culture and genome sequence previously unknow prokaryotes is possible and improves our understanding of functional ties between host and microbe. However, in every sampling opportunity described in this thesis a novel and previously unseen microbiome was unveiled which was common to the group of animals screened indicating that however varied, housing conditions control the microbiome across individuals unlike the differences observed in humans (Gilbert *et al,* 2018). These novel ecological descriptions were found to undergo permanent changes after transit, or experimental interventions. Understanding these minute changes to the degree described by Beresford-Jones *et al* (2022) would be an impractical, lifelong endeavour with benefit to rapidly concluded *in vivo* studies.

This work found that the use of DNA metagenomics and RNA-seq generated too much unconfirmed data, which took too long to analyse in a fast-moving research environment. It showed as other have (Beresford-Jones *et al*, 2022) that this level of taxonomic and functional integration is possible but is impractical without the focus of a pre-existing microbial target or gene of interest. That 16S gene analysis and DNA metagenomics generate divergent data is known (Balvociute & Huson, 2017; Brumfield *et al,* 2020; Durrazi *et al,* 2021; Park *et al,* 2018; Peterson *et al,* 2021; Rausch *et al*, 2019). This is usually as comparisons are made between HiSeq/NextSeq and MiSeq, respectively generating >5x10^6 reads/sample versus 50,000 reads/sample (Peterson *et al,* 2021) or using different databases for each method, pre-loading bias into the analysis (Durrazi *et al,* 2021). In fact, it has been shown that in these comparative studies the choice of sequencing methodology (kit and platform) has more impact on resulting data variability than inter-sample differences (Clooney *et al,* 2015). If it does not make sense to seek consensus in one field of NGS (Pollock *et al,* 2018), then it is futile, to search for it between NGS methods. What proved more sensible was the return to an established method of taxonomic characterisation, improve upon it in a step wise fashion, and rollout its use as an ethical tool for focused microbiome analysis.

7.3 Towards a new idea of health monitoring

By applying a FELASA-based health monitoring approach only animals which are free from disease are used in research. From a veterinary perspective, this ensures that animals may exist free from the threat of microbial disease, while from a scientific perspective, it ensures that experimental data derived from animal studies is not confounded by the advent of a disease outbreak (Mähler *et al*, 2014). The pressure to remove certain microorganisms from rodent colonies via embryonic rederivation has subtly influenced the host microbiota (Franklin & Ericsson, 2017) with each commercial animal supplier generating standardised models and animal lines with FELASA-based health reports. Remarkably, whilst this may imply microbial standardisation, this study illustrated how each model, line, study, and location analysed displayed a unique prokaryotic GI microbial signature. Moreover, these communities were found to shift during acute and chronic interventions such as transport or intervention. This suggests that the ecoevolutionary dynamics of microbiome community structure are subject to ecological (external) changes that drive changes in the microbiome, but we cannot rule out that subtle changes to a single (or small number) focal species can further modify the environment as has been observed in macrobiology (Hendry, 2017). It was recently shown that ecological assembly and dependencies can be predictable in microbiome communities and through modelling and experimental studies that microbe to microbe and microbe to host interactions govern the trajectory of microbiome assembly (Coyte *et al*, 2015).

Identifying bacterial roles from a 16S rRNA amplicon NGS is problematic because high-level classifications (phyla, class) do not provide the level of detail required for species level assignment and there is limited information available for newly discovered and unculturable members of the microbiota were available (Lagkouvardos *et al,* 2016). Although metagenomic sequencing can be used to define both nomenclature and function (Jovel *et al*, 2016), when combined with MG-RAST it was found to be technically difficult and diagnostically inaccurate. RNA-seq work which is mostly used for temporal gene expression analysis also found diagnostic utility through the MG-RAST server (Brumfield *et al*, 2020; Escobar-Zepeda *et al* 2018; Rausch *et al*, 2019;

Tessler *et al*, 2017), but in this study, this failed to generate usable data and suffered from the technical issues identified in the metagenomic work (Chapter 4).

Refining the 16S gene analysis approach and creating a highly specific reference database allowed rapid species level diagnostics to be achieved, and when coupled with literature research enabled an understanding of each species and their potential roles in the host. This approach also is suitable for next generation health monitoring system of animal colonies, as described in this work. This approach is also more inclusive and enables several taxa to be monitored and as data and understanding of systems become better understood may enable early warnings of dysbiosis to be identified. Monitoring these activities can also help explain how the microbiota dynamically interacts with the host.

In 2021, 25,059 articles with the word '*microbiome*' in the title were published globally (https://pubmed.ncbi.nlm.nih.gov/?term=microbiome), which equates to 2.86/hour. Many of the experiments described utilised different bioinformatic pipelines and different reference databases, such that the constant reinvention of better or more illustrative tools and pipelines may be counter-effective and possibly meaningless (Hill, 2018). This study shows that tools and pipelines can be used to answer some questions from some input data but often these approaches do not give appropriate outputs to glean useful information in the context of health monitoring. This work aimed to identify a unifying methodological pipeline for routine deployment in a pharmaceutical research setting. The use of 16S rRNA gene-based phylogeny has been the most widely used method by which ecological niches may be characterised and compared (Yarza *et al,* 2014). The pipeline defined here is simple, but appropriate and can be used with confidence to develop further our understanding of the prokaryotic communities and their dynamics in the mouse in routine health monitoring.

7.4 Can dysbiosis or dysregulation be observed in models of IBD?

Dysbiosis is defined as a change in community structure which negatively impacts the host (Levy *et al*, 2017). The functional redundancy in the microbiota, allows temporary impairments, such that the condition remains subclinical, with the host able to survive in the absence of specific organisms or their metabolites. However, this survival does not equate to wellbeing. The benefits of a functional microbiota are wide ranging, such as the complete breakdown of indigestible carbohydrates (Haller, 2018), generation of vitamins (Mu *et al*, 2018), contributions to enterocyte health (Haller, 2018), inhibition of pathogen adherence and regulation of immune responses (Byndloss *et al*, 2018) and potential contributions to mental wellbeing (Vuong *et al,* 2017). We have also seen that these balanced functions are regulated at innately fine detail by both the host and microbiota (Kiu *et al*, 2020).

In the initial experiments (Chapter 3), significant differences in community structures were observed between GI niches and between sampling locations. The former being driven by physiology and immunoregulation (Hill & Artis, 2010), and the latter being driven by the stress of transport upon the host. It has previously been shown that transit related community disruption regained homeostasis at around five days post-transit, indicating an extended period of flux followed by resolution, if not rebound (Ma *et al*, 2012), which was not observed in this study. Other studies have shown that although acute transit induced disruption does occur, homeostasis is only regained between one- and four-weeks post-transit (Montonye *et al*, 2018). Here it was seen that transport immediately resulted in the loss of almost one third of species found in the GI tract of C57BL/6 mice, creating a new community structure did not return to its former diversity within an acclimatisation period of seven days, or within two weeks of arrival. It is suggested that these fluctuations may contribute to differing study outcomes at distinct locations (Montonye *et al*, 2018). However, in initial experiments and in the subsequent longitudinal NSG study it was shown that a physiological experience endured by the host such as transit impacts prokaryotic diversity at transit and this new community remains stable for at least the four-month period of screening conducted. It has been shown that age (>48 months) eventually affects GI diversity in C57BL/6 mice (Langille *et al*, 2014), indicating that longer term studies may be affected by the overt functional decline in cellular, metabolic, and immunological processes (Bajaj *et al,* 2021). This again illustrates that the autochthonous microbiota is affected by external factors.

Another external driver to overall GI diversity in mice is the systemic use of embryonic rederivation used to remove overt pathogens from breeding colonies (Nicklas & Seidel, 2019), potentially leaving this model species a poor reflection of its wild counterpart, reducing immunostimulatory potential while promoting disease phenotypes (Rosshart *et al*, 2017). Arguably this is beneficial in models of disease but creates a housing environment which promotes the transmission of deleterious microorganisms which maybe subclinical in the immunocompetent and challenged wild population (Becker *et al*, 2007). The limited microbiome diversity of the immunodeficient C.B-17/IcrHsd-PrkcdSCID used in the CD4+ study was seen to gain numbers of specific bacterial species post-arrival at the new location rather than lose diversity. This may be a result of this genetic lines inability to limit colonisation due to the lack of T- and bursal or bone marrow derived (B)-cells (Envigo, 2022). The observed rise in staphylococcal species number and diversity is possibly due to human transmission during cage husbandry in receiving cages (Ferrecchhia *et al*, 2014). However, their innate GI diversity was not seen to alter at the administration of Balb/C CD4+ T-helper cells or thereafter. The expected cellular infiltration of the lumen was therefore shown to not have a downstream effect on microbial diversity possibly as this is a model of cellular activity measured below the lumen, rather than one measured by the destruction of the microvilli from above as in the DSS model (Prattis & Jurjus, 2015).

The C57BL/6 mice used for the DSS study (Chapter 6) showed a significant change in GI diversity post-arrival with a fall in gut Lactobacilli. Hydrogel and DSS are more palatable to mice than water alone and each of these substances were given exclusively at different points of the study which may have played a part in relative body weight changes and microbial diversity. The overt epithelial damage and barrier dysfunction along with tubulerization caused by the administration of DSS into the GI tract firstly affects GI physiology and then the resident microbiota (Eichele *et al*, 2017). In low dose groups (2% and 3%) bacterial species lost at administration were seen to rebound and host health markers were seen to increase when DSS was removed, whereas the 4% dose group suffered from terminal loss of GI function and a highly dysbiotic community structure which saw an exclusive increase in carbohydrate harvesting species possibly utilising the administered DSS.

All the changes to prokaryotic community structure observed here represent dysbiosis driven by external events forced upon the models screened. This work firstly supports the theory that transport initiates a change in the microbiota possibly as a result of stress (Ma *et al*, 2012) but consistently shows that this disruption occurs during transit and remains stable over time thereafter. Secondly, it shows that disease instigation in mouse models and subsequent husbandry choices can drive the level, and species specificity of dysbiosis. This work provides a tool by which the breadth of dysbiosis can be measured in such models, aiding greater understanding of subclinical changes that may occur during research and drug trials.

7.5 Partnering animal welfare and drug discovery

Both IBD studies investigated here were sham or pilot and so no compounds were used. This reduced the number of variables which could have contributed to changes in the prokaryotic community. The CD4+ study was conducted to measure any difference in pathology or out come between a sterile group and non-sterile group as it was though that housing in non-sterile conditions would generate greater microbiome induced upregulation of immune responses post-CD4+ transfer. Here no changes were seen in the microbiota at arrival, at administration or between the sterile and non-sterile groups. The observation that there was no difference between sterile and non-sterile group microbiotas indicates that the continued administration of sterile diet and water could stop in subsequent studies. This would free up time and remove working constraints as husbandry is conducted in sterile change stations which become crowded with the triple wrappers which are used to contain the irradiated goods. Additionally, stopping the use of irradiated diet and water will reduce study costs.

In the DSS pilot study, no antagonistic pharmaceutical agent was administered so that the observed changes in prokaryotic microbiota were attributable solely to the concentration of DSS administered in the drinking water which was used to make gross observations in the health of the mice or possibly the administration of Hydrogel (Figure 6-14) when DSS was removed. These observations or scores were added to those conducted at termination to gauge the effective dose needed to generate a disease model to be used in subsequent elevation studies. The range of outward signs observed in the 4% DSS group indicated that previous literature (Munyaka *et al*, 2016) was correct in using 4% to illicit a significant representation of colitis in the mouse.

However, the 4% dose group was found to lose between 11-19% body weight as the destruction of the GI tract was so severe that functionality had possibly been terminally impaired. In the 2% and 3% dose groups similar but less marked changes were observed in mice which did not lose >10% bodyweight during or after DSS administration and at histological examination. It would therefore be possible to use the lower dose of 3% and have a model that can withstand DSS administration but still provide evaluative data used to measure disease progression and alleviation by the administration of an antagonist over a longer period. By measuring the microbial changes in dose groups throughout the study, it was possible to illustrate significant or relevant changes on a scale in relation to specific bacterial species which corresponded with weight loss and gain and unobservable GI alterations and tubulerization. The rebound in bacterial species, weigh gain and faecal output all indicate that the use of 2% or 3% DSS doses in this study would be possible if routine and timely use of community profiling were employed.

The observation that all mice have a unique community structure according to source, indicates that it would be possible to use 16S amplicon based NGS to take a census of animals on study which could then be refined to use specific bacterial species as markers of disease progression in IBD studies. Both findings illustrate the immediate practical use of 16S amplicon based NGS in animal studies to reduce animal suffering, reduce animal usage, and refine techniques used in the gleaning of data used directly in research groups for decision making thus aiding drugs to market through more refined techniques and granular understanding of the mice models used in pharmaceutical research today.

7.6 Future work

Although this thesis resulted in the successful development of a robust microbiome characterisation method several technical and sampling improvements could be adopted to further develop this method into a highly accurate tool for the improvement of animal welfare and study outcome.

7.6.1 Technical improvements

It is now possible to perform near-full length 16S gene sequencing (Johnson *et al,* 2019). This method uses V1V2, V2V3, V3V4, V4V5, V5V7 and V7V9 primers sets to generate overlapping amplicons covering the variable regions, creating a mosaic of the 16S gene. By applying near-full length 16S gene reads to the novel database method of characterisation described here, resulting identities would be of the highest specificity (Johnson *et al,* 2019). This would allow exceptional diagnostic clarity to be achieved from complex samples removing many of the issues experienced here.

The final workflow described used the NCBI 16S gene curated list as a reference database. Stand-alone analysis of its 21,653 entries was possible on a laptop as the reference scaffolds are around 1500bp long, query sequences are up to 300bp length and total reads do not exceed 100,000. Theoretically, it would be possible to use the NCBI GenBank database (presently holding 474,000 bacterial RefSeq entries) as a reference tool for prokaryotic DNA metagenomic and RNA-seq data sets in the same way. The reference database would be of equally high quality as the 16S rRNA one, and it too would be openly accessible and constantly updated. The only hinderance to applying this alignment method to larger sequencing files may be an increase in processing times, which could necessitate the use of larger processors.

If this increase in processing capacity was tied to the acquisition of better-quality RNA-seq data from faeces, an avenue to understanding microbial gene activity would be possible. It is expected that by utilising an RNA stabilizing agent and a specific processing method such as the RNeasy Powersoil kit that robust RNA-seq data could be generated from murine faeces. If these steps were successful, it would also be possible to use further NCBI directories such as viral, protist and fungal (accessible at http://www. ncbi.nlm.nih.gov.nuccore) to gain high quality diagnostic characterisation of all classes of microorganism from DNA metagenomic and RNA-seq datasets. However, the concurrent identification of all classes of microorganism would never be straightforward by this approach, however, multi-locus sequence typing (MLST) of internal fragments of alternate house-keeping genes could be used (Urwin & Maiden, 2003).

7.6.2 Wider sampling

Due to the ubiquitous nature of faecal matter, and its classification as waste material, any widening of sampling is easy to achieve and would not impact existing study designs or future considerations. It is thought that by sampling all animals at delivery, a characterisation of each commercial supplier would be possible and the subsequent loss of species during transit from each, according to length of travel time. It would be beneficial to sample animal facilities over time to understand whether a facility-specific microbiota exists (Montonye *et al*, 2018) and whether continual imports from multiple suppliers alters this microbial signature over time and if specific community members are always lost in transit. By tethering this advanced health monitoring analysis to complex study outcomes utilising immunocompromised or transgenic models and themes of research (inflammation, cancer therapy etc.) over time, a more granular understanding and awareness of the impact of the murine host's microbiota on drug discovery will become possible.

7.6.3 Reducing attrition

Attrition in drug discovery can be due to unwanted microbial metabolism (Spanogiannopoulos *et al*, 2016). Understanding this activity is possible using the method of characterisation described here by extrapolating accurate taxonomy to functional database searches. Applying this level of analysis in breeding facilities as an extension of health monitoring would build up ecological data. Applying this method to all study types may not always be of use but piggy backing on compound toxicity studies would also provide clear dose graduation versus non-dose variation, a clearer understanding of compound class interactions, and greater statistical power due to the larger number of animals used in this type of investigation. Additionally, toxicity studies are usually run using rats which generate more faecal and GI digseta.

This would allow both technical replicates (validating method reproducibility) and biological replicates (validating dose variations) to be run (Robasky *et al*, 2014). By introducing this evaluative method information would be available to research groups which could illustrate prokaryotic sources of off-target metabolism found in drug metabolism and pharmacokinetic mass spectrometry screening. This would inform the clinic of potential drug interactions, but pre-study screening of rodent colonies or strains would allow the inclusion or exclusion of these metabolic species.

7.6.4 Embedding this work in pharmaceutical research

The embedding of microbiome analysis in any form as a routine evaluation tool remains in its infancy (Cullen *et al*, 2020), despite nearly twenty years of next generation sequencing research and the intention to catalogue the microbial diversity of Earth's entire ecology (Thompson *et al,* 2017). This work illustrates a real world application of this established and mutable technology. However, its wider adoption still rests upon visible and understandable benefits rather than huge, expensive studies with real world applicability. This weight of evidence is only obtainable if researchers can be persuaded to include this type of analysis in their body of research. However, generating yet another method by which a study can be cancelled is not what is usually wanted. The step wise generation of applicable data based on causality can be generated by piggy backing on relevant breeding programs and studies such as those with GI delivered molecules or specific microbial targets may result in this work being accepted for pre-study screening or in-life monitoring of diversity changes in relation to the activity of one key species (Hendry, 2017). Communicating these results in a timely fashion to the researchers in person and the wider animal research community and discovery research groups would hopefully see the consideration of the second genome in drug research becoming routine (Grice & Segre, 2012).

7.7 Conclusion

In summary, this work has succeeded in establishing a robust sampling and 16S gene analysis pipeline for the routine, in-life characterisation of the murine GI microbiota where no previous method existed. Considerable progress was made in understanding the applicability of NGS in this field. This understanding has enhanced routine health monitoring, enabling the characterisation of which prokaryotic species are present in a host animal rather than discerning those which are not. It has begun conversations with the research community about the mutability of the microbiome during drug development studies, how this may be used diagnostically, and how the health of the microbiome is tightly linked to the health of the host which is central to the use of animals in research

Chapter 8: References

Abedon, S.T. (2008) *Bacteriophage ecology – Population growth, evolution, and impact of bacterial viruses*. Cambridge University Press. Cambridge. UK.

Acera, M.G., Patanker, J., Diemund, L., Siegmond, B., Neurath, M., Wirtz, S. & Becker, C. (2021) Comparative transcriptomics of IBD patients indicates induction of type 2 immunity irrespective of disease ideotype. *Frontiers in Medicine*. 8:664045.

Allaire, J., Crowley, S., Law, H., Chang, S., Ko, H. & Vallance, B. (2018) The intestinal epithelium: central coordinator of mucosal immunity. *Trends in Immunology.* 39(9):677695.

Aminov, R.I. (2013) Role of Archaea in human disease. *Frontiers in Cellular and Infection Microbiology.* 3(42):1-4.

Anbazhagen, A., Priyamvada, S., Alrefai, W. & Dudeja, P. (2018) Pathophysiology of IBD associated diarrhea*. Tissue Barriers.* 6(2):e1463897.

Artis, D. (2008) Epithelial-cell recognition of commensal bacteria and maintenance of immune homeostasis in the gut. *Nature Reviews: Immunology.* 8:411-421.

Aziz, R., Bartels, D., Best, A., DeJongh, M., Disz, T., Edwards, R., Formsma, F., Gerdes, S., Glass, E., Kubal, M., Meyer, F., Olsen, G., Olson, R., Osterman, A., Overbeek, R., McNeil, L., Paarman, D., Paczian, T., Parrello, B., Pusch, G., Reich, C., Stevens, R., Vassieva, O., Vonstein, V., Wilke, A. & Zagnitko O. (2008) The RAST server: rapid annotations using subsystems technology. *BMC Genomics*. 9:75.

Bajaj, V., Gadi, N., Spihlman, A.P., Wu, S.C., Choi, C.H. & Moulton, V.R. (2021) Aging, immunity, and COVID-19: how aging influences the host immune response to Coronavirus infections. *Frontiers in Physiology*. doi:103389/fphys2020.571416.

Baltimore, D. (1971) Expression of Animal Virus Genomes. *Bacteriological Reviews*. 35(4):235-241.

Balvociute, M. & Huson, D.H. (2017) SILVA, RDP, Greengenes, NCBI and OTT – how do these taxonomies compare? *BMC Genomics*. DOI 10.1186/s12864-017-3501-4.

Becker, S.D., Bennett, M., Stewart, J.P. & Hurst, J.L. (2007) Serological survey of virus infection among wild house mice (Mus domesticus) in the UK. *Laboratory Animal.* 41(2):229-238.

Belkaid, Y. & Hand, T. (2014) Role of microbiota in immunity and inflammation. *Cell.* 157(1):121-141.

Beresford-Jones, B.S., Fortser, S.C., Stares, M.D., Notely, G., Viciani, E., Browne, H.P., Boehmler, D.J., Soderholm, A.T., Kumar, N., Vervier, K., Cross, J.R., Almeida, A., Lawley, T.D. & Pedicord, V.A. (2022) The mouse gastrointestinal bacteria catalogue enables translation between the mouse and human gut microbiotas via functional mapping. *Cell Host & Microbiome*. 30:124-138.

Bergey, D.H. (1994) *Bergy's Manual of Systemic Bacteriology* (9th edition) Lippincott Williams & Wilkins. Philadelphia. USA.

Bhattacharya, T., Ghosh, T. & Mande, S. (2015) Global profiling of carbohydrate active enzymes in the human gut microbiome. *PLOS ONE*. 10(11):e142038.

Bhinder, G., Sham, H.P., Chgan, J.M., Morompudi, V., Jacobson, K. & Vallance, B.A. (2013) The Citrobacter rodentium mouse model: studying pathogen and host contributions to infectious colitis. *Journal of Visualised Experiments*. 72:e50222.

Bian, X., Wu, W., Lang, L., Lv, L., Wang, Q., Li, Y., Ye, J., Fang, D., Wu, J., Jiang, W., Shi, D. & Li, L. (2019) Administration of Akkermansia muciniphila ameliorates dextran sulfate sodium-induced ulcerative colitis in mice. *Frontiers in Microbiology*. 10:2259.

Bode, L. (2012) Human milk oligosaccharides: every baby needs a sugar mummy. *Glycobiology*. 22(9):1147-1162.

Bravo, J.A., Julio-Pieper, M., Forsythe, P., Kunze, W., Dinan, T.G., Bienenstock, J. & Cryan, J.C. (2012) *Current Opinion in Pharmacology*. 12:667-672.

Bronowski, J. (1973) *The Ascent of Man*. Book Club Associates. London. UK.

Brugiroux, S., Beutler, M., Pfann, C., Garzetti, D., Ruscheweyh, H., Ring, D., Dieli, M., Herp, S., Lötscheer, Y., Hussain, S., Bunk, B., Pukall, R., Huson, D/., Munch, P., McHardy, A., McCoy, K., Macphereson, A., Loy, A., Clavel, T., Berry, D. & Stecher, B. (2017) Genome-guided design of a defined mouse microbiota that confers colonization resistance against *Salmonella enterica* serovar Typhimurium. *Nature Microbiology*. 2:16215.

Brumfield, K.D., Huq, A., Colewell, R.R., Olds, J.L., Leddy, M.B. (2020) Microbial resolution of whole genome and 16S rRNA amplicon metagenomic sequencing using publicly available NEON data. *PLOS ONE*. 15(2):e0228899.

Bunker, J.J. & Bendalec, A. (2018) IgA responses to microbiota. *Immunity.* 49:211-224.

Burke, K.E. & Lamont, J.T. (2014) Clostridium difficile infection: a worldwide disease. *Gut and Liver*. 8(1):1-6.

Byndloss, M.X., Pernitzsch, S.R. & Bäumler, A.J. (2018) Healthy hosts rule within: ecological forces shaping the gut microbiota. *Mucosal Immunology*. 11:1299-1305.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. & Madden, K. (2009) BLAST+: architecture and applications. *BMC Genomics*. 10:421.

Chakravorty, S., Helb, D., Burdy, M., Connell, N & Alland, D. (2007) A detailed analysis of 16S rRNA ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. *Journal of Microbiological Methods*. 69(2):330-339.

Chen, Y., Miao, Z., Yip, K., Cheng, Y., Cheng, Y., Liu, C., Li, L., Lin, C., Wang, J., Wu, D., Cheng, T. & Wang, J. (2020) Gut fecal microbiota transplant in a mouse model of orthopaedic rectal cancer. *Frontiers in Oncology*. 10:568012.

Cho, I., & Blaser, M.J. (2012) The human microbiome: at the interface of health and disease. *Nature Reviews: Genetics*. 13:260-270.

Chou, C., Aydilek, A., Seaggren, E., Maugel, T. (2008) Bacterially induced calicite precipitation via ureolysis*. Environmental Science.*

Clooney, A.G., Fouhy, F., Sleator, R., O'Driscoll, A., Stanton, C., Cotter, P. & Claesson, M.J. (2015) Comparing apples and oranges: next generation sequencing and its impact on microbiome analysis. *PLOS ONE*. 10.1371.

Cole, J.R., Wang, Q., Fish, J.A., Chai, B., McGarrell, M., Sun, Y., Brown, C.T., Porras-Alfaro, A., Kuske, C.R. & Tiedje, J.M. (2013) Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Research.* DOI 10.1093/nar/gkt1244.

Corfild, A. P. (2018) The interaction of the gut microbiota with mucus barrier in health and disease in human. *Microorganisms*. 6:0078.

Cowen, S.T. (1993) *Cowen & Steel's Manual for the identification of medical bacteria*. Cambridge University Press. Cambridge. UK.

Coyte, K.Z., Schluter, J. & Foster, K.R. (2015) The ecology of the microbiome: networks, competition, and stability. *Science.* 350(6261):663-666.

Coyte, K.Z., Rao, C., Rakoff-Nahoum, S. & Foster, K. (2021) Ecological rules for the assembly of microbiome communities. *PLOS Biology*. 10.1371.

Cullen, C.M., Aneja, K.K., Beyhan, S., Cho, C.E., Woloszynek, S., Convertino, M., McCoy, S.J., Zhang, Y., Anderson, M.Z., Alverez-Ponce, D., Smiirnova, E., Karstens, L., Dorrestein, P.C., Li, H., Gupta, A.S., Cheung, K., Powers, J.G., Davenport, E.R., Mizrahi-Ma, O., Michelini, K., Barreiro, L.B., Ober, C. & Gilad, Y. (2014) Seasonal variation in the human gut microbiome composition. *PLOS ONE.* 9(3):e90731.

Dash, S., Shakyawar, S. Shama, M. & Kaushik, S. (2019) Big data in healthcare: management, analysis, and future prospects. *Journal of Big Data.* 6:54. 10.1186.

Davenport, E.M., Sanders, J.G., Song, S.J., Amato, K., Clark, A. & Knight, R. (2017) The human microbiome in evolution. *BMC Biology*. 15:127.

Dennehy, J.J. (2014) What ecologists can tell virologists. *Annual Review of Microbiology*. 68:117-135.

Dethlefson, L. & Relamn, D.A. (2011) Incomplete recovery and individualized responses of the human distal gut microbiota to repeated antibiotic perturbation. *Proceedings of the National Academy of Sciences*. 108:4554-4561.

Dickson, R.P., Erb-Downward, J.R., Martinez, F. & Huffnagle, G.B. (2016) The microbiome and the respiratory tract. *Annual Review of Physiology*. 78:481-504.

Ding, L., Wendl, M.C., Koboldt, D.C. & Mardis, E.R. (2010) Analysis of next generation genomic data in cancer: accomplishments and challenges. *Human Molecular Genetics*. 19(2):188-196.

Durazzi, F., Sala, C., Casterllani, G., Manfreda, G., Remondini, D. & Cesare, A. de, (2021) Comparison between 16S rRNA and shotgun sequencing data for the taxonomic characterisation of the gut microbiota. *Scientific Reports*. 11:3030.

Edgar, R.C. (2017) Updating the 97% identity threshold for 16S ribosomal RNA OTUs. *bioRxiv*. 10.1101/192211.

Eaton, K.A., Dewhurst, F.E., Paster, B.J., Tzellas, N., Coleman, B.E., Paola, J. & Sherding, R. (1996) Prevalence and varieties of Helicobacter species in dogs from random sources and pet dogs: animal and public health implications. *Journal of Clinical Microbiology*. 34(12):3165-3170.

Eck, S. (2018) Challenges in data storage and data management in a clinical diagnostic setting. *Journal of Laboratory Medicine*. 42(2):219-224.

Eichele, D. D. & Kharbanda, K. K. (2017) Dextran sodium sulfate colitis murine model: An indispensable tool for advancing our understanding of inflammatory bowel diseases pathogenesis. *World Journal of Gastroenterology.* 23:6016-6029.

Envigo (2022) SCID mice: C.B-17/IcrHsd-Prkdcscid mutant mice (envigo.com) [Accessed 22nd March 2022]

Ericsson, A.C., Gagliardi, J., Bouhan, D., Spollen, W.G., Givan, S. & Franklin, C.L. (2017) The influence of caging, bedding, and diet on the composition of the microbiota in different regions of the mouse gut. *Nature: Scientific Reports*. 8:4065.

Escobar-Zepeda, A., Godoy-Lorano, E.E., Raggi, L., Segovia, L., Merino, E., Gutierrez-Rios, R., Juarez, K., LIcea-Navarro, A.F., Lopez, L.P. & Sanchez-Flores, A. (2018) Analysis of sequencing strategies and tools for taxonomic annotation: defining standards from progressive metagenomics. *Scientific Reports.* DOI 10.1038/s41598-018-30515-5.

Fabbo, C., Scalabrin, S., Morgante, M. & Giorgi, F.M. (2013) An extensive evaluation of read trimming in Illumina NGS data analysis. *PLOS One.* 8:12/e85024.

Ferrechhia, C.E. Jensen, K. & Andel, R. (2014) Intercage ammonia levels in statci and individually ventilated cages housing C57BL/6 mice on bedding substrates. *Journal of the American Association for Laboratory Animal Science*. 53(2):146-151.

Fiers, W., Contreras, F., Duerinck, F., Haegeman, G., Iserentant, D., Merregaert, J., Min Jou, W., Molemans, F., Raeymaekers, A., Van den Berghe, A., Volckaert, G. & Ysebaert, M. (1976) Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of replicase gene. *Nature*. 260:500-507.

Fischbach, M.A. (2018) Microbiome: focus on causation and mechanism. *Cell.* 174:785-790.

Finkbeiner, S.R., Allred, A.F., Tarr, P.I., Klein, E.J. & Kirkwood, C.D. (2008) Metagenomic analysis of human diarrhea: viral detection and discovery. *Public Library of Science: Pathogens*. 4(2):1-9.

Fler, L. van de, & Clevers, H. (2009) Stem cells, self-renewal, and differentiation in the intestinal epithelium. *Annual Review of Physiology*. 71:241-260.

Flint, H.J., Scott, K.P., Duncan, S.H., Louis, P & Forano, E. (2012) Microbial degradation of complex carbohydrates in the gut. *Gut Microbes*. 3(4):289-306.

Fouladi, F., Bailey, M., Patterson, W., Sioda, M., Blakley, I., Foder, A., Jones, R., Chen, Z., Kim, J., Lurmann, F., Martino, C., Knight, R., Gilliland, F. & Alderate, T. (2020) Air pollution exposure is associated with the gut microbiome as revealed by shotgun metagenomics. *Environmental International*. 138:105604.

Fox, J.G., Gorelick, P.L., Kulberg, M.C., Ge, Z., Dewhirst, F.E. & Ward, J.M. (1999) A novel urease-negative Helicobacter species associated with colitis and typhlitis in IL-10-deficient mice. *Infection & Immunity.* 67(4):1757-1762.

Franklin, C.L. & Ericsson, A.C. (2018) Microbiota and reproducability of rodents. *Laboratory Animal (New York).* 46(4):114-122.

Garza, D.R., Verk, M. van, Huynen, M. & Dutith, B.E. (2018) Towards predicting the environmental metabolome from metagenomics with a mechanistic tool. *Nature Microbiology*. 3:456-460.

Gerbe, F. & Jay, P. (2016) Intestinal tuft cells: epithelial sentinels linking luminal cues to the immune system. *Mucosal Immunology*. 9:1353-1359.

Gilbert, J., Blaser, M.J., Caporaso, J.G., Jansson, J., Lynch, S.V. & Knight, R. (2018) Current understanding of the human microbiome. *Nature Medicine*. 24(4):392-400.

Glaxo Wellcome (1998) *Method of nucleic acid amplification*. WO1998044151.

Gloor, G., Macklaim, J., Pawlowowsky, V. & Egozcue, J. (2017) Microbiome databases are compositional: and this is not optional. *Frontiers in Microbiology*. 8.2224.

Grice, E.A. & Segre, J.A. (2012) The human microbiome: our second genome. *Annual Review of Genomics & Human Genetics*. 13:151-170.

Gregory, A.L., Pensinger, D.A. & Hryckowian, A.J. (2021) A short chain fatty acid-centric view of Clostridioides difficile pathogenesis. *PLOS Pathogens.* 17(10):e1009959.

Guarner, F. & Malagelada, J. (2003) Role of bacteria in experimental colitis. *Best Practice & Research Clinical Gastroenterology*. 17(5):793-804.

Guthrie, L. & Kelly, L. (2019) Bringing microbiome-drug interaction research into the clinic. *EBioMedicine*. 44:708-715.

Haller, D. (eds.) (2018) *The Gut Microbiome in Health and Disease*. Springer. Cham. Switzerland.

Handley, S.A., Thackery, L.B., Zhao, G., Presti, R., Miller, A.D., Droit, L., Abbink, P., Maxfield, L.F., Kambal, A., Duan, E., Stanley, K., Kramer, J., Macri, S.C., Permar, S.R., Schmitz, J.E., Mansfield, K., Brenchley, J.M., Veazey, R.S., Stappenbeck, T.S., Wang, D., Barouch, D.H. & Virgin, H.W. (2012) Pathogenic Simian Immunodeficiency Virus is associated with expansion of enteric virome. *Cell*. 151:253-266.

Harrison, P.W., Ahamed, A., Aslam, R., Alako, B.T.F., Burgin, J., Buso, N., Courtot, M., Fan, J., Gupta, D., Haseeb, M., Holt, S., Ibrahim, T., Ivanov, E., Jayathilaka, S., Balavenkataraman, V., Kumar, M., Lopez, R., Kay, S., Leinonen, R., Liu, X., O'Cathail, C., Pakseresht, A., Park, Y., Pesant, S., Rahman, N., Rajan, J., Sokolov, A., Vijayaraja, S., Waheed, Z., Zyoud, A., Burdett, T., Cochrane, G.. (2021) The European Nucleotide Archive in 2020. *Nucleic Acids Research.* 49(D1):D82-D85.

Hartenstein, V. & Martinez, P. (2019) Structure, development, and evolution of the digestive system. *Cell Tissue Research*. 377(3):289-292.

Hendry, A. P. (2020) *Eco-evolutionary Dynamics*. Princeton University Press. Princeton. USA.

Her Majesty's Home Office (1986) The Animal (Scientific Procedures) Act 1986: Chapter 14. Available at: http://www.legislation.gov.uk/ukpga/1986/14/contents [Accessed 16th September 2021]

Her Majesty's Home Office (2022) Home Office Statistics of Scientific Procedures on Living Animals. Great Britain 2021. Available at: https://www.gov.uk/government/statistics/statistics-of-scientific-procedures-on -living-animals-great-britain-2021 [Accessed: 6th August 2022]

Hickman, J.M. & Davis, I. (2005) Transgenic mice. *Paediatric Respiratory Review*. 7(1):49-53.

Hill, C. (2018) Is it time to stop measuring, and put the 'ology' back into microbiome research? Nature Research Microbiology Community. Available at: https://naturemicrobiologycommunity.nature.com/users/105629-colin-hill. [Accessed 21st March 2020]

Hill, D.A. & Artis, D. (2010) Intestinal bacteria and the regulation of immune cell homeostasis. *Annual Review Immunology*. 28:623-667.

Hiippala, K., Kainulainen, V., Kalliomäki, M., Arkkila, P. & Satokari, R. (2016) Mucosal prevalence and interactions with the epithelium indicate commensalism of Sutterella spp. *Frontiers in Microbiology*. 7:1706.

Hugenholtz, F. & de Vod, W. (2017) Mouse models for human intestinal microbiota research: a critical evaluation. Cellular and Molecular Life Sciences. 75:149-160.

Ilott, N., Bollrath, J., Danne, C., Schiering, C., Shale, M., Adelmann, K., Krausgruber, T., Heger, A., Sims, D. & Powrie, F. (2016). Defining the microbial transcriptional response to colitis through integrated host and microbiome profiling. *The ISME Journal.* 10:2389-2404.

Illumina (2016) 16S Metagenomic sequencing library preparation. https://support.illumina.com/documents/documentation/chemistry_documentation/16s/16s-metagen omic-library-prep-guide-15044223-b.pdf [Accessed: 23rd May 2016]

International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (2009) Available at: http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Multi disciplinary/M3_R2/Step4/M3_R2__Guideline.pdf [Accessed: 23rd May 2016]

Izard, J. & Rivera, M. (2015) *Metagenomics for Microbiology*. Academic Press. Elsevier. London.

Johnson, J.S., Spakowiicz, D.J., Hong, B., Peterson, L.M., Demokoicz, P., Chen, L., Leopold, S.R., Hanson, B.M., Agresta, H.O., Gerstein, M., Sodergren, E. & Weinstock, G.M. (2019) Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. *Nature Communications.* 10:5029.

Joval, J., Patterson, J., Wang, W., Hotte, N., O'Keefe, S., Mitchel, T., Perry, T., Kao, D., Mason, A.L., Madsen, K.I. & Wong, G. (2016) Characterisation of the gut microbiome using 16S rRNA or shotgun metagenomics. *Frontiers in Microbiology.* 7(459):1-16.

Karl, J.P, Meydani, M., Barnett, J.B., Venegas, S.M., Barger, K., Fu, X., Goldin, B., Kane, A., Rasmussen, H., Vangay, P., Knights, D., Jonnalagadda, S.S., Saltzman, E., Roberts, S.B., Meydani, S.N. & Booth, S.L. (2017) Fecal concentrations of bacterially derived vitamin k forms are associated with gut microbiota composition but not plasma or fecal cytokine concentrations in healthy adults. *American Journal of Clinical Nutrition*. 106:1052-1061.

Karst, S.M., Wobus, C.E., ,Lay, M., Davidson, J. & Virgin, H.W. (2003) STAT1-dependant innate immunity to Norwalk-like virus. *Science*. 299:1575-1578.

Kaser, A., Zeissig, S. & Blumberg, R.S. (2010) Inflammatory Bowel Disease. *Annual Review of Immunology*. 28:573-621.

Kasman, L.M. (2005) Barriers to coliphage infection of commensal intestinal microbiota of laboratory mice. *Virology Journal*. 2:34-41.

Kau, A.L., Ahern, P.P., Griffin, N.W., Goodman, A.L. & Gordon, J.I. (2011) Human nutrition, the gut microbiome, and the immune system. *Nature*. 15(7351):327-336.

Khalili, H., Chan, S.S.M., Lochhead, P., Ananthakrishnan, A.N., Hart, A.R. & Chan, A.T. (2018) The role of diet in the aetiopathogenesis of inflammatory bowel disease. *Nature Review of Gastroenterology & Hepatology*. 15:525-535.

Kieser, S., Zdobnov, E. & Trajkovski, M. (2022) Comprehensive mouse microbiota genome catalog reveals major differences to its human counterpart. *PLOS Computational Biology*. 10.1371.

Kim, J., Shajib, S., Manocha, M. & Khan, W. (2012) Investigating intestinal inflammation in DSS-induced model of IBD. *Journal of Visualized Experiments.* 60:e3678.

Kim, N., Kim, C., Yang, S., Park, D., Ha, S. & Lee,I. (2021) MRGM: a mouse reference gut microbiome reveals a large functional disparity for gut bacteria of the same genus between mice and humans. *bioRxiv*. doi.org/10. 1101/2021.10.24.465599.

Kiu, R., Treveil, A., Harnisch, L.C., Caim, S., Leclaire, C., van Sinderin, D., Korcsmaros, T. & Hall, L.J. (2020) Bifidobacterium breve UCC2003 induces a distinct global transcriptome program in neonatal murine intestinal epithelial cells. *iScience*. 23:101336.

Kjellev, S., Lundsgaard, D., Poulsen, S.S. & Markholst, H. (2006) Reconstruction of SCID mice with CD4+CD25- T cells leads to rapid colitis: A model for pharmacologic testing. *International Immunopharmacology*. 6:1341-1354.

Klappenbach, J.A., Saxman, P.R., Cole, J.R. & Schmidt, T.M. (2001) rrndb: the ribosomal RNA operon copy number database. *Nucleic Acids Research*. 29(1):181-184.

Klindworth, A., Pruesse, E., Schweer, T., Peplies, J., Quast, C., Horn, M. & Glochner, F.O. (2013) Evaluation of general 16S rRNA ribosomal RNA gene PCR primers for classical next generation sequencing based diversity studies. *Nucleic Acids Research*. 41(1):e1.

Kuczynski, J., Stombaugh, J., Walters, W.A., González, A., Caporaso, J.G. & Knight, R. (2011) Using QIIME to analyse 16S rRNA gene sequences from microbial communities. *Current Protocols in Bioinformatics*. doi.org/10.1002/0471250953.

Kudesia, G. & Wreghitt, T. (2009) *Clinical & Diagnostioc Virology*. Cambridge University Press. Cambridge. UK.

Kulkarni, P. & Frommolt, P. (2017) Challenges in the setup of scale next generation sequencing analysis tools. *Computational & Structural Biotechnology Journal*. 15:471-477.

Lagier, J.C., Dubourg, G., Million, M., Cadoret, F., Bilen, M., Fenollar, F., Levasseur, A., Rolain, J.M., Fournier, P.E. & Raoult, D. (2018) Culturing the human microbiota and culturomics. *Nature Reviews: Microbiology*. 16:540-550.

Lagkouvardos, I., Pukall, R., Abt, B., Foesel, B.U., Meier-Kolthoff, J.P., Kumar, N., Bresciani, A., Martínez, I., Just, S., Ziegler, C., Brugiroux, S., Garzetti, D., Wenning, M., Bui, T.P., Wang, J., Hugenholtz, F., Plugge, C.M., Peterson, D.A., Hornef, M.W., Baines, J.F., Smidt, H., Walter, J., Kristiansen, K., Nielsen, H.B., Haller, D., Overmann, J., Stecher, B. & Clavel, T. (2016) The Mouse Intestinal Bacterial Collection (miBC) provides host-specific insight into cultured diversity and functional potential of the gut microbiota. *Nature Microbiology*. 1(10):16131.

Lagkouvardos, I., Lesker, T.R., Hitch, T.C.A., Gálvezm E.J.C., Smit, N., Neuhaus, K., Wang, J., Baines, J.F., Abt, B., Stecher, B., Overmann, J., Strowig, T. & Clavel, T. (2019) Sequence and cultivation study of Muribaculaceae reveals novel species, host preference, and functional potential of this yet undescribed family. *Microbiome.* 7(1):28. 10.1186

Langille, M., Meehan, C.J., Koeng, J.E., Dhanani, A.S., Rose, R.A., Howlett, S.E. & Beiko, R. (2014) Microbial shifts in the aging mouse gut. *Microbiome*. 2:50. doi.org/10.1186/s40168-014-0050-9.

LeBlanc, J.G., Chain, F., Martin, R., Bermudez-Humaran, L., Courau, S. & Langella, P. (2017) Beneficial effects on host energy metabolism of short-chain fatty acids and vitamins produced by commensal and probiotic bacteria. *Microbial Cell Factories*. 16:79.

Lepp, P.W., Brinig, M.M., Ouverney, C.C., Pal, K., Armitage, G.C. & Relman, D.A. (2004) Methanogenic Archaea and human periodontal disease. *Proceedings of the National Academy of Science.* 101(16):6176-6181.

Letarov, A. & Kulikov, E. (2009) The bacteriophages in human- and animal body associated microbial communities. *Journal of Applied Microbiology*. 107:1-13.y, R. E., Backhed, F., Turnbaugh, P., Lozupone, C.A., Knight, R.D. & Gorden, J.I. (2005) Obesity alters gut microbial ecology. *PNAS.* 102:11070-11075Le.

Levin, J.Z., Yassour, M., Adiconius, X., Nusbaum, C., Thompson, D.A., Friedman, N., Gnirke, A. & Regev, A. (2010) Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nature Methods*. 7(9):709-715.

Ley, R.E., Bäckhed, F., Turnbaugh, P., Lozupone, C.A., Knight, R.D. & Gordon, J.I. (2005) Obesity alters gut microbial ecology. *Proceedings of the National Academy of Sciences of the United States of America.* 102(31):11070-11075.

Ley, R.E., Peterson, D.A. & Gordon, J.I. (2006) Ecological and evolutionary forces shaping microbial diversity in the human intestine. *Cell.* 124:837-848.

Levy, M., Kolodziejczyk, A.A., Thaiss, C.A. & Elinav, E. (2017) Dysbiosis and the immune system. *Nature Reviews Immunology*. 17:219-232.

Liu, C., Zhou, N., Du, M.X., Sun, Y.T., Wang, K., Wang, Y.J., Li, D.H., Yu, H.Y., Song, Y., Bai, B.B., Xin, Y., Wu, L., Jiang, C.Y., Feng, J., Xiang, H., Zhou, Y., Ma, J., Wang, J., Liu, H.W. & Liu, SJ. (2020) The Mouse Gut Microbial Biobank expands the coverage of cultured bacteria. *Nature Communications.* 11:79.10.1038.

Ma, B.W., Bokulich, N.A., Castillo, P.A., Kananurak, A., Underwood, M.A., Mills, D.A. & Bevins, C.L. (2012) Routine habitat change: a source of unrecognized transient alteration in intestinal microbiota in laboratory mice. *PLOS ONE*. 7(10):e47416.

Magnusdottier, S., Ravcheev, D., Crecy-Lagard, V. & Theile, I. (2015) Systematic genome assessment of B-vitamin biosynthesis suggests co-operation among gut microbes. *Frontiers in Genetics*. 6(148):1-16.

Mähler, M., Berard, M., Feinstein, R., Gallagher, A., Illgen-Wilke, B., Pritchett-Corning, K. & Raspa, M. (2014) FELASA recommendations for the health monitoring of mouse, rat, guinea pig and rabbit colonies in breeding and experimental units. *Laboratory Animals*. Published online February 2014 doi: 10.1177/0023677213516312

Marchesi, J.R. & Ravel, J. (2015) The vocabulary of microbiome research: a propoasl. *Microbiome*. 3(31). DOI 10.1186/s40168-015-0094-5.

Marchesi, J.R., Adams, D., Fava, F., Hemres, G.D., Hirschfield, G.M., Hold, G., Quaraishi, M.N., Kinross, J., Smiidt, H., Touhy, K.M., Thomas, L.V., Zoetendal, E.G. & Hart, A. (2016) The gut microbiota and host health: a clinical frontier. *Gut*. 65(2):330-338.

Marshall, C.R. (2006) Explaining the Cambrian explosion. *Annual Review of Earth & Planetary Sciences*. 34:355-384.

Marston, D.A., McElhinney, L.M., Ellis, R.J., Horton, D.L., Wise, E., Leech, S.L., David, D., Lamballerie, X. & Fooks, A.R. (2013) Next generation sequencing of viral RNA genomes. *BMC Genomics*. 14:444-456.

Martinez, K.B., Leone, V. & Chang, E.B. (2017) Western diets, guy dysbiosis, and metabolic diseases: are they linked? *Gut Microbes*. 8(2):130-142.

Maruo, T., Sakamoto, M., Ito, C., Toda, T. & Benno, Y. (2008) Adlercreutzia equolifaciens gen. nov., sp. nov., an equol-producing bacterium isolated from human faeces, and emended description of the genus Eggerthella. *International Journal of Systematic & Evolutionary Microbiology.* 58:1221-1227.

McDonald, D., Price, M.N., Goodrich, J., Nawrocki, E.P., de Santis, T.Z., Probst, A., Anfewrson, G.L., Knight, R. & Hugenholtz, P. (2012) An improved Greengenes taxonomy with explicit ranks for ecological analysis of bacteria and archaea. *The ISME Journal*. 6:610-1618.

Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E.M., Kubal, M., Paczian, T., Rodriquez, A., Stevens, R., Wilke, A., Wilkening, J. & Edwards, R.A. (2008) The metagenomic RAST server – a public resource for the automated phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*. doi10.1186/1471-2105-9-386.

Mizoguchi, A. (2012) Animal models of inflammatory bowel disease. *Progressive Molecular Biology Translational Science* 105:263-320.

Modi, S.R., Collins, J.J. & Ralman, D.A. (2014) Antibiotics and the gut microbiota. *Journal of Clinical Investigations.* 124(10):4212-4218.

Montonye, D.R., Ericsson, A.C., Bushi, S.B., Lutz, C., Wardwell, K. & Franklin, C.L. (2018) Acclimatisation and institutionalization of the mouse microbiota following transportation. *Frontiers in Microbiology*. 9(1085):1-13.

Moschen, A. R., Tilg, H. & Raine, T. (2019) IL-12, IL-23, and IL-17 in IBD: immunobiology and therapeutic targeting. *Nature Review of Gastroenterology & Hepatology*. 16:185-196.

Mu, Q., Travella, V.J. & Luo, X.M. (2018) Role of Lactobacillus reuterin in human health and diseases. *Frontiers in Microbiology*. doi: 10.3389/fmicb.2018.00757.

Mullis, K., Faloona, F., Scharf, S., Horn, G. & Erlich, H. (1986) Specific enzymic amplification of DNA in vitro: the polymerase chain reaction. *Cold Spring Harbour Symposia on Quantitative Biology.* 51:263-273.

Munyaka, P.M., Rabbi, M.F., Khafipour, E. & Ghia, J. (2016) Acute dextran sulfate sodium (DSS) induced colitis promotes gut microbial dysbiosis in mice. *Journal of Basic Microbiology*. 56(9):986-998.

Mysara, M., Vandamme, P., Props, R., Kerckhof, F., Leys, N., Boon, N., Raes, J. & Monsieurs, P. (2017) Reconciliation between operational taxonomic units and species boundaries. *FEMS Microbiology Ecology*. 93:fix029.

Nguyen, T.L.A., Viera-Silva, S., Liston, A. & Raes, J. (2015) How informative is the mouse for human gut microbiota research? *Disease Model Mechanisms*. 8(1):1-16.

Nicklas, W. & Seidel, K. (2019) Expert information from the working group on hygiene: harmonisation of health monitoring reports. *Laboratory Animals.* 53(2):208-209.

National Institute for Health - Human Microbiome Working Group. (2009) The NIH Human Microbiome Project. *Genome Research*. 19:2317-2323.

Nishiyama, K., Sugiyama, M. & Mukai, T. (2016) Adhesion properties of lactic acid bacteria on intestinal mucin. *Microorganisms*. 4:34.

Norman, J.M., Handley, S.A. & Virgin, H.W. (2014) Kingdom-agnostic metagenomics and their importance of complete characterisation of enteric microbial communities. *Gastroenterology.* 146:1459-1469.

O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., Farrell, C.M., Goldfarb, T., Gupta, T., Haft, D., Hatcher, E., Hlavina, W., Joardar, V.S., Kodali, V.K., Li, W., Maglott, D., Masterson, P., McGarvey, K.M., Murphy, M.R., O'Neill, K., Pujar, S., Rangwala, S.H., Rausch, D., Riddick, L.D., Schoch, C., Shkeda, A., Storz, S.S., Sun, H., Thibaud-Nissen, F., Tolstoy, I., Tully, R.E., Vatsan, A.R., Wallin, C., Webb, D., Wu, W., Landrum, M.J., Kimchi, A., Tatusova, T., DiCuccio, M., Kitts, P., Murphy, T.D., Pruitt, K.D. (2015) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*. D733-D745.

Oh, J., Freeman, A.F., NISC Comparative Sequencing Program, Park, M., Sokolic, R., Candotti, Holland, S.M., Serge, J.A. & Kong, H.H. (2013) The altered landscape of the human skin microbiome in patients with primary immunodeficiencies. *Genome Research.* 23:2103-2114.

Ormerod, K.L., Wood, D.L., Lacher, N., Gelatly, S.L., Daly, J.N., Parsons, J.D., Dal'Molin, C.G.O., Palfreyman, R.W., Nielson, L.K., Cooper, M.A., Morrison, M., Hansbro, PM. & Hugenholtz, P. (2016) Genomic characterisation of the uncultured Bacteriodales family S24-7 inhabiting the guts of homothermic animals. *Microbiome.* 4:36.

Osawa, R., Fujisawa, T. & Pukall, R. (2006) Lactobacillus apodemi sp. Nov., a tannase-producing species isolated from wild mouse faeces. *International Journal of Systemic and Evolutionary Microbiology*. 56(7):1693-1696.

Ost, K.S. & Round, J.L. (2018) Communication between the microbiota the mammalian immunity. *Annual Review of Microbiology*. 72:399-422.

Ostenin, D.V., Bao, J., Koboziev, I., Grey, L., Robinson-Jackson, S.A., Kosloski-Davidson, M., Price, V.H. & Grisham, M.B. (2008) T-cell transfer model of chronic colitis: concepts, considerations, and tricks of the trade. *Gastrointestinal & Liver Physiology*. 296:135-146.

Pabst, O., Cerovic, V. & Hornef, M. (2016) Secretory IgA in the coordination of establishment and maintenance of the microbiota. *Trends in Immunology*. 37(5):287-296.

Paez-Espinio, D., Cghen, I.A., Palaniappan, K., Ratner, A., Chu, K., Szeto, E., Pillay, M., Huang, J., Markowitz, J. & Nielsen, T. (2017) IMG/VR: a database of cultured and uncultured DNA viruses and retroviruses. *Nucleic Acids Research*. 45(D1):457-465.

Pandiyan, P., Bhaskaren, N., Zou, M., Schneider, E., Jayaraman, S. & Huehn (2019) Microbiome dependant regulation of TREGS and T_h17 cells in mucosa. *Frontiers in Immunology*. 10:00426.

Papageorgiou, L., Eleni, P., Raftopoulou, S., Mantaiou, M., Megalooikonomou, V. & Vlachakis, D. (2018) Genomic big data hitting the storage bottleneck. *EMBnet.* 24:e910.

Park, S. & Won, S. (2018) Evaluation of 16S rRNA databases for taxonomic assignments using a mock community. *Genomics & Informatics*. 16(4):e24.

Pasoli, E., Asnicar, F., Manara, S., Zolfo, M., Karcher, N., Armanini, F., Beghini, F., Manghi, P., Tett, A., Ghensi, P., Collado, M.C., Rice, B.L., DuLong, C., Morgan, X.C., Golden, C.D., Quince, C., Huttenhower, C. & Segata, N. (2019) Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. *Cell*. 176:649-662.

Peng, Z., Jin, D., Kim, H., Stratton, C.W., Wu, B., Tang, Y. & Su, X. (2017) Update on antimicrobial resistance in Clostridium difficile: resistance mechanisms and antimicrobial susceptibility testing. *Journal of Clinical Microbiology*. 55(7):1998-2008.

Peterson, D.A., McNulty, N.P., Guruge, J.L. & Gordon, J.I. (2007) IgA response to symbiotic bacteria as a mediator of gut homeostasis. *Cell Host & Microbe.* 2:328-339.

Peterson, D., Bonham, K., Rowland, S., Pattanayak, C., RESONANCE Consortium & Klepac-Ceraj, V. (2021) Comparative analysis of 16S rRNA gene and metagenome sequencing in paediatric gut microbiomes. *Frontiers in Microbiology*. 12:670336.

Pollock, J., Glendinning, L. Wisedchanwet, T. & Watson, M. (2018) The madness of microbiome: attempting to find consensus "best practice" for 16S rRNA microbiome studies. *Applied & Environmental Microbiology*. 84(7):1-12.

Porter, N. & Martrens, E. (2017) The critical roles of polysaccharides in gut microbial ecology and physiology. *Annual Review of Microbiology*. 71:349-369.

Post, S., Jabber, K., Birchenough, G., Arike, L., Akhter, N., Sjovall, H., Johansson, M. & Hansson, G. (2019) Structural weakening of the colonic mucus barrier is an early event in ulcerative colitis. *Gut.* 68:2142-2151.

Prattis, S. & Jurjus, A. (2015) Spontaneous and transgenic rodent models of inflammatory bowel disease. *Laboratory Animal Research*. 31(2):47-68.

Pritchett-Corning, K., Cosentino, J. & Clifford, C.B. (2009) Contemporary prevalence of infectious agents in laboratory mice and rats. *Laboratory Animals.* 43:165-173.

Pruitt, K.D., Tatusoava, T. & Maglott, D.R. (2006) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts, and proteins. *Nucleic Acids Research*. DOI 10.1093/nar/gkl842.

Qiagen-A. DNeasy PowerSoil Pro Kit Handbook – (qiagen.com) [Accessed: 16th September 2021]

Qiagen-B. AllPrep PowerViral DNA/RNA Kit handbook (qiagen.com) [Accessed: 16th September, 2021]

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J. & Glockner, O. (2012) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research.* DOI 10.1093/nar/gks1219.

Rausch, P., Ruhlermann, M., Hermes, B., Doms, S., Dagen, T., Dierking, K., Domin, H. (2019) Comparative analysis of amplicon and metagenomic sequencing methods reveals key features in the evolution of animal metaorganisms. *Microbiome*. 7:133-152.

Rawls, J.F., Mahowald, M.A., Ley, R.E. & Gordon, J.I. (2006) Reciprocal gut microbiota transplants from Zebrafish and mice to germ-free recipients reveal host habitat selection. *Cell.* 127(2):423-433.

Rieder, F. & Fiocchi, C. (2008) Intestinal fibrosis in inflammatory bowel disease: current knowledge and future perspectives. *Journal of Crohn's Disease*. 2:279-290.

Robasky, K., Lewis, N.E. & Church, G.M. (2014) The role of replicates for error mitigation in next-generation sequencing. *Nature Reviews Genetics.* 15:56-62.

Robertson, C.E., Harris, J.K., Spear, J.R. & Pace, N.R. (2005) Phylogenetic diversity and ecology of environmental Archaea. *Current Opinion in Microbiology*. 8:638-642.

Rosshart, S., Vassallo, B., Angeletti, D., Hutchinson, D., Morgan, A., Takeda, K., Hickman, H., McCulloch, J., Badger, J., Ajami, N., Trincheri, G., Villena, F. de, Yewdall, J. & Rehermann, B. (2017) Wild mouse gut microbiota promotes host fitness and improves disease resistance. *Cell.* 171(5):1015-1028.

Sanger, F., Air, G.M., Barrell, B.G., Brown, N.L.,Coulson, A.R., Fiddes, J.C., Hutchinson III, C.A., Slocombe, P.M. & Smith, M. (1977) Nucleotide sequence of bacteriophage ØX174 DNA. *Nature.* 265:687-695.

Santaolalla, R. & Abreu, M.T. (2012) Innate immunity in the small intestine. *Current Opinions in Gastroenterology*. 28(2):124-129.

Scharschimidt, T.C., Vasquez, K.S., Truong, H., Gearty, S.V., Pauli, M.L., Nosbaum, A., Gratz, I.K., Otto, M., Moon, J.J., Liease, J., Abbas, A.K. Fischbach. M.A. (2015) A wave of regulatory T cells into neonatal skin mediates tolerance to commensal microbes. *Immunity*. 43:1011-1021.

Schloss, P.D., Jenior, M.L., Koumpouras, C.C., Westcott, S.L. & Highlander, S.K. (2016) Sequencing 16S rRNA gene fragments using the PacBio SMRT DNA sequencing system*. PeerJ.* 4e1869.

Shannon, C.E. & Weaver, W.W. (1963) *The mathematical theory of communications*. University of Illinois Press. Urbana. USA.

Sharpton, T.J. (2014) An introduction to the analysis of shotgun metagenomic data. *Frontiers in Plant Science*. 5(209):1-14.

Shin, J., Lee, S., Go, M., Lee, S., Kim, S., Lee, C. & Cho, B. (2016) Analysis of the mouse gut microbiome using full length 16S rRNA amplicon sequencing. *Nature Scientific Reports*. 6:29681.

Shkoporov, A.N. & Hill, C. (2019) Bacteriophages of the human gut: the “known unknown” of the microbiome. *Cell & Host Microbe*. 25:195-209.

Smith, E.C. (2017) The no-so-infinite malleability of RNA viruses: viral and cellular determinants of RNA virus mutation rates. *PLOS Pathogens* 13(4):e1006254.

Sousa, T., Paterson, R., Moore, V., Carlsson, A., Abrahamsson, B. & Basit, A.W. (2008) The gastrointestinal microbiota as a site for the biotransformation of drugs. *International Journal of Pharmaceutics.* 363:1-25.

Souza, H. S. de & Fiocchi, C. (2016) Immunopathogenesis of IBD: current state of the art. *Nature Review of Gastroenterology & Hepatology*. 13:13-27.

Spanogiannopoulos, P., Bess, E.N., Carmody, R.N. & Turnbaugh, P.J. (2016) The microbial pharmacists within us: a metagenomic view of xenobiotic metabolism. *Nature Reviews: Microbiology*. 14:273-287.

Spor, A., Koren, O. & Ley, R.E. (2011) Unravelling the effects of the environment and host genotype on the gut microbiome. *Nature Reviews: Microbiology*. 9:279-290.

Tanjo, T., Kawai, Y., Tokunaga, K., Ogasawara, O. & Nagasaki, M. (2021) Practical guide for managing scale human genome data in research. *Journal of Human Genetics*. 66:39-52.

Tessler, M., Neumann, J.S., Afshinnekoo, E., Pineda, M., Hersch, R., Velho, L.F.M., Segovia, B.T., Lansac-Toha, F., Lemke, M., Mason, C.E. & Brugler, M.R. (2017) -scale differences in microbial biodiversity between n16S rRNA amplicon and shotgun sequencing. *Scientific Reports*. doi10.1038/s41598-017-06665-3.

Theinpont, D., Rochette, F. & Vanparijs, O.F.J. (1986) *Diagnosis helminthiasis by coprological examination*. Janseen Research Foundation. Beerse. Belgium.

Thompson, L.R., Sanders, J.G., McDonald, D., Amir, A., Ladau, J., Locey, K.J., Prill, R.J., Tripathi, A., Gibbons, S.M., Ackermann, G., Navas-Molina, J.A., Janssen, S., Kopylova, E., Vázquez-Baeza, Y., González, A., Morton, J.T., Mirarab,, S., Zech, Xu Z., Jiang, L., Haroon, M.F., Kanbar, J., Zhu, Q., Jin

Song, S., Kosciolek, T., Bokulich, N.A., Lefler, J., Brislawn, C.J., Humphrey, G., Owens, S.M., Hampton-Marcell, J., Berg-Lyons, D., McKenzie, V., Fierer, N., Fuhrman, J.A., Clauset, A., Stevens, R.L., Shade, A., Pollard, K.S., Goodwin, K.D., Jansson, J.K., Gilbert, J.A. & Knight, R.; (2017) Earth Microbiome Project Consortium. A communal catalogue reveals Earth's multiscale microbial diversity. *Nature.* 551(7681):457-463.

Tindemans, I., Joose, M.E., & Samsom, J.N. (2020) Dissecting the heterogeneity in T-cell mediated inflammation in IBD. *Cells*. 9010110.

Toga, A. & Dinov, I. (2015) Sharing big biomedical data. *Journal of Big Data*. 2:7. 10.1186.

Tortorella, D., Gewurz, B.E., Furman, M.H., Shust, D.J. & Ploegh, H.L. (2000) Viral subversion of the immune system. *Annual Review of Immunology*. 18:861-926.

Troncoso, L.L., Biancardi, A.L., Moreles, V. de & Zaltman, C. (2017) Ophthalmic manifestations in patients with inflammatory bowel disease: a review. *World of Gastroenterology*. 28(32):5836-5848.

Uniken, W. T., Voskuil, M. D., Dijkstra, G., Weersma, R. K. & Festen, E. A. (2017) The genetic background of inflammatory bowel disease: from correlation to causality. *Journal of Pathology*. 241:146-158.

Urwin, R. & Maiden, M.C.J. (2003) Multi-locus sequence typing: a tool for global epidemiology. *Trends in Microbiology*. 11(10):479-487.

Verhoeven, K., Simonsen, K. & McIntyre, L. (2005) Implementing the false discovery rate control: increasing your power. *Oikos*. 108:643-647.

Vuong, H.E., Yano, J.M., Fung, T.C. & Hsiao, E.Y. (2017) The microbiome and host behaviour. *Annual Review of Neuroscience*. 40:21-49.

Wade, W.G. (2012) The oral microbiome in health and disease. *Pharmaceutical Research.* 69(1):137-143.

Wagner, B., Grunwald, G.., Zerbe, G.., Mikulich-Gilbertson, S., Robertson, C., Zemanick, E. & Harris, J. (2018) On the use of diversity measures in longitudinal sequencing studies of microbial communities. *Frontiers in Microbiology.* 9:1037.

Wallace, J.G., Potts, R., Szamosi, J., Surette, M. & Sloboda, D. (2018) The murine female intestinal microbiota does not shift throughout the oestrous cycle. *PLOS ONE*. 13(7):e0200729.

Walter, J. & Ley, R. (2011) The human gut microbiome: ecology and recent evolutionary changes. *Annual Review of Microbiology*. 65:411-429.

Wang, S., Sun, B., Tu, Jing, T. & Lu, Z. (2016) Improving the microbial community reconstruction at the genus level by multiple 16S rRNA regions. *Journal of Theoretical Biology*. 7:398.1-8.

Waring, M.J., Arrowsmith, J., Leach, A.R., Leeson, P.D., Mandrell, S., Owen, R.M., Pairaudeau, G., Pennie, W.D., Pickett, S.D., Wang, J., Wallace, O. & Weir, A. (2016) An analysis of the attrition of drug candidates from four major pharmaceutical companies. *Nature Review of Drug Discovery.* 14(7):475-485.

Warner, B.B. (2018) The contribution of the gut microbiome to neurodevelopment and neuropsychiatric disorders. *Paediatric Research*. 85:216-224.

Washington I.M. & Payton, M.E. (2016) Ammonia levels and urine-spot characteristics as cage-change indicators for high-density individually ventilated mouse cages. *Journal of the American Association for Laboratory Animal Science*. 55(3):260-267.

Watts, M.N., Leskova, W., Carter, P.R., Zhang, S., Davidson, M., Grisham, M.B. & Harris, N.R. (2013) Ocular dysfunction in a mouse model of chronic gut inflammation. *Inflammatory Bowel Disease*. 19(10):2091-2097.

Wesemann, D.R. & Nagler, C.R. (2016) The microbiome, timing, and barrier function in the context of allergic disease. *Immunity.* 44:728-738.

Wirbel, J., Pyl, P.T., Kartel, E., Zych, K., Kashani, A., Milanase, A., Fleck, J.S., Voigt, A.Y., Palleja, A., Ponnudurai, R., Sunagawa, S., Coelho, L.P., Schrotz-King, P., Vogtmann, E., & Habermann, N. (2019) Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. *Nature Medicine*. 25:679-689.

Woese, C.R. & Fox, G.E. (1977) Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proceedings of the National Academy of Sciences*. 74(11):5088-5090.

Wu, F., Zhao, S., Yu, B., Chen, Y., Wang, W., Song, Z., Hu, Y., Tao, Z., Tian, J., Pai, Y., Yuan, M., Zang, Y., Dai, F., Liu, Y., Wang., Q., Zheng, J., Xu, L., Holmes., E.C. & Zhang, Y. (2020) A new coronavirus associated with human respiratory disease in China. *Nature*. 579:265-269.

Xiao, L., Feng, Q., Liang, S., Sonne, S.B., Xia, Z., Qiu, X., Li, X., Long, H., Zhang, J., Zhang, D., Liu, C., Fang, Z., Chou, J., Glanville, J., Hao, Q., Kotowska, D., Colding, C., Licht, T.R., Wu, D., Yu, J., Sung, J.J., Liang, Q., Li, J., Jia, H., Lan, Z., Tremaroli, V., Dworzynski, P., Nielsen, H.B., Bäckhed, F., Doré, J., Le Chatelier, E., Ehrlich, S.D., Lin, J.C., Arumugam, M., Wang, J., Madsen, L. & Kristiansen, K. (2015) A catalog of the mouse gut metagenome. *Nature Biotechnology*. Oct;33(10):1103-8.

Yamada, S., Kamada, N., Amiya, T., Nakamoto, N., Nakaoka, T., Kimura, M., Siato, Y., Ejima, C., Kanai, T. & Siato, H. (2017) Gut microbiota-mediated generation of saturated fatty acids elicits inflammation in the liver in murine high-fat diet-induced steatohepatitis. *BMC Gastroenterology.* 17:136.

Yarza, P., Yilmaz, P., Pruesse, E., Glockner, F.O., Ludwig, W., Schleifer, K., Whitman, W.B., Euzeby, J., Amann, R. & Rosello-Mora, R. (2014) Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nature Reviews: Microbiology*. 12:635-645.

Yin, Z., Lan, H., Tan, G., Lu, T., Vasilakos, A. & Liu, W. (2017) Computing platforms for big data analysis: perspectives and challenges. *Computational and Structural Biotechnology Journal*. 15:403-411.

Zengler, K. & Zaramela, L.S. (2018) The social network of microorganisms – how auxotrophiles shape complex communities. *Nature Reviews: Microbiology.* 16(6):383-390.

Zhai, R., Xue, X., Zhang, L., Yang, X., Zhao, L. & Zhang, C. (2019) Strain-specific anti-inflammatory properties of Akkermansia muciniphila strains on chronic colitis in mice. *Frontiers in Microbiology*. 9:238.

Zheng, S., Zhao, T., Yuan, S., Yang., Ding, J., Cui, L. & Xu, M. (2019) Immunodeficiency promotes adaptive alterations of the host microbiome: an observational metagenomic study in mice. *Frontiers in Microbiology*. 10:2415.

Zhu, J., Ren, H., Zhong, H., Li, X., Zou, Y., Han, M., Li, M., Madsen, L., Kristiansen, K. & Xiao, L. (2021) An Expanded Gene Catalog of Mouse Gut Metagenomes. *mSphere*. 6(1):e01119-20.

Zimmerman, M., Zimmerman-Kogadeeva, M., Wegmann, R. & Goodman, A. (2019) Separating host and microbiome contributions to drug pharmacokinetics and toxicity. *Science*. 363(6427).

Zuo, T., Liu, Q., Zhang, F., Yeoh, Y., Wan, Y., Zahn, H., Lui, G., Chen, Z., Li, A., Cheung, C., Chen, N., Lv, W., Ng, R., Tso, E., Fung, K., Chan, V., Ling, C., Joyny, G., Hui, D., Chan, F., Chan, P. & Nh, S. (2021) Temporal landscape of human gut RNA and DNA virome in SARS-CoV-2 infection and severity. *Microbiome*. 9:91.

Chapter 9: Appendix

9.1: Microbiome Analysis Standard Operating Procedure (SOP)

The detailed procedure for microbiome analysis of rodent digesta by the GSK Veterinary Microbiology laboratory is as follows:

- Cage faeces will be used for microbiome analysis unless otherwise requested.
- All samples will be taken by Microbiology group.
- Microbiologists will wear suitable PPE while in animal holding or procedure areas.
- Microbiology group will continue to hold appropriate competencies in methods of rodent euthanasia.
- Microbiology group will continue to complete annual Laboratory Animal Allergens assessment.
- Microbiology group will continue to complete annual facility access training for all sites.
- If digesta from other sites is to be analysed, animals will be euthanised using an appropriate method by Microbiologists in an area designated for this task away from other living animals.
- An appropriate secondary method, confirming death will be conducted by Microbiologists in this same area.
- Cadavers will then be taken to a secondary procedure room with downdraft facilities.

- Animal dermis will be wiped down with 70-100% ethanol.
- Cadavers will be opened using pre-sterilised instruments.
- Samples will be taken in a descending order (mouth to anus).
- Cadavers will be disposed of in accordance with local rules.
- Samples for DNA analysis will be taken into PCR-clean safe-lock, 1.5ml Eppendorf tubes on dry ice.
- Samples for RNA analysis will be taken into 2ml RNA-Later tubes on wet ice.
- Samples will be identified and labelled correctly with indelible ink.
- RNA-Later samples will be allowed to equilibrate on wet ice for at least 1hr.
- Samples shall be protected from UV irradiation when being processed out of animal facilities.
- DNA & RNA samples will be stored at -80°C until nucleic extraction takes place.
- Nucleic extraction will be conducted using Qiagen Power-DNA or Power-RNA kits in the Microbiology laboratory.
- Third party outsourcing agreements will be arranged by Procurement team and TPOs stored in a dedicated shared drive.
- At the agreed time of transfer to third parties, extracted nucleic acid will be taken from -80°C freezer and immediately placed in sufficient dry ice to allow temperature maintenance throughout period of transportation.

- Extracted nucleic acids will be sequenced using pre-agreed and quality checked methodologies by designated third party resources.
- Raw .fastq sequencing files shall be uploaded by third parties to the DNAnexus cloud server via an access key generated by the microbiologist.
- Raw .fastq files will be downloaded from DNAnexus in a timely fashion and stored on a suitable GSK cloud-based server with applicable metadata.
- After action reviews of third-party process will take place to ensure quality and reflection on contracted services.
- Files held on DNAnexus server will only then be deleted.
- Compressed .fastq files will be extracted to .fasta format using WinZip or 7-zip and stored on the analysis computer.
- Assemblies will be conducted using Seq-Man-Pro (Lasergene Genomics suite).
- The specific reference taxonomic database will be stored on the analysis computer and selected during assembly on Seq-Man-Pro.
- Percentage identity will be set at 98.5% and auto-trim will be de-selected.
- K-mer size will be left at default seventeen.
- Temporary files will be assigned to the analysis machine's D drive.
- Assembly .astr output files will be assigned to the analysis machine's C drive.
- The D drive will be deleted immediately after assemblies have finished.

- The reference annot.txt generated using Python script file will be amended to the .astr file.
- Only this .astr file and assembly files for each sample will be saved from the C drive output.
- Analysis of .astr files will be conducted using Array Star (Lasergene Genomics suite).
- Quality checking of assemblies will be conducted using Seq-Man-NGen (Lasergene Genomics suite).
- Resulting operational taxonomic unit data will be transferred by selecting and copying all entry lines and copying directly into Excel (Microsoft, USA).
- The mean average will be calculated and set in ascending order.
- Refine count data by changing lowest represented number to zero.
- Counts below the number of samples taken in the study will be hidden.
- Numeric diversity work will be conducted in Excel (Microsoft, USA) while statistical analysis will be conducted in Array Studio (Qiagen, UK).
- Data shall be communicated to requesters in a timely fashion.

9.2: NGS data handling details (4.2.1)

The file size for each method was 1.85Gb (16S rRNA gene analysis), 561Gb (DNA metagenomic), and 593Gb (RNA-seq), totalling >1.12Tb of data. All downloading of the data and uploading to the analysis software was completed on a domestic Wi-Fi. Downloading this data for local analysis was initially attempted onsite at Stevenage using a GSK-build PC. However, single 16S rRNA gene sequence files took >2.5hrs to download only to fail before completion due to data transfer interference. As the DNAnexus server is an external, downloads were then attempted using a domestic internet connection. All data was successfully downloaded over a period of >1 week but this required all GSK network connections to be disabled. To circumvent network constraints, sequence files were finally downloaded onto a non-GSK build laptop and an external hard drive. An attempt was made to upload 16S rRNA gene analysis files (those being the smallest) onto the MG-RAST server from GSK, however, the paucity of upload speed divided data packages too heavily resulting in all uploads being cancelled. Again, file transfer was attempted via a domestic internet connection. The subsequent attempts to upload these raw files onto the MG-RAST server to analyse them was equally ineffectual from an existing intranet environment. It was sometimes possible to upload just one 16S rRNA file per day, however, intermittent connection meant that only one or two files could be successfully uploaded per week. It again become clear that uploading of data would only be possible from a domestic connection. The significant block to this process was the universal focus on download speed in favour of upload. With a maximum domestic upload speed of <10Mps (often not achieved), the process of uploading >1.25Tb of raw data held on 84 individual files took more than six weeks to achieve. Once uploaded, assembly processing and analysis of raw sequence files took <30 seconds to complete each request.