UCC

University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

The Microbiome and Cancer

A thesis presented to the National University of Ireland for the degree of

Doctor of Philosophy

by

Maurice Barrett

(115223349)

School of Microbiology

National University of Ireland, Cork

June 2021

Supervisors

Professor Paul O'Toole

Professor Fergus Shanahan

Doctor Collette Hand

Head of School

Professor Paul O'Toole

To my parents, Kieran, and Catherine Barrett

“The important thing is not to stop questioning. Curiosity has its own reason for existing. One cannot help but be in awe when one contemplates the mysteries of eternity, of life, of the marvellous structure of reality. It is enough if one tries to comprehend only a little of this mystery every day.”

- Albert Einstein

Table of contents

Declaration

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

Signed: __________________________

Maurice Barrett

Chapter 1 – Literature Review

1.1 Introduction to Microbiota research

Microorganisms colonise an impressive array of niches; from the +120°C hydrothermal vent inhabited by *Methanopyrus kandleri* to the -15 °C high Arctic permafrost inhabited by *Pedobacter sp*[1,2]. Thus, the colonization of multicellular metazoans during their evolution is a seemingly inevitable evolutionary event. Indeed, modern *Homo sapiens* are colonised by a vast number of microbes, collectively referred to as the human microbiota. While the term microbiota has been used to refer to the collection of all resident microorganisms within a niche, the term microbiome can be used to describe the collection of genetic material from an environment. However, these terms are often used interchangeably, and a standardization of definitions is still under discussion[3]. Evolution has generated an intimate relationship between humans and the microbiota; indeed the concept of holobiont has been applied to the microbiota-host interaction wherein the microbiota and the host evolve as a discreet unit[4]

The first description of human beings inhabited by microbes dates to 1670s–1680s, when the Dutch scientist Antonie van Leeuwenhoek examined his own oral sample and that of others and noted "…many very little living animalcules, very prettily a-moving". He noted that there were differences between the oral microbiota between people and later noted differences between faecal samples and oral samples. An early piece of work that further established the embryonic field of microbiota research was 'A Flora and Fauna within Living Animals' published by Joseph Leidy in 1853[5].

The microbiota is composed of bacteria, archaea, fungi, protozoa, and viruses. Most studies focus exclusively on the bacterial aspect of the microbiota, sometimes referred to as the Bacteriome. However, there is an increasing focus on other components of the human microbiota such as the virome (total viral community) and the mycobiome (total fungal microbiome) [6,7]. There is an approximately an equal number of bacteria cells relative to host cells and bacteria[8]. An abundance of microbial niches exist on and within humans notably the oral cavity, the stomach, the large intestine, the skin and the nasal cavity.

1.1.1 The Intestinal microbiota

The greatest concentration of microbes in terms of density and absolute numbers reside in the colon with a density of 10^{11} cells/ml and a volume of 0.4L[8]. The colon is by far the most studied human microbial niche. At the phylum level the most represented phyla (accounting for >90% abundancy) are Firmicutes, Actinobacteria, and Bacteroidetes[9] . While a single individual may harbour 250-500 species the total number of bacteria identified in the gut across all individuals studied is multiplies higher[10-12]. Notably, the colon itself is a multifaceted niche with spatial organization.

The colonic microbiota varies along the colon from proximal to distal as well as cross-sectionally from the lumen to the mucosa. Transversally along the colon, bacterial load, pH, oxygen levels, nutrients levels and immune effectors varies[13,14]. The genera Finegoldia, Murdochiella, Peptoniphilus, Porphyromonas, and Anaerococcus are enriched in the distal colon while the taxa Enterobacteriaceae, Bacteroides and Pseudomonas are enriched in the proximal samples[15]. A greater source of variation is the difference between the lumen and the mucosa [15,16]. In the outer mucus, mucin degrading taxa such as *Bacteroides acidifaciens*, B*acteroides fragilis* and *Akkermansia munciniphila* are found to be enriched while oxygen-detoxifying catalase producing taxa such as those in the Acinetobacter spp. and Proteobacteria inhabit the inner mucosal layer[17,18].

Studies of colonic microbiota primarily depend on the nature of the sample taken which typically takes the form of one of two types, namely, stool samples or mucosal biopsy samples. Mucosal sampling can be conducted in two major ways; a pinch biopsy, involving the use of an instrument to takes a sample of colonic tissue, or a mucosal brush that swabs the mucosa. The mucosal brush cover a higher surface area and recover a higher proportion of bacterial DNA to human DNA relative to biopsy samples[19]. However, pinch biopsy would be more suitable when fine scale analysis of the microbiome is needed. A surgical biopsy can also be taken if the clinical setting allows. Faecal samples are used to represent the luminal microbiome. However, transit time and stool consistency have been demonstrated to affect fecal microbiota composition[20]. Rectal swabs may be used to sample the luminal microbiome and have been decribed as a good proxy for the faecal microbiome[21,22]

Louis Pasteur hypothesized that gnotobiotic or germ-free (GF) animals would fail to survive due to their dependence on their co-evolved microbiota. Although viable, GF mice have a number of aberrant features including a shorten lifespan, enlarged caeca, defective immune system and deficiency in both vitamin K and B12[23,24]. Research during the past twenty years has established a clear relationship between the microbiota and normal physiological function and disease[25]. The gut microbiota have been linked to a myriad of diseases in a number of organ systems (Table 1).

Table 1 | Diseases of different organ systems in which the gut microbiota has been implicated. Neoplastic diseases excluded.

Disease	Microbe abundance	Mechanisms
Autism spectrum disorder(ASD)	Increased[26] *Lactobacillus* *Bacteroides* *Desulfovibrio* *Clostridium* Decreased[26] *Bifidobacterium* *Blautia* *Dialister* *Prevotella* *Veillonella* *Turicibacter*	*Lactobacillus* improves social deficits in mice models via Oxytocin signalling through the vagus nerve[27]. The gut microbiota of individuals with ASD has a decrease capacity to degraded toxins. This decrease is correlated with mitochondrial dysfunction[28].
Cardiovascular disease	Increased[29] *Escherichia coli* *Klebsiella spp* *Enterobacter aerogenes* *Streptococcus spp* Decreased[29] *Bacteroides spp* *Faecalibacterium prausnitzii*	Trimethylamine (TMA) is a metabolite produced by the microbial metabolism of phosphatidylcholine and L-carnitine[30,31]. TMA is absorbed into the blood stream and converted by the liver enzyme flavin-containing monooxygenase 3 (FMO3) into TMA N-oxide (TMAO) [32].Studies in both human subjects and mouse models have demonstrated a role of TMAO in cardiovascular disease development[30,31,33,34].
Type 2 diabetes mellitus (T2D)	Increased[35] *Blautia* *Ruminoccus* *Fusobacterium* Decreased[35] *Akkermansia* *Bifidobacterium* *Lactobacillus*	*Bifidobacterium lactis* has been shown to increase the expression of glycogen synthetic genes while decreasing the expression of hepatic gluconeogenesis-related genes[36] *Akkermansia muciniphila* and *Lactobacillus plantarum* have been found to reduce the expression of fmo3 in mouse models. Note that the the knockout of fmo3 attenuates development of hyperglycemia and hyperlipidemia in insulin resistant mice[37]
Inflammatory bowel disease (IBD)	Increased[38] *Ruminococcus gnavus* *Escherichia coli* *Streptococcus parasanguinis* *Blautia product*	*Ruminococcus gnavus* produces inflammatory glucorhamnan polysaccharide. This polysaccharide induces the production TNFα by interacting with the toll-like receptor 4 (TLR4) of innate immune cells such as Dendritic Cells[39].

	Decreased[38] *Coprococcus Catus* *Alistipes finegoldii* *Blautia obeum* *Faecalibacterium prausnitzii* *Gordonibacter Pamelaeae* *Eubacterium rectale*	Adhesive invasive E. coli (AIEC) can replicate in immune cells such as marcophages. Colonisation of marcophages by AIEC has been shown to induce expression of TNFα[40]
Non-alcoholic fatty liver disease (NAFLD)	Increased[41] *Clostridium* *Anaerobacter,* *Streptococcus* *Escherichia* *Lactobacillus* Decreased[41] *Oscillibacter* *Flavonifaractor,* *Odoribacter* *Alistipes spp*	Members of the gut microbiota have the functional capacity to produce ethanol and genotoxic acetaldehyde which contribute to NAFLD development[42,43] The microbiota produced the metabolite phenylacetate which has been shown to contribute to hepatic steatosis [44]

<u>1.2 Sequencing based technologies and microbiome research</u>

The explosion in the Microbiological sub field of microbiome research has been due in no small part to the advancement in next generation sequencing technologies. A significant proportion of microbiome research is based on the ability to survey the microbial members of a niche as a collective and to make assertions and conclusions based on this information. In particular, microbiome surveys have taken one of two forms; 16S ribosomal RNA gene sequencing and shotgun metagenomics. These methodologies depend on the use of high throughput DNA sequencers.

1.2.1 DNA Sequencing

Form fits function is one of the central themes of modern biological research[45]. The function of DNA is to store information in a stable manner which can be interpreted and replicated with fidelity; this is enable by the double helical structure of DNA as first describe by Watson and Crick. The information density of DNA is immense with 455 exabytes per gram of single-stranded DNA[46]. DNA sequencing involves the representation of the four fundamental base pairs as A, T, C and G.

1.2.1.1 Origins of DNA sequencing

DNA sequencing is an ever evolving endeavour and the variation in the theoretical and mechanical basis behind DNA sequencing is reflected in the wide variety of techniques which have been developed over time.

Wu et al published the first length of DNA to be sequenced which was, a 12 base stretch of the overhanging cohesive ends within the Enterobacteria phage λ, partially published in 1968 with the complete sequence reported in 1971[47,48] . In 1973, Gilbert and Maxam reported the sequence consisting of 24 bases of the lactose-repressor binding site using a method known as wandering-spot analysis, a method which was an adaptation of previous techniques used to perform RNA sequencing [49,50].

DNA sequencing took a significant leap forward with the development of the plus and minus system developed by Sanger and Coulson published in 1975[51]. Using this technique, the first ever whole genome sequencing, that of bacteriophage ϕX174 (PhiX), a single stranded DNA genome of 5,375 nucleotides, was published in 1977[52]. In 1977 Maxam and Gilbert reported a new technique of sequencing ‘DNA sequencing by chemical degradation’[53]. This methodology depended on using a series of 4 different chemical reactions to form abasic sites at specific nucleotide locations; One reaction cleaves at both purines (the ‘A + G’ reaction), one preferentially at A (‘A > G’), one at pyrimidines (‘C + T’) and one at cytosines only

(‘C’). These sites would be subsequently cleaved and the fragmented DNA ran out on a polyacrylamide gel in which the length could be used to infer the base sequence. This was more useful than the plus minus method as it could be employed to decipher all sequences including those within homopolymer runs.

A seminal moment in biological research came with the development of Sanger's ‘chain-termination’ or dideoxy technique in 1977[54]. This protocol involved the use of Dideoxynucleotides (ddNTPs) a deoxyribonucleotides (dNTPs) lacking the 3′ hydroxyl group and which cannot form a bond with the 5′ phosphate of the next based to be incorporated. The introduction of this dNTP into the DNA during synthesis would thus terminate synthesis. Four polymerase chain reactions are set up, one of each containing a small fraction of a radio labelled ddNTP analogues to one of the 4 dNTPs. The small fraction of the ddNTPs mean that this reaction will produce a series of amplicons of differing length. Much like the DNA sequencing by chemical degradation method, the amplicons are ran out on a four lane gel and the sequence inferred by the fragment length.

A number of improvements have been made to Sanger sequencing over the years, notably the replacement of dye-labelled primers with four chain-terminating dideoxynucleotides, each carrying a fluorescein dye with a distinct emission spectrum, condensing the reaction from 4 to 1[55].

In 1980 half of the Nobel Prize in Chemistry was awarded jointly to Walter Gilbert and Frederick Sanger "for their contributions concerning the determination of base sequences in nucleic acids". The other half was awarded to Paul Berg "for his fundamental studies of the biochemistry of nucleic acids, with particular regard to recombinant-DNA".

In 1986 Applied Biosystems Incorporated announced the production of the first automated, fluorescence-based Sanger sequencing machines developed by Smith et al[56]. This machine had the capacity of producing 1,000 bases per day[57].

In 1979 Staden developed the concept of shotgun sequencing, a process whereby fragments of a genome are cloned into a cloning vector and sequenced, after which the genome is assembled based of overlapping sequences. Messing et al developed a single-stranded M13 phage cloning vector which was subsequently used to assemble the genome of bacteriophage lambda *de novo* in 1982[58,59].

In 1995, continuing progress and costs reductions in the 90's allowed for the sequencing of the first complete genome of a free-living organism, *Haemophilus influenza* with a genome of over 1.8 million bases[60]. This was followed by the sequencing of the first eukaryotic genome of *Saccharomyces cerevisiae* (~12 Mb, 1996) and first multicellular organism genome of *Caenorhabditis elegans* (~100 Mb, 1998)[61,62]. In 1990 the United States National Institutes of Health (NIH) launched the Human Genome Project (HGP) with the goal of sequencing the haploid human genome. A draft was published in 2001 and a quasi-complete genome was published in 2004[63,64]. Notably the private company Celera led by Craig Venter endeavoured to sequence the human genome in parallel with the HGP using the whole-genome shotgun strategy and published the results in 2001[65].

1.2.1.2 Second generation sequencing

The 1980s and 1990s saw the development of a new range of sequencing technologies. The first of these was Pyrosequencing. The core principle behind Pyrosequencing, developed by Nyrén and Lundin, involves a luminescent method for measuring pyrophosphate synthesis[66]. In this method ATP sulfurylase is used to convert pyrophosphate, produced during DNA synthesis, into ATP which is subsequently used by luciferase producing light proportional to the amount of pyrophosphate produced. In 1993 the first report of the utilization of pyrosequencing was produced combining the principles of the above protocol with that of the solid phase sequencing method which involved the affixing of DNA templates to streptavidin coated magnetic beads[67].

Pyrosequencing was later licensed to 454 Life Sciences, a biotechnology company founded by Jonathan Rothburg, and in 2005 they produced the first commercial SGS instrument the GS20[68]. This machine was constructed with microfabricated microarrays allowing for mass parallelisation of sequencing reactions. This system produced reads of length 400–500 bp. The GS20 was superseded by the 454 GS FLX, which offered a greater number of reads and quality of base calling[68].

The Solexa (Illumina) method is the mode of sequencing that currently dominates the marketplace. Base calling is depended on fluorescent reversible-terminator dNTPs. A fluorescent dye molecule indicates the insertion of a base as DNA synthesis occurs. Both the terminator group and the fluorophore must be removed before the next base is incorporated and the base called. This concept of fluorescent reversible-terminator dNTPs was first envisioned by Bruno Canard and Simon Sarfati at the Pasteur Institute[69]. Work on this concept eventually led to the development of photo-cleavable fluorescent nucleotide reversible terminators for each base[70,71]. This allowed a design where cleavage can be followed by a wash step to remove unincorporated bases. Successive rounds of this allow for the sequence of a template to be determined. Another key concept to the Solexa method is bridge amplification. Bridge amplification enables the production of tight clustering of template copies known as “polonies”, allowing for better base calls[72]. The first Solexa commercial sequencer, the Genome Analyzer (GA) machine, was released in in 2006. This machine outputted 1 GB of data and the reads had a length of 35 bp. However, this method of sequencing involves paired end sequencing in which both ends of the amplified DNA template are sequence. This enables a merged read to be formed from a homologues overlap between the paired reads. In 2007, Solexa was acquired by Illumina. Currently, Illumina currently hold ~75% of the global market share of genetic sequencing. Illumina's premier platform, the NovaSeq 6000 Sequencing System, can output 4800-6000 Gb of data and supports an output of 250bp x 2 read output.

1.2.1.3 Third generation sequencing

While NGS platforms produced by Illumina are continually improving especially when it comes to throughput and cost, these technologies have fundamental drawbacks that limit their use in biological research. One of these issues is the relatively short read length. Illumina platforms usually have an upper limit of 300bp with regard to read length[73]. Furthermore, the Illumina sequencing method depends on an initial polymerase chain reaction (PCR) bridge amplification which can produce a bias with regard to DNA of extremely high guanine-cytosine content (GC) content as these are inefficiently amplified by PCR[73].

We are now seeing the increase usage of what can be described as third generation sequencing (TGS) technologies. There are currently two commercially available TGS technologies; single-molecule real-time (SMRT) sequencing by Pacific Biosciences (PacBio) which is the first viable TGS platform released in 2011 and nanopore sequencing by Oxford Nanopore Technologies (ONT) released in 2014. Both these technologies can produce very long reads with SMRT producing read length N50 values of ~20 kb while Nanopore sequencing can produce read length N50 100 kb[74]. Furthermore, these technologies can be described as real time sequencing as the data is read out continually as each base is deciphered[73].

SMRT involves ligating adapters to the DNA to be sequenced creating a SMRTbell™ library which is a cellular template[75]. These templates are immobilized in wells denoted zero-mode waveguides. A polymerase performs synthesis and incorporation of fluorescently labelled nucleotides is detected, thus SMRT sequencing can be called a SBS method[75].

The core strategy behind ONT platforms involves a motor protein ratcheting DNA through a nanopore in which a current is passed through[76]. Bases are read via interpreting the signal produce by the disruption of the current cause by the base as it passes through the nanopore[76].

Both of these methods can also detect DNA modifications such as 5-methylcytosine (5mC) and 6 methylated adenine. SMRT can do this via measuring the time between nucleotide incorporations is called the ‘interpulse duration’[75]. In essence the length of time between incorporation is indicative of the status of the DNA modification. ONT platforms can detect DNA modifications due to the characteristic disruption they exert on the passing current which is distinguishable from the unmodified base[77].

Sequencing regions of genomes which are repetitive in nature are difficult to delineate using NGS platforms. Such features including centromeres, telomeres and tandem repeats. TGS platforms have the potential to sequence the entirety of a repetitive region thereby avoiding the challenges of assembling these regions using NGS data[78]. TGS have a number of other benefits over NGS such as sequencing RNA isoforms and Haplotype phasing[73].

With respect to microbiology, it is conceptually possible to sequence an entire an entire bacterial genome *de novo*. TGS is also being used in microbial marker gene studies namely 16s rRNA gene sequencing (See section 1.2.2.1).

1.2.2 The 16S ribosomal RNA gene

Ribosomes are ribonucleoprotein structures with the biological function to perform protein synthesis. Ultracentrifugation protocols sediment the bacterial ribosome at 70 Svedberg unit (S) while its constituent parts , the large and small subunit, sediment at 50S and 30S respectively. The large subunit is composed of 33 proteins (Denoted L1–L36) and two rRNAs, the 23S rRNA and the 5S while the small subunit is composed of 21 ribosomal proteins (denoted S1–S21) and a 16S rRNA.

Canonically, the three ribosomal RNAs genes are organised on the Ribosomal RNA Operon in the order 16S-23S-5S. However in some bacteria and archaea the rRNA genes are “unlinked” whereby there is a substantial genomic distance between the 16S and 23S rRNA genes, a phenomenon which is much more prevalent that once believed[79]. However, the unlinked structure does not seem to be present in the gut. In the canonical set up, the three RNAs are all transcribed as one. Within the rRNA Operon there also exist internal transcribed spacer (ITS) regions between 16S and 23S rRNA genes which also contains a DNA sequences encoding for tRNAs. The number of operons in a species can vary considerably with counts from one and twenty one[80].

The median size of the 16s rRNA gene is ~1500 but varies considerably in range[81].

The 16S rRNA gene is thought to be conserved throughout both the bacterial and archaeal domains of life. Although classical dogma would indicate that this conservation is indicative of the essential nature of the 16S rRNA gene, recent research has supported the idea that the evolution rate of a gene is negatively associated with its expression level[82,83].

16S rRNA has a number of functions. The 16s rRNA contains an anti-Shine-Dalgarno sequence which binds to the Shine-Dalgarno sequence in the mRNA sequence and influences translational pausing and codon choice[84]. 16s rRNA also plays a structural role providing a scaffolding in the small subunit.

The sequence structure of the 16S gene can be described as containing nine hypervariable regions (V1–V9) and nine conserved regions (C1-C9). This structural composition is the basis for its use as a taxonomic identifier in 16S rRNA gene sequencing studies.

1.2.2.1 16S ribosomal RNA gene sequencing

Studies which survey the microbiota utilizing sequencing methodologies usually fall into one of two strategies, amplicon-based marker gene surveys or metagenomic whole genome shotgun sequencing (mWGS).

The 16S rRNA gene is a putatively ubiquitous gene in the domains of Archaea and Bacteria. Carl Woese and George E. Fox pioneered the use of the 16s rRNA gene as a phylogenetic marker in their seminal work in which they proposed the three domains of life—Bacteria, Archaea, and Eukarya[85]. Wilson and Blitchington published the first 16S rRNA gene sequences derived from a faecal sample[86]. Suau et al demonstrated that much of the gut microbes captured by the 16S rRNA gene sequences could not be cultured[87]. This work was echoed in the same year with respect to subgingival scrapings by researched carried out by Kroes et al[88]. Although progress has been made with respect to culturing human associated microbes, culture independent sequencing techniques still cast a wider net than culture dependent techniques[89,90]. In 2005 Eckburg et al set a precedence for the scope of microbiome research with their study in which they sequenced 13,355 sequences of the 16S rRNA gene from multiple colonic mucosal sites and faeces from 3 individuals[91]. They reported variation in the microbiome with respect to biogeography as well as significant inter-individual variation. The method of mWGS involves the untargeted sequencing of the genetic contents of a niche. These two strategies have their own inherent advantages and disadvantages (Table 2)

Table 2 | Characteristics of 16S rRNA gene sequencing versus metagenomic whole genome shotgun sequencing

16S rRNA gene sequencing	Pro • Inexpensive (10x cheaper per sample than mWGS) • Computationally less taxing • Less storage space need for data • Selective for archaea and bacteria Cons • Depending on the primers used and other factors, taxonomic resolution usually only goes down to the genus level and occasionally down to species level • Lacks direct functional information • Certain primers can amplify mammalian DNA
Metagenomic whole genome shotgun sequencing	Pro • Complete genomic content • Potential to inspect single nucleotide variant across the genome of an organism • Strain level resolution • Functional information Cons • High read count needed to achieved coverage need to represent through species richness • Samples high in Host DNA such of biopsies can mean the 99% map to the host genome. • Relatively expensive

1.2.2.2 Laboratory aspects of 16S ribosomal RNA gene sequencing

Current 16S experiments often require the production of thousands of 16S reads from hundreds or thousands of samples. The most cost effective and streamlined way of achieving this is to employ NGS namely Illumina paired end sequencing. Amplicon sequences to be analysed are produced by merging paired reads. The Illumina platform most frequently used for 16s is the MiSeq System which depending on the Reagent Kit produces reads of length of 250 or 300bp in length. Taking into consideration the need for a certain number of bases to overlap, the merged amplicon read would be under 600bp, thus only a subsection of the 16s gene can be sequenced. In particular one or more of the variable regions are sequenced. Research has been carried out to determine the most informative primers to use when amplifying a 16S subsection. These primers must best capture the taxonomic diversity while limited to amplifying a section under 600bp. Many such primers pairs have been designed and utilized to study the microbiota. Studies have been conducted to identify the taxonomic diversity these primers capture. Currently, the most prominently used being are the V1-V2 and V3-V4 primer pairs[92].

The polymerase used in 16S gene sequencing experiments are preferably of high fidelity. Taq polymerase has an error rate of 1–20 × 10^{-5} while Phusion® High-Fidelity DNA Polymerase has a 50X increase in fidelity[93].

In current protocols namely those within the Illumina 16S Metagenomic Sequencing Protocol (Illumina, California, USA) sample specific DNA barcodes are added to the sample amplicons in a second PCR known as an index PCR[94]. Previous protocols involved adding these barcodes in the same PCR as the initial amplification step. However it was found that this produced PCR related biases[95].

1.2.2.3 Bioinformatic analysis of 16S rRNA gene sequencing data

Raw data from sequencers must undergo a series of processes before descriptive and statistical analysis can be effectively carried out. A key aspect of this is the assembly of representative sequences. The two premier forms of this are operational taxonomic unit (OTUs) and amplicon sequence variants (ASVs). The generation of OTUs and ASVs both aim to address the issue of incorrect base calling.

With regard to Illumina sequencing, data is output in a fastq format. This format is similar to fasta contains the sequence information but also reporting the corresponding base calling quality in the form of a Phred-like quality score (https://www.illumina.com/science/technology/next-generation-sequencing/plan-experiments/quality-scores.html). The quality score (Q) of a base is calculated by the following equation: $Q = -10\log^{10}(e)$ where e is the estimated probability of the base call being incorrect . For example, a Q score equal to 10 would indicate there is a 1/10 chance of the base being called incorrectly. The maximum score is 40 which equates to an average per base error rate of 1/10000. If one were to take sequencing data unprocessed, difference in sequences due to errors could be inappropriately interpreted as an actual biological difference representing evolutionary divergences.

The term OTUs was coined by Sokal & Sneath referring to groups of closely related individuals being studied[96]. In modern microbiology terms, OTUs are representative sequences based on a threshold of identity, typically 97%[97,98]. There are two methodologies to achieve OTU clustering 1) '*de novo* clustering' and 2) 'Closed-reference OTU clustering. In *de novo* clustering, merged reads are clustered within a dataset based on a certain threshold. The OTUs generated from *de novo* clustering are emergent features of the particular data set which is being studied. Factors such as relative abundances will dictate the generation of the OTUs. Thus de novo OTUs generated from two different datasets cannot be compared. Closed reference OTU cluster merged reads against a reference database. If the same database is shared between two different data-sets, the generated OTUs can be more readily compared against each other. However, biological variation that is not represented in the reference database would lead to a reduction in the diversity detected during assignment to closed-reference OTUs. No matter what the method used for generating OTUs, the clustering methods will lead to the loss of some actual biological variation in the dataset and thus OTU type leads to an under-representation of diversity.

ASVs aim to represent the real biological sequence of the maker-gene. Thus ASVs resolves the data-set to the single nucleotide resolution. The generation of ASVs is dependent on the assumption that biological variants are more likely to be observed in a dataset than those generated by erroneous base calling. In practice, an algorithm needs to generate an error model using read data. Sample ASVs are then inferred by a process known as denoising[99]. At present there are three main software packages utilized for ASV generation, that is, DADA2, UNOISE3, and Deblur[100,101]. DADA has been reported to offer the best sensitivity in terms of number of ASVs detected but perhaps at the cost of specificity[102,103].   Using ASVs to define a microbial community has the potential to overestimate diversity due to intragenomic variation of the 16S gene [104,105].

Taxonomic assignment

An integral aspect of 16S surveys is defining the taxa that are presence in a niche. Both ASVs and OTUs may be assigned to taxonomic rank. A myriad of classification algorithms have been developed including BLAST, IDTAXA, MAPSeq, QIIME, SINTAX, SPINGO, and the RDP Classifier[106]. Furthermore, there exist a number of reference databases of 16S rRNA gene sequences to which the algorithms most popular being SILVA, the Ribosomal Database Project (RDP) and Greengenes[107-110].

Ecological analysis

Methodologies classically used to describe niches of multicellular organisms are also used to describe microbiological niches. In particular alpha diversity (α-diversity) and beta diversity (β-diversity) are frequently used as metrics to describe the overall structure of microbiomes. Alpha diversity describes the richness and evenness of organisms within a niche. There are many indices that are used to calculate alpha diversities each describing richness in different manners (Table 3)

Table 3 | Explanation of alpha-diversity metrics

Alpha diversity metric	**Description**	**References**
Observed species	Counts the number of taxa.	[111]
Chao1	Assumes that the number of observations for a taxa has a Poisson distribution and corrects for variance.	[112]
Simpson's Index	Considers the Evenness of the data. Factors relative abundance of each taxa into the count.	[113]
Shannon index	Much like Simpson's Index, this index considers evenness by adjusting for relative abundances.	[114]
Phylogenetic diversity	This diversity metric considers not only number of taxa but also phylogenetic distance between taxa.	[115]

Beta diversity measurers the difference (or similarity) in microbial composition between samples. Like alpha diversity there are many beta-diversity metrics that can be utilized to describe differences between niches (Table 4).

Table 4 | Explanation of beta-diversity metrics

Beta diversity metric	Description	
Jaccard Index	Calculates similarity base on presence absences. Does not factor abundance.	116
Bray–Curtis dissimilarity	Calculates similarity base on presence absences. And also factors abundance.	117
Unweighted Unifrac distance	Unifrac distance considers phylogenetic between distances between taxa. Unweight considers presences absences.	118,119
Weighted Unifrac distance	This considers not only presences/absences but also abundances of taxa	118,119

Differential abundance

A central goal of many microbiome studies is to identify taxa/ASVs/OTUs that are differentially abundant between groups to a statistically significantly degree. How one achieves this goal is of much debate within the microbiome field. Microbiome data is sparse, complex, and compositional in nature[120,121].

A "classical" test for differential abundance is the Wilcoxon rank-sum test (also called the Mann-Whitney U test) which is a nonparametric test. Microbiome sequence data is compositional in nature[120]. This is simply due to the fact the observation of the genetic data of the microbiome is limited by the number of reads produced by the sequencer. The package ALDEx2 which performs a centred log-ratio (clr) transformation on the data has been argued to be suitable for addressing this the compositional nature of microbiome[122]. Software packages originally developed for RNA-seq such as DESeq2, which employs negative binomial generalized linear model, have also been applied to 16S data sets[123]. Other differential abundance methods have been developed with specific consideration for microbiome data including metagenomeSeq, ANCOM and ANCOM-BC [124-126].

Prediction of gene function

A major limitation of 16S based experiments is that they do not provide direct information on the functional capacities of the microbial community which is being studied. However, there are a number of bioinformatic tools which infer functional capabilities of a community from 16S sequence data. Current softwares include, PICRUSt (The most frequent used), Tax4Fun, Piphillin and PICRUSt2 (the successor to PICRUSt)[127-130]. The core methodology employed by this group of software depends on the alignment of the 16S sequence to functionally annotated reference genomes. Another recently developed tool, IPCO, utilizes a different method which depends on the procedure of double co-inertia analysis involving the RLQ method[131]. Within this method a query data set (16S data set) is co-varied against a paired taxonomic and functional dataset (16S data set and shotgun metagenomics dataset) and the functional data of the query data set inferred from this[131].

1.2.2.4 Future of 16S ribosomal RNA sequencing studies

The development of third generation have led to the possibility of sequencing much larger amplicons compared those possible on Illumina's platforms. Indeed, with TGS it is possible to sequence the whole 16S rRNA gene. Although cost per base continues to declines with these technologies, the viability of their common usage is still restricted by cost. Furthermore, the relatively high error rates of base calling of TGS limits their use in taxonomic delineation. Nonetheless efforts have been made to set up standard operating procedures for the use of TGS in 16S rRNA gene-based surveys.

Possibly the greatest progress has been made with PacBio SMRT sequencing. A method to address high error rate involves the formation of a Circular Consensus Sequences (CCS). A CCS is formed by ligating hairpin adapters that circularize linear DNA molecules and allowing the sequencing polymerase to make multiple passes and producing multiple sub reads. These sub-reads are collapsed into the CCS. Callahan et al used the CCS in conjugation with the denoising algorithm DADA2 to carry out 16S analysis on the mock Zymo community (a commercially available consortium of 8 bacteria and 2 yeasts) and Human Microbiome Project (HMP) mock community (a consortium of 21 microbes developed by the HMP)[132]. This method produced full length (~1.5 kb) 16S rRNA gene reads at an error rate of 4.3×10^{-4} per nucleotide, a comparable error rate to reads produced Illumina on sequencing platforms. This strategy allowed for the identification of intragenomic allelic variation and sub-species classifications. In particular they were able to delineate the enterohemorrhagic O157:H7 clade while the same strain could not even resolve between the Escherichia and Shigella genera when commonly used V3V4 and V4V5 regions were sequenced.

Johnson et al showed that a range of different 16S subsections sequenced using the Circular Consensus Sequences method underperformed verses the whole 16s rRNA gene when it came to capturing diversity[133]. The authors also suggested that clustering at 99% maybe be used to address the issue of over-estimation of diversity due to intragenomic variation between 16S gene copies[133].

Studies utilizing Nanopore sequencing platforms have shown that using the full-length sequence has advantages over sequencing only sub-sections. However the error rate remains too high for appropriate use in 16S rRNA gene sequencing experiments[74].

Future studies may even utilize the whole rrn operon (16S rRNA–ITS–23S rRNA) as this would further increase the resolution with regard to phylogenetic delineation[134]. Current techniques can feasibly address the ~3kb rrn operon[132].

1.2.3 Contamination

Advances in culture-independent next-generation sequencing techniques, namely shotgun sequencing and marker gene PCR based methodologies, have revolutionized our understanding of microbes in numerous niches due to their speed, sensitivity and ever reducing cost. However, the sensitive of these techniques, especially amplification-based methods, have come with the notable downside of detecting DNA sequences which do not belong to the niche under study, that is to say contamination. The challenge of contamination is inversely proportional to the microbial load of the niche under study; studies of high load microbial niches such as luminal faecal matter are less proportionally affected by contamination than low load niches such as glacier ice or brain tissue[135-137]. The problem of contamination has been brought into focus recently and with regard to the human microbiota, reports regarding the placental microbiota have brought notable controversy. This section will discuss the issue of contamination, its origin, its impact on the microbiome field and how it may be addressed.

1.2.3.1 Sources of contamination

The ubiquitous nature of microbes mean that contamination has a plethora of sources including neighbouring niches, sampling equipment, extraction kits, PCR reagents (including polymerase mixtures), laboratory personnel, environments, and equipment.

One of the first sources of contamination that researchers can encounter is contamination from adjacent niches. One can mistakenly sample microbes from a site within close proximity of the niche being investigated. This challenge is especially amplified if the niche under study is of low biomass and the adjacent sites have higher biomass. A seemingly convenient method to sampling the microbiota of the bladder is urine collection. However, this sample type will contain microbes not only from the bladder, but also distal urethra and in the case of women from the vulva and vagina[138]. It is proposed that suprapubic aspiration or transurethral catheterization is required to collect samples directly from the bladder microbiota[138].

Sampling breast tissue microbiota is usually done via surgical resection. However, this process has the potential of acquiring contamination from the skin. Some studies prudently include paired samples of the skin microbiome to control for such cross contamination[139,140]. Pertinent to this thesis, the oesophageal microbiota is in close proximity to the oral cavity and gastric microbiota; both of which are higher biomass than the oesophagus. One should be able to successfully sample the oesophagus via biopsies or swabs.

The methods of extracting nucleic acid for microbiome studies have primarily employed commercially available kits. Although, not overtly non-sterile, trace amount of microbial DNA have long been recognised as been present in the commercial kits[141,142]. Salter et al were arguably the first to study the impact of the impact of kit contamination on high-throughput culture independent methodologies[143]. Using the above techniques, Salter et al studied the effect of serial dilutions on *Salmonella bongori*, 10^8 to 10^3 cells. They found that contaminating reads were present and that this was proportional to the dilution factor of the sample with ~90% of reads belonging to contaminant taxa in the most dilute sample. Furthermore, they found contamination in a range of different commercial kits and to some extent a defined microbiome could be linked to a specific kit.

Glassing et al calculated that there was a presence of 10–15 *E. coli* genome equivalents (70–105 rRNA gene copies) per µl elution buffer from the MoBio PowerSoil Kit. Further 16S rRNA gene sequencing of blank extractions from this kit yielded 81 bacterial genera and 108 tentative species[144].

Marker gene-based genome microbiomes surveys, namely 16S RNA gene based sequencing depend on polymerase chain reaction. PCR master mixes have been identified as sources of contamination[142,145-147]. For the extraction kits and master mixes investigated, Stinson et demonstrated that the PCR master mix was a much greater contributor to contamination that the DNA extraction kits.

1.2.3.2 Resolution of the contamination problem

Knowing the origins of contamination, how it presents itself and when it becomes a considerable factor, one can devise protocols to eliminate or to at least take account the risk of contamination. Indeed, direction and guidelines have been constructed to conduct microbiome research while accounting for contamination[148,149]. Eisenhofer et al proposed a minimal experimental criteria denoted the ‘RIDE’ checklist which they argue should become a “Minimum Standards Checklist for Performing/Reviewing Low microbial Biomass Microbiome Studies”[148].

As contamination is predominantly an issue in low biomass samples, one must endeavour to maximise the cell density of the microbial sample. This may not of course be possible in every study. However, one should quantify starting material microbial load by utilizing methods such as Quantitative PCR (qPCR). For example Salter et al suggested a biomass of over 10^3 to 10^4 cells would be needed to overcome background contamination[143].

As noted above reagents are a major source of contamination. One can use reagents which have an emphasis on the quality of being microbial DNA free. Qiagen produce the ‘QIAamp UCP Pathogen Mini Kit’ which undergoes DNA decontamination processes and is certified as free from contamination. Kirstahler et al produced data that support the hypothesis that such kit reduces contamination[150].

Procedures have been developed to decontaminat PCR reagents[151]. Commercially low contaminant PCR reagents are now available such as MTP Taq DNA Polymerase(MERCK).

In silico methodologies have also been developed to remove contaminating OTUs or ASVs. Firstly, one can simply remove the taxa from one’s data-set which appear in a negative control[152]. Functions such as ‘remove.seqs’ within Mothur allows for such operations. However, this method runs into 2 problems. One, contaminating taxa may overlap with actually biological taxa. Two, the phenonema of index swapping means that reads can be assigned to the incorrect sample whic occurs at a non-negligible rate (0.2 to 6%) [153-155]. Thus, one can mistakenly remove biologically relevant taxa that due to index hopping/swabbing shows up in the negative. Jervis-Bardy et al demonstrated an inverse relationship between relative abundance of contaminating taxa and sample DNA concentration[156]. The open-source R package ‘decontam’ performs such an analysis and identifies contamination[157]. Finally, in the case of well-defined sources of potential contamination, one can use SourceTracker which employs Bayesian modelling to calculate the proportion of potential contaminant taxa within a sample[158].

1.2.3.3 The placental microbiome controversy: a case study

Many anatomical features of humans have long been believed to be sterile including the womb. At the turn of the century the French paediatrician Henry Tissier put forward the model whereby human development occurs initial in the sterile womb and the individual acquires microbes during birthing[159]. In 2014 work published by Aagaard et al provided evidence for a unique placental microbiome. According to Bray-Curtis dissimilarity, this microbiome was most closely associated with the HMP oral dataset. Subsequent studies have been built on these finding, identifying associations between the placental microbiome and excess maternal gestational weight gain, birth weight, pre-eclampsia and gestational diabetes[160-163]. An additional importance of the discovery of a placental microbiome is that it necessarily alters models of the initial genesis and development of an individual’s microbiome. Collado et al formed a framework of microbiome development based on data which included data from placenta and amniotic fluid [164].

However there has been a number of studies challenging the notion of a placental microbiome[165-168]. These studies were designed the experiments to appropriately delineate background contamination from microbes that may exist in the placental samples. These studies could not provide evidence of a placental microbiome which wass was separate from contamination. However, Goffau et did find evidences for the presences of *Streptococcus agalactiae* in ~5% of placental samples studies[167].

1.3 Cancer and the microbiota

Cancer is an umbrella term for an array of diseases which are characterised by the transformation of normal cells into aberrant cells which dispays the qualities of ‘The hallmarks of Cancer[169,170]. This process occurs via somatic evolution fuelled by somatic mutations [171].Worldwide, in 2018, there was an estimated 18.1 million new cancer diagnosis and 9.6 million cancer deaths[172]. The total economic burden of cancer was calculated to be 1.16 trillion USD in 2010[173]. Further, cancer incidence has been projected to double by 2035. An analysis of cancer deaths in the USA between 1969 and 2013 found an age-adjusted decrease in cancer deaths of 17.9% while another study on the US population found a decline in cancer related mortality of 27% between 2007-2016[174,175]. It has been argued that this comparably modest reduction in cancer mortality is due to the lack of support in cancer prevention research[176]. Cancer prevention is relatively under researched when compared to therapeutic development with only 2 to 9% of research funding going towards this area[177].

As stated above, cancers arise due to the accumulation of somatic mutations through time. About 42% of cancer incidences in the US have been attributed to modifiable risk factors, a figure which is reflected in the UK population[178,179]. The International Agency for Research on Cancer (IARC) compiles and evaluates data on known carcinogens. Notable group 1 carcinogens including tobacco smoke, UV light and obesity. These carcinogens promote oncogenesis through a plethora of mechanisms.

Infectious agents are also among well-established carcinogens. There is eleven infectious agents which infect humans that are classified as group 1 carcinogenic agents (Table 5). In 2012, 15.4% of cancer incidence were attributable to ten of eleven of these infectious agents i.e. exclusive of HIV.

Table 5 | Estimated numbers of infection-attributable cancer cases in 2018, by infectious pathogen, cancer subsite, and sex (Data derived from Martel et al, 2020)[180]. These data exclude HIV attributable cancer incidences.

	Men		Women		Total	
	New cases	New cases attributable to infectious pathogens	New cases	New cases attributable to infectious pathogens	New cases	New cases attributable to infectious pathogens
Helicobacter pylori						
Non-cardia gastric cancer	550 000	490 000	300 000	270 000	850 000	760 000
Cardia gastric cancer	130 000	27 000	46 000	8900	180 000	36 000
Non-Hodgkin lymphoma of gastric location	12 000	8700	10 000	7600	22 000	16 000
Human papillomavirus						
Cervix uteri carcinoma	..	..	570 000	570 000	570 000	570 000
Oropharyngeal carcinoma	110 000	34 000	26 000	8100	140 000	42 000
Oral cavity cancer	190 000	3900	91 000	2000	280 000	5900
Larynx cancer*	150 000	3600	22 000	≤1000	180 000	4100
Anus squamous cell carcinoma	9900	9900	19 000	19 000	29 000	29 000
Penis carcinoma*	34 000	18 000	..	..	34 000	18 000
Vagina carcinoma*	..	..	18 000	14 000	18 000	14 000
Vulva carcinoma*	..	..	44 000	11 000	44 000	11 000
Hepatitis B virus						
Hepatocellular carcinoma	490 000	270 000	170 000	90 000	660 000	360 000
Hepatitis C virus						
Hepatocellular carcinoma	490 000	100 000	170 000	40 000	660 000	140 000
Other non-Hodgkin lymphoma	260 000	8700	210 000	7200	480 000	16 000
Epstein-Barr virus						
Nasopharynx carcinoma*	92 000	76 000	35 000	29 000	130 000	110 000
Hodgkin lymphoma*	46 000	24 000	33 000	17 000	80 000	40 000
Burkitt lymphoma	7800	4100	3800	2500	12 000	6600
Human herpesvirus type 8						

Kaposi sarcoma*	28 000	28 000	14 000	14 000	42 000	42 000
Schistosoma haematobium						
Bladder carcinoma	420 000	4000	120 000	1900	550 000	6000
Human T-cell lymphotropic virus						
Adult T-cell leukaemia and lymphoma	1900	1900	1700	1700	3600	3600
Opisthorchis viverrini and Clonorchis sinensis						
Cholangiocarcinoma	69 000	2100	56 000	1300	130 000	3500
All cancer types related to infection	..	1 100 000	..	1 100 000	..	2 200 000

In general, microbiome studies take a more global view of the microbial community. It is unlikely that such studies would identify microbes which contribute a strong odds ratio to cancer. However, these studies offer a framework where one can link global community structures to cancer biology while also preserving the ability to dissect the microbiome to the resolution of species and strains.

1.3.1 Cancer tissue microbiome

The colonic microbiome can exert a biological effect on practically all tissues in the body through a number of mechanisms including communication with the immune system. Hence, the colonic microbiome has been associated with cancers of many tissues not only colorectal cancer[181-183]. Studies have also revealed the existence of microbiomes in non-GI tissue and have been implicated in the cancer biology of host tissue[184] (Table 6). These microbiomes are generally very low biomass in nature and therefore would be susceptible to contamination (See section 1.2.3).

Table 6 | Examples of intratumoral microbiomes and their influences on tumour biology

Cancer type	Example of taxa identified	Comments
Breast	*Enterobacteriaceae* *Bacillus* *Staphylococcus*	*F. nucleatum* is overrepresented in breast tumour samples. Colonization of breast cancer by *F. nucleatum* is facilitated by binding of bacterial Fap2 to breast tissue expressed Gal-GalNAc. Mice models breast cancer demonstrated a role of *F. nucleatum* in promoting tumour growth and metastatic progression. Evidence suggest *that F. nucleatum* does so by suppressing accumulation of tumour infiltrating T cells[185]
Pancreatic Adenocarcinoma (PAC)	Bacteria *Pseudoxanthomonas* *Saccharopolyspora* *Streptomyces* Fungi *Ascomycota* *Basidiomycota* *Malassezia*	Mouse models demonstrate the ability of bacteria to translocate from the gut to the pancreas[186]. Ablating the pancreatic microbiota via germ free models or antibiotic treatment increased infiltration of the tumours with CD4+ T Helper-1 and cytotoxic CD8+ T cells and reduced immunosuppressive myeloid-derived suppressor cells and M2-tumor-associated macrophages Individuals who were classified as long-term survivors had a higher alpha-diversity of the PAC microbiome relative to those who were classified as short term survivors[187]. The abundance of three taxa Pseudoxanthomonas, Saccharopolyspora, and Streptomyces with the species *Bacilus Clausii* is highly predictive of long term survival. The PAC microbiome was associated with long term survival was correlated with recruitment and activation of CD8+ T cells in PADC tissue[187]. In mouse models, gut fungal taxa where observed to translocate from the gut to pancreas. The PDA mycobiome of both humans and mice showed are composed of similar taxa and differed their respective gut microbiome[188]. In mouse models, Fungal ablation protected against oncogenesis while colonisation of the pancreas with the fungal species *Malassezia globose* promoted oncogenesis[188]. Fungal interaction with the mannose-binding lectin may promote oncogenesis by activation of the complement activation[188].

Lung	*Granulicatella* *Abiotrophia* *Streptococcus* *Cyanobacteria*	Higher alpha diversity was observed in tumour tissue and matched healthy tissue compared to healthy controls In particular those tumours with TP53 mutations was enriched with Acidovorax. Cyanobacteria-derived microcystin increase expression of oly (ADP-ribose) polymerase 1 (PARP1) in Non-small cell lung cancer cell models[189].

1.3.2 Fusobacterium nucleatum

Fusobacteria nucleatum is a Gram-negative anaerobic non–spore forming, non-motile bacillus belonging to the genus Fusobacterium. *F. nucleatum* has classically been described as an opportunistic commensal pathogen with a well-established a role in periodontal disease[190]. In recent years *F. nucleatum* has been identified in a range of other human microbiotas and has been associated with an ever-increasing number of diseases including atherosclerosis, liver abscess and most notably cancer [191-194]. In particular there is a growing literature with respect to *F. nucleatum* and its relationship to colorectal cancer oncogenesis and progression.

1.3.2.1 Fusobacterium nucleatum association with colorectal cancer

There is mounting literature regarding an increase higher abundance and of *F. nucleatum* in CRC relative to healthy controls. Initial studies by Castellarin et al and Kostic were among the first to demonstrate this relationship[195,196]. There has since been numerous studies utilizing a myriad of techniques that have corroborate these findings. A recent meta-analysis carried out by Gethings-Behncke et which surveyed the prevalence and abundance of *F. nucleatum* in individuals with colorectal cancer compared with healthy controls in both mucosal and faecal samples found that the signal of the positive association between *F. nucleatum* and CRC was maintained[197]. In particular an odds ratio of *F. nucleatum* DNA being detected in CRC versus healthy controls was 9.01 and 10.06 for faecal and mucosal samples respectively. Further, in individuals who were F. nucleatum positive a consistent increase in abundance in CRC in both sample types was found. Moreover F. nucleatum was seen to have prognostic value with poorer survival in patients with colorectal cancer with high versus low F. nucleatum abundance (Hazard ratio = 1.87)[197].

Another meta-analysis of faecal metagenomes identified *F. nucleatum* adhesion protein A as being overrepresented in CRC versus healthy controls [198]. A prospective analysis on a large American cohort found that prudent diets (rich in whole grains and dietary fiber) were negatively associated with *F. nucleatum* positive tumours[199]. This suggests a complex relationship between diet, the microbiota and CRC.

Fluorescent in situ hybridization using Fusobacterium-specific 16S probes has identified Fusobacterium species cells localized within the crypts of colorectal sections[200]. Furthermore mucosal associated *F. nucleatum* cells have been demonstrated to be viable as they can be cultured from mucosal samples[201].

With regard to the consensus molecular subtypes (CMS) of CRC, *F. nucleatum* was found to be increased in CMS 1, a molecular subtype defined by microsatellite instability and immune cell infiltration as well as poor prognosis[202,203]. There also appears to be variation in the biogeography of *F. nucleatum* colonization, with F. nucleatum-high colorectal cancers gradually increasing from rectum to cecum in an approximately linear reletionship[204].

One of the current models for why *F. nucleatum* is found in the gut is that it transfers constantly from reservoirs in the mouth to the gut via the GI tract. Oral taxa have been found to be enriched on CRC tumour tissue relative to matched healthy tissue[205]. Strain level metagenomic analysis of paired oral-stool samples found extensive and persistent transmission of oral strains to the gut[206]. Furthermore, these analyses found that this transmission was higher in individuals with CRC[206]. Strain typing of cultured *F. nucleatum* from matched mucosal biopsies and oral samples using degenerate primers revealed that these strains were identical between sites within individuals[201]. Another model of how *F. nucleatum* may reach the gut is through the circulatory system. Transient bacteraemia is was observed in individuals up to 15 minutes post tooth brushing[207]. One study found *F. nucleatum* could be cultured from blood samples from individuals who had undergone a dental extraction [207]. In orthotopic rectal CT26 adenocarcinoma, mouse models inoculated with 5×10^6 to 1×10^7 cells of *F. nucleatum* ATCC 23726 via tail vein injection, *F. nucleatum* could be identified in both tumour tissue and healthy control tissue within these mice[208]. In control mice without CRC *F. nucleatum* was not detected indicating that disruption due to CRC development was needed for the translocation via circulatory system.

1.3.2.2 Possible mechanistic relationship between Fusobacterium nucleatum and oncogenesis

The above information dose not demonstrate a direct role for *F. nucleatum* in CRC. However, there are experiments which support an active role of *F. nucleatum*. *F. nucleatum* binds to E-cadherin-expressing CRC cells causing signal transduction cascade through β-catenin leading to the expression of Wnt genes and increased proliferation[209]. Annexin A1 is a mediator of this FadA induced signalling which itself leads to Annexin A1 expression thus leading to a positive feedback loop[210]. Lipopolysaccharides (LPS) produced by *F. nucleatum* can bind to toll-like receptor 4 activating signalling to nuclear factor-kappab leading to the up regulation of the expression of miR-21[211]. The microRNA miR-21 down regulates the RAS GTPase RASA1 whose depletion can lead to the activation of MAPK signalling pathway and proliferation[211].

F. nucleatum can also apparently alter the tumour microenvironment of CRC. Mucosal colonization by *F. nucleatum* in CRC has been shown to promote tumour-infiltrating myeloid cells in *Fusobacterium*-associated colon tumour $Apc^{Min/+}$ mice[212]. Furthermore, *F. nucleatum* was seen to induce the expression of pro-inflammatory cytokines, including TNF, IL-6, IL-8 and IL-1β, via the NF-κB pathway in the mouse models[212]. This immunophenotype is reflected in RNA-seq data derived from Fusobacterium-associated human colon tumour samples[212]. The adhesin Fap2 of *F. nucleatum* binds to a human receptor known as TIGIT that is expressed on natural killer (NK) cells and other tumour-infiltrating lymphocytes[213].

This Fap2- TIGIT interaction inhibits the cytotoxic activities of these immune cells thereby protecting both *F. nucleatum* and CRC tumour cells[213].

Increasing evidence suggests that *F. nucleatum* may play a role in metastasis. Individuals with metastatic CRC have a higher relative abundance of *F. nucleatum* in their mucosa compared to individuals with non-metastatic CRC[214]. Absolute abundance as assessed by qPCR, showed that *F. nucleatum* cell numbers were higher in faecal samples of individuals with metastatic CRC than those with non-metastatic CRC[215]. *F. nucleatum* has been identified at metastatic sites[214,216]. *F. nucleatum* can upregulate Caspase activation and recruitment domain 3 (CARD3) protein which leads to activation of autophagy[214]. This activation of autophagy via CARD3 is a prometastatic pathway[214]. *F. nucleatum* was show to increase trans well migration and lung metastasis in mouse cell models[215]. This metastatic activity was show to be in part induced by the upregulation of the long-noncoding RNA Homo sapiens keratin 7antisense RNA (KRT7-AS) and keratin 7 (KRT7) through the NF-κB signalling pathway[215]. FAP2 dependant colonization of HCT116 cells increased the secretion of IL-8 and CXCL1 and promoted migration[217].

F. nucleatum colonization also appears to promote chemoresistance. The current model for this chemoresistance is that *F. nucleatum* induces the promotion of autophagy which protects against apoptosis[218,219]. This protective autophagy is induced via the TLR4–Myd88 signalling pathway and involves the reduction in the levels of the microRNAs miR-18a∗ and miR-4802 which in turn upregulates autophagy-related proteins including ULK1 and ATG7[218,219].

1.3.2.3 Interventions to control Fusobacterium nucleatum.

The preceding sections have detailed associative and causative relationships between *F. nucleatum* in CRC and other cancers. If one is to take the sum of evidence as sufficient to label it as a cancer promoting microbe, what steps can be taken to prevent *F. nucleatum*-attributable cancer incidence and deaths? Firstly, testing for *F. nucleatum* within subjects may aid in stratifying the population with regard to risk. Including *F. nucleatum* quantification to complement an immunochemical test improves diagnostic capabilities[220]. Secondly, it may be desirable to eliminate *F. nucleatum* from the microbiome of certain individuals. *F. nucleatum* has been shown to be sensitive to a range of antibiotics[221]. CRC xenograft mouse models treated with the antibiotic metronidazole led to a reduction of *Fusobacterium* load and was also linked to reducing cancer cell proliferation and overall tumour growth[216]. However using such a broad spectrum antibiotics may have unforeseen negative side effects due to the targeting other microbes. One solution could be to use predatory bacteria such as *Bdellovibrio bacteriovorus* which can kill *F. nucleatum*[222]. The utilization of bacteriophages to selectively eliminate *F. nucleatum* is also being explored[219,223]. A phage-guided biotic–abiotic hybrid nanosystem was developed which proved to be effective in eliminating intratumoural *F. nucleatum* in mouse models[219]. Furthermore this system was demonstrated to be more effective in reducing tumour growth than with chemotherapy compare to chemotherapy on its own[219].

1.3.2 The microbiota and cancer therapeutics

There is a growing arsenal of therapeutic strategy to treat cancer including immunotherapy and chemotherapy.

1.3.3.1 Immune Checkpoint Inhibitors

Immune checkpoints consists of a system of immunological pathways which modulate self-tolerance and the duration and amplitude of the immune response. These pathways ensure an appropriate response to foreign entities and prevent autoimmunity. Cancer cells may evolve to take advantage of checkpoints and evade immunosurveillance.

Clinical mmune checkpoints inhibitors are typically monoclonal antibodies which target cytotoxic T lymphocyte–associated antigen 4 (CTLA-4) or programmed cell death protein 1 (PD-1) or its ligand (PD-L1) thereby ablating the checkpoint. These ICI have proven to be a breakthrough in the development in cancer therapeutics.

There exists variability with regard to different types of cancers that are susceptible to ICI. ICI have proven effective in treating melanoma, non-small cell lung carcinoma, renal cell carcinoma, small cell carcinoma of the head and neck and urothelial carcinoma[224-228]. Furthermore, there is variation with regard to subtypes of cancer. For instance, ICI have proven effective for MSI high CRC. Resistance to ICI varies inter-individually. In the case of melanoma, a 26% -52% response rates to ICI exist depending on the ICI therapy administered[229]. A number of factors have been identified as modulators of response to immunotherapy including Tumour mutational burden (TMB) and PD-L1 expression [230,231].

The microbiota is now considered a factor which influences ICI efficacy. A seminal set of papers published in *Science* reported significant associations between microbiota features and treatment efficacy in patients undergoing immunotherapy[232-234]. Taxa such as *Akkermansia muciniphila*, *Bifidobacterium longum* and *Faecalibacterium prausnitzii* were found to be enriched in responders. However, these studies did not show a consensus microbial signal with respect to respond. Antibiotics have been reported to impair the efficiency of immune checkpoint inhibitors as measured by overall survival (OS) indicating a role of the microbiome in ICI efficacy[235-238]. These findings lead to the argument that antibiotic therapy should be restricted prior to immunotherapy[239].

There is mechanistic insight into how the microbiota may interact with the immune system thereby enhancing ICI efficiency. Data from both patient and mouse models provide evidence that the levels of short-chain fatty acids (SCFA) namely butyrate and propionate, reduces efficacy of CTLA-4 induced inhibition[240]. However with regard to anti–PD-1, higher levels of faecal SCFA is associated with longer progression-free survival[241]. The purine nucleoside inosine, which is produced via deamination of adenosine, has been demonstrated to augment the efficacy of ICI against CRC in mouse models[209]. Inosine is produced by various microbes such as *Bifidobacterium pseudolongum* and *Akkermansia muciniphila.* Both of these microbes have been found to be more abundant in individuals who responded to ICI relative to nonresponding cancer patients, with the latter found to be statistically significant[242]. Inosine systemic translocation via the colon is thought to be facilitated by perturbation in gut permeability caused by ICI. Inosine activates T helper 1 (TH1) in an adenosine 2A receptor (A2AR)–dependent manner leading to an enhancement of ICI therapeutics[242]. Faecal microbiota transfer (FMT) from ICI responder patients into GF mice has been reported to enhance ICI intervention[233]. Currently, clinical trials are been carried out with respect to the use of FMT as an intervention to augment ICI therapy in humans[243].

1.3.3.2 The microbiota and chemotherapy

The microbiota can biotransform and modulate the efficacy of chemotherapeutic compounds. Streptomyces inactivates doxorubicin by the reduction of the quinone ring of the anthracycline by NADH dehydrogenase[244]. In CRC mouse models, *Mycoplasma* has been demonstrated to inactivate gemcitabine via cytidine deaminase[245]. Mice which lack a microbiota show resistance to Cyclophosphamide[246]. Cyclophosphamide promotes the translocation of intestinal microbes including *Lactobacillus johnsonii*, *Lactobacillus murinus* and *Enterococcus hirae* which stimulate the production of type 17 T helper (TH17) cell and type 1 T helper (TH1) cell[247]. Microbes may also increase the toxicity of chemotherapy. Irinotecan is an anti-cancer prodrug which is converted into its active form in the liver. However, in the gut, β-glucuronidase expressing microbes convert irinotecan into the toxic compound SN-38[248].

Given the growing body of evidence indicating that a variety of tumours contain endogenous bacterial communities, microbiome based profiling of tumours prior to chemotherapeutic intervention has the potential to improve patient outcomes[249,250].

1.4 Mutagenesis by microbe: The role of the microbiota in shaping the cancer genome.

This 1.4 section has been published in the journal *Trends in Cancer*.

Authors:

Maurice Barrett, Collette K Hand, Fergus Shanahan, Thomas Murphy, Paul W O'Toole

Citation:

BARRETT, M., HAND, C. K., SHANAHAN, F., MURPHY, T. & O'TOOLE, P. W. 2020. Mutagenesis by microbe: The role of the microbiota in shaping the cancer genome. Trends in Cancer, 6, 277-287.

Maurice Barrett contributed to this work in the following ways:

- Primary author for this review.

Keywords

microbiota; microbiome; DNA damage; mutational mechanism; mutational signatures

1.4.1 Highlights

The literature describing the differences in microbiota features between individuals with cancer and matched controls has undergone dramatic recent expansion. Mechanistic models for how microbes promote cancer formation and progression are being developed and experimentally tested.

Microbes have been implicated in mutational mechanisms namely in the formation of DNA damage. These mechanisms include the production of crosslinking genotoxic colibactin by Escherichia coli or ectopic expression of activation-induced cytidine caused by Helicobacter pylori infection.

Developments in bioinformatics have allowed for the elucidation of the mutational mechanisms that act upon the cancer genome through oncogenesis, particularly by identifying mutational signatures.

Elucidation of microbe-associated mechanisms will allow for a more complete understanding of the forces behind the etiology of the cancer genome.

1.4.2 Abstract

Cancers arise through the process of somatic evolution fuelled by the inception of somatic mutations. We lack a complete understanding of the sources of these somatic mutations. Humans host a vast repertoire of microbes collectively known as the microbiota. The microbiota plays a role in altering the tumour microenvironment and proliferation. In addition, microbes have been shown to elicit DNA damage which provides the substrate for somatic mutations. An understanding of microbiota-driven mutational mechanism would contribute to a more complete understanding of the origins of the cancer genome.   Here we review the modes by which microbes stimulate DNA damage and the effect of these phenomena upon the cancer genomic architecture, specifically in the form of mutational spectra and mutational signatures.

1.4.3 Origin of the cancer genome and the role of the microbiota

Oncogenesis is driven by the Darwinian selection of somatic mutations (see Glossary) over time [251]. Mutations arise through the formation of genetic aberrations and their subsequent interactions with the DNA repair machinery and cell cycle related pathways including DNA synthesis[252]. Mutational mechanisms alter the DNA in distinguishing manners resulting in genetic patterns known as mutational signatures (Box 1).

Box1 | Mutational signatures

Specific mutational mechanisms produce characteristic patterns in the genome known as mutational signatures. Recent advances in mathematical modelling and bioinformatics have led to great improvements in our ability to identify mutational signatures from cancer genomic data. There are six defined classes of base substitutions: C>A, C>G, C>T, T>A, T>C and T>G [note: In accordance with the Catalogue of Somatic Mutations in Cancer (COSMIC) system, all substitutions are referred to by the pyrimidine of the mutated Watson-Crick base pair]. The incorporation of the 5’ and 3’ bases flanking the mutated base of the six originally defined classes gives an expanded classification system of 96 possible mutations. Utilizing this 96-class system as the framework and applying non-negative matrix factorization and model selection, with input from genomic data from 7042 cancer samples from 31 different cancer types, 21 mutational signatures were initially identified [253]. With the inclusion of more genomes for a heterogeneity of cancers, as well as the consideration of single base insertion/deletions and double base substitutions, the number of mutational signatures has expanded[254]. Currently, the number and type of mutational signatures characterised are as follows: 49 single base substitutions, 11 doublet base substitutions, four clustered base substitutions (DBS), and 17 small insertion and deletion (indels) mutational signatures[254]. Structural variants also occur in cancer genomes and they include insertions, deletions, inversions, balanced or unbalanced translocations, amplifications and complex rearrangements on a scale of >50 bp in size[255]. Efforts have also been made to define the signatures of these events [256]. Mutational signatures provide an insight into the mutational mechanisms that act on a cancer genome over time. Mutational signatures are typically displayed as histogram with the frequency of base substations (or indels or doublet base substitutions) with respect to the genomic context. SBS signature 1 is characterised by C>T transversions at methylated CpG sites within an NpCpG trinucleotide context. The putative mechanisms behind SBS signature 1 is spontaneous or enzymatic deamination of 5-methylcytosine to thymine. This newly formed thymine maybe base-paired with adenine during replication, provided DNA repair is not executed. Many mutational signatures described do not have a known aetiology.

The origin of mutations allows them to be classified into three categories, which is (i) Inherited genetic variants which lead to an increase in the risk of cancer development. (ii) Environmental factors, exogenous factors including UV light, tobacco smoking and diet that mutate the DNA are directly linked to cancer. (ii) Stochastic errors associated with DNA replication. These are seemingly inevitable random mutations which arise due to the intrinsic properties of DNA biology. Seminal work by Tomasetti and Vogelstein showed that about two-thirds of the mutations in the cancer genome originate from stochastic events [257,258].

Lung and cervical adenocarcinoma genomes harbour median values of 33% and 83% stochastic mutations respectively [257]. However, epidemiology evidence indicates that a high portion (~90%) are attributable environmental factors of cases, i.e. tobacco smoking and HPV infection, respectively. The manging of environmental factors is thus crucial is cancer prevention even though stochastic/replicative mechanisms are the major driver (See ref 3 for a more detailed discussion). However a complete catalogue environmental factors that contribute cancer risk is lacking. Note that a great number of known carcinogens promote oncogenesis by causing mutagenesis e.g. ultraviolet light, ethanol, tobacco smoke and radioactive substances.

The human microbiota is increasingly seen as an emerging environmental risk factor. The human microbiota is home to about 3.8 x 10^{13} bacterial cells and it is estimated that the collective metagenome of these bacteria encompasses about 100 times more genes than the human genome [8,10]. Although the majority of studies focus on bacteria, upon which this review is focussed, the human microbiota includes members from all 5 kingdoms of life as well as viruses. A large number of studies demonstrate that microbiota features are involved in the development and progression of a range of cancers. The term ‘oncobiome’ has been coined to describe the relationship between the microbiota and cancers[259]. However, oncobiome research has identified relationships that are primarily correlative rather than causative in nature. With regard to the putative mechanistic role that the microbiota has in cancer development, immune modulation in the form of inflammation caused by the microbiota is an intense area of research [260]. Effort has also been made in defining the role of the microbiota in cell proliferation [261].

The microbiota is known to be involved in a diverse assortment of mutational mechanisms (Table 1). Known variation in cancer risk due to unknown environmental factors could be explained in part by variations in the ability of the microbiota of individual subjects to induce DNA-damage and thus somatic mutations. Here we describe the current state of knowledge on microbes and their ability to compromise the stability of the human genome ultimately leading to cancer.

Table 1. Microbe-Associated Mechanisms and Genomic Consequences

Source	Involvement of microbiota features	Key role in a mutational mechanism	Postulated effected on cancer genomic landscape	Reference
Activation-induced cytidine deaminase (AID)	*Helicobacter pylori* infection cause ectopic expression of AID	Cytosine deamination at specific motifs	Mutational signatures SBS84 and SBS85	[254,262]
Acetaldehyde	Various inhabitants of produce ethanol and are capable metabolic act on it to produces acetaldehyde	N2-ethylidenedeoxyguanosine, Guanine- guanine intrastrand crosslinks	GG-to-TT base substitution. Mutational signature DBS2	[263]
Colibactin	Expressed by *Escherichia coli* containing a *pks* island	Adenine – adenine intra-strand crosslinks, Double strand breaks,	DSBs at an AAWWTT pentanucleotides motif. Mutational signatures SBS28 and SBS41	[264]
Cytolethal distending toxin (CDT)	Produced by various Gram-negative bacteria including enteropathogenic *Escherichia coli*, *Campylobacter* species, *Shigella* species and *Haemophilus ducreyi*	Single strand breaks and Double-strand breaks	Infidelity of DNA repair can lead to structural variants such as indels	[254]
Disruption of DNA mismatch repair	*Helicobacter pylori* and Enteropathogenic *Escherichia coli* can disrupt mismatch repair	Deletion of MMR proteins	Microsatellite instability, Mutational signature SBS6, ID1 and ID2	[253,265,266]
Dinitrogen trioxide	Metabolic activities of the microbiota can produces precursors to N203 e.g. denitrifying bacteria	Nitrosative deamination	Various base substitutions e.g. Adenine nitrosative deamination to Hypoxanthine can lead to T>A substitution	[267,268]

Hypobromous acid	Eosinophil’s produce Hypobromous acid. The microbiota can influence eosinophic biology	8-bromoguanine	G > T primarily but also G > C, G > A, and delG	[269]
Hypochlorous acid	HOCL is produce by Neutrophils. The microbiota can influence neutrophil inflammatory status	Formation of 5-chlorocytosine (5ClC), formation of malondialdehyde	C>T, G >A, G>T substitutions	[270,271]
N-nitroso compounds (NOCs)	Microbes play a role in the production of nitrosating agents and produces biogenic amine	Alkylated DNA base	Various base substitutions e.g O6-methylguanine (O6-MeG) can cause a G(C)>A(T) transition	[272]
Reactive oxygen species	Various metabolic activities	Oxidative Base Lesions	G to T transversion, SBS Mutational signatures 18 and 36	[273]
4-hydroxy-2-nonenal	*Enterococcus faecalis* induces the bystander effect via polarising marcophages. Polarised marcophages produces 4-hydroxy-2-nonenal	Exocyclic HNE-DNA adducts	Chromosomal instability	[274]

In this review we described the microbiota influences on genome integrity through (i) direct DNA damage, (ii) immune cell induced DNA damage, (iii) dietary interaction, and (iv) disruption to the DNA damage response.

1.4.4 Direct DNA Damage

Members of the microbiota can produce proteins, molecules and secondary metabolites that can directly cause DNA damage. These products can interact directly with the host DNA thereby mutating it.

1.4.4.1 Colibactin

Escherichia coli is classified into 4 phylogenetic groups, A, B1, B2, and D. About 30–50% of *E. coli* strains identified in stool microbiota of individuals from high-income nations belong to group B2. Within the B2 group, 35% of isolates possess genomic islands known as *pks* (for polyketide synthase) islands[275]. The 54-kb *pks* island is a biosynthetic gene cluster encoding for a non-ribosomal peptide synthetase (NRPS)–polyketide synthase (PKS) hybrid gene cluster, which encodes for colibactin [276]. Colibactin can cause Double-strand breaks (DSB) in mammalian DNA thereby promoting genome instability and an increase in mutation rate [277,278]. Note, how colibactin is transported to from the outside all the way to the nucleus is currently unknown. The pks+ *E. coli* strains are over-represented in the gut of individuals with colorectal cancer, being detected at a rate 20% in the mucosa of healthy individuals but 55%-67% in patients with colorectal cancer (CRC) [279,280]. Furthermore, pks+ *E. coli* was disproportionally frequently identified in subjects with familial adenomatous polyposis (FAP) compared to healthy controls [281]. Monocolonization of azoxymethane (AOM)–treated IL10−/− mice with pks+ *E. coli* promoted tumorigenesis, while challenge with strains lacking *pks* reduces the frequency of tumorigenesis [279].

Colibactin crosslinks directly with DNA through an electrophilic cyclopropane moiety ‘warhead’ [282]. Liquid chromatography–mass spectrometry-based methodologies have identified that colibactin alkylation of DNA via the cyclopropane warhead resulted in adenine-colibactin adducts [283,284]. This phenomenon was identified in both HeLa cells and in mouse models [284]. Colibactin can also induce DNA inter-strand cross-links and activation of the DNA damage response including Fanconi anemia DNA repair [285]. Recent structural analysis revealed that colibactin contains two conjoined warheads enabling its ability to cause DNA crosslinks [286]. Double strands breaks are not believed to be a direct consequence of colibactin activity but rather occur due to replication stress caused by DNA cross-links [285]. Recent sequencing analysis of sites of colibactin induces DSBs revealed that these DSBs occurred at AT-rich regions and in particularly at the pentanucleotides motif containing the AAWWTT[264]. Single nucleotide variants at the AAWWTT were found to be enriched in a number of cancers including CRC and stomach cancer compared with a WWWWW motif. Two mutational signatures were found to be link with the AAWWTT colibactin motif, SBS28 and SBS41[264]. Mutational signature SBS28 has been associated with POLE mutation while Mutational signature SBS41 has no know etiology.

1.4.4.2 Cytolethal distending toxin (CDT)

The cytolethal distending toxin (CDT) is produced by an array of gram-negative bacteria within the gamma and epsilon classes of the phylum Proteobacteria[287]. It is a heat-labile exotoxin whose properties lead it to be classified as a both a cyclomodulin and a genotoxin. The proteobacteria that can produce CDT are sub-dominant members of the human gut microbiota.

CDT is a heteromultimeric protein comprised of three subunits, CdtA, CdtB and CdtC which are encoded within a bacterial single operon [288,289]. Subunits CdtA and CdtC function to allow delivery and internalization of CDT into target cells[289]. CdtB shares sequence, structural and functional homology with DNase I and is highly conserved among bacteria [290,291]. Furthermore, nuclear localization signals have been identified in CdtB proteins [292]. Studies with ApcMin/+ mice that are genetically susceptible to small bowel cancer found that a *Campylobacter jejuni* strain harbouring the CDT operon promoted colorectal tumorigenesis compared to treatment with non-CDT bacterial controls, while mutation of the cdtB subunit attenuated this phenomenon [293]. CdtB has been shown to promote DSB in *vitro* and in *vivo* [290,294,295]. However, the current model of CdtB activity holds that CdtB acts in a dose-dependent manner and tends not to induce double strand breaks directly [296]. At low to moderate doses, CdtB causes single strand breaks (SSB) which are addressed by Single-strand break repair (SSBR)[297]. If CDT-induced SSBs are not addressed before replication or occur during replication, they may cause a stalled replication fork [296,297]. At high doses, CDT can induce DSB directly by two cuts to the DNA backbone that are juxtaposed to each other [296] .

1.4.4.3 Reactive oxygen species

Reactive oxygen species (ROS) are a chemically reactive family of molecules containing oxygen which include the highly reactive hydroxyl radical (OH−), superoxide radical (O2−), and non-radical hydrogen peroxide (H_2O_2). Reactions of ROS with DNA generates oxidative DNA base lesions. To date, more than 30 oxidative DNA base lesions have been identified(Box 2)[298].

Microbiota activity is known to produce reactive oxygen species through varied means. For example, primary bile acids, cholic acid (CA) and chenodeoxycholic acid; (CDCA) are synthesised by the liver and are secreted into the small intestine from the gall bladder. A small proportion of these bile salts are transformed into secondary bile salts by the gut microbiota. These secondary bile salts are thought to be involved in the production of ROS [299].

Hydrogen sulphide (H_2S) is produced by the metabolic activity of colonic bacteria including taurine desulfonation by *Bilophila wadsworthia*, cysteine degradation by *Fusobacterium nucleatum* and sulfonate degradation by sulfate-reducing bacterium such as *Desulfovibrio desulfuricans*. Increased relative abundance of such bacteria has been linked to CRC development [300,301]. Evidence suggests that H_2S production leads to DNA damage partly due to ROS generation [301,302].

Box 2 | Oxidative DNA Base Lesions

Guanine has the lowest redox potential of the native bases and is thus the most readily oxidised. Two common oxidative base lesions which are generated by the oxidation of Guanine include 8-oxo-7,8-dihydro-2’-deoxyguanosine and 2,6-diamino-4-oxo-5-formamidopyrimidine (FapyG) which occur at an estimated rate of 1000–2000 and 1500–2500 per cell/per day in normal tissues, respectively[303]. Furthermore, the occurrence and the mutagenicity of these oxidative DNA base lesions vary considerable. For example, 7,8-

dihydro-8-oxo-guanine is about four times as mutagenic and four times more frequent in its occurrence than 7,8-dihydro-8-oxo-adenine[303,304]. Replication of DNA containing 8-oxo-7,8-dihydro-2’-deoxyguanosine and 2,6-diamino-4-oxo-5-formamidopyrimidine (FapyG) are shown to induce G:C to T:A (C >A) and G:C to T:A (C >A) respectively[305].

The nucleobases within the cellular nucleotide pool may also undergo oxidation. Misincorporation of these nucleoside triphosphates can induce mutations. The two major products of nucleotide pool oxidation are 8-hydroxy-2′-deoxyguanosine 5′-triphosphate (8-OH-dGTP) and 2-hydroxydeoxyadenosine 5′-triphosphate (2-OH-dATP). 8-OH-dGTP has been demonstrated to induce A:T to C:G transversions when introduced into COS-7 mammalian cells[306]. *In vitro* analysis using HeLa cell extract showed that 2-OH-dATP within the nucleotide pool can led to G·C to A·T (C>T) transitions and G·C to T·A(C>A)[307].

Mutational signatures 18 and 36 have been suggested to be attributed to reactive oxygen species. Mutational signature 36 has been specifically attributed to ROS in the context of MUTYH-Associated Polyposis (MAP) syndrome [273]. MAP syndrome is defined by biallelic germline mutation of MUTYH gene and is a colorectal polyposis which predisposes individuals to CRC. MUTYH DNA glycosylase is coded by the MUTYH gene and functions to prevent 8-Oxoguanine-related mutagenesis by scanning the newlysynthesized daughter strand in order locate and remove incorporated adenine paired with 8-Oxoguanine[305].

1.4.4.4 *Dinitrogen trioxide and nitrosative deamination*

Nitrosative deamination is deamination mediated by dinitrogen trioxide (N_2O_3, nitrous anhydride). In this phenomenon, dinitrogen trioxide can react with nucleotides and induce deamination by nucleophilic aromatic substitution. These events are mutagenic because the resulting deaminated bases may be read incorrectly if not repaired[268].

Dinitrogen trioxide can be generated from the autooxidation of nitric oxide (NO-) or the condensation of nitrous acid (HNO_2)[308]. GIT microbes can produce endogenous nitric oxide and/or nitrous acid by 4 mechanisms, that is, (i) The hemethiolate monooxygenase, nitric oxide synthase (NOS), oxidises L-arginine (Arg) to produce nitric oxide. [309] (ii) Denitrification of nitrate (NO_3^-) to nitrogen (N_2), which is an important part of the nitrogen cycle and is carried out by denitrifying bacteria and plants. During denitrification, nitric oxide is produced by one-electron reduction of nitrite (NO_2^-) by heme or Cu-containing nitrite reductases[267]. (iii) Respiratory nitrite ammonification (also referred to as dissimilatory nitrate reduction to ammonium)[267]. (iv) Acidic non-enzymatic reduction of nitrite to NO which is driven by lactic acid bacteria such as lactobacilli and bifidobacteria[310].

1.4.5 Immune cell induced DNA damage

The microbiota and immune system closely interact from the early stages of human development. In this section we review mechanisms by which the microbiota can influence immune cells to behave in a genotoxic manner.

1.4.5.1 Hypochlorous acid (HOCl) production

Neutrophils, which are a type of polymorphonuclear leukocyte, accumulate at sites of injury with the primary function of promoting inflammation. Neutrophils produce a potent antimicrobial known as hypochlorous acid (HOCl) which is produced by myeloperoxidase using as substrates the chloride ions and hydrogen peroxide (H_2O_2) produced by NADPH oxidase [311]. HOCl is highly reactive and readily interacts with DNA. HOCl has been shown to cause a cytosine to 5-chlorocytosine (5ClC) conversion [270]. This is in turn can cause a C to T transition during replication.

In addition, HOCl can induce the peroxidation of lipids leading to the formation of malondialdehyde (MDA). Studies in both cellular and animal models found that such a production of MDA can lead to a significant increase in the formation of 3-(2-deoxy-β-D-erythro-pentofuranosyl)pyrimido[1,2-α]purin-10(3H)-one (M1dG) , a damaged guanine. [271]. M1dG adducts are mutagenic causing G>T and G >A substitutions.[312]

The microbiota is now known to be a modulator of neutrophilic biology[313]. A recent study in a mouse model demonstrated that neutrophil pro-inflammatory activity correlates positively with neutrophil ageing while in circulation[314]. Furthermore the study found that the microbiota regulates neutrophil ageing by Toll-like receptor and myeloid differentiation factor 88-mediated signalling pathways[314]. A depletion of the microbiota was mirrored in the number of aged neutrophils and an improvement in inflammatory disease.

1.4.5.2 Hypobromous acid production

Eosinophils are granular leukocytes with a multifunctional role in immune biology. Eosinophils secrete eosinophil peroxidase which catalyzes the formation of hypobromous acid (HOBO) from hydrogen peroxide and halide ions (Br^-) in solution. HOBO can also be produced by reaction of HOCl with Br- ions. Like HOCl, HOBO is an oxidant and functions to oxidize the cellular components of invading pathogens; however excess production of HOBO can also lead to host damage including DNA damage, namely the formation of 8-bromo-2′-deoxyguanosine and 5-bromo-2′-deoxycytidine. A SupF forward mutation assay in human cells found that the prominent mutation induced was G >T mutation but HOBO also induces G>C, G>A, and delG [269].

1.4.5.3 Activation-induced cytidine deaminase

Activation-induced cytidine deaminase (AID) is a member of the cytidine deaminase family of enzymes with a role in somatic hypermutation. Immunohistochemistry identified the ectopic overexpression of AID in inflamed tissue derived from patients with Crohn's disease and ulcerative colitis as well as colitis-associated colorectal cancers [315]. The expression of AID in colonic epithelial cell lines induced an increase in the mutation rates in these cells [315]. Knock-out of AID in IL10 null mice attenuated the mutation rate in their colonic cells and also inhibits CRC development[316]. Inflammation seems to be key to this aberrant activity. *H. pylori* infection, which is known to induce inflammation, promotes ectopic expression of AID in non-tumorous epithelial tissues [262]

Whole genome analyses in chronic lymphocytic leukaemia revealed that the activity of AID may produces two types of substitution pattern (i) a 'canonical AID signature' characterised by C to T/G substitutions at WRCY motifs near active transcriptional start sites and (ii) a 'non-canonical AID signature' characterised by A to C mutations at WA (W=A or T) motifs occurring genome-wide in a non-clustered fashion [317]. These mutational processes have been assigned to mutational signatures SBS84 and SBS85[254].

1.4.5.4 By-stander effect and Enterococcus faecalis

Enterococcus faecalis is known to promote CRC oncogenesis in interleukin 10 -/- mice [318]. *E. faecalis* can promote the bystander effect which leads to double-stranded DNA breaks, tetraploidy and chromosomal instability. In this model, *E. faecalis* production of extracellular superoxide induces polarization of macrophages to an M1 phenotype [319-321]. In turn macrophages produce 4-hydroxy-2-nonenal (4-HNE), a diffusible breakdown product of ω-6 polyunsaturated fatty acids whose expression in this context is dependent on

Cyclooxygenase-2[274,322]. Primary murine colon epithelial cells exposed to polarized macrophages or purified 4-HNE undergo transformation [323].

1.4.6 Dietary interaction

The diet of the host and the gut microbiota are inextricably linked. GIT bacteria depend almost exclusively on the host diet for their nutritional substrates (a restricted number of taxa can metabolize mucins and glycoproteins) and indeed the composition of the microbiome is correlated strongly with diet. Diet is a key modulator of cancer risk. In the cases described below, microbiota-diet interactions lead to the formation of genotoxic compounds capable of mutating the host genome.

1.4.6.1 N-nitroso compounds (NOCs)

NOCs, such as nitrosamines and nitrosamide, are known to be potent carcinogens. NOCs are formed by the nitrosation of secondary amines and amides via nitrosating agents, such as N_2O_3 and N_2O_4 [324]. NOCs can be found in foods such as processed meats, smoked/cured fish and German beer[325]. Additional compounds such as nitrate and nitrite which are precursors to nitrosating agents can be found in food including vegetables which may account for 50–70% of an individual’s intake of nitrate and nitrite [326]. Endogenous NOCs are also formed and in many cases, this is because of the activities of microbes. Firstly, bacteria produce nitrosating agents (See Dinitrogen trioxide and nitrosative deamination). Further amines and amides are produced by bacterial decarboxylation of amino acids [326]. Heme has been suggested to catalyse the formation of NOCs[327]. Inhibitors of nitrosation are ingested as part of a diet and include vitamin C, vitamin E and polyphenols[328].

The activated form of NOCs induce a number of methylated DNA adducts, of which over 12 are known, via SN1-nucleophilic substitution[329]. These alkylated DNA bases can be mutagenic if not repaired before replication[272]. SBS mutational signature 11 has been linked to the mutagenic activity of alkylating agents [330].

1.4.6.2 Acetaldehyde

Alcohol is classified as a Group 1 carcinogen (carcinogenic to humans). Worldwide, 3.6% of all cancer deaths and 3.5% of all cancer cases are attributable to alcohol consumption[331]. Ethanol (C_2H_5OH), the psychoactive ingredient in alcoholic beverages, is believed to be the major causative compound of cancer in alcoholic beverages.

Ethanol is introduced into a catabolic pathway where it is broken down and the metabolites expelled via the urinary system. Ethanol is first metabolized by alcohol dehydrogenase (ADH), cytochrome P4502E1 (CYP2E1) and catalase thereby forming acetaldehyde (ethanal). Acetaldehyde is further oxidised by aldehyde dehydrogenase to produce acetate. Aldehydes cause DNA damage in the form of double strand breaks and the Fanconi anaemia pathway is responsible for the repair of this damage [332]. Aldehydes has been demonstrated to cause intrastrand crosslink between adjacent guanine bases[263]. This can lead to the mutagenic event of GG>TT double base substitution which is a characteristics of Mutational signature DBS2 [254,263].

Bacteria can not only produce ethanol but also break it down into acetaldehyde. Oral taxa are known to be able to produce acetaldehyde from ethanol or glucose [333]. In addition, gut microbes can also produce acetaldehyde from sugars [334]. Indeed there have been reports of bacterial autobrewery syndrome (intoxication by ethanol formed by fermentation by microbes in the gut) in which a strain of *Klebsiella pneumoniae* was implicated [42]. This strain was also strongly associated with non-alcoholic fatty liver disease and fatty liver disease symptoms in a mouse model. Mutational signature 16 has been link to alcohol consumption [335].

1.4.7 Disruption to the DNA damage response

Human DNA experiences repeated events of DNA damage throughout the cell cycle. The cell has a complex network of systems whose purpose is to ensure the fidelity of DNA. Known as the DNA damage response, this cellular system is responsible for detecting DNA damage, signalling its presence, promoting DNA repair cell cycle checkpoint and/or apoptosis.

The mismatch repair mechanism is responsible for addressing base-base mismatches and insertion/deletion mispairs generated during DNA replication and recombination[336]. Enteropathogenic *Escherichia coli* was found to promote the depletion of MSH2 and MLH1 proteins, which are crucially important for mismatch repair in cell models[265]. This phenomenon was found to be dependent on the bacterial type-3 secretion effector EspF[265]. Furthermore, mitochondrial targeting of EspF was necessary for this activity. Colonic epithelial cells infected with

Enteropathogenic *E. coli* display an increased mutation rate particularly in microsatellite DNA sequences.

The human gastric pathogen *Helicobacter pylori* also inhibits the expression of MMR gene expression, in part through the modulation of miRNAs [266,337].

Mutational signature 6 is characterised by C>T transitions at an NpCpG trinucleotide context [253]. This mutational signature is associated with small indels (usually 1-3bp) at nucleotide repeats. This indel pattern is equivalent to phenomena known as microsatellite instability. Microsatellite instability is caused by aberrations in the DNA mismatch repair (MMR) machinery. The origin of MMR deficiencies is genetic and/or epigenetic alterations in MMR genes. Microsatellite instability occurs in 15% of CRC genomes; 3% are associated with Lynch syndrome while 12% are associated with sporadic CRC[338]

1.4.8 Mutational signatures as a tool to study the effect of microbes on the human genome

Multiomic experimental designs are supremely placed to delineate the relationship between the microbiota and the architecture of the cancer genome. Population studies in which both cancer genomic and the adjacent microbiome are studied can provide information on the relation between the cancer genetic architecture and its microbiota. However, therein lies a fundamental issue with this type of design. Cancer can take several years to form and mutational mechanisms act at different time spans of the natural history. Furthermore, composition of the microbiota is somewhat dynamic. Thus, a snap shot of the microbiota may not be wholly related to the mutational signatures identified. A prospective study where individual's microbiota are taken at a per-transformation may allow for more direct comparisons between the microbiota and pre-transformation mutational mechanisms. Additionally, individuals with pre-cancer legions such as Barrett's oesophagus may be prime candidates to study due to their increase propensity develop cancer. Studying cancer heterogeneity and evolutionary dynamics can allow for the identification of the timing of mutational mechanisms. Additional recent advancements have allowed for mutational signature extraction from non-cancerous tissue thus allowing elucidation of microbial associated mechanisms prior to transformation [339]. Experiments in which a microbe or a community of microbes are grown in the context of a model such as a cell line or organoids would allow to eliminate confounders and make more direct correlations. Dziubańska-Kusibab et al used model cell lines exposed to colibactin and to identify DNA sequence targets of colibactin. Furthermore this target was cross-referenced with mutational signatures derived in population cancer genomic data to find asssiocateded mutational signatures (See colibactin section)

1.4.9 Concluding Remarks

Cancer prevention is relatively under-researched when compared to therapeutic development, with only 2 to 9% of funding put towards this area [177]. A high proportion of cancer cases and cancer deaths could be avoided through modification of environmental risk factors. About 42% of cancer incidences in the US are estimated as being attributable to modifiable risk factors - this figure is also reflected in the UK population [178]. Evidence is building in favour of the microbiota as an environmental modulator of cancer risk. We outlined the multitude of ways that the metabolic activities of members of the human microbiota can lead to mutations.

Our ability to modulate the microbiota is improving steadily, featuring diet, antibiotics, phage therapy, faecal microbiota transplantation (FMT), prebiotics, probiotics and Live Biotherapeutics[340]. Thus one could plausibly develop strategies to alter the structure of an individual's microbiota in order to reduce its mutagenic potential (see Outstanding Questions).

In order to make informed decisions on therapeutic interventions, a complete catalogue of microbial-associated mutational mechanisms is required. Furthermore, the relative impact of each mutational mechanisms on the cancer genome need to be delineated. Microbial-associated mutational mechanisms which have both been found in a wide range of cancers as well as contributing to a great number of mutations will take priority when deciding what mechanisms need to be addressed first.

We propose to leverage advancements in cancer genomics, namely in the form of mutational signatures, to associate microbes to mutational mechanisms. These can provide qualitative and quantitative information on the mutagenic effect that microbes undoubtedly have.

It is possible that certain aspects of the microbiota activity protect against mutagenesis and cancer. These potential mechanism need to be elucidated to enable the harnessing the microbiota as prophylactic agents.

1.4.10 Acknowledgments

This work was supported by The Health Research Board of Ireland under Grant ILP-POR-2017-034, and by Science Foundation Ireland (APC/SFI/12/RC/2273) in the form of a research centre, the APC Microbiome Institute.

1.4.11 Outstanding Questions Box

- What is the complete repertoire of modes by which the microbiota promotes DNA damage or compromises DNA integrity?
- What is the exact mutational mechanism by which microbes elicit mutations?
- What are the mutational signatures which result in a microbiota-associated mutational mechanism?
- How does the mutagenic potential of the microbiota vary within the population? This would need to take into consideration epidemiological factors such as age, diet, genetics and other modifiers/risk factors.
- How does this variation in the mutagenic capacity of the microbiota contribute to cancer risk?
- What proportion of cancer genomes have microbial influence in their formation? Further, in cancer genomes with microbial influences, what is the quantitative impact it has (frequency per Mbp/ overall abundance)?
- How might the microbiota protect genome stability and prevent cancer?
- What are the necessary interventions that would be required in order to address these microbiota associated mutational mechanisms?

1.4.12 Glossary

Base substitutions: A type of mutation in which one base is replaced by another in DNA.

Chromosomal instability: A phenomena which leads to alterations in chromosome number and/or structure.

DNA adduct: Formed via the addition of a chemical moiety to a DNA base

DNA alkylation: The addition of an alkyl group (C_nH2_{n+1}) to a DNA base

DNA crosslinking: Formation of covalent bonds between two nucleotides. This bond can be formed between nucleotides on the same DNA stand (intrastrand crosslinks) or different strands (interstrand crosslinks)

DNA deamination: The removal of an amino group from a DNA base.

DNA repair: A diverse collection of pathways with the purpose of addressing DNA damage and maintaining genome stability.

Double-strand breaks: This is where both strands of DNA which are juxtaposed to each other

Environmental risk factor: A thing or process which is not inherited that increases the risk for a particular disease.

Microbes: Microorganisms including bacteria, fungi, protists and virus. Usually exist as a single cell organism.

Microbiome: The combined genetic material of the microorganisms in a particular niche.

Microbiota: The collection of organisms in a niche.

Mutational mechanism: Biological phenomena which lead to the generation of mutations. Usually involving DNA damage, DNA repair and DNA replication.

Mutational signature: The characteristic DNA pattern of mutations produced by a mutational mechanism.

Oncogenesis: The transformation of a normal cell into a cancer cell.

Oxidative Base Lesions: DNA Bases that occur due to a reaction with Reactive oxygen species

Somatic mutation: A mutation which occurs in a somatic cell and is thus not heritable.

1.4.13 Colibactin continued

A number of informative papers were released after the publication of *Mutagenesis by Microbe: the Role of the Microbiota in Shaping the Cancer Genome*. The aim of this section is to update and complete the discussion on colibactin for this thesis. In the study by Pleguezuelos-Manzano et al, human intestinal organoids were co-cultured with pks+ *E. coli* and a clbQ knockout strain of pks+ *E. coli* (thus unable to produce colibactin), which was used as a negative control[341]. After a 5-month period whole genome sequence was performed on clones from each arm of the study. Organoids which were exposed to colibactin contained higher numbers of single base substitutions. The genomes of these organoids featured two mutational signatures, a single-base substitution signature and a small indel signature denoted SBS-pks and ID-pks respectively. SBS-pks was characterised by T > N substitution within an ATA, ATT and TTT context (whereby the middle base is the one undergoing substitution). It was found that A was highly represented 3 bp upstream from the mutated SBS-pks T > N site. Moreover this SBS-pks displayed a transcriptional strand bias indicating that the transcription-coupled nucleotide excision repair maybe involved with the repair of colibactin lesions. The ID-pks was characterised by single T deletions at T homopolymers with an enrichment of adenines immediately upstream of the indel containing poly-T stretch. The length of the A polymer was inversely proportional to the T polymer length. Larger indels were also described within the same sequence context of the ID-pks.

Both SBS-pks and ID-pks signatures were found enriched in the cancer genomes of CRC. Indeed, these identified signatures found to occur in recognised CRC driver genes including APC[341]. Other work by Lee-Six et al described a mutational signature appearing in healthy crypts genomes including signatures which correlated with each other denoted SBS-A and ID-A signatures[339]. These signatures were inferred to occur early in life of an individual[339]. SBS-A and ID have also been identified in non-neoplastic IBD-Affected crypts[342]. Note that pks+ *E.coli* occur more frequently in IBD than healthy individuals, 40% versus 20% respectively[279]. SBS-pks and ID-pks show high levels of similarity with SBS-A and ID-A. SBS-pks and ID-pks seem be present early in the evolution of the CRC genome. Yang et al demonstrated that pk+ *E.coli* promoted colorectal carcinogenesis in two mouse models harbouring a complex microbiota[343]. Treatment of mice colonised with pks+ *E.coli* with anti-TNF therapy lead to a decrease in the transcription of the clb island genes and attenuated carcinogenesis[343]. However, mice treated with anti-TNF therapy co-housed with untreated mice no longer displayed protection from CRC development. Due to the coprophagic activity of mice, the ability of anti-TNF therapy to attenuate CRC oncogenesis was inferred to be via the microbiota. Further supporting this was that the findings transplantation of cecum microbiota from anti-TNF treated mice to germ-free Apc min/+ mice protected them from CRC development[343].

1.5 Colorectal cancer

Globally, colorectal cancer (CRC) is the second most common cause of cancer and the second highest cause of cancer mortality[172].In 2018 there were ~1.8 million new diagnosis of colorectal cancer and 881,000 deaths worldwide[172]. CRC is thus the most impactful cancer covered in this thesis in terms of the above metrics. Furthermore, the relationship between CRC and the gut microbiota is the most explored cancer-microbiome interaction. Histologically, more than 95% of CRCs are carcinomas (derived from epithelial cells) while colorectal lymphomas, sarcomas, carcinoids, melanomas and squamous cell carcinomas occur much less frequently[344-348].

The aetiology of colorectal cancer is multifactorial involving, environmental, heritable and stochastics factors. Approximately 60–65% of CRC cases arise sporadically i.e. no known family history, inherited cancer syndrome gene or other inherited genetic mutations.

1.5.1 Evolution of CRC

The development of cancer can occur through 3 described pathways; (1) the conventional adenoma-carcinoma sequence (2) the serrated pathway and (3) inflammatory pathway[349]

The Adenoma-carcinoma sequence is seen as the conventional mode of CRC oncogenesis because 85–90% develop from adenomas[349]. Somatic mutation in the adenomatous polyposis coli (APC) tumour suppressor leading to its inactivation is generally regarded to be the earliest mutation and initiates the adenoma-carcinoma sequence[350]. Inactivation of the APC gene leads to over activation of the Wnt/β-catenin signalling pathway which in turn promotes cellular proliferation[351]. Common somatic mutations acquired subsequently include KRAS, SMAD4 and TP53[350]. The development of chromosomal instability (CIN) occurs frequently along the Adenoma-carcinoma lineage with ~70% occurrence in all sporadic CRC[352].

Approximately 10-15% of CRC arises from serrated polyps[353]. Serrated polys can themselves be furthered histologically classified into traditional serrated Adenoma, sessile serrated adenomas, hyperplastic polyps and mixed polyp[353]. Somatic mutation in BRAF is considered a crucial early initiator of serrated polys[354]. This BRAF mutation leads to constitutive activation of the MAPK signaling cascade and thus aberrant cellular proliferation[355]. The epigenetic molecular phenotype 'CpG island methylation phenotype' (CIMP-H) frequently develops in serrated polys[353]. CIMP-H leads to the silencing of a number of tumour suppressor genes including CDKN2A and the mismatch repair (MMR) gene MLH1. Silencing of MLH1 leads to deficiency in the mismatch repair machinery causing microsatellite instability (MSI).

The Inflammatory pathway of colorectal cancer involves chronic inflammation in the colon of individuals with inflammatory bowel disease (IBD), in particular ulcerative colitis (UC). In a recent population-based cohort study, a hazard ratio of 1·66 (95% CI 1·57–1·76) was calculated[356]. This type of CRC is referred to as colitis-associated CRC (CA-CRC). Polyp formation is not described in this mode of carcinogenesis. Instead the pathogenesis proceeds through to indefinite dysplasia, low-grade dysplasia, high-grade dysplasia and eventually CA-CRC. CA-CRC accounts for less than 2% of all cases of CRC[349]. CA-CRC occurs on average earlier (younger age) in individuals compared with sporadic CRC, 50–60 years versus 65–75 years[357]. CA-CRC is more commonly ‘synchronous’, that is, two primary cancers appearing in the same tissue within 6 months[357]. Mutations in TP53 occur early in the CA-CRC process evident from its clonal ratio and due to the fact it is identified in precancerous neoplasms and non-neoplastic mucosa[342,358-361]. CA-CRC has a higher mutational burden relative to sporadic CRC[361].

1.5.2 Anatomical subtyping of CRC

Albeit originating from the one organ, that is the colon, CRC can be subdivided into two or three types based on anatomical site. In the three-way split, the sections are defined as; proximal colon (caecum, ascending colon, hepatic flexure and transverse colon), distal colon (splenic flexure, descending colon and sigmoid colon) and rectum. Embryologically speaking, the proximal colon develops from the midgut while the distal colon and rectum develop from the hindgut[362]. There is demographic variation in the distribution of CRC along the colon. Proximal cancer is the most frequent subtype observed in the western population, with proximal, distal, rectal having a proportional prevalence of 40%, 22% and 29% respectively (according to US figures)[363]. However, this trend is not globally consistent. In Korea, rectal cancer is the most prevalent, with proximal, distal, rectal having proportional prevalence of 22%, 26% and 52% respectively. There are higher incidences of proximal cancer seen women versus men, 34% versus 25% respectively, in European cohorts[349]. Smoking is associated with increased risk of proximal CRC and rectal CRC but not an increase in distal CRC[364]. Serrated polyps and colitis-associated CRC appear more frequently in the proximal colon.

CRC has been classified based on its molecular characteristics. From this, form consensus molecular subtypes (CMS) have been described[203]. These CMS vary by anatomical prevalence with, CMS1 and CMS3 more prevalent in the proximal colon, while CMS2 and CMS4 are more prevalent in distal and rectal CRC[365]. With regard to therapeutics, proximal CRC is linked to a poorer prognosis in the context of metastasis and these cases more resistant to anti-EGFR therapy[366,367]. However, because that MSI is more inprevalent proximal CRC mean that immune checkpoint inhibitors are more effective in proximal CRC[368].

1.5.3 Inherited risk of CRC

Inherited alterations (genetic or epigenetic in nature) contribute significantly to CRC, risk with calculations for the heritability of CRC ranging from 12% to 40%[369,370]. Indeed 25% of colorectal cancer cases show a family history of CRC which points to the influence of heritabe factors. Furthermore 3–5% of colorectal cancer cases are due to cancer-prone syndromes known as hereditary colorectal cancer syndromes[371]. These cancer syndromes are caused by highly penetrant germline variants which increases dramatically susceptibility to CRC. For example, Lynch syndrome is caused by mutations in mismatch repair genes MLH1, MSH2, MSH6 and PMS2. The life-time risk of developing CRC individuals with Lynch syndrome varies up to 46%[372]. GWAS have also identified a number of less penetrant genes affecting the risk of CRC development[373].

1.5.4 Environmental risk factors

Environmental factors play a significant role in the risk of developing CRC. Wu et al estimated that the risk attributable to extrinsic factors with respect to Colon adenocarcinoma (COAD) was 97.2-97.9%[374]. Developed countries typically have much higher CRC rates than developing countries. The developed world accounts for ~1.2 billion of the world's population but CRC incidence in these regions account for 55% of overall incidences[375]. Strikingly in some case age standardised rate incidence (ASRi) can vary by up to 10 fold such as in the case of Australia and New Zealand (ASRi: 44.8 and 32.2 per 100,000 for men and women respectively) versus western Africa (ASRi: 4.5 and 3.8 per 100,000 for men and women respectively)[350]. One could speculate that this observation might be due to genetic differences between these populations for example people of Europe ancestry might have an increased inherited susceptibility to CRC development. However, two lines of evidence support a model where the environment is the variable which explains this difference. Firstly, it has been recorded that within a particular ethnic population, CRC rates have increase in parallel with economic development and the resulting environmental changes. For instance, in Shanghai, China the ASRi of CRC has increase by ~100% has between the periods of 1972–1977 and 1990–1994[376]. Secondly emigrants who come from low risk countries and live in high risk countries such as people from India moving to the UK are found to have an increased incidence of CRC[377]. CRC incidence rates increase in parallel with economic development. Countries in South America, Asia and Eastern European are predicted to undergo major economic development in the 21st century. Such facts pose a great problem for health systems around the world with respect to CRC. Extrapolating from epidemiological data and taking into account demographic dynamics and economics, the global burden of CRC burden is expected to increase by 60% to more than 2.2 million new cases and 1.1 million cancer deaths by 2030[378]. Notably however, there has been a decrease in CRC rate in The United States of America (USA) with an average decrease of 3.4% per year in the past decade (2001 to 2010)[379]. It is unknown what is causing this decrease, but it has been proposed that public health services in the form of awareness campaigns underline this reduction. It is notable however that there is worrying rise in CRC incidence in individuals under 50 years of age[380].

A myriad of environmental factors that modulate CRC risk have been noted (Table 7). Different subtypes within the colon are differentially affected by risk factors.

Table 7 | summary of the associations between risk or protective factors and colorectal cancer risk by anatomical subsites. ↑↑, convincing risk factor; ↑, probable risk factor; ↓↓, convincing protective factor; ↓, probable protective factor; BMI, body mass index; CI, confidence interval; CRC, colorectal cancer; MET, metabolic equivalent of task; RR, relative risk; WC, waist circumference. A Level of evidence as indicated by WCRF−AICR summary report for CRC9, except for smoking and aspirin (based on evidence from observational studies and randomized controlled trials).b Long latency was required to observe an effect on CRC. These data was derived from Keum and Giovannucci., 2019[381]

Aetiological factors	**level of evidence**	**Unit increase**	**Colorectal cancer RR (95% CI)**	**Colon cancer RR (95% CI)**	**rectal cancer RR (95% CI)**
Obesity	↑↑	5 kg/m^2 in BMI	1.05 (1.03–1.07)	1.07 (1.05–1.09)	1.01 (1.01–1.04)
	↑↑	10 cm in WC	1.02 (1.01–1.03)	1.04 (1.02–1.06	1.02 (1.00–1.03)
Total physical activity	↓↓	5 MET-hours per week	0.97 (0.94–0.99)	0.92 (0.86–0.99)	1.02 (0.95–1.10
Western dietary pattern	↑↑	Highest versus lowest	1.12 (1.01–1.24)	1.30 (1.04–1.63)	1.09 (0.91–1.29)
Prudent dietary pattern	↓↓	Highest versus lowest	0.89 (0.84–0.95)	0.89 (0.80–0.99)	0.96 (0.83–1.10)
Processed meat intake	↑↑	50 g per day	1.16 (1.08–1.26)	1.23 (1.11–1.35)	1.08 (1.00–1.18)
Red meat intake	↑	100 g per day	1.12 (1.00–1.25)	1.22 (1.06–1.39)	1.13 (0.96–1.34)
Total fibre intake	↓	10 g per day	0.93 (0.87–1.00)	0.91 (0.84–1.00)	0.93 (0.85–1.01)
Whole grain intake	↓	90 g per day	0.83 (0.79–0.89)	0.82 (0.73–0.92)	0.82 (0.57–1.16)
Alcohol (as ethanol)	↑↑	10 g per day	1.07 (1.05–1.09)	1.07 (1.05–1.09)	1.08 (1.07–1.10)
Smoking[b]	↑	Current versus never smokers	1.15 (1.00–1.32)	1.10 (0.89–1.36)	1.19 (0.94–1.54)
Aspirin intake	↑↑	75–1200 mg per day versus control	0.76 (0.63–0.94)	0.76 (0.60–0.96)	0.90 (0.63–1.30)
Total calcium[b]	↓	300 mg per day	0.92 (0.89–0.95)	0.91 (0.87–0.96)	0.95 (0.83–1.08)

1.5.5 CRC and Dietary Fibre

According to the CODEX Alimentarius Commission (CAC), dietary fibre is defined by carbohydrate polymers[382] 1) with 10 or more monomeric units 2) which are not hydrolysed by the endogenous enzymes in the small intestine of humans and which belong to the following categories:

1. Edible carbohydrate polymers naturally occurring in the food as consumed.

2. Carbohydrate polymers which have been obtained from food raw material by physical, enzymatic or chemical means and which have been shown to have a physiological effect of benefit to health as demonstrated by generally accepted scientific evidence to competent authorities,

3. Synthetic carbohydrate polymers, which have been shown to have a physiological effect of benefit to health as demonstrated by generally accepted scientific evidence to competent authorities.

Total fibre intake shows a protective effect regarding colorectal cancer; a recent meta-analysis found a relative risk of 0·84 between high consumption and low consumption[383]. This finding is in large agreement with other meta-analysis[384,385]. The protective effect of fibre has shown to exhibit a dose relative effect[383]. Moreover, the protective value of fibre seems to extend after cancer diagnosis with the multivariable Hazard ratio per each 5-g increase in intake per day was 0.78 for CRC-specific mortality[386].

Although the human genetic repertoire dose not encode the ability to break down fibre, the gut microbiota utilizes fibre as a major source of energy. A key metabolite of dietary fibre fermentation by anaerobic gut microbes are short-chain fatty acids (SCFAs). SCFA produced by the microbiome mainly consist of acetate, propionate, and butyrate. The proportions of these produces depend the composition of the microbiota and the type of fibre consumed[387]. SCFAs have been shown to be protective against CRC development[388].

1.5.6 Colorectal cancer and the microbiome

The relationship between the relationships between the microbiome and colorectal cancer has been extensively studied. A number of microbes have been described to be associated with CRC (Table 8)

Table 8 Top 20 Enriched Bacterial Genera and Species in Colorectal Adenoma and CRC Patients. Genera are ordered by rank. Rank is based on the number of studies reporting the association, denoted as hits. These data was derived from Ternes et al, 2020[389]

Genus	Species	Number of hits
Fusobacterium		31
	nucleatum	
	gonidiaformans	
	mortiferum	
	necrophorum	
	peridonticum	
Peptostreptococcus		18
	stomatis	
	anaerobius	
	endodontalis	
Porphyromonas		16
	asaccharolytica	
	uenonis	
	somerae	
Bacteroides		14
	fragilis	
	ovatus	
	caccae	
	dorei	
	eggerthii	
	massiliensis	
	salyersiae	
	splanchnicus	
	vulgatus	
	xylanisolvens	
Parvimonas		13
	micra	
Prevotella		13
	intermedia	
	nigrescens	
Gemella		12
	morbillorum	
Streptococcus		11
	anginosus	
	dysgalactiae	
	constellatus	
	gallolyticus	
	thermophilus	
	tigurinus	
Clostridium		9
	symbiosum	
	hylemonae	
Escherichia		9
	coli	
Bilophila		8
	wadsworthia	

Campylobacter		8
	gracilis	
	rectus	
	showae	
	ureolyticus	
Phascolarctobacterium		8
	succinatutens	
Selenomonas		8
	sputigena	
Ruminococcus		7
	torques	
Shigella		7
Akkermansia		6
	muciniphila	
Desulfovibrio		6
	desulfuricans	
	longreachensis	
	vietnamensis	
Eubacterium		6
	infirmum	
	limosum	
Leptotrichia		6
	hofstadii	
	buccalis	

Many studies report changes in microbial taxa and pathways associated with diseases, including CRC, in a manner one might interpreted that they exert a biological effect in isolation. However, it is important to view these changes in the context of an ecosystem. A number of models have been developed to describe the ecological role these microbes have in CRC oncogenesis. The 'alpha-bug hypothesis' developed by Sears and Pardoll postulates that a key microbe within the microbiota possesses specific virulence factors which enable it to promote oncogenesis while also remodelling the microbial community towards an oncogenic phenotype[390]. Sears and Pardoll use *Enterotoxigenic B. fragilis* (ETBF) as a potential example of such a microbe due to its ability to induce DNA damage and to modify the immune microenvironment[391]. A variant on this model is the driver–passenger model proposed by Tjalsma et al[392]. Like the alpha-bug model, a driver microbe promotes oncogenesis at the early stage. However the changes to a tumour microenvironment allows opportunistic microbes to proliferate in the new niche and ultimately outcompete the driver microbes[392]. These passenger microbes may or may not promote oncogenesis. An example of a putative passenger microbe is *Fusobacterium nucleatum* which has been consistently found enriched on CRC tumours and has also been shown to drive tumour progression. Indeed, the microbiome has been shown to vary between stages within the Adenoma-carcinoma sequence[393,394]. Keystone microbial taxa have been defined as "microbial keystone taxa are highly connected taxa that individually or in a guild exert a considerable influence on microbiome structure and functioning irrespective of their abundance across space and time"[395]. The concept of the keystones species was first proposed by ecologist Robert T. Paine in 1969 and applied to human microbial niches by Hajishengallis et al[396]. An example of a keystone taxon regarding the gut microbiome is *Bacteroides thetaiotaomicron* which has also been shown to be important to the recovery of the microbiome after antibiotic therapy[397,398]. In the context of CRC, a taxon may establish and maintain a pro-oncogenic enviroment. Finally, the hit and run model describes a dynamic whereby a specific microbe induces an insult to the tissue in a manner that promote cancer. The bacterium may not drive further oncogenesis and its presence may be transient. For example colibactin producing pk+ *E.coli* may colonize the colon in an individual causing DNA damage to colonic cells. However, once CRC develops the microbe may no longer be present.

1.6 Oesophageal cancer

Globally, oesophageal cancer is the eleventh most common cancer in terms of incidence (572,000 new cases) and sixth in cancer mortality (509,000 deaths) according to 2018 figures[172]. Histologically, there are two main subtypes of oesophageal cancer; oesophageal adenocarcinoma (OAC) and oesophageal squamous-cell carcinoma (OSCC). Worldwide, OSCC is by far the most prevalent subtype with ~90% of cases[172]. However, there is a dramatic variation in geographical distribution with respect to these two subtypes[399,400]. OSCC shows highest prevalence in developing geographical regions such as China, central Asia and Sub-Saharan Africa, while OAC is the predominant type in developed regions such as Australia, Europe and North America. Indeed, the incidence rate of OAC has seen a dramatic rise in the Western world in the last 30 years, an increase of 600%. In contrast OSCC has seen a decrease in incidence in the past 30 years of over 50%[399,400].

This thesis focuses on OAC, as with other western countries, this is the majority histological presentation within the Irish population. The prognosis for OAC is relatively poor with an overall 5-year survival of <20% for all stages of cancer[401]. The survival rate drops to 5% for the distant disease versus 43% for the localized disease[401].

1.6.1 Natural history of oesophageal adenocarcinoma

The putative natural history of OAC has been well described wherein normal tissue evolves through the gastroesophageal reflux disease – Barrett's oesophagus – oesophageal adenocarcinoma sequence[399,402]. GERD is "a condition that develops when the reflux of stomach contents into the oesophagus causes troublesome symptoms and/or complications" – this causes normal stratified squamous epithelium of the oesophagus to be exposed to acid, bile, and other stomach contents. As a reaction to this chronic exposure the normal epithelium is replaced by metaplastic columnar epithelium which can be described a specialized intestinal metaplasia[399,402,403]. Barrett's oesophagus progresses through low to high grade dysplasia to local OAC and finally metastatic OAC. GERDs is associated with an odds ratio of 12.0 and 4.64 for Barrett's oesophagus and OAC respectively[399]. However, many epidemiological observations have challenged this straight forward series of events. For one, 95% of individuals diagnosed with OAC have no prior diagnosis of Barrett's oesophagus[402]. Individual with Barrett's oesophagus have a risk of developing OAC that is 10-fold to 55-fold higher than that of the general population, however the absolute risk is calculated to be 0.5%, or 1/200 person-years[402,404]. These observation suggest two scenarios; 1) GERDs/Barrett's oesophagus appears in individuals unobserved and/or without symptoms who subsequently develop OAC 2) OAC can develop by mechanisms independent of the described inflammatory-metaplasia-dysplasia-oesophageal adenocarcinoma sequence. However a recent computational model suggests that the most OAC cases arise from Barrett's oesophagus [405]

1.6.1.1 Pathogenesis of Barrett's oesophagus

The tissue of Barrett's oesophagus has a glandular structure comprising of crypts, similar to that of gastric and intestinal tissue. This metaplastic tissue comprises many different types of differentiated cells. These cell types included columnar cells, mucin-secreting gastric foveolar-type cells and goblet cells[406]. The precise cellular origin of Barrett oesophagus is unknown but models have been developed to explain the pathogenesis of Barrett's oesophagus.

In one model oesophageal squamous cells undergo transdifferentiation into metaplastic columnar epithelium[407]. Transdifferentiation is a process where a differentiated cell changes into another differentiated cell[408]. This can occur a response to injury tissue injury but also can be induced artificially in a laboratory setting. This transdifferentiation may occur and directly were squamous cells transdifferentiate directly to columnar epithelium, or indirectly where the conversion occurs through an intermediate.

Transcommitment is a phenomena where immature progenitor cells are reprogramed to alter their differentiation. Where these progenitor cells are derived from are also a matter of research. There is four suspected origins of these progenitors including 1) progenitor oesophageal cells including basal cells of the squamous epithelium or cells of oesophageal submucosal glands and their ducts 2) migrating proximal gastric cardia cells 3) Specialized populations of cells at the Gastro oesophageal junction (GOJ) including residual embryonic cell and transitional basal cell 4)bone marrow progenitor cells.[406]

1.6.2 Environmental risk factors for developing OAC

Because OAC is a disease with a multifactorial pathoetiology, environmental factors have been implicated as risk modifiers (Table 9). In terms of risk factors, GORDS, obesity and tobacco smoking have bern calculated as explaining 80% of OAC cases[409]. GORDs is the strongest factor and is believed to be necessary for the occurrence of Barrett's oesophagus.

Table 9 | Risk factors associated with the development of Oesophageal adenocarcinoma[399,410,411]

Risk factor	**Association with OAC - Odds ratio (95% CI)**
GORD	4.64 (3.28–6.57)
Obesity	2.69 (1.62–4.46)
Tobacco smoking	1.96(1.64-2.34
Helicobacter pylori infection	0.5 (0.4–0.7)
Male Sex	2.2 (1.8–2.5)
High red meat intake	1.91 (1.07-3.38)
NSAID use	0.68(0.56-0.83)
Fruit intake	0.86 (0.80–0.93)

1.6.2.1 Obesity

Obesity is one of the strongest risk factors for developing BO and OAC with a >2 increase in risk in obese individuals versus those of healthy weight[411]. This relationship between BMI and OAC/BE is a linear exposure–response pattern. In particular the distribution of body fat seems to be a particularly important metric regarding risk for BO and OAC. When truncal obesity (excessive abdominal or visceral fat) is controlled for in the form of waist circumference measurements, the relationship between obesity and BO/OAC almost disappears[412]. Obesity during adolescence has also been noted as a particular risk[413,414]. This is a worrying trend as obesity is rising in the teenage population and may give rise to cancer later in life.

Obesity increases the risk of cancer in a wide range of cancers[415]. Obesity seems to exert systemic inflammatory and metabolic alterations[416]. Increased serum levels of insulin and leptin are associated with BO development[417]. In one prospective study, increased levels of leptin and insulin in individuals with BO was positively associated with the development of OAC[418]. In the same study the levels of adipokine adiponectin was inversely associated with OAC devolvement in a non-linear manner[419].

Abdominal fat may act to increase intra-abdominal pressure thereby leading to a relaxation of the lower oesophageal sphincter. This relaxation of the lower oesophageal sphincter may lead to an increased susceptibility in GORDs[420,421].

1.6.3 Formation of the OAC genome

OAC has a very high mutational load relative to other cancers[254,422]. Non-neoplastic BO tissue samples adjacent to OAC samples have a high mutational load with a somatic mutation frequency of 1.3-5.4 mutations per Mb cancers[423]. This level exceeds that found in some cancers such as prostate and breast. The mutational signatures present in OAC has been delineated and OAC tumours may be classified via these signatures[424]. To this end, 3 subgroups of OAC have been defined which include a C>A/T dominant group (comprising Signature 1 and a 18-like mutational signature), DNA Damage Repair (DDR) impaired (BRCA group), and a mutagenic (predominantly Signature 17A or signature S17B) group. The mutagenic group was named due to its statistically highest mutational load. The DDR impaired group exhibit a 4.3-fold enrichment in dysregulation of in homologous recombination (HR) pathways relative to the other groups.

This classification may also inform therapeutic strategies. Tumour mutational burden (TMB) is predictive of clinical response to Immune checkpoint inhibitor[425]. Tumours with a higher TMB have a better response which is putatively due to higher number of tumour neoantigens[426]. Indeed, the mutagenic group had the highest presentation of neoantigens. Treatment of a MFD cell line, with the genetic characteristics of the mutagenic group, with pharmacological inhibitors to the G2/M-phase checkpoint regulators Wee1 and Chk1/2, yielded a 25-fold and 10-fold increased sensitivity relativity to the CAM02 cells which have C>A/T dominant group characteristics. OES127 cells lines, representing the DDR impaired group, experienced cell death when exposed to a combination of Olaparib (Topoisomerase

I inhibitor) and Topotecan (a DNA damaging agent) while the other cell lines did not.

Note that recent analysis has allowed for the separation of Mutational signature 17 into two signatures, that is SBS Signatures 17 A and B. SBS Signatures 17 A is substitutions defined by T>C while SBS Signatures 17 B is defined by T>G substitutions[254].

SBS Signatures 17 A and B are present in a high proportion in both OAC and gastric adenocarcinoma tumours[427]. This may indicate that a common physiological feature such as gastric acid may be a common cause/modifier leading to the signature[254].

The aetiology of SBS Signatures 17 A and B is not known. However, one hypothesis with supporting data involved the stimulation of the production of ROS in Oesophageal cells exposed to acidic bile reflux. ROS has been demonstrated to be generated by both mitochondria and NADPH oxidases[428]. NOX5-S, a truncated variant of NOX5, has been found to produce ROS and to promote DNA damage in a bile acid dependent manner[429,430]. PPIs were found to reduce mRNA levels of NOX5-S in BE mucosa biopsies[431]. Furthermore, NOX1 and NOX2 can also generate ROS in acidic bile salt treated cells[428].

In particular, one explanation for SBS Signatures 17, specifically SBS Signatures 17 B, is the oxidation the nucleotide pool thereby forming 8-hydroxy-2′-deoxyguanosine 5′-triphosphate (8-OH-dGTP) [432,433]. The presence of 8-OH-dGTP in the nucleotide pool has been shown induce A:T to C:G (T>G) base substitutions[434]. These base substitutions are indicative of SBS Signatures 17 B in particular. However mutational signature 18, which has been linked to ROS, does not seem to have a direct link with SBS Signatures 17 B.

Another mechanism by which reflux of bile acid and/or gastric acid promotes DNA damage is thought to be the production of reactive nitrogen species (RNS). Inducible nitric oxide synthase was found to be upregulated in BO and OAC[431,435,436]. Proton pump inhibitors were found to reduce inducible nitric oxide synthase levels in BO tissue but not normal oesophageal tissue[431]. Dinitrogen trioxide can induce adenine nitrosative deamination to hypoxanthine which in turn can lead to T>C substitution during systhesis[431] . This substitution is central to SBS Signatures 17 A.

An early mutation that occurs in OAC oncogenesis is a mutation in the TP53 gene as is evident from the fact it is found in healthy cell populations as well as non-neoplastic BO. However many of the mutations found in non-neoplastic BO are not shared with adjacent tissue.

The transformation BO into OAC can occur via 3 pathways[411,437]. In the traditional pathway a stepwise loss of tumour suppressor genes including CDKN2A and SMAD4 occurs. This is followed by oncogene amplification and MMR deficiency. What is regarded to be a much more frequent mode of evolution is via whole genome duplication[438]. A third mode of genome evolution is through catastrophic genome events including chromothripsis, kataegis and breakage–fusion–bridge[439].

1.6.4 Oesophageal microbiota

Efforts to define the oesophageal microbiota have been made using NGS (Table 10). The oesophageal microbiome is similar to that of other niches of the upper digestive tract such as the oral cavity and the stomach, with the genus Streptococcus being the most dominant taxa, and with other genera such as *Haemophilus, Neisseria and Prevotella* also being dominant taxa.

Table 10 | Studies using NGS technologies to delineate the oesophageal microbiome and its relationship to the cancer development.

Author	Laboratory	cohort	Sample Type	Methods	Findings
Elliott et al., 2017 (The Lancet Gastroenterology & Hepatology)[152]	Rebecca C Fitzgerald (University of Cambridge)	Normal=20 BO=24 HGD=23 OAC=19	Cytosponge, Brush, biopsy	V1-V2 2 × 250 bp	Decreased microbial diversity in OAC tissue compared with controls. Enrichment of acid-tolerant bacteria such as *Lactobacillus fermentum* in OAC samples
Nobel et al., 2018 (Clinical and Translational Gastroenterology)	Julian A. Abrams (University Irving)	GERD=5 BO= 31 Other=11	Two brushings were taken from the following sites: squamous esophagus (3 cm proximal to the squamo-columnar junction), gastric cardia (within 1 cm of the top of the gastric folds), and mid-BE segment in patients with BE. Brush tips were cut using sterile wire cutters and samples	V4 MiSeq 2 × 250 bp Differential abundance: linear discriminant analysis effect size	Subjects were divided into quartiles based on fibre intake. Low fibre intake was associated with increase in the taxa *Aggregatibacter, cardiobacterium, Lautropia, Paludibacter, Prevotella, Neisseria and unclassified Tissierellaceae* High fibre intake was associated with an increased relative abundance of an unclassified genus in family *Pasteurellaceae*
Deshpande et al.,2018 (microbiome)[440]	Nadeem Omar Kaakoush (University of New South Wales)	Normal=59 GERD=29 GM=7 BO=5 EAC=1	Bursh (Biopsies were taken but not analysed)	16S rRNA V4 Shotgun sequencing (Both lumina MiSeq	The esophageal microbiome was found to cluster into functionally distinct community types (esotypes) defined by *Streptococcus* and *Prevotella*

		EoE=1		2 × 250 bp chemistry)	
Okereke et al., 2019 (Scientific Reports)[441]		BO=17	Biopsies of esophageal mucosa were taken from the (1) proximal esophagus, (2) mid-esophagus, (3) distal esophagus, and (4) Barrett's esophagus. Swabs were also taken from the uvula and the endoscope.	Ion Torrent long reads V1-V8	Biopsies samples differed in composition the that of swab samples
Snider et al., 2019 (Cancer Epidemiology, Biomarkers & Prevention)[442]	Julian A. Abrams(University Irving)	16 controls; 14 Barrett's oesophagus without dysplasia (NDBO); 6 low-grade dysplasia (LGD); 5 high-grade dysplasia (HGD); and 4 oesophageal adenocarcinoma (OAC)	See Nobel et al	V4 MiSeq 2 × 250 bp Differential abundance: linear discriminant analysis effect size	Patients with NDBE/LGD had significantly increased *Veillonella*. Paateints with HGD / esophageal had significantly increased *Akkermansia muciniphila*, *Enterobacteriaceae*, *Moraxella*, *Oscillospira* and *Proteus* OAC had reduced alpha diversity as calculated by Simpson Index

The findings summarised in the above table shows numerous studies have tried to find a relationship between the microbiota and oesophageal diseases. However, no consistent microbial signatures have been identified with relationship to the microbiome and oesophageal cancer development. These studies difference in there methodological implementation including primer pairs and sampling procedure. Pinch biopsies as a method of sample collection maybe thought as superior to swabs as they may more effective at collecting mucosal adherent bacteria. From a statistical/bioinformatic perspective, differential abundance analysis is a key aspect of all these microbiome studies. Many of these studies use Linear Discriminant Analysis Effect Size (LEfSe) which has been described by more of a discriminant analysis method than a differential abundance analysis method. Some studies described in the above table have respectable sample size per clinical group study including Elliott et al. However many of these studies included clinical groups composed of cohorts less than 5. Finally few of these studies examine inter-individuals microbiome variation within the oesophagus.

1.7 Aims of this thesis

The research in this thesis worked under the hypothesis that the human microbiome is associated with and plays a role in cancer biology. This thesis contains four projects which, to varying extents, contribute to key areas of cancer research.

In chapter 2, we investigate the microbiome of mucosal biopsies derived from patients along the inflammation-metaplasia-dysplasia-oesophageal adenocarcinoma sequence in an Irish cohort. Furthermore, we collected and analysed multiple biopsy per individuals to examine the intra-individual microbiome variation. Identification of differences in microbiome features between clinical categories would lead to the hypothesis that the oesophageal and/or gastric microbiome modulates the development of oesophageal adenocarcinoma. Moreover, these data would provide information regarding whether these changes expand to the either upper GI tract.

CRC screening programs have been associated with a decrease in CRC incidence and deaths[443]. The microbiome is being explored for its potential to inform the development of new diagnostic tools[444,445]. Previous work has indicated that colorectal cancer is associated with changes in the microbiome throughout the colon and is not restricted to the cancer[394]. In Chapter 3we investigate the spatial organisation of the mucosal colon microbiome in the context of CRC. We sought to identify intra-individual difference in colonic mucosal biopsies in individuals with CRC. To this end, Chapter 2 and Chapter 3 share a core similarity whereby in Chapter 2 inter-individual variation in the oesophageal/gastric microbiome in the context of OAC is being delineated while in Chapter 3 inter-individual variation in the colonic microbiome in the context of CRC is being delineated. This research would add to the discussion on the diagnostic power of non-disease colonic tissue versus diseased tissue.

Even in the context of a robust understanding of cancer risk factors and wide spread screening programs, cancer will occur in society. Immune checkpoint inhibitors (ICI) represent a significant addition to cancer therapeutics. However, a large proportion of individuals do not response to ICI. An increasingly appreciated modulator of response to ICI is the gut microbiota. In chapter 4 we examined the association between the microbiome and clinical responses (response and side effects) to ICI in the context of melanoma. This study was conducted in a geographically different location i.e. Ireland relative to previous studies[446]. These data would allow the examination between consistencies/inconsistencies in microbiome features associated with clinical outcomes to ICI in geographically distinct populations.

Inflammation is known to be a major contributor to oncogenesis[447]. Many inflammatory diseases are known to be risk factors to the development of cancer e.g. ulcerative colitis (UC) is a risk factor for CRC[356]. A major area of microbiome research involves investigating the role of the microbiome in modulating inflammation[448]. One would argue the need to explore the potential of an inflammation-microbiome-cancer axis. Hidradenitis Suppurativa (HS) is a chronic inflammatory skin disease which affects the intertriginous skin[449]. HS is known to increase risk to the development of a range of caners[450]. In Chapter 5 we investigated alterations in the skin and faecal microbiome in individuals with HS. Microbiome features which drive inflammation have the potential to drive oncogenesis. It is thus pertinent to identify microbiome features associated with inflammatory diseases such as HS.

1.8 References

1 Takai, K. *et al.* Cell proliferation at 122 degrees C and isotopically heavy CH4 production by a hyperthermophilic methanogen under high-pressure cultivation. *P Natl Acad Sci USA* **105**, 10949-10954, doi:10.1073/pnas.0712334105 (2008).

2 Goordial, J. *et al.* In Situ Field Sequencing and Life Detection in Remote (79°26′N) Canadian High Arctic Permafrost Ice Wedge Microbial

Communities. *Frontiers in Microbiology* **8**, doi:10.3389/fmicb.2017.02594 (2017).

3 Berg, G. *et al.* Microbiome definition re-visited: old concepts and new challenges. *Microbiome* **8**, 103, doi:10.1186/s40168-020-00875-0 (2020).

4 van Vliet, S. & Doebeli, M. The role of multilevel selection in host microbiome evolution. *Proceedings of the National Academy of Sciences* **116**, 20591, doi:10.1073/pnas.1909790116 (2019).

5 Cebra, J. J. Influences of microbiota on intestinal immune system development. *The American Journal of Clinical Nutrition* **69**, 1046s-1051s, doi:10.1093/ajcn/69.5.1046s (1999).

6 Keen, E. C. & Dantas, G. Close Encounters of Three Kinds: Bacteriophages, Commensal Bacteria, and Host Immunity. *Trends Microbiol* **26**, 943-954, doi:10.1016/j.tim.2018.05.009 (2018).

7 Nilsson, R. H. *et al.* Mycobiome diversity: high-throughput sequencing and identification of fungi. *Nat Rev Microbiol* **17**, 95-109, doi:10.1038/s41579-018-0116-y (2019).

8 Sender, R., Fuchs, S. & Milo, R. Revised Estimates for the Number of Human and Bacteria Cells in the Body. *PLoS Biol* **14**, e1002533, doi:10.1371/journal.pbio.1002533 (2016).

9 Zhernakova, A. *et al.* Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. *Science* **352**, 565-569, doi:10.1126/science.aad3369 (2016).

10 Gilbert, J. A. *et al.* Current understanding of the human microbiome. *Nat Med* **24**, 392-400, doi:10.1038/nm.4517 (2018).

11 Nayfach, S., Shi, Z. J., Seshadri, R., Pollard, K. S. & Kyrpides, N. C. New insights from uncultivated genomes of the global human gut microbiome. *Nature* **568**, 505-510, doi:10.1038/s41586-019-1058-x (2019).

12 Almeida, A. *et al.* A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nature Biotechnology*, doi:10.1038/s41587-020-0603-3 (2020).

13 Donaldson, G. P., Lee, S. M. & Mazmanian, S. K. Gut biogeography of the bacterial microbiota. *Nat Rev Microbiol* **14**, 20-32, doi:10.1038/nrmicro3552 (2016).

14 Tropini, C., Earle, K. A., Huang, K. C. & Sonnenburg, J. L. The Gut Microbiome: Connecting Spatial Organization to Function. *Cell Host Microbe* **21**, 433-442, doi:10.1016/j.chom.2017.03.010 (2017).

15 Flynn, K. J., Ruffin, M. T., Turgeon, D. K. & Schloss, P. D. Spatial Variation of the Native Colon Microbiota in Healthy Adults. *Cancer Prevention Research* **11**, 393, doi:10.1158/1940-6207.CAPR-17-0370 (2018).

16 Lavelle, A. *et al.* Spatial variation of the colonic microbiota in patients with ulcerative colitis and control volunteers. *Gut* **64**, 1553-1561, doi:10.1136/gutjnl-2014-307873 (2015).

17 Pédron, T. *et al.* A crypt-specific core microbiota resides in the mouse colon. *mBio* **3**, doi:10.1128/mBio.00116-12 (2012).

18 Albenberg, L. *et al.* Correlation between intraluminal oxygen gradient and radial partitioning of intestinal microbiota. *Gastroenterology* **147**, 1055-1063.e1058, doi:10.1053/j.gastro.2014.07.020 (2014).

19 Huse, S. M. *et al.* Comparison of brush and biopsy sampling methods of the ileal pouch for assessment of mucosa-associated microbiota of human subjects. *Microbiome* **2**, 5, doi:10.1186/2049-2618-2-5 (2014).

20 Vandeputte, D. *et al.* Stool consistency is strongly associated with gut microbiota richness and composition, enterotypes and bacterial growth rates. *Gut* **65**, 57, doi:10.1136/gutjnl-2015-309618 (2016).

21 Bassis, C. M. *et al.* Comparison of stool versus rectal swab samples and storage conditions on bacterial community profiles. *BMC Microbiology* **17**, 78, doi:10.1186/s12866-017-0983-9 (2017).

22 Reyman, M., van Houten, M. A., Arp, K., Sanders, E. A. M. & Bogaert, D. Rectal swabs are a reliable proxy for faecal samples in infant gut microbiota research based on 16S-rRNA sequencing. *Scientific Reports* **9**, 16072, doi:10.1038/s41598-019-52549-z (2019).

23 Pleasants, J. R. Rearing germfree cesarean-born rats, mice, and rabbits through weaning. *Ann N Y Acad Sci* **78**, 116-126, doi:10.1111/j.1749-6632.1959.tb53099.x (1959).

24 Zheng, D., Liwinski, T. & Elinav, E. Interaction between microbiota and immunity in health and disease. *Cell Res* **30**, 492-506, doi:10.1038/s41422-020-0332-7 (2020).

25 Fan, Y. & Pedersen, O. Gut microbiota in human metabolic health and disease. *Nature Reviews Microbiology*, doi:10.1038/s41579-020-0433-9 (2020).

26 Liu, F. *et al.* Altered composition and function of intestinal microbiota in autism spectrum disorders: a systematic review. *Translational Psychiatry* **9**, 43, doi:10.1038/s41398-019-0389-6 (2019).

27 Sgritta, M. *et al.* Mechanisms Underlying Microbial-Mediated Changes in Social Behavior in Mouse Models of Autism Spectrum Disorder. *Neuron* **101**, 246-259.e246, doi:10.1016/j.neuron.2018.11.018 (2019).

28 Zhang, M. *et al.* A quasi-paired cohort strategy reveals the impaired detoxifying function of microbes in the gut of autistic children. *Sci Adv* **6**, doi:10.1126/sciadv.aba3760 (2020).

29 Jie, Z. *et al.* The gut microbiome in atherosclerotic cardiovascular disease. *Nature Communications* **8**, 845, doi:10.1038/s41467-017-00900-1 (2017).

30 Wang, Z. *et al.* Gut flora metabolism of phosphatidylcholine promotes cardiovascular disease. *Nature* **472**, 57-63, doi:10.1038/nature09922 (2011).

31 Koeth, R. A. *et al.* Intestinal microbiota metabolism of L-carnitine, a nutrient in red meat, promotes atherosclerosis. *Nature medicine* **19**, 576-585, doi:10.1038/nm.3145 (2013).

32 Bennett, B. J. *et al.* Trimethylamine-N-oxide, a metabolite associated with atherosclerosis, exhibits complex genetic and dietary regulation. *Cell Metab* **17**, 49-60, doi:10.1016/j.cmet.2012.12.011 (2013).

33 Tang, W. H. *et al.* Intestinal microbial metabolism of phosphatidylcholine and cardiovascular risk. *N Engl J Med* **368**, 1575-1584, doi:10.1056/NEJMoa1109400 (2013).

34 Zhu, W. *et al.* Gut Microbial Metabolite TMAO Enhances Platelet Hyperreactivity and Thrombosis Risk. *Cell* **165**, 111-124, doi:10.1016/j.cell.2016.02.011 (2016).

35 Gurung, M. *et al.* Role of gut microbiota in type 2 diabetes pathophysiology. *EBioMedicine* **51**, 102590, doi:10.1016/j.ebiom.2019.11.051 (2020).

36 Kim, S. H. *et al.* The anti-diabetic activity of Bifidobacterium lactis HY8101 in vitro and in vivo. *J Appl Microbiol* **117**, 834-845, doi:10.1111/jam.12573 (2014).

37 Plovier, H. *et al.* A purified membrane protein from Akkermansia muciniphila or the pasteurized bacterium improves metabolism in obese and diabetic mice. *Nature Medicine* **23**, 107-113, doi:10.1038/nm.4236 (2017).

38 Schirmer, M., Garner, A., Vlamakis, H. & Xavier, R. J. Microbial genes and pathways in inflammatory bowel disease. *Nature Reviews Microbiology* **17**, 497-511, doi:10.1038/s41579-019-0213-6 (2019).

39 Henke, M. T. *et al.* &lt;em&gt;Ruminococcus gnavus&lt;/em&gt;, a member of the human gut microbiome associated with Crohn’s disease, produces an inflammatory polysaccharide. *Proceedings of the National Academy of Sciences* **116**, 12672, doi:10.1073/pnas.1904099116 (2019).

40 Mirsepasi-Lauridsen, H. C., Vallance, B. A., Krogfelt, K. A. & Petersen, A. M. &lt;em&gt;Escherichia coli&lt;/em&gt; Pathobionts Associated with Inflammatory Bowel Disease. *Clinical Microbiology Reviews* **32**, e00060-00018, doi:10.1128/CMR.00060-18 (2019).

41 Jiang, W. *et al.* Dysbiosis gut microbiota associated with inflammation and impaired mucosal immune function in intestine of humans with non-alcoholic fatty liver disease. *Sci Rep* **5**, 8096, doi:10.1038/srep08096 (2015).

42 Yuan, J. *et al.* Fatty Liver Disease Caused by High-Alcohol-Producing Klebsiella pneumoniae. *Cell Metab* **30**, 675-688 e677, doi:10.1016/j.cmet.2019.08.018 (2019).

43 Le Roy, T. *et al.* Intestinal microbiota determines development of non-alcoholic fatty liver disease in mice. *Gut* **62**, 1787, doi:10.1136/gutjnl-2012-303816 (2013).

44 Hoyles, L. *et al.* Molecular phenomics and metagenomics of hepatic steatosis in non-diabetic obese women. *Nature medicine* **24**, 1070-1080, doi:10.1038/s41591-018-0061-3 (2018).

45 Wainwright, S. A. Form and function in organisms. *American Zoologist* **28**, 671-680 (1988).

46 Church, G. M., Gao, Y. & Kosuri, S. Next-Generation Digital Information Storage in DNA. *Science* **337**, 1628-1628, doi:10.1126/science.1226355 (2012).

47 Wu, R. & Kaiser, A. D. Structure and base sequence in the cohesive ends of bacteriophage lambda DNA. *Journal of Molecular Biology* **35**, 523-537, doi:https://doi.org/10.1016/S0022-2836(68)80012-9 (1968).

48 Wu, R. & Taylor, E. Nucleotide sequence analysis of DNA. II. Complete nucleotide sequence of the cohesive ends of bacteriophage lambda DNA. *J Mol Biol* **57**, 491-511, doi:10.1016/0022-2836(71)90105-7 (1971).

49 Gilbert, W. & Maxam, A. The nucleotide sequence of the lac operator. *P Natl Acad Sci USA* **70**, 3581-3584, doi:10.1073/pnas.70.12.3581 (1973).

50 Sanger, F. Sequences, sequences, and sequences. *Annu Rev Biochem* **57**, 1-28, doi:10.1146/annurev.bi.57.070188.000245 (1988).

51 Sanger, F. & Coulson, A. R. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J Mol Biol* **94**, 441-448, doi:10.1016/0022-2836(75)90213-2 (1975).

52 Sanger, F. *et al.* Nucleotide sequence of bacteriophage φX174 DNA. *Nature* **265**, 687-695, doi:10.1038/265687a0 (1977).

53 Maxam, A. M. & Gilbert, W. A new method for sequencing DNA. *P Natl Acad Sci USA* **74**, 560-564, doi:10.1073/pnas.74.2.560 (1977).

54 Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *P Natl Acad Sci USA* **74**, 5463-5467, doi:10.1073/pnas.74.12.5463 (1977).

55 Prober, J. M. *et al.* A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. *Science* **238**, 336-341, doi:10.1126/science.2443975 (1987).

56 Smith, L. M. *et al.* Fluorescence detection in automated DNA sequence analysis. *Nature* **321**, 674-679, doi:10.1038/321674a0 (1986).

57 Shendure, J. *et al.* DNA sequencing at 40: past, present and future. *Nature* **550**, 345-353, doi:10.1038/nature24286 (2017).

58 Messing, J., Crea, R. & Seeburg, P. H. A system for shotgun DNA sequencing. *Nucleic Acids Research* **9**, 309-321, doi:10.1093/nar/9.2.309 (1981).

59 Sanger, F., Coulson, A. R., Hong, G. F., Hill, D. F. & Petersen, G. B. Nucleotide sequence of bacteriophage lambda DNA. *J Mol Biol* **162**, 729-773, doi:10.1016/0022-2836(82)90546-0 (1982).

60 Fleischmann, R. D. *et al.* Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* **269**, 496-512, doi:10.1126/science.7542800 (1995).

61 Goffeau, A. *et al.* Life with 6000 genes. *Science* **274**, 546, 563-547, doi:10.1126/science.274.5287.546 (1996).

62 Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science* **282**, 2012-2018, doi:10.1126/science.282.5396.2012 (1998).

63 Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921, doi:10.1038/35057062 (2001).

64 Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931-945, doi:10.1038/nature03001 (2004).

65 Venter, J. C. *et al.* The sequence of the human genome. *Science* **291**, 1304-1351, doi:10.1126/science.1058040 (2001).

66 Nyrén, P. Enzymatic method for continuous monitoring of DNA polymerase activity. *Anal Biochem* **167**, 235-238, doi:10.1016/0003-2697(87)90158-8 (1987).

67 Nyrén, P., Pettersson, B. & Uhlén, M. Solid phase DNA minisequencing by an enzymatic luminometric inorganic pyrophosphate detection assay. *Anal Biochem* **208**, 171-175, doi:10.1006/abio.1993.1024 (1993).

68 Heather, J. M. & Chain, B. The sequence of sequencers: The history of sequencing DNA. *Genomics* **107**, 1-8, doi:10.1016/j.ygeno.2015.11.003 (2016).

69 Canard, B. & Sarfati, R. S. DNA polymerase fluorescent substrates with reversible 3'-tags. *Gene* **148**, 1-6, doi:10.1016/0378-1119(94)90226-7 (1994).

70 Ju, J. *et al.* Four-color DNA sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators. *Proceedings of the National Academy of Sciences* **103**, 19635, doi:10.1073/pnas.0609513103 (2006).

71 Turcatti, G., Romieu, A., Fedurco, M. & Tairi, A. P. A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis. *Nucleic Acids Res* **36**, e25, doi:10.1093/nar/gkn021 (2008).

72 Mitra, R. D. & Church, G. M. In situ localized amplification and contact replication of many individual DNA molecules. *Nucleic Acids Res* **27**, e34, doi:10.1093/nar/27.24.e34 (1999).

73 van Dijk, E. L., Jaszczyszyn, Y., Naquin, D. & Thermes, C. The Third Revolution in Sequencing Technology. *Trends Genet* **34**, 666-681, doi:10.1016/j.tig.2018.05.008 (2018).

74 Santos, A., van Aerle, R., Barrientos, L. & Martinez-Urtaza, J. Computational methods for 16S metabarcoding studies using Nanopore sequencing data. *Computational and Structural Biotechnology Journal* **18**, 296-305, doi:https://doi.org/10.1016/j.csbj.2020.01.005 (2020).

75 Ardui, S., Ameur, A., Vermeesch, J. R. & Hestand, M. S. Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. *Nucleic Acids Research* **46**, 2159-2168, doi:10.1093/nar/gky066 (2018).

76 Jain, M., Olsen, H. E., Paten, B. & Akeson, M. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol* **17**, 239, doi:10.1186/s13059-016-1103-0 (2016).

77 Liu, Q. *et al.* Detection of DNA base modifications by deep recurrent neural network on Oxford Nanopore sequencing data. *Nature Communications* **10**, 2449, doi:10.1038/s41467-019-10168-2 (2019).

78 Miga, K. H. *et al.* Telomere-to-telomere assembly of a complete human X chromosome. *Nature* **585**, 79-84, doi:10.1038/s41586-020-2547-7 (2020).

79 Brewer, T. E. *et al.* Unlinked rRNA genes are widespread among bacteria and archaea. *The ISME Journal* **14**, 597-608, doi:10.1038/s41396-019-0552-3 (2020).

80 Lavrinienko, A., Jernfors, T., Koskimäki, J. J., Pirttilä, A. M. & Watts, P. C. Does Intraspecific Variation in rDNA Copy Number Affect Analysis of Microbial Communities? *Trends Microbiol*, doi:10.1016/j.tim.2020.05.019 (2020).

81 Clarridge, J. E., 3rd. Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. *Clinical microbiology reviews* **17**, 840-862, doi:10.1128/CMR.17.4.840-862.2004 (2004).

82 Shen, Y. *et al.* Testing hypotheses on the rate of molecular evolution in relation to gene expression using microRNAs. *Proceedings of the National Academy of Sciences* **108**, 15942-15947 (2011).

83 Managadze, D., Rogozin, I. B., Chernikova, D., Shabalina, S. A. & Koonin, E. V. Negative correlation between expression level and evolutionary rate of long intergenic noncoding RNAs. *Genome biology and evolution* **3**, 1390-1404 (2011).

84 Li, G.-W., Oh, E. & Weissman, J. S. The anti-Shine–Dalgarno sequence drives translational pausing and codon choice in bacteria. *Nature* **484**, 538-541, doi:10.1038/nature10965 (2012).

85 Woese, C. R. & Fox, G. E. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *P Natl Acad Sci USA* **74**, 5088-5090, doi:10.1073/pnas.74.11.5088 (1977).

86 Wilson, K. H. & Blitchington, R. B. Human colonic biota studied by ribosomal DNA sequence analysis. *Appl Environ Microbiol* **62**, 2273-2278 (1996).

87 Suau, A. *et al.* Direct Analysis of Genes Encoding 16S rRNA from Complex Communities Reveals Many Novel Molecular Species within the Human Gut. *Applied and Environmental Microbiology* **65**, 4799-4807, doi:10.1128/aem.65.11.4799-4807.1999 (1999).

88 Kroes, I., Lepp, P. W. & Relman, D. A. Bacterial diversity within the human subgingival crevice. *Proceedings of the National Academy of Sciences* **96**, 14547, doi:10.1073/pnas.96.25.14547 (1999).

89 Browne, H. P. *et al.* Culturing of 'unculturable' human microbiota reveals novel taxa and extensive sporulation. *Nature* **533**, 543-546, doi:10.1038/nature17645 (2016).

90 Lewis, W. H., Tahon, G., Geesink, P., Sousa, D. Z. & Ettema, T. J. G. Innovations to culturing the uncultured microbial majority. *Nature Reviews Microbiology*, doi:10.1038/s41579-020-00458-8 (2020).

91 Eckburg, P. B. *et al.* Diversity of the human intestinal microbial flora. *Science* **308**, 1635-1638, doi:10.1126/science.1110591 (2005).

92 Rausch, P. *et al.* Comparative analysis of amplicon and metagenomic sequencing methods reveals key features in the evolution of animal metaorganisms. *Microbiome* **7**, 1-19 (2019).

93 McInerney, P., Adams, P. & Hadi, M. Z. Error Rate Comparison during Polymerase Chain Reaction by DNA Polymerase. *Molecular Biology International* **2014**, 287430, doi:10.1155/2014/287430 (2014).

94 Amplicon, P., Clean-Up, P. & Index, P. (2013).

95 Berry, D., Ben Mahfoudh, K., Wagner, M. & Loy, A. Barcoded Primers Used in Multiplex Amplicon Pyrosequencing Bias Amplification. *Applied and Environmental Microbiology* **77**, 7846, doi:10.1128/AEM.05220-11 (2011).

96 Sneath, P. H. & Sokal, R. R. *Numerical taxonomy. The principles and practice of numerical classification.* (1973).

97 Westcott, S. L. & Schloss, P. D. De novo clustering methods outperform reference-based methods for assigning 16S rRNA gene sequences to operational taxonomic units. *PeerJ* **3**, e1487, doi:10.7717/peerj.1487 (2015).

98 Kopylova, E. *et al.* Open-Source Sequence Clustering Methods Improve the State Of the Art. *mSystems* **1**, doi:10.1128/mSystems.00003-15 (2016).

99 Callahan, B. J., McMurdie, P. J. & Holmes, S. P. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *The ISME Journal* **11**, 2639-2643, doi:10.1038/ismej.2017.119 (2017).

100 Callahan, B. J. *et al.* DADA2: High-resolution sample inference from Illumina amplicon data. *Nat Methods* **13**, 581-583, doi:10.1038/nmeth.3869 (2016).

101 Amir, A. *et al.* Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns. *mSystems* **2**, e00191-00116, doi:10.1128/mSystems.00191-16 (2017).

102 Nearing, J. T., Douglas, G. M., Comeau, A. M. & Langille, M. G. I. Denoising the Denoisers: an independent evaluation of microbiome sequence error-correction approaches. *PeerJ* **6**, e5364, doi:10.7717/peerj.5364 (2018).

103 Prodan, A. *et al.* Comparing bioinformatic pipelines for microbial 16S rRNA amplicon sequencing. *PLoS One* **15**, e0227434, doi:10.1371/journal.pone.0227434 (2020).

104 Sun, D.-L., Jiang, X., Wu, Q. L. & Zhou, N.-Y. Intragenomic Heterogeneity of 16S rRNA Genes Causes Overestimation of Prokaryotic Diversity. *Applied and Environmental Microbiology* **79**, 5962, doi:10.1128/AEM.01282-13 (2013).

105 Earl, J. P. *et al.* Species-level bacterial community profiling of the healthy sinonasal microbiome using Pacific Biosciences sequencing of full-length 16S rRNA genes. *Microbiome* **6**, 190, doi:10.1186/s40168-018-0569-2 (2018).

106 Murali, A., Bhargava, A. & Wright, E. S. IDTAXA: a novel approach for accurate taxonomic classification of microbiome sequences. *Microbiome* **6**, 140, doi:10.1186/s40168-018-0521-5 (2018).

107 Quast, C. *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* **41**, D590-596, doi:10.1093/nar/gks1219 (2013).

108 Cole, J. R. *et al.* Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Res* **42**, D633-642, doi:10.1093/nar/gkt1244 (2014).

109 McDonald, D. *et al.* An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *The ISME journal* **6**, 610-618 (2012).

110 Balvočiūtė, M. & Huson, D. H. SILVA, RDP, Greengenes, NCBI and OTT—how do these taxonomies compare? *BMC genomics* **18**, 1-8 (2017).

111 Fisher, R. A., Corbet, A. S. & Williams, C. B. The relation between the number of species and the number of individuals in a random sample of an animal population. *The Journal of Animal Ecology*, 42-58 (1943).

112 Chao, A. & Bunge, J. Estimating the number of species in a stochastic abundance model. *Biometrics* **58**, 531-539 (2002).

113 Simpson, E. H. Measurement of diversity. *nature* **163**, 688-688 (1949).

114 Shannon, C. E. A mathematical theory of communication. *The Bell system technical journal* **27**, 379-423 (1948).

115 Swenson, N. G. *et al.* Phylogenetic and functional alpha and beta diversity in temperate and tropical tree communities. *Ecology* **93**, S112-S125 (2012).

116 Jaccard, P. The distribution of the flora in the alpine zone. 1. *New phytologist* **11**, 37-50 (1912).

117 Bray, J. R. & Curtis, J. T. An Ordination of the Upland Forest Communities of Southern Wisconsin. *Ecological Monographs* **27**, 325-349, doi:10.2307/1942268 (1957).

118 Lozupone, C., Lladser, M. E., Knights, D., Stombaugh, J. & Knight, R. UniFrac: an effective distance metric for microbial community comparison. *The ISME journal* **5**, 169-172, doi:10.1038/ismej.2010.133 (2011).

119 Lozupone, C. & Knight, R. UniFrac: a new phylogenetic method for comparing microbial communities. *Appl Environ Microbiol* **71**, 8228-8235, doi:10.1128/AEM.71.12.8228-8235.2005 (2005).

120 Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V. & Egozcue, J. J. Microbiome Datasets Are Compositional: And This Is Not Optional. *Front Microbiol* **8**, 2224, doi:10.3389/fmicb.2017.02224 (2017).

121 Martino, C. *et al.* A novel sparse compositional technique reveals microbial perturbations. *MSystems* **4** (2019).

122 Fernandes, A. D. *et al.* Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. *Microbiome* **2**, 15-15, doi:10.1186/2049-2618-2-15 (2014).

123 Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15**, 550, doi:10.1186/s13059-014-0550-8 (2014).

124 Paulson, J. N., Stine, O. C., Bravo, H. C. & Pop, M. Differential abundance analysis for microbial marker-gene surveys. *Nat Methods* **10**, 1200-1202, doi:10.1038/nmeth.2658 (2013).

125 Mandal, S. *et al.* Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microb Ecol Health Dis* **26**, 27663, doi:10.3402/mehd.v26.27663 (2015).

126 Lin, H. & Peddada, S. D. Analysis of compositions of microbiomes with bias correction. *Nature Communications* **11**, 3514, doi:10.1038/s41467-020-17041-7 (2020).

127 Langille, M. G. I. *et al.* Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nature Biotechnology* **31**, 814-821, doi:10.1038/nbt.2676 (2013).

128 Asshauer, K. P., Wemheuer, B., Daniel, R. & Meinicke, P. Tax4Fun: predicting functional profiles from metagenomic 16S rRNA data. *Bioinformatics* **31**, 2882-2884, doi:10.1093/bioinformatics/btv287 (2015).

129 Douglas, G. M. *et al.* PICRUSt2 for prediction of metagenome functions. *Nature Biotechnology* **38**, 685-688, doi:10.1038/s41587-020-0548-6 (2020).

130 Narayan, N. R. *et al.* Piphillin predicts metagenomic composition and dynamics from DADA2-corrected 16S rDNA sequences. *BMC Genomics* **21**, 56, doi:10.1186/s12864-019-6427-1 (2020).

131 Das, M., Ghosh, T. S. & Jeffery, I. B. IPCO: Inference of Pathways from Co-variance analysis. *BMC Bioinformatics* **21**, 62, doi:10.1186/s12859-020-3404-2 (2020).

132 Callahan, B. J. *et al.* High-throughput amplicon sequencing of the full-length 16S rRNA gene with single-nucleotide resolution. *Nucleic Acids Research* **47**, e103-e103, doi:10.1093/nar/gkz569 (2019).

133 Johnson, J. S. *et al.* Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. *Nature Communications* **10**, 5029, doi:10.1038/s41467-019-13036-1 (2019).

134 Cusco, A., Catozzi, C., Vines, J., Sanchez, A. & Francino, O. Microbiota profiling with long amplicons using Nanopore sequencing: full-length 16S rRNA gene and the 16S-ITS-23S of the rrn operon. *F1000Res* **7**, 1755, doi:10.12688/f1000research.16817.2 (2018).

135 Velasquez-Mejia, E. P., de la Cuesta-Zuluaga, J. & Escobar, J. S. Impact of DNA extraction, sample dilution, and reagent contamination on 16S rRNA gene sequencing of human feces. *Appl Microbiol Biotechnol* **102**, 403-411, doi:10.1007/s00253-017-8583-z (2018).

136 Zhong, Z. P. *et al.* Clean Low-Biomass Procedures and Their Application to Ancient Ice Core Microorganisms. *Front Microbiol* **9**, 1094, doi:10.3389/fmicb.2018.01094 (2018).

137 Branton, W. G. *et al.* Brain microbiota disruption within inflammatory demyelinating lesions in multiple sclerosis. *Sci Rep* **6**, 37344, doi:10.1038/srep37344 (2016).

138 Karstens, L. *et al.* Community profiling of the urinary microbiota: considerations for low-biomass samples. *Nat Rev Urol* **15**, 735-749, doi:10.1038/s41585-018-0104-z (2018).

139 Hieken, T. J. *et al.* The Microbiome of Aseptically Collected Human Breast Tissue in Benign and Malignant Disease. *Sci Rep* **6**, 30751, doi:10.1038/srep30751 (2016).

140 Urbaniak, C. *et al.* The Microbiota of Breast Tissue and Its Association with Breast Cancer. *Appl Environ Microbiol* **82**, 5039-5048, doi:10.1128/AEM.01235-16 (2016).

141 Mohammadi, T., Reesink, H. W., Vandenbroucke-Grauls, C. M. & Savelkoul, P. H. Removal of contaminating DNA from commercial nucleic

acid extraction kit reagents. *J Microbiol Methods* **61**, 285-288, doi:10.1016/j.mimet.2004.11.018 (2005).

142 Corless, C. E. *et al.* Contamination and sensitivity issues with a real-time universal 16S rRNA PCR. *J Clin Microbiol* **38**, 1747-1752 (2000).

143 Salter, S. J. *et al.* Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol* **12**, 87, doi:10.1186/s12915-014-0087-z (2014).

144 Glassing, A., Dowd, S. E., Galandiuk, S., Davis, B. & Chiodini, R. J. Inherent bacterial DNA contamination of extraction and sequencing reagents may affect interpretation of microbiota in low bacterial biomass samples. *Gut Pathog* **8**, 24, doi:10.1186/s13099-016-0103-7 (2016).

145 Rand, K. H. & Houck, H. Taq polymerase contains bacterial DNA of unknown origin. *Mol Cell Probes* **4**, 445-450 (1990).

146 McAlister, M. B., Kulakov, L. A., O'Hanlon, J. F., Larkin, M. J. & Ogden, K. L. Survival and nutritional requirements of three bacteria isolated from ultrapure water. *J Ind Microbiol Biotechnol* **29**, 75-82, doi:10.1038/sj.jim.7000273 (2002).

147 Shen, H., Rogelj, S. & Kieft, T. L. Sensitive, real-time PCR detects low-levels of contamination by Legionella pneumophila in commercial reagents. *Mol Cell Probes* **20**, 147-153, doi:10.1016/j.mcp.2005.09.007 (2006).

148 Eisenhofer, R. *et al.* Contamination in Low Microbial Biomass Microbiome Studies: Issues and Recommendations. *Trends Microbiol* **27**, 105-117, doi:10.1016/j.tim.2018.11.003 (2019).

149 Kim, D. *et al.* Optimizing methods and dodging pitfalls in microbiome research. *Microbiome* **5**, 52, doi:10.1186/s40168-017-0267-5 (2017).

150 Kirstahler, P. *et al.* Genomics-Based Identification of Microorganisms in Human Ocular Body Fluid. *Sci Rep* **8**, 4126, doi:10.1038/s41598-018-22416-4 (2018).

151 Champlot, S. *et al.* An efficient multistrategy DNA decontamination procedure of PCR reagents for hypersensitive PCR applications. *PloS one* **5**, e13042, doi:10.1371/journal.pone.0013042 (2010).

152 Elliott, D. R. F., Walker, A. W., O'Donovan, M., Parkhill, J. & Fitzgerald, R. C. A non-endoscopic device to sample the oesophageal microbiota: a case-control study. *Lancet Gastroenterol Hepatol* **2**, 32-42, doi:10.1016/S2468-1253(16)30086-3 (2017).

153 Costello, M. *et al.* Characterization and remediation of sample index swaps by non-redundant dual indexing on massively parallel sequencing platforms. *BMC Genomics* **19**, 332, doi:10.1186/s12864-018-4703-0 (2018).

154 Larsson, A. J. M., Stanley, G., Sinha, R., Weissman, I. L. & Sandberg, R. Computational correction of index switching in multiplexed sequencing libraries. *Nat Methods* **15**, 305-307, doi:10.1038/nmeth.4666 (2018).

155 Wright, E. S. & Vetsigian, K. H. Quality filtering of Illumina index reads mitigates sample cross-talk. *BMC Genomics* **17**, 876, doi:10.1186/s12864-016-3217-x (2016).

156 Jervis-Bardy, J. *et al.* Deriving accurate microbiota profiles from human samples with low bacterial content through post-sequencing processing of Illumina MiSeq data. *Microbiome* **3**, 19, doi:10.1186/s40168-015-0083-8 (2015).

157 Davis, N. M., Proctor, D. M., Holmes, S. P., Relman, D. A. & Callahan, B. J. Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. *Microbiome* **6**, 226, doi:10.1186/s40168-018-0605-2 (2018).

158 Knights, D. *et al.* Bayesian community-wide culture-independent microbial source tracking. *Nat Methods* **8**, 761-763, doi:10.1038/nmeth.1650 (2011).

159 Funkhouser, L. J. & Bordenstein, S. R. Mom knows best: the universality of maternal microbial transmission. *PLoS Biol* **11**, e1001631, doi:10.1371/journal.pbio.1001631 (2013).

160 Amarasekara, R., Jayasekara, R. W., Senanayake, H. & Dissanayake, V. H. Microbiome of the placenta in pre-eclampsia supports the role of bacteria in the multifactorial cause of pre-eclampsia. *J Obstet Gynaecol Res* **41**, 662-669, doi:10.1111/jog.12619 (2015).

161 Antony, K. M. *et al.* The preterm placental microbiome varies in association with excess maternal gestational weight gain. *Am J Obstet Gynecol* **212**, 653 e651-616, doi:10.1016/j.ajog.2014.12.041 (2015).

162 Zheng, J. *et al.* The Placental Microbiome Varies in Association with Low Birth Weight in Full-Term Neonates. *Nutrients* **7**, 6924-6937, doi:10.3390/nu7085315 (2015).

163 Bassols, J. *et al.* Gestational diabetes is associated with changes in placental microbiota and microbiome. *Pediatr Res* **80**, 777-784, doi:10.1038/pr.2016.155 (2016).

164 Collado, M. C., Rautava, S., Aakko, J., Isolauri, E. & Salminen, S. Human gut colonisation may be initiated in utero by distinct microbial communities in the placenta and amniotic fluid. *Sci Rep* **6**, 23129, doi:10.1038/srep23129 (2016).

165 Leiby, J. S. *et al.* Lack of detection of a human placenta microbiome in samples from preterm and term deliveries. *Microbiome* **6**, 196, doi:10.1186/s40168-018-0575-4 (2018).

166 Lauder, A. P. *et al.* Comparison of placenta samples with contamination controls does not provide evidence for a distinct placenta microbiota. *Microbiome* **4**, 29, doi:10.1186/s40168-016-0172-3 (2016).

167 de Goffau, M. C. *et al.* Human placenta has no microbiome but can contain potential pathogens. *Nature*, doi:10.1038/s41586-019-1451-5 (2019).

168 Theis, K. R. *et al.* Does the human placenta delivered at term have a microbiota? Results of cultivation, quantitative real-time PCR, 16S rRNA gene sequencing, and metagenomics. *Am J Obstet Gynecol* **220**, 267 e261-267 e239, doi:10.1016/j.ajog.2018.10.018 (2019).

169 Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646-674, doi:10.1016/j.cell.2011.02.013 (2011).

170 Hanahan, D. & Weinberg, R. A. The hallmarks of cancer. *Cell* **100**, 57-70 (2000).

171 Jolly, C. & Van Loo, P. Timing somatic events in the evolution of cancer. *Genome Biol* **19**, 95, doi:10.1186/s13059-018-1476-3 (2018).

172 Bray, F. *et al.* Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* **68**, 394-424, doi:10.3322/caac.21492 (2018).

173 McGuire, S. World Cancer Report 2014. Geneva, Switzerland: World Health Organization, International Agency for Research on Cancer, WHO Press, 2015. *Adv Nutr* **7**, 418-419, doi:10.3945/an.116.012211 (2016).

174 Ma, J., Ward, E. M., Siegel, R. L. & Jemal, A. Temporal Trends in Mortality in the United States, 1969-2013. *Jama* **314**, 1731-1739, doi:10.1001/jama.2015.12319 (2015).

175 Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics, 2019. *CA Cancer J Clin* **69**, 7-34, doi:10.3322/caac.21551 (2019).

176 Song, M., Vogelstein, B., Giovannucci, E. L., Willett, W. C. & Tomasetti, C. Cancer prevention: Molecular and epidemiologic consensus. *Science* **361**, 1317-1318, doi:10.1126/science.aau3830 (2018).

177 Eckhouse, S., Lewison, G. & Sullivan, R. Trends in the global funding and activity of cancer research. *Mol Oncol* **2**, 20-32, doi:10.1016/j.molonc.2008.03.007 (2008).

178 Islami, F. *et al.* Proportion and number of cancer cases and deaths attributable to potentially modifiable risk factors in the United States. *CA Cancer J Clin* **68**, 31-54, doi:10.3322/caac.21440 (2018).

179 Alexandrov, L. B. *et al.* Clock-like mutational processes in human somatic cells. *Nat Genet* **47**, 1402-1407, doi:10.1038/ng.3441 (2015).

180 de Martel, C., Georges, D., Bray, F., Ferlay, J. & Clifford, G. M. Global burden of cancer attributable to infections in 2018: a worldwide incidence analysis. *Lancet Glob Health* **8**, e180-e190, doi:10.1016/s2214-109x(19)30488-7 (2020).

181 Eslami-S, Z., Majidzadeh-A, K., Halvaei, S., Babapirali, F. & Esmaeili, R. Microbiome and Breast Cancer: New Role for an Ancient Population. *Front Oncol* **10**, 120-120, doi:10.3389/fonc.2020.00120 (2020).

182 Yu, L.-X. & Schwabe, R. F. The gut microbiome and liver cancer: mechanisms and clinical translation. *Nature Reviews Gastroenterology & Hepatology* **14**, 527-539, doi:10.1038/nrgastro.2017.72 (2017).

183 Half, E. *et al.* Fecal microbiome signatures of pancreatic cancer patients. *Scientific Reports* **9**, 16801, doi:10.1038/s41598-019-53041-4 (2019).

184 Nejman, D. *et al.* The human tumor microbiome is composed of tumor type–specific intracellular bacteria. *Science* **368**, 973, doi:10.1126/science.aay9189 (2020).

185 Parhi, L. *et al.* Breast cancer colonization by Fusobacterium nucleatum accelerates tumor growth and metastatic progression. *Nature Communications* **11**, 3259, doi:10.1038/s41467-020-16967-2 (2020).

186 Pushalkar, S. *et al.* The Pancreatic Cancer Microbiome Promotes Oncogenesis by Induction of Innate and Adaptive Immune Suppression. *Cancer Discov* **8**, 403-416, doi:10.1158/2159-8290.CD-17-1134 (2018).

187 Riquelme, E. *et al.* Tumor Microbiome Diversity and Composition Influence Pancreatic Cancer Outcomes. *Cell* **178**, 795-806.e712, doi:10.1016/j.cell.2019.07.008 (2019).

188 Aykut, B. *et al.* The fungal mycobiome promotes pancreatic oncogenesis via activation of MBL. *Nature* **574**, 264-267, doi:10.1038/s41586-019-1608-2 (2019).

189 Apopa, P. L. *et al.* PARP1 Is Up-Regulated in Non-small Cell Lung Cancer Tissues in the Presence of the Cyanobacterial Toxin Microcystin. *Frontiers in microbiology* **9**, 1757-1757, doi:10.3389/fmicb.2018.01757 (2018).

190 Signat, B., Roques, C., Poulet, P. & Duffaut, D. Fusobacterium nucleatum in periodontal health and disease. *Curr Issues Mol Biol* **13**, 25-36 (2011).

191 Velsko, I. M. *et al.* Fusobacterium nucleatum Alters Atherosclerosis Risk Factors and Enhances Inflammatory Markers with an Atheroprotective Immune Response in ApoE(null) Mice. *PLoS One* **10**, e0129795, doi:10.1371/journal.pone.0129795 (2015).

192 Strauss, J. *et al.* Invasive potential of gut mucosa-derived Fusobacterium nucleatum positively correlates with IBD status of the host. *Inflamm Bowel Dis* **17**, 1971-1978, doi:10.1002/ibd.21606 (2011).

193 Gohar, A., Jamous, F. & Abdallah, M. Concurrent fusobacterial pyogenic liver abscess and empyema. *BMJ Case Rep* **12**, doi:10.1136/bcr-2019-231994 (2019).

194 Vander Haar, E. L., So, J., Gyamfi-Bannerman, C. & Han, Y. W. Fusobacterium nucleatum and adverse pregnancy outcomes: Epidemiological and mechanistic evidence. *Anaerobe* **50**, 55-59, doi:10.1016/j.anaerobe.2018.01.008 (2018).

195 Castellarin, M. *et al.* Fusobacterium nucleatum infection is prevalent in human colorectal carcinoma. *Genome Res* **22**, 299-306, doi:10.1101/gr.126516.111 (2012).

196 Kostic, A. D. *et al.* Genomic analysis identifies association of Fusobacterium with colorectal carcinoma. *Genome Res* **22**, 292-298, doi:10.1101/gr.126573.111 (2012).

197 Gethings-Behncke, C. *et al.* <em>Fusobacterium nucleatum</em> in the Colorectum and Its Association with Cancer Risk and Survival: A Systematic Review and Meta-analysis. *Cancer Epidemiology Biomarkers &amp; Prevention* **29**, 539-548, doi:10.1158/1055-9965.Epi-18-1295 (2020).

198 Wirbel, J. *et al.* Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. *Nat Med* **25**, 679-689, doi:10.1038/s41591-019-0406-6 (2019).

199 Mehta, R. S. *et al.* Association of Dietary Patterns With Risk of Colorectal Cancer Subtypes Classified by Fusobacterium nucleatum in Tumor Tissue. *JAMA Oncol* **3**, 921-927, doi:10.1001/jamaoncol.2016.6374 (2017).

200 McCoy, A. N. *et al.* Fusobacterium is associated with colorectal adenomas. *PLoS One* **8**, e53653, doi:10.1371/journal.pone.0053653 (2013).

201 Komiya, Y. *et al.* Patients with colorectal cancer have identical strains of Fusobacterium nucleatum in their colorectal cancer and oral cavity. *Gut* **68**, 1335-1337, doi:10.1136/gutjnl-2018-316661 (2019).

202 Purcell, R. V., Visnovska, M., Biggs, P. J., Schmeier, S. & Frizelle, F. A. Distinct gut microbiome patterns associate with consensus molecular subtypes of colorectal cancer. *Scientific reports* **7**, 11590-11590, doi:10.1038/s41598-017-11237-6 (2017).

203 Guinney, J. *et al.* The consensus molecular subtypes of colorectal cancer. *Nature Medicine* **21**, 1350-1356, doi:10.1038/nm.3967 (2015).

204 Mima, K. *et al.* Fusobacterium nucleatum in Colorectal Carcinoma Tissue According to Tumor Location. *Clinical and translational gastroenterology* **7**, e200-e200, doi:10.1038/ctg.2016.53 (2016).

205 Flemer, B. *et al.* The oral microbiota in colorectal cancer is distinctive and predictive. *Gut* **67**, 1454-1463, doi:10.1136/gutjnl-2017-314814 (2018).

206 Schmidt, T. S. *et al.* Extensive transmission of microbes along the gastrointestinal tract. *Elife* **8**, doi:10.7554/eLife.42693 (2019).

207 Lockhart, P. B. *et al.* Bacteremia associated with toothbrushing and dental extraction. *Circulation* **117**, 3118-3125, doi:10.1161/CIRCULATIONAHA.107.758524 (2008).

208 Abed, J. *et al.* Fap2 Mediates Fusobacterium nucleatum Colorectal Adenocarcinoma Enrichment by Binding to Tumor-Expressed Gal-GalNAc. *Cell Host Microbe* **20**, 215-225, doi:10.1016/j.chom.2016.07.006 (2016).

209 Rubinstein, M. R. *et al.* Fusobacterium nucleatum promotes colorectal carcinogenesis by modulating E-cadherin/β-catenin signaling via its FadA adhesin. *Cell host & microbe* **14**, 195-206 (2013).

210 Rubinstein, M. R. *et al.* Fusobacterium nucleatum promotes colorectal cancer by inducing Wnt/β-catenin modulator Annexin A1. *EMBO Rep* **20**, doi:10.15252/embr.201847638 (2019).

211 Yang, Y. *et al.* Fusobacterium nucleatum Increases Proliferation of Colorectal Cancer Cells and Tumor Development in Mice by Activating Toll-Like Receptor 4 Signaling to Nuclear Factor-κB, and Up-regulating Expression of MicroRNA-21. *Gastroenterology* **152**, 851-866.e824, doi:10.1053/j.gastro.2016.11.018 (2017).

212 Kostic, A. D. *et al.* Fusobacterium nucleatum potentiates intestinal tumorigenesis and modulates the tumor-immune microenvironment. *Cell Host Microbe* **14**, 207-215, doi:10.1016/j.chom.2013.07.007 (2013).

213 Gur, C. *et al.* Binding of the Fap2 protein of Fusobacterium nucleatum to human inhibitory receptor TIGIT protects tumors from immune cell attack. *Immunity* **42**, 344-355, doi:10.1016/j.immuni.2015.01.010 (2015).

214 Chen, Y. *et al.* Fusobacterium nucleatum Promotes Metastasis in Colorectal Cancer by Activating Autophagy Signaling via the Upregulation of CARD3 Expression. *Theranostics* **10**, 323-339, doi:10.7150/thno.38870 (2020).

215 Chen, S. *et al.* Fusobacterium nucleatum promotes colorectal cancer metastasis by modulating KRT7-AS/KRT7. *Gut Microbes*, 1-15, doi:10.1080/19490976.2019.1695494 (2020).

216 Bullman, S. *et al.* Analysis of &lt;em&gt;Fusobacterium&lt;/em&gt; persistence and antibiotic response in colorectal cancer. *Science* **358**, 1443, doi:10.1126/science.aal5240 (2017).

217 Casasanta, M. A. *et al.* <em>Fusobacterium nucleatum</em> host-cell binding and invasion induces IL-8 and CXCL1 secretion that drives colorectal cancer cell migration. *Science Signaling* **13**, eaba9157, doi:10.1126/scisignal.aba9157 (2020).

218 Yu, T. *et al.* Fusobacterium nucleatum Promotes Chemoresistance to Colorectal Cancer by Modulating Autophagy. *Cell* **170**, 548-563.e516, doi:10.1016/j.cell.2017.07.008 (2017).

219 Zheng, D.-W. *et al.* Phage-guided modulation of the gut microbiota of mouse models of colorectal cancer augments their responses to chemotherapy. *Nature Biomedical Engineering* **3**, 717-728, doi:10.1038/s41551-019-0423-2 (2019).

220 Wong, S. H. *et al.* Quantitation of faecal &lt;em&gt;Fusobacterium&lt;/em&gt; improves faecal immunochemical test in detecting advanced colorectal neoplasia. *Gut* **66**, 1441, doi:10.1136/gutjnl-2016-312766 (2017).

221 Mosca, A., Miragliotta, L., Iodice, M. A., Abbinante, A. & Miragliotta, G. Antimicrobial profiles of Prevotella spp. and Fusobacterium nucleatum isolated from periodontal infections in a selected area of southern Italy. *Int J Antimicrob Agents* **30**, 521-524, doi:10.1016/j.ijantimicag.2007.07.022 (2007).

222 Loozen, G. *et al.* Effect of Bdellovibrio bacteriovorus HD100 on multispecies oral communities. *Anaerobe* **35**, 45-53, doi:10.1016/j.anaerobe.2014.09.011 (2015).

223 Kabwe, M. *et al.* Genomic, morphological and functional characterisation of novel bacteriophage FNU1 capable of disrupting Fusobacterium nucleatum biofilms. *Scientific Reports* **9**, 9107, doi:10.1038/s41598-019-45549-6 (2019).

224 Larkin, J. *et al.* Combined Nivolumab and Ipilimumab or Monotherapy in Untreated Melanoma. *The New England journal of medicine* **373**, 23-34, doi:10.1056/NEJMoa1504030 (2015).

225 Borghaei, H. *et al.* Nivolumab versus Docetaxel in Advanced Nonsquamous Non-Small-Cell Lung Cancer. *N Engl J Med* **373**, 1627-1639, doi:10.1056/NEJMoa1507643 (2015).

226 Rini, B. I. *et al.* Pembrolizumab plus Axitinib versus Sunitinib for Advanced Renal-Cell Carcinoma. *N Engl J Med* **380**, 1116-1127, doi:10.1056/NEJMoa1816714 (2019).

227 Seiwert, T. Y. *et al.* Safety and clinical activity of pembrolizumab for treatment of recurrent or metastatic squamous cell carcinoma of the head and neck (KEYNOTE-012): an open-label, multicentre, phase 1b trial. *Lancet Oncol* **17**, 956-965, doi:10.1016/s1470-2045(16)30066-3 (2016).

228 Bellmunt, J. *et al.* Pembrolizumab as Second-Line Therapy for Advanced Urothelial Carcinoma. *N Engl J Med* **376**, 1015-1026, doi:10.1056/NEJMoa1613683 (2017).

229 Larkin, J. *et al.* Five-Year Survival with Combined Nivolumab and Ipilimumab in Advanced Melanoma. *N Engl J Med* **381**, 1535-1546, doi:10.1056/NEJMoa1910836 (2019).

230 Marabelle, A. *et al.* Association of tumour mutational burden with outcomes in patients with advanced solid tumours treated with pembrolizumab: prospective biomarker analysis of the multicohort, open-label, phase 2 KEYNOTE-158 study. *Lancet Oncol* **21**, 1353-1365, doi:10.1016/s1470-2045(20)30445-9 (2020).

231 Nowicki, T. S., Hu-Lieskovan, S. & Ribas, A. Mechanisms of Resistance to PD-1 and PD-L1 Blockade. *Cancer J* **24**, 47-53, doi:10.1097/PPO.0000000000000303 (2018).

232 Routy, B. *et al.* Gut microbiome influences efficacy of PD-1-based immunotherapy against epithelial tumors. *Science* **359**, 91-97, doi:10.1126/science.aan3706 (2018).

233 Gopalakrishnan, V. *et al.* Gut microbiome modulates response to anti-PD-1 immunotherapy in melanoma patients. *Science* **359**, 97-103, doi:10.1126/science.aan4236 (2018).

234 Matson, V. *et al.* The commensal microbiome is associated with anti-PD-1 efficacy in metastatic melanoma patients. *Science* **359**, 104-108, doi:10.1126/science.aao3290 (2018).

235 Derosa, L. *et al.* Negative association of antibiotics on clinical activity of immune checkpoint inhibitors in patients with advanced renal cell and non-small-cell lung cancer. *Ann Oncol* **29**, 1437-1444, doi:10.1093/annonc/mdy103 (2018).

236 Pinato, D. J. *et al.* Association of Prior Antibiotic Treatment With Survival and Response to Immune Checkpoint Inhibitor Therapy in Patients With Cancer. *JAMA Oncol* **5**, 1774-1778, doi:10.1001/jamaoncol.2019.2785 (2019).

237 Hopkins, A. M., Kichenadasse, G., Karapetis, C. S., Rowland, A. & Sorich, M. J. Concomitant Antibiotic Use and Survival in Urothelial Carcinoma Treated with Atezolizumab. *Eur Urol* **78**, 540-543, doi:10.1016/j.eururo.2020.06.061 (2020).

238 Elkrief, A., Derosa, L., Kroemer, G., Zitvogel, L. & Routy, B. The negative impact of antibiotics on outcomes in cancer patients treated with immunotherapy: a new independent prognostic factor? *Ann Oncol* **30**, 1572-1579, doi:10.1093/annonc/mdz206 (2019).

239 Derosa, L. & Zitvogel, L. Antibiotics impair immunotherapy for urothelial cancer. *Nature Reviews Urology* **17**, 605-606, doi:10.1038/s41585-020-0373-1 (2020).

240 Coutzac, C. *et al.* Systemic short chain fatty acids limit antitumor effect of CTLA-4 blockade in hosts with cancer. *Nature Communications* **11**, 2168, doi:10.1038/s41467-020-16079-x (2020).

241 Nomura, M. *et al.* Association of Short-Chain Fatty Acids in the Gut Microbiome With Clinical Response to Treatment With Nivolumab or Pembrolizumab in Patients With Solid Cancer Tumors. *JAMA Network Open* **3**, e202895-e202895, doi:10.1001/jamanetworkopen.2020.2895 (2020).

242 Mager, L. F. *et al.* Microbiome-derived inosine modulates response to checkpoint inhibitor immunotherapy. *Science* **369**, 1481-1489 (2020).

243 McQuade, J. L., Ologun, G. O., Arora, R. & Wargo, J. A. Gut Microbiome Modulation Via Fecal Microbiota Transplant to Augment Immunotherapy in Patients with Melanoma or Other Cancers. *Current Oncology Reports* **22**, 74, doi:10.1007/s11912-020-00913-y (2020).

244 Westman, E. L. *et al.* Bacterial inactivation of the anticancer drug doxorubicin. *Chem Biol* **19**, 1255-1264, doi:10.1016/j.chembiol.2012.08.011 (2012).

245 Geller, L. T. *et al.* Potential role of intratumor bacteria in mediating tumor resistance to the chemotherapeutic drug gemcitabine. *Science* **357**, 1156-1160, doi:10.1126/science.aah5043 (2017).

246 Viaud, S. *et al.* The intestinal microbiota modulates the anticancer immune effects of cyclophosphamide. *Science* **342**, 971-976, doi:10.1126/science.1240537 (2013).

247 Viaud, S. *et al.* Cyclophosphamide induces differentiation of Th17 cells in cancer patients. *Cancer Res* **71**, 661-665, doi:10.1158/0008-5472.Can-10-1259 (2011).

248 Wallace, B. D. *et al.* Alleviating cancer drug toxicity by inhibiting a bacterial enzyme. *Science* **330**, 831-835, doi:10.1126/science.1191175 (2010).

249 Nejman, D. *et al.* The human tumor microbiome is composed of tumor type–specific intracellular bacteria. *Science* **368**, 973-980 (2020).

250 Kalaora, S. *et al.* Identification of bacteria-derived HLA-bound peptides in melanoma. *Nature*, 1-6 (2021).

251 Williams, M. J., Sottoriva, A. & Graham, T. A. Measuring Clonal Evolution in Cancer with Genomics. *Annu Rev Genom Hum G* **20**, 309-329, doi:10.1146/annurev-genom-083117-021712 (2019).

252 Helleday, T., Eshtad, S. & Nik-Zainal, S. Mechanisms underlying mutational signatures in human cancers. *Nat Rev Genet* **15**, 585-598, doi:10.1038/nrg3729 (2014).

253 Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415-421, doi:10.1038/nature12477 (2013).

254 Alexandrov, L. B. *et al.* The Repertoire of Mutational Signatures in Human Cancer. *bioRxiv*, 322859, doi:10.1101/322859 (2019).

255 Baker, M. Structural variation: the genome's hidden architecture. *Nat Methods* **9**, 133-137, doi:10.1038/nmeth.1858 (2012).

256 Nik-Zainal, S. *et al.* Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534**, 47-54, doi:10.1038/nature17676 (2016).

257 Tomasetti, C., Li, L. & Vogelstein, B. Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention. *Science* **355**, 1330-1334, doi:10.1126/science.aaf9011 (2017).

258 Tomasetti, C. & Vogelstein, B. Cancer etiology. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. *Science* **347**, 78-81, doi:10.1126/science.1260825 (2015).

259 Thomas, R. M. & Jobin, C. The Microbiome and Cancer: Is the 'Oncobiome' Mirage Real? *Trends Cancer* **1**, 24-35, doi:10.1016/j.trecan.2015.07.005 (2015).

260 Belkaid, Y. & Hand, T. W. Role of the microbiota in immunity and inflammation. *Cell* **157**, 121-141, doi:10.1016/j.cell.2014.03.011 (2014).

261 Thomas, J. P., Parker, A., Divekar, D., Pin, C. & Watson, A. The Gut Microbiota Influences Intestinal Epithelial Proliferative Potential. *Gut* **67**, A204-A204, doi:10.1136/gutjnl-2018-BSGAbstracts.407 (2018).

262 Matsumoto, Y. *et al.* Helicobacter pylori infection triggers aberrant expression of activation-induced cytidine deaminase in gastric epithelium. *Nat Med* **13**, 470-476, doi:10.1038/nm1566 (2007).

263 Sonohara, Y. *et al.* Acetaldehyde forms covalent GG intrastrand crosslinks in DNA. *Sci Rep* **9**, 660, doi:10.1038/s41598-018-37239-6 (2019).

264 Dziubańska-Kusibab, P. J. *et al.* Colibactin DNA damage signature indicates causative role in colorectal cancer. *bioRxiv*, 819854, doi:10.1101/819854 (2019).

265 Maddocks, O. D., Scanlon, K. M. & Donnenberg, M. S. An Escherichia coli effector protein promotes host mutation via depletion of DNA mismatch repair proteins. *MBio* **4**, e00152-00113, doi:10.1128/mBio.00152-13 (2013).

266 Kim, J. J. *et al.* Helicobacter pylori impairs DNA mismatch repair in gastric epithelial cells. *Gastroenterology* **123**, 542-553, doi:10.1053/gast.2002.34751 (2002).

267 Kuypers, M. M. M., Marchant, H. K. & Kartal, B. The microbial nitrogen-cycling network. *Nat Rev Microbiol* **16**, 263-276, doi:10.1038/nrmicro.2018.9 (2018).

268 Cao, W. Endonuclease V: an unusual enzyme for repair of DNA deamination. *Cell Mol Life Sci* **70**, 3145-3156, doi:10.1007/s00018-012-1222-z (2013).

269 Shinmura, K. *et al.* Mutation Spectrum Induced by 8-Bromoguanine, a Base Damaged by Reactive Brominating Species, in Human Cells. *Oxid Med Cell Longev* **2017**, 7308501, doi:10.1155/2017/7308501 (2017).

270 Fedeles, B. I. *et al.* Intrinsic mutagenic properties of 5-chlorocytosine: A mechanistic connection between chronic inflammation and cancer. *P Natl Acad Sci USA* **112**, E4571-E4580, doi:10.1073/pnas.1507709112 (2015).

271 Gungor, N. *et al.* Genotoxic effects of neutrophils and hypochlorous acid. *Mutagenesis* **25**, 149-154, doi:10.1093/mutage/gep053 (2010).

272 Shrivastav, N., Li, D. & Essigmann, J. M. Chemical biology of mutagenesis and DNA repair: cellular responses to DNA alkylation. *Carcinogenesis* **31**, 59-70, doi:10.1093/carcin/bgp262 (2010).

273 Viel, A. *et al.* A Specific Mutational Signature Associated with DNA 8-Oxoguanine Persistence in MUTYH-defective Colorectal Cancer. *EBioMedicine* **20**, 39-49, doi:10.1016/j.ebiom.2017.04.022 (2017).

274 Wang, X. *et al.* 4-hydroxy-2-nonenal mediates genotoxicity and bystander effects caused by Enterococcus faecalis-infected macrophages. *Gastroenterology* **142**, 543-551 e547, doi:10.1053/j.gastro.2011.11.020 (2012).

275 Tenaillon, O., Skurnik, D., Picard, B. & Denamur, E. The population genetics of commensal Escherichia coli. *Nat Rev Microbiol* **8**, 207-217, doi:10.1038/nrmicro2298 (2010).

276 Putze, J. *et al.* Genetic Structure and Distribution of the Colibactin Genomic Island among Members of the Family Enterobacteriaceae. *Infect Immun* **77**, 4696-4703, doi:10.1128/Iai.00522-09 (2009).

277 Cuevas-Ramos, G. *et al.* Escherichia coli induces DNA damage in vivo and triggers genomic instability in mammalian cells. *Proc Natl Acad Sci U S A* **107**, 11537-11542, doi:10.1073/pnas.1001261107 (2010).

278 Nougayrede, J. P. *et al.* Escherichia coli induces DNA double-strand breaks in eukaryotic cells. *Science* **313**, 848-851, doi:10.1126/science.1127059 (2006).

279 Arthur, J. C. *et al.* Intestinal inflammation targets cancer-inducing activity of the microbiota. *Science* **338**, 120-123, doi:10.1126/science.1224820 (2012).

280 Buc, E. *et al.* High prevalence of mucosa-associated E. coli producing cyclomodulin and genotoxin in colon cancer. *PLoS One* **8**, e56964, doi:10.1371/journal.pone.0056964 (2013).

281 Dejea, C. M. *et al.* Patients with familial adenomatous polyposis harbor colonic biofilms containing tumorigenic bacteria. *Science* **359**, 592-597, doi:10.1126/science.aah3648 (2018).

282 Vizcaino, M. I. & Crawford, J. M. The colibactin warhead crosslinks DNA. *Nat Chem* **7**, 411-417, doi:10.1038/nchem.2221 (2015).

283 Xue, M., Shine, E., Wang, W., Crawford, J. M. & Herzon, S. B. Characterization of Natural Colibactin-Nucleobase Adducts by Tandem Mass Spectrometry and Isotopic Labeling. Support for DNA Alkylation by Cyclopropane Ring Opening. *Biochemistry* **57**, 6391-6394, doi:10.1021/acs.biochem.8b01023 (2018).

284 Wilson, M. R. *et al.* The human gut bacterial genotoxin colibactin alkylates DNA. *Science* **363**, doi:10.1126/science.aar7785 (2019).

285 Bossuet-Greif, N. *et al.* The Colibactin Genotoxin Generates DNA Interstrand Cross-Links in Infected Cells. *Mbio* **9**, doi:ARTN e02393-17

10.1128/mBio.02393-17 (2018).

286 Xue, M. *et al.* Structure elucidation of colibactin and its DNA cross-links. *Science* **365**, doi:10.1126/science.aax2685 (2019).

287 Jinadasa, R. N., Bloom, S. E., Weiss, R. S. & Duhamel, G. E. Cytolethal distending toxin: a conserved bacterial genotoxin that blocks cell cycle progression, leading to apoptosis of a broad range of mammalian cell lineages. *Microbiol-Sgm* **157**, 1851-1875, doi:10.1099/mic.0.049536-0 (2011).

288 Scott, D. A. & Kaper, J. B. Cloning and sequencing of the genes encoding Escherichia coli cytolethal distending toxin. *Infect Immun* **62**, 244-251 (1994).

289 Lara-Tejero, M. & Galan, J. E. CdtA, CdtB, and CdtC form a tripartite complex that is required for cytolethal distending toxin activity. *Infect Immun* **69**, 4358-4365, doi:Doi 10.1128/Iai.69.7.4358-4365.2001 (2001).

290 Lara-Tejero, M. & Galan, J. E. A bacterial toxin that controls cell cycle progression as a deoxyribonuclease I-like protein. *Science* **290**, 354-357, doi:DOI 10.1126/science.290.5490.354 (2000).

291 Nesic, D., Hsu, Y. & Stebbins, C. E. Assembly and function of a bacterial genotoxin. *Nature* **429**, 429-433, doi:10.1038/nature02532 (2004).

292 Boesze-Battaglia, K., Alexander, D., Dlakic, M. & Shenker, B. J. A Journey of Cytolethal Distending Toxins through Cell Membranes. *Front Cell Infect Microbiol* **6**, 81, doi:10.3389/fcimb.2016.00081 (2016).

293 He, Z. *et al.* Campylobacter jejuni promotes colorectal tumorigenesis through the action of cytolethal distending toxin. *Gut* **68**, 289-300, doi:10.1136/gutjnl-2018-317200 (2019).

294 Elwell, C. A. & Dreyfus, L. A. DNase I homologous residues in CdtB are critical for cytolethal distending toxin-mediated cell cycle arrest. *Mol Microbiol* **37**, 952-963 (2000).

295 Frisan, T., Cortes-Bratti, X., Chaves-Olarte, E., Stenerlow, B. & Thelestam, M. The Haemophilus ducreyi cytolethal distending toxin induces DNA double-strand breaks and promotes ATM-dependent activation of RhoA. *Cell Microbiol* **5**, 695-707 (2003).

296 Fedor, Y. *et al.* From single-strand breaks to double-strand breaks during S-phase: a new mode of action of the Escherichia coli Cytolethal Distending Toxin. *Cellular Microbiology* **15**, 1-15, doi:10.1111/cmi.12028 (2013).

297 Bezine, E. *et al.* Cell resistance to the Cytolethal Distending Toxin involves an association of DNA repair mechanisms. *Sci Rep* **6**, 36022, doi:10.1038/srep36022 (2016).

298 Dizdaroglu, M. Oxidatively induced DNA damage and its repair in cancer. *Mutat Res Rev Mutat Res* **763**, 212-245, doi:10.1016/j.mrrev.2014.11.002 (2015).

299 Bernstein, H., Bernstein, C., Payne, C. M. & Dvorak, K. Bile acids as endogenous etiologic agents in gastrointestinal cancer. *World J Gastroentero* **15**, 3329-3340, doi:10.3748/wjg.15.3329 (2009).

300 Yazici, C. *et al.* Race-dependent association of sulfidogenic bacteria with colorectal cancer. *Gut* **66**, 1983-1994, doi:10.1136/gutjnl-2016-313321 (2017).

301 Ridlon, J. M., Wolf, P. G. & Gaskins, H. R. Taurocholic acid metabolism by gut microbes and colon cancer. *Gut Microbes* **7**, 201-215, doi:10.1080/19490976.2016.1150414 (2016).

302 Attene-Ramos, M. S. *et al.* DNA damage and toxicogenomic analyses of hydrogen sulfide in human intestinal epithelial FHs 74 Int cells. *Environ Mol Mutagen* **51**, 304-314, doi:10.1002/em.20546 (2010).

303 van Loon, B., Markkanen, E. & Hubscher, U. Oxygen as a friend and enemy: How to combat the mutational potential of 8-oxo-guanine. *DNA Repair* **9**, 604-616, doi:10.1016/j.dnarep.2010.03.004 (2010).

304 Tan, X., Grollman, A. P. & Shibutani, S. Comparison of the mutagenic properties of 8-oxo-7,8-dihydro-2'-deoxyadenosine and 8-oxo-7,8-dihydro-2'-deoxyguanosine DNA lesions in mammalian cells. *Carcinogenesis* **20**, 2287-2292 (1999).

305 Whitaker, A. M., Schaich, M. A., Smith, M. R., Flynn, T. S. & Freudenthal, B. D. Base excision repair of oxidative DNA damage: from mechanism to disease. *Front Biosci (Landmark Ed)* **22**, 1493-1522, doi:10.2741/4555 (2017).

306 Satou, K., Kawai, K., Kasai, H., Harashima, H. & Kamiya, H. Mutagenic effects of 8-hydroxy-dGTP in live mammalian cells. *Free Radic Biol Med* **42**, 1552-1560, doi:10.1016/j.freeradbiomed.2007.02.024 (2007).

307 Satou, K., Harashima, H. & Kamiya, H. Mutagenic effects of 2-hydroxy-dATP on replication in a HeLa extract: induction of substitution and deletion mutations. *Nucleic Acids Res* **31**, 2570-2575 (2003).

308 Weiss, B. Evidence for mutagenesis by nitric oxide during nitrate metabolism in Escherichia coli. *J Bacteriol* **188**, 829-833, doi:10.1128/JB.188.3.829-833.2006 (2006).

309 Crane, B. R., Sudhamsu, J. & Patel, B. A. Bacterial nitric oxide synthases. *Annu Rev Biochem* **79**, 445-470, doi:10.1146/annurev-biochem-062608-103436 (2010).

310 Tiso, M. & Schechter, A. N. Nitrate reduction to nitrite, nitric oxide and ammonia by gut bacteria under physiological conditions. *PLoS One* **10**, e0119712, doi:10.1371/journal.pone.0119712 (2015).

311 Klebanoff, S. J., Kettle, A. J., Rosen, H., Winterbourn, C. C. & Nauseef, W. M. Myeloperoxidase: a front-line defender against phagocytosed microorganisms. *J Leukocyte Biol* **93**, 185-198, doi:10.1189/jlb.0712349 (2013).

312 VanderVeen, L. A., Hashim, M. F., Shyr, Y. & Marnett, L. J. Induction of frameshift and base pair substitution mutations by the major DNA adduct of the endogenous carcinogen malondialdehyde. *P Natl Acad Sci USA* **100**, 14247-14252, doi:10.1073/pnas.2332176100 (2003).

313 Zhang, D. & Frenette, P. S. Cross talk between neutrophils and the microbiota. *Blood* **133**, 2168-2177, doi:10.1182/blood-2018-11-844555 (2019).

314 Zhang, D. C. *et al.* Neutrophil ageing is regulated by the microbiome. *Nature* **525**, 528-+, doi:10.1038/nature15367 (2015).

315 Endo, Y. *et al.* Activation-induced cytidine deaminase links between inflammation and the development of colitis-associated colorectal cancers. *Gastroenterology* **135**, 889-898, doi:10.1053/j.gastro.2008.06.091 (2008).

316 Takai, A. *et al.* Targeting activation-induced cytidine deaminase prevents colon cancer development despite persistent colonic inflammation. *Oncogene* **31**, 1733-1742, doi:10.1038/onc.2011.352 (2012).

317 Kasar, S. *et al.* Whole-genome sequencing reveals activation-induced cytidine deaminase signatures during indolent chronic lymphocytic leukaemia evolution. *Nat Commun* **6**, 8866, doi:10.1038/ncomms9866 (2015).

318 Kim, S. C. *et al.* Variable phenotypes of enterocolitis in interleukin 10-deficient mice monoassociated with two different commensal bacteria. *Gastroenterology* **128**, 891-906, doi:10.1053/j.gastro.2005.02.009 (2005).

319 Wang, X. & Huycke, M. M. Extracellular superoxide production by Enterococcus faecalis promotes chromosomal instability in mammalian cells. *Gastroenterology* **132**, 551-561, doi:10.1053/j.gastro.2006.11.040 (2007).

320 Wang, X. M. *et al.* Enterococcus faecalis Induces Aneuploidy and Tetraploidy in Colonic Epithelial Cells through a Bystander Effect. *Cancer Res* **68**, 9909-9917, doi:10.1158/0008-5472.Can-08-1551 (2008).

321 Yang, Y. H. *et al.* Colon Macrophages Polarized by Commensal Bacteria Cause Colitis and Cancer through the Bystander Effect. *Transl Oncol* **6**, 596-+, doi:10.1593/tlo.13412 (2013).

322 Wang, X., Allen, T. D., Yang, Y., Moore, D. R. & Huycke, M. M. Cyclooxygenase-2 generates the endogenous mutagen trans-4-hydroxy-2-nonenal in Enterococcus faecalis-infected macrophages. *Cancer Prev Res (Phila)* **6**, 206-216, doi:10.1158/1940-6207.CAPR-12-0350 (2013).

323 Wang, X., Yang, Y. & Huycke, M. M. Commensal bacteria drive endogenous transformation and tumour stem cell marker expression through a bystander effect. *Gut* **64**, 459-468, doi:10.1136/gutjnl-2014-307213 (2015).

324 Lundberg, J. O., Weitzberg, E., Cole, J. A. & Benjamin, N. Nitrate, bacteria and human health. *Nat Rev Microbiol* **2**, 593-602, doi:10.1038/nrmicro929 (2004).

325 Lijinsky, W. N-Nitroso compounds in the diet. *Mutat Res* **443**, 129-138, doi:10.1016/s1383-5742(99)00015-0 (1999).

326 Habermeyer, M. *et al.* Nitrate and nitrite in the diet: How to assess their benefit and risk for human health. *Mol Nutr Food Res* **59**, 106-128, doi:10.1002/mnfr.201400286 (2015).

327 Bastide, N. M., Pierre, F. H. & Corpet, D. E. Heme iron from meat and risk of colorectal cancer: a meta-analysis and a review of the mechanisms involved. *Cancer Prev Res (Phila)* **4**, 177-184, doi:10.1158/1940-6207.CAPR-10-0113 (2011).

328 Bartsch, H., Pignatelli, B., Calmels, S. & Ohshima, H. Inhibition of nitrosation. *Basic Life Sci* **61**, 27-44, doi:10.1007/978-1-4615-2984-2_3 (1993).

329 Fahrer, J. & Kaina, B. O6-methylguanine-DNA methyltransferase in the defense against N-nitroso compounds and colorectal cancer. *Carcinogenesis* **34**, 2435-2442, doi:10.1093/carcin/bgt275 (2013).

330 Kucab, J. E. *et al.* A Compendium of Mutational Signatures of Environmental Agents. *Cell* **177**, 821-836 e816, doi:10.1016/j.cell.2019.03.001 (2019).

331 Boffetta, P., Hashibe, M., La Vecchia, C., Zatonski, W. & Rehm, J. The burden of cancer attributable to alcohol drinking. *Int J Cancer* **119**, 884-887, doi:10.1002/ijc.21903 (2006).

332 Garaycoechea, J. I. *et al.* Alcohol and endogenous aldehydes damage chromosomes and mutate stem cells. *Nature* **553**, 171-177, doi:10.1038/nature25154 (2018).

333 Moritani, K. *et al.* Acetaldehyde production by major oral microbes. *Oral Dis* **21**, 748-754, doi:10.1111/odi.12341 (2015).

334 Elshaghabee, F. M. *et al.* Ethanol Production by Selected Intestinal Microorganisms and Lactic Acid Bacteria Growing under Different Nutritional Conditions. *Front Microbiol* **7**, 47, doi:10.3389/fmicb.2016.00047 (2016).

335 Letouze, E. *et al.* Mutational signatures reveal the dynamic interplay of risk factors and cellular processes during liver tumorigenesis. *Nat Commun* **8**, 1315, doi:10.1038/s41467-017-01358-x (2017).

336 Li, G. M. Mechanisms and functions of DNA mismatch repair. *Cell Res* **18**, 85-98, doi:10.1038/cr.2007.115 (2008).

337 Santos, J. C. *et al.* Helicobacter pylori infection modulates the expression of miRNAs associated with DNA mismatch repair pathway. *Mol Carcinog* **56**, 1372-1379, doi:10.1002/mc.22590 (2017).

338 Boland, C. R. & Goel, A. Microsatellite instability in colorectal cancer. *Gastroenterology* **138**, 2073-2087 e2073, doi:10.1053/j.gastro.2009.12.064 (2010).

339 Lee-Six, H. *et al.* The landscape of somatic mutation in normal colorectal epithelial cells. *Nature* **574**, 532-537, doi:10.1038/s41586-019-1672-7 (2019).

340 O'Toole, P. W., Marchesi, J. R. & Hill, C. Next-generation probiotics: the spectrum from probiotics to live biotherapeutics. *Nat Microbiol* **2**, 17057, doi:10.1038/nmicrobiol.2017.57 (2017).

341 Pleguezuelos-Manzano, C. *et al.* Mutational signature in colorectal cancer caused by genotoxic pks+ E. coli. *Nature* **580**, 269-273, doi:10.1038/s41586-020-2080-8 (2020).

342 Olafsson, S. *et al.* Somatic Evolution in Non-neoplastic IBD-Affected Colon. *Cell* **182**, 672-684.e611, doi:10.1016/j.cell.2020.06.036 (2020).

343 Yang, Y., Gharaibeh, R. Z., Newsome, R. C. & Jobin, C. Amending microbiota by targeting intestinal inflammation with TNF blockade attenuates development of colorectal cancer. *Nature Cancer* **1**, 723-734, doi:10.1038/s43018-020-0078-7 (2020).

344 Times, M. Colorectal lymphoma. *Clin Colon Rectal Surg* **24**, 135-141, doi:10.1055/s-0031-1285997 (2011).

345 Luna-Perez, P. *et al.* Colorectal sarcoma: analysis of failure patterns. *J Surg Oncol* **69**, 36-40 (1998).

346 Chung, T. P. & Hunt, S. R. Carcinoid and neuroendocrine tumors of the colon and rectum. *Clin Colon Rectal Surg* **19**, 45-48, doi:10.1055/s-2006-942343 (2006).

347 Singer, M. & Mutch, M. G. Anal melanoma. *Clin Colon Rectal Surg* **19**, 78-87, doi:10.1055/s-2006-942348 (2006).

348 Dyson, T. & Draganov, P. V. Squamous cell cancer of the rectum. *World J Gastroenterol* **15**, 4380-4386 (2009).

349 Keum, N. & Giovannucci, E. Global burden of colorectal cancer: emerging trends, risk factors and prevention strategies. *Nat Rev Gastroenterol Hepatol* **16**, 713-732, doi:10.1038/s41575-019-0189-8 (2019).

350 Kuipers, E. J. *et al.* Colorectal cancer. *Nature Reviews Disease Primers*, 15065, doi:10.1038/nrdp.2015.65 (2015).

351 Zhan, T., Rindtorff, N. & Boutros, M. Wnt signaling in cancer. *Oncogene* **36**, 1461-1473, doi:10.1038/onc.2016.304 (2017).

352 Pino, M. S. & Chung, D. C. The chromosomal instability pathway in colon cancer. *Gastroenterology* **138**, 2059-2072, doi:10.1053/j.gastro.2009.12.065 (2010).

353 Nakanishi, Y., Diaz-Meco, M. T. & Moscat, J. Serrated Colorectal Cancer: The Road Less Travelled? *Trends Cancer* **5**, 742-754, doi:10.1016/j.trecan.2019.09.004 (2019).

354 Leggett, B. & Whitehall, V. Role of the serrated pathway in colorectal cancer pathogenesis. *Gastroenterology* **138**, 2088-2100, doi:10.1053/j.gastro.2009.12.066 (2010).

355 Sanz-Garcia, E., Argiles, G., Elez, E. & Tabernero, J. BRAF mutant colorectal cancer: prognosis, treatment, and new perspectives. *Ann Oncol* **28**, 2648-2657, doi:10.1093/annonc/mdx401 (2017).

356 Olén, O. *et al.* Colorectal cancer in ulcerative colitis: a Scandinavian population-based cohort study. *Lancet* **395**, 123-131, doi:10.1016/s0140-6736(19)32545-0 (2020).

357 Choi, C. R., Bakir, I. A., Hart, A. L. & Graham, T. A. Clonal evolution of colorectal cancer in IBD. *Nat Rev Gastroenterol Hepatol* **14**, 218-229, doi:10.1038/nrgastro.2017.1 (2017).

358 Yin, J. *et al.* p53 point mutations in dysplastic and cancerous ulcerative colitis lesions. *Gastroenterology* **104**, 1633-1639, doi:10.1016/0016-5085(93)90639-t (1993).

359 Galandiuk, S. *et al.* Field cancerization in the intestinal epithelium of patients with Crohn's ileocolitis. *Gastroenterology* **142**, 855-864.e858, doi:10.1053/j.gastro.2011.12.004 (2012).

360 Brentnall, T. A. *et al.* Mutations in the p53 gene: an early marker of neoplastic progression in ulcerative colitis. *Gastroenterology* **107**, 369-378, doi:10.1016/0016-5085(94)90161-9 (1994).

361 Baker, A. M. *et al.* Evolutionary history of human colitis-associated colorectal cancer. *Gut* **68**, 985-995, doi:10.1136/gutjnl-2018-316191 (2019).

362 Baran, B. *et al.* Difference Between Left-Sided and Right-Sided Colorectal Cancer: A Focused Review of Literature. *Gastroenterology Res* **11**, 264-273, doi:10.14740/gr1062w (2018).

363 Siegel, R. L. *et al.* Colorectal cancer statistics, 2020. *CA Cancer J Clin* **70**, 145-164, doi:10.3322/caac.21601 (2020).

364 Murphy, N. *et al.* Heterogeneity of Colorectal Cancer Risk Factors by Anatomical Subsite in 10 European Countries: A Multinational Cohort Study. *Clin Gastroenterol Hepatol* **17**, 1323-1331.e1326, doi:10.1016/j.cgh.2018.07.030 (2019).

365 Dekker, E., Tanis, P. J., Vleugels, J. L. A., Kasi, P. M. & Wallace, M. B. Colorectal cancer. *Lancet* **394**, 1467-1480, doi:10.1016/s0140-6736(19)32319-0 (2019).

366 Venook, A. P. *et al.* Impact of primary (1º) tumor location on overall survival (OS) and progression-free survival (PFS) in patients (pts) with metastatic colorectal cancer (mCRC): Analysis of CALGB/SWOG 80405 (Alliance). *Journal of Clinical Oncology* **34**, 3504-3504, doi:10.1200/JCO.2016.34.15_suppl.3504 (2016).

367 Loree, J. M. *et al.* Classifying Colorectal Cancer by Tumor Location Rather than Sidedness Highlights a Continuum in Mutation Profiles and Consensus Molecular Subtypes. *Clin Cancer Res* **24**, 1062-1072, doi:10.1158/1078-0432.Ccr-17-2484 (2018).

368 Ganesh, K. *et al.* Immunotherapy in colorectal cancer: rationale, challenges and potential. *Nat Rev Gastroenterol Hepatol* **16**, 361-375, doi:10.1038/s41575-019-0126-x (2019).

369 Czene, K., Lichtenstein, P. & Hemminki, K. Environmental and heritable causes of cancer among 9.6 million individuals in the Swedish Family-Cancer Database. *Int J Cancer* **99**, 260-266, doi:10.1002/ijc.10332 (2002).

370 Lichtenstein, P. *et al.* Environmental and heritable factors in the causation of cancer--analyses of cohorts of twins from Sweden, Denmark, and Finland. *N Engl J Med* **343**, 78-85, doi:10.1056/NEJM200007133430201 (2000).

371 Peters, U., Bien, S. & Zubair, N. Genetic architecture of colorectal cancer. *Gut* **64**, 1623-1636, doi:10.1136/gutjnl-2013-306705 (2015).

372 Moller, P. *et al.* Cancer incidence and survival in Lynch syndrome patients receiving colonoscopic and gynaecological surveillance: first report from the prospective Lynch syndrome database. *Gut* **66**, 464-472, doi:10.1136/gutjnl-2015-309675 (2017).

373 Law, P. J. *et al.* Association analyses identify 31 new risk loci for colorectal cancer susceptibility. *Nature Communications* **10**, 2154, doi:10.1038/s41467-019-09775-w (2019).

374 Wu, S., Powers, S., Zhu, W. & Hannun, Y. A. Substantial contribution of extrinsic risk factors to cancer development. *Nature*, doi:10.1038/nature16166 (2015).

375 Ferlay, J. *et al.* Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Int J Cancer* **136**, E359-386, doi:10.1002/ijc.29210 (2015).

376 Ji, B. T., Devesa, S. S., Chow, W. H., Jin, F. & Gao, Y. T. Colorectal cancer incidence trends by subsite in urban Shanghai, 1972-1994. *Cancer Epidemiol Biomarkers Prev* **7**, 661-666 (1998).

377 Mohandas, K. M. Colorectal cancer in India: controversies, enigmas and primary prevention. *Indian J Gastroenterol* **30**, 3-6, doi:10.1007/s12664-010-0076-2 (2011).

378 Arnold, M. *et al.* Global patterns and trends in colorectal cancer incidence and mortality. *Gut* **66**, 683-691, doi:10.1136/gutjnl-2015-310912 (2017).

379 Siegel, R., Desantis, C. & Jemal, A. Colorectal cancer statistics, 2014. *CA Cancer J Clin* **64**, 104-117, doi:10.3322/caac.21220 (2014).

380 Akimoto, N. *et al.* Rising incidence of early-onset colorectal cancer—a call to action. *Nature Reviews Clinical Oncology*, 1-14 (2020).

381 Keum, N. & Giovannucci, E. Global burden of colorectal cancer: emerging trends, risk factors and prevention strategies. *Nature reviews Gastroenterology & hepatology* **16**, 713-732 (2019).

382 Jones, J. M. CODEX-aligned dietary fiber definitions help to bridge the 'fiber gap'. *Nutr J* **13**, 34, doi:10.1186/1475-2891-13-34 (2014).

383 Reynolds, A. *et al.* Carbohydrate quality and human health: a series of systematic reviews and meta-analyses. *Lancet* **393**, 434-445, doi:10.1016/S0140-6736(18)31809-9 (2019).

384 Theodoratou, E., Timofeeva, M., Li, X., Meng, X. & Ioannidis, J. P. A. Nature, Nurture, and Cancer Risks: Genetic and Nutritional Contributions to

Cancer. *Annu Rev Nutr* **37**, 293-320, doi:10.1146/annurev-nutr-071715-051004 (2017).

385 Aune, D. *et al.* Whole grain consumption and risk of cardiovascular disease, cancer, and all cause and cause specific mortality: systematic review and dose-response meta-analysis of prospective studies. *BMJ* **353**, i2716, doi:10.1136/bmj.i2716 (2016).

386 Song, M. *et al.* Fiber Intake and Survival After Colorectal Cancer Diagnosis. *JAMA Oncol* **4**, 71-79, doi:10.1001/jamaoncol.2017.3684 (2018).

387 Yang, J., Martínez, I., Walter, J., Keshavarzian, A. & Rose, D. J. In vitro characterization of the impact of selected dietary fibers on fecal microbiota composition and short chain fatty acid production. *Anaerobe* **23**, 74-81, doi:10.1016/j.anaerobe.2013.06.012 (2013).

388 Tian, Y., Xu, Q., Sun, L., Ye, Y. & Ji, G. Short-chain fatty acids administration is protective in colitis-associated colorectal cancer development. *J Nutr Biochem* **57**, 103-109, doi:10.1016/j.jnutbio.2018.03.007 (2018).

389 Ternes, D. *et al.* Microbiome in Colorectal Cancer: How to Get from Meta-omics to Mechanism? *Trends Microbiol* **28**, 401-423, doi:10.1016/j.tim.2020.01.001 (2020).

390 Sears, C. L. & Pardoll, D. M. Perspective: alpha-bugs, their microbial partners, and the link to colon cancer. *J Infect Dis* **203**, 306-311, doi:10.1093/jinfdis/jiq061 (2011).

391 Thiele Orberg, E. *et al.* The myeloid immune signature of enterotoxigenic Bacteroides fragilis-induced murine colon tumorigenesis. *Mucosal Immunology* **10**, 421-433, doi:10.1038/mi.2016.53 (2017).

392 Tjalsma, H., Boleij, A., Marchesi, J. R. & Dutilh, B. E. A bacterial driver–passenger model for colorectal cancer: beyond the usual suspects. *Nature Reviews Microbiology* **10**, 575-582, doi:10.1038/nrmicro2819 (2012).

393 Feng, Q. *et al.* Gut microbiome development along the colorectal adenoma–carcinoma sequence. *Nature Communications* **6**, 6528, doi:10.1038/ncomms7528 (2015).

394 Flemer, B. *et al.* Tumour-associated and non-tumour-associated microbiota in colorectal cancer. *Gut* **66**, 633-643, doi:10.1136/gutjnl-2015-309595 (2017).

395 Banerjee, S., Schlaeppi, K. & van der Heijden, M. G. A. Keystone taxa as drivers of microbiome structure and functioning. *Nature Reviews Microbiology* **16**, 567-576, doi:10.1038/s41579-018-0024-1 (2018).

396 Hajishengallis, G., Darveau, R. P. & Curtis, M. A. The keystone-pathogen hypothesis. *Nat Rev Microbiol* **10**, 717-725, doi:10.1038/nrmicro2873 (2012).

397 Curtis, M. M. *et al.* The gut commensal Bacteroides thetaiotaomicron exacerbates enteric infection through modification of the metabolic

landscape. *Cell Host Microbe* **16**, 759-769, doi:10.1016/j.chom.2014.11.005 (2014).

398 Chng, K. R. *et al.* Metagenome-wide association analysis identifies microbial determinants of post-antibiotic ecological recovery in the gut. *Nature Ecology & Evolution* **4**, 1256-1267, doi:10.1038/s41559-020-1236-0 (2020).

399 Smyth, E. C. *et al.* Oesophageal cancer. *Nat Rev Dis Primers* **3**, 17048, doi:10.1038/nrdp.2017.48 (2017).

400 Thrift, A. P. The epidemic of oesophageal carcinoma: Where are we now? *Cancer Epidemiol* **41**, 88-95, doi:10.1016/j.canep.2016.01.013 (2016).

401 Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics, 2018. *CA Cancer J Clin* **68**, 7-30, doi:10.3322/caac.21442 (2018).

402 Reid, B. J., Li, X., Galipeau, P. C. & Vaughan, T. L. Barrett's oesophagus and oesophageal adenocarcinoma: time for a new synthesis. *Nat Rev Cancer* **10**, 87-101, doi:10.1038/nrc2773 (2010).

403 Hungin, A. P. S., Molloy-Bland, M. & Scarpignato, C. Revisiting Montreal: New Insights into Symptoms and Their Causes, and Implications for the Future of GERD. *Am J Gastroenterol*, doi:10.1038/s41395-018-0287-1 (2018).

404 Cook, M. B. *et al.* Cancer incidence and mortality risks in a large US Barrett's oesophagus cohort. *Gut* **67**, 418-529, doi:10.1136/gutjnl-2016-312223 (2018).

405 Curtius, K., Rubenstein, J. H., Chak, A. & Inadomi, J. M. Computational modelling suggests that Barrett's oesophagus may be the precursor of all oesophageal adenocarcinomas. *Gut*, gutjnl-2020-321598, doi:10.1136/gutjnl-2020-321598 (2020).

406 Que, J., Garman, K. S., Souza, R. F. & Spechler, S. J. Pathogenesis and Cells of Origin of Barrett's Esophagus. *Gastroenterology* **157**, 349-364.e341, doi:10.1053/j.gastro.2019.03.072 (2019).

407 Slack, J. M. W. Metaplasia and transdifferentiation: from pure biology to the clinic. *Nature Reviews Molecular Cell Biology* **8**, 369-378, doi:10.1038/nrm2146 (2007).

408 Jopling, C., Boue, S. & Izpisua Belmonte, J. C. Dedifferentiation, transdifferentiation and reprogramming: three routes to regeneration. *Nat Rev Mol Cell Biol* **12**, 79-89, doi:10.1038/nrm3043 (2011).

409 Olsen, C. M., Pandeya, N., Green, A. C., Webb, P. M. & Whiteman, D. C. Population attributable fractions of adenocarcinoma of the esophagus and gastroesophageal junction. *Am J Epidemiol* **174**, 582-590, doi:10.1093/aje/kwr117 (2011).

410 Coleman, H. G., Xie, S.-H. & Lagergren, J. The Epidemiology of Esophageal Adenocarcinoma. *Gastroenterology* **154**, 390-405, doi:https://doi.org/10.1053/j.gastro.2017.07.046 (2018).

411 Peters, Y. *et al.* Barrett oesophagus. *Nature Reviews Disease Primers* **5**, 35, doi:10.1038/s41572-019-0086-z (2019).

412 Kubo, A. *et al.* Sex-specific associations between body mass index, waist circumference and the risk of Barrett&#039;s oesophagus: a pooled analysis from the international BEACON consortium. *Gut* **62**, 1684, doi:10.1136/gutjnl-2012-303753 (2013).

413 Cook, M. B., Freedman, N. D., Gamborg, M., Sørensen, T. I. A. & Baker, J. L. Childhood body mass index in relation to future risk of oesophageal adenocarcinoma. *British Journal of Cancer* **112**, 601-607, doi:10.1038/bjc.2014.646 (2015).

414 Furer, A. *et al.* Adolescent obesity and midlife cancer risk: a population-based cohort study of 2·3 million adolescents in Israel. *Lancet Diabetes Endocrinol* **8**, 216-225, doi:10.1016/s2213-8587(20)30019-x (2020).

415 Lauby-Secretan, B. *et al.* Body Fatness and Cancer--Viewpoint of the IARC Working Group. *N Engl J Med* **375**, 794-798, doi:10.1056/NEJMsr1606602 (2016).

416 Stone, T. W., McPherson, M. & Gail Darlington, L. Obesity and Cancer: Existing and New Hypotheses for a Causal Connection. *EBioMedicine* **30**, 14-28, doi:10.1016/j.ebiom.2018.02.022 (2018).

417 Chandar, A. K. *et al.* Association of Serum Levels of Adipokines and Insulin With Risk of Barrett's Esophagus: A Systematic Review and Meta-Analysis. *Clinical gastroenterology and hepatology : the official clinical practice journal of the American Gastroenterological Association* **13**, 2241-e2179, doi:10.1016/j.cgh.2015.06.041 (2015).

418 Duggan, C. *et al.* Association between markers of obesity and progression from Barrett's esophagus to esophageal adenocarcinoma. *Clinical gastroenterology and hepatology : the official clinical practice journal of the American Gastroenterological Association* **11**, 934-943, doi:10.1016/j.cgh.2013.02.017 (2013).

419 Duggan, C. *et al.* Association between markers of obesity and progression from Barrett's esophagus to esophageal adenocarcinoma. *Clin Gastroenterol Hepatol* **11**, 934-943, doi:10.1016/j.cgh.2013.02.017 (2013).

420 Ayazi, S. *et al.* Obesity and gastroesophageal reflux: quantifying the association between body mass index, esophageal acid exposure, and lower esophageal sphincter status in a large series of patients with reflux symptoms. *J Gastrointest Surg* **13**, 1440-1447, doi:10.1007/s11605-009-0930-7 (2009).

421 Wu, J. C., Mui, L. M., Cheung, C. M., Chan, Y. & Sung, J. J. Obesity is associated with increased transient lower esophageal sphincter relaxation. *Gastroenterology* **132**, 883-889, doi:10.1053/j.gastro.2006.12.032 (2007).

422 Martínez-Jiménez, F. *et al.* A compendium of mutational cancer driver genes. *Nat Rev Cancer* **20**, 555-572, doi:10.1038/s41568-020-0290-x (2020).

423 Ross-Innes, C. S. *et al.* Whole-genome sequencing provides new insights into the clonal architecture of Barrett's esophagus and esophageal adenocarcinoma. *Nat Genet* **47**, 1038-1046, doi:10.1038/ng.3357 (2015).

424 Secrier, M. *et al.* Mutational signatures in esophageal adenocarcinoma define etiologically distinct subgroups with therapeutic relevance. *Nat Genet* **48**, 1131-1141, doi:10.1038/ng.3659 (2016).

425 Samstein, R. M. *et al.* Tumor mutational load predicts survival after immunotherapy across multiple cancer types. *Nat Genet* **51**, 202-206, doi:10.1038/s41588-018-0312-8 (2019).

426 Schumacher, T. N. & Schreiber, R. D. Neoantigens in cancer immunotherapy. *Science* **348**, 69, doi:10.1126/science.aaa4971 (2015).

427 Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer. *Nature* **578**, 94-101, doi:10.1038/s41586-020-1943-3 (2020).

428 Bhardwaj, V. *et al.* Activation of NADPH oxidases leads to DNA damage in esophageal cells. *Scientific Reports* **7**, 9956, doi:10.1038/s41598-017-09620-4 (2017).

429 Li, D. & Cao, W. Role of intracellular calcium and NADPH oxidase NOX5-S in acid-induced DNA damage in Barrett's cells and Barrett's esophageal adenocarcinoma cells. *Am J Physiol Gastrointest Liver Physiol* **306**, G863-G872, doi:10.1152/ajpgi.00321.2013 (2014).

430 Li, D. & Cao, W. Bile acid receptor TGR5, NADPH Oxidase NOX5-S and CREB Mediate Bile Acid-Induced DNA Damage In Barrett's Esophageal Adenocarcinoma Cells. *Sci Rep* **6**, 31538, doi:10.1038/srep31538 (2016).

431 Li, D., Deconda, D., Li, A., Habr, F. & Cao, W. Effect of Proton Pump Inhibitor Therapy on NOX5, mPGES1 and iNOS expression in Barrett’s Esophagus. *Scientific Reports* **9**, 16242, doi:10.1038/s41598-019-52800-7 (2019).

432 Tomkova, M., Tomek, J., Kriaucionis, S. & Schuster-Böckler, B. Mutational signature distribution varies with DNA replication timing and strand asymmetry. *Genome Biology* **19**, 129, doi:10.1186/s13059-018-1509-y (2018).

433 Christensen, S. *et al.* 5-Fluorouracil treatment induces characteristic T>G mutations in human cancer. *Nature Communications* **10**, 4571, doi:10.1038/s41467-019-12594-8 (2019).

434 Suzuki, T. & Kamiya, H. Mutations induced by 8-hydroxyguanine (8-oxo-7,8-dihydroguanine), a representative oxidized base, in mammalian cells. *Genes Environ* **39**, 2-2, doi:10.1186/s41021-016-0051-y (2016).

435 Clemons, N. J. *et al.* Nitric oxide-mediated invasion in Barrett's high-grade dysplasia and adenocarcinoma. *Carcinogenesis* **31**, 1669-1675, doi:10.1093/carcin/bgq130 (2010).

436 Wilson, K. T., Fu, S., Ramanujam, K. S. & Meltzer, S. J. Increased expression of inducible nitric oxide synthase and cyclooxygenase-2 in

Barrett's esophagus and associated adenocarcinomas. *Cancer Res* **58**, 2929-2934 (1998).

437 Contino, G., Vaughan, T. L., Whiteman, D. & Fitzgerald, R. C. The Evolving Genomic Landscape of Barrett's Esophagus and Esophageal Adenocarcinoma. *Gastroenterology* **153**, 657-673.e651, doi:10.1053/j.gastro.2017.07.007 (2017).

438 Stachler, M. D. *et al.* Paired exome analysis of Barrett's esophagus and adenocarcinoma. *Nat Genet* **47**, 1047-1055, doi:10.1038/ng.3343 (2015).

439 Nones, K. *et al.* Genomic catastrophes frequently arise in esophageal adenocarcinoma and drive tumorigenesis. *Nature Communications* **5**, 5224, doi:10.1038/ncomms6224 (2014).

440 Deshpande, N. P., Riordan, S. M., Castano-Rodriguez, N., Wilkins, M. R. & Kaakoush, N. O. Signatures within the esophageal microbiome are associated with host genetics, age, and disease. *Microbiome* **6**, 227, doi:10.1186/s40168-018-0611-4 (2018).

441 Okereke, I. C. *et al.* Microbiota of the Oropharynx and Endoscope Compared to the Esophagus. *Sci Rep* **9**, 10201, doi:10.1038/s41598-019-46747-y (2019).

442 Snider, E. J. *et al.* Alterations to the Esophageal Microbiome Associated with Progression from Barrett's Esophagus to Esophageal Adenocarcinoma. *Cancer Epidemiology Biomarkers &amp; Prevention*, doi:10.1158/1055-9965.Epi-19-0008 (2019).

443 Cardoso, R. *et al.* Colorectal cancer incidence, mortality, and stage distribution in European countries in the colorectal cancer screening era: an international population-based study. *The Lancet Oncology* (2021).

444 Yu, J. *et al.* Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer. *Gut* **66**, 70-78 (2017).

445 Flemer, B. *et al.* The oral microbiota in colorectal cancer is distinctive and predictive. *Gut*, doi:10.1136/gutjnl-2017-314814 (2017).

446 Gharaibeh, R. Z. & Jobin, C. Microbiota and cancer immunotherapy: in search of microbial signals. *Gut*, doi:10.1136/gutjnl-2018-317220 (2018).

447 Coussens, L. M. & Werb, Z. Inflammation and cancer. *Nature* **420**, 860-867 (2002).

448 Clemente, J. C., Manasson, J. & Scher, J. U. The role of the gut microbiome in systemic inflammatory disease. *Bmj* **360** (2018).

449 Sabat, R. *et al.* Hidradenitis suppurativa. *Nature Reviews Disease Primers* **6**, 1-20 (2020).

450 Jung, J. M. *et al.* Assessment of Overall and Specific Cancer Risks in Patients With Hidradenitis Suppurativa. *JAMA Dermatol* **156**, 844-853, doi:10.1001/jamadermatol.2020.1422 (2020).

3209

Chapter 2- Alterations in the oesophago-gastric mucosal microbiome in patients along the inflammation-metaplasia-dysplasia-oesophageal adenocarcinoma sequence

Draft manuscript

Authors:

Maurice Barrett, Collette K Hand, Fergus Shanahan, Thomas Murphy, Paul W O'Toole

Maurice Barrett contributed to this work as follows:

- Nucleic acid extraction from samples.
- Design of methodologies pertaining to sample processing.
- 16S rRNA gene PCR.
- Next generation sequencing library preparation.
- All bioinformatic analysis including sequence processing, compositional data analysis and statistical analysis.
- Data visualization, i.e., construction of manuscript figures.
- Writing of the manuscript.

2.1 Abstract

The incidence of oesophageal adenocarcinoma (OAC) has risen dramatically in developed countries in the past 40 years, for not completely established reasons. Major modulators of risk for OAC have been identified including obesity and gastro-oesophageal reflux disease. The microbiota has been increasingly recognised as playing a role in cancer biology including gastric and colon cancer, and a role has been proposed in oesophageal cancer. In this study we therefore defined the microbiome in multiple (per patient) gastric and oesophageal biopsies derived from a cohort of individuals with clinical presentations along the OAC transformation sequence. Furthermore, we delineated microbiome differences spatially along the upper digestive tract with respect to these clinical classifications. We identified an ASV assigned to *Fusobacterium nucleatum* that was enriched in oesophageal samples from individuals with or at increased risk of OAC. Further, we identified an ASV assigned to *Fusobacterium necrophorum* that was enriched in gastro-oesophageal junction biopsies derived from individuals who had dysplastic and neoplastic tissue relative to those that did not. These findings provide insight into differences in the oesophago-gastric mucosal microbiome features along the oesophageal adenocarcinoma sequence and may inform diagnostic strategies while also providing information on the pathoaetiology of OAC.

2.2 Introduction

In 2020, oesophageal cancer was the seventh leading cause of cancer (604,000 new cases) and the 6th leading cause of cancer mortality (544,000 deaths) worldwide[1]. As for other cancers, prognosis is dependent on stage of diagnosis[2]. The overall 5-year survival rate for oesophageal cancer is less than 20% in western populations[3]. Two major histological subtypes exist, that is, oesophageal squamous cell carcinoma (OSCC) and oesophageal adenocarcinoma (OAC). There is a distinct geographical distribution in these subtypes, with OAC being the predominant presentation in western countries[4].With regard to western countries, OAC has registered the greatest rise in incidences of all cancers and this trend has continued to increase[5].

OAC is thought to evolve through a defined sequence of histological changes, that is, normal squamous cells -> metaplastic columnar epithelium (Barrett's oesophagus) -> increasing grades of dysplasia -> adenocarcinoma of increasing stages[4]. We refer to this series of events as the OAC sequence. An intestinal metaplastic tissue known as Barrett's oesophagus (BO) develops in the oesophagus, usually near the gastroesophageal junction, as a result of injury due to reflux of gastric and bile acids into the oesophagus. Such reflux is indicative of Gastro-oesophageal Reflux Disease (GORD)[6]. BO is thought to be a precursor to at least a subset of OAC cases, with ~12% of OAC cases having a prior diagnosis of BO[7]. However, recent computational models suggest that nearly all OAC cases evolve through BE[8]. Acquisition of somatic mutations such as those in p53 enable the progression of BO to dysplasia and OAC [9].

A number of environmental risk factors have been identified in the development of OAC with obesity, GORDS and smoking accounting for 70% of cases in western populations[10]. Notably, *Helicobacter pylori* which is known to play a causative role in gastric cancer, is thought to play a protective role against the development of BO and OAC[11,12].

An accumulating body of evidence supports the hypothesis that the microbiota is a modulator of risk for various cancers[13,14]. The colon and its resident microbiota have been extensively studied to identify how this interaction pertains to CRC development and progression[13,15]. Many microbiota features have been linked to CRC oncogenesis in both a correlative and mechanistic manner. Colibactin producing pks+ *Escherichia coli* can induce mutations with a particular nucleotide mutational signature, while *Fusobacterium nucleatum* has been demonstrated to modulate the tumour microenvironment[16,17].

We hypothesise that the microbiome plays a role in the progression of the OAC sequence. Microbes capable of immunomodulation, promoting inflammation or causing DNA damage present in the oesophagus may drive OAC oncogenesis. Even if the microbiome does not play a direct role in OAC oncogenesis, one would expect histological and environmental changes in the oesophagus to be associated with changes in the oesophageal mucosal microbiome. Considering this, we expected differences in microbiome features between different clinical groups along the OAC sequence.

In this study we sought to identify association between features of the microbiota and various stages along the OAC sequence. We also investigated differences in the microbiome between sites within the upper digestive tract with respect to the various stages within the oesophageal adenocarcinoma sequence.

2.3 Methods

2.3.1 Sample collection and clinical classification

The cohort was derived from patients at Mercy University Hospital, Cork undergoing an upper gastrointestinal endoscopy and biopsy examination for the treatment of oesophagitis, Barrett's oesophagus and oesophageal cancer Healthy controls were recruited from patients undergoing upper GI endoscopy for assessment of benign gastroduodenal disorders. Patients who have taken a course of antibiotics in the preceding month were excluded from recruitment. The recruitment period was between the period of April 2016 and January 2020. This study was conducted in accordance with the ethical principles set forth in the current version of the Declaration of Helsinki, the International Conference on Harmonization E6 Good Clinical Practice (ICH-GCP). Ethical approval was granted by The Clinical Research Ethics Committee of the Cork Teaching Hospitals (Cork, Ireland). For each study participant five biopsies were obtained using disposable endoscopic biopsy forceps. One biopsy was taken from the epicentre of the cancer, Barrett's segment or focus of oesophagitis, one on 2-5cm either side of the pathology and one 10-20 cm away from either side of the pathology. Samples were placed in cryotubes and stored in a -80°C freezer. The oesophageal histological presentation of each individual was classified by a consultant pathologist. Patients were classified based on the histology they presented with which represented the latest stage of the OAC sequence e.g. those with metaplastic tissue and dysplastic tissue were classified into the dysplasia clinical group. Adenocarcinomas were located at the distal oesophagus and gastro-oesophageal junction and squamous cell carcinomas were excluded. Demographic variables of general and dental health, physical activity, smoking and alcohol consumption and proton pump inhibitor usage were collected by questionnaire.

2.3.2 Microbial DNA extraction

DNA was extracted from biopsy samples using the AllPrep DNA/RNA Mini Kit (Qiagen, Hilden, Germany) with modifications to include a bead beating step that ensured optimal lysis of microbial cell[18].

2.3.3 16S rRNA gene PCR amplification and sequencing

Amplification was performed using primers for the V3–V4 region (Table 1) of the bacterial 16S rRNA gene with added adapter overhang sequences in accordance with the Illumina 16S Metagenomic Sequencing Protocol (Illumina, California, USA) [19].

Region	**Name**	**F/R**	**Sequence**
V3–V4[20]	S-D-Bact-0341-b-S-17	F	5′ TCG TCG GCA GCG TCA GAT GTG TAT AAG AGA CAG CCT ACG GGN GGC WGC AG
	S-D-Bact-0785-a-A-21	R	5′ GTC TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA G GAC TAC HVG GGT ATC TAA TCC

Table 1. Primers used for 16S rRNA gene amplification.

The initial PCR amplification was performed using the MTP Taq DNA Polymerase (Merck KGaA, Darmstadt, Germany) with the PCR thermocycler protocol as follows: Initiation step of 94 °C for 1 min followed by 35 cycles of 94 °C for 60 s, 55 °C for 45 s, and 72 °C for 30 s, and a final extension step of 72 °C for 5 min. An index PCR was performed to attach dual indices (barcodes) and Illumina sequencing adapters as per Illumina 16S Metagenomic Sequencing Protocol (Illumina, California, USA). DNA concentration was determined using a Qubit fluorometer (Invitrogen) using the ‘High Sensitivity’ assay and samples were pooled at a standardised concentration. The pooled library was sequenced on the Illumina MiSeq platform (Illumina, California, USA) utilising 2 × 300 bp chemistry. Samples were sequenced over 4 batches.

2.3.4 Bioinformatic and biostatistical analysis

Raw nucleotide sequence data was imported into R (v3.6.0). Error model generation, denoising and the generation of an ASV table was performed using the R package DADA2 (v1.12.1)[21]. ASV taxonomy assignment, from phylum to genus level, was performed using mothur[22]. Species level taxonomy assignment was performed using SPINGO [23]. Alpha diversity was calculated using the alpha_diversity.py command within QIIME (v 1.9.1)[24]. Unifrac distance and Bray-Curtis dissimilarity was calculates using the beta_diversity.py within QIIME (v 1.9.1)[24]. The Jaccard index was calculated using the vegdist command within R package (v 2.5.7). Robust Aitchison was calculated using the gemelli auto-rpca command within QIIME2 (version 2020.11.1)[25].. Differential abundance analysis between anatomical sites was performed using the paired wilcoxon test. Differential abundance analysis between clinical classifications was performed using DESeq2. Functional genes and pathways were inferred using the picrust2_pipeline.py command within PICRUSt2[26].

2.3.5 Contamination control

As gastric and oesophageal mucosal biopsies may be considered low biomass, protocols were tailored to address the potential of contamination. Firstly, we used reagents manufactured to be microbial DNA free namely MTP Taq DNA Polymerase and microbial DNA free water (QAIGEN). We performed mock/blank extractions to detect contamination associated with reagents. Further, we also carried out PCR controls i.e. the amplification of microbial DNA free water, to detect contamination specific to the polymerase. With respect to mock extractions, we detected taxa indicative contamination including *Sphingomonas* and *Halobacillus*. However, we did not obtain usable reads with regard to the PCR control. We performed extraction positive controls using the Zymo mock community (Zymo, D6300) at various numbers of cells per extraction. Furthermore we positive amplification control using the ZymoBIOMICS mock community DNA standard (Zymo, D6305) at various DNA amounts. Both these positive controls allowed for the identification of the limit whereby contamination would become detectable in the sequencing data. With respect to extraction positive controls we detected contamination being introduced to the data at $2.8x10^3$ cells per extraction. With respect to positive amplification control we detected contamination being introduced to the data at concentration of 0.0002ng per reaction. Taking these figures, we were reassured that we had sufficient bacterial mass within our gastric and oesophageal mucosal biopsies to employ our protocols

2.4 Results

2.4.1 Patient demographics and oesophageal samples

In this study, we aimed to define the microbiome composition of mucosal biopsies from 5 positions along the upper digestive tract derived from an Irish population cohort (Table 2). These individuals represented defined stages along the OAC sequence including healthy controls, gastro-oesophageal reflux disease (GORD), Barrett's Oesophagus (BO), dysplasia, oesophageal adenocarcinoma (OAC), and metastatic oesophageal adenocarcinoma (metastatic OAC). Individuals were age and sex matched; however, there was a male sex bias. Male sex is a strong risk factor for OAC development[4].

	Controls	**GORD**	**BO**	**Dysplasia**	**OAC**	**Metastatic OAC**	**p value**
Patients (N)	12	30	38	19	36	9	
Age (mean,range)	57.9 (31-78)	57.4 (29-83)	56.8 (35-78)	64.4 (37-87)	61.1 (33-80)	62.2 (53-73)	0.226
Sex (f/m)	8/4	13/17	10/28	2/17	13/23	0/9	0.004
BMI	27.3 (19.5-33.6)	27.5 (20.2-40.5)	30.1 (15.9-58.8)	27.6 (21.0 -37.0)	28.1 (13.3-39.0)	26.0 (22.0-34.6)	0.628
Waist Circumference	90.3 (67-118)	96.8 (75-126)	102.8 (67-146)	98.5 (73-127)	100.5 (53-171)	93.8 (86-111)	0.306

Table 2. Descriptive statistics of the study cohort. Kruskal–Wallis test or χ2 statistic was used to determine significance of difference between clinical groups.

The 5 biopsies were labelled 1 to 5 and represent the following anatomical sites: Biopsy location 3 represent the epicentre of diseased tissue. For example, in the context of Barrett’s oesophagus, biopsy location 3 represents metaplastic tissue. For oesophageal adenocarcinoma, biopsy location 3 represents neoplastic tissue. Due the presentation of diseases along the OAC sequences, biopsy location 3 samples were usually derived from the gastro-oesophageal junction. Biopsy location 1 and 2 were taken approximately 2-5cm and 10-20 cm proximally from the disease epicentre (Biopsy location 3) respectively. Biopsy locations 4 and 5 were taken approximately 2-5cm and 10-20 cm distally from the disease epicentre (Biopsy location 3) respectively. Biopsy locations 4 and 5 were primally gastric in character.

After the quality checks and filtering associated with the bioinformatic pipeline (DADA2) we analysed 649 oesophageal and gastric biopsies from 144 individuals (Table 3).

	Clinical classification						
Biopsy location	**Controls**	**GORD**	**BE**	**Dysplasia**	**OAC**	**Metastatic OAC**	**Total**
1	12	30	36	19	33	8	138
2	12	28	36	17	36	9	138
3	11	30	34	19	35	9	138
4	8	26	23	18	33	8	116
5	9	28	23	18	32	9	119
Total	52	142	152	91	169	43	649

Table 3. Biopsy sample distribution with respect to biopsy location and clinical classification.

In terms of sequencing depth, the mean read number was 12,142 reads per sample with a minimum read depth of 2,077 reads and a maximum of 55,043 reads.

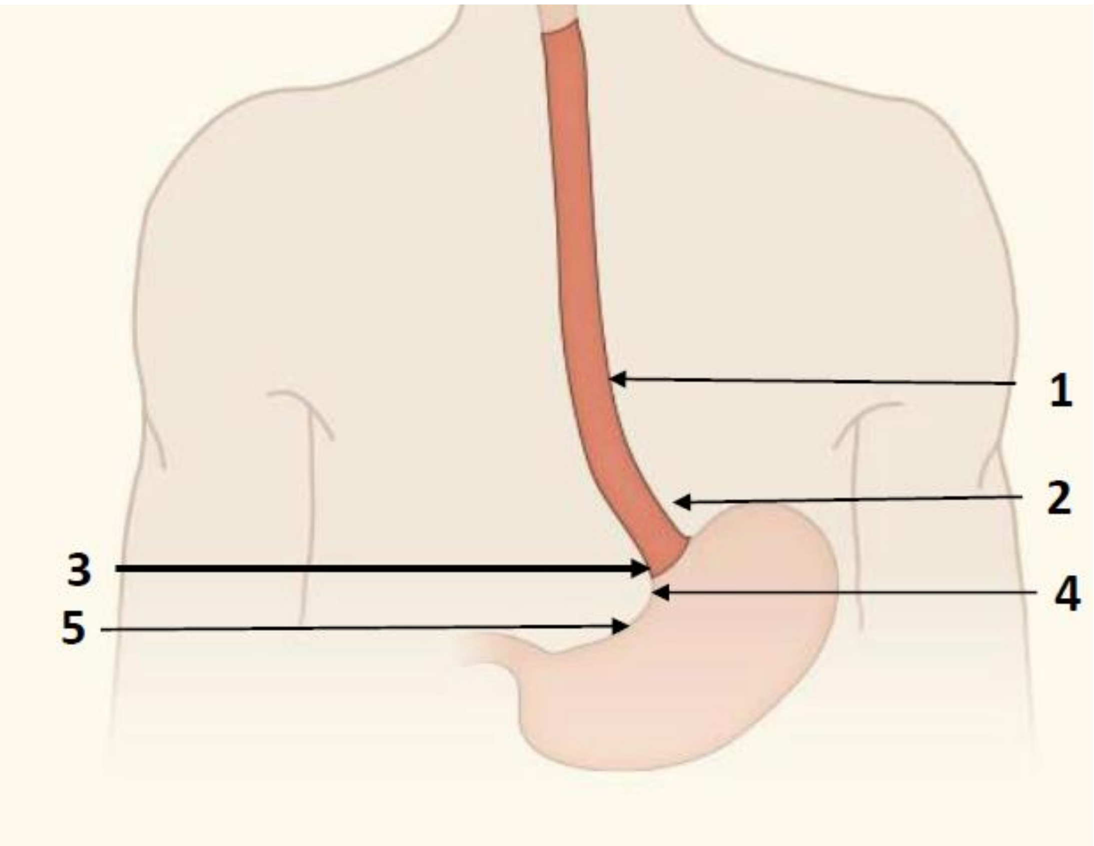
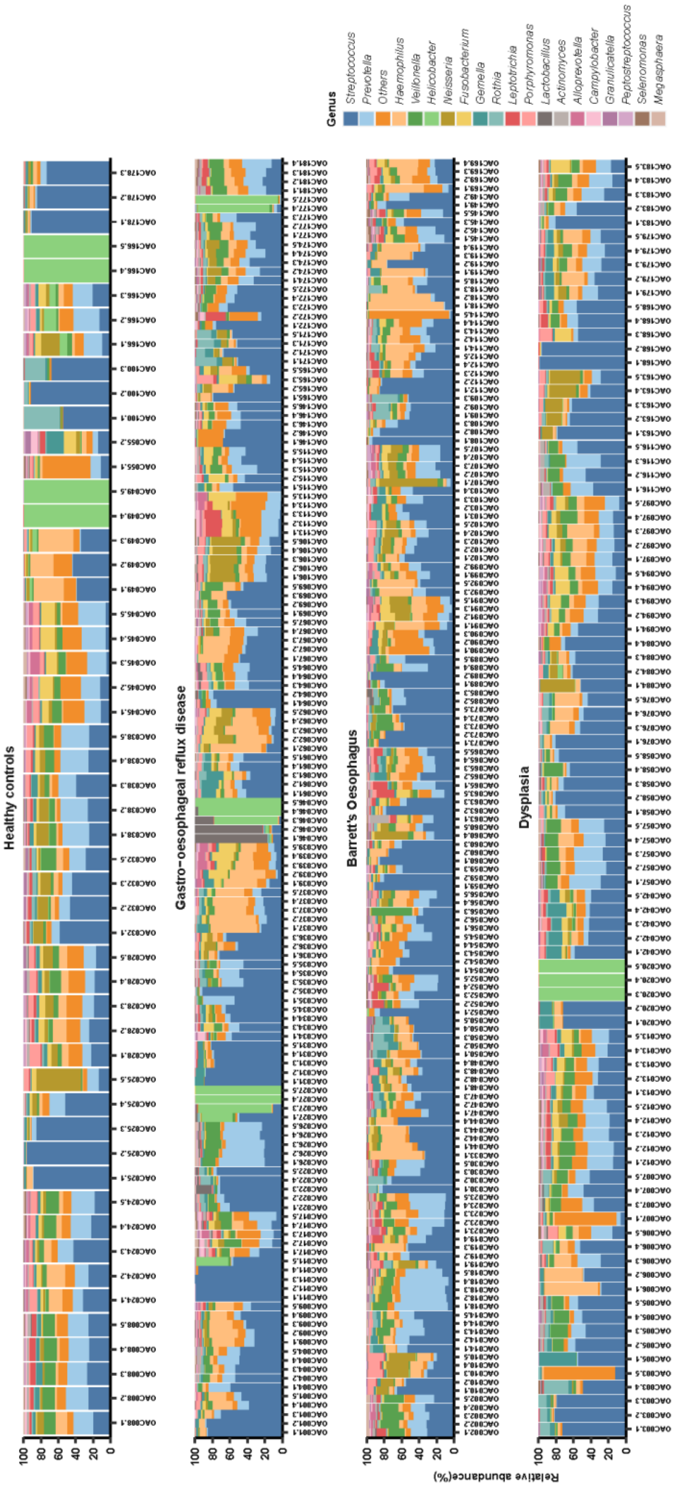


Figure 1. Diagram displaying location from where the biopsy sites and the corresponding number system.

2.4.2 Microbiome alterations with respect to clinical classifications

At the genus level, the microbiome composition of biopsy samples was predominantly composed of *Streptococcus*, *Prevotella* and *Haemophilus*, in line with previous reports describing the gastric and oesophageal microbiome (Supplementary figure 1) [27-29]. This indicated that the measures implemented to deal with potential contamination of low biomass samples were effective.



Supplementary Figure 1. Relative abundance of genera in oesophago-gastric biopsy microbiome. Bar plots of relative abundance of genera in oesophago-gastric biopsies. Samples are organised by clinical classification. Genera with a relative abundance of less 1% across all samples are grouped into 'others' with sequences not classified at the genus level.

We did not detect any major difference in beta-diversity (as measured by Bray–Curtis dissimilarity) between clinical classification groups in biopsies derived from the oesophagus, that is, biopsy location 1 and 2 (Figure1A and B). However, we did identify a significant shift in beta-diversity as measured by Bray–Curtis dissimilarity with respect to clinical classification in biopsies derived from the gastroesophageal junction (biopsy location 3) and the stomach (biopsy location 4 and 5) (Figure 1C, D and E). With respect to biopsy location 3, the anatomical focus of the disease in the respective clinical groups, the microbiome of individuals with OAC and metastatic OAC were seen to cluster while those of healthy controls, individuals with GORD and BO formed a separate cluster and individuals with dysplasia were somewhat intermediate to these two clusters (Figure 1C). Using 4 other microbiome beta-diversity metrics we did not identify any statistically significant differences in clinical groups (Supplementary table 1).

With respect to each biopsy location (1-5), we did not identify any statistically significant difference in alpha diversity with respect to clinical classification (Kruskal Wallis test; data not shown).

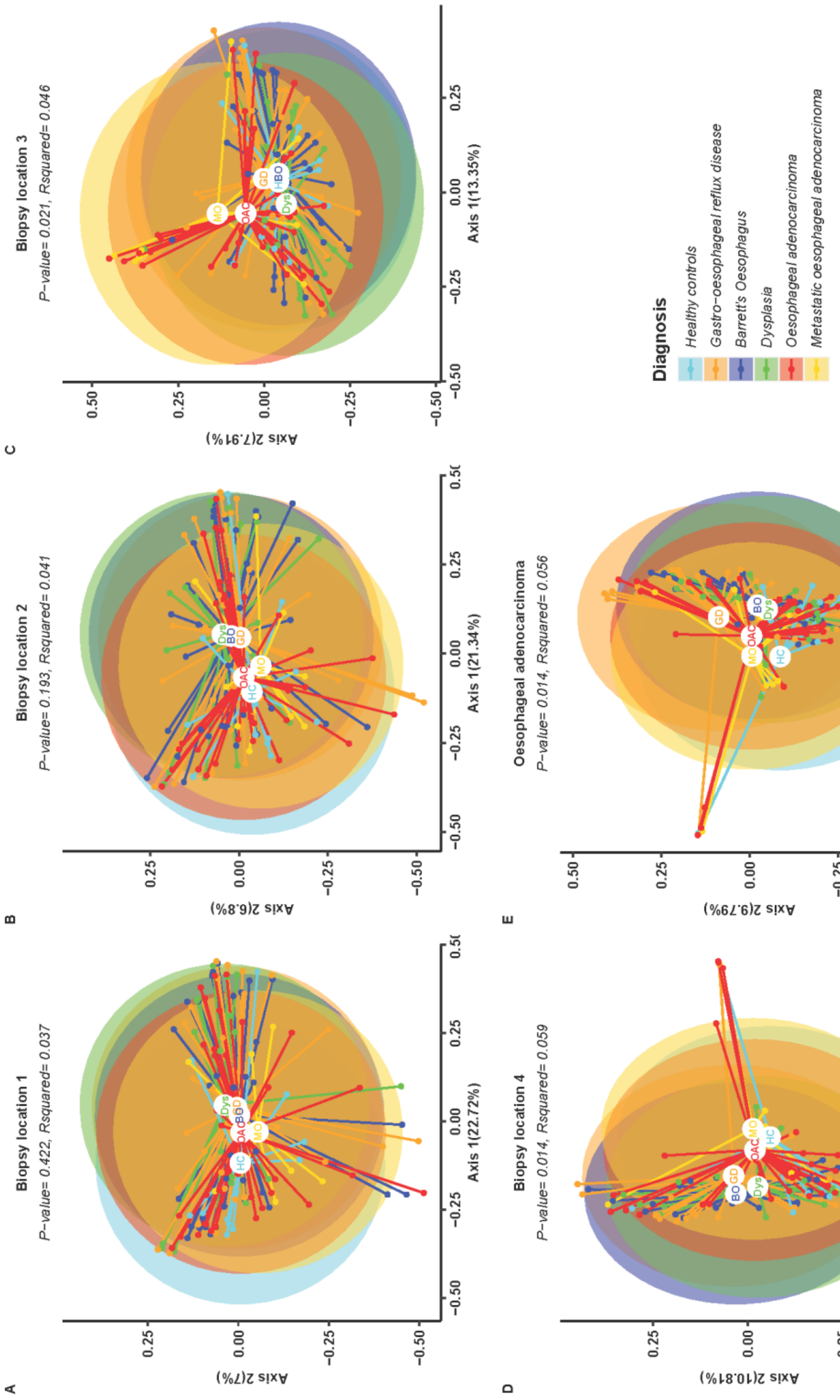


Figure 2. Beta-diversity analysis with respect to clinical classifications. Principal Coordinates Analysis (PCoA) plot representing Bray–Curtis dissimilarity. **(A)** Biopsy location 1, **(B)** Biopsy location 2. **(C)** Biopsy location 3 **(D)** Biopsy location 4 **(E)** Biopsy location 5. Statistical testing performed using Permutational Multivariate Analysis of Variance.

	Difference in Beta-diversity metrics between clinical classification per biopsy location							
Biopsy location	weighted UniFrac		unweighted UniFrac		Jaccard Index		Robust Aitchison	
	P-value	R-squared	P-value	R-squared	P-value	R-squared	P-value	R-squared
1	0.255	0.045	0.328	0.038	0.37	0.037	0.765	0.025
2	0.106	0.056	0.678	0.034	0.183	0.038	0.827	0.022
3	0.2	0.046	0.211	0.04	0.246	0.037	0.611	0.031
4	0.038	0.072	0.333	0.046	0.164	0.045	0.377	0.046
5	0.165	0.054	0.527	0.041	0.33	0.043	0.252	0.053

Supplementary table 1. Analysis of significance between biopsy microbiome beta-diversity metrics with respect to clinical classifications at each of the 5-biopsy location. P-value and R-squared calculated using Permutational multivariate analysis of variance (PERMANOVA).

2.4.3 Differentially abundant ASVs, species and metabolic pathways with respect to clinical classifications

Grouping microbiome data across all biopsy locations with a subject, we performed differential abundance analysis to identifying species and ASVs that are differentially abundant between clinical classifications.

The species *Prevotella denticola* was enriched in the diseased groups (GORD, BO, dysplasia, OAC and metastatic OAC) relative to healthy controls in samples derived from biopsy location 2 and biopsy location 3 (Figure 2 B, C). The species *Bifidobacterium dentium* was enriched in all the disease groups except metastatic relative to healthy controls in samples derived from biopsy location 1,2 and 4 (Figure 2 A, B, D).

In samples derived from the oesophagus, that is, biopsy location 1 and 2, we observed that an ASV, Seq 130, assigned to *Fusobacterium nucleatum* was enriched in biopsies derived from the disease clinical groups relative to healthy controls (Supplementary Figure 2 A, B). However, when all ASVs were binned to the species level, *F. nucleatum* was no longer detected as enriched in the disease groups. (Figure 2A, B)

In samples derived from the gastroesophageal junction, that is, biopsy location 3, we identified an ASV, Seq 52, assigned to *Fusobacterium necrophorum* which was generally enriched in samples derived from clinical groups which are later along the OAC sequence including dysplasia, OAC and metastatic OAC compared to clinical groups which are earlier along the sequence including healthy controls, GORD and BO. (Supplementary figure 2C). This observation was retained when ASVs were binned to the species level (Figure 2C). In samples derived from the stomach (biopsy location 4 and 5) this ASV assigned to *F. necrophorum* was observed to be enriched in BO, dysplasia, OAC and metastatic OAC relative to healthy controls and GORD (Supplementary figure 2D, E). Again, this observation was reflected at the species level (Figure 2D, E).

Using the algorithm DESeq2, a number of microbiome-encoded metabolic pathways were found to be differentially abundant between the clinical groups with respect to each of the biopsy locations. Particular microbiome metabolic pathways were depleted in the metastatic biopsy microbiome with respect to biopsy locations (Supplementary figure 3). In particular the microbial coding capacity for a metabolic pathway involved in Vitamin B12 production (also known as adenosylcobalamin) synthesis was depleted in the microbiome of the metastatic OAC group relative to all other clinical groups, and with respect to all biopsy locations.

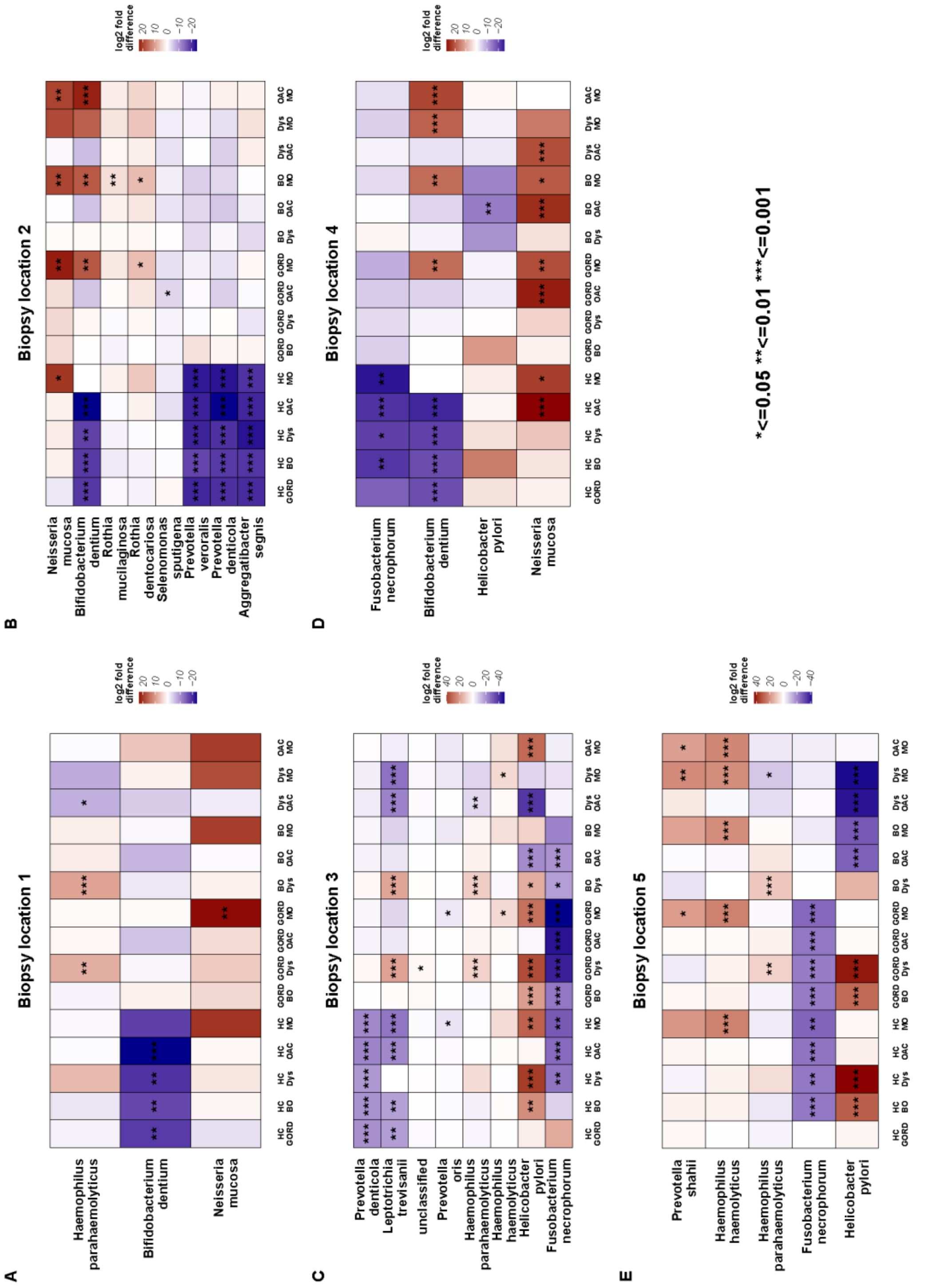
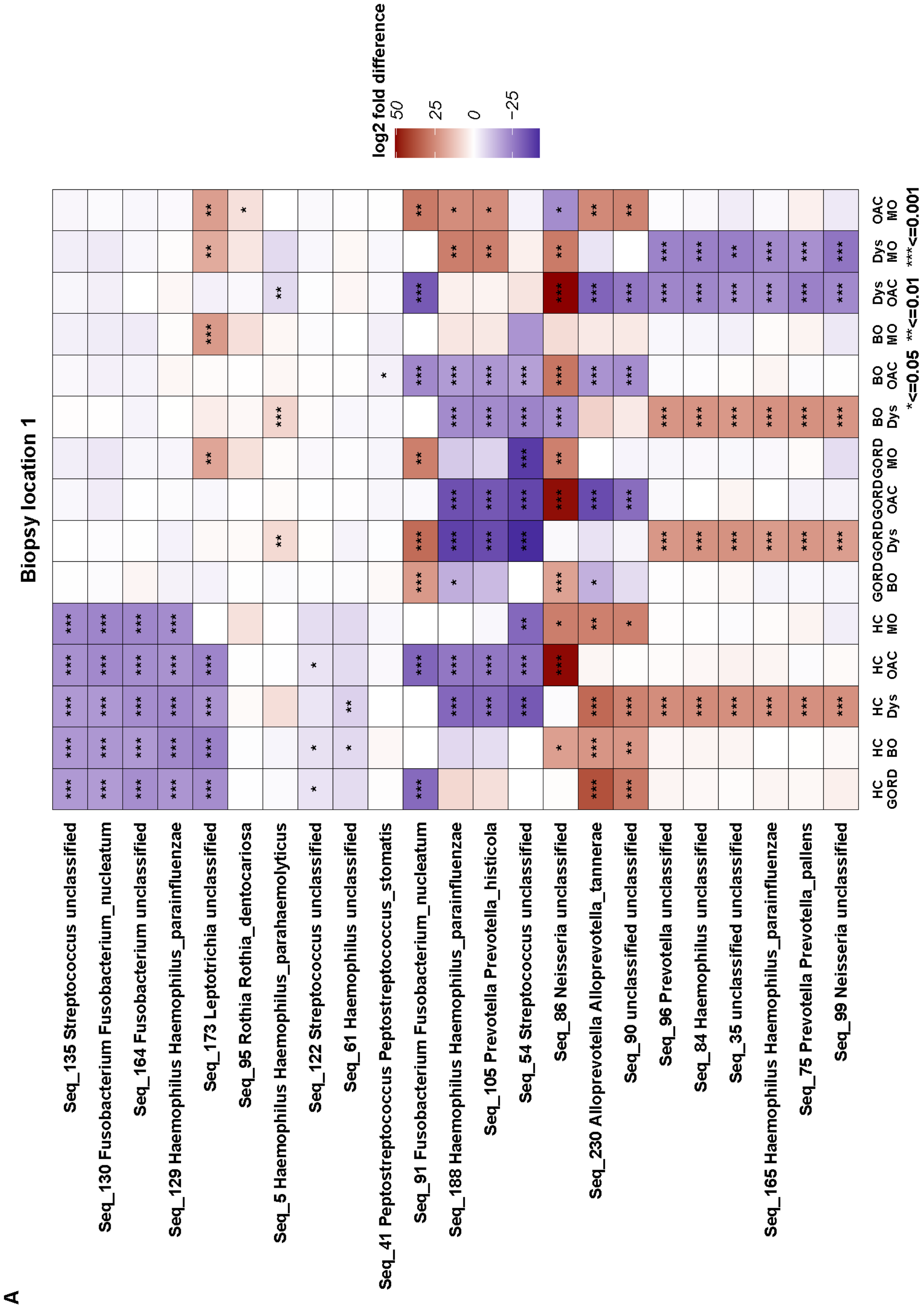
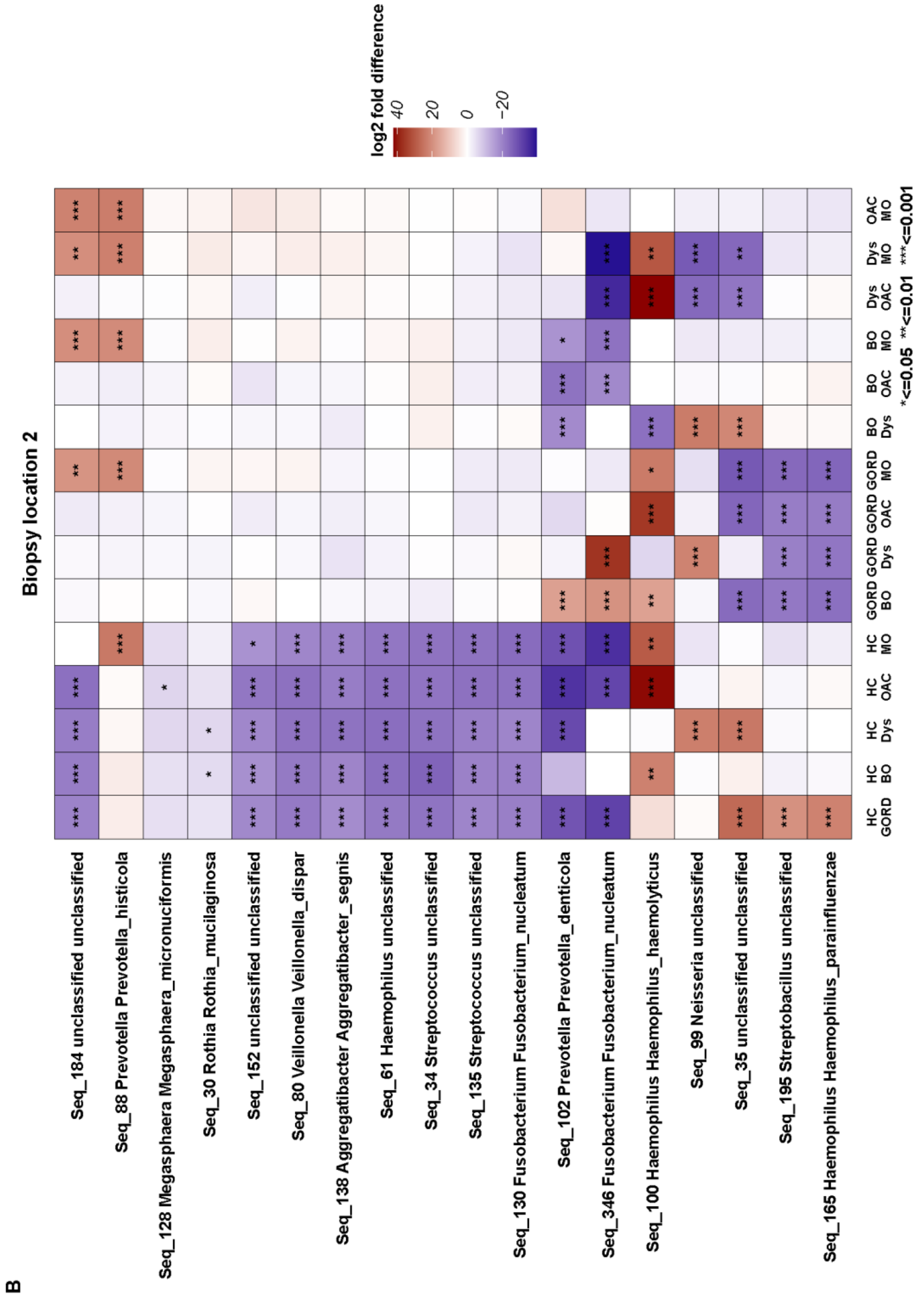
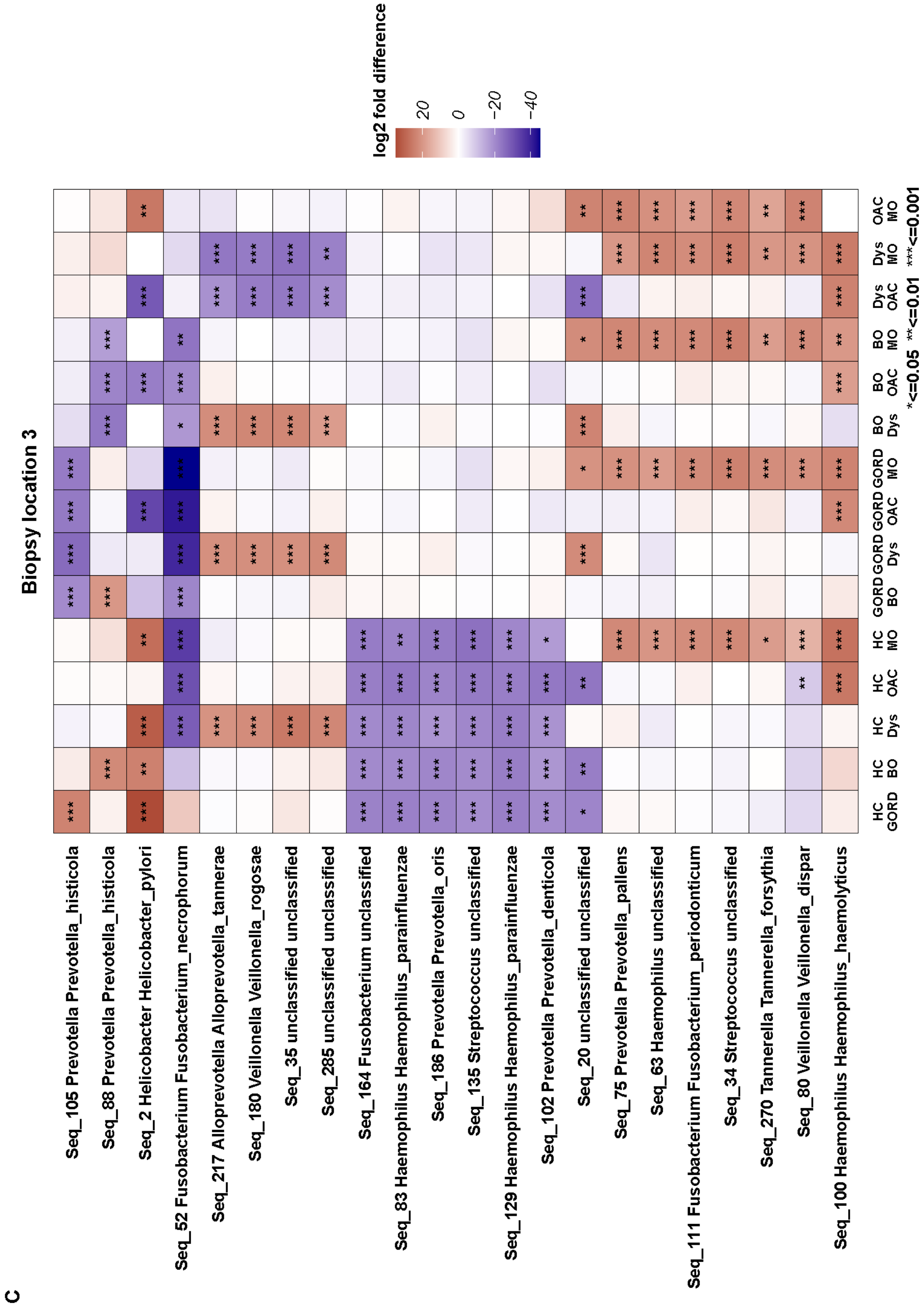
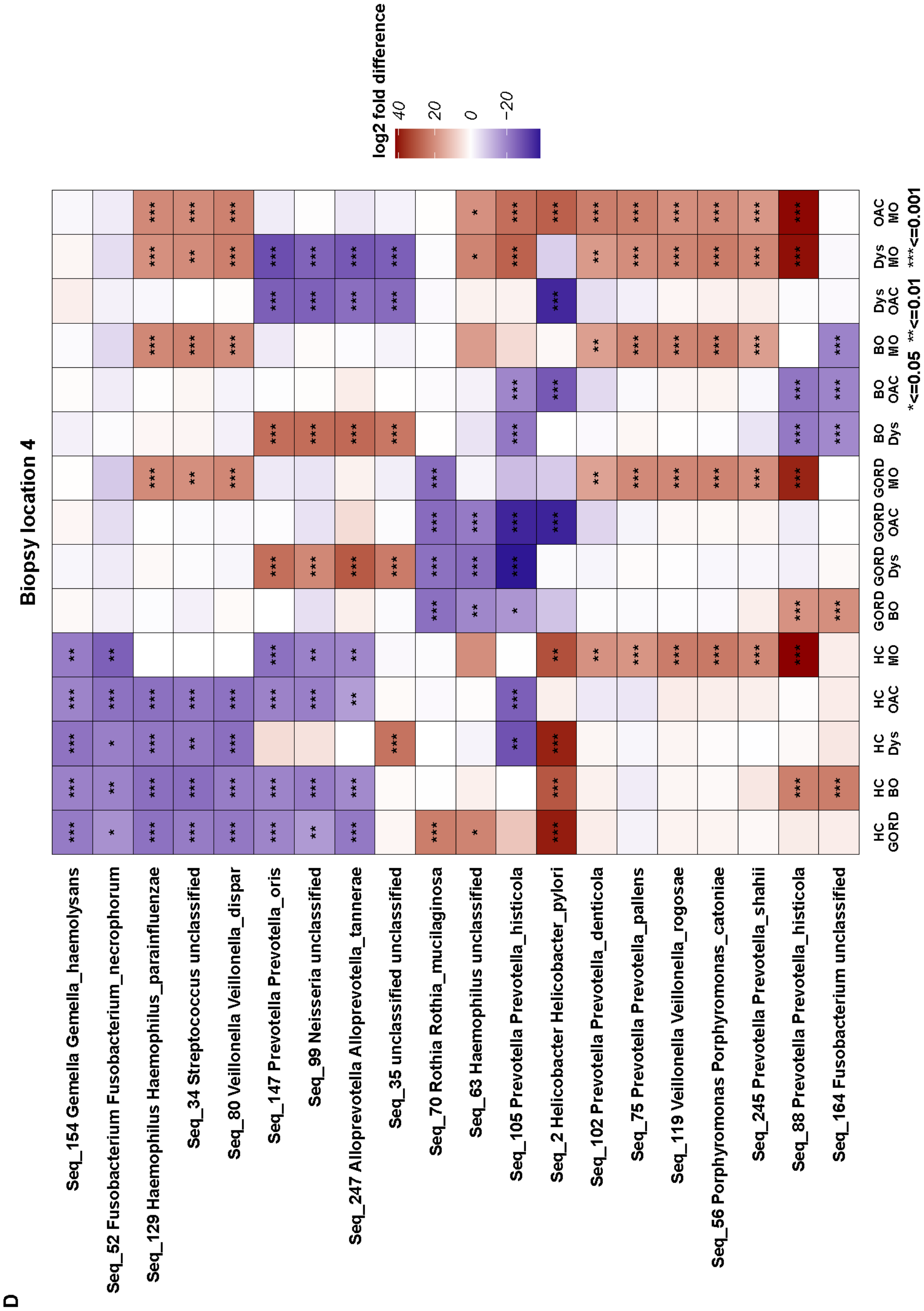


Figure 3. Differentially abundant species in the microbiome of subjects in the studied clinical classifications (A) Biopsy location 1, **(B)** Biopsy location 2. **(C)** Biopsy location 3 **(D)** Biopsy location 4 **(E)** Biopsy location 5. Statistical testing was performed using DESeq2 *<=0.05 **<=0.01 ***<=0.001. HC=Healthy controls, GORD= gastro-oesophageal reflux disease, BO= Barrett's oesophagus, Dys=Dysplasia , OAC= Oesophageal adenocarcinoma, MO= metastatic Oesophageal adenocarcinoma.

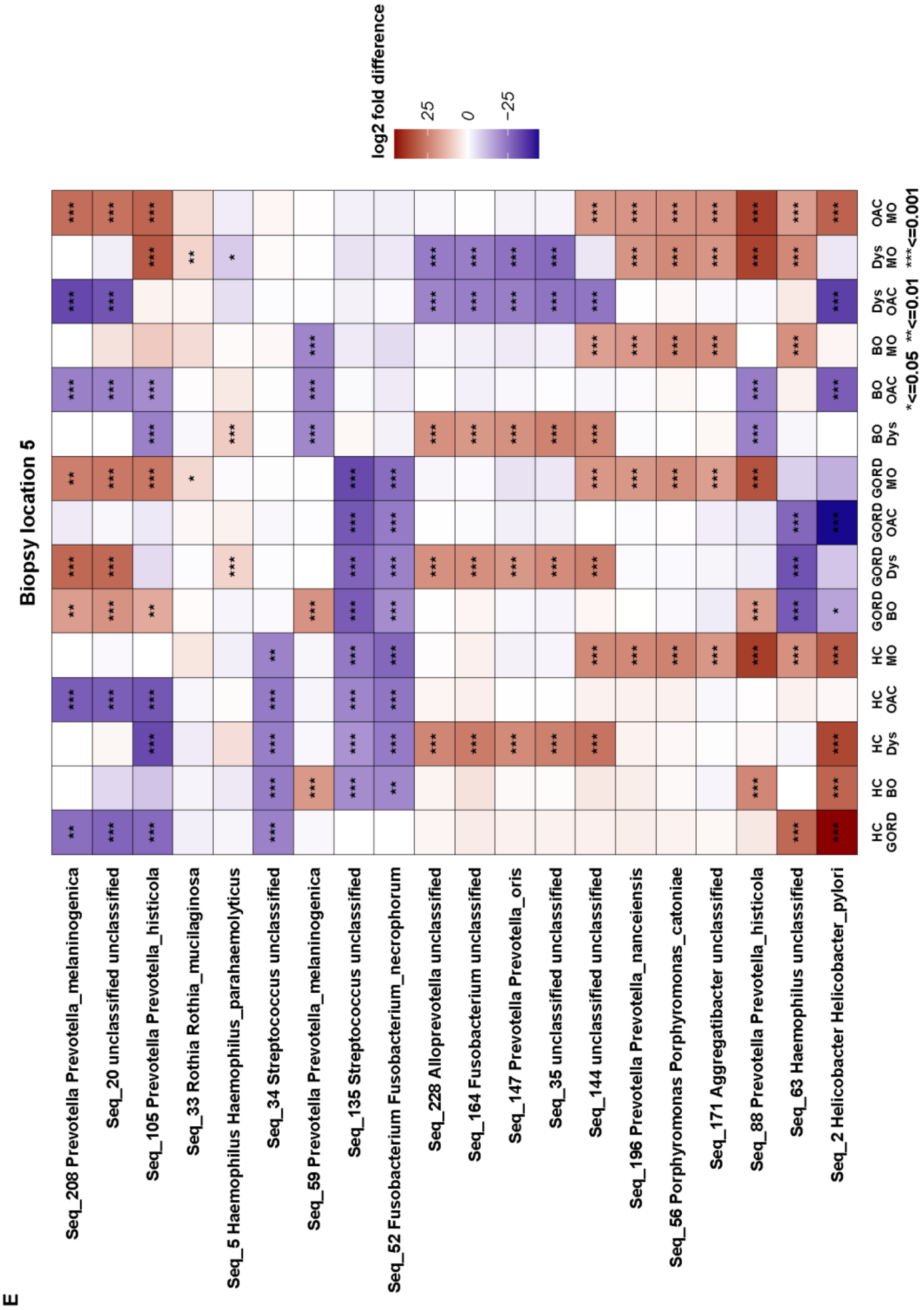

A
Biopsy location 1
log2 fold difference
50
25
0
−25
Seq_135 Streptococcus unclassified
Seq_130 Fusobacterium Fusobacterium_nucleatum
Seq_164 Fusobacterium unclassified
Seq_129 Haemophilus Haemophilus_parainfluenzae
Seq_173 Leptotrichia unclassified
Seq_95 Rothia Rothia_dentocariosa
Seq_5 Haemophilus Haemophilus_parahaemolyticus
Seq_122 Streptococcus unclassified
Seq_61 Haemophilus unclassified
Seq_41 Peptostreptococcus Peptostreptococcus_stomatis
Seq_91 Fusobacterium Fusobacterium_nucleatum
Seq_188 Haemophilus Haemophilus_parainfluenzae
Seq_105 Prevotella Prevotella_histicola
Seq_54 Streptococcus unclassified
Seq_86 Neisseria unclassified
Seq_230 Alloprevotella Alloprevotella_tannerae
Seq_90 unclassified unclassified
Seq_96 Prevotella unclassified
Seq_84 Haemophilus unclassified
Seq_35 unclassified unclassified
Seq_165 Haemophilus Haemophilus_parainfluenzae
Seq_75 Prevotella Prevotella_pallens
Seq_99 Neisseria unclassified
HC GORD
HC BO
HC Dys
HC OAC
HC MO
GORD BO
GORD Dys
GORD OAC
GORD MO
BO Dys
BO OAC
BO MO
Dys OAC
Dys MO
OAC MO
*<=0.05 **<=0.01 ***<=0.001


B
Biopsy location 2
log2 fold difference
40
20
0
−20
Seq_184 unclassified unclassified
Seq_88 Prevotella Prevotella_histicola
Seq_128 Megasphaera Megasphaera_micronuciformis
Seq_30 Rothia Rothia_mucilaginosa
Seq_152 unclassified unclassified
Seq_80 Veillonella Veillonella_dispar
Seq_138 Aggregatibacter Aggregatibacter_segnis
Seq_61 Haemophilus unclassified
Seq_34 Streptococcus unclassified
Seq_135 Streptococcus unclassified
Seq_130 Fusobacterium Fusobacterium_nucleatum
Seq_102 Prevotella Prevotella_denticola
Seq_346 Fusobacterium Fusobacterium_nucleatum
Seq_100 Haemophilus Haemophilus_haemolyticus
Seq_99 Neisseria unclassified
Seq_35 unclassified unclassified
Seq_195 Streptobacillus unclassified
Seq_165 Haemophilus Haemophilus_parainfluenzae
HC GORD
HC BO
HC Dys
HC OAC
HC MO
GORD BO
GORD Dys
GORD OAC
GORD MO
BO Dys
BO OAC
BO MO
Dys OAC
Dys MO
OAC MO
*<=0.05 **<=0.01 ***<=0.001

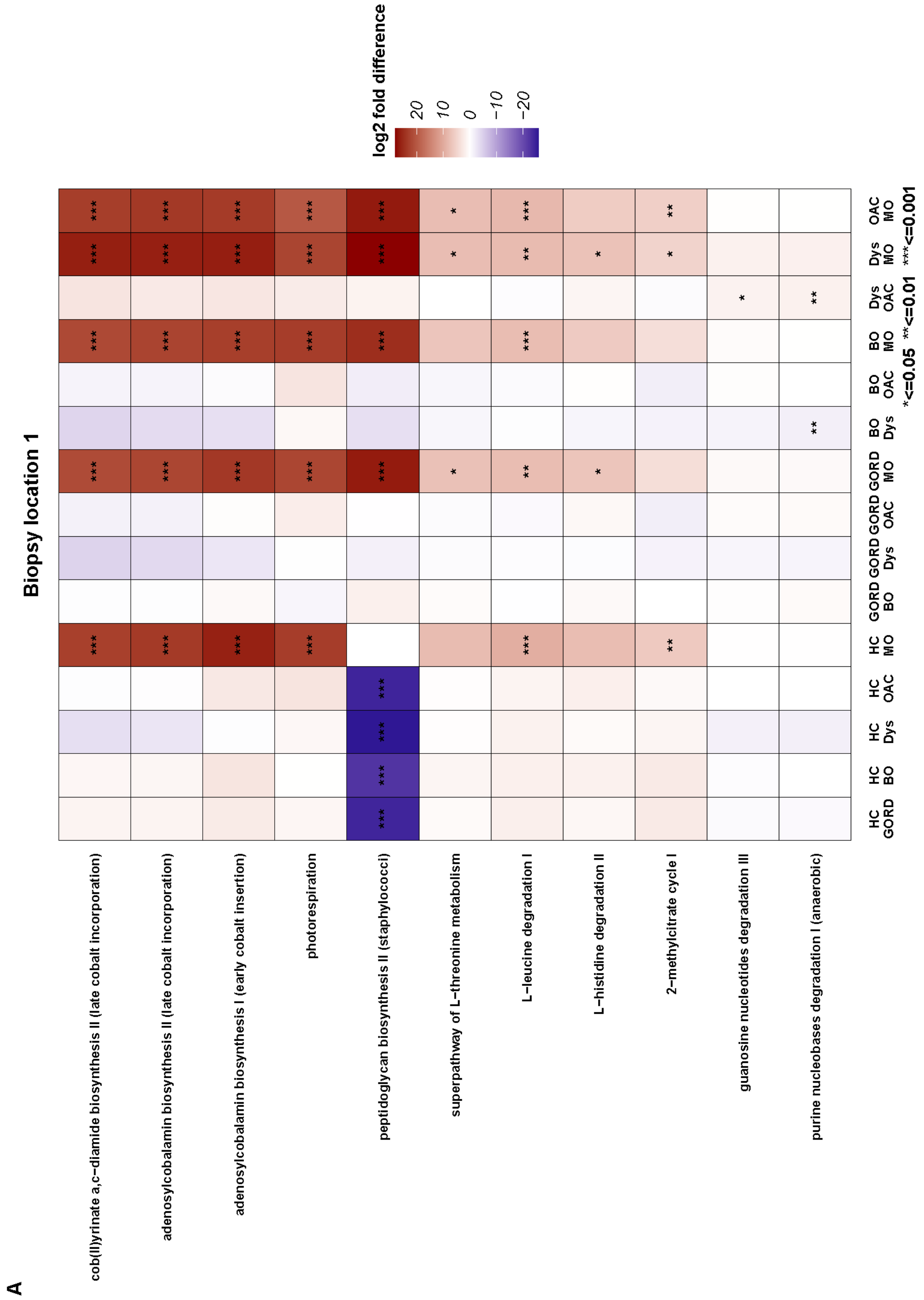
C
Biopsy location 3
log2 fold difference
20
0
−20
−40
Seq_105 Prevotella Prevotella_histicola
Seq_88 Prevotella Prevotella_histicola
Seq_2 Helicobacter Helicobacter_pylori
Seq_52 Fusobacterium Fusobacterium_necrophorum
Seq_217 Alloprevotella Alloprevotella_tannerae
Seq_180 Veillonella Veillonella_rogosae
Seq_35 unclassified unclassified
Seq_285 unclassified unclassified
Seq_164 Fusobacterium unclassified
Seq_83 Haemophilus Haemophilus_parainfluenzae
Seq_186 Prevotella Prevotella_oris
Seq_135 Streptococcus unclassified
Seq_129 Haemophilus Haemophilus_parainfluenzae
Seq_102 Prevotella Prevotella_denticola
Seq_20 unclassified unclassified
Seq_75 Prevotella Prevotella_pallens
Seq_63 Haemophilus unclassified
Seq_111 Fusobacterium Fusobacterium_periodonticum
Seq_34 Streptococcus unclassified
Seq_270 Tannerella Tannerella_forsythia
Seq_80 Veillonella Veillonella_dispar
Seq_100 Haemophilus Haemophilus_haemolyticus
HC GORD
HC BO
HC Dys
HC OAC
HC MO
GORD BO
GORD Dys
GORD OAC
GORD MO
BO Dys
BO OAC
BO MO
Dys OAC
Dys MO
OAC MO
*<=0.05 **<=0.01 ***<=0.001

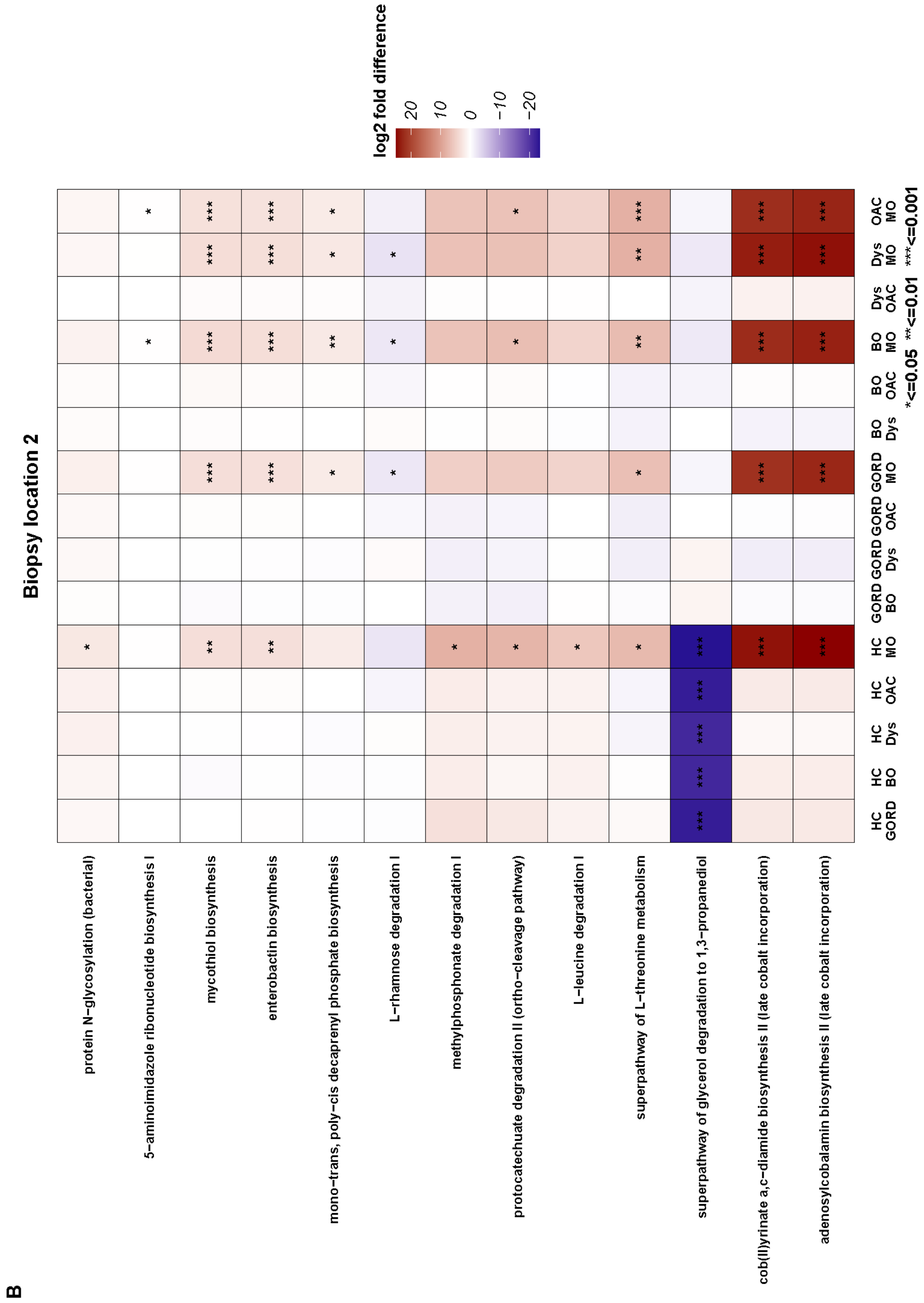
D
Biopsy location 4
log2 fold difference
40
20
0
−20
HC GORD
HC BO
HC Dys
HC OAC
HC MO
GORD BO
GORD Dys
GORD OAC
GORD MO
BO Dys
BO OAC
BO MO
Dys OAC
Dys MO
OAC MO
*<=0.05 **<=0.01 ***<=0.001
Seq_154 Gemella Gemella_haemolysans
Seq_52 Fusobacterium Fusobacterium_necrophorum
Seq_129 Haemophilus Haemophilus_parainfluenzae
Seq_34 Streptococcus unclassified
Seq_80 Veillonella Veillonella_dispar
Seq_147 Prevotella Prevotella_oris
Seq_99 Neisseria unclassified
Seq_247 Alloprevotella Alloprevotella_tannerae
Seq_35 unclassified unclassified
Seq_70 Rothia Rothia_mucilaginosa
Seq_63 Haemophilus unclassified
Seq_105 Prevotella Prevotella_histicola
Seq_2 Helicobacter Helicobacter_pylori
Seq_102 Prevotella Prevotella_denticola
Seq_75 Prevotella Prevotella_pallens
Seq_119 Veillonella Veillonella_rogosae
Seq_56 Porphyromonas Porphyromonas_catoniae
Seq_245 Prevotella Prevotella_shahii
Seq_88 Prevotella Prevotella_histicola
Seq_164 Fusobacterium unclassified

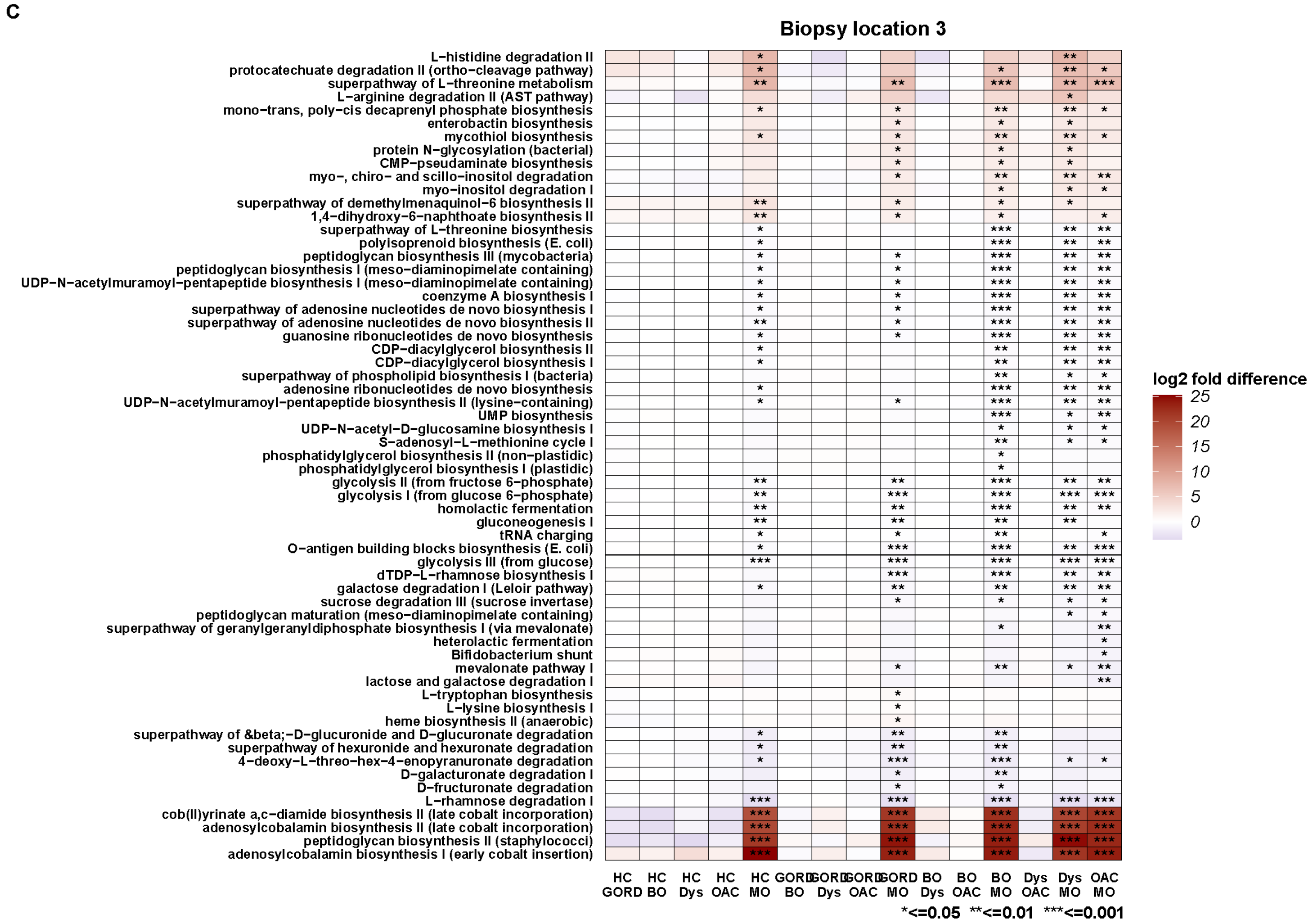
E

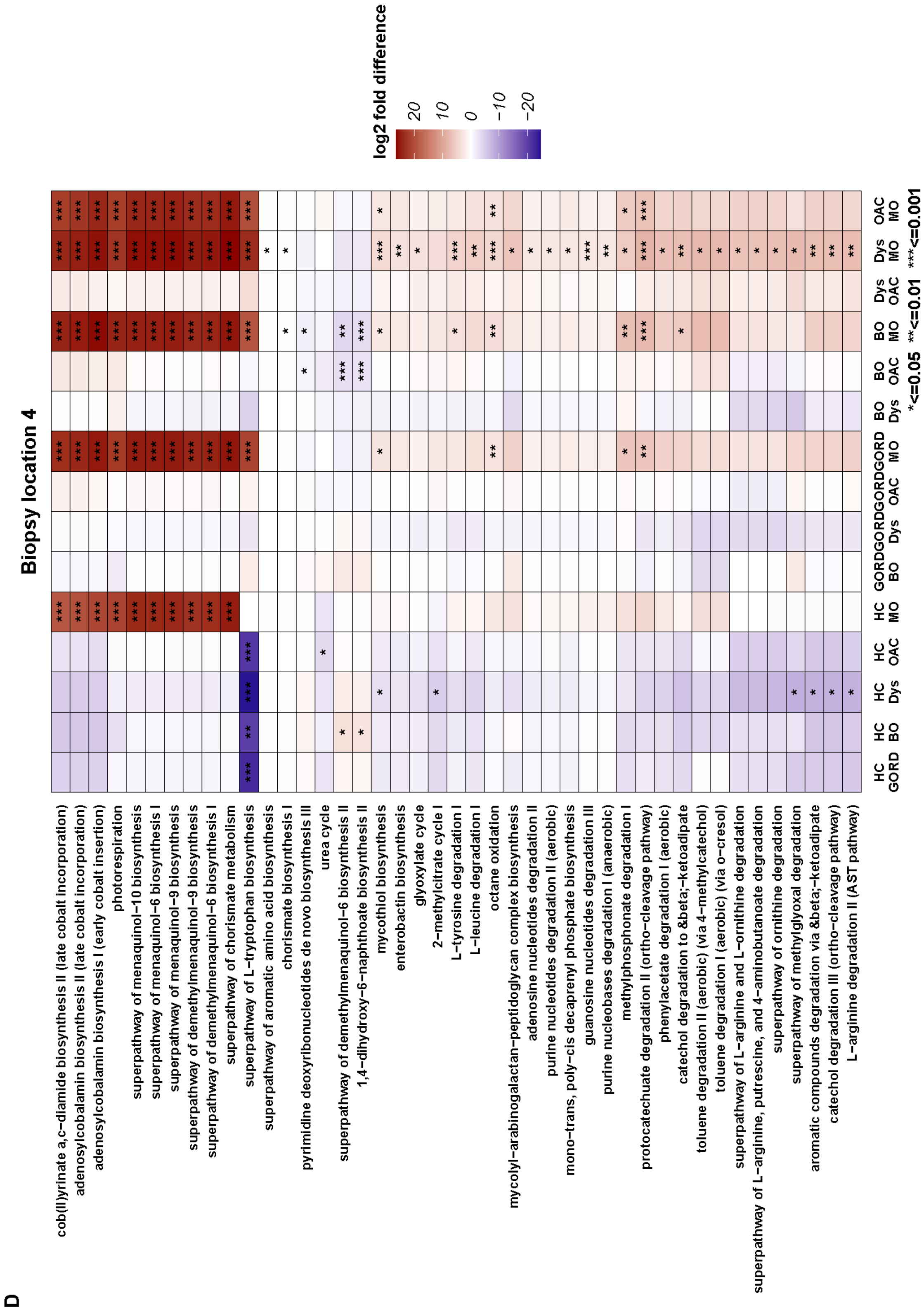


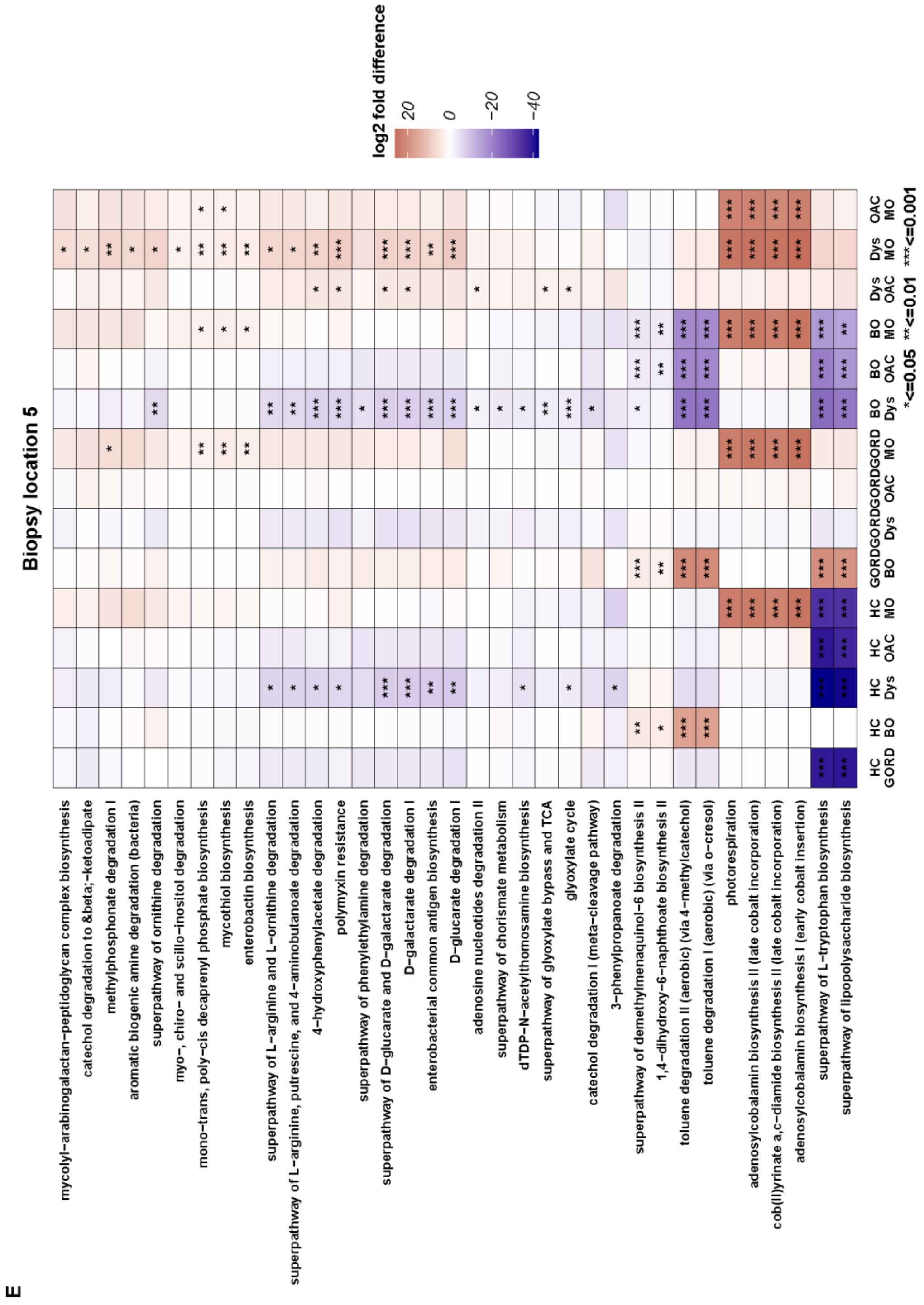
Supplementary figure 2. Differentially abundant ASVs between clinical classifications (A) Biopsy location 1, **(B)** Biopsy location 2. **(C)** Biopsy location 3 **(D)** Biopsy location 4 **(E)** Biopsy location 5. Statistical testing was performed using DESeq2 *<=0.05 **<=0.01 ***<=0.001. HC=Healthy controls, GORD= gastro-oesophageal reflux disease, BO= Barrett's oesophagus, Dys=Dysplasia , OAC= Oesophageal adenocarcinoma, MO= metastatic Oesophageal adenocarcinoma.


A
Biopsy location 1
log2 fold difference
20
10
0
−10
−20
cob(II)yrinate a,c−diamide biosynthesis II (late cobalt incorporation)
adenosylcobalamin biosynthesis II (late cobalt incorporation)
adenosylcobalamin biosynthesis I (early cobalt insertion)
photorespiration
peptidoglycan biosynthesis II (staphylococci)
superpathway of L−threonine metabolism
L−leucine degradation I
L−histidine degradation II
2−methylcitrate cycle I
guanosine nucleotides degradation III
purine nucleobases degradation I (anaerobic)
HC GORD
HC BO
HC Dys
HC OAC
HC MO
GORD BO
GORD Dys
GORD OAC
GORD MO
BO Dys
BO OAC
BO MO
Dys OAC
Dys MO
OAC MO
*<=0.05 **<=0.01 ***<=0.001


B
Biopsy location 2
log2 fold difference
20
10
0
−10
−20
protein N−glycosylation (bacterial)
5−aminoimidazole ribonucleotide biosynthesis I
mycothiol biosynthesis
enterobactin biosynthesis
mono−trans, poly−cis decaprenyl phosphate biosynthesis
L−rhamnose degradation I
methylphosphonate degradation I
protocatechuate degradation II (ortho−cleavage pathway)
L−leucine degradation I
superpathway of L−threonine metabolism
superpathway of glycerol degradation to 1,3−propanediol
cob(II)yrinate a,c−diamide biosynthesis II (late cobalt incorporation)
adenosylcobalamin biosynthesis II (late cobalt incorporation)
HC GORD
HC BO
HC Dys
HC OAC
HC MO
GORD BO
GORD Dys
GORD OAC
GORD MO
BO Dys
BO OAC
BO MO
Dys OAC
Dys MO
OAC MO
*<=0.05 **<=0.01 ***<=0.001


C
Biopsy location 3
L−histidine degradation II
protocatechuate degradation II (ortho−cleavage pathway)
superpathway of L−threonine metabolism
L−arginine degradation II (AST pathway)
mono−trans, poly−cis decaprenyl phosphate biosynthesis
enterobactin biosynthesis
mycothiol biosynthesis
protein N−glycosylation (bacterial)
CMP−pseudaminate biosynthesis
myo−, chiro− and scillo−inositol degradation
myo−inositol degradation I
superpathway of demethylmenaquinol−6 biosynthesis II
1,4−dihydroxy−6−naphthoate biosynthesis II
superpathway of L−threonine biosynthesis
polyisoprenoid biosynthesis (E. coli)
peptidoglycan biosynthesis III (mycobacteria)
peptidoglycan biosynthesis I (meso−diaminopimelate containing)
UDP−N−acetylmuramoyl−pentapeptide biosynthesis I (meso−diaminopimelate containing)
coenzyme A biosynthesis I
superpathway of adenosine nucleotides de novo biosynthesis I
superpathway of adenosine nucleotides de novo biosynthesis II
guanosine ribonucleotides de novo biosynthesis
CDP−diacylglycerol biosynthesis II
CDP−diacylglycerol biosynthesis I
superpathway of phospholipid biosynthesis I (bacteria)
adenosine ribonucleotides de novo biosynthesis
UDP−N−acetylmuramoyl−pentapeptide biosynthesis II (lysine−containing)
UMP biosynthesis
UDP−N−acetyl−D−glucosamine biosynthesis I
S−adenosyl−L−methionine cycle I
phosphatidylglycerol biosynthesis II (non−plastidic)
phosphatidylglycerol biosynthesis I (plastidic)
glycolysis II (from fructose 6−phosphate)
glycolysis I (from glucose 6−phosphate)
homolactic fermentation
gluconeogenesis I
tRNA charging
O−antigen building blocks biosynthesis (E. coli)
glycolysis III (from glucose)
dTDP−L−rhamnose biosynthesis I
galactose degradation I (Leloir pathway)
sucrose degradation III (sucrose invertase)
peptidoglycan maturation (meso−diaminopimelate containing)
superpathway of geranylgeranyldiphosphate biosynthesis I (via mevalonate)
heterolactic fermentation
Bifidobacterium shunt
mevalonate pathway I
lactose and galactose degradation I
L−tryptophan biosynthesis
L−lysine biosynthesis I
heme biosynthesis II (anaerobic)
superpathway of &beta;−D−glucuronide and D−glucuronate degradation
superpathway of hexuronide and hexuronate degradation
4−deoxy−L−threo−hex−4−enopyranuronate degradation
D−galacturonate degradation I
D−fructuronate degradation
L−rhamnose degradation I
cob(II)yrinate a,c−diamide biosynthesis II (late cobalt incorporation)
adenosylcobalamin biosynthesis II (late cobalt incorporation)
peptidoglycan biosynthesis II (staphylococci)
adenosylcobalamin biosynthesis I (early cobalt insertion)
HC GORD
HC BO
HC Dys
HC OAC
HC MO
GORD BO
GORD Dys
GORD OAC
GORD MO
BO Dys
BO OAC
BO MO
Dys OAC
Dys MO
OAC MO
*<=0.05 **<=0.01 ***<=0.001
log2 fold difference
25
20
15
10
5
0


D
Biopsy location 4
log2 fold difference
20
10
0
−10
−20
HC GORD
HC BO
HC Dys
HC OAC
HC MO
GORD BO
GORD Dys
GORD OAC
GORD MO
BO Dys
BO OAC
BO MO
Dys OAC
Dys MO
OAC MO
*<=0.05 **<=0.01 ***<=0.001
cob(II)yrinate a,c−diamide biosynthesis II (late cobalt incorporation)
adenosylcobalamin biosynthesis II (late cobalt incorporation)
adenosylcobalamin biosynthesis I (early cobalt insertion)
photorespiration
superpathway of menaquinol−10 biosynthesis
superpathway of menaquinol−6 biosynthesis I
superpathway of menaquinol−9 biosynthesis
superpathway of demethylmenaquinol−9 biosynthesis
superpathway of demethylmenaquinol−6 biosynthesis I
superpathway of chorismate metabolism
superpathway of L−tryptophan biosynthesis
superpathway of aromatic amino acid biosynthesis
chorismate biosynthesis I
pyrimidine deoxyribonucleotides de novo biosynthesis III
urea cycle
superpathway of demethylmenaquinol−6 biosynthesis II
1,4−dihydroxy−6−naphthoate biosynthesis II
mycothiol biosynthesis
enterobactin biosynthesis
glyoxylate cycle
2−methylcitrate cycle I
L−tyrosine degradation I
L−leucine degradation I
octane oxidation
mycolyl−arabinogalactan−peptidoglycan complex biosynthesis
adenosine nucleotides degradation II
purine nucleotides degradation II (aerobic)
mono−trans, poly−cis decaprenyl phosphate biosynthesis
guanosine nucleotides degradation III
purine nucleobases degradation I (anaerobic)
methylphosphonate degradation I
protocatechuate degradation II (ortho−cleavage pathway)
phenylacetate degradation I (aerobic)
catechol degradation to &beta;−ketoadipate
toluene degradation II (aerobic) (via 4−methylcatechol)
toluene degradation I (aerobic) (via o−cresol)
superpathway of L−arginine and L−ornithine degradation
superpathway of L−arginine, putrescine, and 4−aminobutanoate degradation
superpathway of ornithine degradation
superpathway of methylglyoxal degradation
aromatic compounds degradation via &beta;−ketoadipate
catechol degradation III (ortho−cleavage pathway)
L−arginine degradation II (AST pathway)



Supplementary figure 3. Differentially abundant microbiome metabolic pathways between clinical classifications (A) Biopsy location 1, **(B)** Biopsy location 2. **(C)** Biopsy location 3 **(D)** Biopsy location 4 **(E)** Biopsy location 5. Statistical testing was performed using DESeq2 *<=0.05 **<=0.01 ***<=0.001. HC=Healthy controls, GORD= gastro-oesophageal reflux disease, BO= Barrett's oesophagus, Dys=Dysplasia , OAC= Oesophageal adenocarcinoma, MO= metastatic Oesophageal adenocarcinoma.

2.4.4 Microbiome alterations with respect to biopsy location

Separating samples by each of the defined clinical classifications, we sought to identify differences in global ecological measures including alpha-diversity and beta-diversity between biopsy locations. Samples derived from individuals with BO showed the most significant difference in alpha-diversity with respect to biopsy site (Figure 3C). In particular, metaplastic tissue derived from the GOJ (biopsy location 3) had a higher alpha diversity than oesophageal (biopsy location 1 and 2) and gastric biopsies (biopsy location 4 and 5). Differences were observed in various alpha diversity indices between samples sites within the other clinical classifications, but a particular trend was not apparent (Figure 3). With respect to samples derived from individuals with dysplasia, gastric sample microbiomes (biopsy location 4 and 5) had higher alpha-diversity, as measured by Shannon diversity and Simpson's diversity, relative to oesophageal samples (biopsy location 1 and 3) (Figure 3D).

Aggregating samples across all stages of the of the oesophageal adenocarcinoma sequence, alpha diversity was statistically significantly higher in GOJ and gastric biopsies relative to oesophageal biopsies as measured by Simpson and Shannon diversity (paired Wilcoxon) (Supplementary figure 4)

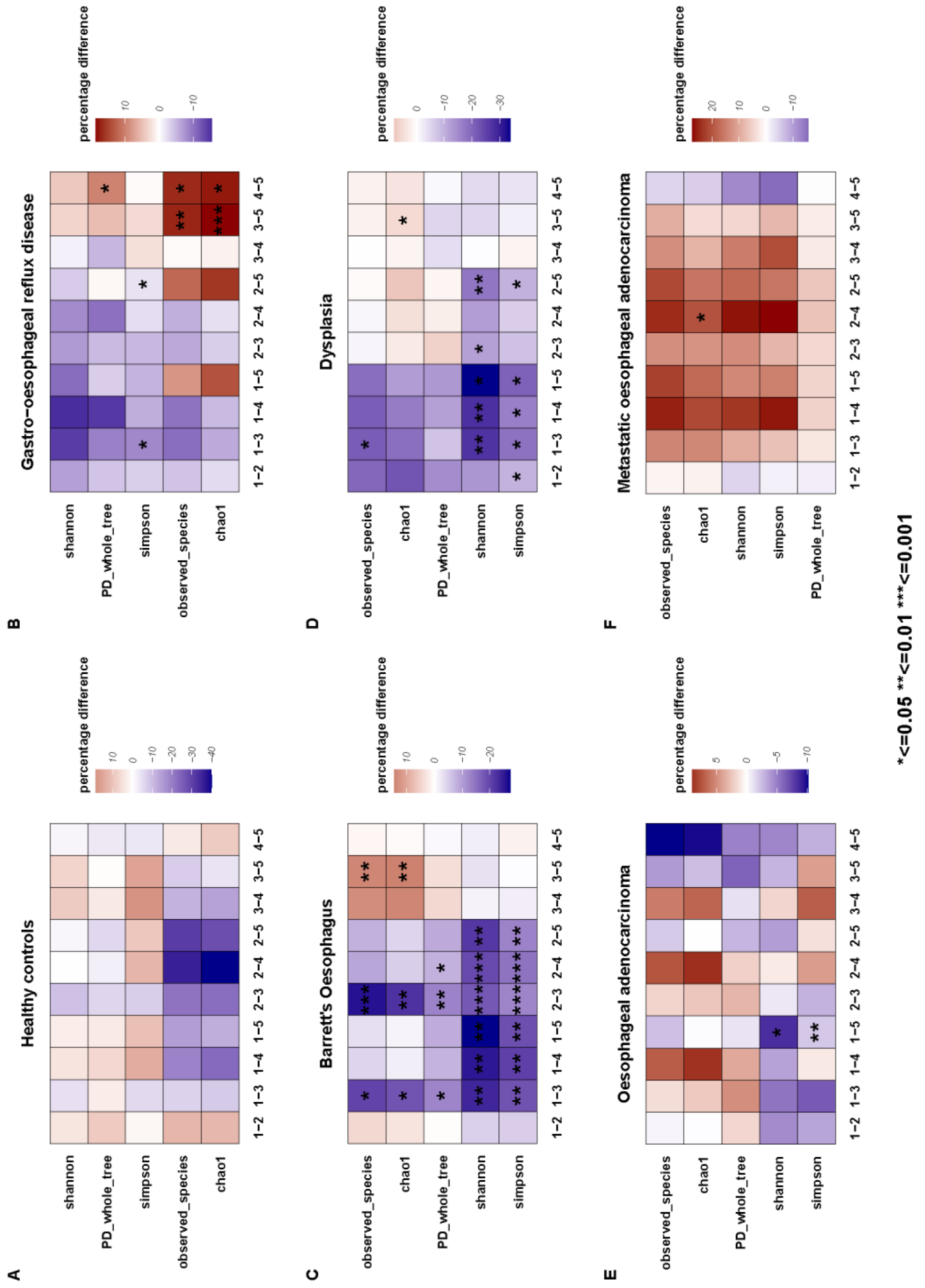
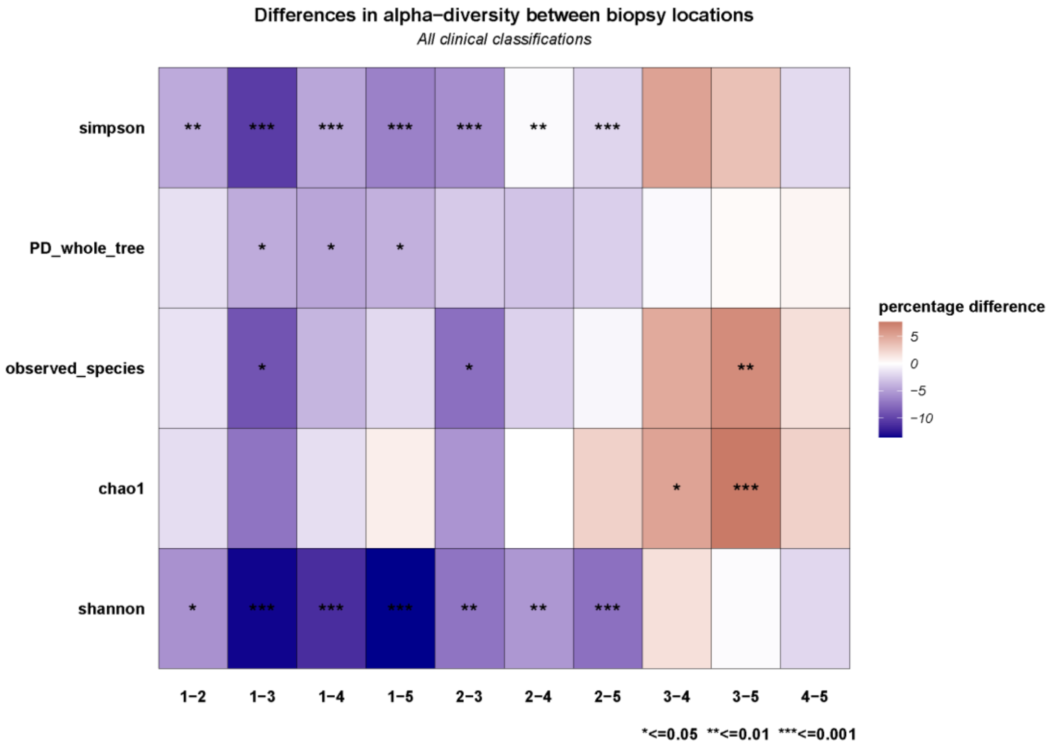


Figure 4. Differences in Alpha-diversity between biopsy location with respect to clinical classification. Heat-plot representing differences in alpha diversity indices between each pair of biopsy location. Statistical testing performed using paired Wilcoxon. **(A)** Data derived from Healthy controls. **(B)** Data derived from individuals with GORD. **(C)** Data derived from individuals with BO. **(D)** Data derived from individuals with Dysplasia. **(E)** Data derived from individuals with OAC. **(F)** Data derived from individuals with metastatic OAC.



Supplementary figure 4. Differences in Alpha-diversity with respect to biopsy location. Heat-plot representing differences in various alpha-diversity measurements. Data was derived from biopsies from all clinical classifications. Statistical testing performed using paired Wilcoxon.

A significant difference in beta-diversity was observed between biopsy location in the context of each group on the oesophageal adenocarcinoma sequence (Figure 4, Supplementary data 2). The microbiomes of biopsies which were anatomically closer together tended to cluster closer together.

	Beta-diversity metrics with respect to biopsy location							
Clinical classification	**Bray–Curtis**		**unweighted UniFrac**		**Jaccard Index**		**Robust Aitchison**	
	P-value	**R-squared**	**P-value**	**R-squared**	**P-value**	**R-squared**	**P-value**	**R-squared**
Healthy Controls	0.002	0.049	0.243	0.054	0.327	0.047	0.105	0.035
GORD	0.002	0.015	0.226	0.014	0.049	0.016	0.001	0.021
BO	0.001	0.025	0.008	0.02	0.069	0.014	0.003	0.015
Dysplasia	0.001	0.051	0.077	0.029	0.181	0.026	0.001	0.052
OAC	0.001	0.02	0.062	0.015	0.062	0.011	0.073	0.009
Metastatic OAC	0.116	0.049	0.665	0.03	0.283	0.039	0.453	0.016

Supplementary table 2. Analysis of difference in beta-diversity metrics with respect to biopsies location in the context of each clinical classification. P-value and R-squared calculated using Permutational multivariate analysis of variance (PERMANOVA).

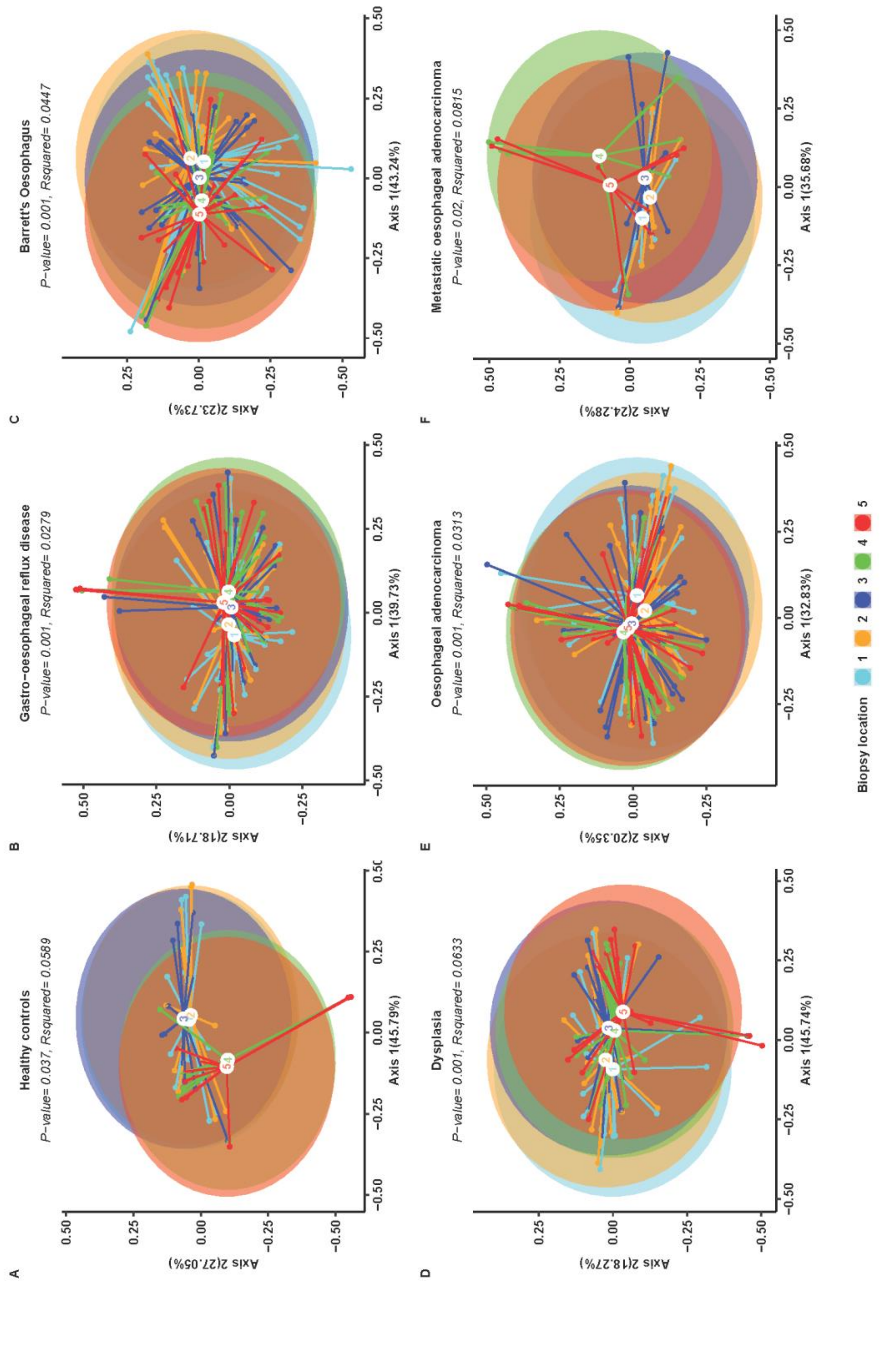


Figure 5. Difference in beta-diversity with respect to biopsy location. Principal Coordinates Analysis (PCoA) plot representing weighted UniFrac distance. Data was derived from biopsies from all clinical classifications. Statistical testing performed using Permutational Multivariate Analysis of Variance. (A) Data derived from Healthy controls. (B) Data derived from individuals with GORD. (C) Data derived from individuals with BO. (D) Data derived from individuals with Dysplasia. (E) Data derived from individuals with OAC. (F) Data derived from individuals with metastatic OAC.

2.4.5 Differentially abundant ASVs, species and metabolic pathways

Differential abundance analysis on paired samples was performed to identify differential species and ASVs between biopsy location within clinical classification groups.

Samples derived from individuals with BO had the highest number of differentially abundant species and ASVs (Figure 5C, Supplementary figure 5C). In people with BO, *Fusobacterium nucleatum*, a putative oncobacterium, was enriched on metaplastic tissue (biopsy location 3) relative to an adjacent oesophageal tissue (biopsy location 2). Similarly, in individuals with dysplasia, *F. nucleatum* was found to be enriched on dysplastic tissue relative to adjacent oesophageal tissue (Figure 5D). With respect to individuals with OAC, only one species, *Veillonella atypica*, differed in abundance between neoplastic tissue and at only one site, biopsy location 5 (Figure 5E). At the ASV level, an ASV, Seq 62, assigned to *F. nucleatum* was enriched in biopsy location 3 relative to biopsy location 4 (Supplementary figure 5E).

Generally, sample sites which were physically closer together (e.g., 1 versus 2 and 4 versus 5) had fewer differentially abundant taxa. Notably, at the species level, no differentially abundant taxa were observed between sites for the metastatic group.

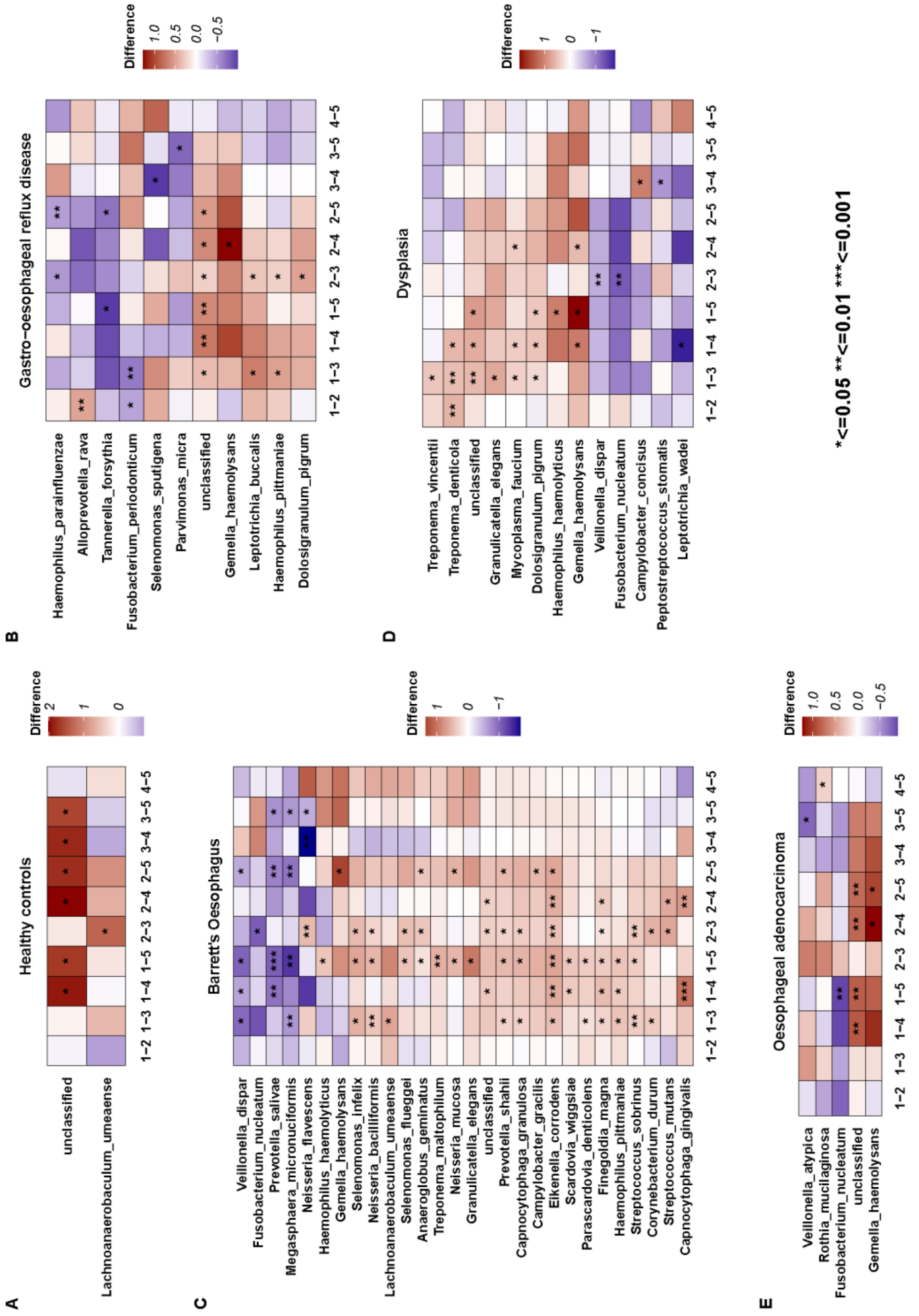
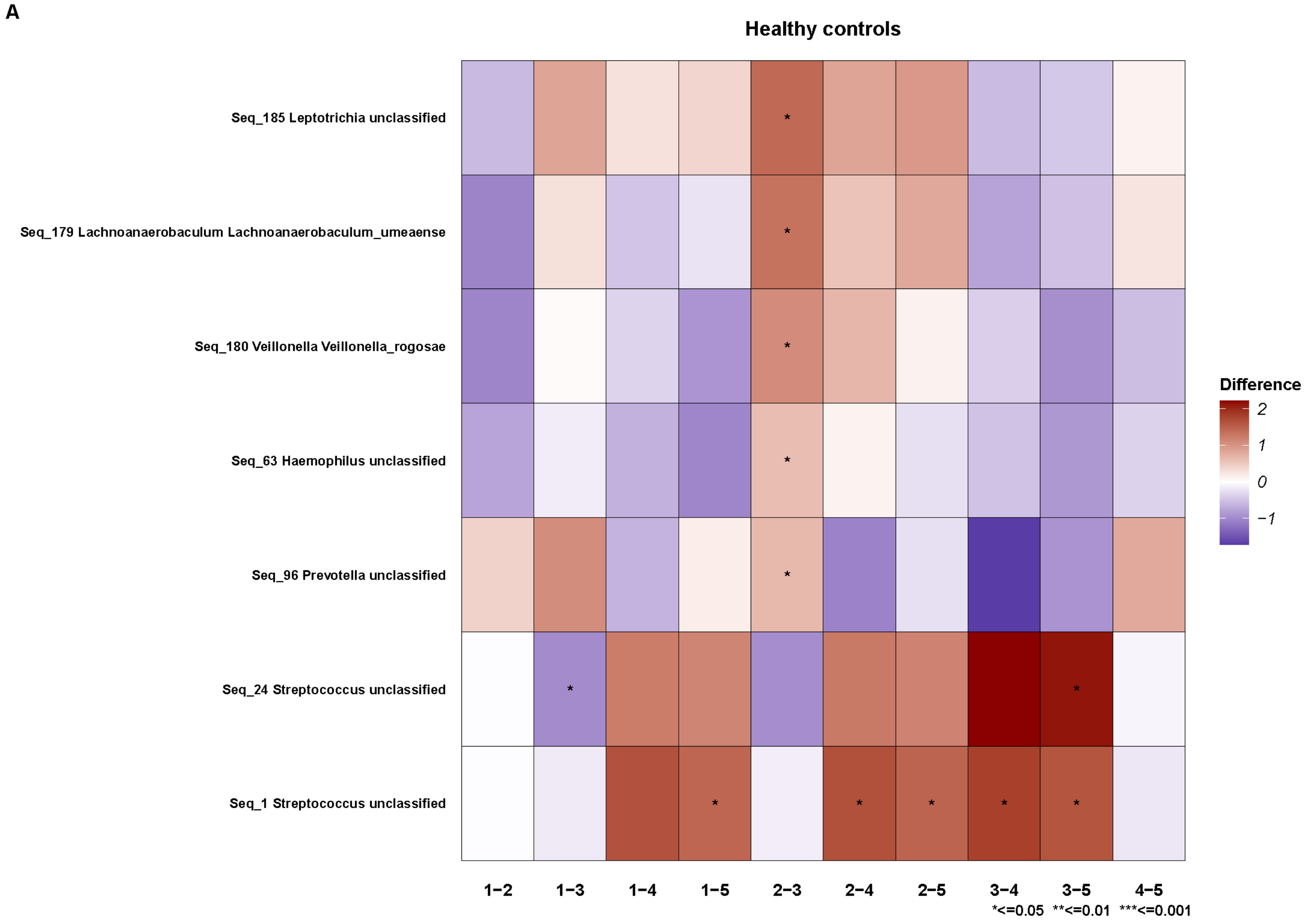
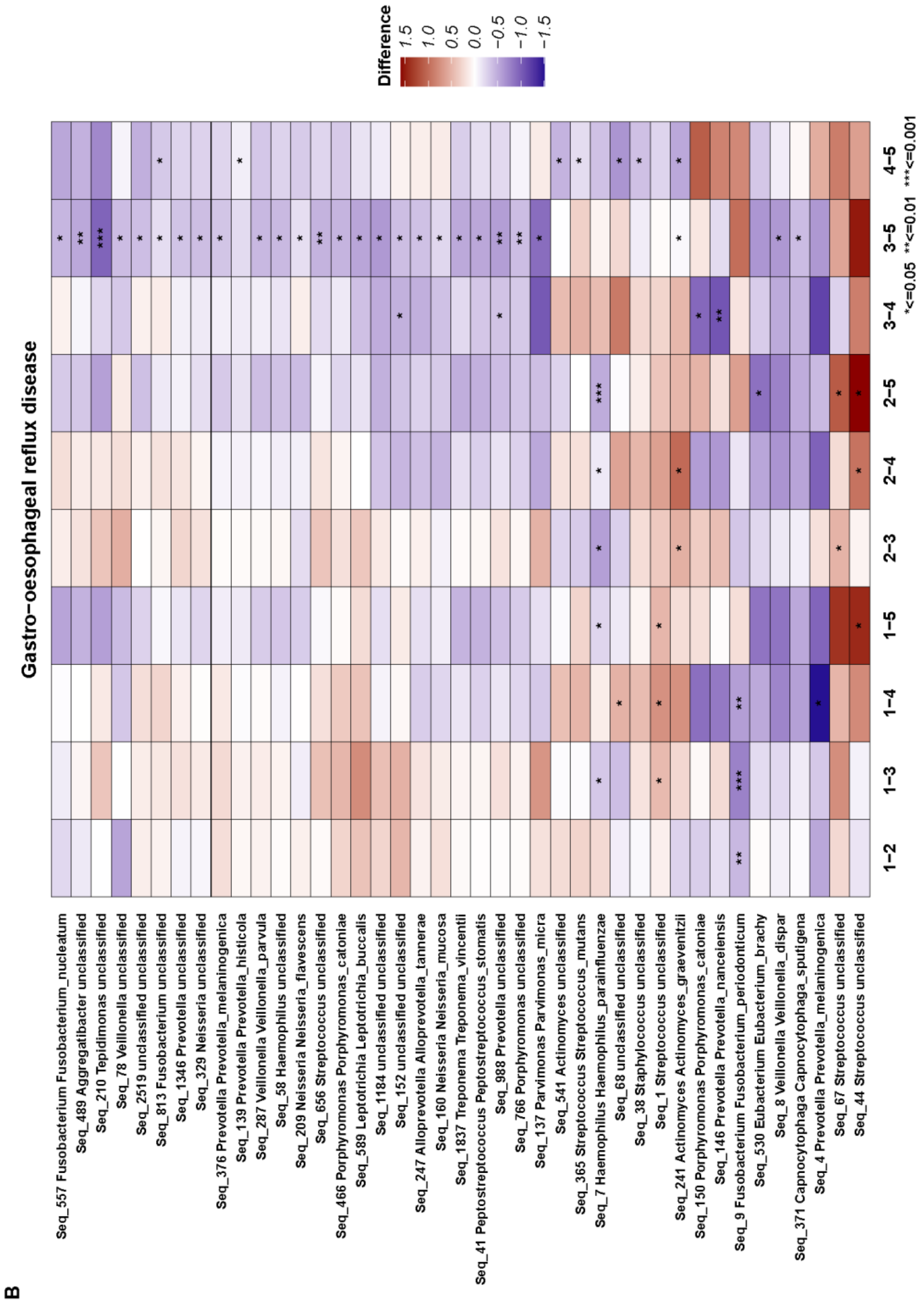
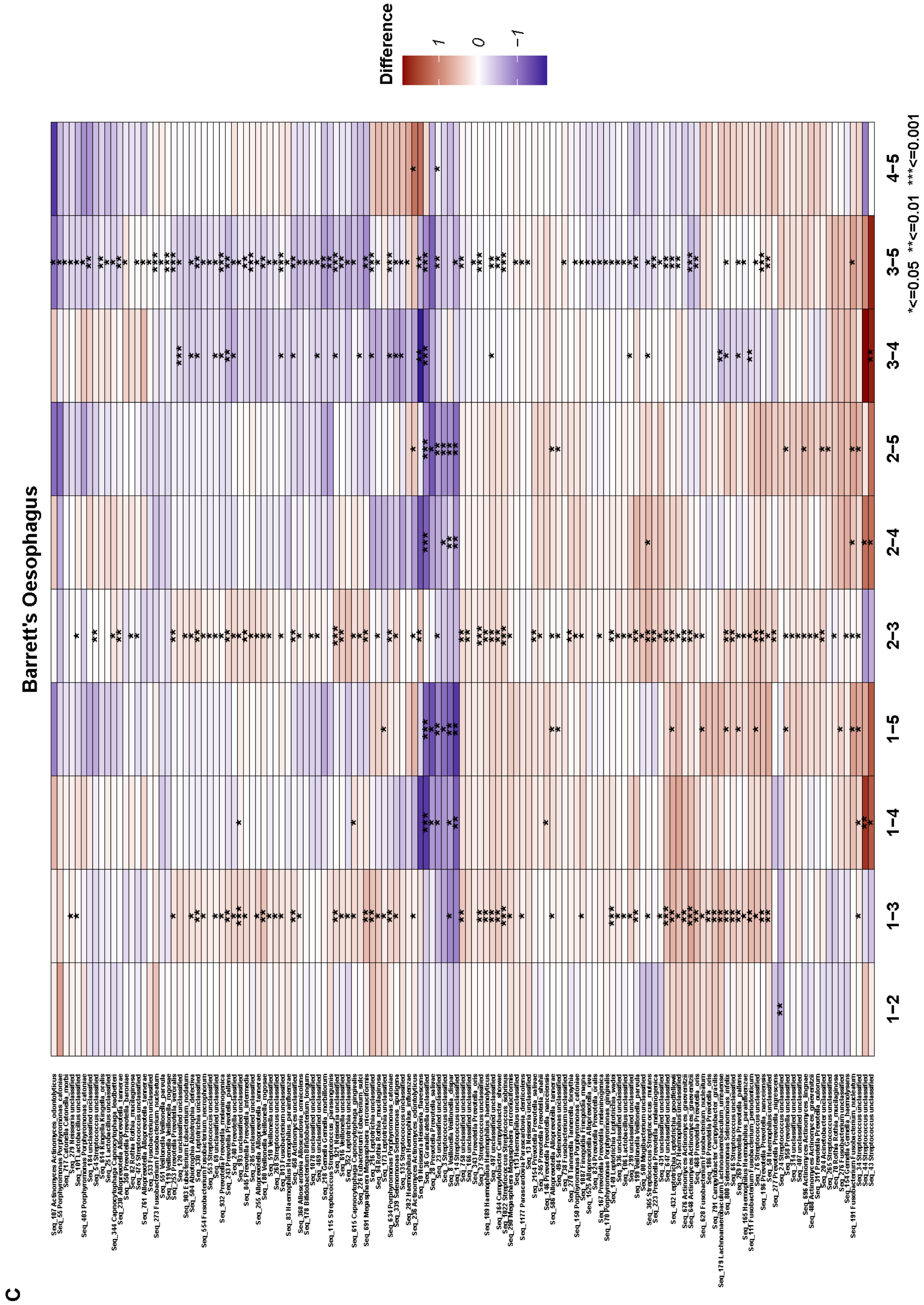
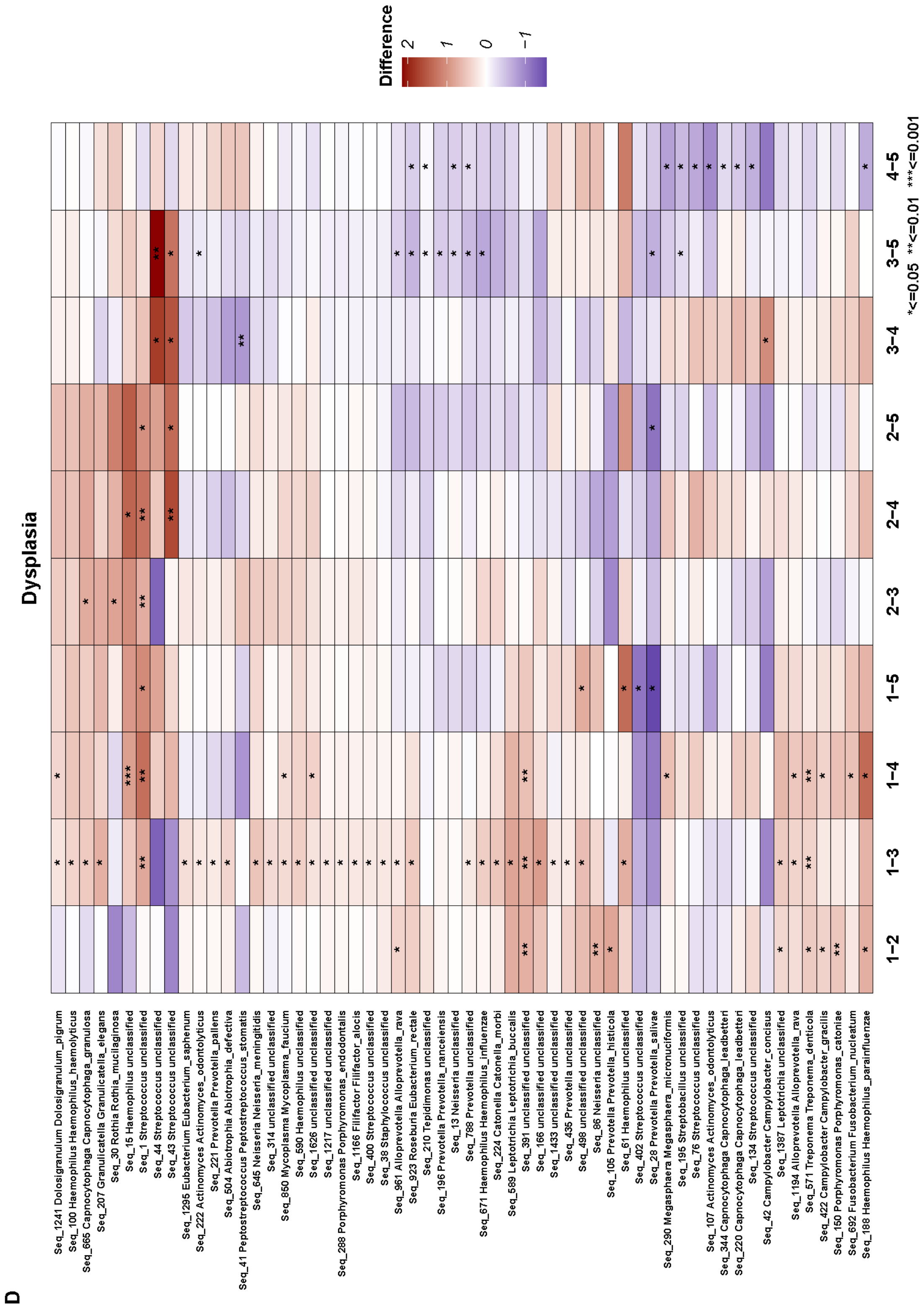


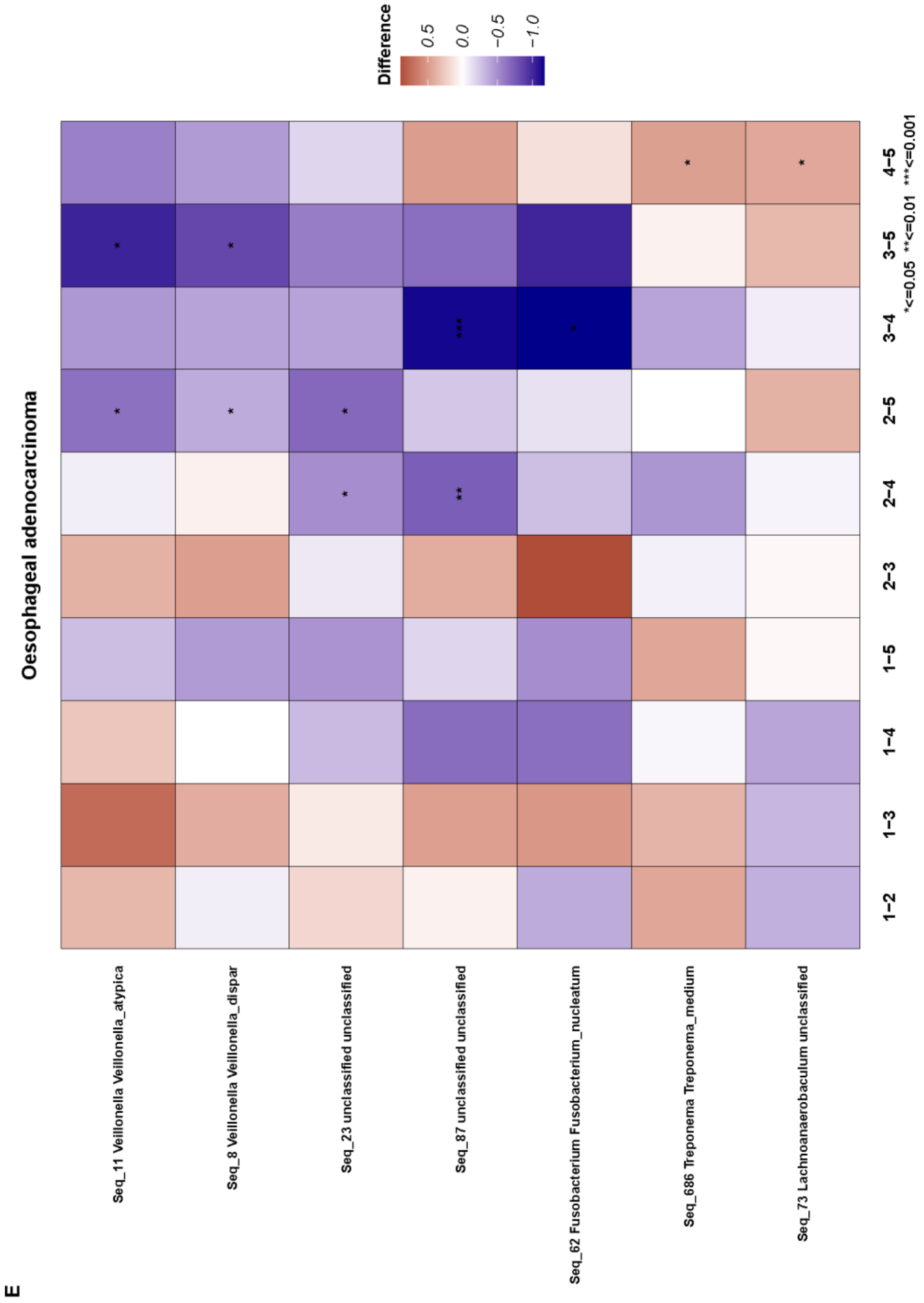
Figure 6. Differentially abundant species between biopsy location within clinical classification groups. Heat-map of differential species between each pair of biopsy location per clinical classification. **(A)** Data derived from Healthy controls. **(B)** Data derived from individuals with GORD. **(C)** Data derived from individuals with BO. **(D)** Data derived from individuals with Dysplasia. **(E)** Data derived from individuals with OAC. Statistical testing was using paired Wilcoxon.

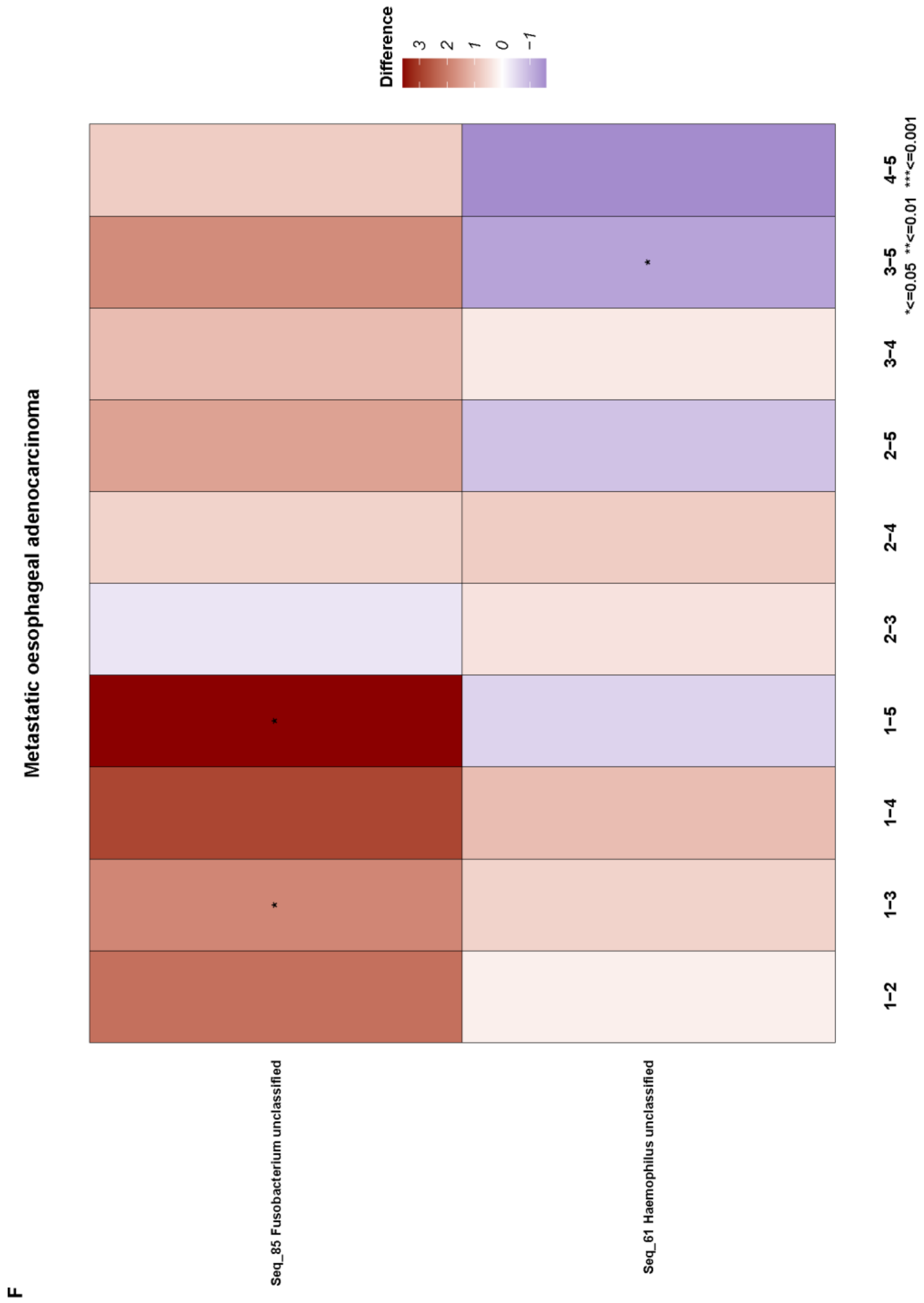

A
Healthy controls
Seq_185 Leptotrichia unclassified
Seq_179 Lachnoanaerobaculum Lachnoanaerobaculum_umeaense
Seq_180 Veillonella Veillonella_rogosae
Seq_63 Haemophilus unclassified
Seq_96 Prevotella unclassified
Seq_24 Streptococcus unclassified
Seq_1 Streptococcus unclassified
1−2
1−3
1−4
1−5
2−3
2−4
2−5
3−4
3−5
4−5
*<=0.05 **<=0.01 ***<=0.001
Difference
2
1
0
−1


B
Gastro−oesophageal reflux disease
Seq_557 Fusobacterium Fusobacterium_nucleatum
Seq_489 Aggregatibacter unclassified
Seq_210 Tepidimonas unclassified
Seq_78 Veillonella unclassified
Seq_2519 unclassified unclassified
Seq_813 Fusobacterium unclassified
Seq_1346 Prevotella unclassified
Seq_329 Neisseria unclassified
Seq_376 Prevotella Prevotella_melaninogenica
Seq_139 Prevotella Prevotella_histicola
Seq_287 Veillonella Veillonella_parvula
Seq_58 Haemophilus unclassified
Seq_209 Neisseria Neisseria_flavescens
Seq_656 Streptococcus unclassified
Seq_466 Porphyromonas Porphyromonas_catoniae
Seq_589 Leptotrichia Leptotrichia_buccalis
Seq_1184 unclassified unclassified
Seq_152 unclassified unclassified
Seq_247 Alloprevotella Alloprevotella_tannerae
Seq_160 Neisseria Neisseria_mucosa
Seq_1837 Treponema Treponema_vincentii
Seq_41 Peptostreptococcus Peptostreptococcus_stomatis
Seq_988 Prevotella unclassified
Seq_766 Porphyromonas unclassified
Seq_137 Parvimonas Parvimonas_micra
Seq_541 Actinomyces unclassified
Seq_365 Streptococcus Streptococcus_mutans
Seq_7 Haemophilus Haemophilus_parainfluenzae
Seq_68 unclassified unclassified
Seq_38 Staphylococcus unclassified
Seq_1 Streptococcus unclassified
Seq_241 Actinomyces Actinomyces_graevenitzii
Seq_150 Porphyromonas Porphyromonas_catoniae
Seq_146 Prevotella Prevotella_nanceiensis
Seq_9 Fusobacterium Fusobacterium_periodonticum
Seq_530 Eubacterium Eubacterium_brachy
Seq_8 Veillonella Veillonella_dispar
Seq_371 Capnocytophaga Capnocytophaga_sputigena
Seq_4 Prevotella Prevotella_melaninogenica
Seq_67 Streptococcus unclassified
Seq_44 Streptococcus unclassified
1−2
1−3
1−4
1−5
2−3
2−4
2−5
3−4
3−5
4−5
*<=0.05 **<=0.01 ***<=0.001
Difference
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5


C
Barrett's Oesophagus
Difference
1
0
−1
1−2
1−3
1−4
1−5
2−3
2−4
2−5
3−4
3−5
4−5
*<=0.05 **<=0.01 ***<=0.001

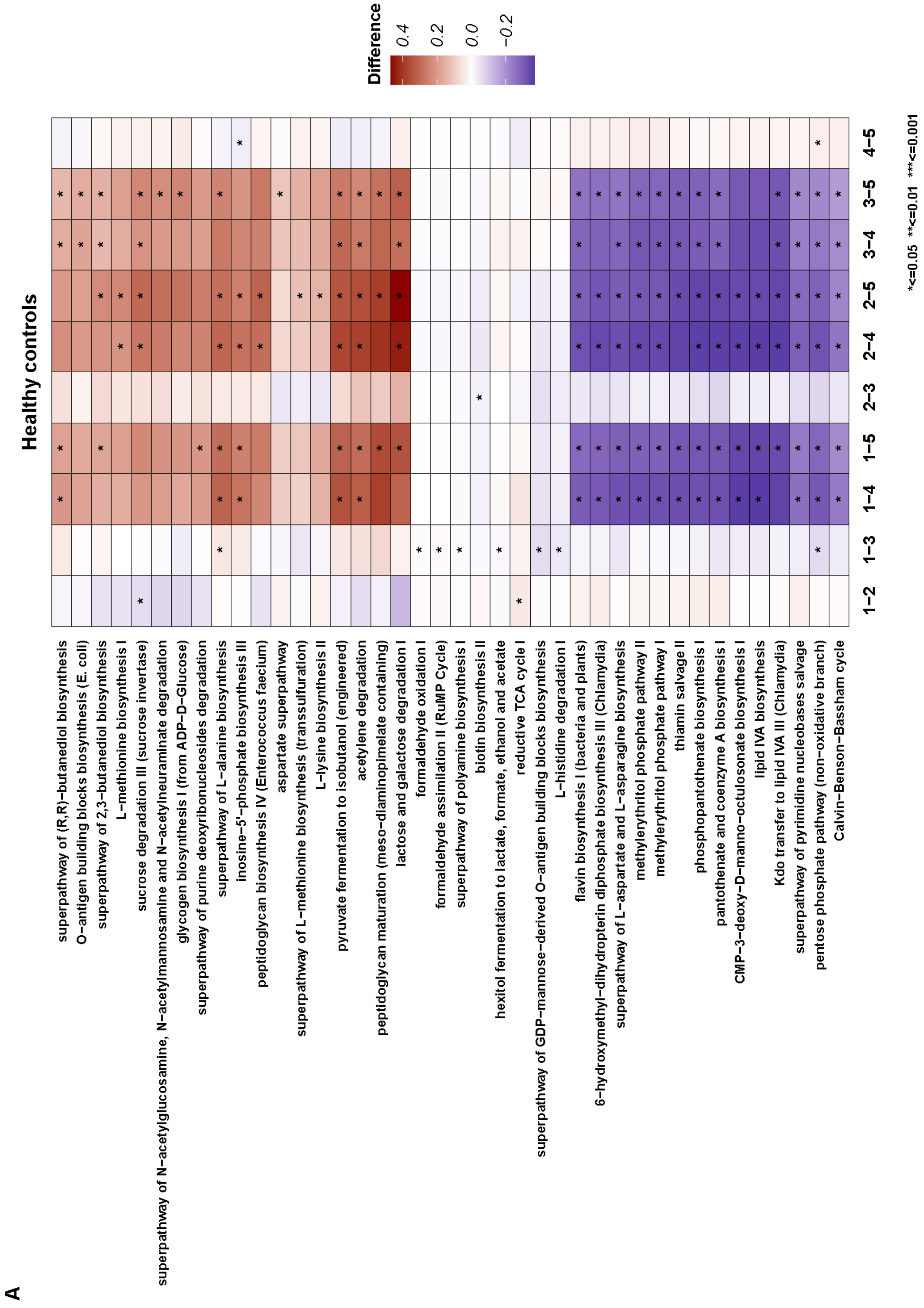

D
Dysplasia
Seq_1241 Dolosigranulum Dolosigranulum_pigrum
Seq_100 Haemophilus Haemophilus_haemolyticus
Seq_665 Capnocytophaga Capnocytophaga_granulosa
Seq_207 Granulicatella Granulicatella_elegans
Seq_30 Rothia Rothia_mucilaginosa
Seq_15 Haemophilus unclassified
Seq_1 Streptococcus unclassified
Seq_44 Streptococcus unclassified
Seq_43 Streptococcus unclassified
Seq_1295 Eubacterium Eubacterium_saphenum
Seq_222 Actinomyces Actinomyces_odontolyticus
Seq_221 Prevotella Prevotella_pallens
Seq_504 Abiotrophia Abiotrophia_defectiva
Seq_41 Peptostreptococcus Peptostreptococcus_stomatis
Seq_645 Neisseria Neisseria_meningitidis
Seq_314 unclassified unclassified
Seq_850 Mycoplasma Mycoplasma_faucium
Seq_590 Haemophilus unclassified
Seq_1626 unclassified unclassified
Seq_1217 unclassified unclassified
Seq_288 Porphyromonas Porphyromonas_endodontalis
Seq_1166 Filifactor Filifactor_alocis
Seq_400 Streptococcus unclassified
Seq_38 Staphylococcus unclassified
Seq_961 Alloprevotella Alloprevotella_rava
Seq_923 Roseburia Eubacterium_rectale
Seq_210 Tepidimonas unclassified
Seq_196 Prevotella Prevotella_nanceiensis
Seq_13 Neisseria unclassified
Seq_788 Prevotella unclassified
Seq_671 Haemophilus Haemophilus_influenzae
Seq_224 Catonella Catonella_morbi
Seq_589 Leptotrichia Leptotrichia_buccalis
Seq_391 unclassified unclassified
Seq_166 unclassified unclassified
Seq_1433 unclassified unclassified
Seq_435 Prevotella unclassified
Seq_498 unclassified unclassified
Seq_86 Neisseria unclassified
Seq_105 Prevotella Prevotella_histicola
Seq_61 Haemophilus unclassified
Seq_402 Streptococcus unclassified
Seq_28 Prevotella Prevotella_salivae
Seq_290 Megasphaera Megasphaera_micronuciformis
Seq_195 Streptobacillus unclassified
Seq_76 Streptococcus unclassified
Seq_107 Actinomyces Actinomyces_odontolyticus
Seq_344 Capnocytophaga Capnocytophaga_leadbetteri
Seq_220 Capnocytophaga Capnocytophaga_leadbetteri
Seq_134 Streptococcus unclassified
Seq_42 Campylobacter Campylobacter_concisus
Seq_1387 Leptotrichia unclassified
Seq_1194 Alloprevotella Alloprevotella_rava
Seq_571 Treponema Treponema_denticola
Seq_422 Campylobacter Campylobacter_gracilis
Seq_150 Porphyromonas Porphyromonas_catoniae
Seq_692 Fusobacterium Fusobacterium_nucleatum
Seq_188 Haemophilus Haemophilus_parainfluenzae
1–2
1–3
1–4
1–5
2–3
2–4
2–5
3–4
3–5
4–5
*<=0.05 **<=0.01 ***<=0.001
Difference
2
1
0
−1

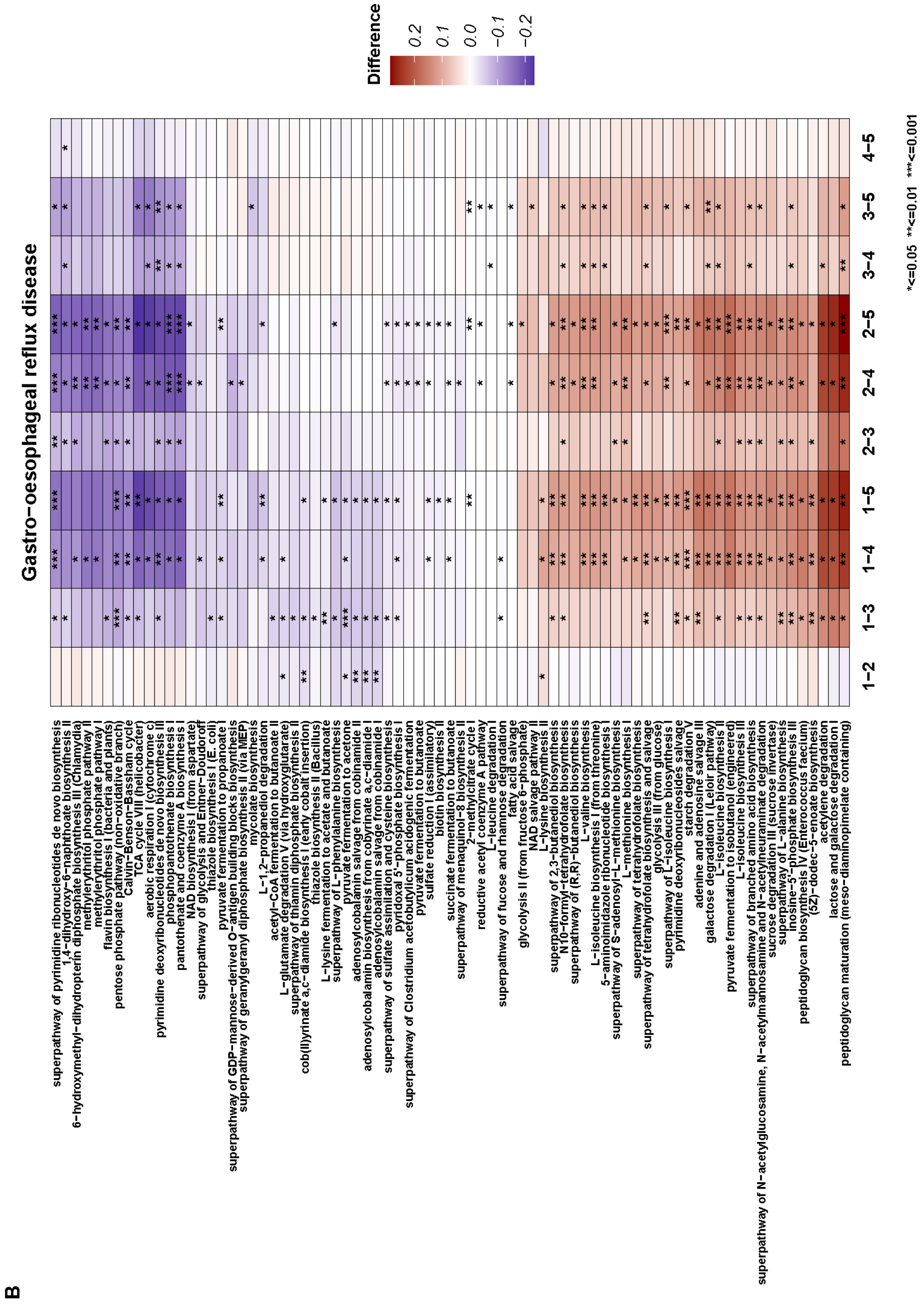

E
Oesophageal adenocarcinoma
Seq_11 Veillonella Veillonella_atypica
Seq_8 Veillonella Veillonella_dispar
Seq_23 unclassified unclassified
Seq_87 unclassified unclassified
Seq_62 Fusobacterium Fusobacterium_nucleatum
Seq_686 Treponema Treponema_medium
Seq_73 Lachnoanaerobaculum unclassified
1−2
1−3
1−4
1−5
2−3
2−4
2−5
3−4
3−5
4−5
*<=0.05 **<=0.01 ***<=0.001
Difference
0.5
0.0
−0.5
−1.0

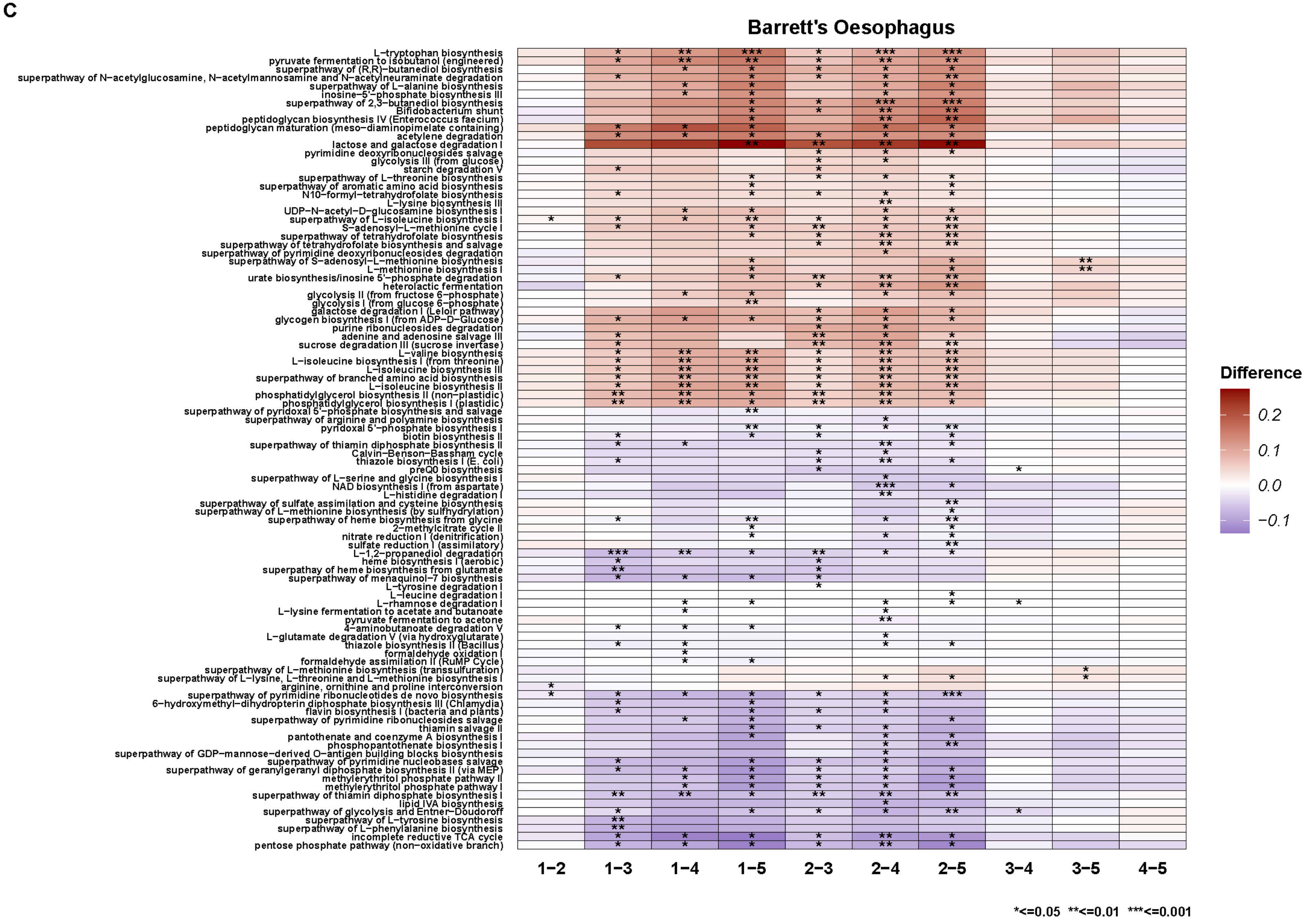


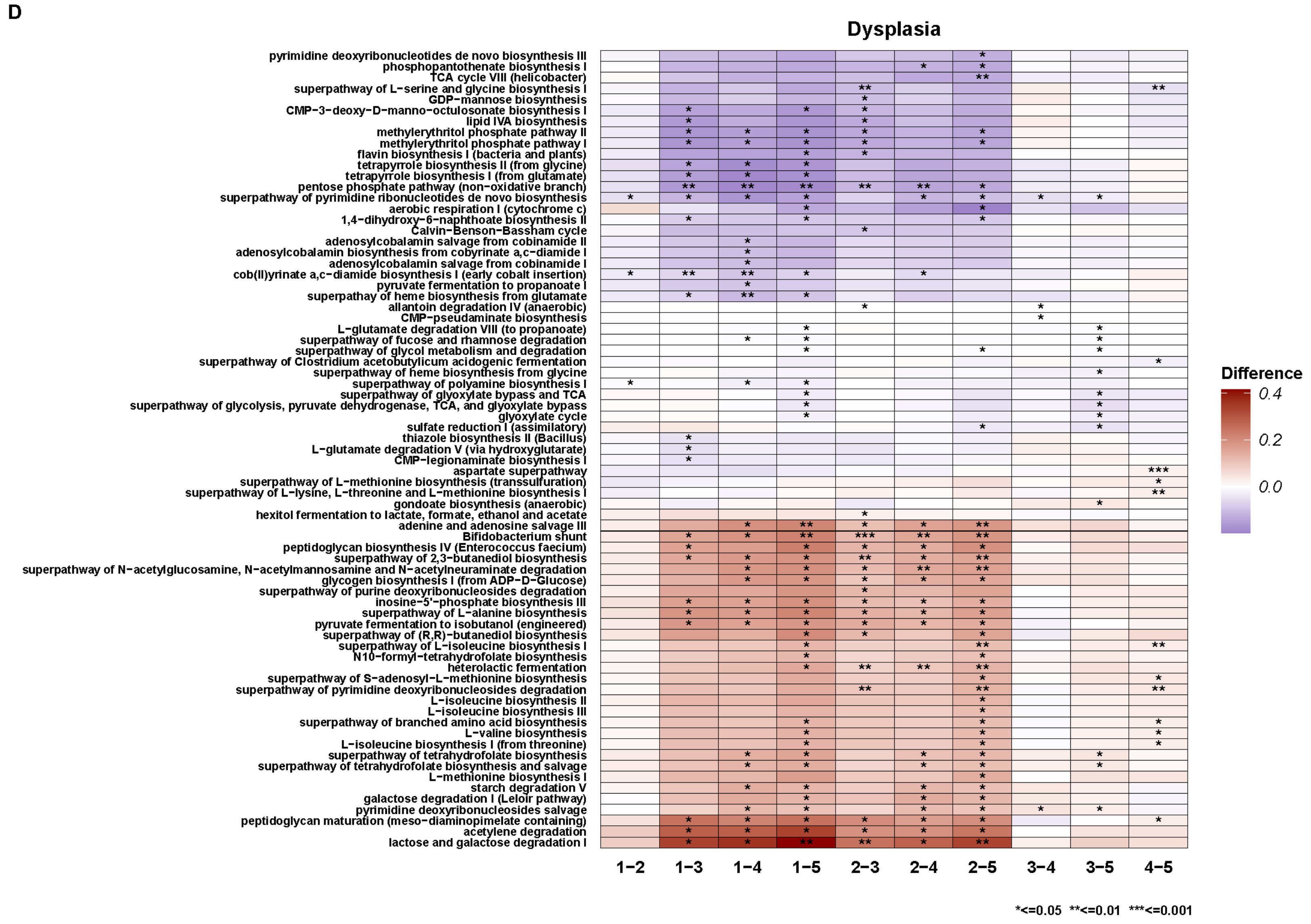
Supplementary figure 5. Differentially abundant ASVs between biopsy location. Heat-map of differential ASVs between each pair of biopsy location per clinical classification. Statistical testing was using paired Wilcoxon. **(A)** Data derived from Healthy controls. **(B)** Data derived from individuals with GORD. **(C)** Data derived from individuals with BO. **(D)** Data derived from individuals with Dysplasia. **(E)** Data derived from individuals with OAC. **(F)** Data derived from individuals with metastatic OAC.

Using the algorithm PICRUSt2, we inferred metabolic pathways from ASV data. A number of pathways were found to be differential abundant between biopsy sites derived from all clinical classifications (Supplementary figure 6). In line with differential species and ASVs, the number of metabolic pathways that were statistically different increases the further the sites were physically distant from each other.

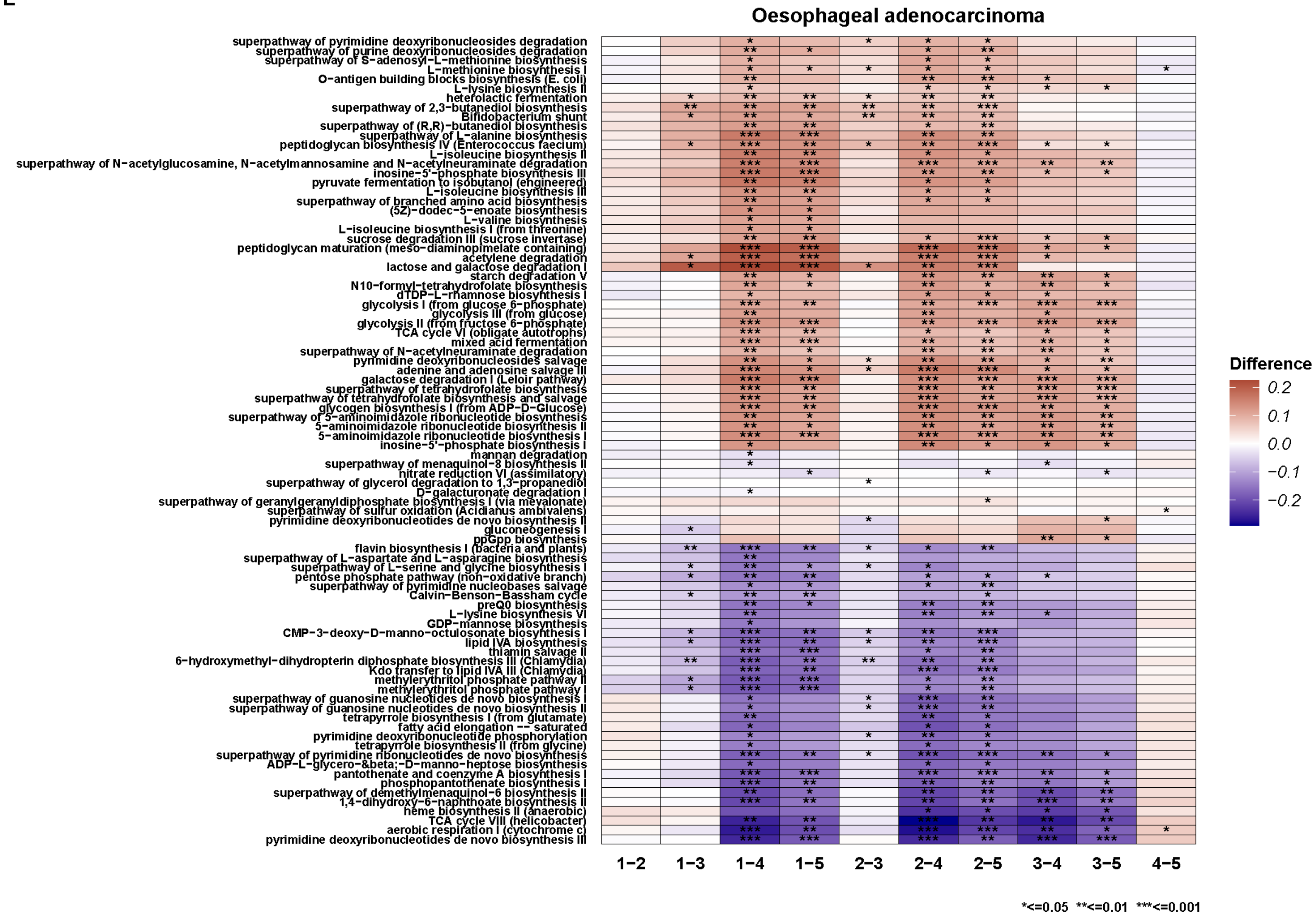

A
Healthy controls
superpathway of (R,R)−butanediol biosynthesis
O−antigen building blocks biosynthesis (E. coli)
superpathway of 2,3−butanediol biosynthesis
L−methionine biosynthesis I
sucrose degradation III (sucrose invertase)
superpathway of N−acetylglucosamine, N−acetylmannosamine and N−acetylneuraminate degradation
glycogen biosynthesis I (from ADP−D−Glucose)
superpathway of purine deoxyribonucleosides degradation
superpathway of L−alanine biosynthesis
inosine−5'−phosphate biosynthesis III
peptidoglycan biosynthesis IV (Enterococcus faecium)
aspartate superpathway
superpathway of L−methionine biosynthesis (transsulfuration)
L−lysine biosynthesis II
pyruvate fermentation to isobutanol (engineered)
acetylene degradation
peptidoglycan maturation (meso−diaminopimelate containing)
lactose and galactose degradation I
formaldehyde oxidation I
formaldehyde assimilation II (RuMP Cycle)
superpathway of polyamine biosynthesis I
biotin biosynthesis II
hexitol fermentation to lactate, formate, ethanol and acetate
reductive TCA cycle I
superpathway of GDP−mannose−derived O−antigen building blocks biosynthesis
L−histidine degradation I
flavin biosynthesis I (bacteria and plants)
6−hydroxymethyl−dihydropterin diphosphate biosynthesis III (Chlamydia)
superpathway of L−aspartate and L−asparagine biosynthesis
methylerythritol phosphate pathway II
methylerythritol phosphate pathway I
thiamin salvage II
phosphopantothenate biosynthesis I
pantothenate and coenzyme A biosynthesis I
CMP−3−deoxy−D−manno−octulosonate biosynthesis I
lipid IVA biosynthesis
Kdo transfer to lipid IVA III (Chlamydia)
superpathway of pyrimidine nucleobases salvage
pentose phosphate pathway (non−oxidative branch)
Calvin−Benson−Bassham cycle
1−2
1−3
1−4
1−5
2−3
2−4
2−5
3−4
3−5
4−5
Difference
0.4
0.2
0.0
−0.2
*<=0.05 **<=0.01 ***<=0.001

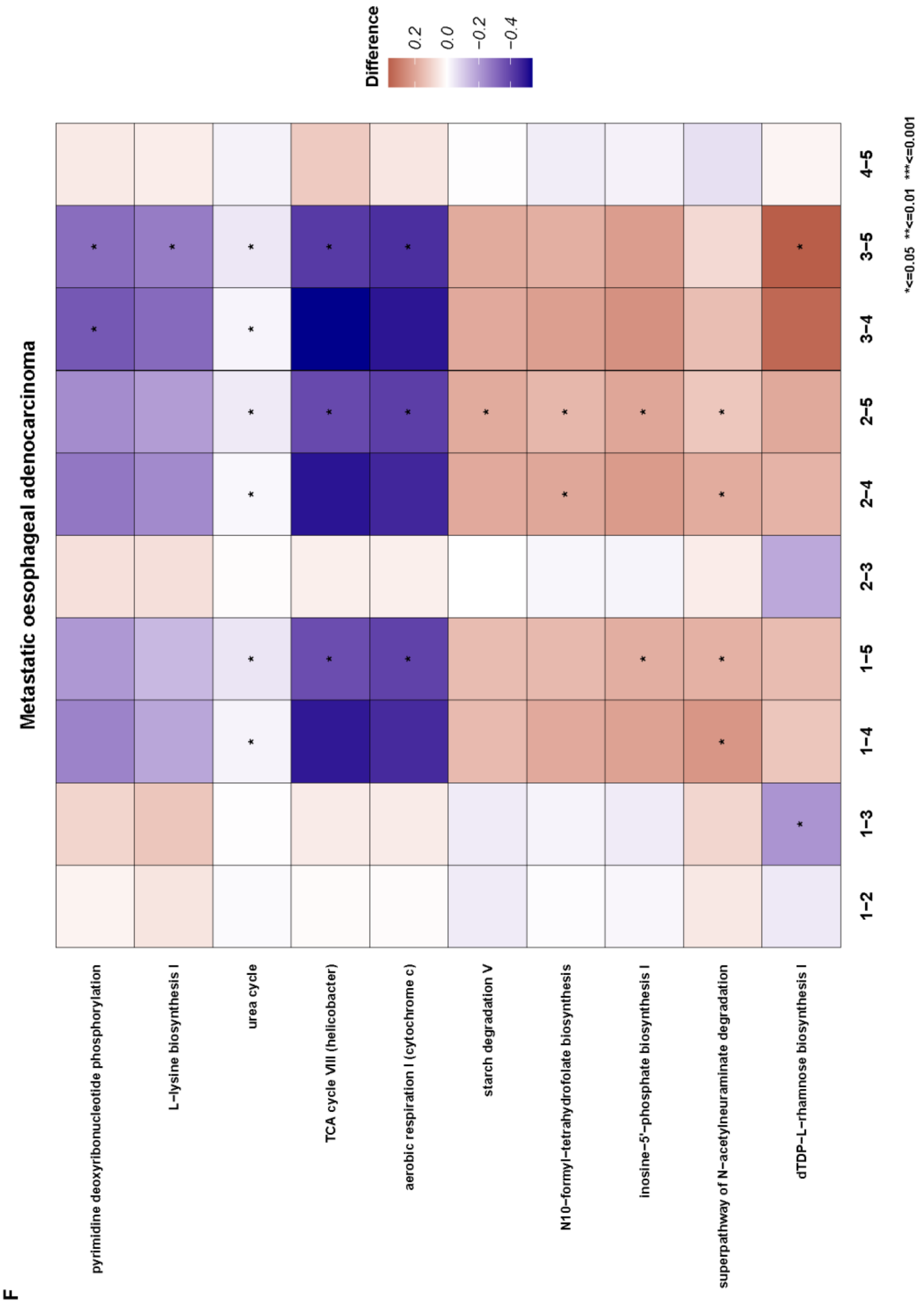

B
Gastro−oesophageal reflux disease
Difference
0.2
0.1
0.0
−0.1
−0.2
1−2
1−3
1−4
1−5
2−3
2−4
2−5
3−4
3−5
4−5
*<=0.05 **<=0.01 ***<=0.001
superpathway of pyrimidine ribonucleotides de novo biosynthesis
1,4−dihydroxy−6−naphthoate biosynthesis II
6−hydroxymethyl−dihydropterin diphosphate biosynthesis III (Chlamydia)
methylerythritol phosphate pathway II
methylerythritol phosphate pathway I
flavin biosynthesis I (bacteria and plants)
pentose phosphate pathway (non−oxidative branch)
Calvin−Benson−Bassham cycle
TCA cycle VIII (helicobacter)
aerobic respiration I (cytochrome c)
pyrimidine deoxyribonucleotides de novo biosynthesis III
phosphopantothenate biosynthesis I
pantothenate and coenzyme A biosynthesis I
NAD biosynthesis I (from aspartate)
superpathway of glycolysis and Entner−Doudoroff
thiazole biosynthesis I (E. coli)
pyruvate fermentation to propanoate I
superpathway of GDP−mannose−derived O−antigen building blocks biosynthesis
superpathway of geranylgeranyl diphosphate biosynthesis II (via MEP)
mycolate biosynthesis
L−1,2−propanediol degradation
acetyl−CoA fermentation to butanoate II
L−glutamate degradation V (via hydroxyglutarate)
superpathway of thiamin diphosphate biosynthesis II
cob(II)yrinate a,c−diamide biosynthesis I (early cobalt insertion)
thiazole biosynthesis II (Bacillus)
L−lysine fermentation to acetate and butanoate
superpathway of L−phenylalanine biosynthesis
pyruvate fermentation to acetone
adenosylcobalamin salvage from cobinamide II
adenosylcobalamin biosynthesis from cobyrinate a,c−diamide I
adenosylcobalamin salvage from cobinamide I
superpathway of sulfate assimilation and cysteine biosynthesis
pyridoxal 5'−phosphate biosynthesis I
superpathway of Clostridium acetobutylicum acidogenic fermentation
pyruvate fermentation to butanoate
sulfate reduction I (assimilatory)
biotin biosynthesis II
succinate fermentation to butanoate
superpathway of menaquinol−8 biosynthesis II
2−methylcitrate cycle I
reductive acetyl coenzyme A pathway
L−leucine degradation I
superpathway of fucose and rhamnose degradation
fatty acid salvage
glycolysis II (from fructose 6−phosphate)
NAD salvage pathway II
L−lysine biosynthesis II
superpathway of 2,3−butanediol biosynthesis
N10−formyl−tetrahydrofolate biosynthesis
superpathway of (R,R)−butanediol biosynthesis
L−valine biosynthesis
L−isoleucine biosynthesis I (from threonine)
5−aminoimidazole ribonucleotide biosynthesis I
superpathway of S−adenosyl−L−methionine biosynthesis
L−methionine biosynthesis I
superpathway of tetrahydrofolate biosynthesis
superpathway of tetrahydrofolate biosynthesis and salvage
glycolysis III (from glucose)
superpathway of L−isoleucine biosynthesis I
pyrimidine deoxyribonucleosides salvage
starch degradation V
adenine and adenosine salvage III
galactose degradation I (Leloir pathway)
L−isoleucine biosynthesis II
pyruvate fermentation to isobutanol (engineered)
L−isoleucine biosynthesis III
superpathway of branched amino acid biosynthesis
superpathway of N−acetylglucosamine, N−acetylmannosamine and N−acetylneuraminate degradation
sucrose degradation III (sucrose invertase)
superpathway of L−alanine biosynthesis
inosine−5'−phosphate biosynthesis III
peptidoglycan biosynthesis IV (Enterococcus faecium)
(5Z)−dodec−5−enoate biosynthesis
acetylene degradation
lactose and galactose degradation I
peptidoglycan maturation (meso−diaminopimelate containing)


C
Barrett's Oesophagus
L−tryptophan biosynthesis
pyruvate fermentation to isobutanol (engineered)
superpathway of (R,R)−butanediol biosynthesis
superpathway of N−acetylglucosamine, N−acetylmannosamine and N−acetylneuraminate degradation
superpathway of L−alanine biosynthesis
inosine−5'−phosphate biosynthesis III
superpathway of 2,3−butanediol biosynthesis
Bifidobacterium shunt
peptidoglycan biosynthesis IV (Enterococcus faecium)
peptidoglycan maturation (meso−diaminopimelate containing)
acetylene degradation
lactose and galactose degradation I
pyrimidine deoxyribonucleosides salvage
glycolysis III (from glucose)
starch degradation V
superpathway of L−threonine biosynthesis
superpathway of aromatic amino acid biosynthesis
N10−formyl−tetrahydrofolate biosynthesis
L−lysine biosynthesis III
UDP−N−acetyl−D−glucosamine biosynthesis I
superpathway of L−isoleucine biosynthesis I
S−adenosyl−L−methionine cycle I
superpathway of tetrahydrofolate biosynthesis
superpathway of tetrahydrofolate biosynthesis and salvage
superpathway of pyrimidine deoxyribonucleosides degradation
superpathway of S−adenosyl−L−methionine biosynthesis
L−methionine biosynthesis I
urate biosynthesis/inosine 5'−phosphate degradation
heterolactic fermentation
glycolysis II (from fructose 6−phosphate)
glycolysis I (from glucose 6−phosphate)
galactose degradation I (Leloir pathway)
glycogen biosynthesis I (from ADP−D−Glucose)
purine ribonucleosides degradation
adenine and adenosine salvage III
sucrose degradation III (sucrose invertase)
L−valine biosynthesis
L−isoleucine biosynthesis I (from threonine)
L−isoleucine biosynthesis III
superpathway of branched amino acid biosynthesis
L−isoleucine biosynthesis II
phosphatidylglycerol biosynthesis II (non−plastidic)
phosphatidylglycerol biosynthesis I (plastidic)
superpathway of pyridoxal 5'−phosphate biosynthesis and salvage
superpathway of arginine and polyamine biosynthesis
pyridoxal 5'−phosphate biosynthesis I
biotin biosynthesis II
superpathway of thiamin diphosphate biosynthesis II
Calvin−Benson−Bassham cycle
thiazole biosynthesis I (E. coli)
preQ0 biosynthesis
superpathway of L−serine and glycine biosynthesis I
NAD biosynthesis I (from aspartate)
L−histidine degradation I
superpathway of sulfate assimilation and cysteine biosynthesis
superpathway of L−methionine biosynthesis (by sulfhydrylation)
superpathway of heme biosynthesis from glycine
2−methylcitrate cycle II
nitrate reduction I (denitrification)
sulfate reduction I (assimilatory)
L−1,2−propanediol degradation
heme biosynthesis I (aerobic)
superpathay of heme biosynthesis from glutamate
superpathway of menaquinol−7 biosynthesis
L−tyrosine degradation I
L−leucine degradation I
L−rhamnose degradation I
L−lysine fermentation to acetate and butanoate
pyruvate fermentation to acetone
4−aminobutanoate degradation V
L−glutamate degradation V (via hydroxyglutarate)
thiazole biosynthesis II (Bacillus)
formaldehyde oxidation I
formaldehyde assimilation II (RuMP Cycle)
superpathway of L−methionine biosynthesis (transsulfuration)
superpathway of L−lysine, L−threonine and L−methionine biosynthesis I
arginine, ornithine and proline interconversion
superpathway of pyrimidine ribonucleotides de novo biosynthesis
6−hydroxymethyl−dihydropterin diphosphate biosynthesis III (Chlamydia)
flavin biosynthesis I (bacteria and plants)
superpathway of pyrimidine ribonucleosides salvage
thiamin salvage II
pantothenate and coenzyme A biosynthesis I
phosphopantothenate biosynthesis I
superpathway of GDP−mannose−derived O−antigen building blocks biosynthesis
superpathway of pyrimidine nucleobases salvage
superpathway of geranylgeranyl diphosphate biosynthesis II (via MEP)
methylerythritol phosphate pathway II
methylerythritol phosphate pathway I
superpathway of thiamin diphosphate biosynthesis I
lipid IVA biosynthesis
superpathway of glycolysis and Entner−Doudoroff
superpathway of L−tyrosine biosynthesis
superpathway of L−phenylalanine biosynthesis
incomplete reductive TCA cycle
pentose phosphate pathway (non−oxidative branch)
1−2
1−3
1−4
1−5
2−3
2−4
2−5
3−4
3−5
4−5
Difference
0.2
0.1
0.0
−0.1
*<=0.05 **<=0.01 ***<=0.001


D
Dysplasia
pyrimidine deoxyribonucleotides de novo biosynthesis III
phosphopantothenate biosynthesis I
TCA cycle VIII (helicobacter)
superpathway of L−serine and glycine biosynthesis I
GDP−mannose biosynthesis
CMP−3−deoxy−D−manno−octulosonate biosynthesis I
lipid IVA biosynthesis
methylerythritol phosphate pathway II
methylerythritol phosphate pathway I
flavin biosynthesis I (bacteria and plants)
tetrapyrrole biosynthesis II (from glycine)
tetrapyrrole biosynthesis I (from glutamate)
pentose phosphate pathway (non−oxidative branch)
superpathway of pyrimidine ribonucleotides de novo biosynthesis
aerobic respiration I (cytochrome c)
1,4−dihydroxy−6−naphthoate biosynthesis II
Calvin−Benson−Bassham cycle
adenosylcobalamin salvage from cobinamide II
adenosylcobalamin biosynthesis from cobyrinate a,c−diamide I
adenosylcobalamin salvage from cobinamide I
cob(II)yrinate a,c−diamide biosynthesis I (early cobalt insertion)
pyruvate fermentation to propanoate I
superpathay of heme biosynthesis from glutamate
allantoin degradation IV (anaerobic)
CMP−pseudaminate biosynthesis
L−glutamate degradation VIII (to propanoate)
superpathway of fucose and rhamnose degradation
superpathway of glycol metabolism and degradation
superpathway of Clostridium acetobutylicum acidogenic fermentation
superpathway of heme biosynthesis from glycine
superpathway of polyamine biosynthesis I
superpathway of glyoxylate bypass and TCA
superpathway of glycolysis, pyruvate dehydrogenase, TCA, and glyoxylate bypass
glyoxylate cycle
sulfate reduction I (assimilatory)
thiazole biosynthesis II (Bacillus)
L−glutamate degradation V (via hydroxyglutarate)
CMP−legionaminate biosynthesis I
aspartate superpathway
superpathway of L−methionine biosynthesis (transsulfuration)
superpathway of L−lysine, L−threonine and L−methionine biosynthesis I
gondoate biosynthesis (anaerobic)
hexitol fermentation to lactate, formate, ethanol and acetate
adenine and adenosine salvage III
Bifidobacterium shunt
peptidoglycan biosynthesis IV (Enterococcus faecium)
superpathway of 2,3−butanediol biosynthesis
superpathway of N−acetylglucosamine, N−acetylmannosamine and N−acetylneuraminate degradation
glycogen biosynthesis I (from ADP−D−Glucose)
superpathway of purine deoxyribonucleosides degradation
inosine−5'−phosphate biosynthesis III
superpathway of L−alanine biosynthesis
pyruvate fermentation to isobutanol (engineered)
superpathway of (R,R)−butanediol biosynthesis
superpathway of L−isoleucine biosynthesis I
N10−formyl−tetrahydrofolate biosynthesis
heterolactic fermentation
superpathway of S−adenosyl−L−methionine biosynthesis
superpathway of pyrimidine deoxyribonucleosides degradation
L−isoleucine biosynthesis II
L−isoleucine biosynthesis III
superpathway of branched amino acid biosynthesis
L−valine biosynthesis
L−isoleucine biosynthesis I (from threonine)
superpathway of tetrahydrofolate biosynthesis
superpathway of tetrahydrofolate biosynthesis and salvage
L−methionine biosynthesis I
starch degradation V
galactose degradation I (Leloir pathway)
pyrimidine deoxyribonucleosides salvage
peptidoglycan maturation (meso−diaminopimelate containing)
acetylene degradation
lactose and galactose degradation I
1−2
1−3
1−4
1−5
2−3
2−4
2−5
3−4
3−5
4−5
Difference
0.4
0.2
0.0
*<=0.05 **<=0.01 ***<=0.001

E


Oesophageal adenocarcinoma
superpathway of pyrimidine deoxyribonucleosides degradation
superpathway of purine deoxyribonucleosides degradation
superpathway of S−adenosyl−L−methionine biosynthesis
L−methionine biosynthesis I
O−antigen building blocks biosynthesis (E. coli)
L−lysine biosynthesis II
heterolactic fermentation
superpathway of 2,3−butanediol biosynthesis
Bifidobacterium shunt
superpathway of (R,R)−butanediol biosynthesis
superpathway of L−alanine biosynthesis
peptidoglycan biosynthesis IV (Enterococcus faecium)
L−isoleucine biosynthesis II
superpathway of N−acetylglucosamine, N−acetylmannosamine and N−acetylneuraminate degradation
inosine−5'−phosphate biosynthesis III
pyruvate fermentation to isobutanol (engineered)
L−isoleucine biosynthesis III
superpathway of branched amino acid biosynthesis
(5Z)−dodec−5−enoate biosynthesis
L−valine biosynthesis
L−isoleucine biosynthesis I (from threonine)
sucrose degradation III (sucrose invertase)
peptidoglycan maturation (meso−diaminopimelate containing)
acetylene degradation
lactose and galactose degradation I
starch degradation V
N10−formyl−tetrahydrofolate biosynthesis
dTDP−L−rhamnose biosynthesis I
glycolysis I (from glucose 6−phosphate)
glycolysis III (from glucose)
glycolysis II (from fructose 6−phosphate)
TCA cycle VI (obligate autotrophs)
mixed acid fermentation
superpathway of N−acetylneuraminate degradation
pyrimidine deoxyribonucleosides salvage
adenine and adenosine salvage III
galactose degradation I (Leloir pathway)
superpathway of tetrahydrofolate biosynthesis
superpathway of tetrahydrofolate biosynthesis and salvage
glycogen biosynthesis I (from ADP−D−Glucose)
superpathway of 5−aminoimidazole ribonucleotide biosynthesis
5−aminoimidazole ribonucleotide biosynthesis II
5−aminoimidazole ribonucleotide biosynthesis I
inosine−5'−phosphate biosynthesis I
mannan degradation
superpathway of menaquinol−8 biosynthesis II
nitrate reduction VI (assimilatory)
superpathway of glycerol degradation to 1,3−propanediol
D−galacturonate degradation I
superpathway of geranylgeranyldiphosphate biosynthesis I (via mevalonate)
superpathway of sulfur oxidation (Acidianus ambivalens)
pyrimidine deoxyribonucleotides de novo biosynthesis II
gluconeogenesis I
ppGpp biosynthesis
flavin biosynthesis I (bacteria and plants)
superpathway of L−aspartate and L−asparagine biosynthesis
superpathway of L−serine and glycine biosynthesis I
pentose phosphate pathway (non−oxidative branch)
superpathway of pyrimidine nucleobases salvage
Calvin−Benson−Bassham cycle
preQ0 biosynthesis
L−lysine biosynthesis VI
GDP−mannose biosynthesis
CMP−3−deoxy−D−manno−octulosonate biosynthesis I
lipid IVA biosynthesis
thiamin salvage II
6−hydroxymethyl−dihydropterin diphosphate biosynthesis III (Chlamydia)
Kdo transfer to lipid IVA III (Chlamydia)
methylerythritol phosphate pathway II
methylerythritol phosphate pathway I
superpathway of guanosine nucleotides de novo biosynthesis I
superpathway of guanosine nucleotides de novo biosynthesis II
tetrapyrrole biosynthesis I (from glutamate)
fatty acid elongation −− saturated
pyrimidine deoxyribonucleotide phosphorylation
tetrapyrrole biosynthesis II (from glycine)
superpathway of pyrimidine ribonucleotides de novo biosynthesis
ADP−L−glycero−&beta;−D−manno−heptose biosynthesis
pantothenate and coenzyme A biosynthesis I
phosphopantothenate biosynthesis I
superpathway of demethylmenaquinol−6 biosynthesis II
1,4−dihydroxy−6−naphthoate biosynthesis II
heme biosynthesis II (anaerobic)
TCA cycle VIII (helicobacter)
aerobic respiration I (cytochrome c)
pyrimidine deoxyribonucleotides de novo biosynthesis III
1−2
1−3
1−4
1−5
2−3
2−4
2−5
3−4
3−5
4−5
Difference
0.2
0.1
0.0
−0.1
−0.2
*<=0.05 **<=0.01 ***<=0.001



Supplementary figure 6. Differentially abundant microbiome-encoded metabolic pathways between biopsy locations. Heat-map of differential species between each pair of biopsy location per clinical classification. (A) Data derived from Healthy controls. (B) Data derived from individuals with GORD. (C) Data derived from individuals with BO. (D) Data derived from individuals with Dysplasia. (E) Data derived from individuals with OAC. (F) Data derived from individuals with metastatic OAC. Statistical testing was using paired Wilcoxon.

2.5 Discussion

In this study we identified a number of microbiome features which differed between clinical classifications along the oesophageal adenocarcinoma sequence. We also identified microbiome differences within the upper digestive tract of the individuals along the OAC sequence.

Although the microbiome did not dramatically differ with respect to biopsy location, a shift in the microbiome was detected as measured by beta-diversity. We observed that the microbiome of metaplastic tissue derived from individuals with BO had a higher alpha diversity relative to that of the adjacent tissue. Metaplastic tissue has a crypt structure similar to that of the intestine. This structure could possibly allow for growth of a more diverse range of bacterial taxa. We did not observe a significant difference in alpha diversity between clinical groups. This is in contrast to previous reports by Elliott et al who identified a lower alpha diversity in cancer samples relative to BO samples and healthy control samples[27]. Differences in sample depth may explain this disparity as the current study has more than a 3X greater minimum sequencing depth relative to the Elliott et al study.

Previous studies have identified an enrichment of *F. nucleatum* on tumour samples relative to matched normal tissue in the context of colorectal cancer and breast cancer[30,31]. We did not find any conclusive evidence for the enrichment of *F. nucleatum* on OAC tissue relative to matched healthy controls. However, we did find a particular ASV to be enriched on adenocarcinoma samples relative to adjacent gastric samples.

At a macroecological level, the microbiome associated with the various stages along the oesophageal adenocarcinoma sequence did not differ dramatically as measured by alpha and beta-diversity. We did see a difference in beta-diversity in samples derived from the GOJ (Biopsy location 3) as well as gastric biopsies (Biopsy location 4 and 5).

A history of periodontal disease has been associated with OAC, with a 43% and 52% increased risk[32]. *P. denticola and Bifidobacterium dentium* have been implicated in the development of dental caries[33,34]. We observed these taxa to be enriched in disease groups relative to healthy controls in a number of biopsy sites. Previous work by Elliott et al identified *Lactobacillus fermentum,* also a caries-associated taxon, as being enriched in the oesophageal microbiome of individuals with OAC[35]. The acid resistant nature of these taxa may provide a selective advantage to grow in an oesophagus with abnormally low pH due to acid reflux.

We identified an ASV assigned to *F. nucleatum* to be enriched in the disease groups relative to healthy controls in oesophageal-derived samples. A growing body of work has linked *F. nucleatum* to CRC oncogenesis both by association, but also and mechanistically[36]. An enrichment of *F. nucleatum* in oesophageal samples has been previously reported to be associated with a poorer prognosis as it relates to oesophageal squamous cell carcinoma[37,38].

At the GOJ, *F. necrophorum* was observed to be enriched in subjects/biopsies with dysplastic and neoplastic presentations versus those without[39]. In a recent meta-analysis, an enrichment of *F. necrophorum* in the colon microbiome was associated with colorectal cancer. *F. necrophorum* can be described as an opportunistic pathogen which is a canonical resident of the human alimentary canal. *F.*

necrophorum is a causative agent of Lemierre's syndrome which is characterised by a septic thrombophlebitis of the internal jugular vein[40]. Furthermore, *Fusobacterium necrophorum* is known to cause other infections of the head and neck including non-streptococcal tonsillitis and peritonsillar abscess[41,42]. What drives the progression of the oesophageal adenocarcinoma sequence remains an area of intense research. An inflammatory response to chronic colonization by *F. necrophorum* may promote oncogenesis. However, one could not rule out a model where *F. necrophorum* opportunistically grows in the setting of diseased tissue.

We found microbiome-encoded pathways relating to B12 synthesis to be depleted in the metastatic OAC cohort. Increased levels of serum B12 has been previously associated with increased mortality in the context of cancer[43]. One might speculate that an increasing level of B12 in the environment of a microbe would lead to the down regulation of B12 synthetic pathways as the need for microbes to synthesise their own B12 would be attenuated.

A number of limitations within this study should be noted. Some of the clinical groups within this study, particular the healthy control group and metastatic OAC group, have low numbers of individuals. As noted, there is a bias in terms of sex in the clinical groups with those of the male sex being more frequent in the later stages of the OAC sequence. As sex is known to associate with differences in gut microbiome this sex driven variation may be also found in the oesophagus[44]. No quantitative microbiome data was gathered during this study. One might expect significant variation in microbial load between clinical groups. As mentioned the crypt like structure of BO may provide a niche which allows a higher alpha diversity but may also allow a greater bacterial load. Furthermore, one would expect the gastric microbiota to have a higher biomass than oesophageal microbiota. It has been previously reported that the use of different primer pairs led to different levels of off target amplification of human DNA[45]. In this study we used the V3 V4 primer pair which has been observed to amplify Human DNA more than the V1 V3 primer pair.

Characterising the upper digestive tract microbiome in the context of oesophageal adenocarcinoma may provide information pertaining to OAC oncogenesis, detection, and therapeutic development. Bacterial taxa which promote inflammation including those associated with periodontal disease may provide a tumorigenic microenvironment which promotes cancer development. Even if these taxa do not directly drive oncogenesis, their abundance may be directly associated with the oncogenesis process. Thus, taxa associated with adenocarcinoma process may provide diagnostic or prognostic information. A microbe such as *F. necrophorum* may provide prognostic data with respect to delineating which individuals with BO will go on to develop OAC. Recently, adjuvant immune checkpoint inhibitor treatment was demonstrated to increase disease-free survival in patients with OAC[46]. The gut microbiome has been associated with the efficacy of immune checkpoint inhibitors[47,48]. It is possible that the local microbiome of the oesophagus may modulate the immune microenvironment of OAC and thus the efficiency of immune checkpoint inhibitors.

Further research such as longitudinal studies and mechanistic assays will be needed to further validate the findings of this study and the accompanying inferences.

2.6 Acknowledgments

This work was supported by The Health Research Board of Ireland under Grant ILP-POR-2017-034, and by Science Foundation Ireland (APC/SFI/12/RC/2273_P2) in the form of a research centre, APC Microbiome Ireland.

2.7 References

1 Sung, H. *et al.* Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*, doi:10.3322/caac.21660 (2021).

2 Neal, R. D. *et al.* Is increased time to diagnosis and treatment in symptomatic cancer associated with poorer outcomes? Systematic review. *British journal of cancer* **112 Suppl 1**, S92-S107, doi:10.1038/bjc.2015.48 (2015).

3 Anderson, L. A. *et al.* Survival for oesophageal, stomach and small intestine cancers in Europe 1999-2007: Results from EUROCARE-5. *Eur J Cancer* **51**, 2144-2157, doi:10.1016/j.ejca.2015.07.026 (2015).

4 Smyth, E. C. *et al.* Oesophageal cancer. *Nat Rev Dis Primers* **3**, 17048, doi:10.1038/nrdp.2017.48 (2017).

5 Thrift, A. P. Global burden and epidemiology of Barrett oesophagus and oesophageal cancer. *Nature Reviews Gastroenterology & Hepatology*, doi:10.1038/s41575-021-00419-3 (2021).

6 Richter, J. E. & Rubenstein, J. H. Presentation and Epidemiology of Gastroesophageal Reflux Disease. *Gastroenterology* **154**, 267-276, doi:https://doi.org/10.1053/j.gastro.2017.07.045 (2018).

7 Tan, M. C. *et al.* Systematic review with meta-analysis: prevalence of prior and concurrent Barrett's oesophagus in oesophageal adenocarcinoma patients. *Aliment Pharmacol Ther* **52**, 20-36, doi:10.1111/apt.15760 (2020).

8 Curtius, K., Rubenstein, J. H., Chak, A. & Inadomi, J. M. Computational modelling suggests that Barrett’s oesophagus may be the precursor of all oesophageal adenocarcinomas. *Gut*, gutjnl-2020-321598, doi:10.1136/gutjnl-2020-321598 (2020).

9 Stachler, M. D. *et al.* Detection of Mutations in Barrett's Esophagus Before Progression to High-Grade Dysplasia or Adenocarcinoma. *Gastroenterology* **155**, 156-167, doi:10.1053/j.gastro.2018.03.047 (2018).

10 Thrift, A. P. Global burden and epidemiology of Barrett oesophagus and oesophageal cancer. *Nature Reviews Gastroenterology & Hepatology*, 1-12 (2021).

11 Erőss, B. *et al.* Helicobacter pylori infection reduces the risk of Barrett's esophagus: A meta-analysis and systematic review. *Helicobacter* **23**, e12504, doi:10.1111/hel.12504 (2018).

12 Wang, Z. *et al.* Helicobacter pylori Infection Is Associated With Reduced Risk of Barrett's Esophagus: An Analysis of the Barrett's and Esophageal Adenocarcinoma Consortium. *Am J Gastroenterol* **113**, 1148-1155, doi:10.1038/s41395-018-0070-3 (2018).

13 Barrett, M., Hand, C. K., Shanahan, F., Murphy, T. & O'Toole, P. W. Mutagenesis by microbe: The role of the microbiota in shaping the cancer genome. *Trends in cancer* **6**, 277-287 (2020).

14 Sepich-Poore, G. D. *et al.* The microbiome and human cancer. *Science* **371** (2021).

15 Ternes, D. *et al.* Microbiome in colorectal cancer: how to get from meta-omics to mechanism? *Trends in microbiology* **28**, 401-423 (2020).

16 Kostic, A. D. *et al.* Fusobacterium nucleatum potentiates intestinal tumorigenesis and modulates the tumor-immune microenvironment. *Cell Host Microbe* **14**, 207-215, doi:10.1016/j.chom.2013.07.007 (2013).

17 Pleguezuelos-Manzano, C. *et al.* Mutational signature in colorectal cancer caused by genotoxic pks+ E. coli. *Nature* **580**, 269-273, doi:10.1038/s41586-020-2080-8 (2020).

18 Flemer, B. *et al.* Tumour-associated and non-tumour-associated microbiota in colorectal cancer. *Gut* **66**, 633-643, doi:10.1136/gutjnl-2015-309595 (2017).

19 Amplicon, P., Clean-Up, P. & Index, P. (2013).

20 Klindworth, A. *et al.* Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Res* **41**, e1, doi:10.1093/nar/gks808 (2013).

21 Callahan, B. J. *et al.* DADA2: High-resolution sample inference from Illumina amplicon data. *Nat Methods* **13**, 581-583, doi:10.1038/nmeth.3869 (2016).

22 Schloss, P. D. *et al.* Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* **75**, 7537-7541, doi:10.1128/AEM.01541-09 (2009).

23 Allard, G., Ryan, F. J., Jeffery, I. B. & Claesson, M. J. SPINGO: a rapid species-classifier for microbial amplicon sequences. *BMC Bioinformatics* **16**, 324, doi:10.1186/s12859-015-0747-1 (2015).

24 Caporaso, J. G. *et al.* QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* **7**, 335-336, doi:10.1038/nmeth.f.303 (2010).

25 Martino, C. *et al.* A Novel Sparse Compositional Technique Reveals Microbial Perturbations. *mSystems* **4**, e00016-00019, doi:10.1128/mSystems.00016-19 (2019).

26 Douglas, G. M. *et al.* PICRUSt2 for prediction of metagenome functions. *Nature Biotechnology* **38**, 685-688, doi:10.1038/s41587-020-0548-6 (2020).

27 Elliott, D. R. F., Walker, A. W., O'Donovan, M., Parkhill, J. & Fitzgerald, R. C. A non-endoscopic device to sample the oesophageal microbiota: a case-control study. *Lancet Gastroenterol Hepatol* **2**, 32-42, doi:10.1016/S2468-1253(16)30086-3 (2017).

28 Deshpande, N. P., Riordan, S. M., Castano-Rodriguez, N., Wilkins, M. R. & Kaakoush, N. O. Signatures within the esophageal microbiome are associated with host genetics, age, and disease. *Microbiome* **6**, 227, doi:10.1186/s40168-018-0611-4 (2018).

29 Nardone, G., Compare, D. & Rocco, A. A microbiota-centric view of diseases of the upper gastrointestinal tract. *The Lancet Gastroenterology & Hepatology* **2**, 298-312 (2017).

30 Castellarin, M. *et al.* Fusobacterium nucleatum infection is prevalent in human colorectal carcinoma. *Genome Res* **22**, 299-306, doi:10.1101/gr.126516.111 (2012).

31 Parhi, L. *et al.* Breast cancer colonization by Fusobacterium nucleatum accelerates tumor growth and metastatic progression. *Nature Communications* **11**, 3259, doi:10.1038/s41467-020-16967-2 (2020).

32 Lo, C.-H. *et al.* Periodontal disease, tooth loss, and risk of oesophageal and gastric adenocarcinoma: a prospective study. *Gut* **70**, 620-621 (2021).

33 Zhang, L. *et al.* Quantitative analysis of salivary oral bacteria associated with severe early childhood caries and construction of caries assessment model. *Scientific reports* **10**, 1-8 (2020).

34 Nakajo, K., Takahashi, N. & Beighton, D. Resistance to acidic environments of caries-associated bacteria: Bifidobacterium dentium and Bifidobacterium longum. *Caries research* **44**, 431-437 (2010).

35 Elliott, D. R. F., Walker, A. W., O'Donovan, M., Parkhill, J. & Fitzgerald, R. C. A non-endoscopic device to sample the oesophageal microbiota: a case-control study. *The lancet Gastroenterology & hepatology* **2**, 32-42 (2017).

36 Brennan, C. A. & Garrett, W. S. Fusobacterium nucleatum - symbiont, opportunist and oncobacterium. *Nature reviews. Microbiology* **17**, 156-166, doi:10.1038/s41579-018-0129-6 (2019).

37 Yamamura, K. *et al.* Human microbiome Fusobacterium nucleatum in esophageal cancer tissue is associated with prognosis. *Clinical Cancer Research* **22**, 5574-5581 (2016).

38 Yamamura, K. *et al.* Intratumoral Fusobacterium nucleatum levels predict therapeutic response to neoadjuvant chemotherapy in esophageal squamous cell carcinoma. *Clinical Cancer Research* **25**, 6170-6179 (2019).

39 Thomas, A. M. *et al.* Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. *Nature medicine* **25**, 667-678 (2019).

40 Kuppalli, K., Livorsi, D., Talati, N. J. & Osborn, M. Lemierre's syndrome due to Fusobacterium necrophorum. *The Lancet infectious diseases* **12**, 808-815 (2012).

41 Atkinson, T. P. *et al.* Analysis of the tonsillar microbiome in young adults with sore throat reveals a high relative abundance of Fusobacterium necrophorum with low diversity. *PloS one* **13**, e0189423 (2018).

42 Ehlers Klug, T., Rusan, M., Fuursted, K. & Ovesen, T. Fusobacterium necrophorum: most prevalent pathogen in peritonsillar abscess in Denmark. *Clinical Infectious Diseases* **49**, 1467-1472 (2009).

43 Arendt, J. F. H., Farkas, D. K., Pedersen, L., Nexo, E. & Sørensen, H. T. Elevated plasma vitamin B12 levels and cancer prognosis: A population-based cohort study. *Cancer epidemiology* **40**, 158-165 (2016).

44 Zhang, X. *et al.* Sex-and age-related trajectories of the adult human gut microbiota shared across populations of different ethnicities. *Nature Aging* **1**, 87-100 (2021).

45 Walker, S. P. *et al.* Non-specific amplification of human DNA is a major challenge for 16S rRNA gene sequence analysis. *Scientific Reports* **10**, 16356, doi:10.1038/s41598-020-73403-7 (2020).

46 Kelly, R. J. *et al.* Adjuvant nivolumab in resected esophageal or gastroesophageal junction cancer. *New England Journal of Medicine* **384**, 1191-1203 (2021).

47 Baruch, E. N. *et al.* Fecal microbiota transplant promotes response in immunotherapy-refractory melanoma patients. *Science* **371**, 602-609 (2021).

48 Davar, D. *et al.* Fecal microbiota transplant overcomes resistance to anti–PD-1 therapy in melanoma patients. *Science* **371**, 595-602 (2021).

Chapter 3 - Mapping the colorectal tumour microbiota.

This work has been accepted for publication in the journal *Gut Microbes*.

Authors:

Clodagh L Murphy*, Maurice Barrett*, Paola Pellanda, Shane Kileen, Morgan McCourt, Micheal O'Riordain, Fergus Shanahan, Paul W O'Toole

*Joint first authorship: These authors contributed equally to this work.

Citation:

MURPHY, C., BARRETT, M., PELLANDA, P., KILLEEN, S., MCCOURT, M., ANDREWS, E., O'RIORDAIN, M., SHANAHAN, F. & O'TOOLE, P. 2021. Mapping the colorectal tumor microbiota. *Gut Microbes*, 13**,** 1-10.

Maurice Barrett contributed to this work as follows:

- All bioinformatic analysis including sequence processing, compositional data analysis and statistical analysis.
- Data visualization i.e., construction of publication figures.
- Writing over half of the manuscript.

3.1 Abstract

The gut microbiome in patients with colorectal cancer (CRC) is different than that of healthy controls. Previous studies have profiled the CRC tumor microbiome using a single biopsy. However, since the morphology and cellular subtype vary significantly within an individual tumor, the possibility of sampling error arises for the microbiome within an individual tumor. To test this hypothesis, seven biopsies were taken from representative areas on and off the tumor in five patients with CRC. The microbiome composition was strikingly similar across all samples from an individual. The variation in microbiome alpha-diversity was significantly greater between individuals' samples then within individuals. This is the first study, to our knowledge, that shows that the microbiome of an individual tumor is spatially homogeneous. Our finding strengthens the assumption that a single biopsy is representative of the entire tumor, and that microbiota changes are not limited to a specific area of the neoplasm.

Keywords: colorectal, cancer, microbiome, tumor, gut

3.2 Introduction

Colorectal cancer (CRC) is the second largest cause of cancer death in the United States[1]. Sporadic CRC arises after a series of cumulative genetic mutations[2], with a ten year progression from adenoma to CRC[3]. The microbiome is distinctly different in biopsies of CRC and adenomatous polyps[4] [5], leading to an updated hypothesis that microbial changes[6] and secondary consequences for immunological cell signalling[7] may play a role in tumor progression. Bacteria are an established risk factor for cancer, such as *H. pylori*-related MALT lymphoma and gastric carcinoma[8,9]. In particular, several individual microbes such as *Fusobacterium nucleatum*[10] *and Escherichia coli*[11] *have been implicated in the pathogenesis of colorectal cancer, but a cause-effect relationship has not been established; rather,* microbes and their metabolome represent complex collections of gene networks that interact bidirectionally with cancer cells[12].

CRC-associated microbiota is characterized by a reduced alpha diversity compared with healthy controls[13]. Patients with CRC[4,14] or adenomatous polyps[4,15] show also distinct qualitative differences in both the microbiome and metabolome in fecal[16,17] and biopsy samples[4,14] compared with healthy controls. In these studies, the microbiota associated with cancerous and non-cancerous tissues within the same individual did not differ significantly[4,14] which suggests that in CRC, a global microbial ecosystem change occurs throughout the colon[4,18]. However, the microbial alterations differ between proximal and distal cancers[4]. These compositional changes often represent a relative over-abundance of oral bacteria, which are hypothesized to organize into biofilm-like structures[19] on the tumor and on the right side of the colon[4,20]. We have previously described that CRC patients can be stratified into four groups based on bacterial co-abundance groups (CAGs) that link distinct mucosal gene-expression profiles[4] with similar networks of oral-based bacteria found on the gut mucosa and oral mucosa[18,20,21].

Distinct morphological and phenotypical differences exist within and between colorectal tumours[22]. Classification systems such as NICE[23], Paris[24] and Kudo[25] use macroscopically visible differences in lesions to stratify malignant potential[24] or stage neoplastic tumors[26] detected at the time of endoscopy. Similarly, the World Health Organization (WHO) has classified the appearances of colorectal tumors at surgery into four groups: exophytic, endophytic, diffusely infiltrative and annular, with the recognition that significant overlap occurs between these categories[27]. Macroscopic phenotypes may also be an overall predictor of genetic alterations and DNA methylation in a colorectal tumor[28]. Intra-tumoral heterogeneity for both genetic and epigenetic factors in CRC are also evident[29].

Untargeted colonoscopy biopsies or untargeted segments of resected tumors has been used in most studies of CRC microbiota[4,14,30,31]. Given the histologic and genetic intra-tumoral heterogeneity[32] of CRC, topographic variance in the microbiota of a single tumor may be a confounding factor. Therefore, we undertook the first study aims to investigate the intra-tumoral microbial heterogeneity and its comparison with adjacent proximal and distal non-cancerous tissue.

3.3 Results

Five patients were recruited to the study, four males and one female, with a mean age of 72 ± 6.7 years as shown in Table 1. All patients had a diagnosis of colonic adenocarcinoma within the previous 1-2 months. Seven samples were obtained from each individual comprising normal tissue proximal to the tumor (biopsy 6), normal tissue distal to the tumor (biopsy 5), a central tumoral biopsy (biopsy 5) and four peripheral tumor biopsies (biopsies 1-4). The tissue microbiome was profiled by 16S rRNA gene amplicon sequencing.

Patient	GT 001	GT 007	GT 009	GT 010	GT 011
Type of neoplasm	Adenocarcinoma	Adenocarcinoma	Adenocarcinoma	Adenocarcinoma	Adenocarcinoma
Tumor location	rectum	transverse colon	sigmoid colon	caecum	ascending colon
Stage of neoplasm	T3N0M0	T3N0M0	T3N1M0	T3N1M0	T3N0M0
Time since diagnosis (months)	1	1	1	1	2
Type of surgery	Anterior resection	Right hemi-colectomy	Anterior resection	Right hemi-colectomy	Right hemi-colectomy
Bowel Prep	Moviprep	Moviprep	Moviprep	Moviprep	Moviprep
Alcohol intake per weeks	10 units	none	3 units	none	none
Smoking status	Current (2/day)	Ex-smoker (10/day x20years)	Ex-smoker (20/day x40years)	Non smoker	Ex-Smoker (10/day x35 years)
Probiotic use	No	No	No	No	No
Antibiotic exposure	No	Yes	Yes	Yes	Yes
Antibiotic regime used at surgery	N/A	IV co-amoxiclav and metronidazole	Oral metonidazole and neomycin	Oral metonidazole and neomycin	IV co-amoxiclav and metronidazole
Diverticulae	no	no	no	no	no
Medical comorbidites	none	Hypertension, NIDDM	NIDDM, obstructive uropathy	Hypertension, anemia	Epilepsy, NIDDM, hypertension, hyperlipidemia
Medications	nil	aspirin, ramipril, esomprazole, atorvastatin, empagliflozin, metformin	atorvastatin	ramipril, lercanidipine, ferrous fumerate	bisoprolol, ezetimibe, rosuvastatin, hyoscine butylbromide, esomprazole, lercanidipine, carbamazepinesitagliptin, metformin

Table 1. Patient characteristics. Footnote: n = 4 males, 1 female, with a mean age of 72 ± 6.7 years

The microbiome composition was highly similar among samples within a particular individual (Figure 1A). The genus level composition differed significantly between patients (Figure 1A) but was remarkably similar within a single subject, both on (biopsy 1-5) and off the tumor site (biopsy 6 and 7). This was reflected in beta diversity distance metrics wherein samples clustered by individual rather that biopsy site as represented in Principal Co-ordinate Analysis (PCoA) plots (Figure 1B).

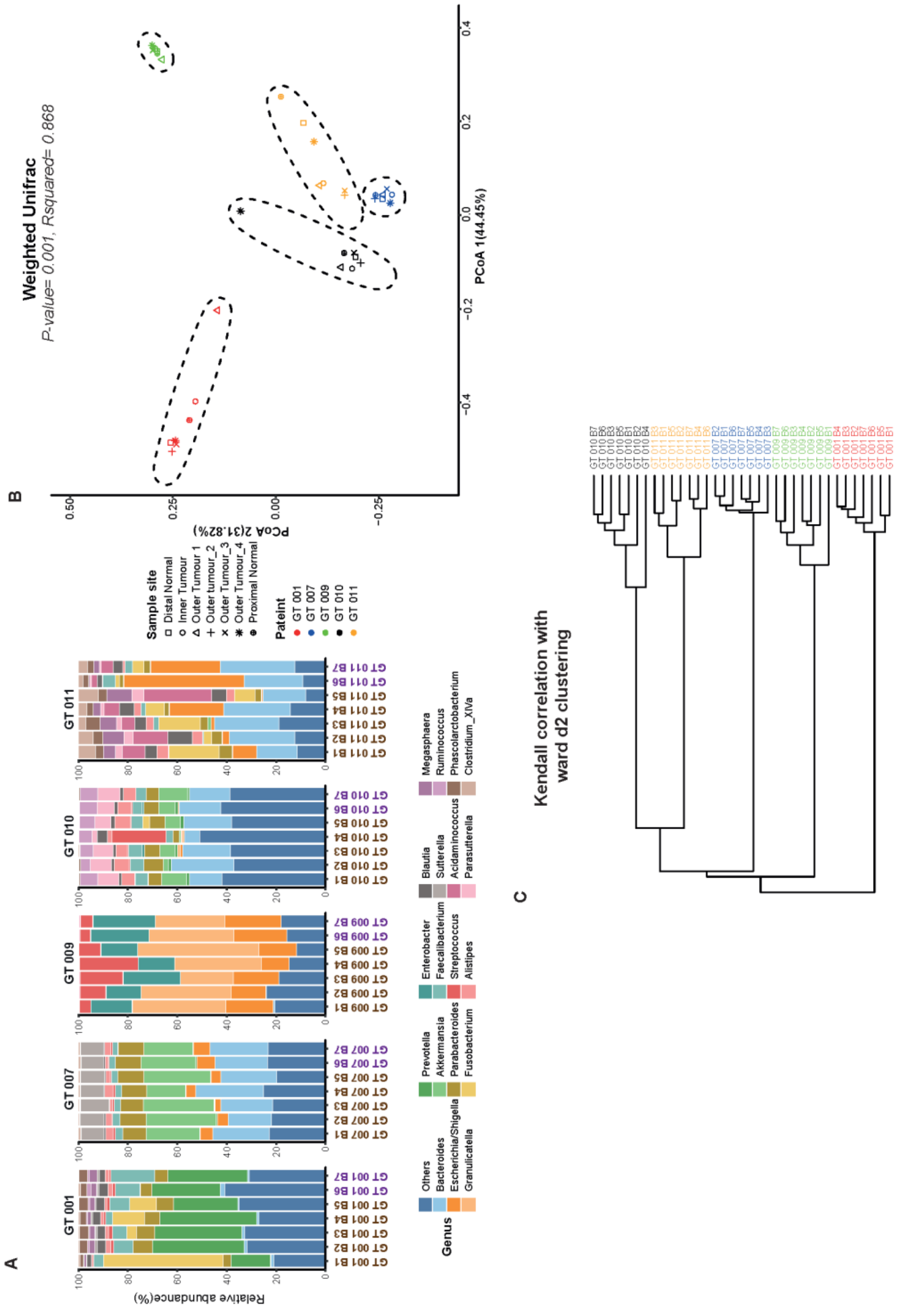


Figure 1. Microbiome relatedness of biopsies within Individuals. (A) Taxonomic bar plot of the proportional relative abundance of genera. “Others” is a grouping of genera with less than 1% abundancy across the samples as well as unclassified genera (B) PCoA plot representing weighted Unifrac distances. Biopsy location is represented by shapes while colours represent individual patients. Utilising the R package ggforce v0.3.1, ellipses were estimated using the Khachiyan algorithm. R-squared (R^2) and p-values were calculated using Permutational Multivariate Analysis of Variance (PERMANOVA) via the R package vegan v2.4-2. (C) Dendrogram representing Kendall correlation with ward d2 clustering. Samples are coloured by individual.

The identity of the patient from whom the biopsy was taken was associated with the top four PCoA axes which collectively explained >90% of variance (see Supplementary figure 1, Supplementary table 1). However, there was no association between any of the top ten PCoA axes, which collectively explained ~99% of the variance, and sample site (Supplementary table 2). We employed Permutational multivariate analysis of variance (PERMANOVA) to calculate the association between sample meta-data factors and the global microbiome structure as defined by the beta-diversity distance matrixes. A strong association between the biopsy patient origin and the microbiome was identified (Figure 1B, Supplementary table 3). However, we did not detect any statistically significant association between global microbiome structure and sample site (Supplementary table 4). We next performed a patient-specific rank sum normalization on all samples to reduce the impact of patient bias. We performed a PERMANOVA on this transformed data to test for a significant association between location and the beta diversity metrics. However, we did not find a significant association (Supplementary table 5).

The beta diversity clustering data were supported by hierarchical clustering in which the topology of the dendrogram was clearly dictated by the subject identity rather than biopsy site (Figure 1C). Within subjects, there was no reproducible pattern of microbiota relatedness by anatomical origin that was replicated across subjects (Figure 1C).

PCoA axis	P-value
1	0.0000017603
2	0.0000023873
3	0.0000266130
4	0.0000196660
5	0.1453059839
6	0.2208569369
7	0.9189331198
8	0.8767527693
9	0.7447672631
10	0.9402627020

Supplementary table 1. Association between PCoA axes and Patient ID. P value calculated using Kruskal–Wallis test

PCoA axis	P-value
1	0.9997487
2	0.9977087
3	0.9787245
4	0.9781814
5	0.2440231
6	0.1786758
7	0.5677194
8	0.4210597
9	0.3520217
10	0.1317221

Supplementary table 2. Association between PCoA axes and sample site. P value calculated using Kruskal–Wallis test

Beta-diversity metric	P-value	R squared
Weighted unifrac	0.001	0.868
Unweighted unifrac	0.001	0.715
Bray Curtis dissimilarity	0.001	0.852
Jaccard similarity	0.001	0.721

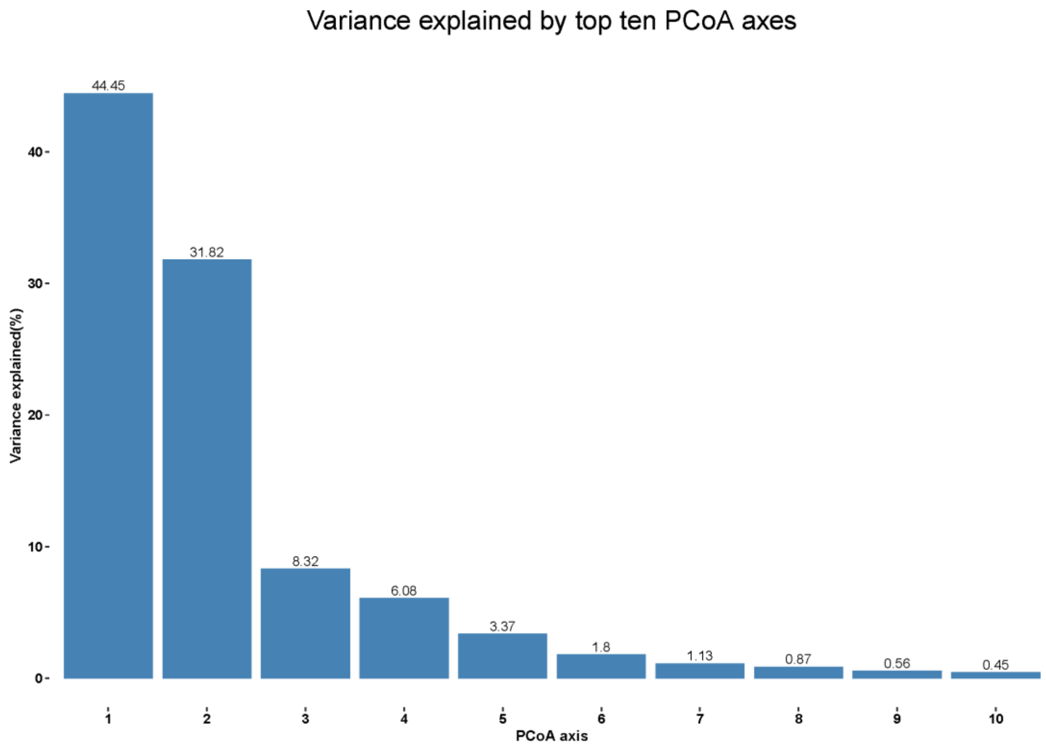
Supplementary table 3. Association between beta diversity metrics and Patient ID. P-value and R squared calculated using PERMANOVA.

Beta-diversity metric	P-value	R squared
Weighted unifrac	1	0.033
Unweighted unifrac	1	0.047
Bray Curtis dissimilarity	1	0.032
Jaccard similarity	1	0.058

Supplementary table 4. Association between beta diversity metrics and Patient ID. P-value and R squared calculated using PERMANOVA.

Beta-diversity metric	P-value	R squared
Bray Curtis dissimilarity	1	0.03151
Jaccard similarity	1	0.05768

Supplementary table 5. Association between beta diversity metrics and Patient ID with rank-sum normalization. P-value and R squared calculated using PERMANOVA.



Supplementary figure 1. Variance explained by first 10 PCoA axes. Bar plot displaying level of variance of variance explained by each access with regard to unweighted Unifrac distance.

Samples were pooled based on biopsy site and pairwise analysis was performed for each sample pair within the biopsy site. Differential ASV abundance was not detected with respect to anatomical site when we applied paired sample Wilcoxon test with Benjamini-Hochberg adjustment for multiple comparisons (Supplementary table 6). We next utilized DESeq2 which has been demonstrated to be sensitive when applied to small sample sizes[33,34]. We identified a number of differentially abundant ASVs between sample-sites while controlling for which patient the biopsy originated from (Figure 2). Notably, a number of ASVs assigned to the oral species *Fusobacterium nucleatum*, were observed to be enriched on tumor samples relative to undiseased disease (distal normal and proximal normal). In particular Seq 31 was identified to be enriched in 5/5 proximal tumor biopsies relative to the healthy distal biopsy and 4/5 tumor biopsies relative to the healthy distal biopsy.

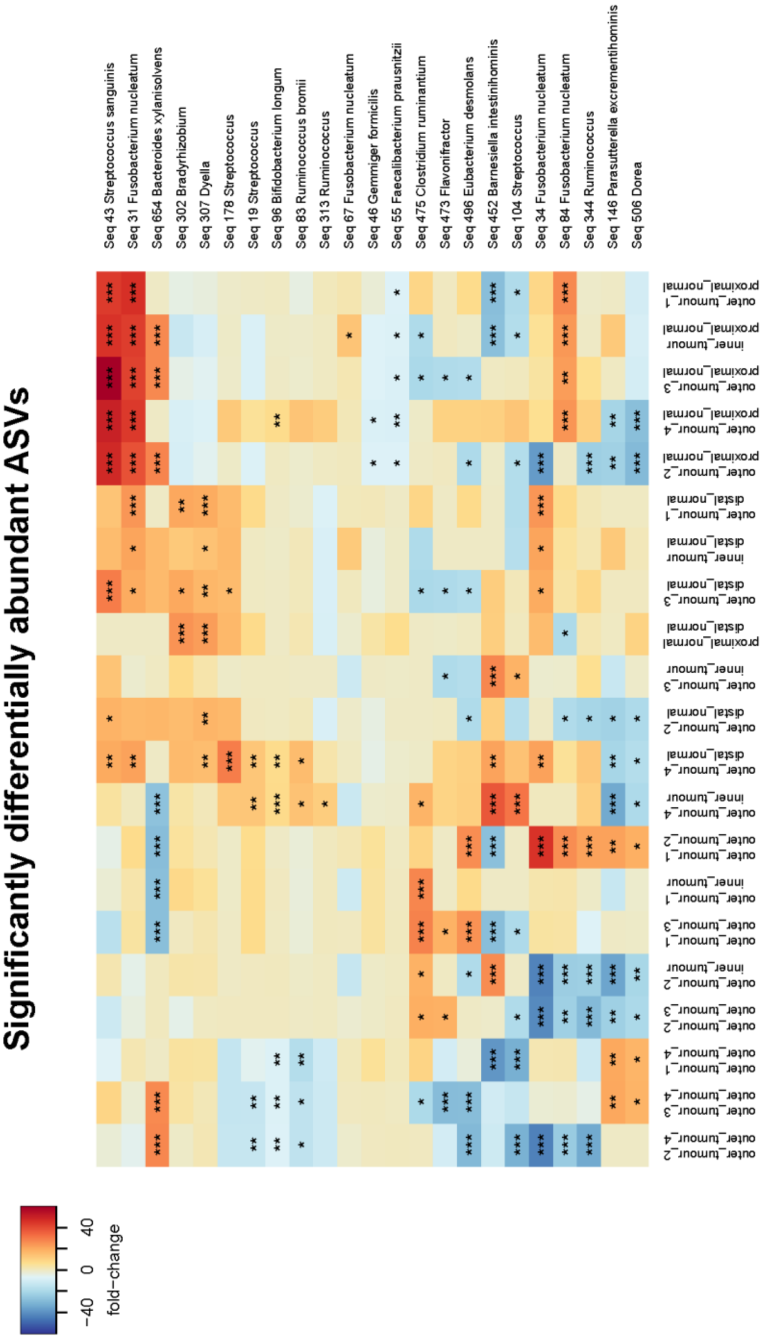


Figure 2. Differentially abundant ASVs. Heat plot displaying differentially ASVs between each pairwise comparison of every sample sit. Column names indicate which pairwise comparison. Row names display ASVs with which taxa it was assigned too. Only ASVs which could be assigned to the Genus level displayed. Stars indicate P-value. *<0.05, **<0.01 and ***<0.001

Previous studies have indicated that oral microbes can translocate from the oral cavity to the gut[35]. Furthermore, CRC tumor microbiota is enriched with oral taxa[20]. For these reasons, the buccal swab microbiota composition was analyzed and compared to that of the respective subjects' biopsy sites as a function of beta diversity distance (Figure 3A, 3B, Supplementary Figure 2). This analysis revealed that the microbiota of all the biopsies were equally distance from the oral microbiota in all the subjects.

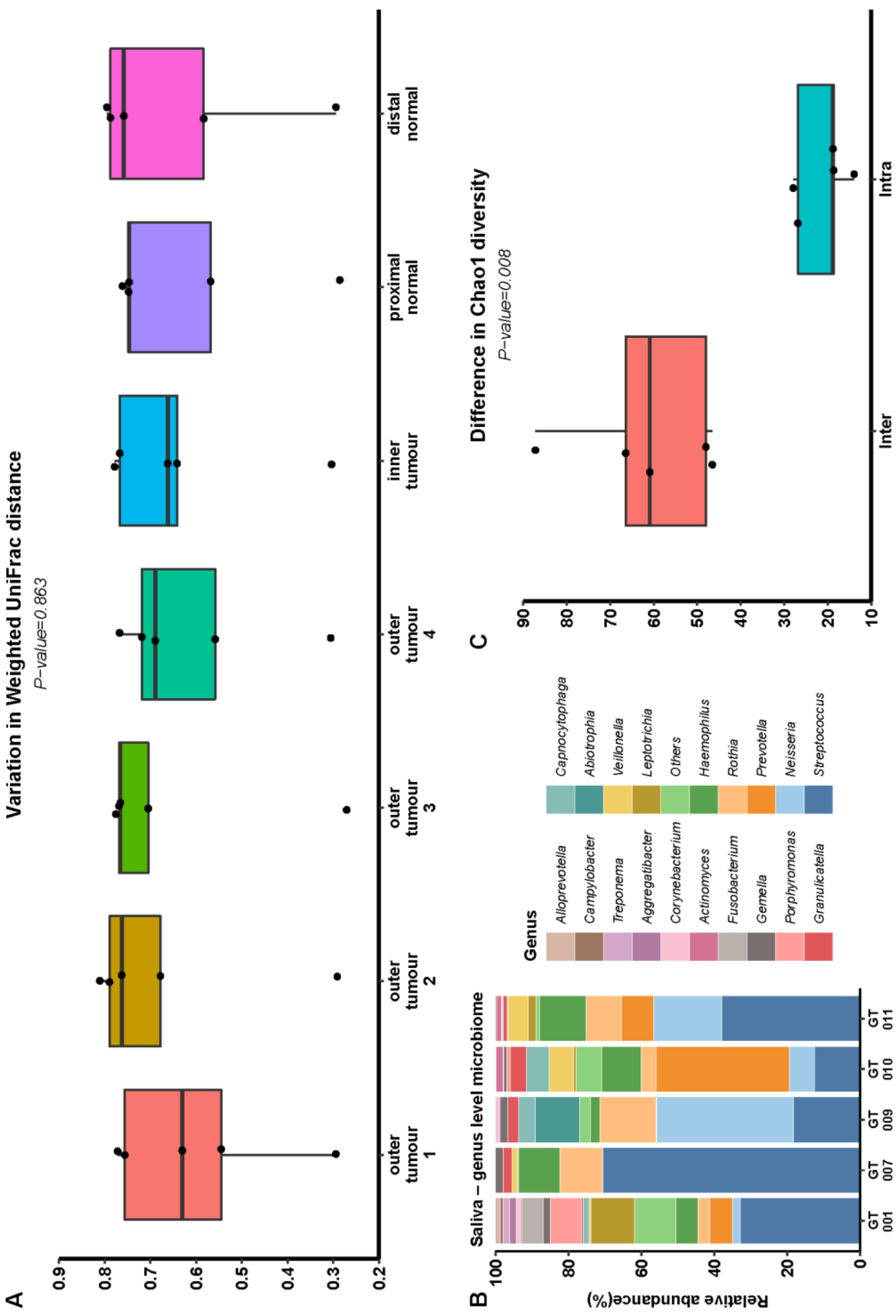
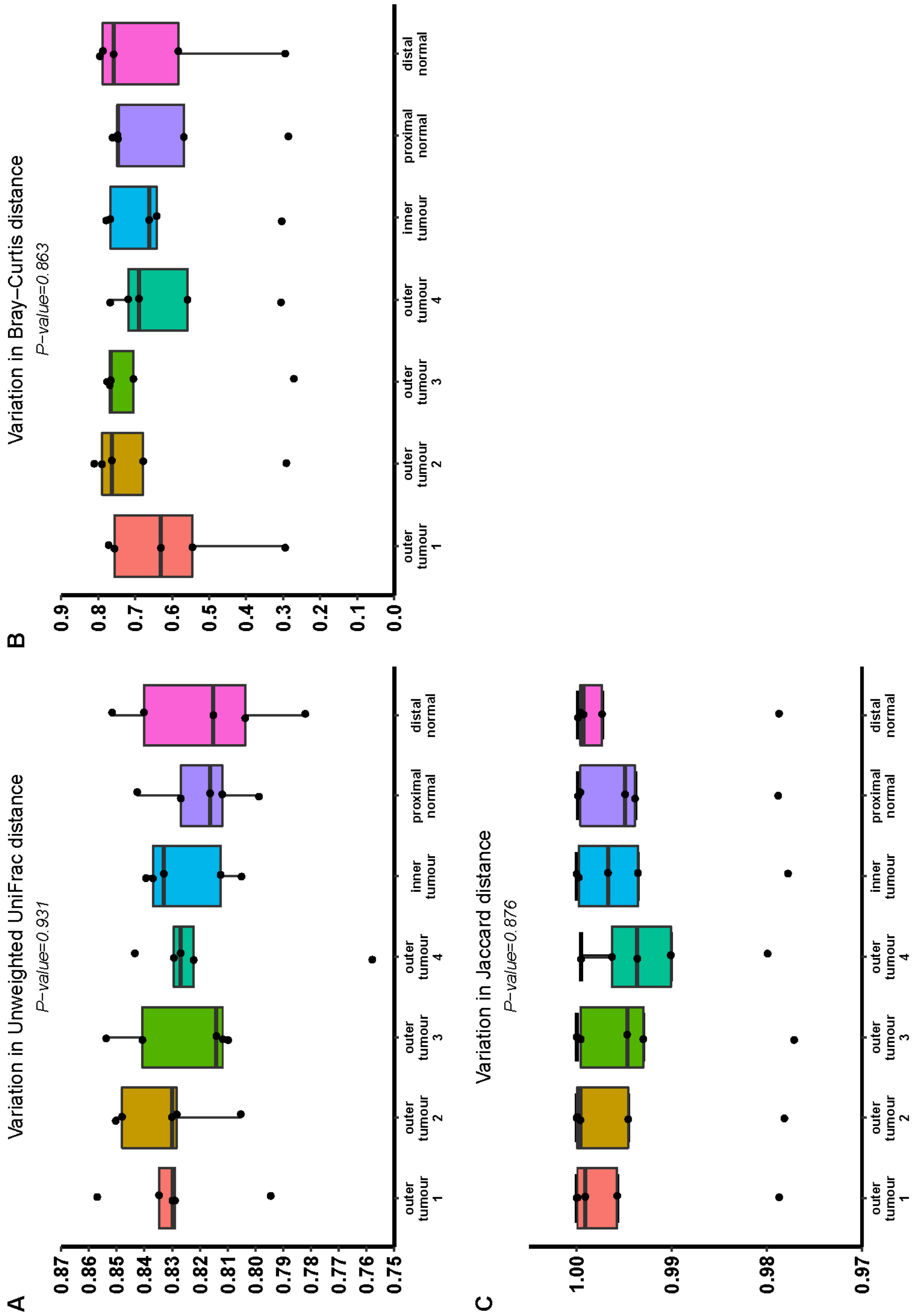
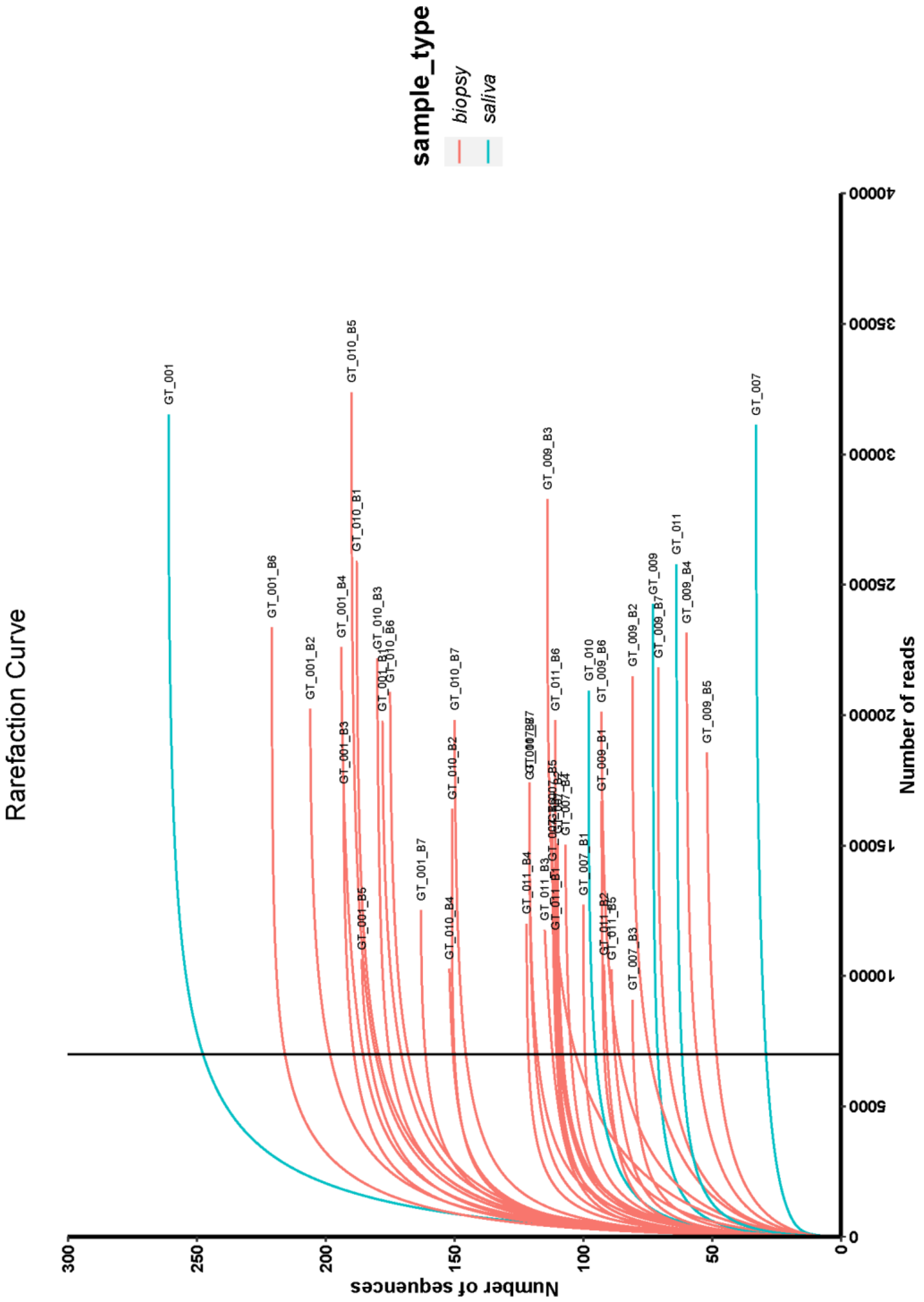


Figure 3. (A)Bar plot of the difference in Beta-diversity distance between the microbiota of indicated biopsy sites and paired buccal swab microbiota from the same subject. Kruskal–Wallis test was used to calculate p-values. (B) Taxonomic bar plot of the proportional relative abundance of genera of oral samples. “Others” is a grouping of genera with less than 0.25% abundancy across the samples as well as unclassified genera (C) Bar plot displaying the difference between Inter-individuals versus Intra-individual variation in alpha-diversity (Chao1).

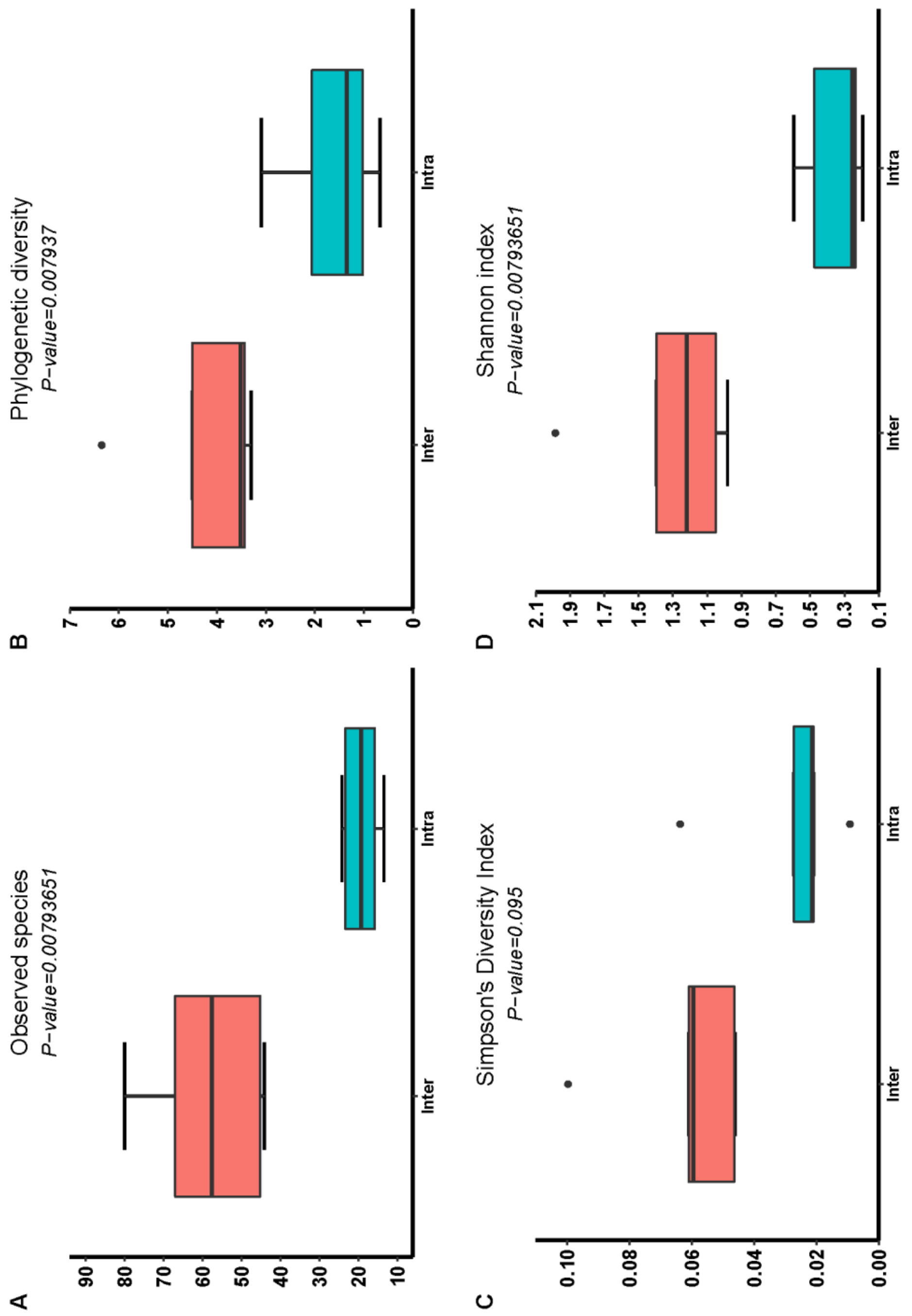


Supplementary figure 2. Bar plot of the difference in Beta-diversity distance between the microbiota of indicated biopsy sites and paired buccal swab microbiota from the same subject. (A) Unifrac distance (B) Bray–Curtis (C) Jaccard. Kruskal–Wallis test was used to calculate p-values

The sequencing depth of the samples allowed for a thorough investigation of alpha diversity, that is microbial richness and evenness (Supplementary table 7, Supplementary Figure 3). Considering all biopsies from each sample sites examined, the difference in alpha diversity of the biopsy microbiota datasets as measured by 5 different indices was significantly greater between any two individuals then it was within individuals (Figure 3C, Supplementary figure 4).



Supplementary figure 3. Rarefaction Curve. Number of reads on x-axis. Number of unique ASV sequences. Blue lines indicate saliva samples. Red lines indicate colonic biopsy samples.



Supplementary figure 4. Bar plot displaying the difference between Inter-individuals versus Intra-individual variation in alpha-diversity (A) Observed species (B) Phylogenetic diversity (C) Simpson's Diversity Index (D) Shannon index

3.4 Discussion

Many studies have profiled the microbiome in CRC using cancer tissue[4,14,30,31] from a single biopsy assuming that the microbiome profiled on this single specimen was representative of the tumor as a whole. This study confirms that this is a valid assumption.

Given the macroscopic and microscopic heterogeneity of CRC tumors, it may seem surprising that the microbiome of an individual tumor is very similar throughout the entire tumor tissue, as shown in this study. In contrast, significant differences were noted in the genus level abundance of particular taxa in the microbiota sequenced from biopsy samples from five individuals in the study. These variations are probably due to the differences of tumor location (Figure 1) as has been previously reported[4,30], as well as to other factors such as antibiotic exposure[36] and diet[37], which are known to alter the baseline microbiome.

Interestingly, as we showed in a previous study[4], paired samples of un-diseased tissue proximal and distal to the tumor harbored the same microbiota with respect to dominant taxa and their relative abundance. Previous work has demonstrated the presence of anaerobic oral bacteria on the colorectal tumor mucosa[20,31] consistent with the notion of a biofilm of pathologic bacteria forming[38] and seeding on the tumor. In the current study, various distance metrics did not show that any particular site was closer to the oral microbiome. However, we did detect specific oral-associated taxa such as *Fusobacterium nucleatum* and *Streptococcus sanguinis* overrepresented on tumor sample sites. Indeed, from the growing catalogue of microbes associated with CRC many of these microbes belong to oral-associated taxa including *Fusobacterium*, *Porphyromonas*, *Gemella*, *Streptococcus* and *Leptotrichia*[39]. Two routes of translocation of oral microbes to the colon have been proposed: 1) though the gastrointestinal tract and 2) through circulatory system[35,40]. Both *Fusobacterium nucleatum* and *Streptococcus sanguinis* have been observed to cause endocarditis demonstrating the potential to travel through the circulatory system[41,42]. *Fusobacterium nucleatum* is of particular note due to the growing body of evidence of its mechanistic role in the oncogenesis of CRC[42].

There are some limitations to this study. The sample size of five patients is small, but tumor tissue within each individual was extensively biopsied to capture macroscopically morphologically different areas such as ulcerated and non-ulcerated tissue. Four individuals were treated with antibiotics prior to or during the procedure as per hospital protocol. Similarly, all patients took a bowel preparation on the day prior to their surgery which is known to alter the microbiome[43]. However, in this study each individual was taken as a separate entity therefore acting as an internal control and comparator and it is assumed that these modifiers of the microbiome affected the microbiome as a whole.

The global burden of CRC is increasing and this disease is a significant contributor to cancer deaths[1]. Prospective trials are ongoing that incorporate microbiota analysis with other factors as part of the investigative assessment and staging of cancer[44] and to predict CRC outcomes[45]. Through demonstration of microbial homogeneity within an individual tumor and in the adjacent normal tissue, this study helps validate the methodology of sampling tissue going forward for these and other indications.

3.5 Patients and Methods/Materials and Methods

3.5.1 Patient recruitment

A total of five patients who were scheduled for colonic resection for colorectal cancer as part of their standard of care at Cork University Hospital and Mercy University Hospital, Cork were recruited to the study. Patients were labelled as GT (Geography of Tumor) 001,007,009,010 and 011. Recruitment to the study took place from February 2019 to June 2019. Ethical approval was granted by The Clinical Research Ethics Committee of the Cork Teaching Hospitals (Cork, Ireland). The study was conducted in accordance with the ethical principles set forth in the current version of the Declaration of Helsinki, the International Conference on Harmonization E6 Good Clinical Practice (ICH-GCP). Exclusion criteria included a history of inflammatory bowel disease or irritable bowel syndrome, a significant acute or chronic coexisting illness and neoadjuvant chemotherapy or radiotherapy. All patients received a macrogol preparation pre-operatively. A single dose of oral metronidazole and neomycin were administered to two patients pre-operatively and two other patients received intraoperative intravenous co-amoxiclav and metronidazole as per hospital protocol. The fifth patient took no antibiotics. None of the patients had probiotic exposure pre-operatively.

A mouth swab was taken from patients in the pre-operative room prior to anesthetic and snap frozen. Immediately after removal from the patient, the ex-vivo specimen was anatomically orientated, was dissected and the tumor was exposed. A representative tissue biopsy from each of the four quadrants of the tumor was taken in a clockwise manor starting at 12 o’clock. Tissue from a central area of tumor plus two biopsies of adjacent macroscopically normal tissue 10 cm proximal and distal to the tumor were taken. A different set of sterile instruments was used for every biopsy taken and for each individual. This ensured there was no transfer of bacterial material from sample to sample within or between individuals. Samples were snap frozen in cryotubes and transferred immediately for storage at -80° C.

3.5.2 DNA extraction and 16S RNA amplicon sequencing

Genomic DNA from biopsies was extracted using the AllPrep DNA kit from Qiagen. When preparing each sample, approximately 20mg in total of tissue was dissected in small fragments from around the biopsy and pooled. These pooled fragments were then added to a bead beating tube containing sterile beads and 600 µl of buffer RLT plus was added. Samples were then homogenized for two 15 sec at full speed pulses in a MagnaLyzer (Roche, Penzberg, Germany) with rests on ice between pulses. The rest of the DNA extraction was carried out according to the Qiagen AllPrep DNA/RNA extraction kit. Oral genomic DNA was extracted using Qiagen DNeasy PowerSoil Kit following the manufacturer’s instruction.

3.5.3 Library preparation and sequencing

The 16S rRNA gene was amplified using primers for the V3-V4 region; forward, TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCTACGGGNGGCWGCAG-3′ and reverse, 5′-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGACTACHVGGGTATCTAATCC-3'. DNA was normalized to a concentration of 10ng/µl and 10 µl DNA was added per 30 µl PCR reaction. The PCR thermocycler protocol was as follows:

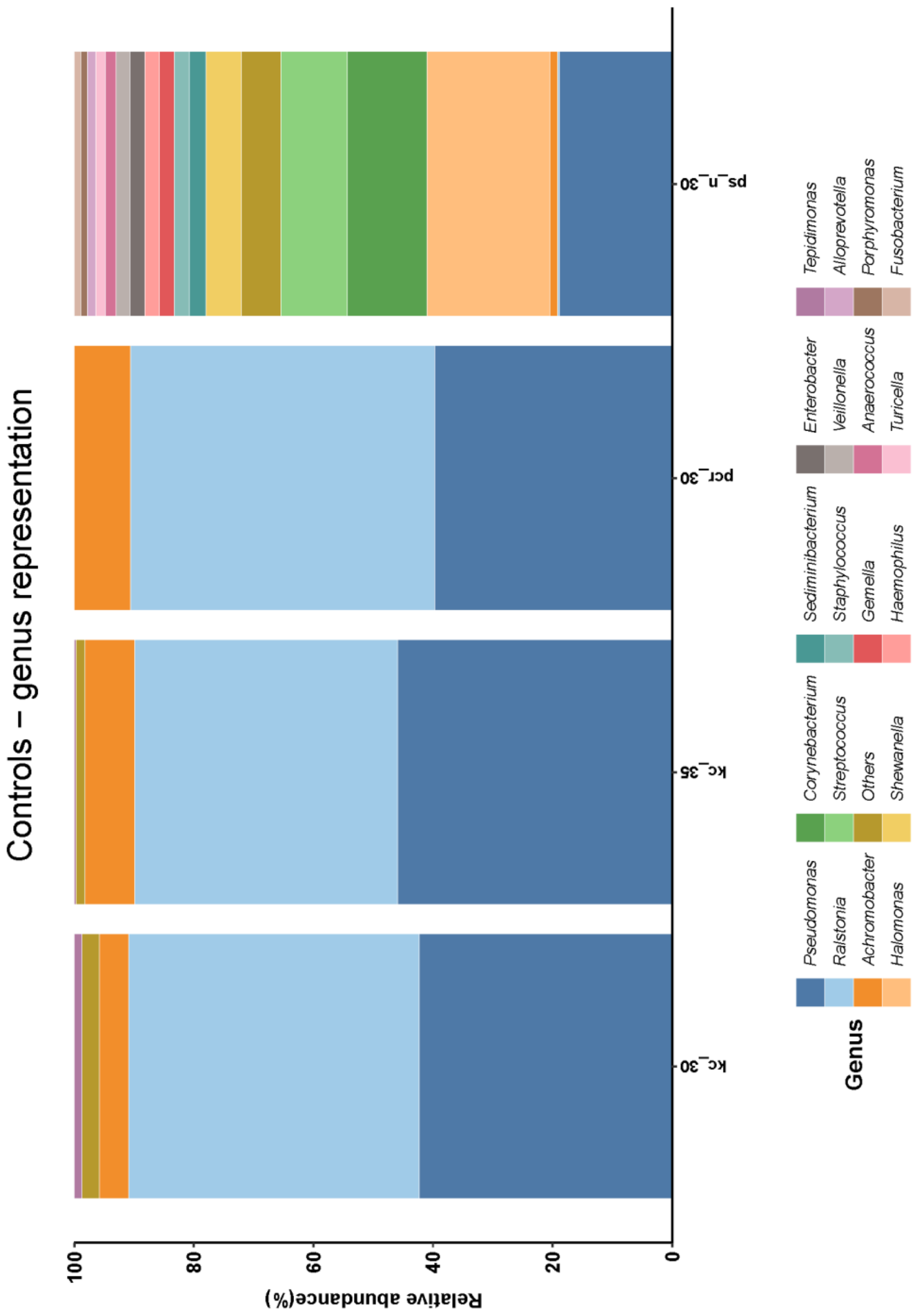
Initiation step of 98 °C for 3 min followed by 30 cycles of 98 °C for 30 s, 55 °C for 60 s, and 72 °C for 20 s, and a final extension step of 72 °C for 5 min. Indexes were subsequently added to the purified amplicons according to Illumina 16S Metagenomic Sequencing Protocol (Illumina, CA, USA). Libraries DNA concentration was quantified using a Qubit fluorometer (Invitrogen) using the 'High Sensitivity' assay and samples were pooled at a standardized concentration (80 ng of each sample). The pooled library was sequenced at Eurofins Genomics/GATC Biotech (Konstanz, Germany) on the Illumina MiSeq platform using 2×300 bp chemistry. All samples in this study were prepared in the same library and sequenced together.

3.5.4 Bioinformatics analyses

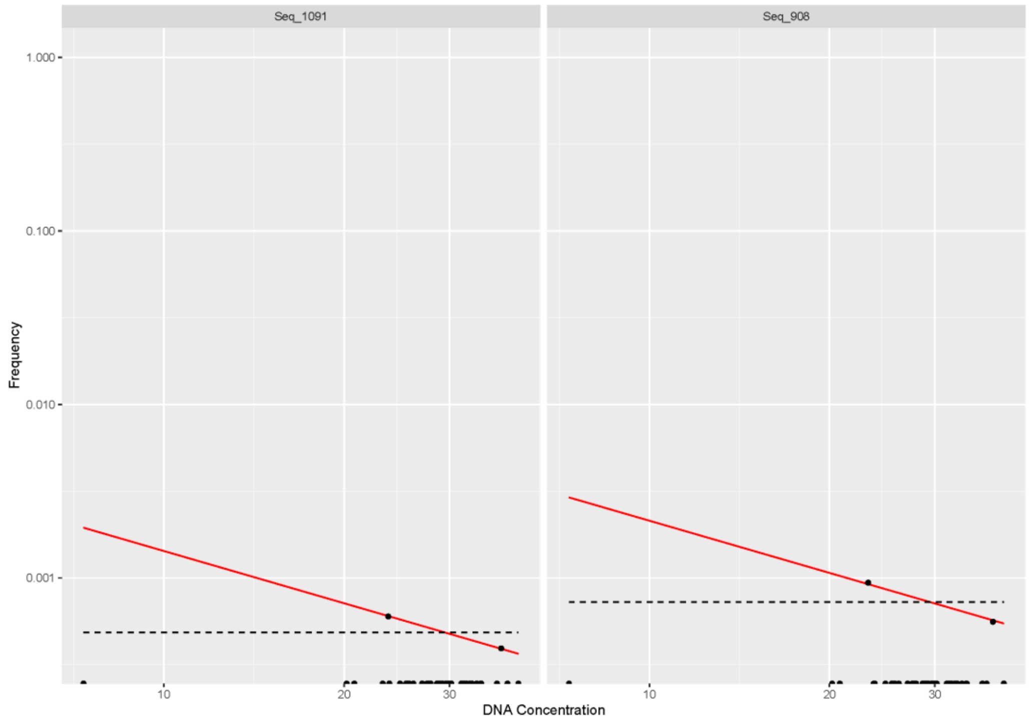
Raw data was imported into R v3.5.3 for processing and analysis. Paired reads were quality filtered, trimmed, merged and Amplicon Sequence Variants (ASV) inferred using the R package dada2 v1.12.1. The following parameters were used for the filterAndTrim function; filtRs,trimLeft=c(19,21) ,maxEE=c(2,2), truncLen=c(260,230). Taxonomic classification was performed using the RDP naive Bayesian Classifier within the dada2 against the Silva v132 database. Alpha diversity was calculated from the ASV table using QIIME v1.9.1 as previously described in Kuczynski et al[46]. Samples were rarefied to 7000 reads in order to calculate alpha-diversity. QIIME v1.9.1 and the R package vegan v2.5.6 were used to infer β-diversity metrics[47]. β-diversity was visualized via principal coordinates analysis (PCoA) plots whose coordinates were identified using with the Ape package v5.1. The adonis() function within the R package vegan (v2.4-2) was used to perform Permutational multivariate analysis of variance (PERMANOVA) Difference in paired biopsy-buccal distance was assessed using paired Wilcoxon test. DESeq2 (v1.28.1) was used to identify differentially abundant taxa from the microbiota dataset.[33] Differences between inter and intra alpha-diversity was tested using Wilcoxon signed-rank test.

3.5.6 Contamination control

We first carried out mock extractions to detect reagent-associated contamination from the two kits used in this study (Supplementary figure 5). Further, we also carried out PCR controls i.e. water, to detect contamination specific to the polymerase (Supplementary figure 5). These negative controls underwent 5-10 additional PCR cycles relative to biological specimens to capture low levels of bacterial template. We utilized both the Frequency and Prevalence method within the R package decontam (v 1.8.0) to identify contaminating ASVs[48]. Using the "frequency" method, isContaminant(phyloseq_object, method="frequency", conc="qubit",threshold = 0.05), two ASVs were identified (Supplementary figure 6). However, these ASVs were present at a very low abundance and only present in 2 samples. Furthermore, these ASVs were assigned to Clostridiales and Burkholderiales which are known gut taxa and not indicative of contamination (Supplementary table 8). Using the "prevalence" method, isContaminant(phyloseq_object, method="prevalence", neg="is.neg",threshold=0.05), we identified 7 contaminating ASVs (Supplementary table 9). However, these ASVs were only identified in three of our samples and only contributed between 2-6 reads to the samples. Thus, we treated them as negligibly.



Supplementary figure 5. Taxonomic bar plot of the proportional relative abundance of genera within controls samples. KC-30 denotes AllPrep DNA kit mock extraction followed by 30 cycle 16s gene PCR amplification. KC-35 denotes AllPrep DNA kit mock extraction followed by 35 cycle 16s gene PCR amplification. pcr-30 denotes mock amplfcation (just water) of the 16s gene. ps-n-30 denotes DNeasy PowerSoil Kit mock extraction followed by 30 cycle 16s gene PCR amplification.



Supplementary figure 6. Decontam frequency graph. X axis equals concentration of sample before normalization. Y-axis equals frequency of ASV. Each dot represents a sample.

ASV	Order	Genus	Species
Seq 908	Burkholderiales	Sutterella	Sutterella stercoricanis
Seq 1091	Clostridiales	unclassified	unclassified

Supplementary table 8. ASV identified as contamination using the "frequency" method within decontam.

ASV	Genus	Species
Seq_4	Ralstonia	Ralstonia insidiosa
Seq_6	Pseudomonas	unclassified
Seq_8	Pseudomonas	unclassified
Seq_30	Achromobacter	unclassified
Seq_243	Tepidimonas	unclassified
Seq_311	Pseudomonas	unclassified
Seq_626	Propionibacterium	Propionibacterium acnes

Supplementary table 9 ASV identified as contamination using the "prevalence" method within decontam.

3.6 Acknowledgements

Work in PWOTs laboratory is supported by Science Foundation Ireland through a Centre award (APC/SFI/12/RC/2273_P2) to APC Microbiome Ireland, and by The Health Research Board of Ireland under Grant ILP-POR-2017-034. We thank the patients who consented to participate in this study, and the support staff of relevant departments in Cork University Hospital and the Mercy University Hospital Cork.

3.7 Reference

1 Bibbins-Domingo, K. *et al.* Screening for Colorectal Cancer: US Preventive Services Task Force Recommendation Statement. *JAMA* **315**, 2564-2575, doi:10.1001/jama.2016.5989 (2016).

2 Fearon, E. R. & Vogelstein, B. A genetic model for colorectal tumorigenesis. *Cell* **61**, 759-767 (1990).

3 Stryker, S. J. *et al.* Natural history of untreated colonic polyps. *Gastroenterology* **93**, 1009-1013 (1987).

4 Flemer, B. *et al.* Tumour-associated and non-tumour-associated microbiota in colorectal cancer. *Gut* **66**, 633-643, doi:10.1136/gutjnl-2015-309595 (2017).

5 Thomas, A. M. *et al.* Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. *Nat Med* **25**, 667-678, doi:10.1038/s41591-019-0405-7 (2019).

6 Raskov, H., Burcharth, J. & Pommergaard, H. C. Linking Gut Microbiota to Colorectal Cancer. *J Cancer* **8**, 3378-3395, doi:10.7150/jca.20497 (2017).

7 Sommer, F. & Bäckhed, F. The gut microbiota--masters of host development and physiology. *Nat Rev Microbiol* **11**, 227-238, doi:10.1038/nrmicro2974 (2013).

8 Marshall, B. J. & Warren, J. R. Unidentified curved bacilli in the stomach of patients with gastritis and peptic ulceration. *Lancet* **1**, 1311-1315 (1984).

9 Wang, F., Meng, W., Wang, B. & Qiao, L. Helicobacter pylori-induced gastric inflammation and gastric cancer. *Cancer Lett* **345**, 196-202, doi:10.1016/j.canlet.2013.08.016 (2014).

10 Bullman, S. *et al.* Analysis of. *Science* **358**, 1443-1448, doi:10.1126/science.aal5240 (2017).

11 Pleguezuelos-Manzano, C. *et al.* Mutational signature in colorectal cancer caused by genotoxic pks. *Nature*, doi:10.1038/s41586-020-2080-8 (2020).

12 Louis, P., Hold, G. L. & Flint, H. J. The gut microbiota, bacterial metabolites and colorectal cancer. *Nat Rev Microbiol* **12**, 661-672, doi:10.1038/nrmicro3344 (2014).

13 Peters, B. A. *et al.* The gut microbiota in conventional and serrated precursors of colorectal cancer. *Microbiome* **4**, 69, doi:10.1186/s40168-016-0218-6 (2016).

14 Hibberd, A. A. *et al.* Intestinal microbiota is altered in patients with colon cancer and modified by probiotic intervention. *BMJ Open Gastroenterol* **4**, e000145, doi:10.1136/bmjgast-2017-000145 (2017).

15 Lu, Y. *et al.* Mucosal adherent bacterial dysbiosis in patients with colorectal adenomas. *Sci Rep* **6**, 26337, doi:10.1038/srep26337 (2016).

16 Weir, T. L. *et al.* Stool microbiome and metabolome differences between colorectal cancer patients and healthy adults. *PLoS One* **8**, e70803, doi:10.1371/journal.pone.0070803 (2013).

17 Ahn, J. *et al.* Human gut microbiome and risk for colorectal cancer. *J Natl Cancer Inst* **105**, 1907-1911, doi:10.1093/jnci/djt300 (2013).

18 Nakatsu, G. *et al.* Gut mucosal microbiome across stages of colorectal carcinogenesis. *Nat Commun* **6**, 8727, doi:10.1038/ncomms9727 (2015).

19 Drewes, J. L. *et al.* High-resolution bacterial 16S rRNA gene profile meta-analysis and biofilm status reveal common colorectal cancer consortia. *NPJ Biofilms Microbiomes* **3**, 34, doi:10.1038/s41522-017-0040-3 (2017).

20 Flemer, B. *et al.* The oral microbiota in colorectal cancer is distinctive and predictive. *Gut*, doi:10.1136/gutjnl-2017-314814 (2017).

21 Yang, Y. *et al.* Prospective Study of Oral Microbiome and Colorectal Cancer Risk in Low-income and African American Populations. *Int J Cancer*, doi:10.1002/ijc.31941 (2018).

22 Punt, C. J., Koopman, M. & Vermeulen, L. From tumour heterogeneity to advances in precision treatment of colorectal cancer. *Nat Rev Clin Oncol* **14**, 235-246, doi:10.1038/nrclinonc.2016.171 (2017).

23 Hewett, D. G. *et al.* Validation of a simple classification system for endoscopic diagnosis of small colorectal polyps using narrow-band imaging. *Gastroenterology* **143**, 599-607.e591, doi:10.1053/j.gastro.2012.05.006 (2012).

24 The Paris endoscopic classification of superficial neoplastic lesions: esophagus, stomach, and colon: November 30 to December 1, 2002. *Gastrointest Endosc* **58**, S3-43 (2003).

25 Kudo, S. *et al.* Diagnosis of colorectal tumorous lesions by magnifying endoscopy. *Gastrointest Endosc* **44**, 8-14 (1996).

26 Kudo, S. *et al.* Colorectal tumours and pit pattern. *J Clin Pathol* **47**, 880-885 (1994).

27 S.R., H. & L.A., A. *World Health Organization Classification of Tumours. Pathology and Genetics of Tumours of the Digestive System. .* 108 (IARC Press, 2000).

28 Konda, K. *et al.* Distinct molecular features of different macroscopic subtypes of colorectal neoplasms. *PLoS One* **9**, e103822, doi:10.1371/journal.pone.0103822 (2014).

29 Jones, H. G. *et al.* Genetic and Epigenetic Intra-tumour Heterogeneity in Colorectal Cancer. *World J Surg* **41**, 1375-1383, doi:10.1007/s00268-016-3860-z (2017).

30 Wu, Y. *et al.* Microbiota Diversity in Human Colorectal Cancer Tissues Is Associated with Clinicopathological Features. *Nutr Cancer* **71**, 214-222, doi:10.1080/01635581.2019.1578394 (2019).

31 Warren, R. L. *et al.* Co-occurrence of anaerobic bacteria in colorectal carcinomas. *Microbiome* **1**, 16, doi:10.1186/2049-2618-1-16 (2013).

32 Harada, T. *et al.* Surface microstructures are associated with mutational intratumoral heterogeneity in colorectal tumors. *J Gastroenterol*, doi:10.1007/s00535-018-1481-z (2018).

33 Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15**, 550, doi:10.1186/s13059-014-0550-8 (2014).

34 Weiss, S. *et al.* Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome* **5**, 27, doi:10.1186/s40168-017-0237-y (2017).

35 Schmidt, T. S. *et al.* Extensive transmission of microbes along the gastrointestinal tract. *Elife* **8**, doi:10.7554/eLife.42693 (2019).

36 Modi, S. R., Collins, J. J. & Relman, D. A. Antibiotics and the gut microbiota. *J Clin Invest* **124**, 4212-4218, doi:10.1172/JCI72333 (2014).

37 Ghosh, T. S. *et al.* Mediterranean diet intervention alters the gut microbiome in older people reducing frailty and improving health status: the NU-AGE 1-year dietary intervention across five European countries. *Gut*, doi:10.1136/gutjnl-2019-319654 (2020).

38 Tomkovich, S. *et al.* Human colon mucosal biofilms from healthy or colon cancer hosts are carcinogenic. *J Clin Invest* **130**, 1699-1712, doi:10.1172/JCI124196 (2019).

39 Ternes, D. *et al.* Microbiome in Colorectal Cancer: How to Get from Meta-omics to Mechanism? *Trends Microbiol* **28**, 401-423, doi:10.1016/j.tim.2020.01.001 (2020).

40 Abed, J. *et al.* Fap2 Mediates Fusobacterium nucleatum Colorectal Adenocarcinoma Enrichment by Binding to Tumor-Expressed Gal-GalNAc. *Cell Host Microbe* **20**, 215-225, doi:10.1016/j.chom.2016.07.006 (2016).

41 Di Filippo, S. *et al.* Current patterns of infective endocarditis in congenital heart disease. *Heart* **92**, 1490-1495, doi:10.1136/hrt.2005.085332 (2006).

42 Brennan, C. A. & Garrett, W. S. Fusobacterium nucleatum - symbiont, opportunist and oncobacterium. *Nature reviews. Microbiology* **17**, 156-166, doi:10.1038/s41579-018-0129-6 (2019).

43 Jalanka, J. *et al.* Effects of bowel cleansing on the intestinal microbiota. *Gut* **64**, 1562-1568, doi:10.1136/gutjnl-2014-307240 (2015).

44 Murphy, C. L., O’Toole, P. W. & Shanahan, F. The Gut Microbiota in Causation, Detection, and Treatment of Cancer. *Am J Gastroenterol*, doi:10.14309/ajg.0000000000000075 (2019).

45 Veziant, J. *et al.* Prognostic value of a combination of innovative factors (gut microbiota, sarcopenia, obesity, metabolic syndrome) to predict surgical/oncologic outcomes following surgery for sporadic colorectal

cancer: a prospective cohort study protocol (METABIOTE). *BMJ Open* **10**, e031472, doi:10.1136/bmjopen-2019-031472 (2020).

46 Kuczynski, J. *et al.* Using QIIME to analyze 16S rRNA gene sequences from microbial communities. *Curr Protoc Microbiol* **Chapter 1**, Unit 1E.5., doi:10.1002/9780471729259.mc01e05s27 (2012).

47 Caporaso, J. G. *et al.* QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* **7**, 335-336, doi:10.1038/nmeth.f.303 (2010).

48 Davis, N. M., Proctor, D. M., Holmes, S. P., Relman, D. A. & Callahan, B. J. Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. *Microbiome* **6**, 226, doi:10.1186/s40168-018-0605-2 (2018).

Chapter 4 - Association between the microbiome and treatment outcomes in patients with metastatic melanoma treated with Immunotherapy

This chapter is currently under review in the journal *British Journal of Cancer*

Authors:

Clodagh L Murphy*, Maurice Barrett*, Paola Pellanda, Fergus Shanahan, Derek G Power, Paul W O'Toole

*Joint first authorship: These authors contributed equally to this work.

Maurice Barrett contributed to this work as follows:

- All bioinformatic analysis including sequence processing, compositional data analysis and statistical analysis.
- Data visualization, that is, the construction of manuscript figures.
- Writing of half of the manuscript.

4.1 Abstract

Background

The development of immune checkpoint inhibitors has contributed significantly to cancer therapeutics. However, treatment efficacy is limited by both non-responsiveness and side effects in certain patients. Mounting evidence indicates that the gut microbiome modulates both treatment response and immune-mediated side effects, but no single microbiome feature or universal signature has been linked to these clinical outcomes. Since ethnic and geographic factors influence microbiome variance, we studied treatment outcomes as a function of microbiome composition in a cohort of caucasian Irish patients with melanoma undergoing treatment with checkpoint inhibitors.

Methods

We recruited 37 patients with metastatic melanoma, 21 commencing on immunotherapy *de novo* and 16 who were already established on treatment. Furthermore, we recruited 30 healthy controls to provide a reference microbiome. We profiled their faecal microbiome by 16S rRNA gene amplicon sequencing.

Results

We did not observe any significant difference in alpha or beta diversity with respect to response or side effects. We identified 15 sequence-based bacterial taxa that were differentially abundant between responders and non-responders. Consistent with previous work, the taxa showing higher relative abundance in responders included *Akkermansia muciniphila* and *Bifidobacterium longum*. Further, we identified previously unreported taxa associated with response including *Barnesiella intestinihominis* and *Clostridium disporicum*. *Faecalibacterium prausnitzii* was found to be associated with non-response, contradicting previous findings. We identified nine differentially abundant sequence-based bacterial taxa pertaining to side-effects including *Oscillibacter* which is negatively associated with inflammation. Using bioinformatic prediction of bacterial pathways, we identified a number of differentially abundant proteins (in the form of KEGG Orthologues) between response groups and side effect groups. These included proteins involved in

exopolysaccharide biogenesis that were enriched in both responders and individuals with no side effects.

Conclusions

Significant differences in microbial features were associated with both treatment response and protection against moderate and severe side effects in patients with stage four metastatic melanoma. Identification of these microbiome features can point to biomarkers to stratify cancer patients, inform microbial based therapeutics and provide insight into the basic biology of immune checkpoint inhibitors.

4.2 Background

Harnessing the immune system to destroy cancer cells has revolutionized cancer treatment [1]. Certain cancers develop immune resistance by upregulating immune checkpoint molecules such as PD-1 ligand (PD-L1) on the cancer cell, and its ligation to PD-1 on antigen-specific CD8(+) T cells[2]. Prolonged antigen exposure from cancer tissue can also cause exhaustion of T cells leading to decreased proliferation and release of cytokines[3]. These mechanisms inhibit apoptosis of the tumour cell and promote peripheral T effector cell ineffectiveness[4]. CTLA-4 (Cytotoxic T lymphocyte-associated molecule-4) is a cell surface molecule expressed on CD4^{+} and CD8^{+} T cells[5] which halts potentially autoreactive T cell activation at the naïve stage[6]. Checkpoint Inhibitors are monoclonal antibodies that inhibit these pathways to reactivate T cells, enhancing adaptive immune cell function allowing response to tumor antigens[7]. Ipilimumab is representative of a growing panel of such antibodies, and is directed against human CTLA-4[8]. Pembrolizumab and nivolumab are PD-1-blocking monoclonal antibodies used in metastatic melanoma and other malignancies[9]. Atezolizumab and avelumab are PD-L1-targeted immunotherapies for lung cancer, hepatocellular carcinoma, urothelial cancer, and merkel cell carcinoma[9].

Unprecedented overall survivals (OS) have been reported with immunotherapy in historically ‘difficult to treat’ cancers, e.g., 52% 5-years OS in metastatic melanoma[10],34% 3-year OS in advanced non-small cell lung cancer[11]. While much research is focused on biomarkers for treatment efficacy and methods to increase drug potency[1], the microbiome has emerged as an important modifier of the efficacy of immunotherapy[12]. No uniformly diagnostic species correlating with treatment success has been reported to date[13]. In murine models, the germ-free state significantly decreases the efficacy of certain immunotherapy in cancer models [14] [15]. Similarly, antibiotics not only interfere with immunotherapy efficacy[16] but also decrease overall and progression-free survival[17]. Gut microbiome abundance of *Ruminococcus* and *Alistipes* was found to enhance response to CpG-oligonucleotide treated mice, with a Lactobacillus predominant microbiota impairing response[15]. In murine studies, the efficacy of CTLA-4 blockade was linked to T cell responses specific for Bacteroides (*B. thetaiotaomicron* or *B. fragilis*)[18]. In patients treated with anti-PD-1 or PDL-1 immunotherapy, higher abundance of certain microbes was associated with treatment success. *Faecalibacterium prausnitzii*[19], *Akkermansia muciniphilia*[17] and *Bifidobacterium longum*[20] were associated with treatment responders in three separate studies. When microbiome composition data were reanalyzed using the same methodology there was a statistically significant difference in beta diversity in responders in two out of the three cohorts[21].

Immune system stimulation can lead to inflammatory side effects in patients receiving immunotherapy[22]. The exact mechanism for this immune toxicity is emerging. Macrophage-mediated toxicity, baseline low-level self-reactive T cells production of antibodies by activated B cells [23] , and cytotoxic T cells[24] are postulated to be involved. As with many autoimmune disorders, some patients may have a genetic predisposition to development of drug-related side effects[25]. Whether the development of immune-mediated side effects with use of immune checkpoint inhibitors correlates with improved antitumor immunity due to greater immunologic activation is unclear [22].

Serious or life threatening adverse side effects (CTCAE grade II-IV, Common terminology Criteria for Adverse Events v5.0 [31])have been reported to occur in up to 30% of patients on CTLA-4 and 16% of patients on PD-1 immunotherapy, and in up to 55% of patients receiving combined treatment with ipilimumab and nivolumab [26]. Common toxicities affect the endocrine[27], dermatologic, gastrointestinal, musculoskeletal, dermatological and neurological systems [26]. Many side effects are self-limiting but fatalities attributable specifically to drug toxicity rather than the underlying malignancy have been reported[28]. Early diagnosis of immune checkpoint inhibitor toxicity with investigations such as endoscopy and CT scan and subsequent early treatment appear to be beneficial[29]. Immunosuppression with glucocorticoids or other agents is occasionally required[30]. Immune-mediated side effects may occur at any time during treatment. However, cumulative exposure to immunotherapy does not appear to increase risk of development of side effects[31].. Long term immunological consequences are unknown[22].

Colitis is one of the most common immune-mediated side effect leading to discontinuation of treatment in 3-25% of patients[26]. A combination of lack of regulatory T cell depletion and accumulation of cytotoxic and proliferative CD8 T cells contribute to immune-mediated colitis [24]. The microbiome may also be involved because a microbiome-dependent subclinical colitis can be induced in specific pathogen free mice or germfree mice treated with CTLA-4 Ab [18]. Similarly, histopathological signs of colitis-induced by CTLA-4 blockade could also be reduced by introduction of *B. fragilis* and *Burkholderia cepacia* in antibiotic treated mice [18].

The present study profiles the gut microbiome of a cohort of patients from a single large tertiary referral cancer centre with stage 4 melanoma that were treated with immune checkpoint inhibitor therapy. The results show that the abundance of specific microbial species is linked with response to treatment and development of side effects.

4.3 Methods

4.3.1 Recruitment

Patients with metastatic melanoma, aged over 18 years, commencing (n=21) or established (n=16) on immune checkpoint inhibitor treatment in Cork University Hospital Cancer Centre, Cork, Ireland were recruited to this study. The study was conducted in accordance with the ethical principles set forth in the current version of the Declaration of Helsinki, the International Conference on Harmonization E6 Good Clinical Practice (ICH-GCP). Ethical approval was granted by The Clinical Research Ethics Committee of the Cork Teaching Hospitals (Cork, Ireland). The study was conducted from October 2017 to January 2019.

Forty one patients with metastatic melanoma receiving immune checkpoint inhibitors were identified at weekly multidisciplinary team meetings (MDT) and subsequently recruited through the oncology outpatient clinics. After giving informed consent, patients were given a sealed, sterile pack for faecal collection as well as a detailed patient-adapted standard operating procedure for safe collection of samples. A baseline pre-treatment faecal sample was collected from each patient commencing on therapy. Patients who were already on immune checkpoint inhibitor therapy at the time of study commencement were also asked to provide a faecal sample. The patients brought the faecal sample to the hospital during a routine planned appointment as part of their standard of care. Samples were passed in the morning and kept in cool bags for transfer to the hospital. Patients were met at the hospital appointment by a co-investigator. Samples were coded and stored immediately at -80 degrees Celsius for future processing. There were no changes in the conduct of the study or planned analyses or no adverse and serious adverse events throughout the study.

Demographic, clinical data and, medication history were obtained by direct questioning. Data collection of standard clinicopathologic parameters, clinical outcomes including treatment response, toxicity, duration of response, progression free survival and overall survival were collected sequentially for each patient.

Patients were stratified into two groups, the treatment response (R) group versus treatment non-response (NR) group. Treatment response was defined as radiological stability or decrease of disease burden or disease resolution at six months as defined by the standardized iRECIST (Immune Response Evaluation Criteria in Solid Tumours) criteria[32].

Patients were also stratified into groups based on documented immune checkpoint inhibitor related side effects. Toxicity was graded by oncology clinicians at the time of occurrence using the standardized National Cancer Institute CTCAE (Common Terminology Criteria for Adverse Events) v.5 system[33]. Patients were stratified into two groups, mild or no side effects (NSE) versus side effects (SE). Patients were included in the SE group if they met the criteria of having CTCAE grade 3 (severe adverse event), grade 4 (life threatening or disabling adverse event) or grade 5 (death related to adverse event) side effects.

Healthy controls were also obtained from the population to offer a reference microbiome. These control group were aged between 18-64 with no chronic disease, on no regular medication and had no antibiotics in the preceding month.

4.3.2 DNA extraction from human faeces

Extraction of total microbial DNA was achieved using the repeat bead beating technique with modifications as previously described[34].

16S rRNA gene library preparation and sequencing

Genomic DNA underwent 16s rRNA gene PCR. The 16S rRNA gene was amplified using primers for the V3-V4 region; forward, TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCTACGGGNGGCWGCAG-3′ and reverse, 5′-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGACTACHVGGGTATCTAATCC-3'. The PCR thermocycler protocol was as follows: Initiation step of 98 °C for 3 min followed by 30 cycles of 98 °C for 30 s, 55 °C for 60 s, and 72 °C for 20 s, and a final extension step of 72 °C for 5 min. Indexes were subsequently added to amplicons according to Illumina 16S Metagenomic Sequencing Protocol (Illumina, CA, USA). Libraries DNA concertation was quantified using a Qubit fluorometer (Invitrogen) using the ‘High Sensitivity’ assay and samples were pooled at a standardised concentration. The pooled library was sequenced at Eurofins Genomics/GATC Biotech (Konstanz, Germany) on the Illumina MiSeq platform utilising 2×300 bp chemistry.

4.3.3 Bioinformatic and biostatistical analysis

Raw data was imported into R (v3.6.0) for processing and analysis. Paired reads were quality filtered, trimmed, merged and Amplicon Sequence Variants (ASV) inferred using the R package dada2 (v1.12.1)[35]. Taxonomic classification was performed using the RDP Classifier within Mothur in conjugation with SPINGO, a species-level classifier[36]. A confidence cut of 80% was used for taxonomic assignment. QIIME v1.9.1 and the R package vegan v2.5.6 were used to calculate β-diversity metrics[37]. β-diversity was visualized via principal coordinates analysis (PCoA) plots whose coordinates were identified using with the Ape package v5.1. R-squared (R^2) and p-value were calculated using Permutational Multivariate Analysis of Variance (PERMANOVA) via the R package vegan (v2.4.2). Differential abundance analysis was carried out using DEseq2 (v1.22.2)[38]. Genomic functionality was inferred using PICRUSt2 with the command picrust2_pipeline.py with default parameters[39]. Differential abundance of KOs was performed using DESeq2.

4.4 Results

4.4.1 Patient characteristics and treatment responses

All patients from Cork University Hospital Cancer Centre with malignant metastatic melanoma established or commencing on immunotherapy during the study period were considered eligible for recruitment. Forty one patients were enrolled but four patients were excluded due to frailty or inability to provide samples. Therefore 37 patient samples were analysed. Twenty one patients were commencing on immunotherapy therapy de novo and 16 patients were established on treatment. All patients had stage four metastatic melanoma.

By iRECIST criteria[32], 21 patients were classified as immune checkpoint inhibitor responders (R) (7 de novo and 14 established treatment patients) and 16 as non-responders (NR) (14 de novo patients and 2 established treatment patients). (Table 1) The two groups were comparable in terms of median age and included patients on differing immunotherapy drugs including combination therapy. Seventeen of the responder group either remained on treatment or had successfully completed their treatment protocol at time of analysis. The remaining 4 patients in the responder group developed side effects and had therapy discontinued however still had treatment response at 6 months. None of the non-responder patients remained on therapy at the time of analysis. Ten of the non-responder group had treatment discontinued due to disease progression (n=9) or protocol (n=1) and a further 5 patients had treatment discontinued due to side effects.

Using the CTCAE v. 5 criteria [33], 11 patients had one or more severe side effects (SE) (9 de novo patients, 2 established patients) and 26 patients had mild or no side effects (NSE) (12 de novo patients and 14 established patients). (Table 2). The median age of patients who developed mild or no side effects was 7.8 years older than those who suffered severe side effects. There were seven different immune-mediated conditions recorded in the patient cohort, with three patients experiencing several side effects concurrently (Table 3).None of the 11 patients who had side effects remained on immunotherapy but 14 of the 26 patients in the no side effect category continued treatment at the time of analysis.

Table 1. Demographics of Treatment Responders Versus Non- Responders

Demographics		**Treatment responders (n=21)**	**Treatment non-responders (n=16)**
Mean Age (st deviation)		54(14.5)	57 (10.5)
Sex			
	Male	9	10
	Female	12	6
Type of melanoma			
	Cutaneous	19	14

	Choroidal	1	1
	Unknown primary	1	0
	Gastric	0	1
Median time since diagnosis		66 months	29 months
Treatment Type			
	Pembrolizumab	13	6
	Nivolumab	6	6
	Pembrolizumab/ Ipilimumab	2	1
	Nivolimumab/ ipilimumab	0	3
Treatment ongoing			
	Yes	15	0
	Stopped due to side effect	5	5
	Stopped due to protocol	1	0
	Stopped due to disease progression	0	11

Previous radiotherapy			
	Yes	4	6
	No	17	10
Prior treatment			
	No	11	9
	Short course Ipilimumab	6	1
	Short course Ipilimumab/ nivolumab	1	0
	Dabrafenib/ trametinib	1	4
	Carboplatin /gemcitabine	1	0
	Electro-chemotherapy	1	2
Antibiotic treatment in last 6 weeks			
	No	16	14
	Oral cephalexin	1	0
	Oral Co-amoxiclav	2	0

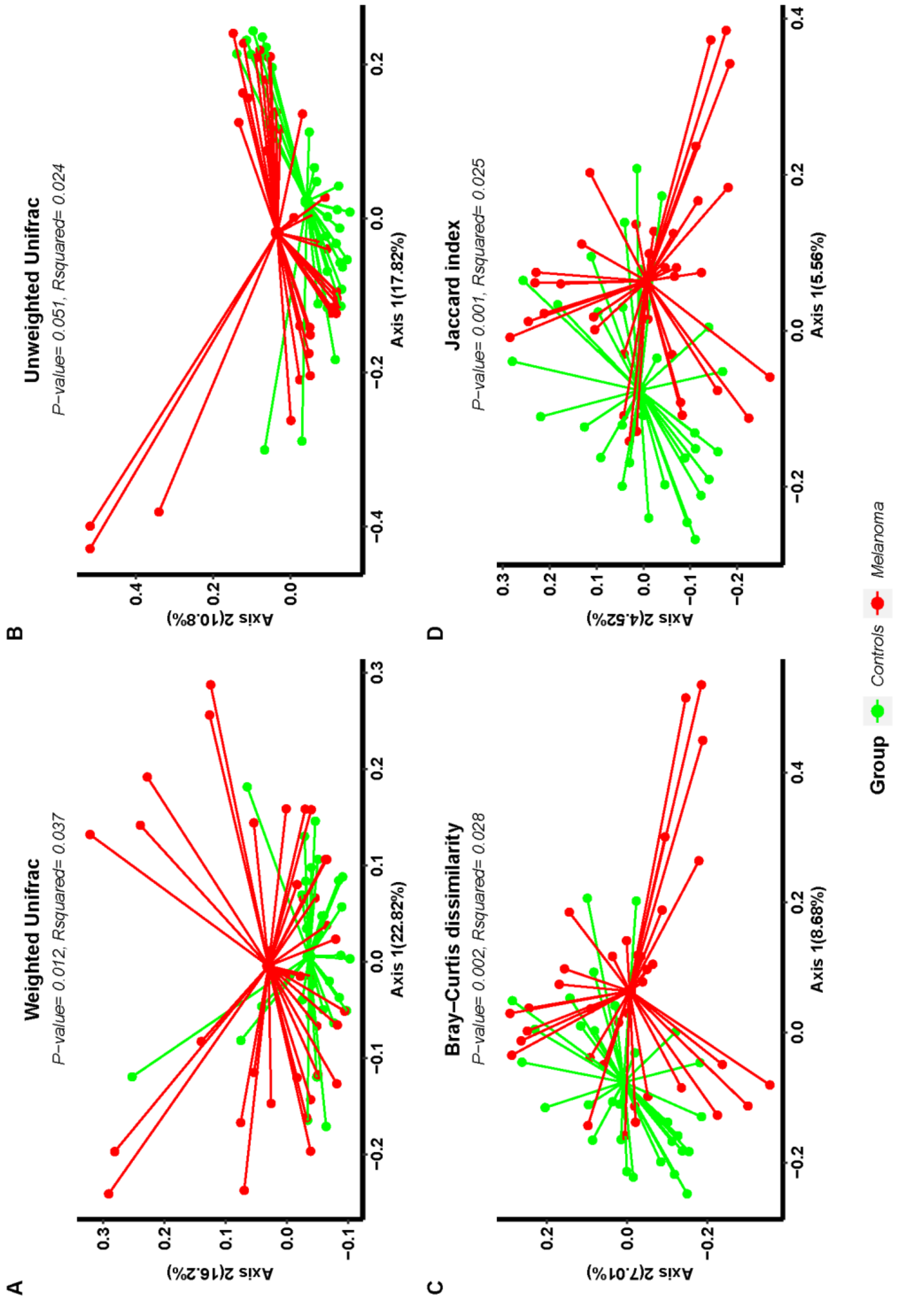
	Oral Penicillin	1	0
	IV Vancomycin	0	1
	Unknown antibiotic	1	1
Alcohol intake			
	None	17	9
	1-5 units per week	2	6
	5-10 units per week	1	1
	10-15 units per week	1	0
Smoking status			
	Current smoker	1	2
	Ex-smoker	4	4
	Non-smoker	16	10
Deaths			
	Yes	1	13
	No	20	1

Table 3. Side effects of ICI therapy with attributable medications

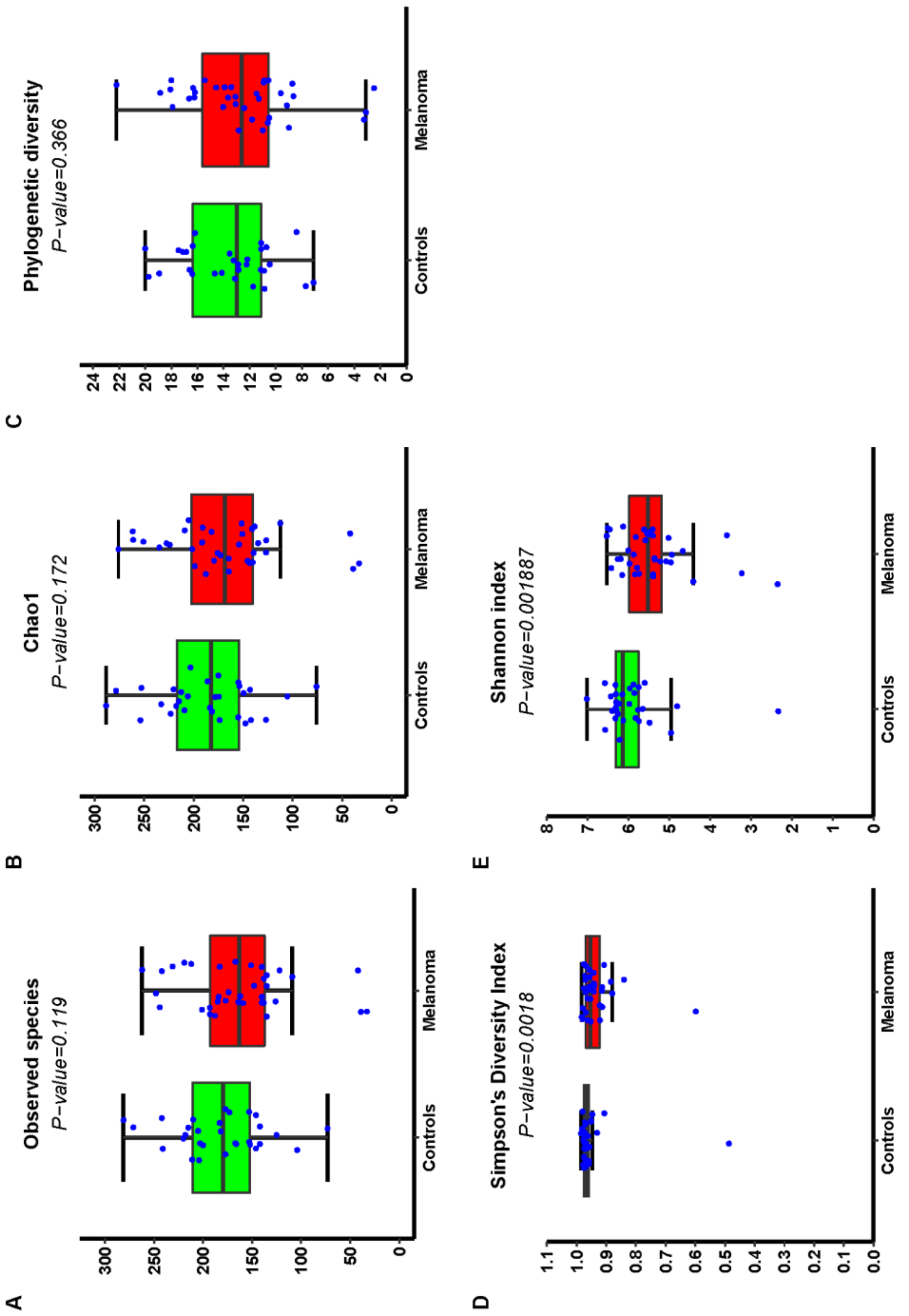
Side Effect (CTCAE grade 3/4)	**Number of patients**	**Attributable medication**
Hypophysitis	3	Pembrolizumab n=2 Pembolizumab/ipilimumab n=1
Hepatitis	2	Nivolumab/ipilimumab n=1 Ipilimumab/pembrolizumab n=1
Rash	1	Pembrolizumab
Colitis	4	Nivolumab/ipilimumab n=1 Pembrolizumab n=2 Nivolumab n=1
Neurotoxicity	1	Nivolumab
Cellulitis	1	Pembrolizumab
Diabetic ketoacidosis	1	Pembrolizumab

4.4.2 Microbiota features associated with therapy outcomes

Global microbiome structure as measured by beta diversity differed slightly between melanoma patients and healthy controls (Supplementary figure 1). There was no significant difference in Alpha diversity (a measure of species richness) as measured by observed species, chao1 and phylogenetic diversity indices between melanoma patients and healthy controls (Supplementary figure 2 A, B, C). A significant reduction in alpha diversity as measured by Simpson's and Shannon index (a measure of diversity evenness) was observed in melanoma patients relative to healthy controls (Supplementary figure 2 D, E).



Supplementary Figure 1. Principal Coordinates Analysis representation of beta diversity comparing patients with melanoma versus healthy controls. (A) Weighted Unifrac **(B)** Unweighted Unifrac **(C)** Bray–Curtis dissimilarity **(D)** Jaccard index. R-Squared and P-value calculated using Permutational Multivariate Analysis of Variance (PERMANOVA).



Supplementary Figure 2. Bar plots showing the difference in alpha diversity metrics between individuals with melanoma and healthy controls. (A) Observed species index **(B)** Chao1 **(C)** Phylogenetic diversity **(D)** Simpson's Diversity Index **(E)** Shannon index. Statistical testing was performed using Wilcoxon signed-rank test

Pairwise comparison of beta diversity with respect to response demonstrated that both the responder group and non-responder group had a significantly different beta-diversity compared to healthy controls (Figure 1A, Supplementary table 1). However, there was no significant difference in beta diversity between responders and non-responders. With respect to side effects, individuals with no side effects differed significantly compared to healthy controls; however no other pairwise comparison differed significantly including the observation that individuals with no side effects did not differ significantly from individuals with side effects. (Figure 1B, Supplementary table 2)

Alpha diversity did not differ significantly between responders versus non responders nor between individuals with non-side do effects versus individuals with side effects (figure 1C, D, Supplementary figure 3, Supplementary figure 4).

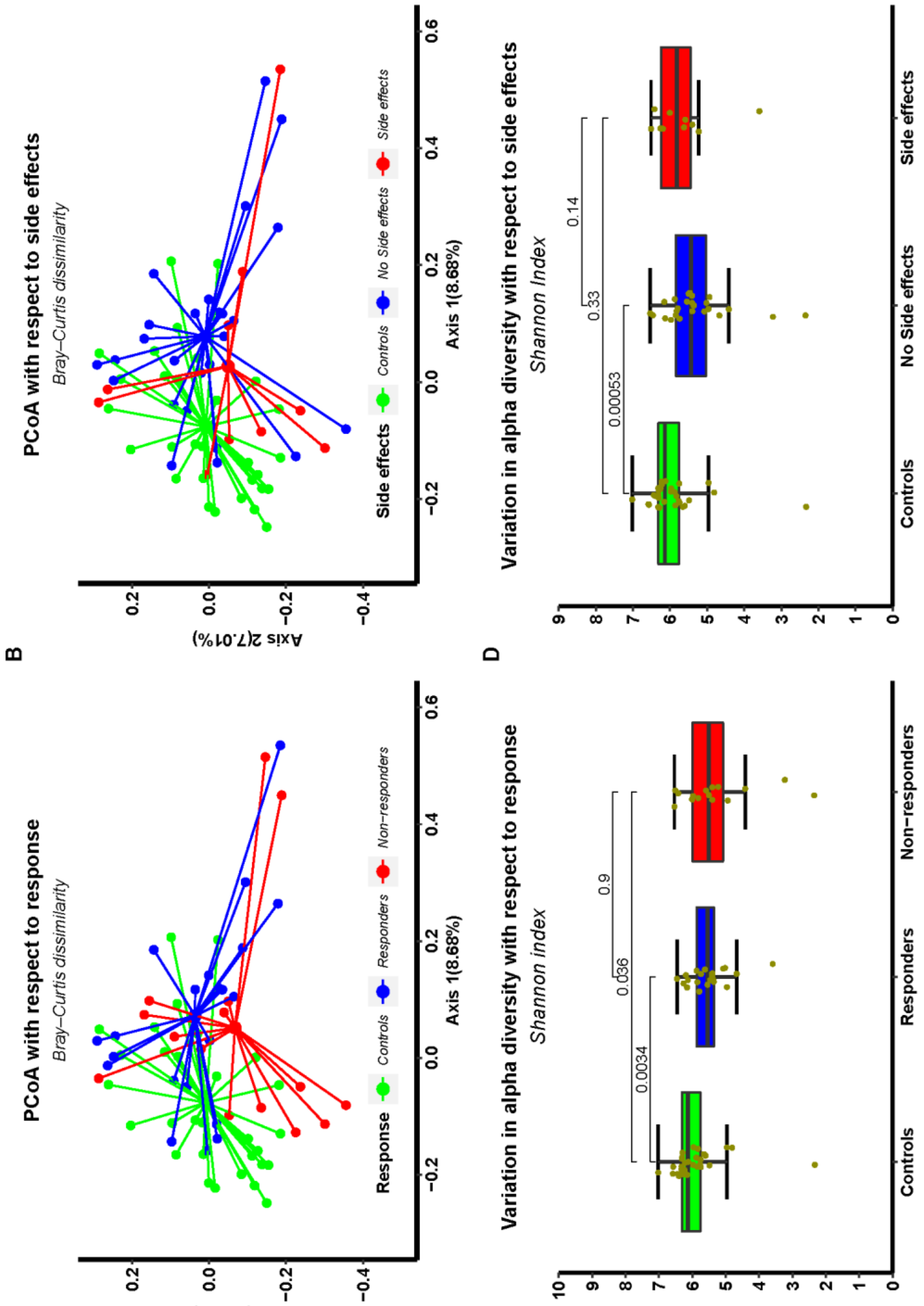


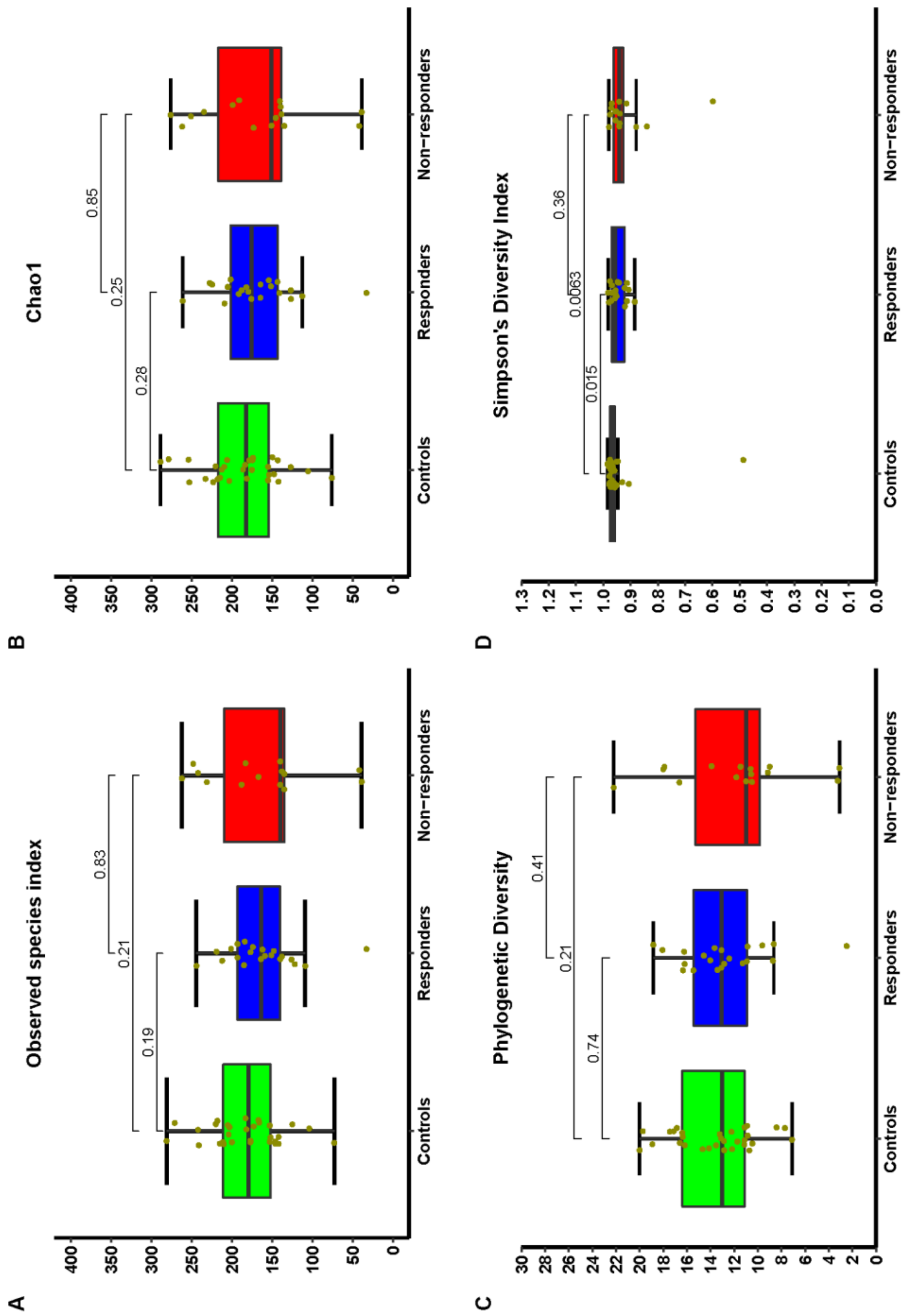
Figure 1. Comparisons of microbiome ecological metrics between immunotherapy outcome groups. (A) Principal Coordinates Analysis representation of Beta diversity (Bray–Curtis dissimilarity) between controls, responders and non-responders. **(B)** Principal Coordinates Analysis representation of Beta diversity (Bray–Curtis dissimilarity) between controls, no side effects and side effects. **(C)** Boxplot comparing alpha diversity (Shannon index) between controls, responders and non-responders. **(D)** Boxplot comparing alpha diversity (Shannon index) between controls, individuals with no side effects and individuals with side effects. Statistical testing of alpha-diversity was performed using Wilcoxon signed-rank test

Pairwise PERMANOVA with respect to responds			
Weighted unifrac			
	p-value	R squared	Adjusted p-value
Controls versus Responders	0.017982	0.043342	0.034466
Controls versus non responders	0.022977	0.048418	0.034466
Responders versus Non_responder	0.509491	0.025559	0.509491
Unweighted unifrac			
	p-values	R squared	Adjusted p-value
Controls versus Responders	0.063936	0.029749	0.095904
Controls versus non responder	0.041958	0.036071	0.095904
Responders versus non responder	0.293706	0.031309	0.293706
Bray–Curtis dissimilarity			
	p-values	R-squared	Adjusted p-value
Controls versus Responders	0.001998	0.035139	0.004496
Controls versus non responders	0.002997	0.038529	0.004496
Responders versus non responder	0.123876	0.03438	0.123876
Jaccard index			
	P-values	rsquared	Adjusted p-value
Controls versus Responders	0.000999	0.031047	0.001499
Controls versus non responders	0.000999	0.033836	0.001499
Responders versus non responders	0.17982	0.031066	0.17982

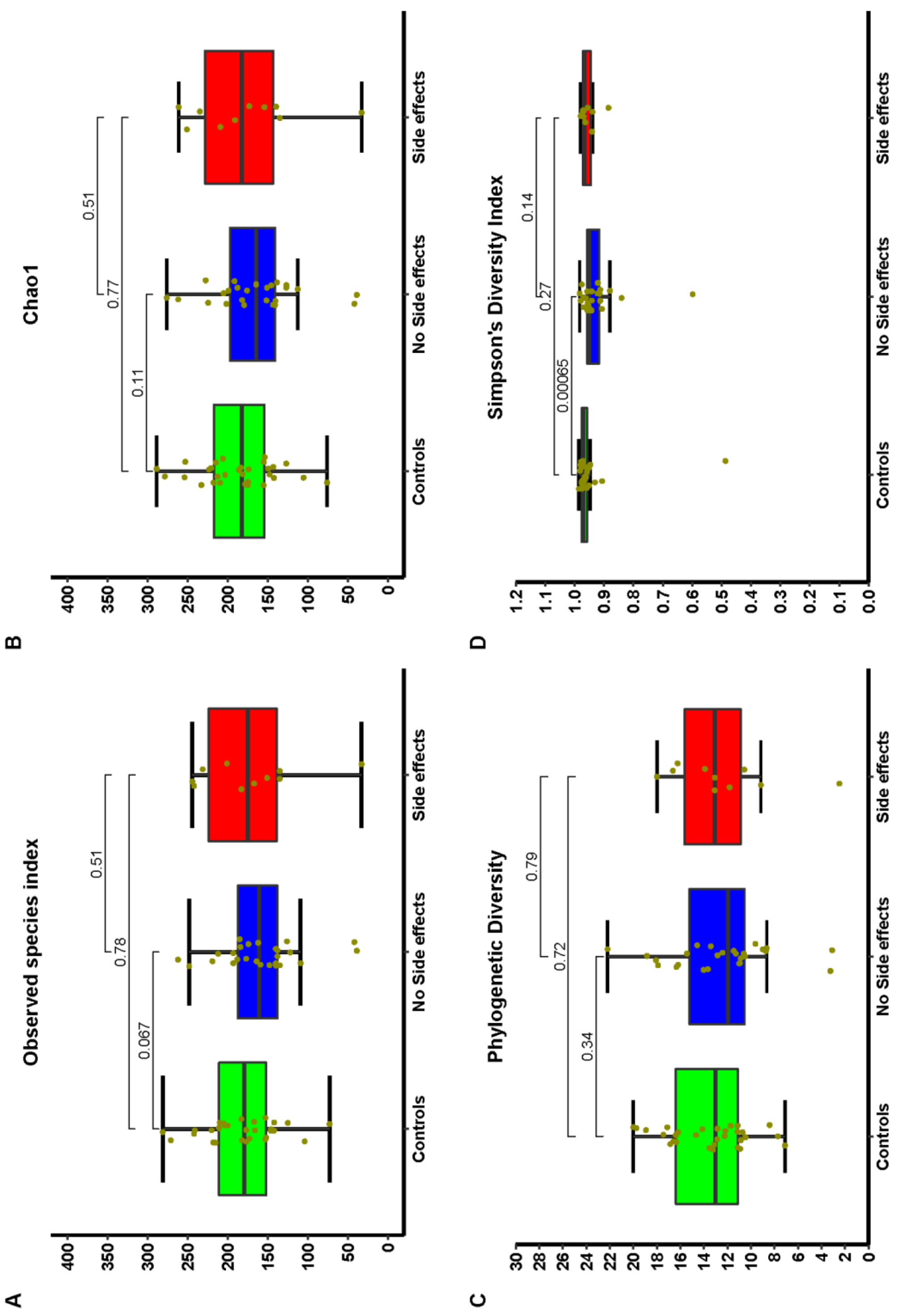
Supplementary table 1. Pairwise comparisons with respect to response. P-value calculated using Permutational multivariate analysis of variance (PERMANOVA). Multiple correction performed using Benjamini–Hochberg procedure.

Pairwise PERMANOVA with respect to side effects			
	p-values	R squared	Adjusted p-value
Controls versus no side effects	0.008991	0.047504	0.026973
Controls versus side effects	0.453546	0.024113	0.68032
No side effects versus side effects	0.896104	0.014881	0.896104
Unweighted unifrac			
	p-values	R squared	Adjusted p-value
Controls versus no side effects	0.037962	0.027729	0.113886
Controls versus side effects	0.235764	0.030116	0.353646
No side effects versus side effects	0.896104	0.020161	0.896104
Bray–Curtis dissimilarity			
	p-values	R squared	Adjusted p-value
Controls versus no side effects	0.000999	0.034634	0.002997
Controls versus side effects	0.367632	0.0263	0.551449
No side effects versus side effects	0.967033	0.020871	0.967033
Jaccard index			
	p-values	R squared	Adjusted p-value
Controls versus no side effects	0.000999	0.030479	0.002997
Controls versus side effects	0.133866	0.028443	0.200799
No side effects versus Side effects	0.935065	0.024676	0.935065

Supplementary table 2. Pairwise comparisons with respect to side effects. P-value calculated using Permutational multivariate analysis of variance (PERMANOVA). Multiple correction performed using Benjamini–Hochberg procedure.



Supplementary Figure 3 Bar plots showing the difference in alpha diversity metrics between controls, responders and non-responders. (A) Observed species index **(B)** Chao1 **(C)** Phylogenetic diversity **(D)** Simpson's Diversity Index. Statistical testing was performed using Wilcoxon signed-rank test



Supplementary Figure 4. Bar plots showing the alpha diversity metrics of controls, individuals with no side effects and individuals with side effects. **(A)** Observed species index **(B)** Chao1 **(C)** Phylogenetic diversity **(D)** Simpson's Diversity Index. Statistical testing was performed using Wilcoxon signed-rank test

In this study we used the denoising DADA2 algorithm to rationalise microbial sequence data to the single nucleotide resolution in the form of amplicon sequence variants (ASVs) [35]. Differential abundance analysis of ASVs was performed using DESeq2. We identified 15 ASVs that were significantly differentially abundant between responders and non-responders while 9 ASVs were significantly differentially abundant between individuals with no-side effects versus individuals with side effects (Figure 2). ASVs assigned to the species *Ruminococcus bromii*, *Bifidobacterium longum*, *Akkermansia muciniphila*, *Gemmiger formicilis* and *Prevotella copri* were found to be enriched in responders relative to non-responders which is consistent with previous findings [20,40-42]. In a recent meta-analysis *A. muciniphila* and *R. bromii* were found to be consistently over-represented in responders[21]. ASVs assigned to responder associated species including *R.bromii* and *B.longum* significantly more enriched in responders versus healthy controls (Supplementary figure 5A).However, healthy controls were observed to be enriched in ASVs assigned to responder associated species versus non-responders, that is, *A. muciniphila*, *G. formicilis* and *P. copri* (Supplementary figure 5B).

A number of ASVs found to be enriched in individuals with no side effects relative to healthy controls over-lapped with those enriched in responders including ASVs assigned to the species *A. muciniphila* and *B. intestinihominis* (Figure 2). Of note, an ASV assigned to the genus Oscillibacter was uniquely enriched in individuals with no side effects. Further, a number of ASVs were differentially abundant between individuals with and without side effects and both of these versus healthy controls (Supplementary Figure 6).

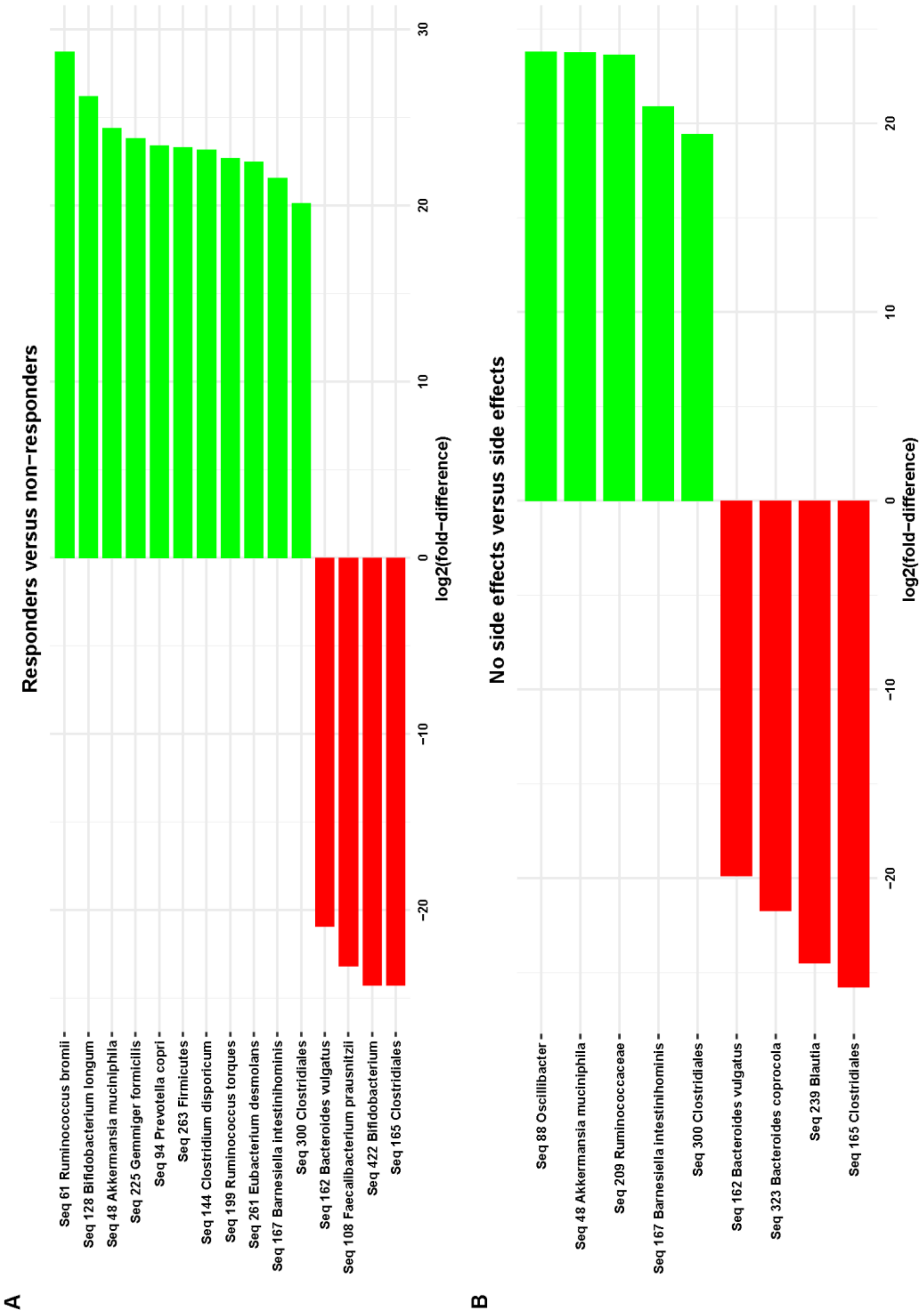
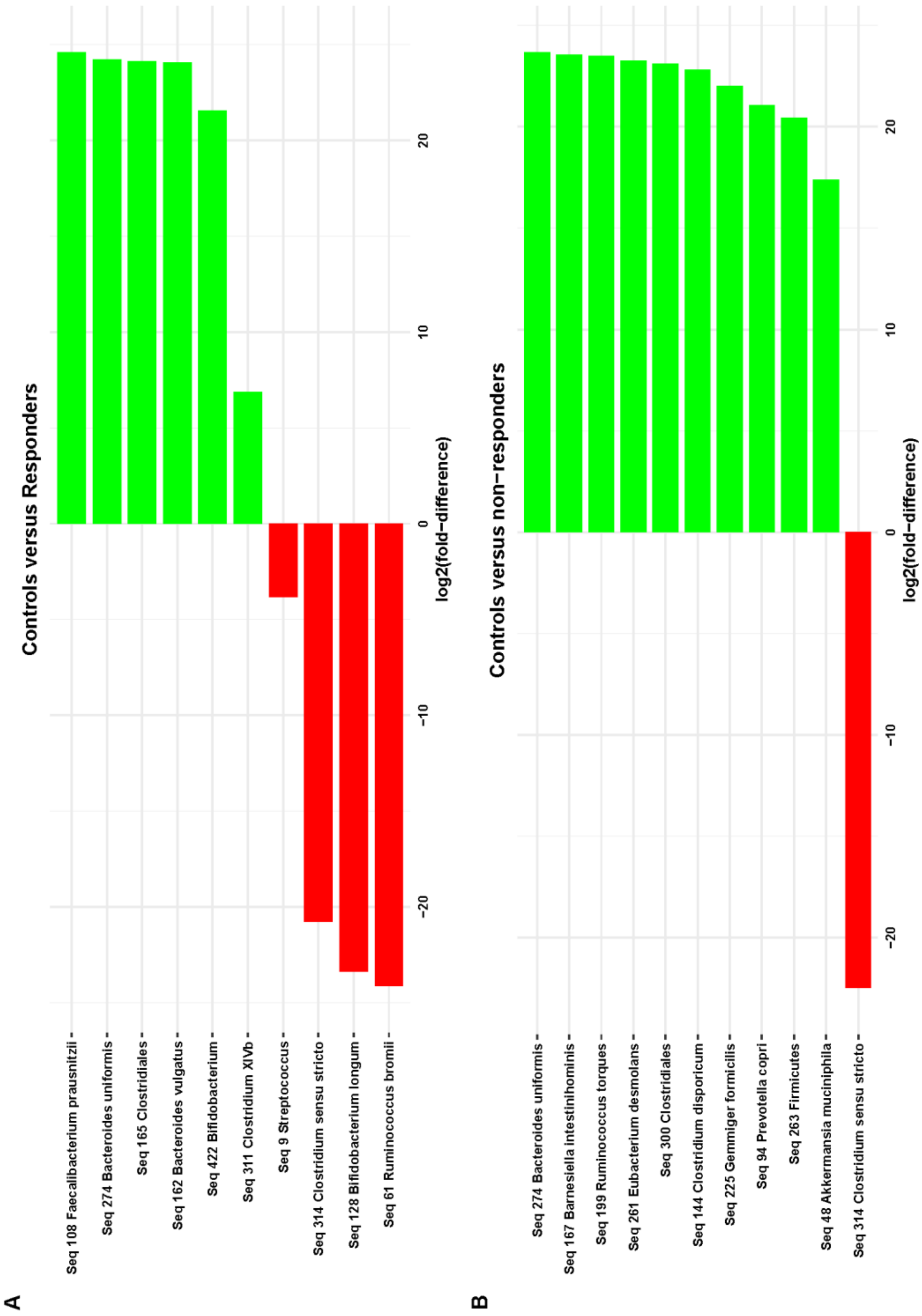
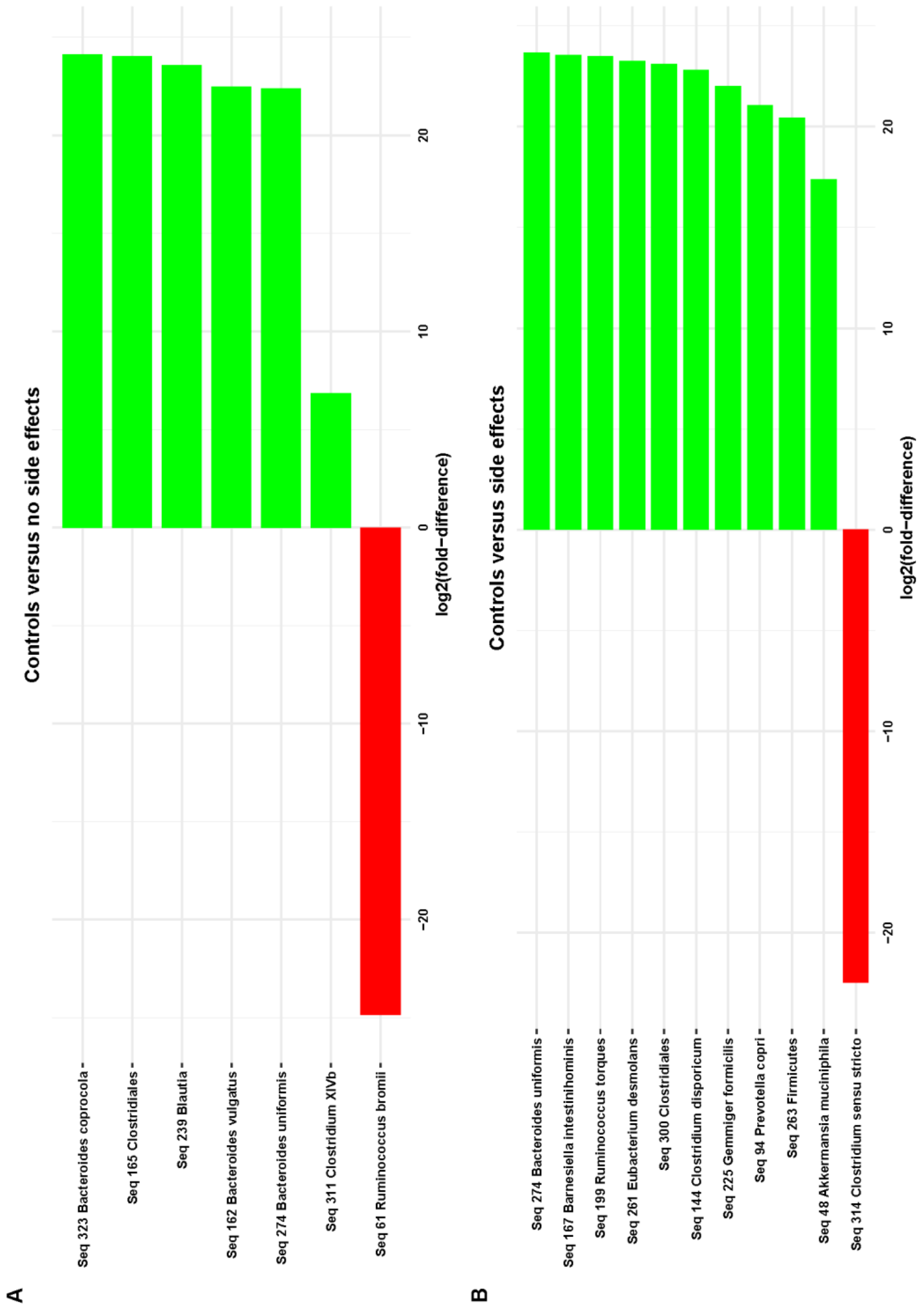


Figure 2. Differentially abundant ASVs associated with immunotherapy response and side effects. (A) Significantly differentially abundant ASVs with respect to response**.** ASVs over-represented in responders in green. ASVs over-represented in non-responders in red **(B)** Significantly differentially abundant ASVs with respect to side effects. ASVs over-represented in individuals with no side effects in green. ASVs over-represented in individuals with side effects in red. Statistical testing was performed using DESeq2, p-value < 0.05.



Supplementary Figure 5. Differentially abundant ASVs between controls and responder groups. (A) Significantly differentially abundant ASVs between controls and responders. **(B)** Significantly differentially abundant ASVs between controls and non-responders. ASVs over-represented in controls in green. ASVs over-represented in non-responders in red. Statistical testing was performed using DESeq2, p-value < 0.05.



Supplementary Figure 6. Differentially abundant ASVs between controls and side effect groups. (A) Significantly differentially abundant ASVs between controls and individuals with no side effects. ASVs over-represented in controls in green. ASVs over-represented in individuals with no side effects in red **(B)** Significantly differentially abundant ASVs between controls and individuals with side effects. ASVs over-represented in controls in green. ASVs over-represented in individuals with side effects in red. Statistical testing was performed using DESeq2, p-value < 0.05.

Previous studies have identified that the abundance of particular gut microbiota proteins, represented by gene counts of KEGG orthologues, to be differentially abundant in responders and non-responders[21]. We inferred functional genomic capabilities of the microbiome composition datasets with the software PICRUSt2 [39], and then DESeq2 was used to identify differential abundant KEGG orthologues (KOs). A number of KOs were thus found to be differentially abundant between responders and non-responders as well as between individuals with no side effects versus individuals with side effects (Figure 3). A galactosyltransferase and glycosyltransferase were overrepresented in the microbiome of responders relative to non-responders. These enzymes are involved in the production of Exopolysaccharides (EPS), a bacterial polymer which in some bacteria has immunomodulatory properties [43]. A number of membrane associated proteins including methyl‒accepting chemotaxis protein IV, type III secretion protein, sensor histidine kinase EvgS and proteins relating to EPS production were also found to be enriched in individuals with no side effects.

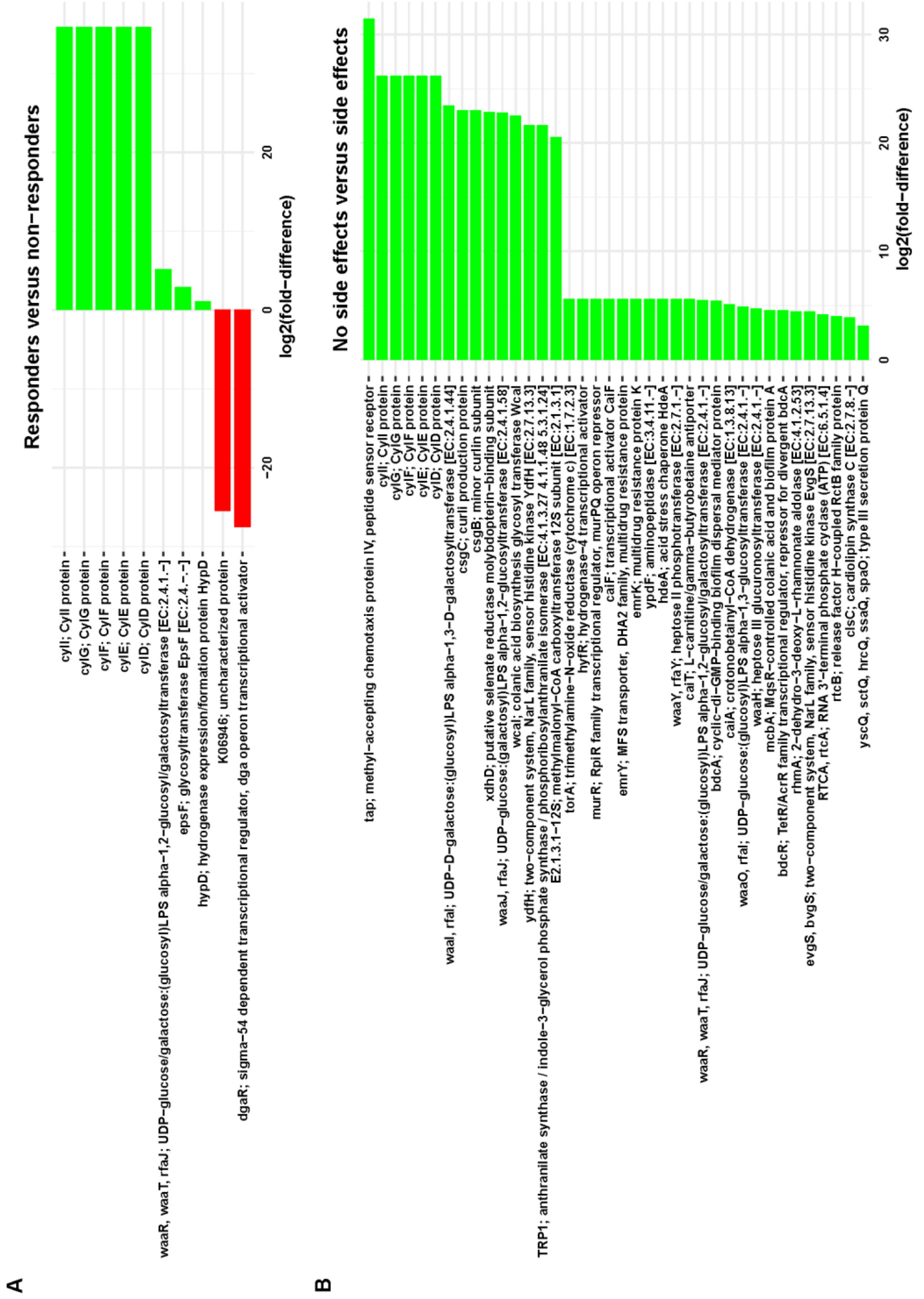


Figure 3. Differentially abundant KEGG Orthologs (KOs) associated with immunotherapy response and side effects. (A) Significantly differentially abundant KOs with respect to response. KOs over-represented in responders in green. KO s over-represented in non-responders in red **(B)** Significantly differentially abundant KOs with respect to side effects. KOs over-represented in individuals with no side effects in green. KOs over-represented in in individuals with side effects in red Statistical testing was performed using DESeq2, p-value < 0.05.

4.5 Discussion

We identify significant differences in gut microbiota species abundance associated with both treatment response, and protection against moderate and severe side effects in patients with stage four metastatic melanoma. There was no major difference in global ecological structure as measured by alpha and beta-diversity. Consistent with previous reports, we found a number of taxa which were over-represented in treatment responders including *A. muciniphila*. One mechanism by which *A. muciniphila* may modulate response is through the production of the purine nucleoside inosine which can activates T helper 1 (TH1) in an adenosine 2A receptor (A2AR)–dependent manner[44]. *A. muciniphila* has been previously identified to be negatively correlated with overweight/obese individuals[45,46]. However, although not conclusive, the current data indicates that there is an association between Progression-free survival and overweight/obese individuals[47-49]. We also identified ASVs assigned to species which have not been previously reported as over-represented in treatment responders including *Clostridium disporicum, Ruminococcus torques, Eubacterium desmolans* and *Barnesiella intestinihominis.* Notably, *B. intestinihominis* has been demonstrated in mice models to augment the chemo-immunotherapeutic drug Cyclophosphamide via promoting recruitment of Type 1 CD8+ T cells and type 1 CD4+ T helper cells to the colon and the restoration of intratumoral interferon-γ (IFN-γ) producing gamma delta (γδ) T cells [50]. It was shown that a consortium of 11 microbes promote the anti-cancer effect of immune checkpoint inhibitors by promoting the production of CD8+ IFN-γ+ T cells[51]. Thus *B. intestinihominis* may modulate ICI response via a mechanism involving IFN-γ production.

Curiously, in contrast to previous reports, an ASV assigned to *Faecalibacterium prausnitzii*, a bacterium usually associated with putative health-promoting properties [52], was elevated in non-responders. However certain strains of *Faecalibacterium prausnitzii* cause distinct effects on immune cells when compared to other strains [53], and so the ASV identified may represent strain differences compared to previous findings.

Discontinuation of immunotherapy as a result of side effects [26] despite treatment efficacy is an unfortunate reality for many patients on immune checkpoint inhibitors. Few studies have examined differences in microbiome composition in patients with and without side effects. We report, for the first time, to our knowledge, a number of ASVs as associated with mild or no side effects relative to patients who developed side effects. Further, while these reports focused on immune checkpoint inhibitor colitis, our study addressed all side effects associated with immunotherapy. It is uncertain whether the mechanism of colitis and that for other side effects differs. Individuals with no side effects were observed to have an increased relative abundance of an ASV assigned to the Oscillibacter, a genus known to produce anti-inflammatory compounds and reduce intestinal TH17 cell expansion in mice models[41]. A recent study reported an enrichment of Oscillibacter in inactive Crohn’s disease relative to active Crohn’s disease [54]. Together this might suggest that Oscillibacter has a role in preventing immune related side effects.

We identified a number of proteins which were differential abundant between responders and non-responders as well between side effect groups. We identified proteins involved in Exopolysaccharides (EPS) biogenesis enriched in both responders and individuals with no side effects. EPS are polymers produced by lactic acid bacteria including *Lactobacillus* and *Bifidobacterium*. Some EPS types has been demonstrated to have immune-stimulatory, immune-modulating and anti-inflammatory qualities [43,55]. Furthermore, EPS has been shown to have cytotoxic affects against cancer cells[56]. It is possible that EPS molecules can help augment the actions of ICI while preventing an excessive/aberrant immune response. The other surface proteins which were found to be enriched in individuals with no side-effects may also offer mechanistic insight to modulating the immune system in the context of immunotherapy.

A limitation of the present study is a relatively small cohort size but this is offset by the inclusion of subjects of the same ethnicity and geographic region. While we used the validated iRESIST criteria to assess disease state and treatment response at six months[32] long-term longitudinal data will be required to identify microbiota composition linked with overall progression-free survival.

Identification of microbes associated with treatment response and protection against side effects in patients receiving immunotherapy raises questions about methods of microbiome manipulation to induce a favourable microbial state. Faecal microbiota transplant (FMT) is an effective treatment of recurrent and refractory *C.difficile* infection[57] which has led to the investigation of FMT to change the gut microbiome in mice treated with immunotherapy, with promising preliminary results[17,19,20,58]. Recent reports have highlighted the potential of FMT in overcoming resistance in patients with melanoma receiving immunotherapy[59,60] However, FMT poses the risk of transmissible infection [61] and the possibility of transfer of inflammatory, metabolic or behavioural phenotypes [62]. FMT using defined microbial consortia, so-called artificial stool, may represent a safer method of replacing the "missing microbe". [63] Therefore robust, adequately powered trials are also required in patients receiving immunotherapy to evaluate the best methods of microbiome manipulation [13].

Acknowledgements: We thank the patients who consented to participate in this study, and the support staff of relevant departments in Cork University Hospital and the Mercy University Hospital Cork. We would like to thank Dr. Dearbhaile Collins for her assistance with sample collection

Authors' contributions: CLM, MB, PP, DGP, FS, PWOT contributed equally to this paper. All authors conceived the work that led to the submission and played an important role in its completion, drafted and revised the paper, approved the final version and agreed to be accountable for all aspects of the work.

Ethics approval and consent to participate: Ethical approval was granted by The Clinical Research Ethics Committee of the Cork Teaching Hospitals (Cork, Ireland) October 2017 under the study reference APC081. The study was conducted in accordance with the ethical principles set forth in the current version of the Declaration of Helsinki, the International Conference on Harmonization E6 Good Clinical Practice (ICH-GCP)

Consent for publication: Consent for publication was obtained from all study participants.

Data availability: Datasets available on request

Competing Interests: The authors declare no conflict of interest

Funding information: Work in PWOTs laboratory is supported by Science Foundation Ireland through a Centre award (APC/SFI/12/RC/2273_P2) to APC Microbiome Ireland, and by The Health Research Board of Ireland under Grant ILP-POR-2017-034.

4.6 Reference

1 Couzin-Frankel, J. Breakthrough of the year 2013. Cancer immunotherapy. *Science* **342**, 1432-1433, doi:10.1126/science.342.6165.1432 (2013).

2 Tumeh, P. C. *et al.* PD-1 blockade induces responses by inhibiting adaptive immune resistance. *Nature* **515**, 568-571, doi:10.1038/nature13954 (2014).

3 Botticelli, A. *et al.* Cross-talk between microbiota and immune fitness to steer and control response to anti PD-1/PDL-1 treatment. *Oncotarget* **8**, 8890-8899, doi:10.18632/oncotarget.12985 (2017).

4 Francisco, L. M. *et al.* PD-L1 regulates the development, maintenance, and function of induced regulatory T cells. *J Exp Med* **206**, 3015-3029, doi:10.1084/jem.20090847 (2009).

5 McCoy, K. D. & Le Gros, G. The role of CTLA-4 in the regulation of T cell immune responses. *Immunol Cell Biol* **77**, 1-10, doi:10.1046/j.1440-1711.1999.00795.x (1999).

6 Buchbinder, E. I. & Desai, A. CTLA-4 and PD-1 Pathways: Similarities, Differences, and Implications of Their Inhibition. *Am J Clin Oncol* **39**, 98-106, doi:10.1097/COC.0000000000000239 (2016).

7 Sharma, P. & Allison, J. P. The future of immune checkpoint therapy. *Science* **348**, 56-61, doi:10.1126/science.aaa8172 (2015).

8 Leach, D. R., Krummel, M. F. & Allison, J. P. Enhancement of antitumor immunity by CTLA-4 blockade. *Science* **271**, 1734-1736 (1996).

9 Pitt, J. M. *et al.* Resistance Mechanisms to Immune-Checkpoint Blockade in Cancer: Tumor-Intrinsic and -Extrinsic Factors. *Immunity* **44**, 1255-1269, doi:10.1016/j.immuni.2016.06.001 (2016).

10 Larkin, J. *et al.* Five-Year Survival with Combined Nivolumab and Ipilimumab in Advanced Melanoma. *N Engl J Med* **381**, 1535-1546, doi:10.1056/NEJMoa1910836 (2019).

11 Herbst, R. S. *et al.* Long-Term Outcomes and Retreatment Among Patients With Previously Treated, Programmed Death-Ligand 1–Positive, Advanced Non–Small-Cell Lung Cancer in the KEYNOTE-010 Study. *J Clin Oncol* **38**, 1580-1590, doi:10.1200/JCO.19.02446 (2020).

12 Pitt, J. M. *et al.* Fine-Tuning Cancer Immunotherapy: Optimizing the Gut Microbiome. *Cancer Res* **76**, 4602-4607, doi:10.1158/0008-5472.CAN-16-0448 (2016).

13 Murphy, C. L., O’Toole, P. W. & Shanahan, F. The Gut Microbiota in Causation, Detection, and Treatment of Cancer. *Am J Gastroenterol*, doi:10.14309/ajg.0000000000000075 (2019).

14 Paulos, C. M. *et al.* Microbial translocation augments the function of adoptively transferred self/tumor-specific CD8+ T cells via TLR4 signaling. *J Clin Invest* **117**, 2197-2204, doi:10.1172/JCI32205 (2007).

15 Iida, N. *et al.* Commensal bacteria control cancer response to therapy by modulating the tumor microenvironment. *Science* **342**, 967-970, doi:10.1126/science.1240527 (2013).

16 Iglesias-Santamaría, A. Impact of antibiotic use and other concomitant medications on the efficacy of immune checkpoint inhibitors in patients with advanced cancer. *Clin Transl Oncol*, doi:10.1007/s12094-019-02282-w (2020).

17 Routy, B. *et al.* Gut microbiome influences efficacy of PD-1-based immunotherapy against epithelial tumors. *Science*, doi:10.1126/science.aan3706 (2017).

18 Vétizou, M. *et al.* Anticancer immunotherapy by CTLA-4 blockade relies on the gut microbiota. *Science* **350**, 1079-1084, doi:10.1126/science.aad1329 (2015).

19 Gopalakrishnan, V. *et al.* Gut microbiome modulates response to anti-PD-1 immunotherapy in melanoma patients. *Science* **359**, 97-103, doi:10.1126/science.aan4236 (2018).

20 Matson, V. *et al.* The commensal microbiome is associated with anti-PD-1 efficacy in metastatic melanoma patients. *Science* **359**, 104-108, doi:10.1126/science.aao3290 (2018).

21 Gharaibeh, R. Z. & Jobin, C. Microbiota and cancer immunotherapy: in search of microbial signals. *Gut*, doi:10.1136/gutjnl-2018-317220 (2018).

22 Postow, M. A., Sidlow, R. & Hellmann, M. D. Immune-Related Adverse Events Associated with Immune Checkpoint Blockade. *N Engl J Med* **378**, 158-168, doi:10.1056/NEJMra1703481 (2018).

23 Chen, D. S. & Mellman, I. Oncology meets immunology: the cancer-immunity cycle. *Immunity* **39**, 1-10, doi:10.1016/j.immuni.2013.07.012 (2013).

24 Luoma, A. M. *et al.* Molecular Pathways of Colon Inflammation Induced by Cancer Immunotherapy. *Cell* **182**, 655-671.e622, doi:10.1016/j.cell.2020.06.001 (2020).

25 Seldin, M. F. The genetics of human autoimmune disease: A perspective on progress in the field and future directions. *J Autoimmun* **64**, 1-12, doi:10.1016/j.jaut.2015.08.015 (2015).

26 Khoja, L., Day, D., Wei-Wu Chen, T., Siu, L. L. & Hansen, A. R. Tumour- and class-specific patterns of immune-related adverse events of immune checkpoint inhibitors: a systematic review. *Ann Oncol* **28**, 2377-2385, doi:10.1093/annonc/mdx286 (2017).

27 Zhai, Y. *et al.* Endocrine toxicity of immune checkpoint inhibitors: a real-world study leveraging US Food and Drug Administration adverse events reporting system. *J Immunother Cancer* **7**, 286, doi:10.1186/s40425-019-0754-2 (2019).

28 Wang, D. Y. *et al.* Fatal Toxic Effects Associated With Immune Checkpoint Inhibitors: A Systematic Review and Meta-analysis. *JAMA Oncol* **4**, 1721-1728, doi:10.1001/jamaoncol.2018.3923 (2018).

29 Shivaji, U. N. *et al.* Immune checkpoint inhibitor-associated gastrointestinal and hepatic adverse events and their management. *Therap Adv Gastroenterol* **12**, 1756284819884196, doi:10.1177/1756284819884196 (2019).

30 Michot, J. M. *et al.* Immune-related adverse events with immune checkpoint blockade: a comprehensive review. *Eur J Cancer* **54**, 139-148, doi:10.1016/j.ejca.2015.11.016 (2016).

31 Topalian, S. L. *et al.* Survival, durable tumor remission, and long-term safety in patients with advanced melanoma receiving nivolumab. *J Clin Oncol* **32**, 1020-1030, doi:10.1200/JCO.2013.53.0105 (2014).

32 Seymour, L. *et al.* iRECIST: guidelines for response criteria for use in trials testing immunotherapeutics. *Lancet Oncol* **18**, e143-e152, doi:10.1016/S1470-2045(17)30074-8 (2017).

33 Institute, N. C. *Common Terminology Criteria for Adverse Events (CTCAE) V5.0*, <https://ctep.cancer.gov/protocolDevelopment/electronic_applications/ctc.htm> (2020).

34 Ghosh, T. S. *et al.* Mediterranean diet intervention alters the gut microbiome in older people reducing frailty and improving health status: the NU-AGE 1-

year dietary intervention across five European countries. *Gut* **69**, 1218, doi:10.1136/gutjnl-2019-319654 (2020).

35 Callahan, B. J. *et al.* DADA2: High-resolution sample inference from Illumina amplicon data. *Nat Methods* **13**, 581-583, doi:10.1038/nmeth.3869 (2016).

36 Allard, G., Ryan, F. J., Jeffery, I. B. & Claesson, M. J. SPINGO: a rapid species-classifier for microbial amplicon sequences. *BMC Bioinformatics* **16**, 324, doi:10.1186/s12859-015-0747-1 (2015).

37 Caporaso, J. G. *et al.* QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* **7**, 335-336, doi:10.1038/nmeth.f.303 (2010).

38 Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15**, 550, doi:10.1186/s13059-014-0550-8 (2014).

39 Douglas, G. M. *et al.* PICRUSt2: An improved and extensible approach for metagenome inference. *bioRxiv*, 672295, doi:10.1101/672295 (2019).

40 Gopalakrishnan, V. *et al.* Gut microbiome modulates response to anti–PD-1 immunotherapy in melanoma patients. *Science* **359**, 97-103, doi:10.1126/science.aan4236 (2018).

41 Routy, B. *et al.* Gut microbiome influences efficacy of PD-1-based immunotherapy against epithelial tumors. *Science* **359**, 91-97, doi:10.1126/science.aan3706 (2018).

42 Chaput, N. *et al.* Baseline gut microbiota predicts clinical response and colitis in metastatic melanoma patients treated with ipilimumab. *Ann Oncol* **28**, 1368-1379, doi:10.1093/annonc/mdx108 (2017).

43 Fanning, S. *et al.* Bifidobacterial surface-exopolysaccharide facilitates commensal-host interaction through immune modulation and pathogen protection. *Proc Natl Acad Sci U S A* **109**, 2108-2113, doi:10.1073/pnas.1115621109 (2012).

44 Mager, L. F. *et al.* Microbiome-derived inosine modulates response to checkpoint inhibitor immunotherapy. *Science* **369**, 1481, doi:10.1126/science.abc3421 (2020).

45 Liu, R. *et al.* Gut microbiome and serum metabolome alterations in obesity and after weight-loss intervention. *Nature medicine* **23**, 859-868 (2017).

46 Cani, P. D. & de Vos, W. M. Next-generation beneficial microbes: the case of Akkermansia muciniphila. *Frontiers in microbiology* **8**, 1765 (2017).

47 Cortellini, A. *et al.* A multicenter study of body mass index in cancer patients treated with anti-PD-1/PD-L1 immune checkpoint inhibitors: when overweight becomes favorable. *Journal for immunotherapy of cancer* **7**, 1-11 (2019).

48 Cortellini, A. *et al.* Baseline BMI and BMI variation during first line pembrolizumab in NSCLC patients with a PD-L1 expression≥ 50%: a

multicenter study with external validation. *Journal for immunotherapy of cancer* **8** (2020).

49 Indini, A. *et al.* Impact of BMI on Survival Outcomes of Immunotherapy in Solid Tumors: A Systematic Review. *International journal of molecular sciences* **22**, 2628 (2021).

50 Daillère, R. *et al.* Enterococcus hirae and Barnesiella intestinihominis Facilitate Cyclophosphamide-Induced Therapeutic Immunomodulatory Effects. *Immunity* **45**, 931-943, doi:10.1016/j.immuni.2016.09.009 (2016).

51 Tanoue, T. *et al.* A defined commensal consortium elicits CD8 T cells and anti-cancer immunity. *Nature* **565**, 600-605, doi:10.1038/s41586-019-0878-z (2019).

52 Lopez-Siles, M., Duncan, S. H., Garcia-Gil, L. J. & Martinez-Medina, M. Faecalibacterium prausnitzii: from microbiology to diagnostics and prognostics. *ISME J* **11**, 841-852, doi:10.1038/ismej.2016.176 (2017).

53 Rossi, O. *et al.* Faecalibacterium prausnitzii A2-165 has a high capacity to induce IL-10 in human and murine dendritic cells and modulates T cell responses. *Sci Rep* **6**, 18507, doi:10.1038/srep18507 (2016).

54 Metwaly, A. *et al.* Integrated microbiota and metabolite profiles link Crohn's disease to sulfur metabolism. *Nat Commun* **11**, 4322, doi:10.1038/s41467-020-17956-1 (2020).

55 Strisciuglio, C. *et al.* Bifidobacteria Enhance Antigen Sampling and Processing by Dendritic Cells in Pediatric Inflammatory Bowel Disease. *Inflamm Bowel Dis* **21**, 1491-1498, doi:10.1097/mib.0000000000000389 (2015).

56 Tukenmez, U., Aktas, B., Aslim, B. & Yavuz, S. The relationship between the structural characteristics of lactobacilli-EPS and its ability to induce apoptosis in colon cancer cells in vitro. *Scientific Reports* **9**, 8268, doi:10.1038/s41598-019-44753-8 (2019).

57 Lai, C. Y. *et al.* Systematic review with meta-analysis: review of donor features, procedures and outcomes in 168 clinical studies of faecal microbiota transplantation. *Aliment Pharmacol Ther* **49**, 354-363, doi:10.1111/apt.15116 (2019).

58 Sivan, A. *et al.* Commensal Bifidobacterium promotes antitumor immunity and facilitates anti-PD-L1 efficacy. *Science* **350**, 1084-1089, doi:10.1126/science.aac4255 (2015).

59 Davar, D. *et al.* Fecal microbiota transplant overcomes resistance to anti-PD-1 therapy in melanoma patients. *Science* **371**, 595-602, doi:10.1126/science.abf3363 (2021).

60 Baruch, E. N. *et al.* Fecal microbiota transplant promotes response in immunotherapy-refractory melanoma patients. *Science* **371**, 602-609, doi:10.1126/science.abb5920 (2021).

61 Cammarota, G. *et al.* European consensus conference on faecal microbiota transplantation in clinical practice. *Gut* **66**, 569-580, doi:10.1136/gutjnl-2016-313017 (2017).

62 Collins, S. M., Kassam, Z. & Bercik, P. The adoptive transfer of behavioral phenotype via the intestinal microbiota: experimental evidence and clinical implications. *Curr Opin Microbiol* **16**, 240-245, doi:10.1016/j.mib.2013.06.004 (2013).

63 Murphy, C. L., Zulquernain, S. A. & Shanahan, F. Faecal Microbiota Transplantation (FMT) - classical bedside-to-bench clinical research. *QJM*, doi:10.1093/qjmed/hcz181 (2019).

Chapter 5 - Altered Skin and Gut Microbiome in Hidradenitis Suppurativa

The following chapter has been accepted for publication in the journal *Journal of Investigative Dermatology*

Authors:

Siobhan McCarthy*, Maurice Barrett*, Shivashini Kirthi, Paola Pellanda, Klara Vlckova, Anne-Marie Tobin, Michelle Murphy, Fergus Shanahan, Paul W O'Toole

*Joint first authorship: These authors contributed equally to this work.

Maurice Barrett contributed to this work in the following ways:

- Design of methodologies to collect and process samples.
- All bioinformatic analysis including sequence processing, compositional data analysis and statistical analysis.
- Data visualization i.e., construction of publication figures.
- Writing of over half of the manuscript.

5.1 Abstract

Hidradenitis suppurativa (HS) is a chronic inflammatory skin disease characterized by the formation of nodules, abscesses, and fistula at intertriginous sites. The skin-gut axis is an area of emerging research in inflammatory skin disease and is a potential contributory factor to the pathogenesis of HS. 59 patients with HS provided fecal samples, nasal and skin swabs of affected sites for analysis. 30 healthy controls provided fecal samples and 20 healthy controls provided nasal and skin swabs. We performed bacterial 16S rRNA gene amplicon sequencing on total DNA derived from the samples. Microbiome alpha diversity was significantly lower in the fecal, skin and nasal samples of individuals with HS which may be secondary to disease biology or related to antibiotic usage. *Ruminococcus gnavus* was more abundant in the fecal microbiome of individuals with HS, which is also reported in Crohn’s disease (CD), suggesting comorbidity due to shared gut microbiota alterations. *Finegoldia magna* was over-abundant in HS skin samples relative to healthy controls. It is possible local inflammation is driven by F. magna through promoting the formation of neutrophil extracellular traps (NET). These alterations in both the gut and skin microbiome in HS warrant further exploration, and therapeutic strategies including fecal microbiota transplant (FMT) or bacteriotherapy could be of benefit.

5.2 Introduction

Hidradenitis suppurativa (HS) is a chronic, debilitating, follicular skin disease presenting with deep-seated, painful, inflammatory nodules of the axillary, inframammary, inguinal, and anogenital regions[1]. These lesions can spontaneously rupture or coalesce to form painful deep dermal abscesses which often heal with scar formation. A population prevalence of up to 4% has been reported in the literature[1-3]. There is a female predominance, with onset often around puberty[1,4,5]. Smoking, obesity are recognised associations, and a genetic predisposition has been reported[4,6,7]. A broad range of comorbidities have been identified in patients with HS including spondyloarthropathy, metabolic syndrome and inflammatory bowel disease (IBD), particularly Crohn’s disease[5,8,9]. As well as significant morbidity, HS is associated with increased mortality, in particular due to cardiovascular events, with increased cancer risk also recorded[10,11]. Depression, anxiety and substance misuse is also common among its sufferers[12,13].

The cause of HS is incompletely understood, with follicular occlusion, dysregulated inflammatory response of cytokines such as tumour necrosis factor (TNF)-a, interleukin(IL)-1ß and IL-17, and an altered microbiota all thought to play a role[7,14-18].

HS and IBD share common manifestations characterised by sterile abscesses, scarring and sinus tract formation[19]. Similar inflammatory pathways are activated in Crohn’s disease and HS, with elevated production of the innate immune mediators IL-1, IL-6, IL-17, IL-23 and TNF-alpha[20-22]. Smoking and obesity are common associations, and HS and IBD respond to TNF-alpha inhibitor therapy[7,23-25].

Extensive research supports the role of the gut microbiota in IBD and other inflammatory conditions including rheumatoid arthritis, psoriasis and psoriatic arthritis[26-31]. Although the skin microbiota in HS is an area of expanding research, the gut microbiota or the 'gut-skin axis' in HS deserves greater consideration[14,32,33]. One study investigated *Faecalibacterium prausnitzii* and *Escherichia coli* levels in patients with psoriasis, concomitant psoriasis and IBD, HS, and concomitant HS and IBD[32]. Increased levels of *E. coli* and decreased levels of *F. prausnitzii* was noted in patients with psoriasis. A significant difference in abundance of *E. coli* or *F. prausnitzii* was not noted in patients with HS[32]. Since an altered gut microbiota has been associated with various pathophysiologies involving immune dysregulation, it may play a role in the development of HS.

In this study we tested for an association between microbiota alteration in the skin, nasal mucosa, and feces and HS. The microbiota across the various niches was compared to that of healthy controls.

5.3 Results

5.3.1 Descriptive statistics of the study population

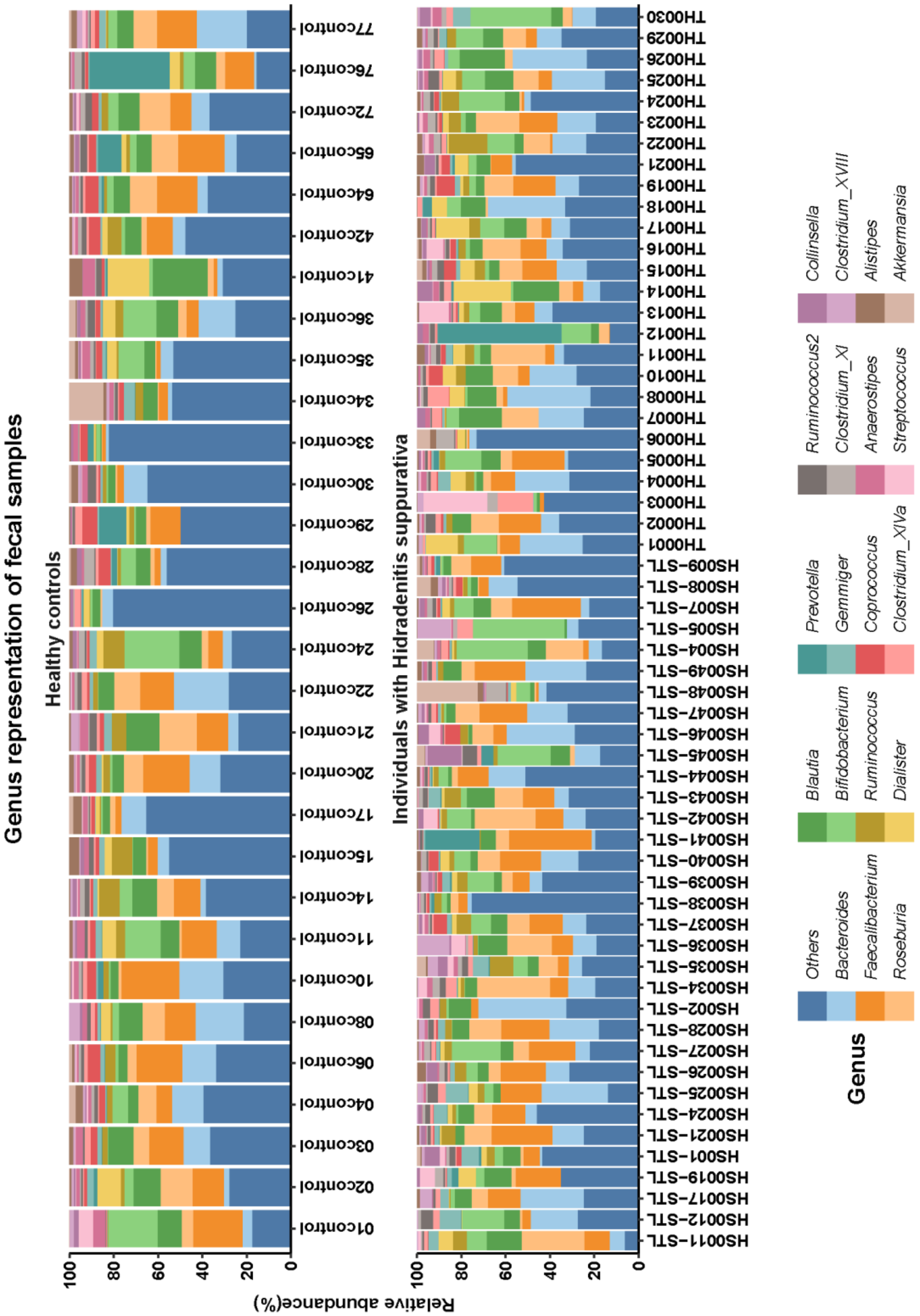
We collected 322 samples including fresh fecal samples, nasal swabs and skin swabs from 4 different locations including the axilla, inframammary area, buttock and groin (Table 1). 59 patients with HS were recruited providing fecal samples, skin swabs and nasal swabs. 30 healthy controls provided fecal samples (Planned 2:1 ratio) and 20 healthy controls provided skin and nasal swabs (Planned 3:1 ratio). Mean body mass index (BMI) in the HS group was 31.5, and 28.2 in the fecal control group and 28.06 in the skin control group. 4 (6.8%) patients in the HS group had a history of Crohn's disease. Of the 59 patients with HS, 18 (30.5%) were Hurley Stage 1 (abscess formation without sinus tracts and cicatrisation), 32 (54.2%) were Hurley Stage 2 (recurrent abscesses with tract formation and scars) and 9 (15.3%) were Hurley Stage 3 (multiple interconnected tracts and abscesses throughout an entire area).

	HS	Controls	Significance
Fecal			
N (patients)	59	30 (fecal) 20 (skin)	n/a
Gender (Female/Male)	45/14	20/10 (fecal) 15/5 (skin)	NS
Age (mean, range)	37, 21-62	38, 19-62 (fecal) 41, 24-68 (skin)	NS
BMI (mean, range)	31.5, 19.6-45.0	28.2, 18.7-46.3 (fecal) 28.06 20.3-40 (skin)	0.023
Crohn’s Disease (yes/no)	4/59	0/30 (fecal)	NS
TNF-α inhibitor therapy	9/59	0/30 (fecal)	
Nasal			
N (patients)	25	17	n/a
Gender (Female/Male)	22/3	12/5	NS
Age (mean, range)	41, 24-54	36, 24-68	NS
BMI (mean, range)	33, 20.4-45.0	29	NS
Axilla			
N (patients)	19	6	n/a
Gender (Female/Male)	16/3	2/4	NS
Age (mean, range)	39, 24-54	37, 29-52	NS
BMI (mean, range)	31, 19.6-45.0	29, 22.5-39.2	NS
Groin			
N (patients)	15	17	n/a
Gender (Female/Male)	12/3	12 /5	NS
Age (mean, range)	35, 24-52	40, 24-68	NS
BMI (mean, range)	30, 20.4-44.9	29, 20.9-40.0	NS
Breast			
N (patients)	5	13	n/a
Gender (Female/Male)	5/0	13/0	NS
Age (mean, range)	39, 25-52	39, 24-54	NS
BMI (mean, range)	30.5, 19.6-38.0	28.3, 23.6-40.0	NS
Buttock			
N (patients)	4	19	n/a
Gender (Female/Male)	2/2	14/5	NS
Age (mean, range)	36, 28-39	40, 24-68	NS
BMI (mean, range)	31 30.0-31.6	28, 20.9-40.0	NS

Table 1. Subject characteristics: HS subjects and healthy controls. Comparison of variables between HS cohort and healthy controls. Wilcoxon signed-rank test or χ2 statistic was used to determine significance.

5.3.2 Overall structure of the fecal microbiota is altered in HS

We examined the microbiome composition of 59 and 30 fecal specimens from individuals with HS and healthy controls respectively (eFigure 1). Ecological metrics showed a difference between the microbiome of individuals with HS and healthy controls (Figure 1). Alpha-diversity, a marker of microbial species richness or variation within a sample, was significantly lower in individuals with HS (Figure 1A). This reduction was also observed for four other metrics of alpha-diversity including Shannon and phylogenetic diversity (eFigure 2). A microbiome separation in beta-diversity, (a comparison of global microbial composition in all the samples), between the HS and healthy controls was observed across all metrics tested (Figure 1B) (eFigure 3), noting also less clustering within the HS samples. Hierarchical clustering replicated and reinforced this separation as seen by the presence of a cluster composed exclusively of patients with HS (Figure 1C).



eFigure 1: Taxonomic representation within faecal specimens. Bar plots displaying the relative abundance of genera within faecal samples. Genera with a relative abundance of less 1% across all samples grouped into ‘others’ with sequences not classified at the genus level

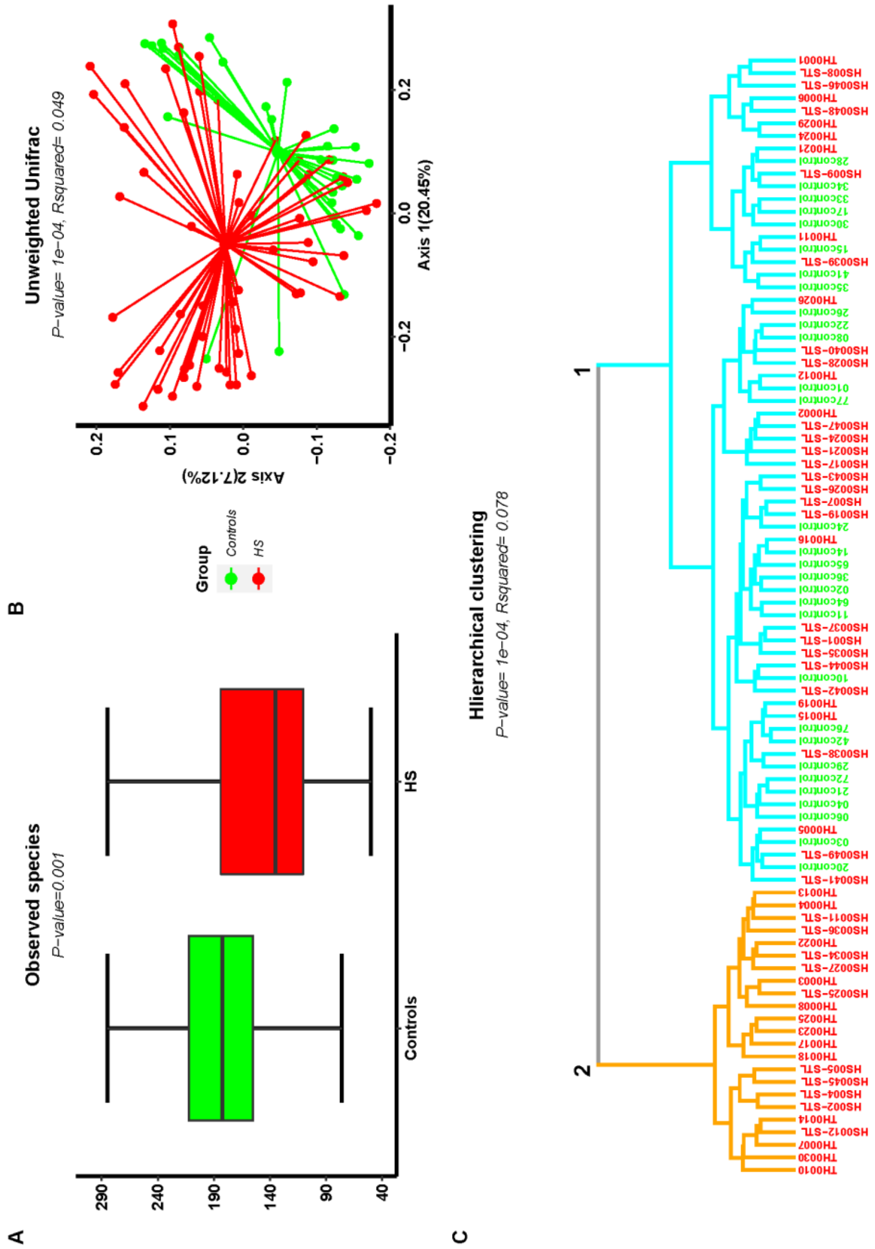
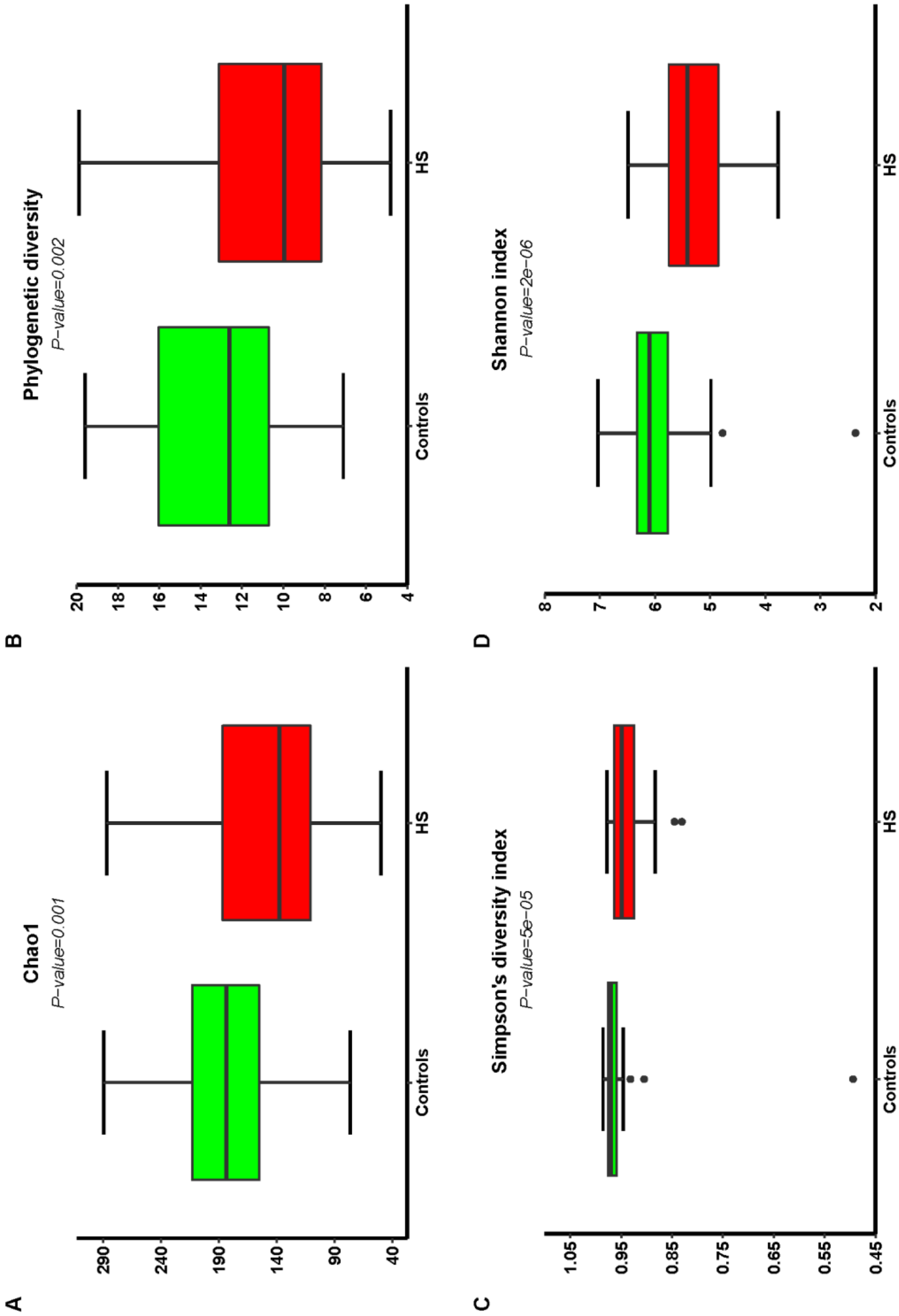
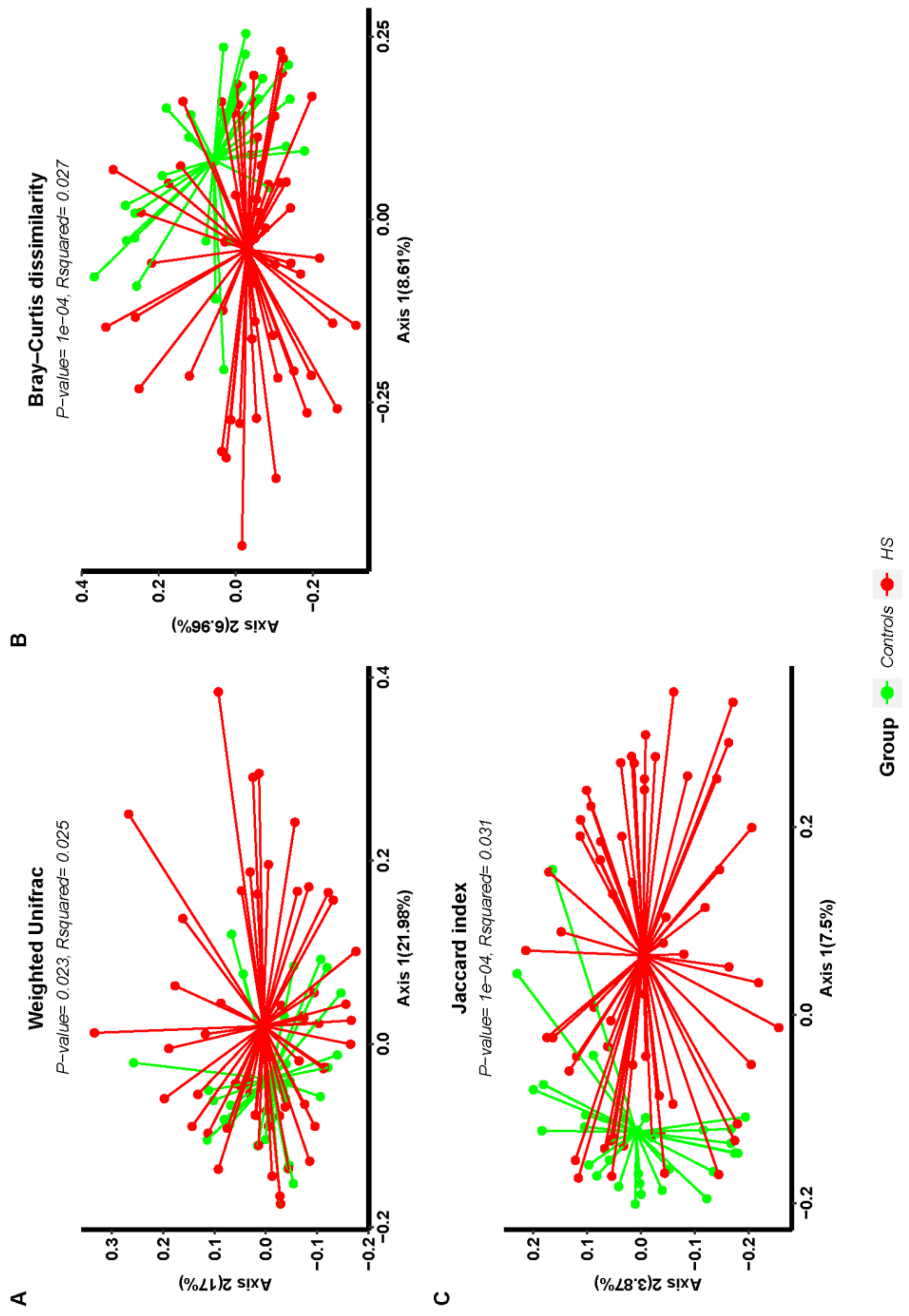


Figure 1: Ecological overview of fecal microbiome data. (A) Alpha-diversity. Boxplot comparing alpha diversity (observed species index) between individuals with HS versus healthy controls. Wilcoxon signed-rank test was used to calculate p-values **(B) Beta diversity**. Respective distance between samples based on their microbiome composition was calculated using unweighted Unifrac distance. Principal Coordinates Analysis (PCoA) was performed to obtain the coordinates of the first two PCoA and were plotted. Statistical testing was performed using Permutational Multivariate Analysis of Variance (PERMANOVA). **(C) Hierarchical clustering.**The closeness between subjects based of microbiome composition was calculated using Spearman's rank correlation coefficient. Hierarchical clustering was performed using the Ward2 method and the results plotted as a dendrogram. . Statistical testing was performed using Permutational Multivariate Analysis of Variance (PERMANOVA).

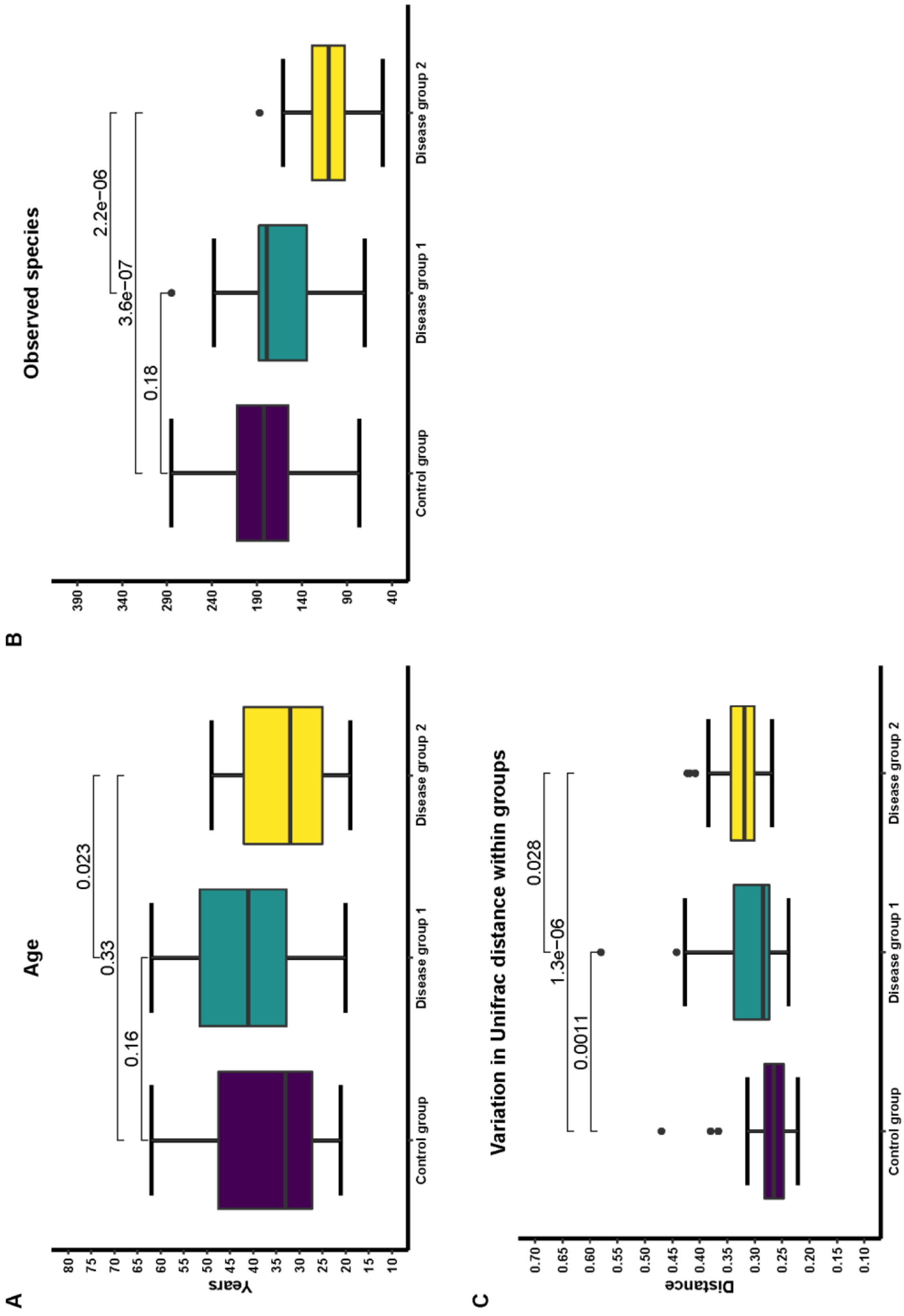


eFigure 2: Bar plots of alpha diversity metrics regarding faecal samples. (A) Chao1. (B) Phylogenetic diversity. (C) Simpson's Diversity Index. (D) Shannon index. Wilcoxon signed-rank test was used to calculate p-values.



eFigure 3: PCoA representing Beta diversity metrics regarding faecal samples. (A) Weighted Unifrac. (B) Bray–Curtis Dissimilarity. (C) Jaccard index. Statistical testing was performed using Permutational Multivariate Analysis of Variance (PERMANOVA).

We further investigated this cohort by grouping the individuals informed by the hierarchical clustering that is, control group (control samples, cyan branch, branch No.1), disease group 1 (cyan branch, branched No.1) and disease group 2 (orange branch, branched No.2). We found that disease group 2 was composed of significantly younger subjects than disease group 1 but not the controls (eFigure 4A). Alpha diversity was lower in disease group 2 compared to the disease group 1 and the control group (eFigure 4B). There was greater within-group microbiome variation, evidenced by higher levels of Unifrac distance between samples, within disease group 2 compared to the other two groups (eFigure 4C).



eFigure 4: Difference in groups informed by clustering. (A) Bar plot of age differences (B) Bar plots of observed species. (C) Differences in the Variation in Unifrac distance within groups. Wilcoxon signed-rank test was used to calculate p-values.

5.3.3 Differentially abundant ASVs in the fecal microbiome

Microbial amplicon sequencing data can be rationalised in terms of amplicon sequence variants (ASVs) which allows the data to be resolved down to single-nucleotide difference[34]. A number of ASVs were found to be differentially abundant between the fecal microbiome of patients with HS and healthy controls (Figure 2A). With regard to log2 fold differences, the ASVs assigned to the taxa *Ruminococcus callidus* and *Eubacterium rectale* were the most enriched in individuals with HS relative to healthy controls. However, with respect to proportional abundance, the greatest difference was detected in ASVs assigned to the taxa *Streptococcus spp*. (an average relative abundance of 0.19% in the control cohort versus 0.95% in the HS cohort) and *Ruminococcus gnavus* (average relative abundance values of 0.01% in the control cohort versus 0.7% in the HS cohort).

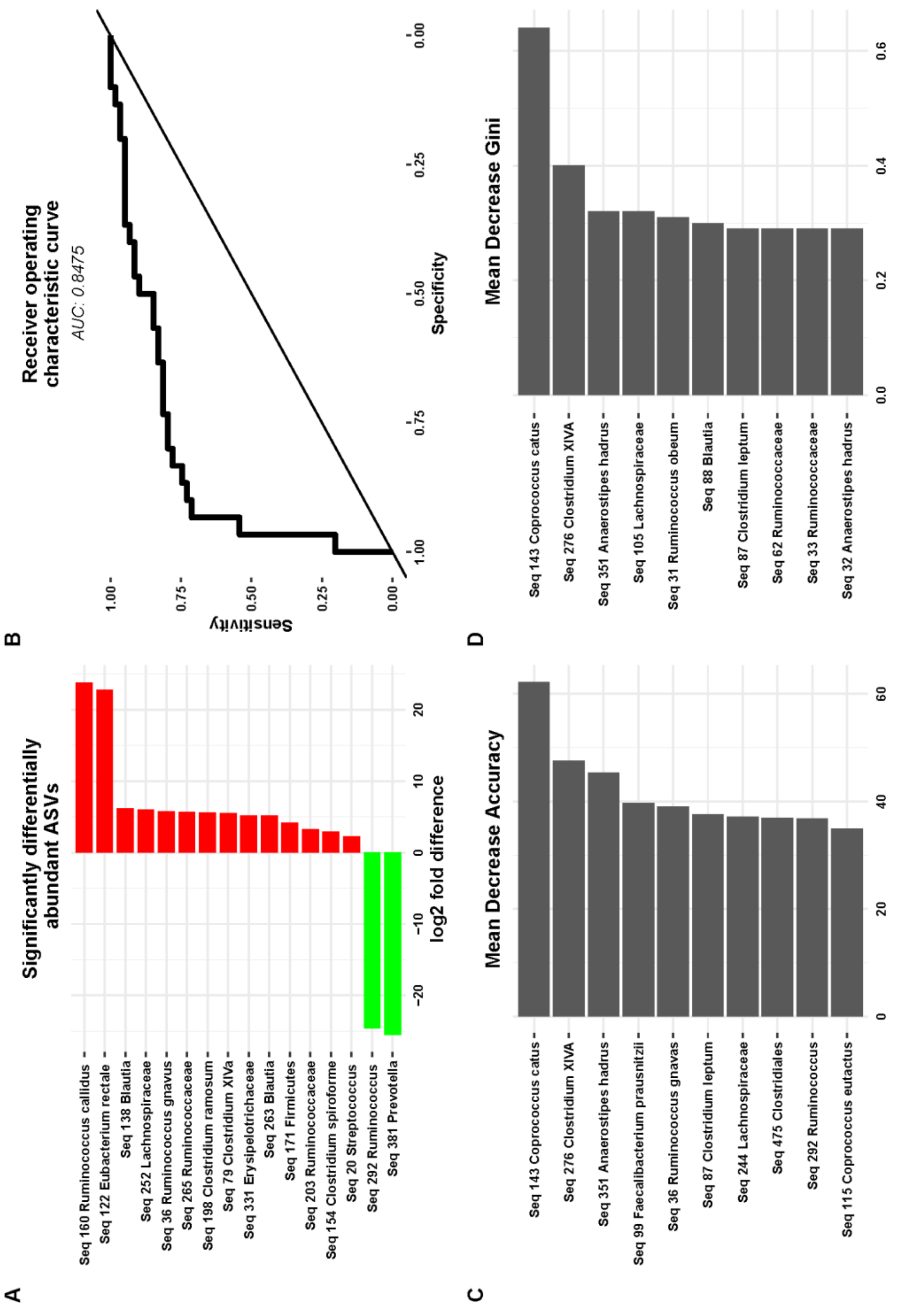
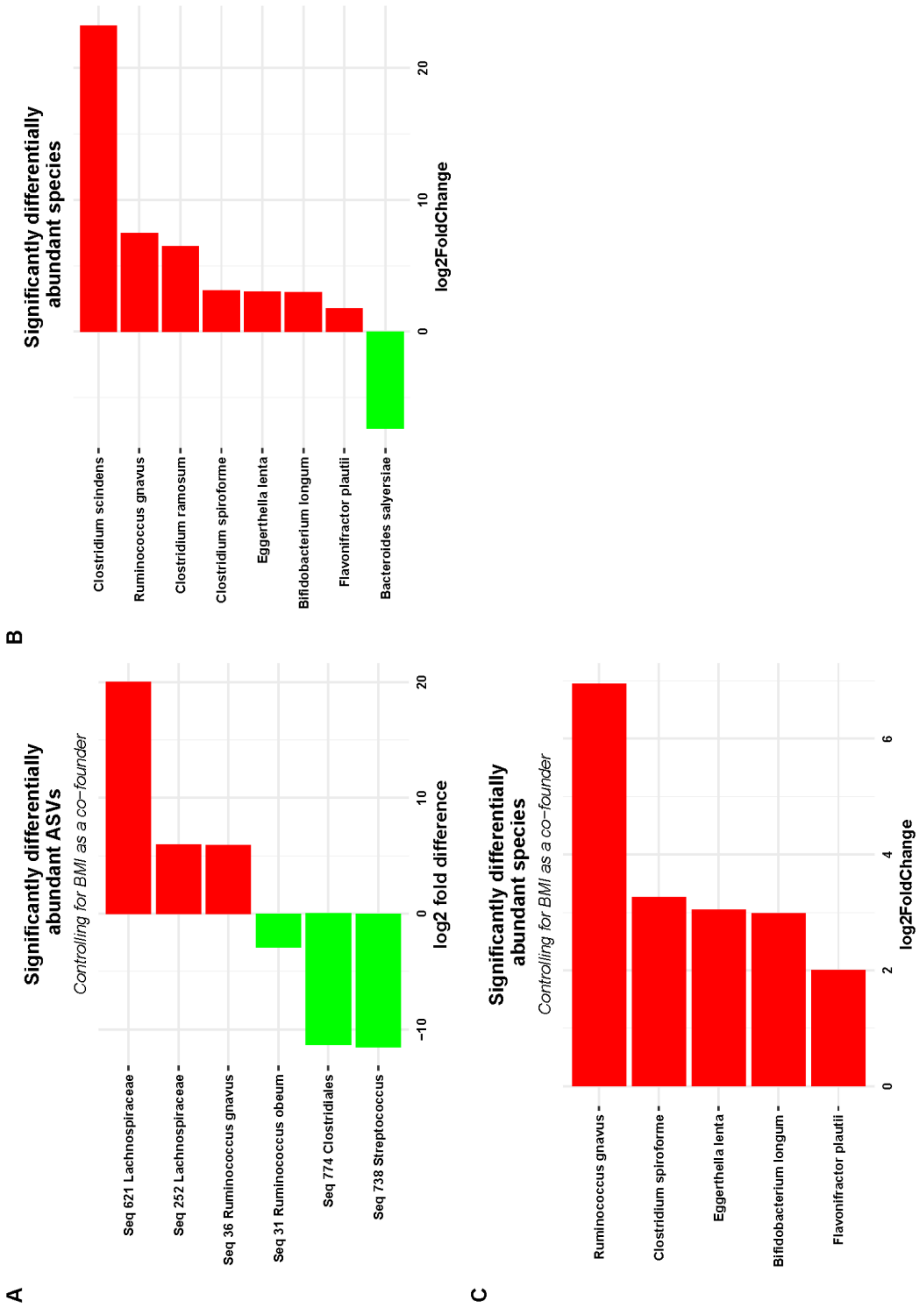


Figure 2: Differentially abundant ASVs and machine learning classification. (A) Bar plot of differential abundant ASVs as expressed by log fold difference. Negative values indicates overrepresented in controls. Positive values indicates overrepresented in individuals with HS **(B)** Receiver operating characteristic curves (ROC) **(C,D)** Top discriminatory ASVs with regard to discriminating the HS subjects from healthy controls. **(C)** Mean Decrease Accuracy. **(D)** Mean Decrease Gini.

As BMI was significantly different between the groups, the DESeq2 model was re-run to adjust for BMI (eFigure 5). The differential over-abundance of *Ruminococcus gnavus* remained statistically significant in the HS cohort. However, an ASV assigned to *Ruminococcus obeum* was revealed to be depleted in individuals with HS. ASVs were also collapsed to the species level and differential abundance of species determined with and without BMI as a confounder (eFigure 5). *R. gnavus* was retained as being significantly over-abundant in the HS cohort in both analyses.



eFigure 5: Bar plots of differential abundant ASVs and species in fecal samples. (A) Significantly differentially abundant ASVs with BMI integrated into the model. (B) Significantly differentially abundant species. (C) Significantly differentially abundant species with BMI integrated into the model. DESeq2 used to for statistical analysis. ASVs/species enriched in HS samples in red. ASVs/species enriched in control samples in green.

Antibiotic usage for the previous year was recorded in the HS cohort. There were no significant ASVs that were differently abundant between those who had received antibiotic therapy in the last year and those who did not.

5.3.4 Machine learning identification of HS-related microbiota members

The machine learning classifier random forest (RF) was employed to test if ASVs could discriminate the HS patients from the healthy control cohort. The RF classifier performed reasonably with an area under the curve (AUC) of 0.8458 (Figure 2B). A number of ASVs identified as discriminatory (i.e. as contributing to the RF model) were taxonomically assignable to butyrate-producing bacterial species including *Faecalibacterium prausnitzii*, *Coprococcus eutactus*, *Coprococcus catus* and *Anaerostipes hadrus* (Figure 2C). A number of ASVs that we had identified using the DESeq2 model as being differentially abundant were also identified including *Ruminococcus gnavus* and *Ruminococcus obeum*.

5.3.5 Changes in predicted metabolic function of the fecal microbiota

Metagenomic functionality was inferred using the algorithm PICRUSt2, which is based on the metabolic pathways of reference microbiota data to which a test-set of 16S data is compared. Several metabolic pathways were thus predicted to be differentially abundant between HS and control metagenomes (Figure 3A).

Metabolic pathways for D-glucarate degradation and D-galactarate degradation, which are associated with a poor prognosis in CD, were overrepresented in individuals with HS relative to healthy controls[35].

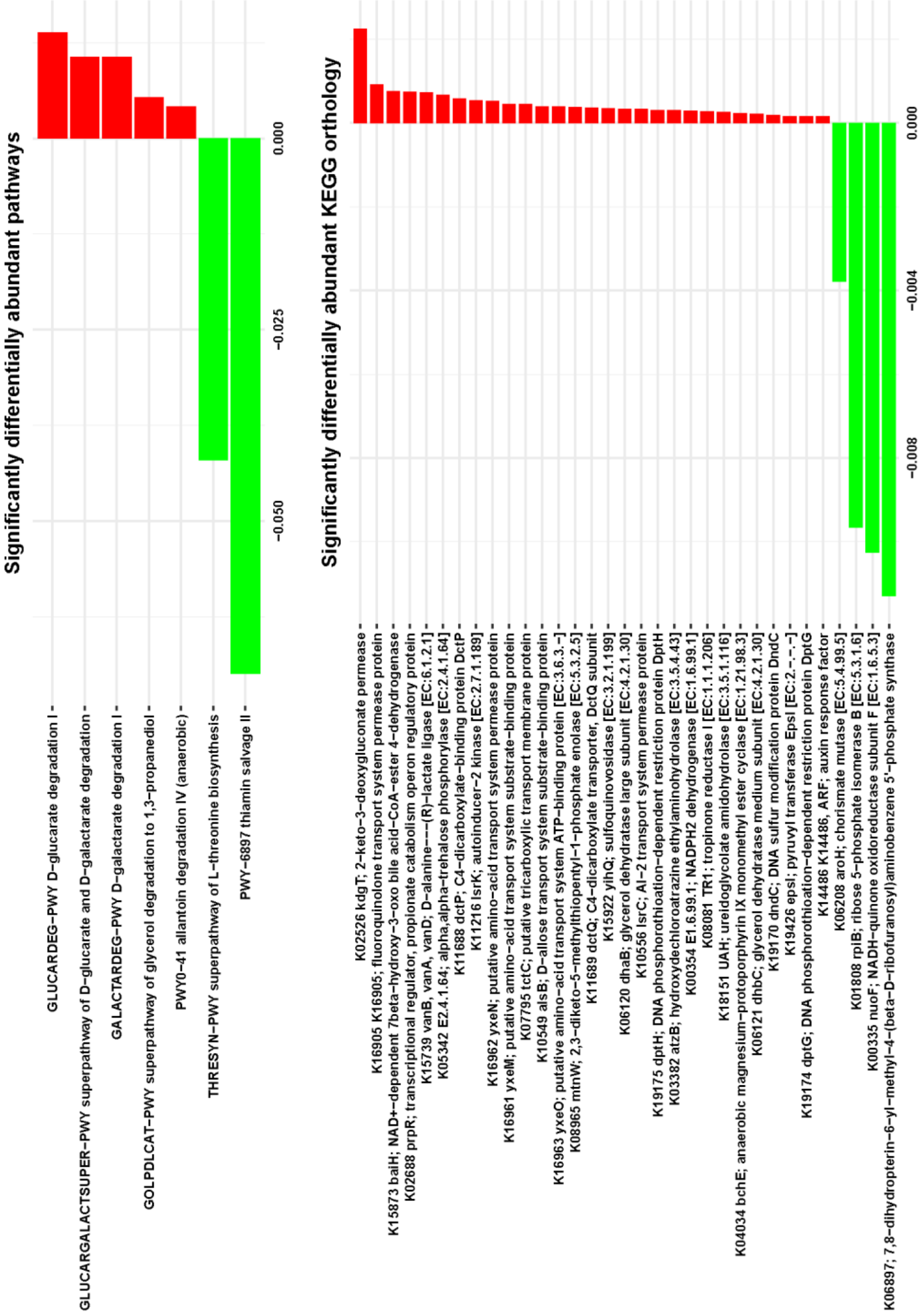
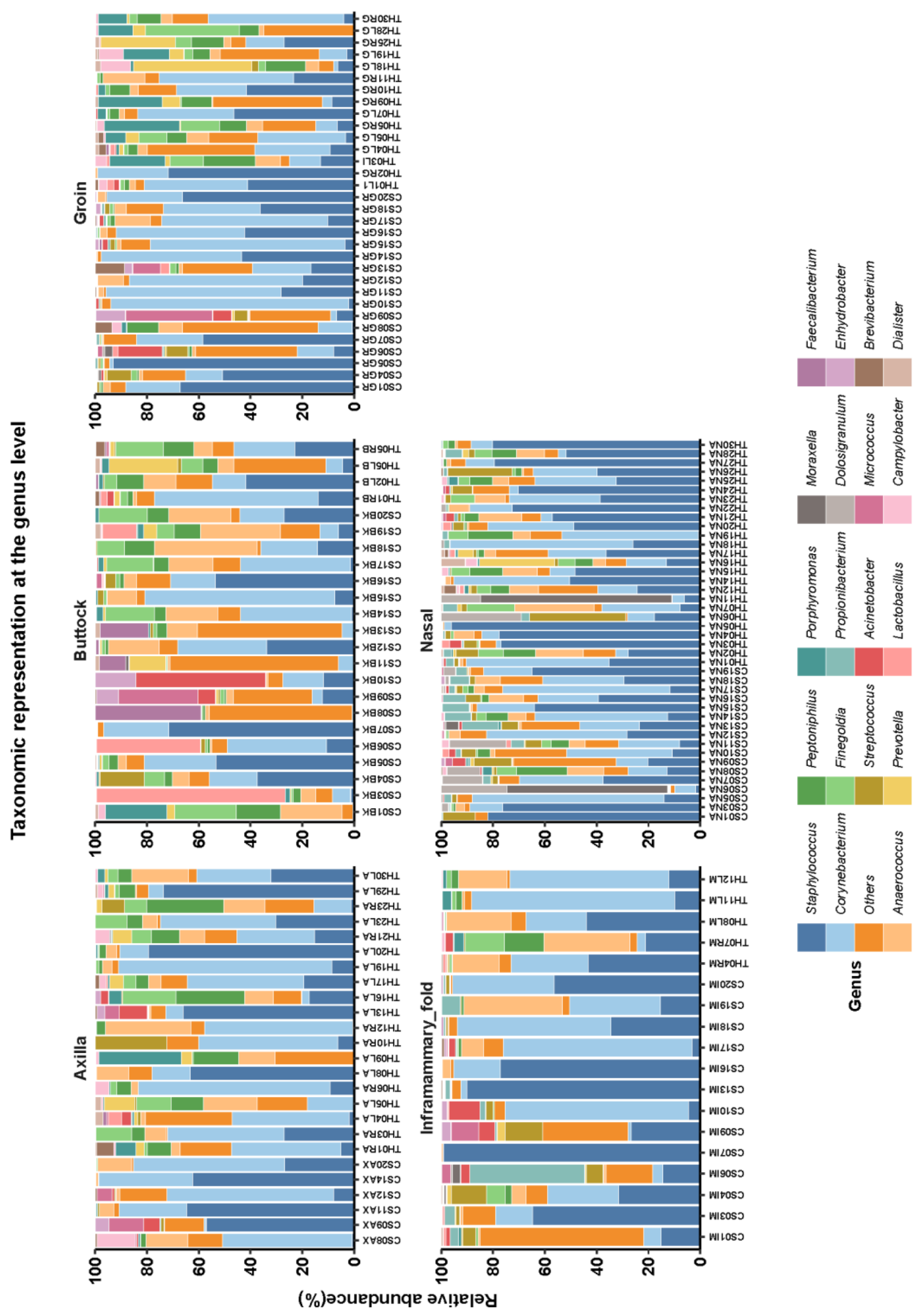


Figure 3: Differentially abundant metabolic pathways and KEGG orthologs. (A) Bar plot of differential abundant MetaCyc as expressed by difference in mean proportional abundance. MetaCyc pathways enriched in HS samples in red. MetaCyc pathways enriched in control samples in green. **(B)** Bar plot of differential abundant KOs as expressed by difference in mean proportional abundance. . KOs enriched in HS samples in red. KOs enriched in control samples in green. Wilcoxon signed-rank test was used to calculate p-values.

5.3.6 Ecological structure is altered in nasal and skin microbiome

The overall microbiome composition of nasal and skin samples was typical of what has been previously described, that is, mainly composed of the genera Staphylococcus and Corynebacterium (efigure 6). Both nasal and skins swabs showed a reduction in alpha-diversity in the HS cohort (Figure 4)[36]. However, only nasal swabs reached statistical significant decrease. This was also true for other alpha-diversity metrics including Observed species and Chao1 but not for Simpson or Shannon indices (eFigure 7). The number of subjects that contributed samples to some sites was low, with a low control number, thus the statistical power was reduced and significance difficult to capture. There was a statistically significant separation in beta-diversity with respect to axilla, groin, and nasal microbiota datasets (eFigure 8), showing that different microbiome communities are present at these body sites.



eFigure 6: Taxonomic representation within nasal and skin specimens. Bar plots displaying the relative abundance of genera within nasal and skin samples. Genera with a relative abundance of less than 0.5% across all samples grouped into ‘others’ with sequences not classified at the genus level.

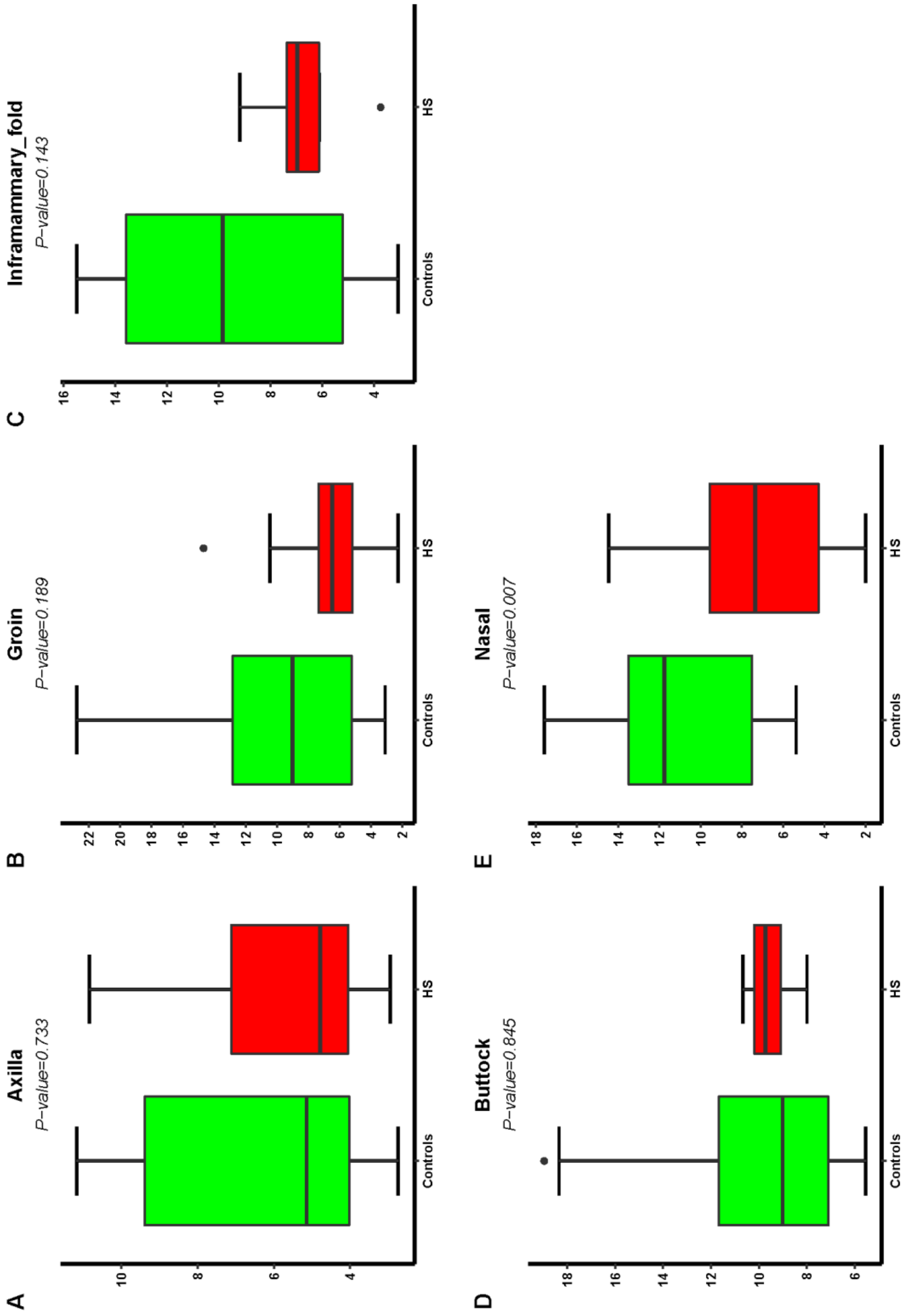
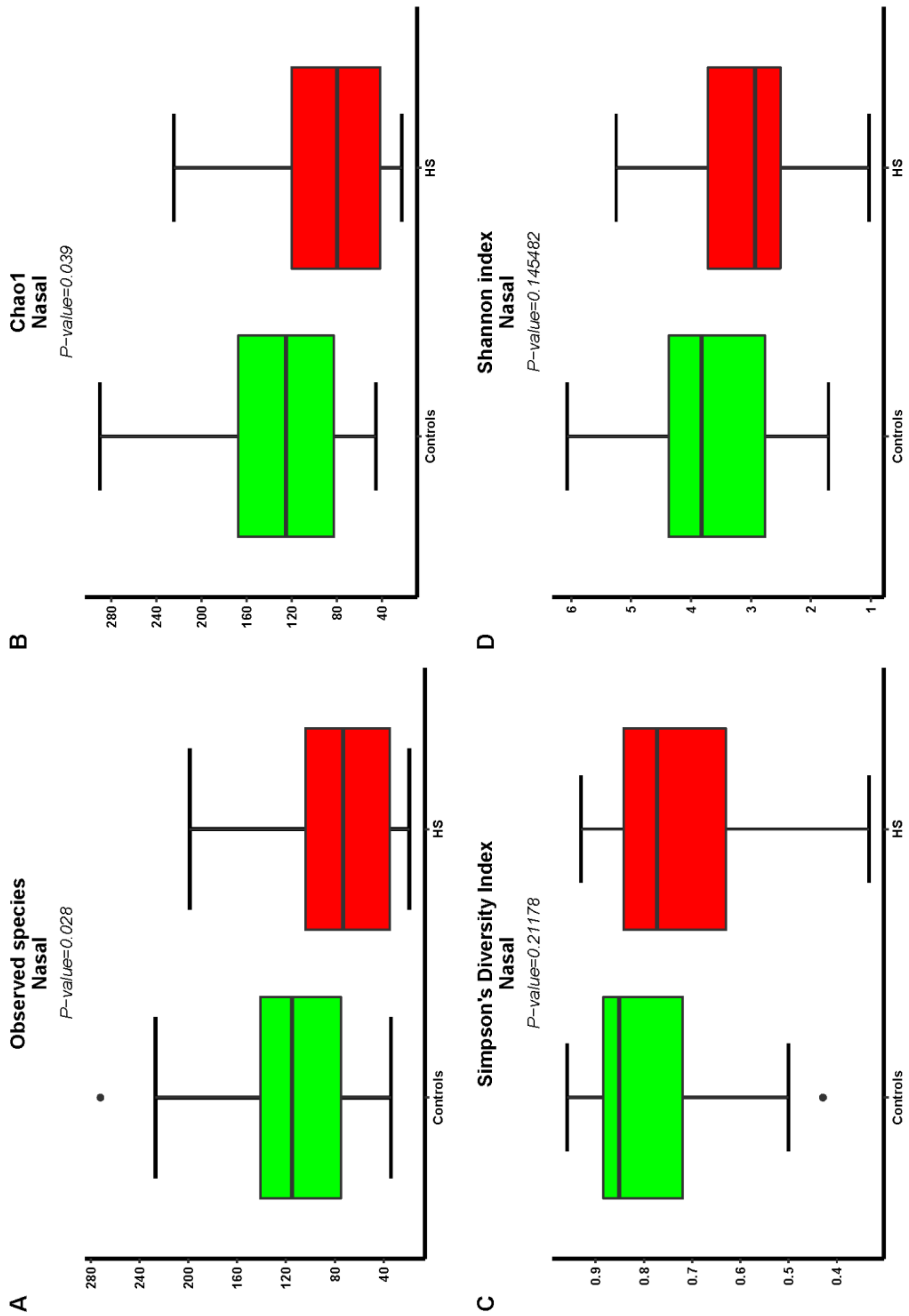
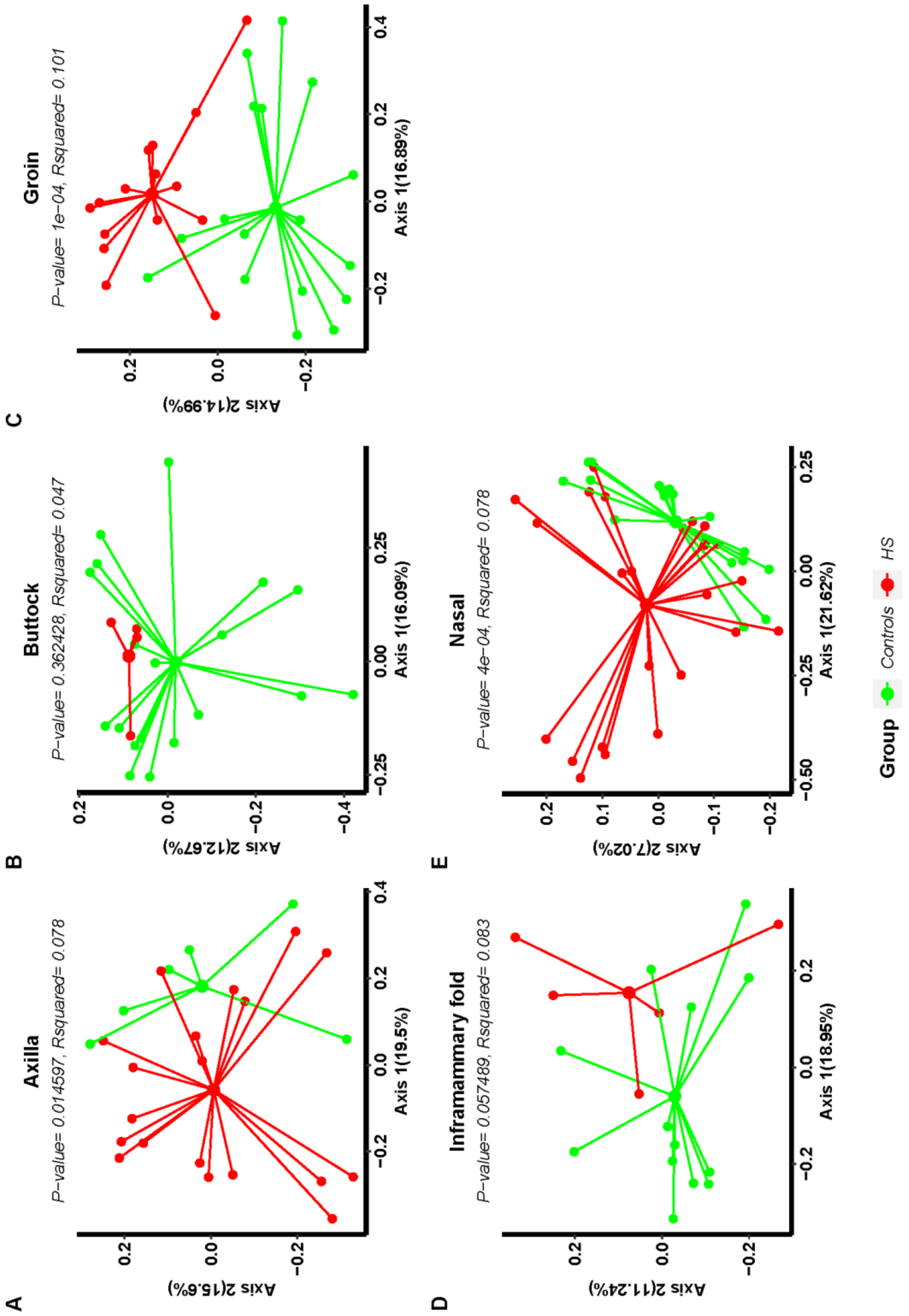


Figure 4: Alpha-diversity comparisons across nasal and skin. Bar plots of alpha-diversity (Phylogenetic diversity) comparing healthy controls versus individuals with HS. Wilcoxon signed-rank test was used to calculate p-values.



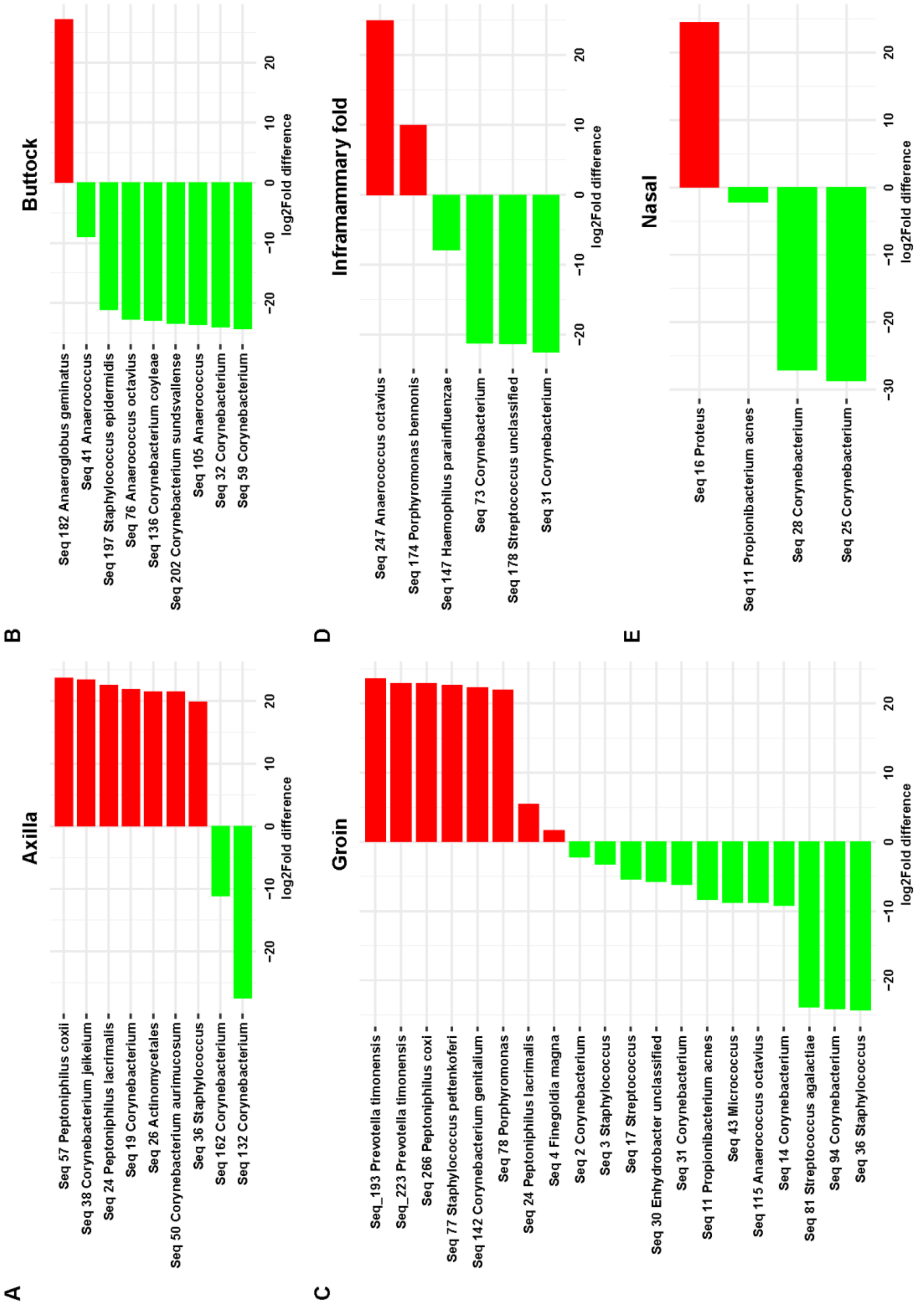
eFigure 7: Bar plots of alpha diversity metrics regarding nasal samples. (A) Chao1. (B) Phylogenetic diversity. (C) Simpson's Diversity Index. (D) Shannon index. Wilcoxon signed-rank test was used to calculate p-values.



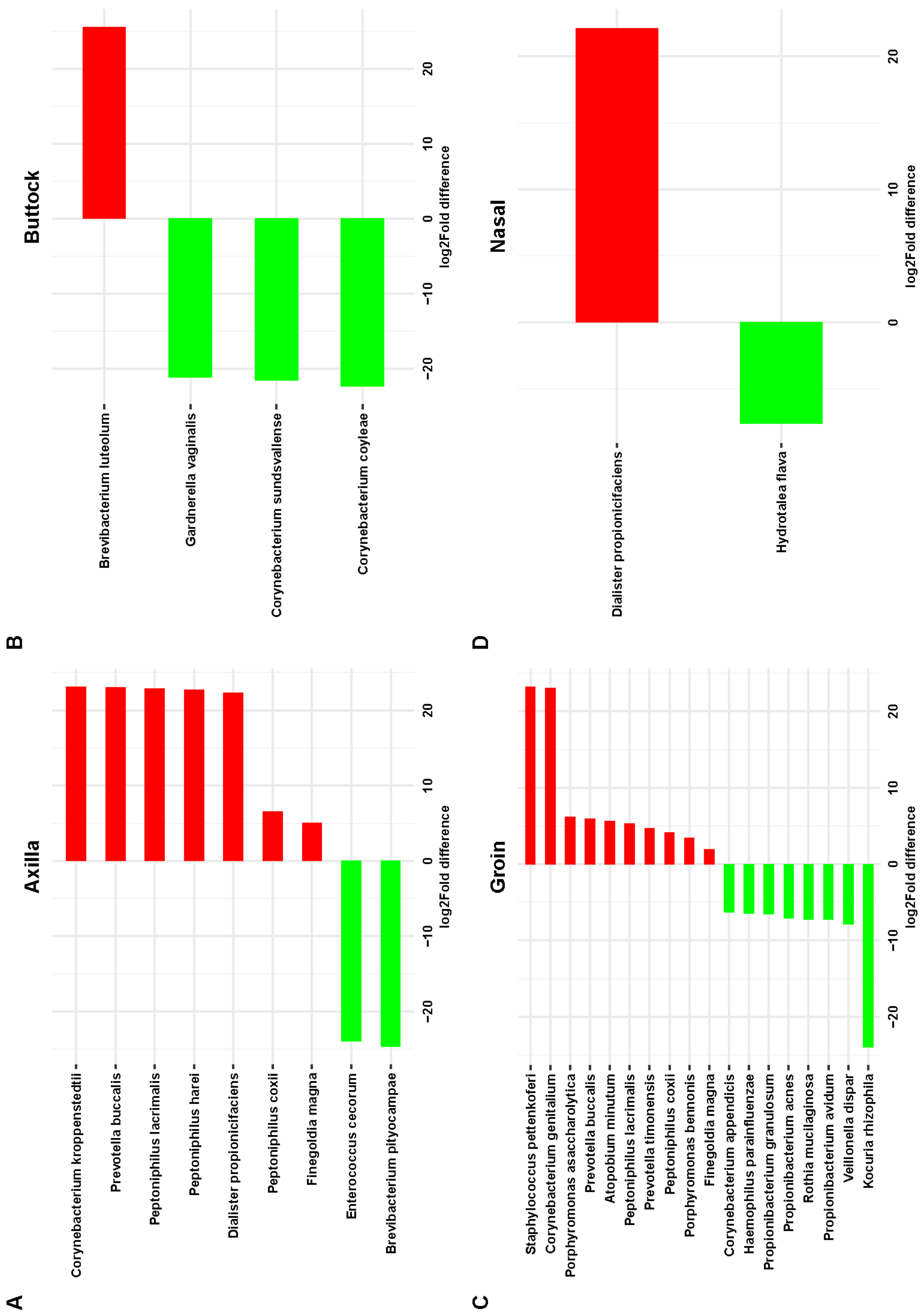
eFigure 8: Beta-diversity comparisons of nasal and skin microbiome. Principal Coordinates Analysis representation of Beta diversity (Unweighted Unifrac) between individuals with HS versus healthy controls. Statistical testing was performed using Permutational Multivariate Analysis of Variance.

5.3.6 Differentially abundant ASVs and metabolic pathways in the nasal and skin microbiome of HS patients

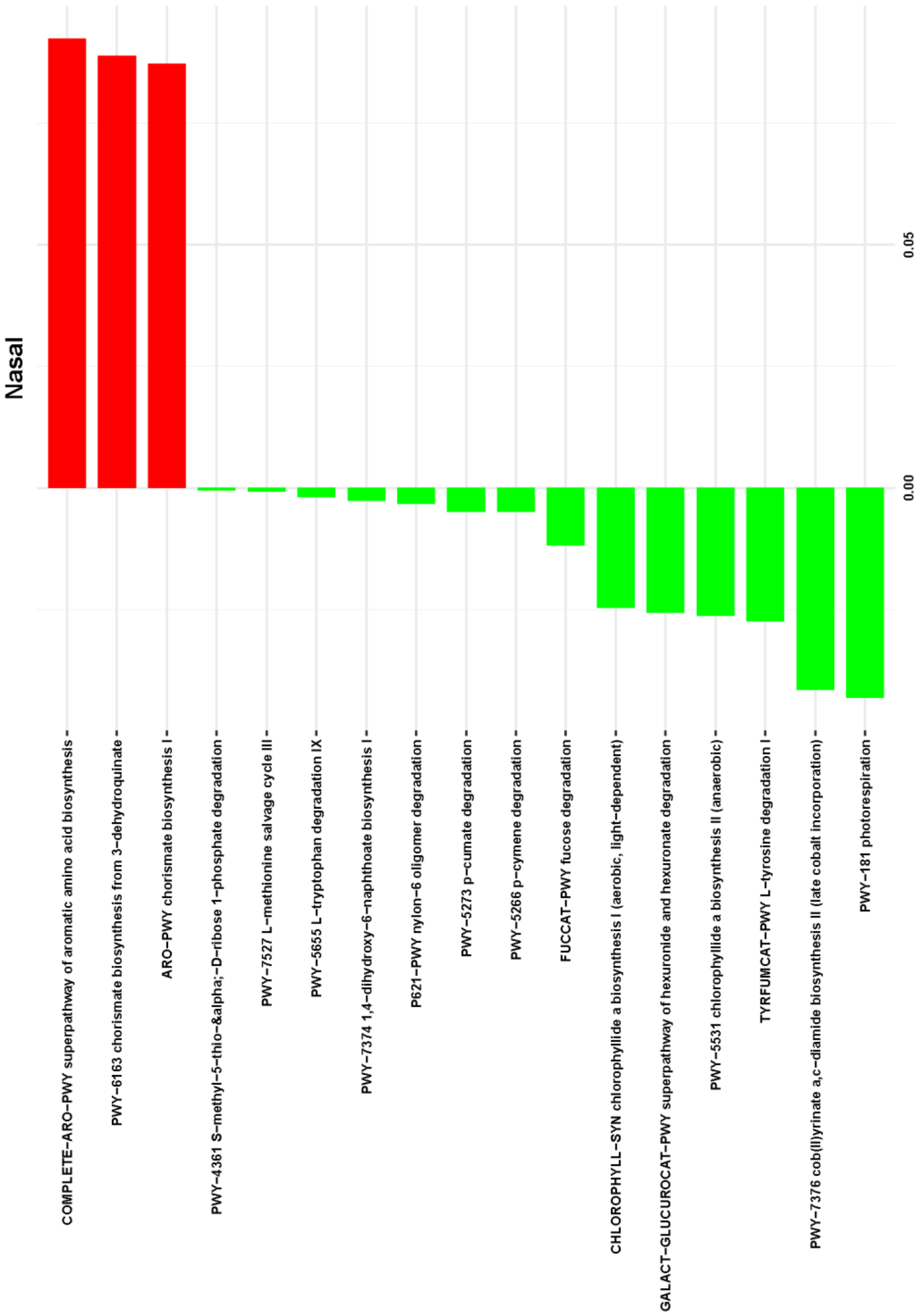
We identified differentially abundant ASVs in the nasal microbiome and all skin sites studied (eFigure 9). ASVs were collapsed to the species level and differential species abundance delineated (eFigure 9). An ASV assigned to *Finegoldia magna* had a significantly higher abundance at the groin site in individuals with HS relative to healthy controls. Furthermore, at the species level, *F. magna* was more abundant in HS relative to healthy controls in groin and axilla samples (eFigure 10). Significantly differentially abundant pathways were found in relation to the nasal microbiome and one pathway in the groin microbiome (eFigure 11).



eFigure 9: Bar plots of differential abundant ASVs across nasal and skin swab samples. DESeq2 used to for statistical analysis. ASVs enriched in HS samples in red. ASVs enriched in control samples in green.



eFigure 10: Bar plots of differential species across nasal and skin sites. DESeq2 used to for statistical analysis. Species enriched in HS samples in red. Species enriched in control samples in green.



eFigure 11: Differential MetaCyc pathways across skin sites. Nasal. Wilcoxon signed-rank test was used to calculate p-values. MetaCyc pathways enriched in HS samples in red. MetaCyc pathways enriched in control samples in green.

5.4. Discussion

We identified a number of differences in microbiome configuration in fecal and swab samples from across a number of body sites in individuals with HS compared to healthy controls. Alpha diversity was lower in subjects with HS across most of these sites suggesting a reduction in the richness of the gut and skin microbiota compared to controls. Decreases in alpha diversity in skin, and nasal microbiota have been previously reported in atopic dermatitis, with conflicting results for the gut microbiota[37-40]. Alpha-diversity has also been observed to be lower in skin samples from individuals with psoriasis, with even more variable results in the gut[41-43].

5.4.1 Gut Microbiome in HS

Elevated levels of *Ruminococcus gnavus* and *Clostridium ramosum* were among the greatest differences in relative abundance between patients with HS and healthy control microbiomes in this study. *R. gnavus* has been consistently found to be overrepresented in subjects with Crohn’s Disease and has also been associated with spondyloarthritis, and irritable bowel syndrome[28,44-49]. *R. gnavus* has also been linked to development of eczema and other allergic diseases in infants, thought to be due to its effect on the host immune system development[37,49]. A mechanistic role of *R. gnavus* in Crohn’s disease has been experimentally supported namely the production of a potent proinflammatory polysaccharide which induces the production TNF-α via interacting with the toll-like receptor 4 (TLR4) of innate immune cells such as dendritic cells[50]. The production of this polysaccharide could be a contributor to the pathogenesis of HS. It is possible that the diseases that are comorbid with HS have a common aetiology, due, in part, to the activity of *R. gnavus*. The abundance of *C. ramosum* has also been reported to be increased in Crohn's disease and obese individuals[26,51]. In a previous study, *R. obeum* was strongly enriched in controls relative to individuals with IBD, as seen in this study[52].

We found that pathways related to galactarate and glucarate degradation were more abundant in the fecal microbiome from individuals with HS. These metabolic pathways have also been implicated in Crohn’s disease clinical outcome and could be linked to systemic inflammation. In a recent paired whole exome shotgun metagenomics study comparing individuals with IBD and healthy controls, immune-related gene CABIN1 was associated with an increase of D-glucarate degradation[53]. Both D-glucarate degradation and D-galactarate degradation were overrepresented in individuals with Crohn’s Disease with a poor prognosis relative to those with a good prognosis[35]. Antibiotics are known to induce a host-mediated elevation in the levels of galactarate and glucarate in the gut and increased expression of microbial genes responsible for galactarate and glucarate degradation may be a response to this[54]. Mouse model studies demonstrated that antibiotic treatment leads to an increase in galactarate and glucarate through increased expression of inducible nitric oxide synthase (iNOS)[54].

5.4.2 Skin Microbiome

In a previous study, Ring *et al* examined the difference in the skin microbiome between subjects with hidradenitis suppurativa and healthy controls using skin biopsies[55]. Our analysis corroborates and extends the number of taxa found to differentially abundant. In particular there is agreement that *Peptoniphilus lacrimalis, Finegoldia magna*, *Peptoniphilus coxii*, *Anaerococcus murdochii, and Anaerococcus obesiensis* are more abundant in HS, with a higher abundance of *Cutibacterium acnes* in healthy controls. Colonisation and proliferation of certain strains of *C. acnes* are thought to play an important role in the pathogenesis of acne vulgaris[56,57]. The depletion of *C. acnes* in individuals with HS suggests that it does not play a similar mechanistic role in HS as it does in acne. However, a decrease in *C. acnes* may alter the microbial ecological of skin in a manner that promotes HS pathogenesis. Alpha diversity was reduced in nasal and groin samples. This is also reflected in findings for atopic dermatitis; however, the corresponding increase in *Staphylococcus aureus* typically seen in atopic dermatitis was not detected in these patients with HS[39,58]. Similarly, in psoriasis a reduction in alpha diversity is seen compared to healthy controls, featuring elevated *Streptococcus* and reduction in *Propionibacterium* that was not seen in this cohort with HS[43]. Higher numbers of bacteria with pathogenic capability namely *F. magna* was noted in the current study. *F. magna* has been shown to have immune modulating activities; in particular it can promote the formation of neutrophil extracellular traps (NET). These NETs feature prominently in HS lesions and their abundance is correlated with disease severity, as measured by Hurley staging[59]. Thus *F. magna* may contribute to HS disease biology by stimulating NET formation. *F. magna* has also been shown to activate mast cells and basophils which in turn produce proinflammatory histamine and cytokines[60,61]. In a recent study *F. magna* was demonstrated to activate proinflammatory neutrophils mediated by virulence factors protein L and FAF (*F. magna* adhesion factor)[62].

5.4.3 Potential impact

There is a lack of high-quality evidence for the best treatment options in HS[1,6]. A multidisciplinary approach with a combination of medical and surgical treatment is often needed, combined with lifestyle measures: smoking cessation and weight loss[1,6]. Antibiotics remain the initial treatment for most patients, with TNF-alpha inhibitors in those who fail to respond[1,6]. The use of antibiotics in HS, for their anti-inflammatory rather than anti-microbial effect, may play a role in the reduction in alpha-diversity seen in this study; however we detected no significant difference in ASVs in those who received antibiotics in the preceding year compared to those who did not. Microbiota based therapies may have potential benefits in HS, in particular targeted microbial supplementation to increase diversity and richness. Furthermore, the selective depletion of certain microbes such as *F. magna* and *R. gnavus*, which may play a pathogenic role, may prove another target for evolving therapies.

We have characterised the gut and skin microbiota in patients with HS compared to healthy controls. We have provided evidence for a possible microbial link between IBD and HS, with *R. gnavus* abundant in both conditions. The identification of particular taxa that may contribute to HS pathogenesis, such as *R. gnavus* and *F. magna*, could inform future microbiota-based therapeutic strategies.

5.5. Material and Me thods

5.5.1 Study Population

Adult patients with a confirmed clinical diagnosis of hidradenitis suppurativa made by a consultant dermatologist in two tertiary referral centres in Ireland were invited to participate in the study. Ethical approval was obtained (Cork Research Ethics Committee and Tallaght University Hospital Ethics Committee). Exclusion criteria included topical or oral antibiotic usage in the preceding four weeks. Data including age, gender, smoking status, weight, height, body mass index, presence of co-morbidities and severity of disease (Hurley Score) were recorded[63]. Healthy adult controls were recruited from the general population, and were age and gender matched.

5.5.2 Sample collection

Fresh (<24 hours) fecal samples were provided by patients and controls and stored at -80°C prior to microbial DNA extraction. In patients with HS, skin swabs (DNA-free) were taken from affected sites (axilla, inframammary, inguinal and perineal/buttocks) using a buffer solution with firm swabbing for 30-60 seconds. Affected sites varied by number and location in HS patients. Participants did not bathe for at least 24-48 hours prior to taking swabs and were asked to not apply antiperspirants or emollients on the skin in that time. Skin swabs were taken from corresponding sites in controls (axilla, inframammary, inguinal and perineal/buttocks) using the same technique. Nasal swabs were also taken from both groups.

5.5.3 Microbial DNA extraction

Microbial DNA was extracted from stool samples using the repeated bead beating method as previously described, with some modifications.(Ghosh et al., 2020) Nasal and skin swabs were extracted using QIAamp UCP Pathogen Mini Kit (Qiagen, Hilden, Germany) as per manufacturer's instructions.

5.5.4 Library Preparation and 16S rRNA gene sequencing

Total community DNA extracted from clinical samples underwent 16S rRNA gene PCR. The 16S rRNA gene was amplified using primers for the V3-V4 region; forward, TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCTACGGGNGGCWGCAG-3′ and reverse, 5′-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGACTACHVGGGTATCTAATCC-3'.(Klindworth et al., 2013)

Fecal microbial genomic DNA was amplified using Phusion High-Fidelity DNA Polymerase (Thermo Scientific, Massachusetts, USA) with the PCR thermocycler protocol as follows: Initiation step of 98 °C for 3 min followed by 25 cycles of 98 °C for 30 s, 55 °C for 60 s, and 72 °C for 20 s, and a final extension step of 72 °C for 5 min.

Microbial genomic DNA extracted from skin swab samples was amplified using MTP Taq DNA Polymerase (Merck KGaA, Darmstadt, Germany) with the PCR thermocycler protocol as follows: Initiation step of 94°C for 1 min followed by 35 cycles of 94°C for 60 s, 55 °C for 45 s, and 72 °C for 30 s, and a final extension step of 72 °C for 5 min.

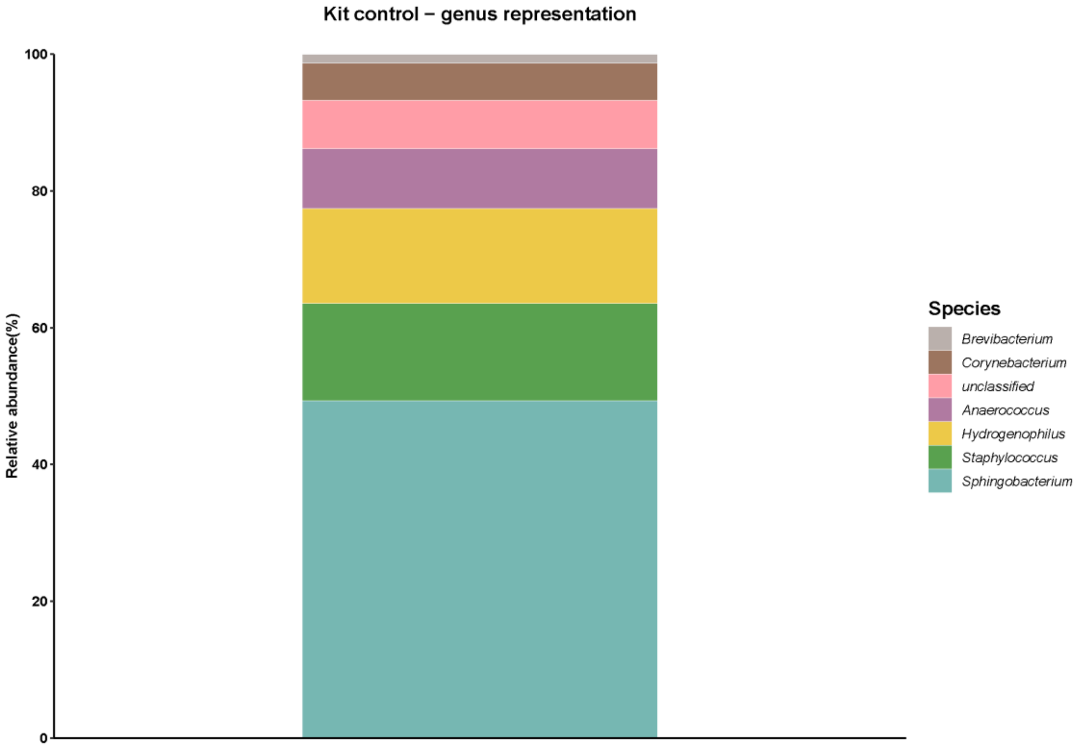
A subsequent indexing PCR was carried out to add unique sample-specific DNA barcodes to the generated amplicons in accordance with the Illumina 16S Metagenomic Sequencing Protocol (Illumina, California, USA).(Illumina, n.d.) Libraries DNA concentration was quantified using a Qubit fluorimeter (Invitrogen) using the 'High Sensitivity' assay and samples were pooled at a standardised concentration.(Illumina, n.d.) The pooled library was sequenced on the Illumina MiSeq platform (Illumina, California, USA) utilising 2×300 bp chemistry.

5.5.5 Bioinformatic and biostatistical analysis

The majority of the analysis was performed in R (v3.6.0). Paired reads were quality filtered, trimmed, merged and Amplicon Sequence Variants (ASV) inferred using the R package dada2 (v1.12.1)[34]. Taxonomic classification was performed using the RDP Classifier within Mothur in conjugation with SPINGO, a species-level classifier[64]. A confidence cut of 80% was used for taxonomic assignment. QIIME v1.9.1 and the R package vegan v2.5.6 were used to calculate β-diversity metrics[65]. β-diversity was visualized via principal coordinates analysis (PCoA) plots whose coordinates were identified using with the Ape package v5.1. R-squared (R^2) and p-value were calculated using Permutational Multivariate Analysis of Variance (PERMANOVA) via the R package vegan (v2.4.2). Differential abundance analysis was carried out using DEseq2 (v1.22.2)[66]. Random forest was performed in R using the package randomForest (v4.6.14) Genomic functionality was inferred using PICRUSt2 with the command picrust2_pipeline.py with default[67].

5.5.6 Identification of potential microbial DNA contamination

Because skin samples are considered low biomass with respect to bacterial load, we included protocols to mitigate the potential impact of contamination. Reagents used were selected based on their quality of being putatively microbial DNA free including the QIAamp Ultraclean production Pathogen Mini Kit (Qiagen, Hilden, Germany) and MTP Taq DNA Polymerase (Merck). A negative control was run within the same sequencing batch to detect potential contamination from reagents. This negative control was dominated by taxa typical of contamination including *Sphingobacterium* and *Hydrogenophilus*(eFigure12)[68]. Furthermore the negative sample had atypically low DNA concentration (0.521ng/µl) relative to extracted clinical samples as measured by the qubit (eTable 1). Other non-contaminant taxa were found in the kit control but we posit that this is due to index swapping[69]. We further utilized the R package decontam to detect contaminating ASVs[70]. Using the 'frequency method' we identified 15 ASVs that reached the threshold (eTable 2). These ASVs contributed only modestly to the samples, median=0, mean=0.01881(eFigure 13). Filtering these ASVs from the ASV table had no effect on differential abundance analysis.



eFigure 12: Taxonomic representation within mock extraction. Bar plots displaying the relative abundance of genera within the kit control.

	patient	disease_state	qubit_values	skin_site	sample_type
Blank	Blank	blank	0.521	blank	blank
CS01BK	CS01	control	29.6	Buttock	swab
CS01GR	CS01	control	30.4	Groin	swab
CS01IM	CS01	control	2.63	Inframammary_fold	swab
CS01NA	CS01	control	30.7	Nasal	swab
CS03BK	CS03	control	35.1	Buttock	swab
CS03IM	CS03	control	7.25	Inframammary_fold	swab
CS03NA	CS03	control	31.7	Nasal	swab
CS04BK	CS04	control	28.6	Buttock	swab
CS04GR	CS04	control	24.9	Groin	swab
CS04IM	CS04	control	26.3	Inframammary_fold	swab
CS05BK	CS05	control	15.1	Buttock	swab
CS05GR	CS05	control	22.4	Groin	swab
CS05NA	CS05	control	8.53	Nasal	swab
CS06BK	CS06	control	28	Buttock	swab
CS06GR	CS06	control	7.55	Groin	swab
CS06IM	CS06	control	3.59	Inframammary_fold	swab
CS06NA	CS06	control	23.8	Nasal	swab
CS07BK	CS07	control	17.3	Buttock	swab
CS07GR	CS07	control	16.6	Groin	swab
CS07IM	CS07	control	16.1	Inframammary_fold	swab
CS07NA	CS07	control	40.2	Nasal	swab
CS08AX	CS08	control	18.6	Axilla	swab
CS08BK	CS08	control	25.3	Buttock	swab
CS08GR	CS08	control	24.3	Groin	swab
CS08NA	CS08	control	27.5	Nasal	swab
CS09AX	CS09	control	25.1	Axilla	swab
CS09BK	CS09	control	30.6	Buttock	swab
CS09GR	CS09	control	37.3	Groin	swab
CS09IM	CS09	control	26.4	Inframammary_fold	swab
CS09NA	CS09	control	20.2	Nasal	swab
CS10BK	CS10	control	3.55	Buttock	swab
CS10GR	CS10	control	23	Groin	swab
CS10IM	CS10	control	23.9	Inframammary_fold	swab
CS10NA	CS10	control	15.9	Nasal	swab
CS11AX	CS11	control	11	Axilla	swab
CS11BK	CS11	control	24	Buttock	swab
CS11GR	CS11	control	16.3	Groin	swab
CS11NA	CS11	control	19.6	Nasal	swab
CS12AX	CS12	control	15.9	Axilla	swab
CS12BK	CS12	control	24.5	Buttock	swab
CS12GR	CS12	control	24	Groin	swab

CS12NA	CS12	control	30.3	Nasal	swab
CS13BK	CS13	control	18.5	Buttock	swab
CS13GR	CS13	control	14.8	Groin	swab
CS13IM	CS13	control	23.8	Inframammary_fold	swab
CS13NA	CS13	control	18.9	Nasal	swab
CS14AX	CS14	control	2.98	Axilla	swab
CS14BK	CS14	control	18.2	Buttock	swab
CS14GR	CS14	control	32.7	Groin	swab
CS14NA	CS14	control	16.9	Nasal	swab
CS15BK	CS15	control	24.2	Buttock	swab
CS15GR	CS15	control	28.4	Groin	swab
CS15NA	CS15	control	19.2	Nasal	swab
CS16BK	CS16	control	5.02	Buttock	swab
CS16GR	CS16	control	13.9	Groin	swab
CS16IM	CS16	control	8.03	Inframammary_fold	swab
CS16NA	CS16	control	31.6	Nasal	swab
CS17BK	CS17	control	15.3	Buttock	swab
CS17GR	CS17	control	16.2	Groin	swab
CS17IM	CS17	control	19.2	Inframammary_fold	swab
CS17NA	CS17	control	13.6	Nasal	swab
CS18BK	CS18	control	16.2	Buttock	swab
CS18GR	CS18	control	9.69	Groin	swab
CS18IM	CS18	control	9.25	Inframammary_fold	swab
CS18NA	CS18	control	6.42	Nasal	swab
CS19BK	CS19	control	24	Buttock	swab
CS19IM	CS19	control	21.4	Inframammary_fold	swab
CS19NA	CS19	control	19.8	Nasal	swab
CS20AX	CS20	control	20.6	Axilla	swab
CS20BK	CS20	control	32	Buttock	swab
CS20GR	CS20	control	32.2	Groin	swab
CS20IM	CS20	control	27.1	Inframammary_fold	swab
TH01L1	TH01	HS	20.6	Groin	swab
TH01NA	TH01	HS	35.9	Nasal	swab
TH01RA	TH01	HS	32.5	Axilla	swab
TH01RB	TH01	HS	32.8	Buttock	swab
TH02LB	TH02	HS	20.7	Buttock	swab
TH02NA	TH02	HS	38.5	Nasal	swab
TH02RG	TH02	HS	36.2	Groin	swab
TH03LI	TH03	HS	30.7	Groin	swab
TH03NA	TH03	HS	2.76	Nasal	swab
TH03RA	TH03	HS	48	Axilla	swab
TH04LA	TH04	HS	2.95	Axilla	swab
TH04LG	TH04	HS	23.9	Groin	swab
TH04NA	TH04	HS	25.6	Nasal	swab
TH04RM	TH04	HS	20.1	Inframammary_fold	swab

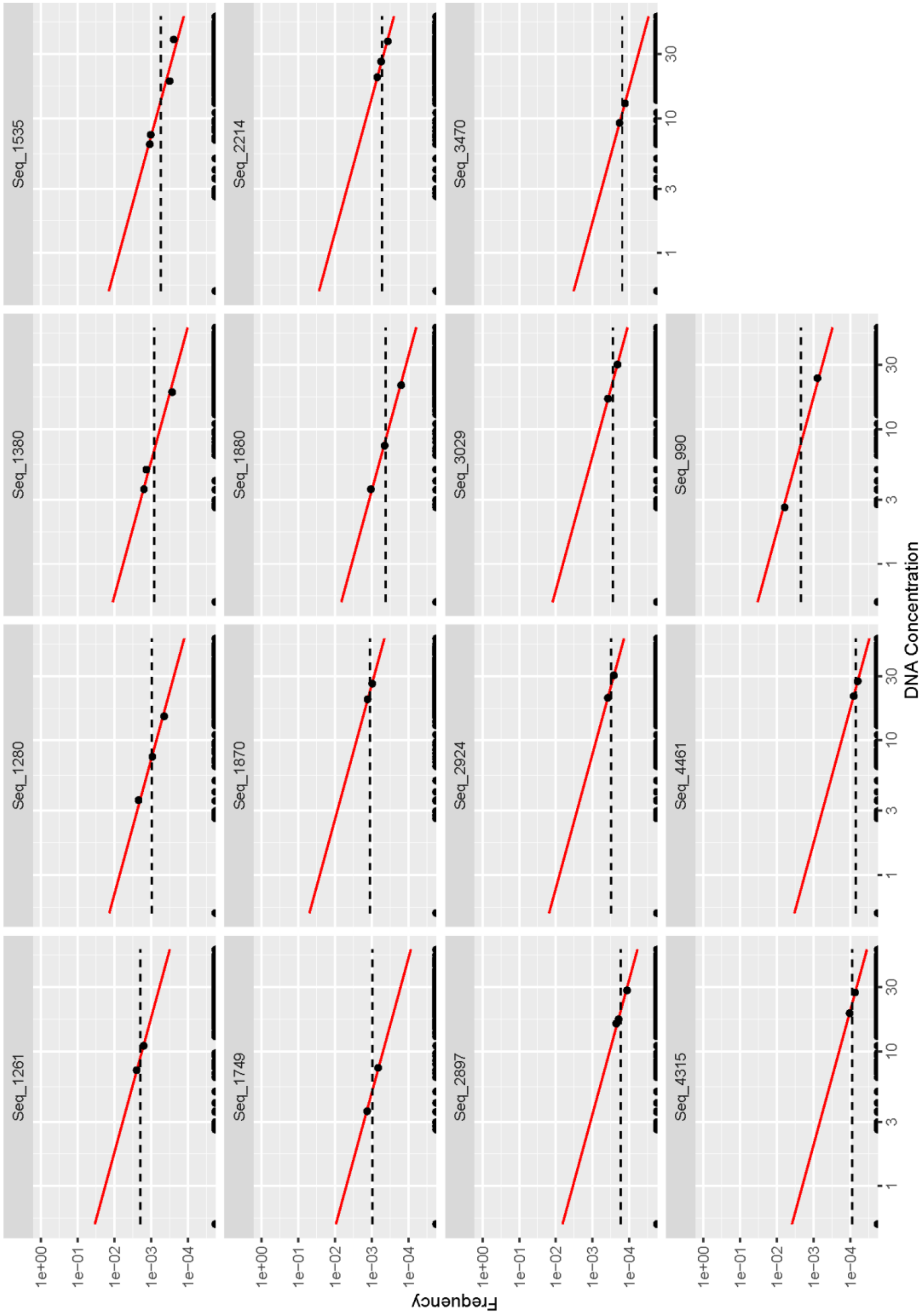
TH05LA	TH05	HS	21.8	Axilla	swab
TH05LB	TH05	HS	36.5	Buttock	swab
TH05LG	TH05	HS	18.1	Groin	swab
TH05NA	TH05	HS	22.1	Nasal	swab
TH05RB	TH05	HS	29.2	Buttock	swab
TH05RG	TH05	HS	44.8	Groin	swab
TH06NA	TH06	HS	30.8	Nasal	swab
TH06RA	TH06	HS	32	Axilla	swab
TH07LG	TH07	HS	43.1	Groin	swab
TH07NA	TH07	HS	45.8	Nasal	swab
TH07RM	TH07	HS	33.5	Inframammary_fold	swab
TH08LA	TH08	HS	39.7	Axilla	swab
TH08LM	TH08	HS	44.9	Inframammary_fold	swab
TH08SP	TH08	HS	28.7	Suprapubic	swab
TH09LA	TH09	HS	31	Axilla	swab
TH09RG	TH09	HS	24.6	Groin	swab
TH10RA	TH10	HS	28	Axilla	swab
TH10RG	TH10	HS	33.7	Groin	swab
TH11LM	TH11	HS	47.8	Inframammary_fold	swab
TH11NA	TH11	HS	37.1	Nasal	swab
TH11RG	TH11	HS	18.8	Groin	swab
TH12LM	TH12	HS	33.1	Inframammary_fold	swab
TH12NA	TH12	HS	28.6	Nasal	swab
TH12RA	TH12	HS	52	Axilla	swab
TH13LA	TH13	HS	29.9	Axilla	swab
TH14NA	TH14	HS	57	Nasal	swab
TH15NA	TH15	HS	27.4	Nasal	swab
TH16LA	TH16	HS	27.9	Axilla	swab
TH16NA	TH16	HS	31.8	Nasal	swab
TH16RAB	TH16	HS	35.6	Abdomen	swab
TH17LA	TH17	HS	2.78	Axilla	swab
TH17NA	TH17	HS	44.9	Nasal	swab
TH18LG	TH18	HS	28.6	Groin	swab
TH18NA	TH18	HS	40.3	Nasal	swab
TH19LA	TH19	HS	19.7	Axilla	swab
TH19LG	TH19	HS	40.2	Groin	swab
TH19NA	TH19	HS	23.3	Nasal	swab
TH20LA	TH20	HS	37.9	Axilla	swab
TH20NA	TH20	HS	47.3	Nasal	swab
TH21NA	TH21	HS	22.9	Nasal	swab
TH21RA	TH21	HS	4.14	Axilla	swab
TH22NA	TH22	HS	32.1	Nasal	swab
TH22SP	TH22	HS	30.4	Suprapubic	swab
TH23LA	TH23	HS	38.9	Axilla	swab
TH23NA	TH23	HS	26.5	Nasal	swab

TH23RA	TH23	HS	34.7	Axilla	swab
TH24LSP	TH24	HS	27.5	Suprapubic	swab
TH24NA	TH24	HS	12.9	Nasal	swab
TH25NA	TH25	HS	31.1	Nasal	swab
TH25RG	TH25	HS	45.6	Groin	swab
TH26AB	TH26	HS	43.2	Abdomen	swab
TH26NA	TH26	HS	49.5	Nasal	swab
TH27NA	TH27	HS	31.1	Nasal	swab
TH28LG	TH28	HS	6.89	Groin	swab
TH28NA	TH28	HS	27.9	Nasal	swab
TH29LA	TH29	HS	21.3	Axilla	swab
TH29LSP	TH29	HS	40.7	Suprapubic	swab
TH30LA	TH30	HS	34.7	Axilla	swab
TH30NA	TH30	HS	34.1	Nasal	swab
TH30RG	TH30	HS	30.3	Groin	swab

Etable 1| DNA concentrations of samples post library preparation.

	Phylum	Class	Order	Family	Genus	Species
Seq_990	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Blautia	Blautia luti
Seq_1261	Cyanobacteria/Chloroplast	Chloroplast	Chloroplast	Streptophyta	unclassified	unclassified
Seq_1280	Actinobacteria	Actinobacteria	Actinomycetales	Micrococcaceae	Kocuria	unclassified
Seq_1380	Proteobacteria	Alphaproteobacteria	Sphingomonadales	Sphingomonadaceae	Sphingomonas	unclassified
Seq_1535	Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus	Lactobacillus delbrueckii
Seq_1749	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Pseudomonadaceae	Pseudomonas	Pseudomonas rhizosphaerae
Seq_1870	Bacteroidetes	Flavobacteriia	Flavobacteriales	Flavobacteriaceae	Wautersiella	Wautersiella falsenii
Seq_1880	Proteobacteria	Alphaproteobacteria	Rhizobiales	Xanthobacteraceae	Xanthobacter	unclassified
Seq_2214	Candidatus_Saccharibacteria	unclassified	unclassified	unclassified	unclassified	unclassified
Seq_2897	Proteobacteria	Alphaproteobacteria	Rhizobiales	Methylobacteriaceae	Methylobacterium	Methylobacterium aquaticum
Seq_2924	Firmicutes	Erysipelotrichia	Erysipelotrichales	Erysipelotrichaceae	unclassified	unclassified
Seq_3029	Firmicutes	Bacilli	Bacillales	Bacillaceae_1	Bacillus	unclassified
Seq_3470	Bacteroidetes	Bacteroidia	Bacteroidales	Porphyromonadaceae	Tannerella	Tannerella forsythia
Seq_4315	Acidobacteria	Acidobacteria_Gp4	Blastocatella	unclassified	unclassified	unclassified
Seq_4461	Proteobacteria	Epsilonproteobacteria	Campylobacterales	Campylobacteraceae	Campylobacter	Campylobacte ureolyticus

Etable 2| Taxonomic assignment of ASVs identified as contamination.



eFigure 13: Decontam frequency graph. X axis equals concentration of sample before normalization. Y-axis equals frequency of ASV. Each dot represents a sample.

5.5.7 Storage of sequencing data

Datasets related to this article can be found at https://www.ebi.ac.uk/ena/browser/home, hosted at European Nucleotide Archive, accession number PRJEB43835.(https://www.ebi.ac.uk/ena/browser/view/PRJEB43835, Accessed 03/26/2021.)

5.6 Acknowledgement

Access to Data and Data Analysis

Dr Siobhan McCarthy and Mr. Maurice Barrett had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.

5.7 Authors Contribution statement

SM, MM, AMT, PWOT & FS conceived and designed the study. SM, SK, MB, KV, and PP were involved in data curation. MB, PWOT performed the formal analysis. MM, FS and PWOT were involved in funding acquisition. SM, MB, and PWOT wrote the original draft. AMT, SK, KV, PP, FS, MM, and PWOT reviewed and edited the paper for publication.

Conflicts of Interest: None declared

Funding/Support

Dr Siobhan McCarthy received a grant from the City of Dublin Skin Cancer Hospital

Charity (CDSCHC) and a grant in the form of salary support from the Irish Centre for Arthritis Research and Education (ICARE).

Work in Professor Paul W. O'Toole's laboratory is supported by Science Foundation Ireland through a Centre award (APC/SFI/12/RC/2273_P2) to APC Microbiome Ireland.

Role of Funder/Sponsor Statement

Sponsors CDSCHC, ICARE and SFI had no role in design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication

5.8 References

1 Zouboulis, C. *et al.* European S1 guideline for the treatment of hidradenitis suppurativa/acne inversa. *Journal of the European Academy of Dermatology and Venereology* **29**, 619-644 (2015).

2 Jemec, G. B. & Kimball, A. B. Hidradenitis suppurativa: epidemiology and scope of the problem. *Journal of the American Academy of Dermatology* **73**, S4-S7 (2015).

3 Miller, I. M., McAndrew, R. J. & Hamzavi, I. Prevalence, risk factors, and comorbidities of hidradenitis suppurativa. *Dermatologic clinics* **34**, 7-16 (2016).

4 Revuz, J. E. *et al.* Prevalence and factors associated with hidradenitis suppurativa: results from two case-control studies. *Journal of the American Academy of Dermatology* **59**, 596-601 (2008).

5 Sabat, R. *et al.* Increased prevalence of metabolic syndrome in patients with acne inversa. *PloS one* **7**, e31810 (2012).

6 Ingram, J. R. Interventions for hidradenitis suppurativa: updated summary of an original Cochrane Review. *JAMA dermatology* **153**, 458-459 (2017).

7 Sabat, R. *et al.* Hidradenitis suppurativa. *Nature Reviews Disease Primers* **6**, 1-20 (2020).

8 Chen, W.-T. & Chi, C.-C. Association of Hidradenitis Suppurativa With Inflammatory Bowel Disease: A Systematic Review and Meta-analysis. *JAMA Dermatol* **155**, 1022-1027, doi:10.1001/jamadermatol.2019.0891 (2019).

9 Richette, P. *et al.* Hidradenitis suppurativa associated with spondyloarthritis—results from a multicenter national prospective study. *The Journal of rheumatology* **41**, 490-494 (2014).

10 Egeberg, A., Gislason, G. H. & Hansen, P. R. Risk of major adverse cardiovascular events and all-cause mortality in patients with hidradenitis suppurativa. *JAMA dermatology* **152**, 429-434 (2016).

11 Jung, J. M. *et al.* Assessment of Overall and Specific Cancer Risks in Patients With Hidradenitis Suppurativa. *JAMA Dermatol* **156**, 844-853, doi:10.1001/jamadermatol.2020.1422 (2020).

12 Garg, A., Papagermanos, V., Midura, M., Strunk, A. & Merson, J. Opioid, alcohol, and cannabis misuse among patients with hidradenitis suppurativa: A population-based analysis in the United States. *Journal of the American Academy of Dermatology* **79**, 495-500. e491 (2018).

13 Patel, K. R. *et al.* Association between hidradenitis suppurativa, depression, anxiety, and suicidality: a systematic review and meta-analysis. *Journal of the American Academy of Dermatology* **83**, 737-744 (2020).

14 Balato, A. *et al.* Human microbiome: composition and role in inflammatory skin diseases. *Archivum immunologiae et therapiae experimentalis* **67**, 1-18 (2019).

15 Guet-Revillet, H. *et al.* Bacterial pathogens associated with hidradenitis suppurativa, France. *Emerging infectious diseases* **20**, 1990 (2014).

16 Kelly, G. *et al.* Dysregulated cytokine expression in lesional and nonlesional skin in hidradenitis suppurativa. *British Journal of Dermatology* **173**, 1431-1439 (2015).

17 Laffert, M. v. *et al.* Hidradenitis suppurativa (acne inversa): early inflammatory events at terminal follicles and at interfollicular epidermis. *Experimental dermatology* **19**, 533-537 (2010).

18 Nikolakis, G. *et al.* Bacteriology of hidradenitis suppurativa/acne inversa: a review. *Journal of the American Academy of Dermatology* **73**, S12-S18 (2015).

19 van der Zee, H. H., Horvath, B., Jemec, G. B. & Prens, E. P. The association between hidradenitis suppurativa and Crohn’s disease: in search of the missing pathogenic link. *Journal of Investigative Dermatology* **136**, 1747-1748 (2016).

20 Abraham, C. & Cho, J. H. Mechanisms of disease. *N Engl J Med* **361**, 2066-2078 (2009).

21 Duerr, R. H. *et al.* A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. *science* **314**, 1461-1463 (2006).

22 Schlapbach, C., Hänni, T., Yawalkar, N. & Hunger, R. E. Expression of the IL-23/Th17 pathway in lesions of hidradenitis suppurativa. *Journal of the American Academy of Dermatology* **65**, 790-798 (2011).

23 Baumgart, D. C. & Sandborn, W. J. Crohn's disease. *The Lancet* **380**, 1590-1605 (2012).

24 Seksik, P., Nion-Larmurier, I., Sokol, H., Beaugerie, L. & Cosnes, J. Effects of light smoking consumption on the clinical course of Crohn's disease. *Inflammatory bowel diseases* **15**, 734-741 (2009).

25 Tuvlin, J. A. *et al.* Smoking and inflammatory bowel disease: trends in familial and sporadic cohorts. *Inflammatory bowel diseases* **13**, 573-579 (2007).

26 Clooney, A. G. *et al.* Ranking microbiome variance in inflammatory bowel disease: a large longitudinal intercontinental study. *Gut* **70**, 499-510 (2021).

27 Clooney, A. G. *et al.* Whole-virome analysis sheds light on viral dark matter in inflammatory bowel disease. *Cell host & microbe* **26**, 764-778. e765 (2019).

28 Lloyd-Price, J. *et al.* Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* **569**, 655-662, doi:10.1038/s41586-019-1237-9 (2019).

29 Scher, J. U. *et al.* Decreased bacterial diversity characterizes the altered gut microbiota in patients with psoriatic arthritis, resembling dysbiosis in inflammatory bowel disease. *Arthritis Rheumatol* **67**, 128-139, doi:10.1002/art.38892 (2015).

30 Schirmer, M., Garner, A., Vlamakis, H. & Xavier, R. J. Microbial genes and pathways in inflammatory bowel disease. *Nature Reviews Microbiology* **17**, 497-511, doi:10.1038/s41579-019-0213-6 (2019).

31 Zhang, X. *et al.* The oral and gut microbiomes are perturbed in rheumatoid arthritis and partly normalized after treatment. *Nat Med* **21**, 895-905, doi:10.1038/nm.3914 (2015).

32 Eppinga, H. *et al.* Similar depletion of protective Faecalibacterium prausnitzii in psoriasis and inflammatory bowel disease, but not in hidradenitis suppurativa. *Journal of Crohn's and Colitis* **10**, 1067-1075 (2016).

33 Kam, S., Collard, M., Lam, J. & Alani, R. M. Gut microbiome perturbations in patients with hidradenitis suppurativa: a case series. *The Journal of investigative dermatology* **141**, 225-228. e222 (2021).

34 Callahan, B. J. *et al.* DADA2: High-resolution sample inference from Illumina amplicon data. *Nat Methods* **13**, 581-583, doi:10.1038/nmeth.3869 (2016).

35 Park, S.-k. *et al.* Differentially Abundant Bacterial Taxa Associated with Prognostic Variables of Crohn's Disease: Results from the IMPACT Study. *Journal of clinical medicine* **9**, 1748 (2020).

36 Byrd, A. L., Belkaid, Y. & Segre, J. A. The human skin microbiome. *Nature Reviews Microbiology* **16**, 143 (2018).

37 Clausen, M.-L. *et al.* Association of disease severity with skin microbiome and filaggrin gene mutations in adult atopic dermatitis. *JAMA dermatology* **154**, 293-300 (2018).

38 Fyhrquist, N. *et al.* Microbe-host interplay in atopic dermatitis and psoriasis. *Nature communications* **10**, 1-15 (2019).

39 Kong, H. H. *et al.* Temporal shifts in the skin microbiome associated with disease flares and treatment in children with atopic dermatitis. *Genome research* **22**, 850-859 (2012).

40 Petersen, E., Skov, L., Thyssen, J. & Jensen, P. Role of the gut microbiota in atopic dermatitis: a systematic review. *Acta dermato-venereologica* **99**, 5-11 (2019).

41 Alekseyenko, A. V. *et al.* Community differentiation of the cutaneous microbiota in psoriasis. *Microbiome* **1**, 1-17 (2013).

42 Sikora, M. *et al.* Gut microbiome in psoriasis: an updated review. *Pathogens* **9**, 463 (2020).

43 Yerushalmi, M., Elalouf, O., Anderson, M. & Chandran, V. The skin microbiome in psoriatic disease: A systematic review and critical appraisal. *Journal of translational autoimmunity* **2**, 100009 (2019).

44 Breban, M. *et al.* Faecal microbiota study reveals specific dysbiosis in spondyloarthritis. *Annals of the Rheumatic Diseases* **76**, 1614, doi:10.1136/annrheumdis-2016-211064 (2017).

45 Hall, A. B. *et al.* A novel Ruminococcus gnavus clade enriched in inflammatory bowel disease patients. *Genome Med* **9**, 103, doi:10.1186/s13073-017-0490-5 (2017).

46 Jeffery, I. B. *et al.* Differences in fecal microbiomes and metabolomes of people with vs without irritable bowel syndrome and bile acid malabsorption. *Gastroenterology* **158**, 1016-1028. e1018 (2020).

47 Joossens, M. *et al.* Dysbiosis of the faecal microbiota in patients with Crohn's disease and their unaffected relatives. *Gut* **60**, 631-637 (2011).

48 Nishino, K. *et al.* Analysis of endoscopic brush samples identified mucosa-associated dysbiosis in inflammatory bowel disease. *Journal of gastroenterology* **53**, 95-106 (2018).

49 Zheng, H. *et al.* Altered gut microbiota composition associated with eczema in infants. *PloS one* **11**, e0166026 (2016).

50 Henke, M. T. *et al.* Ruminococcus gnavus, a member of the human gut microbiome associated with Crohn’s disease, produces an inflammatory polysaccharide. *Proceedings of the National Academy of Sciences* **116**, 12672-12677 (2019).

51 Le Chatelier, E. *et al.* Richness of human gut microbiome correlates with metabolic markers. *Nature* **500**, 541-546, doi:10.1038/nature12506 (2013).

52 Franzosa, E. A. *et al.* Gut microbiome structure and metabolic activity in inflammatory bowel disease. *Nat Microbiol* **4**, 293-305, doi:10.1038/s41564-018-0306-4 (2019).

53 Hu, S. *et al.* Whole exome sequencing analyses reveal gene–microbiota interactions in the context of IBD. *Gut* **70**, 285-296 (2021).

54 Faber, F. *et al.* Host-mediated sugar oxidation promotes post-antibiotic pathogen expansion. *Nature* **534**, 697-699 (2016).

55 Ring, H. C. *et al.* The follicular skin microbiome in patients with hidradenitis suppurativa and healthy controls. *JAMA dermatology* **153**, 897-905 (2017).

56 Beylot, C. *et al.* Propionibacterium acnes: an update on its role in the pathogenesis of acne. *Journal of the European Academy of Dermatology and Venereology* **28**, 271-278 (2014).

57 Tuchayi, S. M. *et al.* Acne vulgaris. *Nature reviews Disease primers* **1**, 1-20 (2015).

58 Li, W. *et al.* Inverse association between the skin and oral microbiota in atopic dermatitis. *Journal of Investigative Dermatology* **139**, 1779-1787. e1712 (2019).

59 Byrd, A. S. *et al.* Neutrophil extracellular traps, B cells, and type I interferons contribute to immune dysregulation in hidradenitis suppurativa. *Science translational medicine* **11** (2019).

60 Genovese, A. *et al.* Bacterial Immunoglobulin Superantigen Proteins A and L Activate Human Heart Mast Cells by Interacting with Immunoglobulin E. *Infection and Immunity* **68**, 5517, doi:10.1128/IAI.68.10.5517-5524.2000 (2000).

61 Patella, V., Casolaro, V., Björck, L. & Marone, G. Protein L. A bacterial Ig-binding protein that activates human basophils and mast cells. *The Journal of Immunology* **145**, 3054-3061 (1990).

62 Neumann, A., Björck, L. & Frick, I.-M. Finegoldia magna, an Anaerobic Gram-Positive Bacterium of the Normal Human Microbiota, Induces Inflammation by Activating Neutrophils. *Frontiers in microbiology* **11**, 65 (2020).

63 Hurley Jr, H. Hidradenitis suppurativa. *Dermatology in General Medicine* **1**, 761-766 (1993).

64 Allard, G., Ryan, F. J., Jeffery, I. B. & Claesson, M. J. SPINGO: a rapid species-classifier for microbial amplicon sequences. *BMC Bioinformatics* **16**, 324, doi:10.1186/s12859-015-0747-1 (2015).

65 Caporaso, J. G. *et al.* QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* **7**, 335-336, doi:10.1038/nmeth.f.303 (2010).

66 Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550, doi:10.1186/s13059-014-0550-8 (2014).

67 Douglas, G. M. *et al.* PICRUSt2 for prediction of metagenome functions. *Nature Biotechnology* **38**, 685-688, doi:10.1038/s41587-020-0548-6 (2020).

68 Eisenhofer, R. *et al.* Contamination in Low Microbial Biomass Microbiome Studies: Issues and Recommendations. *Trends Microbiol* **27**, 105-117, doi:10.1016/j.tim.2018.11.003 (2019).

69 Costello, M. *et al.* Characterization and remediation of sample index swaps by non-redundant dual indexing on massively parallel sequencing platforms. *BMC Genomics* **19**, 332, doi:10.1186/s12864-018-4703-0 (2018).

70 Davis, N. M., Proctor, D. M., Holmes, S. P., Relman, D. A. & Callahan, B. J. Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. *Microbiome* **6**, 226, doi:10.1186/s40168-018-0605-2 (2018).

Chapter 6- General discussion and future perspectives

6.1 The role of microbiology in cancer research in the 21st century

For much of human history, infections by microbes were the leading causes of death. Microbes such as *Mycobacterium tuberculosis* (Tuberculosis), *influenza* (flu) and *Plasmodium falciparum* (malaria) have killed innumerable individuals throughout human existence. However, research into microbes in the 20th century allowed us to combat infectious diseases through medical innovations including antibiotics and vaccines. This is particularly the case in developed countries, with developing countries still suffering considerably from infectious agents[1]. In the second half of the 20th century and during the 21st century, non-communicable diseases including cancer and heart disease have become the leading cause of mortality. Cancer is now the leading cause of death in high-income countries[2]. This shift is due to many factors including lifestyle changes such as diet and a longer lifespan. Cancer research is obviously a major effort within the overall field of biomedical research, with billions of US dollars being spent a year worldwide[3].

Research into the relationship between human biology and the resident human microbiota has experienced a renaissance over the past 15 years. As an aspect of this endeavour, a complex model describing the interaction between cancer and the microbiota is currently being formed. Knowledge of this interaction has informed practically all areas of cancer research including oncogenesis, diagnostics, prognostics and therapeutics. Thus, research into microbes may be integral to combating and hopefully eliminating cancer in the 21st century, saving even greater millions of lives above those saved in the 20th century.

6.2 Categorization of areas of cancer research

One might divide cancer research areas into three categories.

1) *The cause of cancer:* Key to combatting cancer is understanding the underlying mechanisms by which normal healthy cells transform into cancer cells. This includes knowledge of all factors modulating the risk of this phenomenon, predominantly environmental and genetic risks. A high proportion of cancers are believed to be avoidable through risk aversion measures. Bearing in mind the wisdom of the Dutch philosopher Desiderius Erasmus - 'prevention is better than cure', comprehensive models of the origin of cancer would enable strategies to reduce cancer incidences.
2) *Diagnostics and prognostics:* Quick, cheap, sensitive and specific tests are needed to identified individuals with cancer and to determine the likely course of disease progression. Early detection of certain cancers such as colon, liver and lung cancer can improve survival rates[4]. Furthermore, many cancers have identified pre-cancer lesions from which the develop including Barrett's Oesophagus and colonic polyps which are the precursors of oesophageal adenocarcinoma and colorectal cancer, respectively. Stratification of individuals with precancerous lesions into those who are likely and are not likely to develop cancer is needed to save lives.
3) *Therapeutics:* The presence of cancer in a population is all but inevitable. Although a significant proportion of cancer related deaths are avoidable through the modification of risk factors, cancer will arise in population. Furthermore, the elimination of environmental risk factors for cancer such as smoking, or obesity does not seem likely in the near future. Even with our

current arsenal of cancer therapeutics, the survival rate of many cancers remains poor. The overall 5-year survival rate for pancreatic cancer and oesophageal cancer is <7% and <20% respectively[5,6]. This is particularly the case if cancer has metastasised, known as distant disease or distant metastasis.

I contend that the contents of this thesis provide arguments and evidence for the contributory role of microbiota research into all three areas.

6.2.1 The cause of cancer and the microbiota

The question “What is the cause of cancer?” is a captivating question for researchers and non-researchers alike. In modern molecular biology, oncogenesis involves the Darwinian natural selection of somatic mutations within somatic cells[7]. Mutated cells may evolve to acquire the phenotypes known as the Hallmarks of Cancer[8,9]. It is important to recognise that the fitness associated with a mutation, somatic or otherwise, is dependent on the environment in which it occurs[10,11]. Cells in a healthy tissue environment are under purifying selection[12]. It is therefore integral to consider the changes to the tissue environment in which these mutations occur, because change to a tissue environment may itself be a major driver of cancer[10,11]. Accumulation of mutations with age, as well as age related changes to the tissue microenvironment, explain why old age is the strongest risk factor for cancer development. What are the other factors which modulate the generations of somatic mutations and changes to tissue microenvironment? In this thesis, the microbiota is considered as such a factor. It may be important to distinguish two microbiotas which influence cancer biology: the gut microbiota and the intratumoral bacteria specific to the cancer tissue. The gut microbiota contains approximately 97% of the bacterial cells found in the human body[13]. Due to its metabolic range and size, it can exert an influence all tissues in the body through communication with the immune system and release of metabolites into the bloodstream. Increasing number of reports describe the existence of an intratumoral microbiome[14-16]. Theses microbiomes may act locally to modulate the tumour microenvironment.

In chapter one of this thesis, I provided a comprehensive discussion of the role of microbes in mutational mechanisms. The mechanism by which colibactin producing *E. coli* generates mutational signatures is supported by the most robust evidence. However, future studies will need to investigate the global structure of the microbiota and how it relates to the mutational portrait of cancers rather than individual microbes and their related mutational mechanism. Such research would hopefully allow researchers discover microbially driven mutational mechanisms in a more systematic manner rather than one microbe, one metabolite, one mutational signature at the time.

Beyond initiation of cancer evolution though mutational mechanisms, the microbiota can drive tumorigenesis through mechanisms that alter the tissue microenvironment. Another way of looking at this question is, how might the microbiota influence the purifying selection that somatic cells are under?

Shanahan & O'Toole hypothesize that a difference in microbial load and content may in part explain the differences in the rates of cancers between the proximal gut (small intestine) and distal gut (large intestine)[17]. This hypothesis has been recently supported by work by Kadosh et al[18]. The phenotype expressed by mutations in Trp53 (the gene that encodes p53 in mice) varied from tumour-suppressive to oncogenic depending on the tissue environment. In particular Kadosh et al demonstrated that, in the context of WNT-driven intestinal cancer mouse models, p53 had a pro-oncogenic effect in the distal gut while it exerted a tumour suppressive effect in the foregut Such a switching between genetic functionality was found to be dictated by microbiota-derived gallic acid[18]

Chronic inflammation increases the risk of cancer ,with 30% of cancers incidences being linked to chronic inflation[19]. Environmental factors such as tobacco smoking promote cancer in part by promoting chronic inflammation. Microbial causes of inflammation included *Helicobacter pylori* and hepatitis B virus (HBV) or C (HCV) which promote the development of gastric cancer and hepatocellular carcinoma, respectively[20,21].Inflamtion can be regarded as a pro-carcinogenic environment for cancer development. Many diseases characterised by chronic inflammation are associated with an increased risk of cancer. Ulcerative colitis, pelvic inflammatory disease and celiac disease are linked to increases in colorectal cancer, ovarian cancer, and intestinal lymphoma respectively[22-24]. Current models of the pathogenesis of these inflammatory diseases include, to varying degrees, the microbiota playing a role[25-27].

In chapter 5 of this thesis, I describe changes in the skin and faecal microbiota that are linked to hidradenitis suppurativa. Individuals with hidradenitis suppurativa have an increased risk for a variety of cancers. Relevant to this thesis, individuals with hidradenitis suppurativa have a reported increased risk of colorectal cancer of ~45%[28]. Individuals with HS have a higher rate of IBD and in particular Crohn’s Disease relative to the general population. The prevalence of IBD in the general population is 0.3% and 0.5% for Crohn’s disease and ulcerative colitis, respectively, while in the HS population the rates are 0.8–2.5% and 0.8–1.3% for Crohn's disease and ulcerative colitis, respectively. Individuals with Crohn's disease have an increased risk of colorectal cancer (~40% increase) and small bowel cancer (~1000% increase)[29,30]. Microbiome features associated with HS and Crohn's, both shared and otherwise, may at least in part explain this increase in cancer risk. We found *Ruminococcus gnavus,* a microbe commonly found to be enriched in individuals with Crohn's disease, to be enriched in individuals with HS. *Ruminococcus gnavus* has been demonstrated to have pro-inflammatory activities. Thus, particular incidences of colorectal cancer could be explained using a model involving microbially-induced inflammation. Such models have been experimentally supported. Lung adenocarcinoma development was found to be promoted by lung microbiota driven inflammation through the activation of interleukin-17 and interleukin-23 producing γδ T cells[31]. Such findings will have to be replicated in other geographical settings and with larger cohorts. Furthermore, methodologies which offer a more in depth interrogation of the gut microbiome namely shotgun metagenomic sequencing should be employed. There is a growing selection of methods which enable the engineering of microbiome features including faecal microbiota transplant, phage therapy, bacteriocins and dietary medication[32-34]. Such strategies are being developed to treat inflammatory bowel diseases[35]. One could envisage the opportunity to take advantage of such efforts in order to treat HS. For example the development of a phage based therapeutic strategy to target *Ruminococcus gnavus* with the purpose of treating IBD may be repurposed to treat HS.

6.2.2 Diagnostic and prognostic potential of microbiota data

Certain microbial signatures that are correlated with specific cancers hold the potential to be exploited for diagnostic and prognostic purposes. Currently available methods to detect colorectal cancer include the faecal occult blood test which allows non-evasive detection. Colonoscopy in conjunction with biopsy collection are used as more comprehensive yet more invasive forms of CRC detection. A microbiome-based test could replace or, more likely, complement such procedures. Flemer et al identified an enrichment of taxa that typically colonize the oral cavity in individuals with polyps and CRC[36]. These results were supported by work by Thomas et al who found an increase in the abundance of oral species in individuals with CRC relative to healthy controls[37]. Using machine learning classifiers on oral and/or stool microbiome data, a number of studies have demonstrated that individuals with CRC can be distinguished from control cohorts[36-39].

In our study described in Chapter 3 we established that mucosal biopsies derived from different areas of a single excised tumour harboured largely the same microbiota and were similar to undisease tissue from the same individual. This might suggest that samples taken during colonoscopy from the colon for microbiome analysis would be equally informative regardless of the location from which the sample was taken. However, we did find certain microbes enriched on tumour samples relative to non-diseased tissue. In particular *Fusobacterium nucleatum* was found to be enriched. *F.nucleatum* has been identified as predictive within these models. Thus, it is possible that the predictive power of samples derived from non-tumour samples may be reduced.

The identification of individuals that will develop cancer is a key strategy in early detection and prevention. As discussed in preceding chapters, individuals with Barrett's Oesophagus have a 10-fold to 55-fold higher risk of developing oesophageal adenocarcinoma. However only about 0.1%-1% of individuals with Barrett's Oesophagus go on to develop OAC. Thus the question arises, what are the biological mechanisms that determine progression of Barrett's oesophageal to oesophageal adenocarcinoma? Furthermore, can we predict those individuals with Barrett's oesophageal disease that will go on to develop oesophageal adenocarcinoma? Biomarkers in the form of genomic and epigenetic including p53 expression, DNA-methylation changes, copy number instability and clonal diversity[40-43].

Changes in the oesophageal tract may predict defined histological progression along the oesophageal adenocarcinoma sequence.

In Chapter 2, we defined a number of differences in microbial features between clinical groups within the oesophageal adenocarcinoma sequence. Pertinent to the above discussion we found that, with respect to biopsy samples derived from the gastro-oesophageal junction, an enrichment occurred of *Fusobacterium necrophorum* in dysplastic and neoplastic tissue relative to normal stratified epithelium and metaplastic tissue. The relative/absolute abundance of *Fusobacterium necrophorum* may thus be predictive of the transformation of metaplastic tissue to dysplastic and neoplastic tissue. However, the cross-sectional nature of the study design in chapter 2 limits what one can infer with regard to the microbiome dynamics during the oesophageal adenocarcinoma sequence. For example, the abundance of *Fusobacterium necrophorum* may simply increase in parallel with histological transformation. Longitudinal studies are required to provide greater predictive power using microbiome data when it comes to transformation of metaplastic tissue.

6.2.3 The role of the microbiota in cancer therapeutics

The use of microbes in the treatment of cancer is an ancient endeavour. A treatment attributed to the Egyptian physician Imhotep (~2600 BCE) involved causing an infection to reduce tumours (swellings)[44]. In 1891, William B. Coley injected heat-killed *streptococcal organism* [sic] and *Serratia marcescens* (Coley's Toxin) into individuals with cancer with the hope of eliciting an anti-tumorigenic immune response[45]. This treatment demonstrated some level of success with a >10-year disease-free survival in ~30%[44]. Thus, this not only the first demonstration of immunotherapy but also the first (recorded) example of microbially directed immunotherapy

The microbiota is now considered an important modulator of immune checkpoint inhibitor (ICI)-based therapeutics. In chapter 4 of this thesis, we described the difference between the faecal microbiota of responders versus non-responders and individuals with no-side effects versus individuals with side effects, with regards to ICI, in an Irish cohort. With respect to responders, there was notably an overlap in taxa found to be associated with responders between our study and previous studies. Inter-individual variation in gut microbiome composition is strongly influenced by geography[46-48]. Thus, the reproducibility of response associated taxa is strengthened by the fact the data come from geographically distinct locations. Functional genomic features have failed to reveal a similar consistency. In the study described in chapter 4, microbiome functional features associated with response included those involved in exopolysaccharide synthesis. Exopolysaccharides have been reported to have immunomodulatory activity. In chapter 4 I also explored the relationships between the faecal microbiome and ICI induced side-effects.

Data derived from microbiome studies can be used to inform and enhance ICI therapeutics. In a broad sense one can change the microbiota of an individual to that corresponding of those of patients who responded to ICIs. Early-stage clinical trials regarding the use of FMTs in ICIs therapy has shown preliminary promise[49,50]. The use of single microbes in the form of probiotics such as *Bifidobacterium* to complement ICI therapy is being explored[51,52]. The introduction of living bacteria may not be even necessary. In chapter 4 we reported that *Akkermansia muciniphila* was associated with individuals who did not exhibit side effects when treated with ICI. In mouse models, the introduction of pasteurized *A. muciniphila* or Amuc_1100 (Type IV pili protein and agonist to Toll-like receptor 2) attenuated azoxymethane induced colitis and colon carcinogenesis[53,54]. This anti-inflammatory effect was reported to be achieved though effecting a reduction in infiltrating macrophages and CD8+ cytotoxic T lymphocytes in the colon[53] . Flagellin derived from *E. gallinarum* was shown to be an immunostimulant by interacting the toll like receptor 5[55].These studies demonstrate that abiotic microbial derived materials may augment ICI therapy.

6.2.3.1 Cancer vaccines

No fields of productive scientific research happen in isolation from society at large. Currently, there is a global pandemic caused by Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Vaccines can be argued to be the single greatest innovation in medical history as measured by lives saved. Vaccines are being employed to tackle the current pandemic and are seen as the most promising avenue to exit this pandemic. These vaccines have been developed as a result of great scientific effort backed by appropriate funding. Such an endeavour would hopefully have spinoffs to other biomedical fields including cancer research in the same manner that space exploration has lent itself to society-changing spinoff technologies.

Therapeutic cancer vaccines are currently under development. These are distinct from prophylactic cancer vaccines such as those directed against hepatitis B virus and human papillomavirus, which are the causes of hepatocellular carcinoma and cervical cancer, respectively[56]. Therapeutic cancer vaccines are designed to target antigens of two general classes: tumour-associated antigens and tumour-specific antigen[57]. Tumour-associated antigens are self-antigens that are either preferentially or abnormally expressed in tumour cells. Tumour-associated antigens are expressed to some extent on normal healthy cells; thus, vaccines developed against these antigens encounter the problems of low immunologically reactivity and (in the cases where they do work) autoimmune reactions[57]. Vaccines developed against tumour-specific antigen, antigens expressed exclusively by tumours, hold the potential to train the immune system to selectively destroy tumour cells.

A mutated variant of isocitrate dehydrogenase is commonly identified in astrocytomas, a type of brain cancer and is presented on the major histocompatibility complex (MHC) class II[58]. Such an antigen can be defined as a ‘shared neoantigen’ as it is a tumour-specific antigen which is shared across tumours from many individuals[57]. Recent vaccines against this antigen have proven safe and preliminary effective in phase I clinical trials[59]. Therapeutic cancer vaccines could work in conjunction with other therapeutics, namely ICIs. Indeed, several studies have explored the possibility of this synergistic interaction and their results have shown promise[60-62].

Intratumoral bacteria may provide tumour-specific antigens which vaccines could be developed against. In recent work by Kalaora et al, melanoma cells were found to present bacterially- derived peptides on human leukocyte antigens (HLAs)[63]. As there is growing evidence for a resident tumour microbiome, a range of cancers may be targeted by developing vaccines for tumour specific bacteria.

Using bacterially derived tumour-specific antigen as targets for vaccines faces a number of obstacles. First, one must ensure that such bacterial antigens are truly tumour specific. Previous studies regarding intratumoral bacteria reported taxa such as *Fusobacterium* and *Staphylococcus* that are readily found elsewhere in the body[14,63]. Thus, a vaccine targeting these taxa are likely to have off-target effects. In the context of situations where vaccines are used to augment ICI, the unintended targeting of responder-associated taxa could possibly have detrimental outcomes.

Furthermore, why does the immune system attack this non-self-entity without therapeutic intervention? Like cancer cells, bacterial cells are under evolutionary pressure to evade the immune system. *Fusobacterium nucleatum* has been demonstrated to supress immune surveillance by the binding of its surface protein Fap2 to the TIGIT receptor of tumour-infiltrating lymphocytes[64]. Thus, attenuating the immune suppressive activity of the intratumoral microbiota may be crucial to harnessing the full potential of cancer vaccines.

Finally, the issue of contamination comes into focus when dealing with the intratumoral microbiota. In chapter 2, 3 and 5 I carefully applied methodologies to mitigate the effect of contamination on the microbiome data under study. The intratumoral microbiome of almost all cancers surveyed would necessarily be derived from samples with a lower biomass than oesophageal biopsies, colonic biopsies, and skin swabs. Studies reporting taxa which compose the intratumoral microbiome report taxa indicative of contamination such as *Pseudomonas, Sphingomonas*, *Shewanella,* and *Photobacterium,* even though these studies seek to address contamination[14,63]. Still other studies do not give sufficient care to the issue of contamination. A number of recent reports have published data that one would regard as clear indications of contamination. Thyagarajan et al reported that, in terms of relative abundance, *Ralstonia* was the dominant bacterial genus in biopsies in derived from breast tumours and healthy breast tissue[65]. In another study which aimed to define the microbiome of three adipose tissue deposits as well as the liver and plasma of morbidly obese individuals, the authors went so far to propose "… environmental bacteria—and/or their fragments—that are present in food and water can accumulate in the MAT[mesenteric adipose tissue] and may affect blood glucose regulation"[66]. A more in-depth critical analysis to rule out contamination is needed before one can start discussing the potential mechanistic implications of the presence of microbes in certain tissues.

Further methodological improvements need to be developed and implemented to address the issue of contamination. Negative controls in the form of mock extractions may not be suitable to detect reagent related contamination. DNA extraction protocols using the silica-based method, such as those included in commercially available QIAGEN kits, can be limited in their efficacy by very low quantise of starting template DNA. Although a biopsy sample might contain relatively small microbiota levels, this would usually have an abundance of human

DNA. Thus, a clinical sample used in a DNA extraction may be more prone to suffering from contamination. The use of carrier DNA has been shown to increase DNA extraction efficacy and has been utilized to address contamination in relation to ancient DNA analysis[67]. Thus, carrier DNA may enhance the ability of negative kit controls to detect contamination.

6.3 Concluding remarks.

The research undertaken in this thesis hopefully contributes to our understanding of the relationship between the microbiome and cancer. Chapter 2 offers one of the most in-depth studies describing the oesophago-gastric mucosal microbiome in the context of the oesophageal adenocarcinoma sequence. Information gleaned may provide avenues to develop diagnostic tools but also to provide the associations needed to inform mechanistic studies. In chapter 3, we further strengthen the hypothesis that colorectal cancer is associated with a change along the whole colon, and it is not restricted to the tumour. However, taxa such as *F. nucleatum* can be observed to be differentially abundant between tumour and matched healthy tissue. In chapter 3 we identified a number of taxa associated with response to ICI. While this thesis presented multiple novel findings, the global thesis findings also corroborate other studies which were carried out in other geographical settings. Taken together this suggests a level of robustness in the microbiome alterations identified. In chapter 5 we identified microbiome changes which may explain, in part, the inflammatory phenotype observed in HS while also providing a microbiome-based explanation for the comorbidity between HS and Crohn's disease. Further, due to the link between inflammation and cancer, the difference alteration in the microbiome of individuals with HS may explain the increase relative risk of cancer. As is with the nature of science, these chapters open more questions which will need to be answered by future studies.

6.4 References

1 Reid, M. J. *et al.* Building a tuberculosis-free world: The Lancet Commission on tuberculosis. *The Lancet* **393**, 1331-1384 (2019).

2 Dagenais, G. R. *et al.* Variations in common diseases, hospital admissions, and deaths in middle-aged adults in 21 countries from five continents (PURE): a prospective cohort study. *The Lancet* **395**, 785-794 (2020).

3 Eckhouse, S., Lewison, G. & Sullivan, R. Trends in the global funding and activity of cancer research. *Mol Oncol* **2**, 20-32, doi:10.1016/j.molonc.2008.03.007 (2008).

4 Hawkes, N. (2019).

5 Kleeff, J. *et al.* Pancreatic cancer. *Nature reviews Disease primers* **2**, 1-22 (2016).

6 Thrift, A. P. Global burden and epidemiology of Barrett oesophagus and oesophageal cancer. *Nature Reviews Gastroenterology & Hepatology*, 1-12 (2021).

7 Gerlinger, M. *et al.* Cancer: evolution within a lifetime. *Annual review of genetics* **48**, 215-236 (2014).

8 Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646-674, doi:10.1016/j.cell.2011.02.013 (2011).

9 Hanahan, D. & Weinberg, R. A. The hallmarks of cancer. *Cell* **100**, 57-70 (2000).

10 Rozhok, A. I. & DeGregori, J. Toward an evolutionary model of cancer: Considering the mechanisms that govern the fate of somatic mutations. *Proceedings of the National Academy of Sciences* **112**, 8914-8921 (2015).

11 Laconi, E., Marongiu, F. & DeGregori, J. Cancer as a disease of old age: Changing mutational and microenvironmental landscapes. *British journal of cancer* **122**, 943-952 (2020).

12 Rozhok, A. I. & DeGregori, J. The evolution of lifespan and age-dependent cancer risk. *Trends in cancer* **2**, 552-560 (2016).

13 Sender, R., Fuchs, S. & Milo, R. Revised Estimates for the Number of Human and Bacteria Cells in the Body. *PLoS Biol* **14**, e1002533, doi:10.1371/journal.pbio.1002533 (2016).

14 Nejman, D. *et al.* The human tumor microbiome is composed of tumor type–specific intracellular bacteria. *Science* **368**, 973-980 (2020).

15 Dohlman, A. B. *et al.* The cancer microbiome atlas: a pan-cancer comparative analysis to distinguish tissue-resident microbiota from contaminants. *Cell Host & Microbe* **29**, 281-298. e285 (2021).

16 Poore, G. D. *et al.* Microbiome analyses of blood and tissues suggest cancer diagnostic approach. *Nature* **579**, 567-574 (2020).

17 Shanahan, F. & O'toole, P. W. Host–microbe interactions and spatial variation of cancer in the gut. *Nature Reviews Cancer* **14**, 511-512 (2014).

18 Kadosh, E. *et al.* The gut microbiome switches mutant p53 from tumour-suppressive to oncogenic. *Nature* **586**, 133-138 (2020).

19 Grivennikov, S. I., Greten, F. R. & Karin, M. Immunity, inflammation, and cancer. *Cell* **140**, 883-899 (2010).

20 McColl, K. E. Helicobacter pylori infection. *New England Journal of Medicine* **362**, 1597-1604 (2010).

21 Perz, J. F., Armstrong, G. L., Farrington, L. A., Hutin, Y. J. & Bell, B. P. The contributions of hepatitis B virus and hepatitis C virus infections to cirrhosis and primary liver cancer worldwide. *Journal of hepatology* **45**, 529-538 (2006).

22 Olén, O. *et al.* Colorectal cancer in ulcerative colitis: a Scandinavian population-based cohort study. *The Lancet* **395**, 123-131 (2020).

23 Piao, J., Lee, E. J. & Lee, M. Association between pelvic inflammatory disease and risk of ovarian cancer: An updated meta-analysis. *Gynecologic oncology* **157**, 542-548 (2020).

24 Catassi, C., Bearzi, I. & Holmes, G. K. Association of celiac disease and intestinal lymphomas and other cancers. *Gastroenterology* **128**, S79-S86 (2005).

25 Caruso, R., Lo, B. C. & Núñez, G. Host–microbiota interactions in inflammatory bowel disease. *Nature Reviews Immunology* **20**, 411-426 (2020).

26 Brunham, R. C., Gottlieb, S. L. & Paavonen, J. Pelvic inflammatory disease. *New England Journal of Medicine* **372**, 2039-2048 (2015).

27 Valitutti, F., Cucchiara, S. & Fasano, A. Celiac disease and the microbiome. *Nutrients* **11**, 2403 (2019).

28 Jung, J. M. *et al.* Assessment of Overall and Specific Cancer Risks in Patients With Hidradenitis Suppurativa. *JAMA Dermatol* **156**, 844-853, doi:10.1001/jamadermatol.2020.1422 (2020).

29 Olén, O. *et al.* Colorectal cancer in Crohn's disease: a Scandinavian population-based cohort study. *The Lancet Gastroenterology & Hepatology* **5**, 475-484 (2020).

30 Zhao, R. *et al.* Crohn’s disease instead of UC might increase the risk of small bowel cancer. *Gut* **70**, 809-810 (2021).

31 Jin, C. *et al.* Commensal microbiota promote lung cancer development via γδ T cells. *Cell* **176**, 998-1013. e1016 (2019).

32 Heilbronner, S., Krismer, B., Brötz-Oesterhelt, H. & Peschel, A. The microbiome-shaping roles of bacteriocins. *Nature Reviews Microbiology*, 1-14 (2021).

33 Bilinski, J. *et al.* Fecal microbiota transplantation in patients with acute and chronic graft-versus-host disease-spectrum of responses and safety profile. Results from a prospective, multicenter study. *American journal of hematology* **96**, E88-E91 (2021).

34 Dalmasso, M. *et al.* Three new Escherichia coli phages from the human gut show promising potential for phage therapy. *PloS one* **11**, e0156773 (2016).

35 Sokol, H. *et al.* Fecal microbiota transplantation to maintain remission in Crohn's disease: a pilot randomized controlled study. *Microbiome* **8**, 1-14 (2020).

36 Flemer, B. *et al.* The oral microbiota in colorectal cancer is distinctive and predictive. *Gut* **67**, 1454-1463, doi:10.1136/gutjnl-2017-314814 (2018).

37 Thomas, A. M. *et al.* Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. *Nature medicine* **25**, 667-678 (2019).

38 Wirbel, J. *et al.* Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. *Nat Med* **25**, 679-689, doi:10.1038/s41591-019-0406-6 (2019).

39 Ghosh, T. S., Das, M., Jeffery, I. B. & O'Toole, P. W. Adjusting for age improves identification of gut microbiome alterations in multiple diseases. *Elife* **9**, e50240 (2020).

40 Sikkema, M. *et al.* Aneuploidy and overexpression of Ki67 and p53 as markers for neoplastic progression in Barrett's esophagus: a case-control study. *American Journal of Gastroenterology* **104**, 2673-2680 (2009).

41 Jin, Z. *et al.* A multicenter, double-blinded validation study of methylation biomarkers for progression prediction in Barrett's esophagus. *Cancer research* **69**, 4112-4115 (2009).

42 Killcoyne, S. *et al.* Genomic copy number predicts esophageal cancer years before transformation. *Nature medicine* **26**, 1726-1732 (2020).

43 Martinez, P. *et al.* Dynamic clonal equilibrium and predetermined cancer risk in Barrett's oesophagus. *Nature communications* **7**, 1-10 (2016).

44 Sepich-Poore, G. D. *et al.* The microbiome and human cancer. *Science* **371** (2021).

45 McCarthy, E. F. The toxins of William B. Coley and the treatment of bone and soft-tissue sarcomas. *The Iowa orthopaedic journal* **26**, 154 (2006).

46 Ghosh, T. S., Arnoux, J. & O'Toole, P. W. Metagenomic analysis reveals distinct patterns of gut lactobacillus prevalence, abundance, and geographical variation in health and disease. *Gut microbes* **12**, 1822729 (2020).

47 Keohane, D. M. *et al.* Microbiome and health implications for ethnic minorities after enforced lifestyle changes. *Nature Medicine* **26**, 1089-1095 (2020).

48 He, Y. *et al.* Regional variation limits applications of healthy gut microbiome reference ranges and disease models. *Nature medicine* **24**, 1532-1535 (2018).

49 Baruch, E. N. *et al.* Fecal microbiota transplant promotes response in immunotherapy-refractory melanoma patients. *Science* **371**, 602-609 (2021).

50 Davar, D. *et al.* Fecal microbiota transplant overcomes resistance to anti–PD-1 therapy in melanoma patients. *Science* **371**, 595-602 (2021).

51 Sun, S. *et al.* Bifidobacterium alters the gut microbiota and modulates the functional metabolism of T regulatory cells in the context of immune checkpoint blockade. *Proceedings of the National Academy of Sciences* **117**, 27509-27515 (2020).

52 Lee, S.-H. *et al.* Bifidobacterium bifidum strains synergize with immune checkpoint inhibitors to reduce tumour burden in mice. *Nature Microbiology* **6**, 277-288 (2021).

53 Wang, L. *et al.* A purified membrane protein from Akkermansia muciniphila or the pasteurised bacterium blunts colitis associated tumourigenesis by modulation of CD8+ T cells in mice. *Gut* **69**, 1988-1997 (2020).

54 Plovier, H. *et al.* A purified membrane protein from Akkermansia muciniphila or the pasteurized bacterium improves metabolism in obese and diabetic mice. *Nature Medicine* **23**, 107-113, doi:10.1038/nm.4236 (2017).

55 Lauté-Caly, D. L. *et al.* The flagellin of candidate live biotherapeutic Enterococcus gallinarum MRx0518 is a potent immunostimulant. *Scientific reports* **9**, 1-14 (2019).

56 Stanley, M. Tumour virus vaccines: hepatitis B virus and human papillomavirus. *Philosophical Transactions of the Royal Society B: Biological Sciences* **372**, 20160268 (2017).

57 Hollingsworth, R. E. & Jansen, K. Turning the corner on therapeutic cancer vaccines. *npj Vaccines* **4**, 1-10 (2019).

58 Schumacher, T. *et al.* A vaccine targeting mutant IDH1 induces antitumour immunity. *Nature* **512**, 324-327 (2014).

59 Platten, M. *et al.* A vaccine targeting mutant IDH1 in newly diagnosed glioma. *Nature*, 1-6 (2021).

60 Sahin, U. *et al.* An RNA vaccine drives immunity in checkpoint-inhibitor-treated melanoma. *Nature* **585**, 107-112 (2020).

61 Shekarian, T. *et al.* Repurposing rotavirus vaccines for intratumoral immunotherapy can overcome resistance to immune checkpoint blockade. *Science translational medicine* **11** (2019).

62 Ali, O. A., Lewin, S. A., Dranoff, G. & Mooney, D. J. Vaccines combined with immune checkpoint antibodies promote cytotoxic T-cell activity and tumor eradication. *Cancer immunology research* **4**, 95-100 (2016).

63 Kalaora, S. *et al.* Identification of bacteria-derived HLA-bound peptides in melanoma. *Nature*, 1-6 (2021).

64 Gur, C. *et al.* Binding of the Fap2 protein of Fusobacterium nucleatum to human inhibitory receptor TIGIT protects tumors from immune cell attack. *Immunity* **42**, 344-355 (2015).

65 Thyagarajan, S. *et al.* Comparative analysis of racial differences in breast tumor microbiome. *Scientific reports* **10**, 1-13 (2020).

66 Anhê, F. F. *et al.* Type 2 diabetes influences bacterial tissue compartmentalisation in human obesity. *Nature Metabolism* **2**, 233-242 (2020).

67 Xu, Z. *et al.* Improving the sensitivity of negative controls in ancient DNA extractions. *Electrophoresis* **30**, 1282-1285 (2009).

Appendix 1-Comparative exome analysis of mutational processes in colorectal cancers from patients harbouring two divergent gut microbiota types.

7.1 Abstract

Like other cancers, colorectal cancers (CRC) develop through a process that involves Darwinian selection acting upon somatic mutations in cancer cells. There is mounting evidence of a significant role for the colonic microbiota in the development, progression and treatment of CRC. Previously we defined six microbiota subtypes whose abundance was differentially associated with CRC or healthy controls. To explore the microbiota as an environmental driver of mutation, CRC exome sequence data was generated from six subjects, three from each of two distinct colonic microbiota subtypes dominated by either phylum Firmicutes or genus Prevotella. No significant differences in the somatic exonic mutational landscape were identified between the microbiota-defined groups. However, there was a non-significantly higher mutational burden and greater representation of mutational signature 5 in the Prevotella microbiota subtype tissue samples, which may reflect an underlying biological mechanism.

7.2 Introduction

Colorectal cancer kills almost 700,000 people a year worldwide, making it the 4th leading cause of cancer morbidity.[1] As for all cancers, the development of CRC is an evolutionary process enabled by somatic mutation.[2] Tomasetti and Vogelstein showed that a major source of mutations (~66%) is stochastic DNA replication errors, [3,4] and indeed hydrolytic deamination of 5-methylcytosine, tautomeric mispairs and anionic mispairs are seemingly inevitable aspects of DNA biology.[5-8] Nonetheless, dietary and lifestyle variables including wholegrain consumption, alcohol, calcium intake, smoking and consumption of processed meat and red meat are other plausible sources of mutagens, either directly or as a consequence of gut bacterial processing.[4,9-11]

The human microbiota is increasingly recognised as playing a role in human health and disease.[12] The greatest density of microbiota resides in the colon with an estimated $9x10^{10}$ bacteria per gram of wet stool.[13] A growing body of evidence implicates the colonic microbiota in CRC development.[14] Using hierarchical clustering techniques, we previously identified six mucosal-associated bacterial co-abundance groups (CAGs) that are differentially represented in CRC patients compared to controls.[15,16] These CAGs resemble previously described enterotypes.[17,18] Categorization of the gut microbiome into subtypes as described by CAGs or enterotypes allows for the separation/stratification of cohorts into defined groups. These groupings enable study design to interrogate the microbiota configuration as a whole rather than focusing on individual elements such as taxa.

The mutagenic influence of the microbiota occurs through multiple mechanisms.[19] It is reasonable to hypothesize that microbiota subtypes may contribute varying degrees of risk/protection by varying the extent to which they promote or protect against somatic mutation. Gastrointestinal microbe-derived genotoxins such as cytolethal distending toxin (produced by an array of gram-negative bacteria within the gamma and epsilon classes of the phylum Proteobacteria) and colibactin (produced by pks+ strains of *Escherichia coli*) induce double stand breaks.[20-24] The immune system may be stimulated by microbes in a manner that leads to DNA damage. *Enterococcus faecalis*-generated superoxide radicals can activate macrophage cyclooxygenase-2 expression leading to the production of genotoxic trans-4-hydroxy-2-nonenal, which in turns causes chromosomal instability (CIN). [25] Finally, intestinal microbes have been shown to influence DNA damage repair.[26] *Helicobacter pylori* and enteropathogenic *E. coli* both down-regulate the expression of mismatch repair proteins including MSH2 and MLH1, thus compromising host genome integrity.[19,27-29] The gut microbiota may modulate stochastically generated DNA aberrations by influencing their repair.

The characteristics of the mutations in a cancer genome are indicative of the mutational mechanisms which caused those mutations. For example, C>T transversions at CpG dinucleotides are indicative of spontaneous deamination of 5-methylcytosine.[30] Thus, interrogation of the cancer genome can yield information on the different mutational mechanisms which acted upon the cancer genome during its evolution. Recent developments in methods, namely those designed to extract so-called mutational signatures, allow an in-depth interrogation of the cancer genome regarding mutational mechanism.[31]

In this pilot study, we performed whole exome sequencing on paired cancer/ normal colorectal biopsy samples. Samples were derived from 6 individuals, 3 individuals from each of the two well categorized microbiota configurations, Firmicutes subtype and Prevotella subtype.[15,16] We investigated the genomic architecture of these two groups in terms of somatic nucleotide variants (SNVs) and copy number alterations (CNA). The data-sets are a preliminary resource for studying the relationship between the gut microbiome and host genome stability while providing supportive impetus for further investigations of the microbiota-host genome interaction in cancer.

7.3 Results

We previously identified consortia of gut microbial taxa whose abundances co-vary, labelled co-abundance groups (CAGs). [15,16] Such definition of the structure of the colonic microbiota allows reduction of dimensionality in microbiota research. We have thus used this methodology to categorise individuals into microbiota subtypes.

The relative abundance of these CAGs within the colonic mucosal microbiota distinguished colorectal cancer cases from those of controls.[15] With explicit relevance to this study, ‘Firmicutes 1’ CAG was over-represented in healthy individuals while the ‘Prevotella’ CAG was over-represented in individuals with colorectal cancer. We sought to determine if subjects with cancer belonging to these two microbiota subtypes had relevant mutational difference in their genomes. We identified 3 individuals whose colonic mucosal microbiota was dominated by either ‘Firmicutes 1’ CAG and 3 dominated by CAG ‘Prevotella’ CAG (**Figure 1**). These individuals were chosen to represent the most typical bacterial taxonomic profiles for the respective CAGs. Such a selection likely compensates for the limiting effect of small numbers through minimises within group variance.

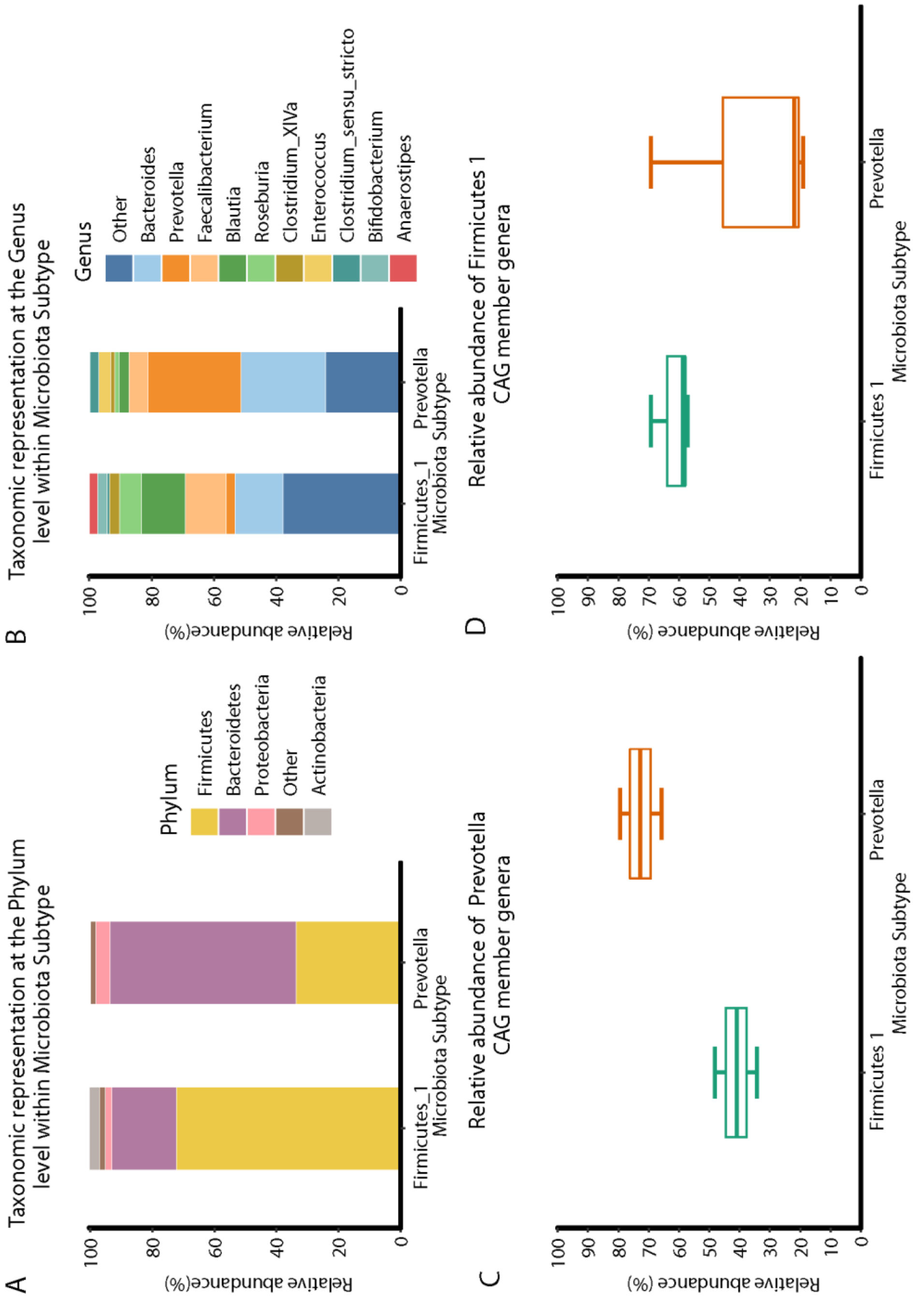


Figure 1| Mucosal microbiota composition in CRC patients of divergent microbiota subtype. Data shown are proportional abundance of the indicated bacterial taxa in the colonic mucosal microbiota averaged across the 3 subjects per microbiota subtype. Bar plot of the relative proportions of indicated taxa at the phylum (panel A) and genus level (panel B) in each microbiota subtype group. Box plot summing the contribution of members of the genera of Prevotella (panel C) and Firmicutes (panel D).

Mutational burden in tumours from different microbiota subtypes

Tumour mutational burden (TMB), the total number of mutations per coding megabase of a tumour genome, is recognised as an indicator of cancer history as well as a prognostic marker, particularly in relationship to immunotherapy.[32-35] Recent studies have identified the gut microbiota as a modulator of immunotherapy.[36-39] Given these links, we sought to examine the relationship between defined microbiota subtypes and TMB. We analysed whole exome sequence data from paired tumour/normal tissue. Somatic mutations were called, filtered and quantified per exome. We performed a bivariate analysis on TMB versus microbiota subtypes taking into account sequencing depth. Although this analysis revealed a higher TMB in subjects from the Prevotella group relative to the Firmicutes 1 group, the difference did not reach statistical significance due to low sample number and within-subtype variation (**Figure 2.A**).

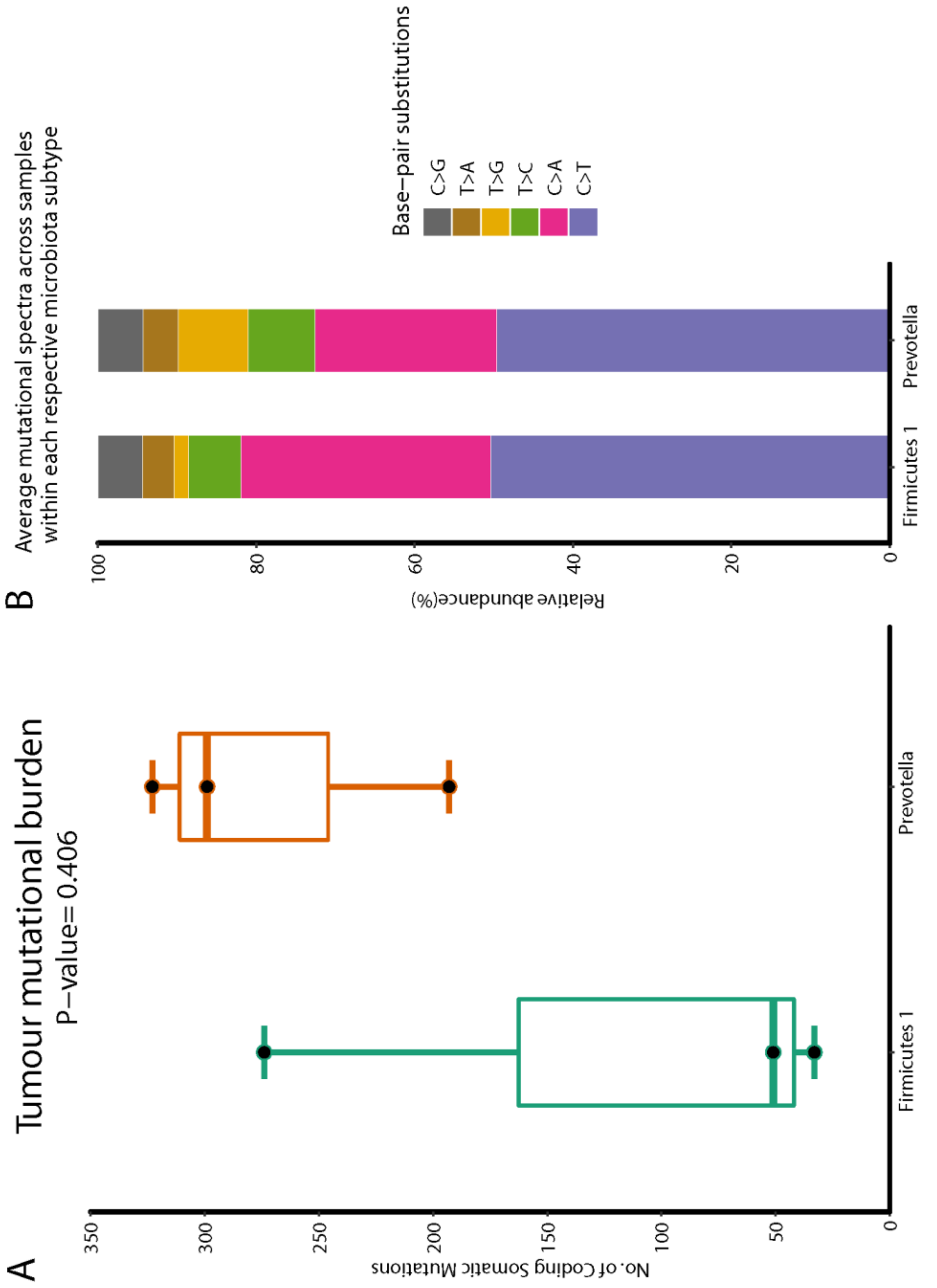


Figure 2| Tumour mutational burden and proportional representation of base substitutions between microbiota subtypes. A| Box plots of the abundance and distribution of the TMB with in each microbiota subtype. Y- axis shows absolute count of somatic base substitutions within the exome. **B|** Bar plots indicate the relative abundance of each base substitution type [note: In accordance with the Catalogue of Somatic Mutations in Cancer (COSMIC) system all substitutions are referred to by the pyrimidine of the mutated Watson-Crick base pair].[30]

Microbiota associations of mutational spectra and signatures

To generate an overview of the genomic dynamics of the tumours, we identified and compared the mutation spectra of the samples in this study (**Figure 2.B**). Overall the mutational spectra obtained were typical of previously described spectra for CRC.[40] C>T transitions were slightly more represented in the Firmicutes 1 subtype tumours while T>G transversions were more common in the Prevotella group.With respect to the six classes of base pair substitutions, and the microbiota subtypes, we identified no gross difference in mutation spectra.

We fitted the mutational matrices of the samples to previously defined COSMIC mutational signatures (**Figure 3.A**) limiting them to signatures previously described in CRC which are known to act in a clock-like manner (signature 1 and 5).[31,41] Further, we identified the fit of the model of contributions and the residuals (**Figure 3.B**). The relative contributions of the mutational signatures were somewhat typical of previous reports.[31,41] We did not detect any statistically significant difference between the association of the microbiota subtypes and the fitted COSMIC mutational signatures. Mutational signature 5 showed a non-statically significant increased relative frequency in the Prevotella group.

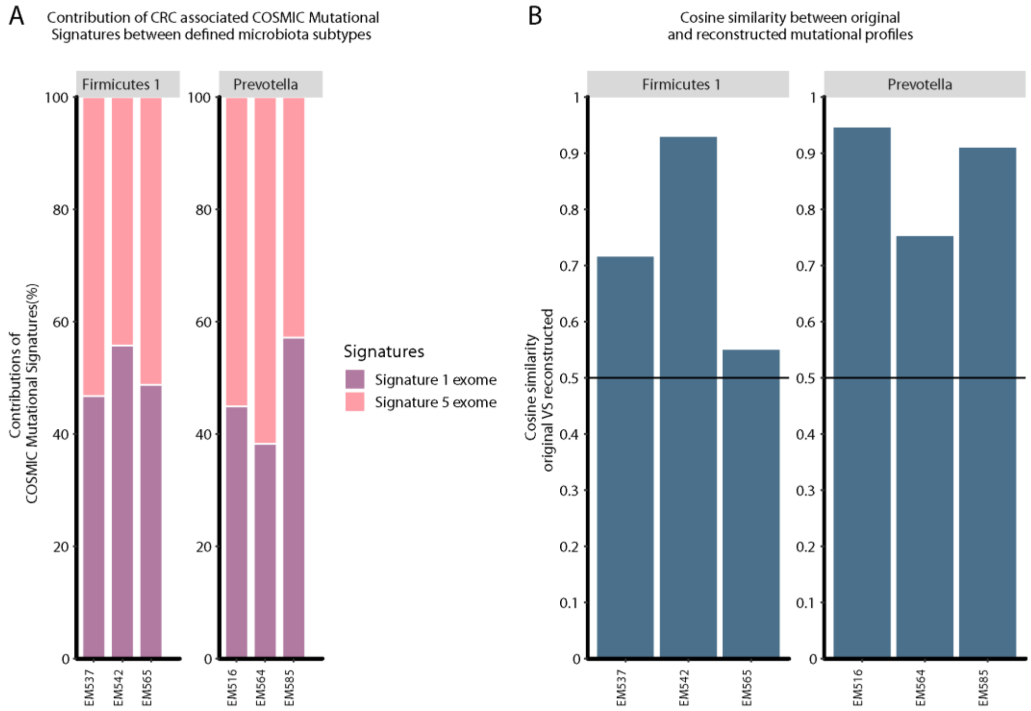


Figure 3: Proportional representation of Mutational signatures and cosine similarity between original and reconstructed mutational profiles. A| Bar plots of relative contribution of fitted COSMIC mutational signatures. **B|** Bar plots showing the level in which the samples’ mutational matrices can be recreated with fitted signatures.

We sought to cluster samples based on the contributions of the defined COSMIC mutational signatures to the mutational profile of the samples. In brief, we measured the ability of COSMIC Mutational Signatures to explain the 96 trinucleotide mutation matrix of the tumour genomes by calculating cosine similarity. Cosine similarity was used to perform complete clustering and the results are visualized on a heat-map (**Figure 4**). This clustering provides an easy method to visualize the similarities between samples with regard to their mutational portrait. Samples clustered based on the number of somatic variants present. Clustering based on mutational signature did not co-segregate with clustering based on microbiota subtype.

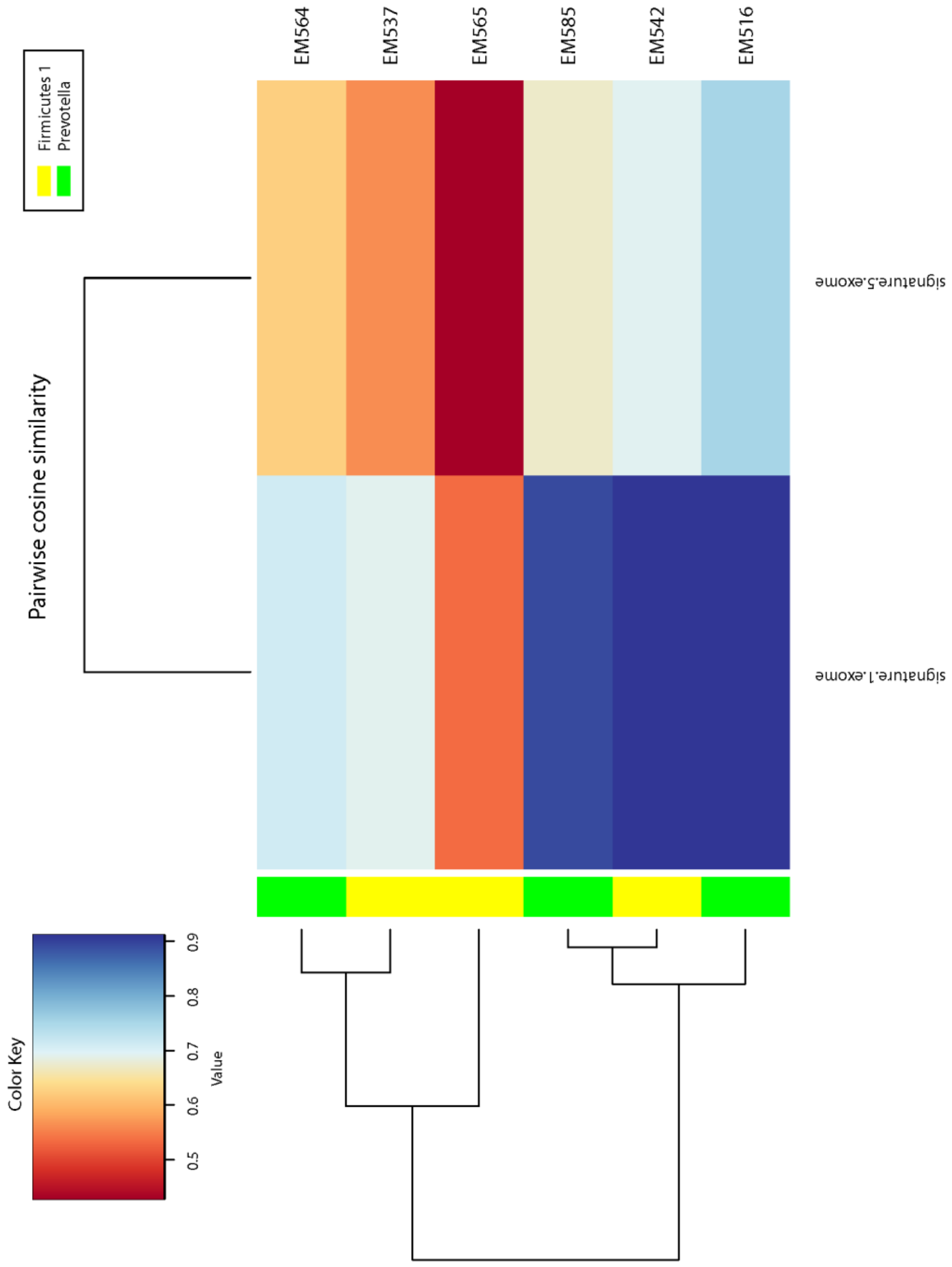
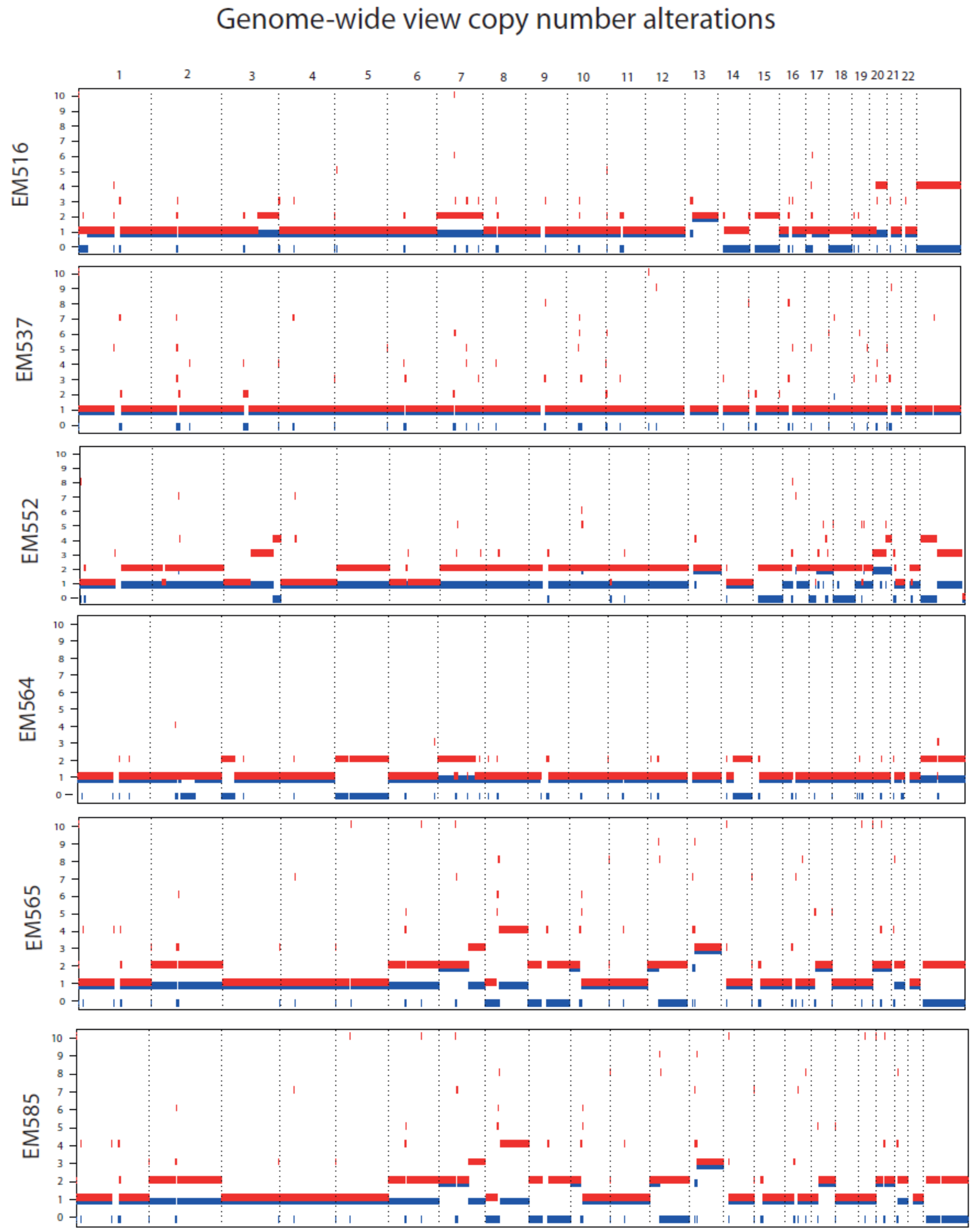


Figure 4| Heat map of pairwise cosine similarity between mutational profiles. The degree to which the 96 trinucleotide mutation count was attributed to the COSMIC mutational was obtained through calculating Cosine similarity. Hierarchical clustering of the cosine similarity was performed using complete linkage clustering. Colour bars to left of figure indicate Firmicutes 1 (yellow) and Prevotella (green).

Copy number variation is independent of microbiota subtype

Aneuploidy is a feature of the majority of CRC genomes and has been identified as a prognostic marker.[42] The R package Sequenza was used to infer copy-number alteration from the exome sequencing data. CNV did not statistically vary with respect to microbiota subtype (**Supplmenetary Figure 1**).



Supplementary figure 1 | Genome-wide view of CNA

Discussion

This analysis set out to test for interaction between the genetic architecture of colorectal cancer and the neighbouring colonic mucosal microbiota. We investigated the relationship between various features of the cancer genome including TMB, mutational signatures and copy number alterations. Although none of these features separate to an extent that reached significance, we did observe suggestive trends for of microbiota-host-genome interaction. Most notably was the trend towards higher TMB in the Firmicutes subtype.

Recent studies have clearly identified the intestinal microbiota as modifying the efficacy of cancer immunotherapy. [36-39] The increased abundance of certain taxa (Ruminococcaceae , Faecalibacterium, Bifidobacteria, Alistipes, Enterococci, Collinsella) and higher microbiota diversity have been linked with positive response to immune checkpoint blockade treatment. Neoplasms develop through the accumulations of somatic variants, particularly in driver genes, which stimulate the evolution of healthy cells to cancer cells. It is possible that individuals that have a dominance of the Prevotella CAG in their gut microbiota experience increased mutagenic stress on their colonic cells and a correspondingly increased level of TMB. Thus the Prevotella CAG would be overrepresented in individuals with CRC. Somewhat paradoxically, provided a sufficient mutational effect, these individuals may have a better response rate to cancer immunotherapy because of higher level production of neoantigens.

None of the COSMIC signatures we identified exhibited a bias of representation with regard to microbiota defined groups. An increased contribution of signature 5 to the mutational portrait was observed (though was not significant) in patients whose tumour microbiota was dominated by the Prevotella CAG. The most prominent feature of mutational signature 5 is a transcriptional strand bias for T>C substitutions at the ApTpN context. A current model for the origin of mutational signature 5 involves deletion of the FHIT gene which leads to the down regulation of Thymidine Kinase 1 (TK1) expression and a reduction in thymidine triphosphate pool levels.[43] Such a decrease in of dTTP levels would lead to an increased ratio of dUTP:TTP thereby increasing the likelihood of dUTP misincorporation (U:A) in place of TTP. An abasic site may then arise during the base excision repair (BER) pathway. Certain translesion polymerase activity could incorporate guanine or a cytosine across from an abasic site, ultimately leading to T>C or a T>G base substitution during subsequent S phase. Notably, the intestinal microbiota is known to influence the activity of various host enzymes and proteins. In one mouse study examining the differential activity of various enzymes between germ free and normal mice, it was found that the presence of a microbiota reduced the activity of thymidine kinase by 50%.[44] Thus, it is reasonable to postulate that the metabolic activity of the intestinal microbiota influences genome instability such as that induced by FHIT deletion. In terms of candidate mechanisms, it is also possible that certain microbiota compositions have specific or greater magnitude of influence upon the regulation of expression of particular colonic cell proteins. Individuals in the current study whose microbiota was dominated by Prevotella may have experienced greater dysregulation of genome integrity leading to an increased prevalence of COSMIC mutational signature 5.

This study provides suggestive evidence for the interaction between the gut microbiota and host genome stability. The gut microbiota is readily accessible to observation as well as intervention. The interrogation of the gut microbiota has been shown as a credible method for diagnosing CRC.[15,16,45] It could also be possible to derive added information from the microbiota with regard to the genomic architecture of a tumour. This data would inform the choice of further testing as well as therapeutics such as immunotherapy. Moreover, provided there is a direct causative effect of the microbiome in shaping the cancer genome and thus oncogenesis, one could devise strategies to intervene and alter the microbiota in a prophylactic manner. Finally, cancer therapeutics strategies have been devised that target DNA damage response (DDR).[46-48] Gastrointestinal microbes are known to localise to CRC tumours as well as to interact with host DDR.[19,49-51] It is conceptually possible to use microbes as a DDR centric therapeutic.

7.4 methods

7.4.1 Recruitment and sample acquisition

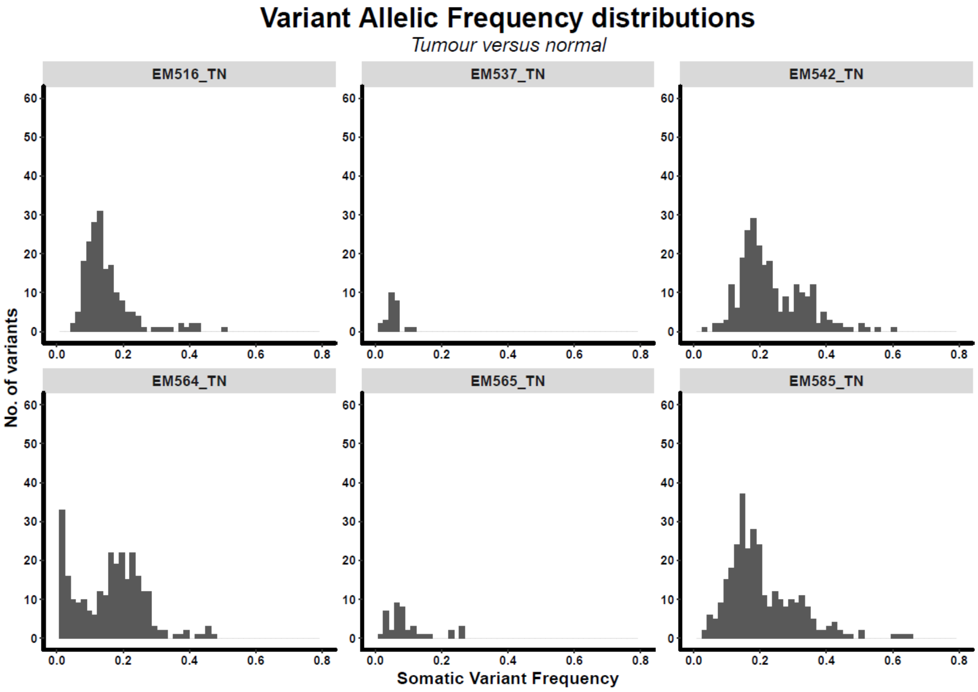
Biological samples were obtained as described in previous studies.[15,16] In brief, individuals were recruited from a cohort scheduled to undergo colonic resection at Mercy University Hospital, Cork, Ireland. Exclusion criteria included no personal history of Irritable Bowel Syndrome or Inflammatory Bowel Disease and no treatment with antibiotics in the past month. Neoplasms and healthy samples were dissected from surgical restricted colon. Samples were placed in 3 mL RNAlater, stored at 4°C for 12 h and then stored at −20°C post-surgery.

7.4.2 DNA extraction and whole exome sequencing (WES)

Genomic DNA was extracted from biopsies using the AllPrep DNA/RNA kit from Qiagen as previously described[15]. DNA concentration was quantified by measuring the 260/280 nm and 260/230 nm ratios with an ND1000 spectrophotometer (Nanodrop Technologies, ThermoFisher). Exome capture was performed using Sureselect Human All Exon V5. Pair-end reads of length 101bp were produced on the Illumina HiSeq4000 platform (mean/median 100X raw data coverage).

7.4.3 WES pipeline: somatic SNV calling

WES reads were aligned to the reference human genome GRCh37 using BWA MEM-mem.[52] Using the Picard tools (v.2.6.0), BAM files were sorted and duplicates marked thereby producing analysis-ready files (**Supplementary figure 2**). The somatic variant caller Mutect2, within the Genome Analysis Toolkit (GATK, v3.7) suite of tools, was used to call somatic variants by comparing BAM files from tumour and matched normal samples.[53] The confidence of somatic variants was weighted within the calling, using the Single Nucleotide Polymorphism Database (dbSNP, v138) and the Catalogue of Somatic Mutations in Cancer (COSMIC, v54).[54,55] SNVs were further filtered on the criteria that at least 3 reads supported the variant in the tumour sample and at least 10 reads covered the variant in in both tumour and normal samples.



Supplementary figure 2 | Variant allele frequency distribution plot.

7.4.4 Mutational signature analysis

Mutational signature analysis was performed using the R package MutationalPatterns (version 1.6.1).[56] Known COSMIC mutational signatures which occur in CRC were fitted to the mutational profile of the samples. Trinucleotide counts within COSMIC mutational signatures were normalized by the number of times each trinucleotide context was observed in the exome region relative to the whole genome.

7.4.5 Copy number variation

Copy number variation was derived from the exome sequence data using Sequenza.[57] Further, tumour purities and ploidies were calculated using Sequenza with default parameters.

7.5 Acknowledgments

This work was supported by The Health Research Board of Ireland under Grant ILP-POR-2017-034, and by Science Foundation Ireland (APC/SFI/12/RC/2273) in the form of a research centre, the APC Microbiome Institute. We thank Prof Ian Sanderson (QMUL) for facilitating the collaboration between the Microbial Genomics Laboratory (UCC) and the Evolution and Cancer Laboratory (QMUL).

7.6 Disclosure of interest

No potential conflict of interest was reported by the authors.

7.7 References

1 Ferlay, J. *et al.* Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Int J Cancer* **136**, E359-386, doi:10.1002/ijc.29210 (2015).

2 Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458**, 719-724, doi:10.1038/nature07943 (2009).

3 Tomasetti, C. & Vogelstein, B. Cancer etiology. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. *Science* **347**, 78-81, doi:10.1126/science.1260825 (2015).

4 Tomasetti, C., Li, L. & Vogelstein, B. Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention. *Science* **355**, 1330-1334, doi:10.1126/science.aaf9011 (2017).

5 Duncan, B. K. & Miller, J. H. Mutagenic deamination of cytosine residues in DNA. *Nature* **287**, 560-561 (1980).

6 Lewis, C. A., Jr., Crayle, J., Zhou, S., Swanstrom, R. & Wolfenden, R. Cytosine deamination and the precipitous decline of spontaneous mutation during Earth's history. *Proc Natl Acad Sci U S A* **113**, 8194-8199, doi:10.1073/pnas.1607580113 (2016).

7 Kimsey, I. J., Petzold, K., Sathyamoorthy, B., Stein, Z. W. & Al-Hashimi, H. M. Visualizing transient Watson-Crick-like mispairs in DNA and RNA duplexes. *Nature* **519**, 315-320, doi:10.1038/nature14227 (2015).

8 Kimsey, I. J. *et al.* Dynamic basis for dG*dT misincorporation via tautomerization and ionization. *Nature* **554**, 195-201, doi:10.1038/nature25487 (2018).

9 Theodoratou, E., Timofeeva, M., Li, X., Meng, X. & Ioannidis, J. P. A. Nature, Nurture, and Cancer Risks: Genetic and Nutritional Contributions to Cancer. *Annu Rev Nutr* **37**, 293-320, doi:10.1146/annurev-nutr-071715-051004 (2017).

10 Cheng, J. *et al.* Meta-analysis of prospective cohort studies of cigarette smoking and the incidence of colon and rectal cancers. *Eur J Cancer Prev* **24**, 6-15, doi:10.1097/CEJ.0000000000000011 (2015).

11 Wolin, K. Y., Yan, Y., Colditz, G. A. & Lee, I. M. Physical activity and colon cancer prevention: a meta-analysis. *Br J Cancer* **100**, 611-616, doi:10.1038/sj.bjc.6604917 (2009).

12 Lynch, S. V. & Pedersen, O. The Human Intestinal Microbiome in Health and Disease. *N Engl J Med* **375**, 2369-2379, doi:10.1056/NEJMra1600266 (2016).

13 Sender, R., Fuchs, S. & Milo, R. Revised Estimates for the Number of Human and Bacteria Cells in the Body. *PLoS Biol* **14**, e1002533, doi:10.1371/journal.pbio.1002533 (2016).

14 Tilg, H., Adolph, T. E., Gerner, R. R. & Moschen, A. R. The Intestinal Microbiota in Colorectal Cancer. *Cancer Cell* **33**, 954-964, doi:10.1016/j.ccell.2018.03.004 (2018).

15 Flemer, B. *et al.* Tumour-associated and non-tumour-associated microbiota in colorectal cancer. *Gut* **66**, 633-643, doi:10.1136/gutjnl-2015-309595 (2017).

16 Flemer, B. *et al.* The oral microbiota in colorectal cancer is distinctive and predictive. *Gut* **67**, 1454-1463, doi:10.1136/gutjnl-2017-314814 (2018).

17 Arumugam, M. *et al.* Enterotypes of the human gut microbiome. *Nature* **473**, 174-180, doi:10.1038/nature09944 (2011).

18 Costea, P. I. *et al.* Enterotypes in the landscape of gut microbial community composition. *Nat Microbiol* **3**, 8-16, doi:10.1038/s41564-017-0072-8 (2018).

19 Chumduri, C., Gurumurthy, R. K., Zietlow, R. & Meyer, T. F. Subversion of host genome integrity by bacterial pathogens. *Nat Rev Mol Cell Biol* **17**, 659-673, doi:10.1038/nrm.2016.100 (2016).

20 Bezine, E. *et al.* Cell resistance to the Cytolethal Distending Toxin involves an association of DNA repair mechanisms. *Sci Rep* **6**, 36022, doi:10.1038/srep36022 (2016).

21 Bezine, E., Vignard, J. & Mirey, G. The cytolethal distending toxin effects on Mammalian cells: a DNA damage perspective. *Cells* **3**, 592-615, doi:10.3390/cells3020592 (2014).

22 Jinadasa, R. N., Bloom, S. E., Weiss, R. S. & Duhamel, G. E. Cytolethal distending toxin: a conserved bacterial genotoxin that blocks cell cycle progression, leading to apoptosis of a broad range of mammalian cell lineages. *Microbiology* **157**, 1851-1875, doi:10.1099/mic.0.049536-0 (2011).

23 Bossuet-Greif, N. *et al.* The Colibactin Genotoxin Generates DNA Interstrand Cross-Links in Infected Cells. *Mbio* **9**, doi:ARTN e02393-17

10.1128/mBio.02393-17 (2018).

24 Vizcaino, M. I. & Crawford, J. M. The colibactin warhead crosslinks DNA. *Nat Chem* **7**, 411-417, doi:10.1038/nchem.2221 (2015).

25 Wang, X., Yang, Y. & Huycke, M. M. Commensal bacteria drive endogenous transformation and tumour stem cell marker expression through a bystander effect. *Gut* **64**, 459-468, doi:10.1136/gutjnl-2014-307213 (2015).

26 Sahan, A. Z., Hazra, T. K. & Das, S. The Pivotal Role of DNA Repair in Infection Mediated-Inflammation and Cancer. *Front Microbiol* **9**, 663, doi:10.3389/fmicb.2018.00663 (2018).

27 Kim, J. J. *et al.* Helicobacter pylori impairs DNA mismatch repair in gastric epithelial cells. *Gastroenterology* **123**, 542-553 (2002).

28 Koeppel, M., Garcia-Alcalde, F., Glowinski, F., Schlaermann, P. & Meyer, T. F. Helicobacter pylori Infection Causes Characteristic DNA Damage

Patterns in Human Cells. *Cell Rep* **11**, 1703-1713, doi:10.1016/j.celrep.2015.05.030 (2015).

29 Maddocks, O. D., Scanlon, K. M. & Donnenberg, M. S. An Escherichia coli effector protein promotes host mutation via depletion of DNA mismatch repair proteins. *MBio* **4**, e00152-00113, doi:10.1128/mBio.00152-13 (2013).

30 Helleday, T., Eshtad, S. & Nik-Zainal, S. Mechanisms underlying mutational signatures in human cancers. *Nat Rev Genet* **15**, 585-598, doi:10.1038/nrg3729 (2014).

31 Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415-421, doi:10.1038/nature12477 (2013).

32 Yarchoan, M., Hopkins, A. & Jaffee, E. M. Tumor Mutational Burden and Response Rate to PD-1 Inhibition. *N Engl J Med* **377**, 2500-2501, doi:10.1056/NEJMc1713444 (2017).

33 Pai, S. G. *et al.* Correlation of tumor mutational burden and treatment outcomes in patients with colorectal cancer. *J Gastrointest Oncol* **8**, 858-866, doi:10.21037/jgo.2017.06.20 (2017).

34 Germano, G. *et al.* Inactivation of DNA repair triggers neoantigen generation and impairs tumour growth. *Nature* **552**, 116-120, doi:10.1038/nature24673 (2017).

35 Hellmann, M. D. *et al.* Nivolumab plus Ipilimumab in Lung Cancer with a High Tumor Mutational Burden. *N Engl J Med* **378**, 2093-2104, doi:10.1056/NEJMoa1801946 (2018).

36 Gopalakrishnan, V. *et al.* Gut microbiome modulates response to anti-PD-1 immunotherapy in melanoma patients. *Science* **359**, 97-103, doi:10.1126/science.aan4236 (2018).

37 Matson, V. *et al.* The commensal microbiome is associated with anti-PD-1 efficacy in metastatic melanoma patients. *Science* **359**, 104-108, doi:10.1126/science.aao3290 (2018).

38 Routy, B. *et al.* Gut microbiome influences efficacy of PD-1-based immunotherapy against epithelial tumors. *Science* **359**, 91-97, doi:10.1126/science.aan3706 (2018).

39 Zitvogel, L., Ma, Y., Raoult, D., Kroemer, G. & Gajewski, T. F. The microbiome in cancer immunotherapy: Diagnostic tools and therapeutic strategies. *Science* **359**, 1366-1370, doi:10.1126/science.aar6918 (2018).

40 Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214-218, doi:10.1038/nature12213 (2013).

41 Alexandrov, L. B. *et al.* Clock-like mutational processes in human somatic cells. *Nat Genet* **47**, 1402-1407, doi:10.1038/ng.3441 (2015).

42 Danielsen, H. E., Pradhan, M. & Novelli, M. Revisiting tumour aneuploidy - the place of ploidy assessment in the molecular era. *Nat Rev Clin Oncol* **13**, 291-304, doi:10.1038/nrclinonc.2015.208 (2016).

43 Volinia, S., Druck, T., Paisie, C. A., Schrock, M. S. & Huebner, K. The ubiquitous 'cancer mutational signature' 5 occurs specifically in cancers with deleted FHIT alleles. *Oncotarget* **8**, 102199-102211, doi:10.18632/oncotarget.22321 (2017).

44 Whitt, D. D. & Savage, D. C. Influence of indigenous microbiota on activities of alkaline phosphatase, phosphodiesterase I, and thymidine kinase in mouse enterocytes. *Appl Environ Microbiol* **54**, 2405-2410 (1988).

45 Purcell, R. V., Visnovska, M., Biggs, P. J., Schmeier, S. & Frizelle, F. A. Distinct gut microbiome patterns associate with consensus molecular subtypes of colorectal cancer. *Sci Rep* **7**, 11590, doi:10.1038/s41598-017-11237-6 (2017).

46 O'Connor, M. J. Targeting the DNA Damage Response in Cancer. *Mol Cell* **60**, 547-560, doi:10.1016/j.molcel.2015.10.040 (2015).

47 Pearl, L. H., Schierz, A. C., Ward, S. E., Al-Lazikani, B. & Pearl, F. M. Therapeutic opportunities within the DNA damage response. *Nat Rev Cancer* **15**, 166-180, doi:10.1038/nrc3891 (2015).

48 Gavande, N. S. *et al.* DNA repair targeted therapy: The past or future of cancer treatment? *Pharmacol Ther* **160**, 65-83, doi:10.1016/j.pharmthera.2016.02.003 (2016).

49 Bullman, S. *et al.* Analysis of Fusobacterium persistence and antibiotic response in colorectal cancer. *Science* **358**, 1443-1448, doi:10.1126/science.aal5240 (2017).

50 Castellarin, M. *et al.* Fusobacterium nucleatum infection is prevalent in human colorectal carcinoma. *Genome Res* **22**, 299-306, doi:10.1101/gr.126516.111 (2012).

51 Kostic, A. D. *et al.* Genomic analysis identifies association of Fusobacterium with colorectal carcinoma. *Genome Res* **22**, 292-298, doi:10.1101/gr.126573.111 (2012).

52 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760, doi:10.1093/bioinformatics/btp324 (2009).

53 McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-1303, doi:10.1101/gr.107524.110 (2010).

54 Forbes, S. A. *et al.* The Catalogue of Somatic Mutations in Cancer (COSMIC). *Curr Protoc Hum Genet* **Chapter 10**, Unit 10 11, doi:10.1002/0471142905.hg1011s57 (2008).

55 Smigielski, E. M., Sirotkin, K., Ward, M. & Sherry, S. T. dbSNP: a database of single nucleotide polymorphisms. *Nucleic Acids Res* **28**, 352-355 (2000).

56 Blokzijl, F., Janssen, R., van Boxtel, R. & Cuppen, E. MutationalPatterns: comprehensive genome-wide analysis of mutational processes. *Genome Med* **10**, 33, doi:10.1186/s13073-018-0539-0 (2018).

57 Favero, F. *et al.* Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. *Ann Oncol* **26**, 64-70, doi:10.1093/annonc/mdu479 (2015).

Appendix 2-Non-specific amplification of human DNA is a major challenge for 16S rRNA gene sequence analysis.

The following chapter has been accepted for publication in the journal Scientific Reports.

Authors:

Sidney P. Walker *, Maurice Barrett *, Glenn Hogan, Yensi Flores Bueso, Marcus J. Claesson, Mark Tangney

*Joint first authorship: These authors contributed equally to this work.

Citation:

WALKER, S. P., BARRETT, M., HOGAN, G., BUESO, Y. F., CLAESSON, M. J. & TANGNEY, M. 2020. Non-specific amplification of human DNA is a major challenge for 16S rRNA gene sequence analysis. *Scientific reports,* 10**,** 1-7.

8.1 Abstract

The targeted sequencing of the 16S rRNA gene is one of the most frequently employed techniques in the field of microbial ecology, with the bacterial communities of a wide variety of niches in the human body have been characterised in this way. This is performed by targeting one or more hypervariable (V) regions within the 16S rRNA gene in order to produce an amplicon suitable in size for next generation sequencing. To date, all technical research has focused on the ability of different V regions to accurately resolve the composition of bacterial communities.

We present here an underreported artefact associated with 16S rRNA gene sequencing, namely the off-target amplification of human DNA. By analysing 16S rRNA gene sequencing data from a selection of human sites we highlighted samples susceptible to this off-target amplification when using the popular primer pair targeting the V3-V4 region of the gene. The most severely affected sample type identified (breast tumour samples) were then re-analysed using the V1-V2 primer set, showing considerable reduction in off target amplification.

Our data indicate that human biopsy samples should preferably be amplified using primers targeting the V1-V2 region. It is shown here that these primers result in on average 80% less human genome aligning reads, allowing for more statistically significant analysis of the bacterial communities residing in these samples.

8.2 Introduction

This communication highlights off-target amplification of human DNA in 16S rRNA gene sequencing, detailing the circumstances necessary for this to occur, and the effects on ensuing research. Such artefacts are not a universal problem, and only occur in samples containing an overwhelming ratio of human to bacterial DNA. This leaves stool samples and skin samples which contain less than 10% and 90% human DNA respectively, unaffected, but can critically impact on analysis of human biopsy samples, where over 97% of the DNA present is of human origin [1]. Given the increased use of human biopsies from a number of body sites in microbiome research [2-5], this communication serves as a timely and, to our knowledge, unique methodological warning and remedy, particularly as only one mention of this issue can currently be found in the literature [6].

Currently, comparisons of primer pairs and the hypervariable regions they target in the 16S rRNA gene have focused exclusively on differing levels of taxonomic resolution and specificity [7,8]. The degree to which bacterial resolution is lost to the production human-derived amplicons has, so far, received no attention. This is because workflows for the analysis of 16S rRNA gene sequencing data typically remove reads falling too far from the mean or median sequence length, or if they are not classified taxonomically as originating from bacterial DNA. This is effective in ensuring that the presence of amplified human DNA does not have any impact on downstream analysis. Unaddressed is the fact that in a sequencing experiment yielding a finite amount of data (~13.5 Gb on a typical Miseq run [9]), a significant proportion of these can be wasted due to this off target amplification. This affects sequencing studies in two ways:

- Prospectively: If this loss of data is anticipated, fewer samples can be sequenced on a given sequencing run, adding to the expense which is already prohibitive for smaller labs.
- Retrospectively: If this loss if data is not anticipated, insufficient bacterial reads may be yielded to accurately characterise the samples being sequenced, particularly if attempting to identify the prevalence of rare taxa between different treatment groups.

Here, we show that the most commonly-used primer set for 16S rRNA sequencing, targeting the V3-V4 hypervariable regions, is particularly susceptible to this off-target amplification, while another commonly used primer set, targeting the V1-V2 primer region, shows almost no off-target amplification, as outlined in Figure 1 below. While this off-target amplification does not appear to affect research using stool or skin swab samples, we would urge all groups carrying out metataxonomic analysis of low microbial biomass human biopsy samples using high throughput sequencing to use the V1-V2 primer set in future.

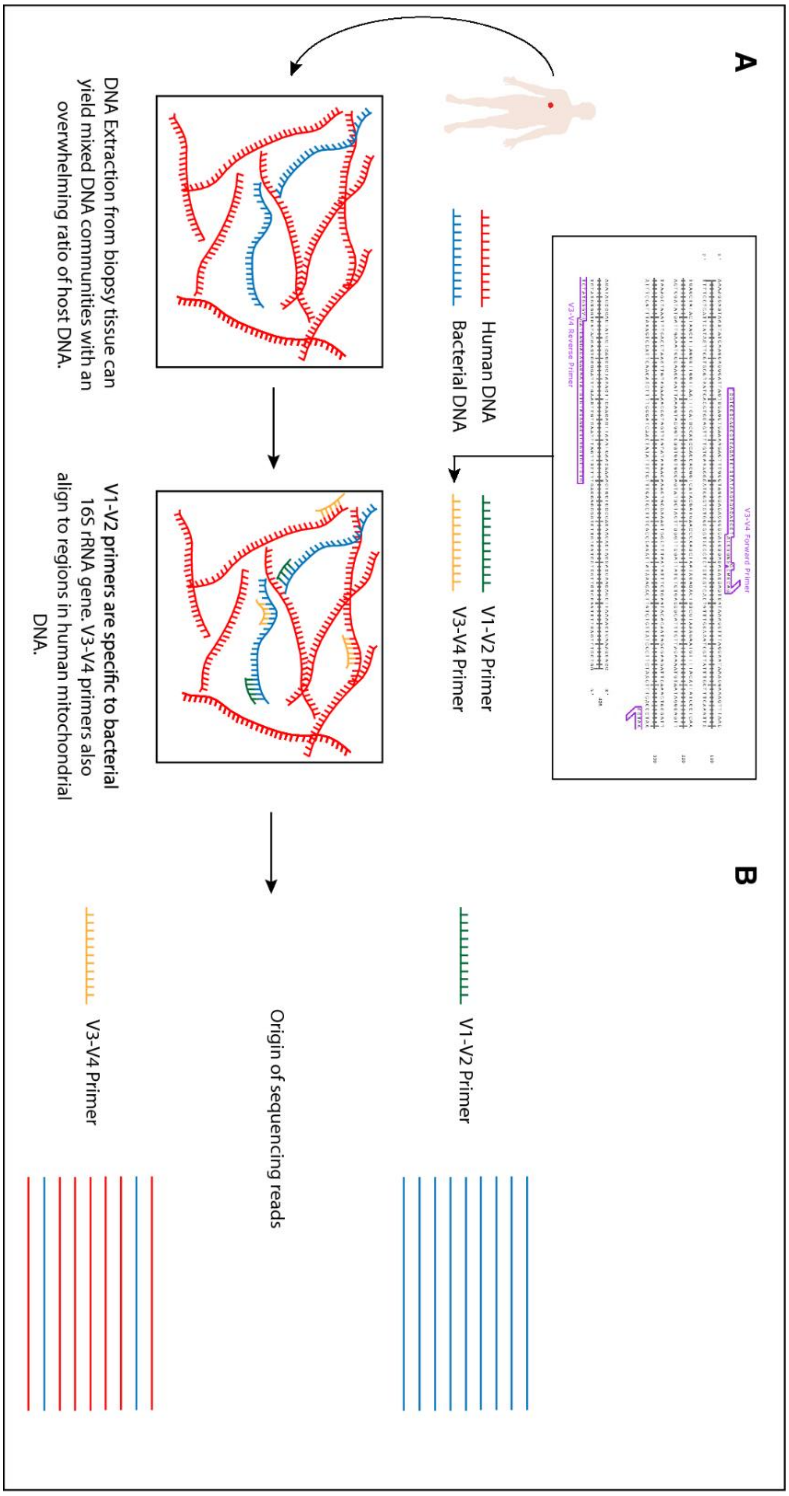

A
Human DNA
Bacterial DNA
V1-V2 Primer
V3-V4 Primer
V3-V4 Forward Primer
V3-V4 Reverse Primer
DNA Extraction from biopsy tissue can yield mixed DNA communities with an overwhelming ratio of host DNA.
V1-V2 primers are specific to bacterial 16S rRNA gene. V3-V4 primers also align to regions in human mitochondrial DNA.
B
V1-V2 Primer
Origin of sequencing reads
V3-V4 Primer

Figure 1| Proposed mechanism for off target amplification of mammalian DNA by V3–V4 primers, as opposed to V1–V2. (A) DNA extracted from human biopsies is known to contain large proportions of human DNA. In these circumstances V3–V4 degenerate primers, which also align to region in human mitochondrial DNA as shown can bind and amplify human DNA. There is no such alignment for V1–V2 degenerate primers. (B) Off target amplification significantly alters the 16S rRNA gene sequencing profile of a sample.

7.3 Materials/Methods

7.3.1 Sample Collection

Breast tissue was collected from women undergoing breast surgery at Cork University Hospital, Cork, Ireland. Breast tumour core-biopsies were aseptically resected using an Achieve 14G Breast Biopsy System (Iskus Health, UT, USA). The specimens were transported in sterile PBS to the lab, where they were flash-frozen and kept at -80^{o}C until further processing. DNA from the specimens was purified following the protocol and reagents provided in the Ultra Deep Microbiome Prep (Molzym, GmbH & Co. KG., Bremen, Germany) and eluted in 100 µl of Tris-HCl.

7.3.2 DNA Purification

Samples were processed and DNA purified following the procedures specified in protocols listed in Table 1. In all cases, DNA was eluted in Tris-HCl buffer and stored at -20^{o}C until further analysis.

Sample	**DNA extraction strategy**
Breast: Tumour and Normal	Molzym Ultradeep Microbiome (Molzym, Bremen, Germany)
Oesophageal biopsies	AllPrep DNA/RNA Mini Kit (Qiagen, Hilden, Germany) with modifications [10].
Skin Swab samples	QIAamp UCP Pathogen Mini Kit (Qiagen, Hilden, Germany)

Stool samples	Repeated bead beating method as previously described, with modifications[11,12]

Table 1. Samples and corresponding DNA extraction strategy.

7.3.3 16S rRNA gene sequencing Library Preparation.

Genomic DNA was amplified by PCR with primers targeting the hypervariable V1-V2 region or the V3–V4 region of the 16S rRNA gene. Table 2 details the primers sequences (underlined) included for compatibility with the Illumina 16S Metagenomic Sequencing Protocol (Illumina, CA, USA).

Region	**Name**	**F/R**	**Sequence**
V1 – V2 [13,14]	S-D-Bact-0027-b-S-20	F	5′-TCG TCG GCA GCG TCA GAT GTG TAT AAG AGA CAG AGM GTT YGA TYM TGG CTC AG
	S-D-Bact-0338-a-A-18	R	5’-GTC TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA G GCT GCC TCC CGT AGG AGT
V3 – V4 [15]	S-D-Bact-0341-b-S-17	F	5′ TCG TCG GCA GCG TCA GAT GTG TAT AAG AGA CAG CCT ACG GGN GGC WGC AG

	S-D-Bact-0785-a-A-21	R	5′ GTC TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA G GAC TAC HVG GGT ATC TAA TCC

Table 2. Primers used for 16S rRNA gene sequencing analysis.

For Breast Tumour and Normal Adjacent samples, amplification was performed in 50 µl reactions, containing 1X NEBNext High Fidelity 2X PCR Master Mix (NEB, USA), 0.5 µM of each primer, 8 µl template (5-15 ng/µl) and 12 µl nuclease free water. The thermal profile included an initial 98 ^{o}C x 30 sec denaturation, followed by 25 cycles of denaturation at 98 ^{o}C x 10 sec, annealing at 55 ^{o}C x 30 sec for V3-V4 or 62^{o}C x 30 sec for V1-V2 and extension at 72 ^{o}C x 30 sec. Plus a final extension at 72 ^{o}C x 5 min. Amplification was confirmed by running 5 µl of PCR product on a 2 % agarose gel, by visualisation of a ≈310 bp band for V1-V2 and ≈460 bp band for V3-V4

Faecal microbial genomic DNA was amplified using Phusion High-Fidelity DNA Polymerases (Thermo Scientific, Massachusetts, USA) with the PCR thermocycler protocol as follows: Initiation step of 98 °C for 3 min followed by 25 cycles of 98 °C for 30 s, 55 °C for 60 s, and 72 °C for 20 s, and a final extension step of 72 °C for 5 min.

Oesophageal biopsies and skin swab samples microbial genomic DNA was amplified using MTP Taq DNA Polymerase (Merck KGaA, Darmstadt, Germany) with the PCR thermocycler protocol as follows: Initiation step of 94°C for 1 min followed by 35 cycles of 94°C for 60 s, 55 °C for 45 s, and 72 °C for 30 s, and a final extension step of 72 °C for 5 min.

An index PCR was performed to add sample specific DNA barcodes to sample amplicons in accordance with the Illumina 16S Metagenomic Sequencing Protocol (Illumina, California, USA)[16]. Libraries DNA concertation was quantified using a Qubit fluorometer (Invitrogen) using the 'High Sensitivity' assay and samples were pooled at a standardised concentration[16]. The pooled library was sequenced on the Illumina MiSeq platform (Illumina, California, USA) utilising 2×300 bp chemistry.

7.3.4 16S rRNA sequence analysis

The quality of the paired-end sequencing data was visualised using FastQC v(0.11.9), and trimmed using Trimmomatic v(0.39) ensuring a minimum average quality of 25. Reads were then imported into R environment v(3.6.3)[17] to be resolved into Amplicon Sequence Variants by the DADA2 package v(1.12).

7.3.5 Contamination Control

In all samples a contamination control strategy was implemented in keeping with the RIDE checklist as proposed by Eisenhofer et al[18], incorporating aseptic techniques and a variety of negative controls from different stages of the sample-to-sequence data process. Retrospective contamination assessment and removal based on sequencing data from negative controls was also performed following published guidelines[19].

7.3.6 Retrospective Bioinformatics based removal of human amplicons

Sequencing reads aligning to the human genome (*GRCh38*) within the fasta file generated by DADA2 were identified using bowtie2[20]. To confirm reads mapped to the human genome were not erroneously aligned bacterial reads, all human aligning reads were classified with Mothur[21], using the RDP database v(11.4) as a reference.

7.3.7 Statistical analysis and data visualisation

All statistical analysis was carried out in the R environment, using the following libraries: Phyloseq v(1.30), Vegan v(2.5.6), ggplot2 v(3.3.0), reshape2 v(1.4.3).

7.4 Results and Discussion

All three sampled biopsy sites where an overwhelming ratio of host DNA was expected (breast, breast tumour and oesophageal) showed significant off target amplification of human DNA when amplified using the V3-V4 primer set (Figure 2).

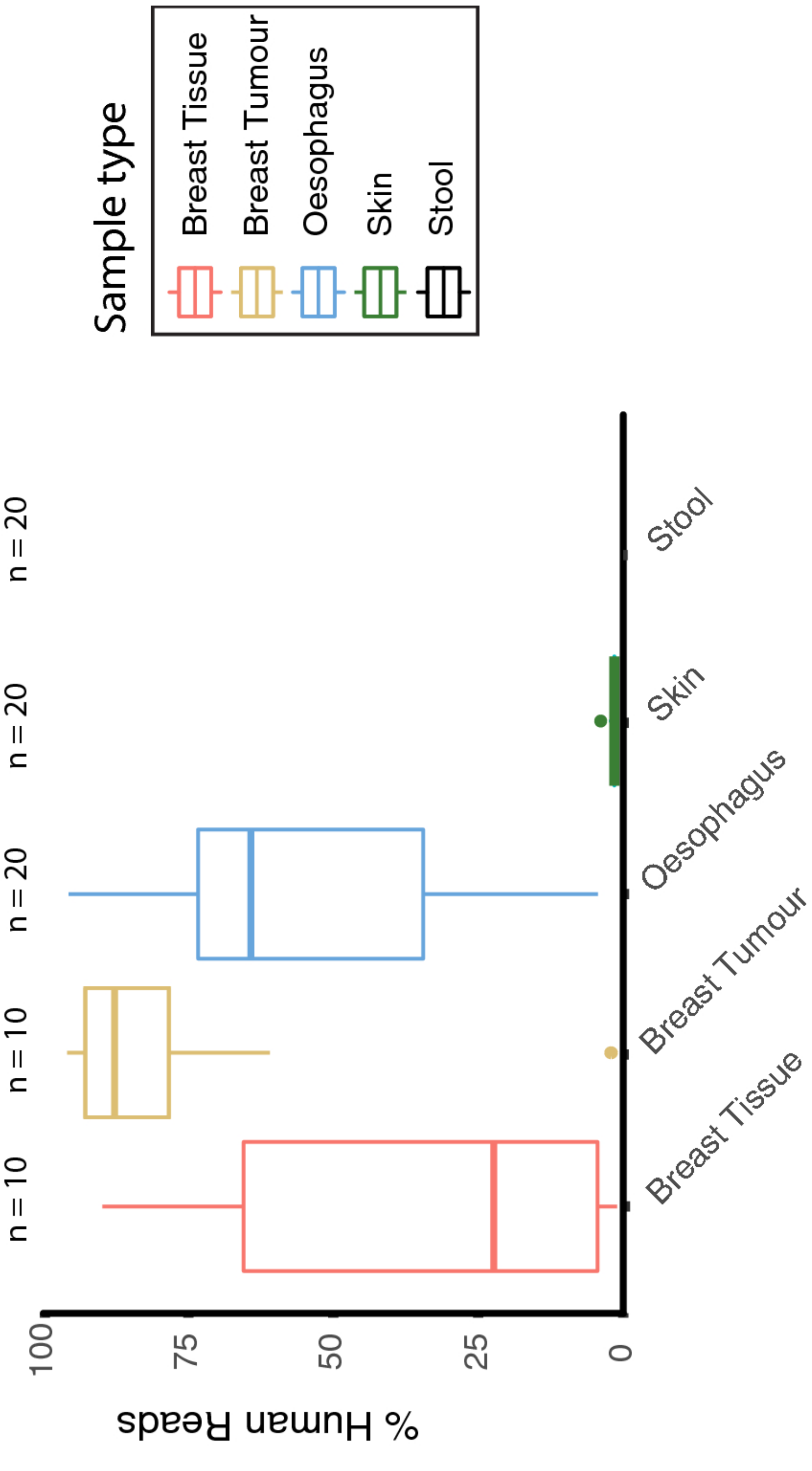


Figure 2| The scale of the problem of off-target amplification. % of sequencing reads produced by Miseq 2 × 300 bp sequencing of amplicons produced by primers targeting the V3–V4 regions shown to align to the human genome.

This was not seen when sequencing samples with lower levels of human DNA, such as skin swabs and stool samples. An average of 34.1% of all Amplicon Sequence Variants (ASV) detected in normal breast tissue samples were shown to align to the human genome GRCh38 using bowtie2.This included the most prevalent ASV, which was identified further using BLAST as *Homo sapiens haplogroup H8 mitochondrion, complete genome* (Accession no. MN986463.1) with an E-value of 7e-138 and 100% identity. In the breast tumour samples, 77.2% of all ASV's detected aligned to the human genome, with the most prevalent ASV again being identified as *Homo sapiens haplogroup H8 mitochondrion, complete genome* (Accession no. MN986463.1) with an E-value of 7e-138 and 100% identity. This situation was identical in Oesophageal biopsies, with a 55.6% of ASVs aligning to the human genome (*Homo sapiens haplogroup H8 mitochondrion, complete genome* (Accession no. MN986463.1) with an E-value of 7e-138 and 100% identity) . The skin swab samples showed a much lower level of amplification of human DNA, but these reads aligned to chromosomal DNA, most frequently *Homo sapiens chromosome 17, clone RP11-646F1, complete sequence* and were present in very low levels.

While human contamination is a very common problem in amplification-free shotgun metagenomic sequencing strategies [22], it is under reported as an issue for 16S rRNA gene sequencing, due to the use of bacteria/archaea specific primers. However, degenerate primers are routinely used for 16S rRNA sequencing [23]. This increases coverage, in terms of the number of 16S rRNA sequences matched by at least one primer, but also allows for off target amplification of non-bacterial DNA.

Figure 1A shows that the V3-V4 primers align to a region within the human mitochondrial DNA. We show here that when the ratio of host:bacterial DNA is overwhelming, human mitochondrial DNA can be amplified by primers targeting the 16S rRNA gene region. To ensure the validity of the results, reads identified as aligning to the human genome using Bowtie2 were classified using the Mothur [21] classifier trained on the RDP database. In all cases the reads identified as aligning to the human genome could not be classified when screened against the RDP database as shown in Table 3 below.

Sample	**% reads unclassified at Kingdom Level**	**% reads unclassified at Phylum level**
Oesophageal samples	99.5373235	0.4626765
Normal adjacent samples	98.867576	1.132424
Tumour samples	98.710027	1.289973
Skin samples	99.8588468	0.1411532

Table 3. Summary of Mothur output when classifying reads identified as aligning to the human genome by Bowtie2.

The most heavily affected sample type in our study (breast tumour tissue) was reanalysed by performing a pairwise comparison of samples amplified with the V3-V4 and V1-V2 primer sets (Figure 3).

Looking initially at the rarefaction curves produced by the sequencing data corresponding to the previously mentioned paired V1-V2 and V3-V4 primer pair amplified breast tumour sample there is a clear difference between the two groups. This is done by plotting new species against number of reads per sample. Figure 3A below shows that the distribution of samples in this 2D plane appears to be stochastic prior to the removal of human reads. Figure 3B, following removal of human reads, shows clearly that samples amplified with the V1-V2 primer pair consistently yield more observable species, a greater number of reads per sample, and a plateauing of the rarefaction curve which suggests sufficient sampling depth is available for accurate characterisation.

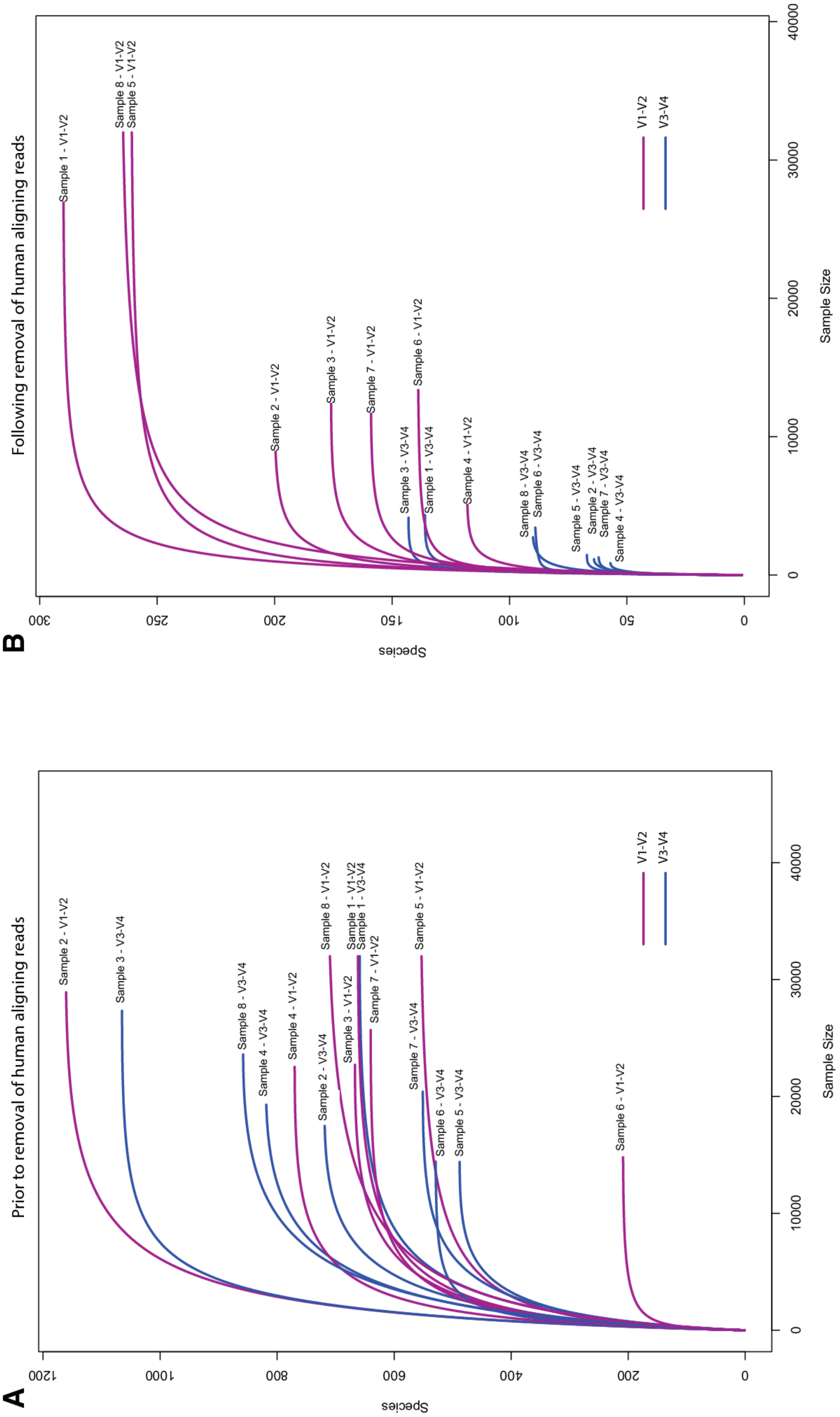


Figure 3| Rarefaction curve generated by plotting observed species vs read depth on a per sample basis. (A) Rarefaction curve prior to removal of human genome aligning reads. (B) Rarefaction curve following removal of human genome aligning reads.

The community structure in samples amplified with V1-V2 primers was visually similar to those amplified with V3-V4 primers(Figure 4A) and no bacterial family was found to be significantly elevated using one primer set over the other as per Wilcoxon signed-rank test, once p-values had been corrected for multiple testing using the FDR method (Supplementary table 1). There was also no significant difference in terms of Shannon diversity (Figure 4B), indicating choice of primers did not have any adverse effect on the downstream results. Of considerable interest to any groups carrying out low biomass research in the future, is the huge discrepancy in the number of reads yielded once human contamination had been filtered out. As can be seen in Figure 4C, samples amplified with primers targeting the V1-V2 region have a consistently and significantly higher number of ASVs per sample following the removal of ASV’s aligning to the human genome.

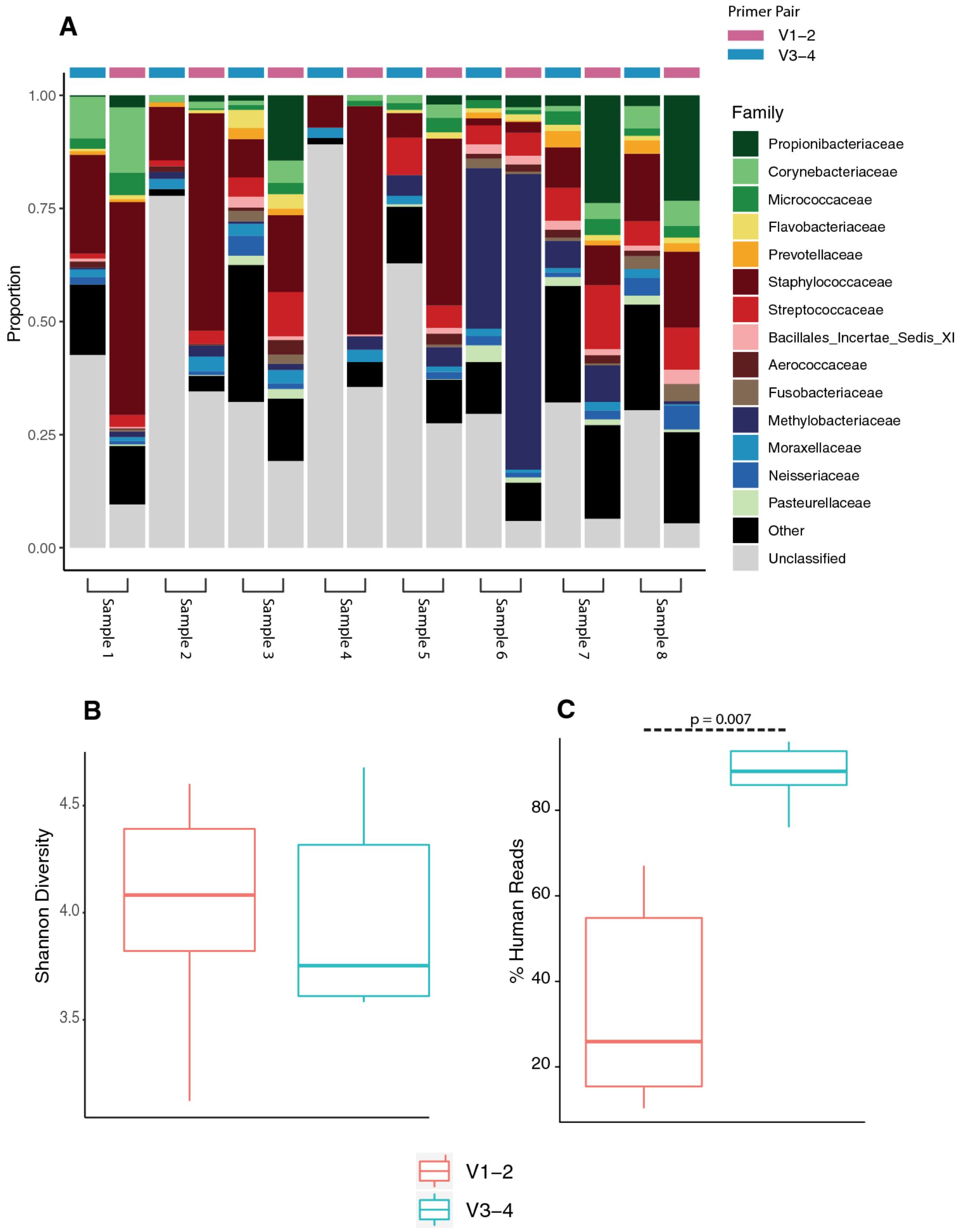


Figure 4| Pairwise comparison of matched samples using primers targeting the V1–V2 and V3–V4 regions of the 16S rRNA gene fragment. (A) Sample composition at the family level of paired samples. (B) Average Shannon Diversity comparison between samples amplified using V1–V2 primers (red) and V3–V4 primers (blue). (C) Percentage of total sequencing reads aligning to human genome. In both (B) and (C) statistical testing is performed using Wilcoxon signed-rank test.

7.5 Future Perspectives

Third generation sequencing technologies, such as those produced by Oxford Nanopore Technologies and Pacific BioSiences are now being utilised in 16S rRNA gene sequencing experiments. The Pacific BioSciences SMRT platform has seen the greatest promise in this regard with the implementation of "Circular Consensus Sequencing" in conjunction with denoising algorithms, allowing for the production of long reads of high quality[24]. Earl et al showed that this new method using degenerate primers targeting the entire 16S rRNA gene, still resulted in off target amplification of the human genome[25]. This study also noted that this off target amplification was related to the ratio of human to bacterial DNA. The human genome must be considered when designing or choosing primers now and in the future.

7.6 Acknowledgements

The authors acknowledge the contribution of Prof. Paul O'Toole, Microbiology Department, UCC, in providing access to sequencing data used in this study.

7.7 Declarations

The authors declare no competing interests. All procedures in this study were performed in accordance to national ethical guidelines, following ethical approval from the University College Cork Clinical Research Committee. Patients provided written informed consent for sample collection and subsequent analyses.

7.8 Reference

1 Pereira-Marques, J. *et al.* Impact of Host DNA and Sequencing Depth on the Taxonomic Resolution of Whole Metagenome Sequencing for Microbiome Analysis. *Frontiers in microbiology* **10**, doi:10.3389/fmicb.2019.01277 (2019).

2 Deshpande, N. P., Riordan, S. M., Castaño-Rodríguez, N., Wilkins, M. R. & Kaakoush, N. O. Signatures within the esophageal microbiome are associated with host genetics, age, and disease. *Microbiome* **6**, 227, doi:10.1186/s40168-018-0611-4 (2018).

3 Riquelme, E. *et al.* Tumor Microbiome Diversity and Composition Influence Pancreatic Cancer Outcomes. *Cell* **178**, 795-806.e712, doi:10.1016/j.cell.2019.07.008 (2019).

4 Grice, E. A. & Segre, J. A. The skin microbiome. *Nat Rev Microbiol* **9**, 244-253, doi:10.1038/nrmicro2537 (2011).

5 Urbaniak, C. *et al.* The Microbiota of Breast Tissue and Its Association with Breast Cancer. *Applied and Environmental Microbiology* **82**, 5039, doi:10.1128/AEM.01235-16 (2016).

6 Davis, N. M., Proctor, D. M., Holmes, S. P., Relman, D. A. & Callahan, B. J. Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. *Microbiome* **6**, 226, doi:10.1186/s40168-018-0605-2 (2018).

7 Pinna, N. K., Dutta, A., Monzoorul Haque, M. & Mande, S. S. Can Targeting Non-Contiguous V-Regions With Paired-End Sequencing Improve 16S rRNA-Based Taxonomic Resolution of Microbiomes?: An In Silico Evaluation. *Frontiers in Genetics* **10**, doi:10.3389/fgene.2019.00653 (2019).

8 Johnson, J. S. *et al.* Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. *Nature Communications* **10**, 5029, doi:10.1038/s41467-019-13036-1 (2019).

9 Illumina. *Specifications for the Miseq system*, <https://www.illumina.com/systems/sequencing-platforms/miseq/specifications.html> (

10 Flemer, B. *et al.* Tumour-associated and non-tumour-associated microbiota in colorectal cancer. *Gut* **66**, 633-643, doi:10.1136/gutjnl-2015-309595 (2017).

11 Yu, Z. & Morrison, M. Improved extraction of PCR-quality community DNA from digesta and fecal samples. *Biotechniques* **36**, 808-812, doi:10.2144/04365st04 (2004).

12 Costea, P. I. *et al.* Towards standards for human fecal sample processing in metagenomic studies. *Nat Biotechnol* **35**, 1069-1076, doi:10.1038/nbt.3960 (2017).

13 Browne, H. P. *et al.* Culturing of ‘unculturable’ human microbiota reveals novel taxa and extensive sporulation. *Nature* **533**, 543-546, doi:10.1038/nature17645 (2016).

14 Elliott, D. R. F., Walker, A. W., O'Donovan, M., Parkhill, J. & Fitzgerald, R. C. A non-endoscopic device to sample the oesophageal microbiota: a case-control study. *The Lancet Gastroenterology & Hepatology* **2**, 32-42, doi:10.1016/S2468-1253(16)30086-3 (2017).

15 Klindworth, A. *et al.* Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic acids research* **41**, e1-e1, doi:10.1093/nar/gks808 (2013).

16 Illumina. *Amplicon, P. C. R., Clean-Up, P. C. R. & Index, P. C. R. 16S Metagenomic Sequencing Library Preparation*, <https://www.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/16s/16s-metagenomic-library-prep-guide-15044223-b.pdf (2013).> (

17 R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing, Vienna, 2019).

18 Eisenhofer, R. *et al.* Contamination in Low Microbial Biomass Microbiome Studies: Issues and Recommendations. *Trends in Microbiology* **27**, 105-117, doi:https://doi.org/10.1016/j.tim.2018.11.003 (2019).

19 Walker, S. P., Tangney, M. & Claesson, M. J. Sequence-Based Characterization of Intratumoral Bacteria—A Guide to Best Practice. *Frontiers in Oncology* **10**, doi:10.3389/fonc.2020.00179 (2020).

20 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-359, doi:10.1038/nmeth.1923 (2012).

21 Schloss, P. D. *et al.* Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and environmental microbiology* **75**, 7537-7541, doi:10.1128/AEM.01541-09 (2009).

22 Marotz, C. A. *et al.* Improving saliva shotgun metagenomics by chemical host DNA depletion. *Microbiome* **6**, 42, doi:10.1186/s40168-018-0426-3 (2018).

23 Sambo, F. *et al.* Optimizing PCR primers targeting the bacterial 16S ribosomal RNA gene. *BMC Bioinformatics* **19**, 343-343, doi:10.1186/s12859-018-2360-6 (2018).

24 Callahan, B. J. *et al.* High-throughput amplicon sequencing of the full-length 16S rRNA gene with single-nucleotide resolution. *Nucleic Acids Research* **47**, e103-e103, doi:10.1093/nar/gkz569 (2019).

25 Earl, J. P. *et al.* Species-level bacterial community profiling of the healthy sinonasal microbiome using Pacific Biosciences sequencing of full-length 16S rRNA genes. *Microbiome* **6**, 190, doi:10.1186/s40168-018-0569-2 (2018).

Acknowledgements

After approximately 6 years and ~70,000 words one may not be so enthused to keep on writing. Unfortunately, I am told gratitude might require a modicum of effort and many people deserve much more than a modicum (whatever that is). Firstly, I would like to thank my supervisor Prof. Paul O'Toole for the opportunity to taken on a PhD. I would further like to thank him for his guidance during the PhD including his effort to polish my skill sets and rid me of my not so beneficial attributes. I would also like to give thanks to my co-supervisor Fergus Shanahan and advisor Collette Hand for their guidance and input throughout my PhD. I would like to thank my lab colleagues over the number of years. Many of you were not only a source of invaluable expertise but also friendship. I would like to give thanks to the institutions and people who form the environment I worked in, namely, the school of microbiology and APC Microbiome Ireland. I found this environment to be friendly, engaging and well suited for the growth of a scientific researcher.

I would have to mention the particular friends I made during my PhD including Adam Clooney, David Mullins, Feargal Ryan, Mrinmoy Das, Ross Holohan, Sidney Walker and Tom Sutton. I would not be able to complete my PhD without you not only due to your friendship but also the technical help.

I would like to thank my family. I am grateful to an immeasurable level for the love and support my parents, Kieran and Catherine, have given during my life. I would like to thank my siblings Siobhán, Keara, Tomás and Eoin, for their encouragement and help. Finally, and foremost, I would like to thank my dog, Patsey, for his eternal guardianship of reality from the encroachment of oblivion, without which, we would be all doomed to nothingness.